Investigating metadata discrepancies between OpenAlex and the Web of Science: the case of Library and Information Sciences

2025-08-03

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, a bibliographic database launched in 2022, has distinguished itself for its openness and extensive coverage. While OpenAlex may reduce or eliminate barriers to accessing bibliometric data, one of the concerns that hinders its broader adoption for research and research evaluation is the quality of its metadata. This study aims to assess metadata quality in OpenAlex and WoS, focusing on document type, publication year, language, and number of authors. By addressing discrepancies and misattributions in metadata, this research seeks to enhance awareness of data quality issues that could impact bibliometric research and evaluation outcomes.

# Introduction

Bibliometrics has been used in research and data-driven research evaluation for approximately half a century (Narin, 1976), originally supported by the development of the Web of Science (WoS), which was conceptualized in 1955 and launched in 1964 (Clarivate, n.d.). It took more than 30 years for other players to emerge: Elsevier’s Scopus was founded in 1996, Crossref in 1999, Google Scholar in 2004, Microsoft Academic (now integrated into OpenAlex) and Dimensions in 2018, and OpenAlex in 2022[[1]](#footnote-20). Myriad other databases exist to support specific bibliometric work or topical analysis; PubMed, for example, primarily indexes life sciences and biomedical works and The Lens provides a view into innovation through its coverage of patent knowledge[[2]](#footnote-27).

The bibliometrics field has long paid attention to the coverage and quality of these data sources and investigated differences in evaluative outcomes produced from them[[3]](#footnote-30). Bruce & Hillmann (2004) note that the library community’s efforts to define quality in bibliographic records are linked, in part, to their aim to enforce it. The advent of OpenAlex- an open 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 and standardizer of data from Microsoft Academic Graph (MAG), Crossref, Open Researcher and Contributor ID (ORCID), Research Organization Registry (ROR), Directory of Open Access Journals (DOAJ), Unpaywall, PubMed, and several other sources[[4]](#footnote-31), its coverage and quality have been of interest regarding its efficacy for quantitative analysis. This scrutiny has resulted in observations that OpenAlex might contain erroneous metadata, or that it is still largely incomplete. For example, changing processes for matching and disambiguating data have demonstrated the importance of persistent identifiers and remaining cautious when creating new open data sources dependent on OpenAlex (Mongeon et al., 2024).

## Coverage of OpenAlex versus Web of Science

As a nascent data source, numerous recent studies have compared the coverage of OpenAlex to more established databases. 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 & Eck (2024), though they observed OpenAlex’s coverage of publication and author information was high compared to WoS and Scopus. Simard et al. (2025) 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[[5]](#footnote-33). In a coverage comparison of six databases, Ortega & Delgado-Quirós (2024) found that OpenAlex indexes more retracted works than WoS, Scopus, and PubMed.

## Metadata quality of OpenAlex versus Web of Science

Donner (2017) found document type (articles, reviews, letters, and others) discrepancies between WoS and journal websites or article full text for a sample of 791 publications, with letters and reviews particularly affected. Different average page and reference counts for errors were also observed for arising discrepancies, hinting at the potential development of semi-automatic methods using these findings.

Other studies focused on the quality of OpenAlex data 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, potentially explained by Ortega & Delgado-Quirós (2024) as resulting from the database’s reliance on Crossref’s less granular system of classification.

## Research objectives

The rise of open science has promoted the creation, sustainability, and use of open infrastructure; the completeness and accuracy of such sources are compared frequently to proprietary, curated databases which tout more exclusive standards for indexing, but make their data curation processes less transparent. However, in describing how the Citation Index Project (an early version of what is now the Web of Science) could be used for evaluation exercises, Eugene Garfield emphasized how such information could be used “as a basis for comparative studies on the efficacy of various indexing techniques” (Garfield & Sher, 1963, p. 195). Indeed, studies of this kind thus serve as “vital statistics” which allude to major implications of bibliometric analyses conducted using different data sources. For example, Liu et al. (2021) found major discrepancies between publication records for the same journal indexed in WoS and Scopus, respectively, forcing users into the dilemma of Segal’s law[[6]](#footnote-37). Thus, this study aims to compare discrepancies between OpenAlex, a massive open bibliographic database increasingly utilized by the bibliometrics community since its launch, and WoS, an entrenched source of data regarded for its high-quality curation and standardization and recently found to be the most reliable in terms of data completeness and accuracy (Singh et al., 2024).

While OpenAlex’s coverage has been under recent intense investigation to evaluate its usability for accurate analyses, more needs to be understood about the comparability of its metadata to other data sources. Our work thus aims to assess the quality of metadata in OpenAlex and WoS. More specifically, it addresses the following research questions:

* **RQ1**: How frequent are discrepancies between WoS and OpenAlex records in terms of document type, language, publication year, and number of authors?
* **RQ3**: What share of records with discrepancies are correct in WoS, OpenAlex, neither, or both??
* **RQ4**: What explains these discrepancies?

