



OpenAlex in focus: Metadata quality of publication type and language fields in an open peer review corpus

Güleda Doğan and Ayça Nur Sezen

DOI: <https://doi.org/10.47989/ir31iConf64207>

Abstract

Introduction. OpenAlex is widely used as a free bibliographic database for bibliometric and scholarly communication research. Despite its openness and coverage, its metadata contains inconsistencies that require systematic cleaning. This study examines metadata quality in OpenAlex, focusing on publication type and language.

Method. Publications on open peer review were retrieved from OpenAlex. After filtering and deduplication, 6,640 records were manually checked. Document type and language fields were cross-verified with publisher sources, while Crossref publication types were also collected for comparison.

Analysis. Manual classification was harmonised across categories to ensure comparability. The main focus was to evaluate the agreement between OpenAlex and manual classifications of type and language, and to assess the consistency of Crossref publication types with both.

Results. Of 6,640 records, 2,878 (43%) showed publication type discrepancies, with ‘Article’ most often misused. Crossref aligned more closely with OpenAlex in broad categories but diverged from manual verification. Additionally, 222 records (3.3%) had language mismatches, often English labels wrongly assigned to non-English works.

Conclusion. OpenAlex is a valuable infrastructure, yet its metadata for publication type and language shows notable inconsistencies. Researchers should apply systematic cleaning and validation before using OpenAlex or similar databases.

Introduction

Open access, open science, and open data have emerged as central pillars of contemporary scholarly communication, aiming to democratise knowledge, enhance transparency, and accelerate innovation by reducing barriers to access and participation. The rise of these ‘open’ movements reflects both technological developments and a growing recognition of the societal value of accessible and reproducible research (Gelfand & Lin, 2020; Giannini & Molino, 2020). At the same time, their implementation raises policy, ethical, and infrastructural challenges that vary across regions and disciplines, with persistent inequities observed in data-sharing practices (Kar & Rath, 2025).

Within this context, the sustainability and effectiveness of open movements depend heavily on robust infrastructures capable of operationalising openness at scale. OpenAlex, launched in 2022 as the successor to the Microsoft Academic Graph (MAG), represents one of the most prominent examples of such infrastructures. By structuring the research ecosystem around works, authors, venues, institutions, and concepts, OpenAlex provides comprehensive metadata linking scholarly outputs to their broader contexts (Priem et al., 2022). Its metadata are generated through the aggregation and standardisation of information from multiple authoritative scholarly infrastructures, using open linking, identifier-based matching, and continuous updating mechanisms. These sources include Microsoft Academic Graph, Crossref, identifier registries such as ORCID and ROR, open-access and indexing platforms (e.g., DOAJ, Unpaywall, PubMed, PubMed Central), registry services (ISSN International Centre), institutional and subject repositories (e.g., arXiv, Zenodo), web crawls, and digitised collections from the Internet Archive (About the Data, n.d.; About Us, n.d.).

Compared to commercial databases such as Web of Science (WoS) and Scopus, OpenAlex offers broad coverage, free accessibility, and openness, and has already been integrated into widely used evaluative systems such as the CWTS Leiden Ranking (Zhang et al., 2024). These characteristics have made it an increasingly popular data source for bibliometric and open science monitoring (Hval et al., 2023). At the same time, a growing body of research has documented metadata limitations across OpenAlex and other large bibliographic databases, including missing institutional affiliations (Zhang et al., 2024), inconsistencies in open access status (Jahn et al., 2023), misclassification of retracted publications (Hauschke & Nazarovets, 2025; Ortega & Delgado-Quirós, 2024), and inaccuracies in document type and language assignments (Alperin et al., 2024; Thelwall & Jiang, 2025). While OpenAlex often provides broader inclusivity than WoS or Scopus (particularly for preprints and data journals) this inclusivity may come at the cost of metadata reliability (Jiao et al., 2023; Simard et al., 2024).

Taken together, these findings highlight both the opportunities and challenges of using OpenAlex for bibliometric research and underscore the need for systematic validation prior to analysis. Beyond conceptual scoping, review scope is commonly delimited using study characteristics such as document type and language, which play a central role in filtering, eligibility criteria, and comparability across bibliometric analyses and systematic reviews (Gusenbauer & Gauster, 2025). Accordingly, the present study examines metadata quality in OpenAlex with a focus on document type and language. These fields were selected not because they exhaust metadata quality concerns, but because they constitute cross-cutting elements that directly shape inclusion and exclusion decisions and downstream analytical validity. The dataset was initially retrieved from OpenAlex to study open peer review; however, metadata inconsistencies identified during data preparation prompted a secondary, focused analysis. While this necessarily constrains the scope and generalisability of the findings, it enables a concrete and empirically grounded examination of specific metadata issues in OpenAlex.