Scientific literature engages with several concepts in various combinations in relation to data sources, including *quality*, *accuracy*, *validity*, *completeness*, *reliability*, and *consistency*. Bruce and Hillman (2004) describe the most recognized characteristics of data quality as completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility[[7]](#footnote-38). In this study, we largely refer to quality broadly, largely as accuracy and completeness, and we consider the relationship of both of these concepts in relation to their consistency. Thus, the examination of discrepancies reveals not only if and which errors are occurring, but also in what frequency. Findings regarding consistency may implicate the reliability of the data, potentially proving especially useful to those running large-scale analyses. In aggregate, this study’s findings may also point to possible reasons for such errors which may have useful implication for the validation of this data and avoidance of future error through the updating of database processes, considering that these errors are difficult to identify (Meester et al., 2016).

# Data and 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). We collected all WoS records with a DOI published between 2021 and 2023 (N = 7,661,474). We removed 30,3094 (0.4%) WoS records with multiple document types to avoid complications with the analysis. Of the remaining 7,631,080 WoS records, 6,599,479 (86.5%) had a DOI match in the February 2024 snapshot of OpenAlex accessed through Google Big Query (see Mazoni & Costas, 2024). We used the subset of 6,594,747 Web of Science records with a single DOI match in OpenAlex for our analysis. In the OST database, every journal is assigned to one of 143 specialties and 14 disciplines of the National Science Foundation (NSF) classification. For our analysis, we grouped the disciplines into four groups: Arts and Humanities (AH), Biomedical Research (BM), Natural Science and Engineering (NSE), and Social Sciences (SS).

## Identification of discrepancies

For each matching WoS and OpenAlex record, we compared the following four metadata elements: 1) document type, 2) language, 3) publication year, and 4) number of authors.

For the document type, we did not consider discrepancies where a record is a review according to WoS and an article according to OpenAlex, and vice versa (i.e., we consider articles and reviews as the same document type). 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 considered discrepancies for which the record is identified as an article or a review in either WoS or OpenAlex, but is identified as neither in the other source. We also excluded discrepancies for which the record is a meeting abstract in WoS and an article in OpenAlex. Overall, we found 427,593 discrepancies that met these criteria (6.5% of all records in the dataset).

The identification of discrepancies in language, publication year, and number of authors was limited to the subset of 5,924,459 articles and reviews with no discrepancy in document type (i.e., the records that are articles or reviews in both WoS and OpenAlex). The publication language is recorded in its long form (e.g., English) in our WoS data and its ISO code (e.g., EN) in OpenAlex, so we created a list of all language combinations (e.g., English-EN, English-FR), determined which ones constitute discrepancies, and identified 33,322 discrepancies (0.6% of all articles and reviews) in publication language between WoS and OpenAlex. The publication year and number of authors being numeric indicators, identifying discrepancies in these metadata fields is easily done by calculating the difference between the two values and flagging all non-zero values. We identified a total of 469,098 discrepancies (7.9% of all articles and reviews) in publication years and 70,590 discrepancies (1.2% of all articles and reviews) in the number of authors. We recorded them both as dichotomous variables (discrepancy vs no discrepancy) and as differences between the two values, which allowed us to measure both the frequency and the strength of the discrepancies.

## Investigation of discrepancies

### Discrepancies Sample

For the publication year and number of authors, we also stratified using six bins based on the difference between the values in WoS and OpenAlex (-4 or less, -2 to -3, -1, 1, 2 to 3, 4 or more) to ensure that a range of cases are represented in the data.

### Investigation Process

The sample discrepancies were manually investigated by looking at the article on the journal’s website and the full text when necessary and available. We recorded whether the WoS or OpenAlex record was correct and, when possible, explained the discrepancy. For the publication year, it is typical for articles to have two publication dates: the date of the first online publication and the date of the publication of the issue. For cases in which the online publication year was recorded in one database and the issue year was recorded in the other, we did not consider either database to be correct.

For author discrepancies, the landing page and, when available, the published version was examined for author counts. Where available, author declarations, acknowledgements, and CRediT declarations were also considered to clarify authorship where groups or consortia were involved in the production of the work. For languages, the full text was the primary source, followed by the publisher’s landing page for recording observations on language assignment. We recorded multiple language full texts if available, as well as translated abstracts available on the landing page. There are limitations to using the publisher landing page with a web browser.  The language encoding in the head of the HTML could be used by the GET request to display your preferred language type, if the server is configured to respond to such a request.

This process allowed us to go beyond the simple counting of differences between the databases and gain insight into the different factors that can cause discrepancies and estimate the percentage of erroneous records in each database. One caveat, of course, is that because our process relies on discrepancies in identifying errors, we are not considering cases where both databases contain the same error.

# Results

We structured our results section by type of discrepancy. For each, we first show the distribution of values in the Web of Science and the OpenAlex datasets. Note that these are not the distributions in the entire Web of Science OpenAlex databases, but the distributions in the subset of records with a DOI match between the databases that are included in our analysis. Second, we present descriptive statistics related to the discrepancies identified and the sample that was used for investigations. Finally, we present the results of these investigations.

## Document type

Table 2 and Table 3 present the distribution of records across document types in WoS and OpenAlex, respectively, to provide a general picture of the databases’ content and 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). Most documents in both data sources are articles and reviews.

| oa\_type | n | pct |
| --- | --- | --- |
| article | 5,831,736 | 88.43 |
| review | 511,688 | 7.76 |
| letter | 136,052 | 2.06 |
| editorial | 59,645 | 0.90 |
| erratum | 48,177 | 0.73 |
| retraction | 4,705 | 0.07 |
| book-chapter | 1,332 | 0.02 |
| preprint | 872 | 0.01 |
| paratext | 301 | 0.00 |
| book | 95 | 0.00 |
| other | 94 | 0.00 |
| total | 6,594,697 | 100.00 |

Tables 4 and 5 show that the vast majority 301,850 of the 310,843 discrepancies (97%) are cases where a record is an article or review in OpenAlex but not WoS. We find that in both tables the majority of discrepancies indicate a misclassification of a record as an article or review. However, WoS is much more accurate at detecting misclassification in OpenAlex (93.5% of the sample) than in WoS (68.8% of the sample). This generally points to a much larger number of OpenAlex records misclassified as articles or review compared to WoS.

| Type in OpenAlex | Discrepancies | | Sample | | Errors | |
| --- | --- | --- | --- | --- | --- | --- |
| n | pct | n | pct | n | pct |
| letter | 5,007 | 55.7 | 331 | 56.7 | 219 | 66.2 |
| editorial | 1,341 | 14.9 | 87 | 14.9 | 63 | 72.4 |
| book-chapter | 1,303 | 14.5 | 85 | 14.6 | 85 | 100.0 |
| preprint | 840 | 9.3 | 54 | 9.2 | 14 | 25.9 |
| paratext | 237 | 2.6 | 14 | 2.4 | 14 | 100.0 |
| erratum | 149 | 1.7 | 8 | 1.4 | 6 | 75.0 |
| other | 83 | 0.9 | 4 | 0.7 | 0 | 0.0 |
| retraction | 28 | 0.3 | 1 | 0.2 | 1 | 100.0 |
| book | 5 | 0.1 | 0 | 0.0 | 0 | 0.0 |
| total | 8,993 | 100.0 | 584 | 100.0 | 402 | 68.8 |

| Type in WoS | Discrepancies | | Sample | | Errors | |
| --- | --- | --- | --- | --- | --- | --- |
| n | pct | n | pct | n | pct |
| editorial material | 160,789 | 53.3 | 350 | 54.3 | 308 | 88.0 |
| book review | 67,165 | 22.3 | 146 | 22.6 | 146 | 100.0 |
| letter | 30,234 | 10.0 | 65 | 10.1 | 65 | 100.0 |
| correction | 19,206 | 6.4 | 40 | 6.2 | 40 | 100.0 |
| news item | 11,327 | 3.8 | 23 | 3.6 | 23 | 100.0 |
| biographical-item | 5,906 | 2.0 | 11 | 1.7 | 11 | 100.0 |
| retraction | 3,711 | 1.2 | 7 | 1.1 | 7 | 100.0 |
| other | 3,512 | 1.2 | 3 | 0.5 | 3 | 100.0 |
| total | 301,850 | 100.0 | 645 | 100.0 | 603 | 93.5 |

## Publication Language

### Distribution of Records by Language in WoS and OpenAlex

Table 5 and Table 6 show the distribution of articles and reviews across languages in WoS and OpenAlex, respectively, limited to the top 10 most frequent languages. Unsurprisingly, the quasi totality of records are in English in both databases. It is important to note, once again, that our results are note meant to reflect the compositon of the entire databases, but only the subset of documents included in both, which expectedly underestimate the representation of non-English literature especially on the OpenAlex side. Also important to reiterate here is that we our WoS data includes only the SCI-E, SSCI, and A&HCI, and not the ESCI which would include more non-English records.

| Language in WoS | Records | |
| --- | --- | --- |
| n | pct |
| English | 5,884,089 | 99.32 |
| German | 14,440 | 0.24 |
| Spanish | 7,822 | 0.13 |
| French | 6,069 | 0.10 |
| Chinese | 3,898 | 0.07 |
| Russian | 1,960 | 0.03 |
| Portuguese | 1,509 | 0.03 |
| Polish | 870 | 0.01 |
| Japanese | 772 | 0.01 |
| Turkish | 580 | 0.01 |
| Other | 2,450 | 0.04 |
| Total | 5,924,459 | 100.00 |

Aside from the dominance of English in the dataset, we also observe differences in the languages that compose the top 10 in each database, as well as in their ranking. Spanish, French, and Portugese are much more prevalent in the OpenAlex records than in the WoS. On the other hand, Chinese is the 5th most prevalent language in the WoS data, but does not make the top 10 in OpenAlex. Since these are supposed to be the same set of records, these differences point to significant discrepancies in the language between the two databases.

| Language in OpenAlex | Records | |
| --- | --- | --- |
| n | pct |
| English | 5,875,530 | 99.17 |
| Spanish | 13,629 | 0.23 |
| German | 13,201 | 0.22 |
| French | 10,324 | 0.17 |
| Portugese | 3,184 | 0.05 |
| Polish | 1,045 | 0.02 |
| Turkish | 909 | 0.02 |
| Hungarian | 841 | 0.01 |
| Croatian | 573 | 0.01 |
| Romanian | 537 | 0.01 |
| Other | 3,615 | 0.06 |
| Unknown | 1,071 | 0.02 |
| Total | 5,924,459 | 100.00 |

### Language discrepancies between WoS and OpenAlex

For this part of the analysis, we focus on articles and review that are in English in Web of Science or in OpenAlex, and in a language other than English in the other database. In other words, we do not investigate discrepancies between Portuguese and Spanish or French and German, but only discrepancies for which one of the languages is English (e.g., English-German, English-Chinese).