Methodology

Data collection

The data were collected from OpenAlex on 26 November 2024 using a full-text search for the term 'open peer review.' To ensure query stability, the search was repeated using the variant 'open peer-review,' yielding identical results. The OpenAlex interface returned 9,315 records, which were exported in CSV format, resulting in a dataset of 9,275 items. Because OpenAlex indexes records that reference 'open peer review' in unrelated contexts, a manual data cleaning process was required prior to analysis.

Initial filtering by document type

As a first step, the dataset was filtered by document type to exclude records not relevant to our study. In total, 2,184 records were removed, including 2,013 items labelled as 'peer review,' 10 libguides, 37 paratexts, 35 datasets, 4 errata, 56 reference entries, and 29 items classified as 'other.' After this filtering process, the dataset was reduced from 9,275 to 7,090 records. The remaining set included only the publication types considered in scope for the study: article, book, book chapter, discussion, dissertation, editorial, letter, preprint, proceeding paper/article, report, review, thesis, and proceeding book.

Deduplication

After initial filtering, the remaining records were screened for duplicates using DOI identifiers. In this step, duplicate entries arising from multiple versions of the same work (most commonly preprints and their subsequently published versions) were identified and removed, with the published version retained where available. Additional duplicate records and test entries without accessible full text were also excluded. Following deduplication, the final dataset consisted of 6,640 unique publications.

Quality control workflow

The remaining dataset underwent a manual verification workflow to ensure accuracy of metadata regarding publication type and language. The publication language and type were verified to determine whether they differed from the values indicated in OpenAlex metadata. For non-English publications, the accuracy of the publication type was checked by reviewing an English translation of the relevant webpage and/or article text using Google automatic translation or ChatGPT. In addition, ChatGPT was used to identify non-English languages when necessary. In determining the publication type, we generally relied on the type specified on the publisher's webpage. Except in cases where the type was not provided at all or was clearly incorrect, we did not make additional determinations based on the publication's full content.

Document type classification

In OpenAlex, the publications in our dataset were originally assigned to the following document types: article, book, book chapter, dissertation, editorial, letter, preprint, report, and review. During our verification process, however, we encountered cases where the assigned type did not fully match the information provided on the publisher's website. To address these discrepancies, we developed a refined classification scheme that considered the corrected publication types. This expanded scheme included: article, blog post, book, book chapter, book review, editorial, peer review, proceeding paper, opinion/commentary, report, review, thesis/dissertation, and other. The categories are presented below in descending order of frequency in the dataset and interpreted as follows:

Other: A residual category for heterogeneous items that could not be confidently mapped to a standard scholarly type. It encompassed proceeding abstracts, poster presentations, news items, data notes, magazines, popular articles, interviews, and other marginal or non-research formats.

Article: Represented conventional research outputs. It included research articles, case reports, software tool articles, method articles, research notes, and less frequent subtypes such as original articles, case studies, and meta-research articles.

Opinion/commentary: Covered non-research contributions such as perspectives, opinion articles, commentaries, viewpoints, correspondences, comments, and letters to the editor. It also included less frequent formats like open letters, forums, essays, and discussion pieces.

Review: Comprised scholarly reviews of the literature, including systematic reviews, narrative reviews, minireviews, and overview reviews.

Editorial: Encompassed journal editorials, forewords, and introductory notes, often difficult to separate in OpenAlex metadata from commentary-type contributions.

Peer review: Contained referee reports published in open peer review settings.

Report: This category encompassed a wide variety of academic and professional documents distinct from standard research articles. It included study protocols, brief reports, conference reports, meeting reports, annual meeting reports, recommendations, white papers, research reports, technical reports, working papers, policy briefs, survey reports, and guides.

Proceeding paper: This category included only full conference papers. Shorter formats such as extended abstracts or poster summaries were not classified under this category but were instead included in the Other category.

Book review: Reviews of scholarly books and related media, which OpenAlex generally assigned to Article.

Blog post: Blog-style contributions published on scholarly or publisher platforms. OpenAlex frequently subsumed these under Article.

Preprint: Manuscripts deposited on preprint servers, sometimes duplicated with published versions.

Book chapter: Individual chapters in edited scholarly volumes, occasionally misclassified as ‘*proceeding papers*’ in OpenAlex.