Table 7 and Table 8 respectively display the number of discrepancies found for articles in English in WoS and articles in English in OpenAlex. They also display the number of records manually investigated to determine whether the discrepancies stem from mislabeling non-English content as English, as well as the result of that investigation. To understand how an incorrect assignment may have occurred, we looked at the abstracts and full text on the landing page as well as the full text when it was available. Most sampled works seem to have multi-language abstracts either on the landing page or in the PDF of the full text. Also, some articles are published in multiple language, meaning that for some of the discrepancies observed, both OpenAelx and WoS are correct.

A first observation is that WoS appears to be more accurate than OpenALex in terms of publication language. Table Only 20,086 English publications in the WoS were in another language according to OpenAlex, and in slightly less than half (48.2%) of the verified cases, the article was in fact in English.

| Language in OpenAlex | Discrepancies | | Sample | | Errors | |
| --- | --- | --- | --- | --- | --- | --- |
| n | pct | n | pct | n | pct |
| Spanish | 7,019 | 34.9 | 125 | 37.9 | 42 | 33.6 |
| French | 5,538 | 27.6 | 101 | 30.6 | 70 | 69.3 |
| Portugese | 2,110 | 10.5 | 35 | 10.6 | 9 | 25.7 |
| German | 1,566 | 7.8 | 26 | 7.9 | 19 | 73.1 |
| Romanian | 529 | 2.6 | 8 | 2.4 | 8 | 100.0 |
| Hungarian | 520 | 2.6 | 9 | 2.7 | 0 | 0.0 |
| Turkish | 464 | 2.3 | 6 | 1.8 | 1 | 16.7 |
| Croatian | 382 | 1.9 | 3 | 0.9 | 3 | 100.0 |
| Polish | 277 | 1.4 | 3 | 0.9 | 0 | 0.0 |
| Other | 1,681 | 8.4 | 14 | 4.2 | 7 | 50.0 |
| Total | 20,086 | 100.0 | 330 | 100.0 | 159 | 48.2 |

However, we notice that for some languages, such as French, German, Romanian and Croatian, the discrepancies correctly flagged articles mislabelled as English in WoS., although the sample sizes for Romanian and Croatian were quite small. On the other hand, the majority of articles labelled as English in WoS and as Spanish, Portuguese, or Turkish in OpenAlex were actually found to be in English.

## Publication year

### Distribution of publication years in WoS and OpenALex

Now turning our attention to discrepancies in publication years for articles in WoS and OpenAlex, we first present a statistical summary of the

| **Database** | **n** | **mean** | **sd** | **var** | **q1** | **median** | **q3** | **min** | **max** | **skew** | **kurtosis** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| OpenAlex | 5,924,459 | 2,021.904 | 0.952 | 0.906 | 2,021 | 2,022 | 2,023 | 1,915 | 2,025 | -7.449 | 538.571 |
| WoS | 5,924,459 | 2,021.974 | 0.812 | 0.660 | 2,021 | 2,022 | 2,023 | 2,021 | 2,023 | 0.048 | -1.482 |

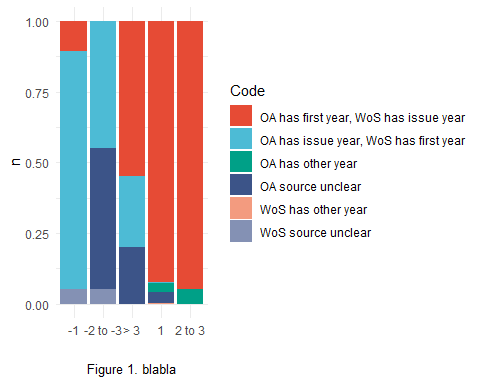
### Publication Year Discrepancies Between WoS and OpenAlex

We observe a high prevalence of publication year discrepancies in articles and reviews in the WoS and OpenAlex, with 470,107 cases identified (see Table 10). We grouped the cases bases on the the direction and extent of the discrepancies by substracting the OpenAlex publication year from the WoS publication year. A difference of 1 indicates that according to WoS the article was published one year later than accoridng to OpenAlex. Inversely, a difference of -1 indicates that the articles was published one year ealier accoring to WoS than according to OpenAlex.

| **Difference in publication year (WoS - OpenAlex)** | **Discrepancies** | | **Sample** | |
| --- | --- | --- | --- | --- |
| **n** | **pct** | **n** | **pct** |
| -2 to -3 | 783 | 0.2 | 20 | 5.3 |
| -1 | 50,632 | 10.8 | 38 | 10.2 |
| 1 | 393,548 | 83.7 | 276 | 73.8 |
| 2 to 3 | 23,222 | 4.9 | 20 | 5.3 |
| > 3 | 1,922 | 0.4 | 20 | 5.3 |
| Total | 470,107 | 100.0 | 374 | 100.0 |

### Explanation for publication year discrepancies

Publication year was divided between ‘first published year’ and ‘issue year’ as both dates do not have to be the same. In the majority of works, OpenAlex correctly identified the first publication year and WoS assigned the issue year (476 works), followed by the opposite condition where WoS used the first published year and OpenAlex used the issue year (155 works). The remainder of the table includes explanations that identify where sources were unclear, not included, do not match, or were incorrect, or neither date was available. 255 works did not have a publication date, and 11 did not have the issue date.