Thesis/Dissertation: Graduate theses and dissertations, rarely indexed in OpenAlex and inconsistently labelled.

Book: Books were straightforward to identify but were underrepresented in the dataset.

This refined classification allowed us to distinguish research articles from other types of scholarly output that OpenAlex frequently grouped together. By correcting misclassifications and aiming to represent the actual publication types as accurately as possible, we ensured a clearer and more reliable picture of document type diversity in the dataset.

OpenAlex employs a broad work-type classification in which most research outputs (including journal articles, conference papers, and posted content) are grouped under the general type ‘*article*,’ while records primarily hosted on preprint servers or explicitly identified as such may be labelled as ‘*preprint*.’ In this study, we did not adopt a hierarchical view in which preprints are treated as a subset of research articles. Instead, we distinguished publication states analytically, treating preprints and peer-reviewed journal articles as conceptually separate categories. Accordingly, within our framework, a classification was considered correct when a record labelled as ‘*Preprint*’ corresponded to content disseminated prior to formal journal publication (e.g., submitted versions), whereas records linked to peer-reviewed journal venues were treated as articles, regardless of their underlying research genre (OpenAlex Technical Documentation, 2025).

Results

After the filtering and deduplication steps described in the methodology, the final dataset consisted of 6,640 publications. Within this corpus, 2,878 records (43%) had a publication type that differed from what was indicated in OpenAlex, and 222 records (3.3%) had a language discrepancy.

Figure 1 illustrates the differences between the Assigned Type (OpenAlex) and the Verified Type (Manual Check) after our manual validation. Among the 2,878 records with type misclassifications, 1,719 (60%) were assigned as 'Article' in OpenAlex. Within this group, 732 (43%) were reclassified as Other, 312 (18%) as Opinion/Commentary, and 247 (14%) as Editorial.

A considerable number of misclassifications were also observed in works labelled as 'Preprint' in OpenAlex. Out of the 795 preprints identified among the misclassified records (28% of all type errors), 405 (51%) were verified as Article, and 147 (18.5%) as Review.

For items indexed as 'Review' in OpenAlex, 298 records were included in the misclassified group (10% of all type errors). Manual verification revealed that 197 (66%) of these were in fact Peer Review reports.

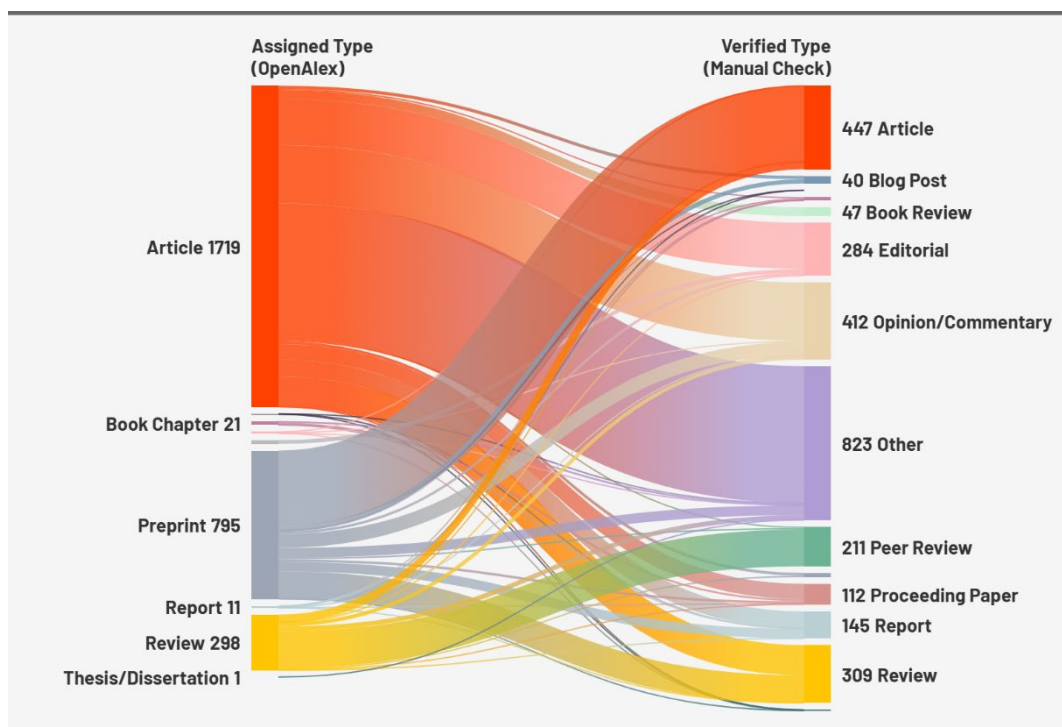


Figure 1. Reclassification of publication types: assigned in OpenAlex vs. verified manually.