## Number of authors

Author discrepancies can be ambiguous to determine a ‘correct’ number. Authorships were manually counted on the landing page, and where available, the full text, which we considered the authoritative number, given that the authors would have reviewed and approved it. Additionally, Crossref metadata for authors was also considered as a possible explanation, particularly as OpenAlex draws heavily from this data source. The following table contains explanations determined by the investigators as to whether the OpenAlex or WoS count was correct, and if not, why the information might not be available. In rare cases, some articles contained over 1000 authors and were set aside for the time as being ‘too many to count’ manually.

In Table 6, the majority were found to have OpenAlex as the correct author reporting at 594 counts out of 1107 (53.7%), and WoS author counts were correct 368 (33.2%). Neither was found to be correct 91 times (8.2%). Bad URLs (meaning they are non-resolving, behind a paywall with no landing page, or were links to XML pages meant for similarity checking, among other possibilities) were found 2.3% of the time, followed by the lack of landing pages for works 1.4% of the time. Only 2 (0.2%) were found to be unregistered at the time, and only 1 (0.1%) was shown as ‘withdrawn’ by the publisher. In Table 7, author count explanations are examined regarding disciplines.

There are no large discrepancies biased towards one discipline or the other. Social Sciences had slightly more correct authors assigned by OpenAlex than the other disciplines, whereas BM had slightly more attributions by WoS as correct. BM also had more where neither was correct, followed by NSE, which is to be expected as both of these disciplines typically have more consortia, research groups, or teams attributed as authors, which can be tricky to determine if not explicitly declared by the authors. AH and NSE had more bad URLs, and AH contained the majority of those with no landing page for the work, making it beyond the scope of our methods. NSE had the majority of those over 1000 authors, which is not surprising given the large teams in fields such as Physics.

# Discussion

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 concerning document type. Our findings support those of past studies that have found metadata quality in OpenAlex to be in need of improvement lower in comparison to WoS. As Haupka et al. (2024) have indicated, document type classifications are critical for bibliometric research and evaluation, and so improved accuracy of document types is desired. We also echo that calibrating diverse document types across disciplines and different bibliographic databases remains a significant 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.

## Document-type discrepancies[[MH1]](#Xd8c3846978d47bdf5af6a8f09caf94b87f551a5)

## Publication year discrepancies[[MH2]](#X8399f62e9d7f114b86a09bf8b19c27794216230)

Pinning down a single published date value can possibly yield different dates that are valid. To gain a little more insight into the sources for both OpenAlex and WoS, we examined an API call[[1]](#Xc217c44f9d6f9d647df03178eb8b64fb8f54252) for all 10 dates available in the Crossref metadata (posted, issued, updated-to, approved, indexed, accepted, published, published-print, published-online, deposited – three of these have to do with the published date). As an example, the DOI 10.1093/isle/isaa156 has three different unique dates that align with the OA and WoS values. The published and the published-online dates are 2020-11-30, and this agrees with the OpenAlex published date. But the published-print is 2021-12-13, and this agrees with the WoS values. So, both are correct. Yet, this work’s metadata was not deposited in Crossref until 2024-08-16. A single date, then, is tricky to pin down and does not completely capture the journey of a work through its scholarly lifecycle.

Overall, we found that published date values were preferred by OpenAlex, and WoS tends to use the issue date (476 works). 155 works were found that WoS used the published date, and OA used the issue date. The total of these (631 works) accounts for 67% of the works sampled, and the remainder have an omission, incorrect value, or the source was unclear with the date provided by either WoS or OpenAlex. Further work is needed to understand how local practices within publishers or other limitations may be affecting how publishers provide this information and make it available to scholarly databases.

## Language discrepancies[[MH3]](#X244de5aaa2bcbc79f59e5cff9542183a9452a48) [[MH4]](#X736af059264e1ed797b38ce72679914516493ca) [[MH5]](#Xf199a2fc6aeb224f3820b0674358b97f71538ae)

Reducing the language down to a single value for analysis is challenging and comes with many caveats. We observed that multi-language abstracts are not uncommon, though it is unknown if this is increasing.

Overall, there is an inconsistency in the single language value that is limited by both the OpenAlex and WoS metadata schema as many versions had multiple languages in the landing page, abstracts, and in the full text. This makes declaring a single language assignment difficult or even misrepresenting.

## Author discrepancies[[MH6]](#X3b054053523ea588a8e79f82fb90674084f0f05) [[MH7]](#Xf93d60c3c4cf4ef48a65543a0c3fcbad738eaa4)

The author discrepancies subset was selected based on those in the dataset that had a difference between ‘oa\_n\_authors’ and ‘wos\_n\_authors’. As such, we find a few publishers that are over-represented as their local practices or limitations affect whether authors are entered into the metadata at all.