For interactive version and data: <https://public.flourish.studio/visualisation/25116016/>

The CSV dataset downloaded from OpenAlex also included Crossref-derived document type information. To enable a meaningful comparison across sources, we examined the alignment between Crossref types, OpenAlex-assigned types, and our manually verified types. In this process, Crossref items labelled as journal-article were mapped to the Article category, while those labelled as posted-content were mapped to the Preprint category. This normalisation was necessary to enable a consistent comparison of document types across the three datasets.

Table 1 summarises the results for the five most frequent Crossref categories. Publication types with lower frequencies were not included in the table.

Crossref type	Total	Matches with OpenAlex (n / %)	Matches with Manual (n / %)
Article	2,368	1,649 (69.6%)	364 (15.4%)
Peer Review	196	0 (0.0%)	196 (100.0%)
Preprint	193	181 (93.8%)	8 (4.1%)
Proceeding Paper	76	0 (0.0%)	52 (68.4%)
Book Chapter	22	20 (90.9%)	1 (4.5%)

Table 1. Comparison of Crossref types with OpenAlex-assigned and manually verified types.

These figures show that Crossref types overlapped most strongly with OpenAlex in the broad categories of Article and Preprint. By contrast, for Peer Review all of the overlap was observed with the manually verified types, while in OpenAlex these works had been labelled as ‘Article’ rather than Peer Review. Proceeding Papers showed a similar pattern, with no matches to OpenAlex but a substantial share (68.4%) matching the manually verified categories. For Book Chapters, Crossref was closely aligned with OpenAlex, with only a small number of discrepancies against the manual verification.

In addition to publication type inconsistencies, we examined discrepancies in language metadata. Among the 6,640 records analysed, 222 publications (3.3%) showed a mismatch between the language assigned in OpenAlex and our manual verification. While this proportion is relatively small and falls within an expected margin of error for large-scale bibliographic databases, it nonetheless highlights residual classification noise that may be relevant for studies relying on language-based filtering.

verified as another language. In addition, records with unspecified language in OpenAlex were often identifiable through manual checks, suggesting that language coverage is incomplete and inconsistently recorded.

Our findings are consistent with and extend earlier research on OpenAlex metadata quality. Céspedes et al. (2025), for example, reported that nearly 15% of records in OpenAlex had misclassified language information, substantially higher than the 3.3% observed in our dataset. This difference likely reflects both variations in corpus selection and methodology. In terms of publication type, Haupka et al. (2025) highlighted that OpenAlex overuses the Article category, grouping heterogeneous content under a single label. Our analysis substantiates this claim, demonstrating that many items labelled as Article in OpenAlex were in fact Editorials, Opinion/Commentaries, or Other categories. Moreover, comparative analyses show that OpenAlex's inclusivity, which covers more data journals and preprints than WoS or Scopus (Jiao et al., 2023; Simard et al., 2024), often comes at the cost of metadata accuracy, a trade-off we also observed in both document type and language. Finally, in line with Alperin et al. (2024) and Thelwall & Jiang (2025), our results confirm that while OpenAlex's openness and inclusivity are valuable, its metadata inconsistencies significantly limit its direct usability without systematic cleaning and validation.

The additional comparison with Crossref provided further perspective but was not the primary aim of this study. Crossref showed closer alignment with OpenAlex in broad categories such as Article and Preprint, yet it also diverged sharply from manual verification. In categories like Peer Review and Proceedings, Crossref was consistent with manual checks but not with OpenAlex, highlighting how database-specific approaches to classification contribute to observed discrepancies.

Limitations and implications

This study is limited to two metadata fields (publication type and language) and does not assess other elements such as affiliations, references, or open access status. The dataset was originally constructed as part of an open peer review corpus, which may limit the generalisability of the findings beyond this subject area. In addition, manual classification relied on publisher-provided metadata, which may itself contain inaccuracies. Nevertheless, the systematic examination of a large dataset provides empirical evidence on the nature and scale of metadata errors in OpenAlex. The findings indicate that OpenAlex should not be used 'as is' for bibliometric analyses without data cleaning and validation, and they point to the need for database providers to refine classification systems, reduce the overuse of default categories such as Article, and improve language metadata completeness. While publication type inconsistencies are not unexpected in large-scale bibliographic infrastructures, they are practically relevant because they affect document filtering, eligibility decisions, and the interpretation of results in applied bibliometric and systematic review contexts.