Overall, we see that OpenAlex was most correct on author counts on 53.7% of the subset of works, whereas WoS was correct on 33.2% of the subset. Regarding disciplines, the distribution is fairly even across disciplines for both OpenAlex and Wos, but BM and NSE are more affected by the ‘neither is correct’ condition, which is most likely affected by those authorships using consortia or other group identifiers. We also looked to see if there were similarities between OpenAlex and Crossref author counts and found that 50.3% of works contained the same authors, but alarmingly, we also found that 25.3% had author list errors, and 14.4% contained no authors at all. These omissions account for a large proportion of the works, and while both OpenAlex and WoS have other methods for extracting authors from landing pages and full text, publisher-supplied metadata should align with the published work.

The use of group terminology for authorships, such as consortia, research groups, study groups, teams, etc., seems to be common in the BM and NSE fields. There is great inconsistency in how these groups are accounted for, credit is given for authorship, how members are identified, or the language used to signify how authors work on behalf of or as part of a consortium. A few works in BM utilized the CRediT declaration and provided complete author lists, which clearly defined who was an author/contributor to the work. We also observed that group names may be listed as authors on the first page of the work or the landing page but were absent from the CRediT declaration. While the CRediT declaration made our job easier and established a clear authority by the authors on contributions, it is applied inconsistently. More work needs to be done to document these local practices to enable or encourage more consistency in author reporting and responsibilities.

We also observed consistently that OpenAlex excludes organization names as contributors, even if they exist in the Crossref metadata. Their process does not seem to affect singular names, such as those used in non-Western naming conventions, but it does seem to exclude consortia and other group identifiers in the author listing.

Conversely, WoS seems to be consistently overcounting author names, even if the Crossref metadata seems to align with the published version of the work. In contrast to the ‘lazy’ method of OpenAlex by excluding organization/group names, any pattern matching used by WoS seems to be greedy, accepting many more than there are listed in the work or author declarations. This seems to work, as most of the times that WoS was correct on the author count, the Crossref metadata was incorrect.

Translated works were an exception to the greedy approach by WoS, as their counts consistently excluded the number of translators on a work, mostly affecting works in the AH discipline. OpenAlex included both the author and translator on works, and this aligned with the Crossref metadata that also included translators as authors/contributors.

## Advantages and disadvantages in using each source[[MH8]](#X328b2bef0308b79a5bed9f015889eba939487f7)

Haunschild and Bornmann (2024) offer useful insights from their use of OpenAlex data to produce overlay maps, one being that the main advantage of using OpenAlex data is that it can be used without restrictions (i.e., free availability). These are offset by concurrent limitations, such as in partial correctness or erroneous metadata, in their case, in OpenAlex assigned “concepts”[[2]](#X424ae1dfb199fdaa161d17bdce77f94503db3d6) raising questions about the reliability and validity of OpenAlex’s processes (Haunschild & Bornmann, 2024).

# Conclusion[[MH9]](#Xa4e94a85b473b4b409b2a6bb45d6ffb197149f4) [[MH10]](#Xf9f1c43a5b694918a63def0a745f268d3c58013)

The idiosyncrasies of databases should be accounted for when researchers or research evaluators decide which to use for their data analysis tasks, as they affect the discoverability of works as well as metrics (Barbour et al., 2025). Differences among data sources regarding indexing coverage are, to a degree, largely understood and part of decision-making, Barbour et al. (2025) explain how the known selectivity of WoS and Scopus contrast with OpenAlex, which has a global and comprehensive aim. Metadata discrepancies, however, are more granular peculiarities and difficult to track.

The Paris Conference on Open Research Information and the Barcelona Declaration on Open Research Information are two initiatives that emphasize 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. This carries implications for OpenAlex, as they look to the research community for feedback on necessary improvements to their metadata, as well as for those conducting research and research evaluation using open sources, who must remain apprised of findings related to limitations of the data they use.

## Limitations

## Future work

Many open bibliographic data sources such as OpenAlex are downstream from metadata providers such as Crossref; as a crucial source for PubMed, OpenAIRE, OpenCitations, and, among other metadata digital infrastructure organizations, attention should be focused on the provenance of metadata, and their sustainability and reliability as a data stream for scholarly metadata (van Eck & Waltman, 2025). In noting that the quality of bibliographic metadata is quantifiable and measurable, Bruce and Hillman (2004) also note that enforcing standards must be done at the level of the community. Liu et al. (2021) make recommendations for remedying discrepancies: they recommend the use of official publication dates to avoid different publication years; this would also require the alignment of policies across databases. They also suggest databases work closely with journal publishers to mitigate document omission and metadata errors, double checking them frequently (Liu et al., 2021). Finally, they recommend the examination of data processing flows to identify potential causes when it comes to duplicate entries (Liu et al., 2021). To this list we also add…

# Conflicts of interest

The authors have no conflicts of interest to report.

# Author contributions

[Conceptualization](http://credit.niso.org/contributor-roles/conceptualization/) (PM), [Data curation](http://credit.niso.org/contributor-roles/data-curation/) (PM), [Formal analysis](http://credit.niso.org/contributor-roles/formal-analysis/) (PM, PR), Investigation (PM, MH, GK, RM, PR, RT, SW), [Project administration](http://credit.niso.org/contributor-roles/project-administration/) (PM), (PM), [Visualization](http://credit.niso.org/contributor-roles/visualization/) (RT, PR), [Writing – original draft](http://credit.niso.org/contributor-roles/writing-original-draft/) (PM, MH, PR), [Writing – review & editing](http://credit.niso.org/contributor-roles/writing-review-editing/) (PM, MH, GK, RM, PR, RT, SW)

# Data availability

# References

# Notes

[[1]](#Xf4c8113f07524640efcbf4d5822e1fc633a9e60) <https://api.crossref.org/works/?select=DOI,posted,issued,update-to,approved,indexed,accepted,published,published-print,published-online,deposited&filter=doi:10.1093/isle/isaa156> This API call selects all 10 date elements available in the Crossref REST API.

[[2]](#X6ae359e89ea24432127ac5e463063b1e05d509f) OpenAlex Concepts have been deprecated in favor of Topics: <https://docs.openalex.org/api-entities/concepts>

[[MH1]](#X8d324fc7bb1d799b5d36acc94a659546b34a12f)[**https://link.springer.com/article/10.1007/s11192-017-2483-y#Sec6**](https://link.springer.com/article/10.1007/s11192-017-2483-y#Sec6)

**VERY useful for doc types.**

[[MH2]](#Xbda551afffb27f0f4d17f8fe448104c2da4c05a)<https://arxiv.org/abs/1505.00796>

[[MH3]](#X897e5763d5ede7cd4b0155df172d51a1ddcf296)From Shi et al. (2025): The lack of standardized and widely adopted Romanization schemes for many languages

itself results in errors and inconsistencies: localized standards may be developed and used

in isolation; when multiple schemes exist like this, guidance may be referenced and applied

inconsistently (Park, 2007); or Romanized forms may be decided on independent of any guidance.

Moreover, the choice to record Romanizations only may preclude access to resources

by users unfamiliar with such schemes or who would transcribe or transliterate differently

(Rigby, 2015). This raises further ethical questions about who metadata caters to when rendered

only in translation, transcription, or transliteration.d

·          [[MH4]](#X85abeb476d0392f07125e4c76435600885dff70)Vera-Baceta, M.-A., Thelwall, M., & Kousha, K. (2019). Web of Science and Scopus language coverage. *Scientometrics*, *121*(3), 1803–1813. <https://doi.org/10.1007/s11192-019-03264-z>  
  
English is by far the dominant language in both Scopus (at >92%) and WoS (>95%). Generally, WoS has less coverage of most (non-English) languages than Scopus, and in particular, has less than 1/10 as many works in Chinese (which comes in second, at 2.76%, in Scopus). WoS does have higher coverage of some languages, including Spanish. These numbers differ somewhat from studies before 2015, as WoS has since added the Emerging Sources Citation Index to its core collection, expanding coverage of non-English works.

·         Language coverage varies widely across subject areas.

·          [[MH5]](#X7e603da9b74bb38b1f0a636e3364aa6b45d0279)Cespedes:  
  
Find that even accounting for errors, works in OpenAlex have far greater linguistic diversity than Web of Science. WoS is more than 95% English, whereas the OpenAlex sample is identified as such by the metadata 75% of the time.

·         Manual examination of the works themselves (vs. metadata) brings this number closer to 68%. English is thus overestimated, as are other Western European languages, including French and German. Chinese and Russian are underreported.

·         The language field is populated for 98% of the documents, and is true to the underlying document for more than 85%.

·         Lack or presence of DOIs does not seem to impact accuracy of language.

·         Point out that the OpenAlex language field is algorithmically derived from the metadata (title & abstract), and that the *langdetect* algorithm is limited to 55 languages.

[[MH6]](#Xffd31f161222bc13310af6148f97a188ff9539e)Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author‐based citation analysis? Journal of the American Society for Information Science and Technology, 63(9), 1820–1833. https://doi.org/10.1002/asi.22695

·          [[MH7]](#X7573be7a9f3fadbde314fe785a23b9df9f31396)Without complex author disambiguation, even in a single field (stem cell research) over a short period (2004-2009), more than a third of “top authors” wind up being conflated groups of hundreds of individuals sharing a last name + first initial. Chinese and Korean names are particularly affected by this. The differences between author co-citation analyses with and without advanced disambiguation is significant when such authors are present, and the effect is generally a complex one, rather than simply the conflation of pairs of authors.

·         Mentions that Web of Science could be a problem, given the possibility of omitting the last author (which is generally easier to disambiguate, given the lower number of name collisions, and lower prevalence of Chinese/Korean names in this position).