Following metadata quality updates announced by OpenAlex in October 2025, the findings reflect the November 2024 data snapshot and may not capture subsequent improvements (Jason, 2025).

About the authors

Güleda Doğan is an Associate Professor in the Department of Information Management at Hacettepe University, Ankara, Türkiye. She received her Ph.D. in information science from Hacettepe University. Her research interests include bibliometrics, scholarly communication, research evaluation, open science, and the quality and use of large-scale bibliographic databases. She can be contacted at gduzyol@hacettepe.edu.tr

Ayça Nur Sezen is a PhD candidate in the Department of Information Management at Hacettepe University, Ankara, Türkiye. She has professional experience in the scholarly publishing domain and is currently working at The Scientific and Technological Research Council of Türkiye (TÜBİTAK), National Academic Network and Information Center (ULAKBİM). Her research interests include scholarly publishing, research infrastructures, bibliographic metadata, and open science. She can be contacted at aycanursezen@hacettepe.edu.tr

References

- About the data. (n.d.). OpenAlex. Retrieved January 7, 2026, from <https://help.openalex.org/hc/en-us/articles/24397285563671-About-the-data>
- About us. (n.d.). OpenAlex. Retrieved January 7, 2026, from <https://help.openalex.org/hc/en-us/articles/24396686889751-About-us>
- Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses (No. arXiv:2404.17663). arXiv. <https://doi.org/10.48550/arXiv.2404.17663>
- Céspedes, L., Kozłowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimi, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, 76(6), 884–895. <https://doi.org/10.1002/asi.24979>
- Gelfand, J. M., & Lin, A. (2020). How Open Science Influences Next Developments in Grey Literature. *Grey Journal (TGJ)*, 16(1), 34–48.
- Giannini, S., & Molino, A. (2020). Open Access—A Never-Ending Transition? *Grey Journal (TGJ)*, 16(1), 6–26.
- Gusenbauer, M., & Gauster, S. P. (2025). How to search for literature in systematic reviews and meta-analyses: A comprehensive step-by-step guide. *Technological Forecasting and Social Change*, 212, 123833. <https://doi.org/10.1016/j.techfore.2024.123833>
- Hauptka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2025). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar (No. arXiv:2406.15154). arXiv. <https://doi.org/10.48550/arXiv.2406.15154>
- Hauschke, C., & Nazarovets, S. (2025). (Non-)retracted academic papers in OpenAlex. *Journal of Information Science*. <https://doi.org/10.1177/01655515251322478>
- Hval, G., Harboe, I., Johansen, M., Larsen, M., & Næss, G. (2023). Evaluation of OpenAlex. Folkehelseinstituttet. <https://www.fhi.no/en/publ/2023/Evaluation-of-OpenAlex/>
- Jahn, N., Hauptka, N., & Hobert, A. (2023). Scholarly Communication Analytics: Analysing and reclassifying open access information in OpenAlex. https://subugoe.github.io/scholcomm_analytics/posts/oalex_oa_status/
- Jason. (2025, October 1). OpenAlex rewrite enters beta! OpenAlex Blog. <https://blog.openalex.org/openalex-rewrite-enters-beta-%f0%9f%8e%89/>
- Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of the Web of Science, Scopus, Dimensions, and OpenAlex (No. arXiv:2307.09704). arXiv. <https://doi.org/10.48550/arXiv.2307.09704>

- Kar, S., & Rath, D. S. (2025). Open Data in Social Sciences: Growth, Impact, and Equity in Data Paper Publishing | DESIDOC Journal of Library & Information Technology. DESIDOC Journal of Library & Information Technology, 45(4), 350–366.
- 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, 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 (No. arXiv:2205.01833). arXiv.
<https://doi.org/10.48550/arXiv.2205.01833>
- Simard, M.-A., Basson, I., Hare, M., Lariviere, V., & Mongeon, P. (2024). The open access coverage of OpenAlex, Scopus and Web of Science (No. arXiv:2404.01985). arXiv.
<https://doi.org/10.48550/arXiv.2404.01985>
- Thelwall, M., & Jiang, X. (2025). Is OpenAlex suitable for research quality evaluation and which citation indicator is best? <https://doi.org/10.1002/asi.70020>
- Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*, 129(10), 5869–5891.
<https://doi.org/10.1007/s11192-023-04923-y>
- OpenAlex Technical Documentation. (2025, December 14). Work object.
<https://docs.openalex.org/api-entities/works/work-object>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).