[[MH8]](#X36a6426aa41780641f78dd6b2706f9bf519565c)<https://osf.io/preprints/metaarxiv/smxe5_v2>

[[MH9]](#X242b26289c443ef0ee1d29f3c3769bc66c66078)Oncepossibledatabaseerrorsareidentified,theycanbenotifiedtothedatabasestaffthroughdedicatedsupport/feedback mechanisms. We have noticed that Scopus and WoS are both very responsive to these feedbacks (Meester et al., 2016). As regards database administrators, at the risk of being repetitive, we renew our exhortation to improve in terms of data cleaning. We remark that all the database errors analyzed and classified in this research were preventable: in fact, all the citations omitted by one database are, by definition, correctly indexed by the other one. Weareawarethatthecitation-matchingalgorithmsusedbydatabaseswillneverbeinfallible,astheystruggletofindthe optimalbalancebetween(i)theriskoffailingtoidentifyauthenticcitations(falsenegatives)and(ii)thatofassigningphantom citations (false positives). Nevertheless, we believe that databases could introduce additional (automated) controls on the results of the citation mapping process (e.g., not to reinvent the wheel, the automated algorithm presented in (Franceschini et al., 2013)). This would be much more effective than waiting for the feedbacks from users, with an important benefit in terms of image. In the interest of the entire scientific community and their own one, we hope that Scopus and WoS will invest in such improvements.

Alonso-Alvarez, P., & Eck, N. J. van. (2024). Coverage and metadata availability of african publications in OpenAlex: A comparative analysis. *arXiv Preprint arXiv:2409.01120*. <https://doi.org/10.48550/arXiv.2409.01120>

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

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

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 Preprint arXiv:2409.10633*. <https://doi.org/10.48550/arXiv.2409.10633>

Clarivate. (n.d.). The history of ISI and the work of eugene garfield. In *The Institute for Scientific Information*. <https://clarivate.com/academia-government/the-institute-for-scientific-information/history/>

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of web of science. *Scientometrics*, *113*, 219–236. <https://doi.org/10.1007/s11192-017-2483-y>

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 Preprint arXiv:2406.15154*. <https://doi.org/10.48550/arXiv.2406.15154>

Liu, W., Huang, M., & Wang, H. (2021). Same journal but different numbers of published records indexed in scopus and web of science core collection: Causes, consequences, and solutions. *Scientometrics*, *126*(5), 4541–4550. <https://doi.org/10.1007/s11192-021-03934-x>

Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2024). Geographical and disciplinary coverage of open access journals: OpenAlex, scopus and WoS. *arXiv Preprint arXiv:2411.03325*. <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>

Meester, W. J. N., Colledge, L., & Dyas, E. E. (2016). A response to “the museum of errors/horrors in scopus” by franceschini et al. *Journal of Informetrics*, *10*(2), 569–570. <https://doi.org/10.1016/j.joi.2016.04.011>

Mongeon, P., Bowman, T., Costas, R., & Arroyo-Machado, W. (2024). *An update on the scholars on twitter dataset*. <https://www.leidenmadtrics.nl/articles/an-update-on-the-scholars-on-twitter-dataset>

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. *arXiv Preprint arXiv:2205.01833*. <https://doi.org/10.48550/arxiv.2205.01833>

Schares, E. (2024). Comparing funder metadata in OpenAlex and dimensions. *Iowa State University*. <https://doi.org/10.31274/b8136f97.ccc3dae4>

Simard, M.-A., Basson, I., Hare, M., Larivière, V., & Mongeon, P. (2025). Examining the geographic and linguistic coverage of gold and diamond open access journals in OpenAlex, scopus and web of science. *Quantitative Science Studies*, 1–29. <https://doi.org/10.1162/QSS.a.1>

Singh, P., Singh, V. K., & Kanaujia, A. (2024). Exploring the publication metadata fields in web of science, scopus and dimensions: Possibilities and ease of doing scientometric analysis. *Journal of Scientometric Research*, *13*(3), 715–731. <https://doi.org/10.5530/jscires.20041144>

Wikipedia contributors. (2024). *Segal’s law*. <https://en.wikipedia.org/wiki/Segal%27s_law>; Wikipedia, The Free Encyclopedia.

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>

1. <https://www.webofscience.com/wos/> (originally named the Science Citation Index); <https://www.scopus.com/>; <https://www.crossref.org/>;  <https://scholar.google.com/>;  <https://www.dimensions.ai/>;  <https://openalex.org/>. [↑](#footnote-ref-20)
2. <https://pubmed.ncbi.nlm.nih.gov/>; <https://www.lens.org/>. [↑](#footnote-ref-27)
3. Turgel and Chernova (2024) usefully point out that “citation”, “bibliometric”, and “Scientometric” databases are all distinct. Citation databases focus on citation connections between documents; similarly, bibliometric databases focus on publications and their citation metrics. Scientometric databases possess a range of metadata and indicators for more robust analyses of the structure and dynamics or the research landscape. They define Web of Science and OpenAlex as Scientometric databases (pg. 4). [↑](#footnote-ref-30)
4. See OpenAlex’s data sources page: <https://help.openalex.org/hc/en-us/articles/24397285563671-About-the-data> [↑](#footnote-ref-31)
5. <https://docs.openalex.org/api-entities/works/work-object#language> [↑](#footnote-ref-33)
6. Segal’s law states that “A man with a watch knows what time it is. A man with two watches is never sure” (Wikipedia contributors, 2024) [↑](#footnote-ref-37)
7. See Bruce & Hillmann (2004) for a definition of each of these characteristics. [↑](#footnote-ref-38)