






Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar

Nick Haupka ¹, Jack H. Culbert ², Alexander Schniedermann
³, Najko Jahn ¹, and Philipp Mayr ²

¹Göttingen State and University Library, University of Göttingen

²GESIS – Leibniz Institute for the Social Sciences

³German Centre for Higher Education Research and Science
Studies, DZHW

June 2024

Abstract

This study compares and analyses publication and document types in the following bibliographic databases: OpenAlex, Scopus, Web of Science, Semantic Scholar and PubMed. The results demonstrate that typologies can differ considerably between individual database providers. Moreover, the distinction between research and non-research texts, which is required to identify relevant documents for bibliometric analysis, can vary depending on the data source because publications are classified differently in the respective databases. The focus of this study, in addition to the cross-database comparison, is primarily on the coverage and analysis of the publication and document types contained in OpenAlex, as OpenAlex is becoming increasingly important as a free alternative to established proprietary providers for bibliometric analyses at libraries and universities.

Keywords: Publication Analysis, Publication Types, Coverage, Bibliometric databases, Open Scholarly Metadata

1 Introduction

The shift towards open science implies an emerging role of open metadata for scientometric analyses and has led to the development of open bibliographic

databases. Consequently, an increasing body of research has focused on comparing these open resources to established commercial data sources in terms of coverage and quality (Culbert et al., 2024; Waltman & Larivière, 2020). In this large-scale study, we examine the classification of publication and document types in the open data sources OpenAlex, Semantic Scholar, and PubMed in comparison to the proprietary databases Scopus and Web of Science.

Despite the varied typologies employed by these data sources, the extent of their differences remains unclear. Understanding these variations is crucial because the classification of publication and document types by bibliometric data sources influences the calculation of indicators. For example, some bibliometric studies exclude editorials when calculating average citation rates for journals to ensure more meaningful and robust indicators. Conversely, the calculation of the original Journal Impact Factor can be manipulated by journals altering their document types, because this metric includes citations to all documents in the nominator but only articles and reviews in the denominator (Moed & Van Leeuwen, 1995).

Bibliometric data sources use different curation strategies to assign publication and document types. For instance, WoS tags a document as a review if it has been classified as a review by a journal or if the word *review* is included in the title as well as in the text and has at least one reference¹. In contrast, PubMed’s classification system also involves human indexers who apply the standardised Medical Subject Headings (MeSH) (van Buskirk, 1984). OpenAlex reuses and aggregates types from Crossref² and has recently started to extend the existing classification to certain types such as reviews and preprints³. The accuracy of these assignments is uncertain, with recent studies pointing to possible errors (Alperin, Portenoy, Demes, Larivière, & Haustein, 2024; Visser, van Eck, & Waltman, 2021). An analysis by Donner (2017) on document types in WoS and Scopus has shown that around 17 percent of publications in WoS had an incorrect document type classification. Similarly, Mokhnacheva (2023) compared document types of 3,843 publications by Russian authors between 2010 and 2020 in Scopus and WoS and found differences in the typification of documents from the databases and publisher websites.

In the following study, we first provide a conceptual overview of our interpretation of document, publication, and study types. Building on these definitions, this study will then focus on a comparison of document and publication types between OpenAlex, Scopus, WoS, Semantic Scholar and PubMed. In addition to a general analysis of publication types, we examine and compare document types of specific items from journals from 2012 to 2022 to assess the quality of the classification of the corresponding data sources.

¹<https://webofscience.help.clarivate.com/en-us/Content/document-types.html>

²https://docs.openalex.org/api-entities/works/work-object#type_crossref

³<https://groups.google.com/g/openalex-users/c/YujaIIjY02A>

2 Background

Document-type classification can be used to represent the epistemic characteristics of research publications. Considering all scientific domains, most of the research results are published as journal articles, thereby this document type is known as the gold standard for bibliometric research (van Raan, 2004). Review articles aggregate and synthesise prior research and are important for agenda setting or field formation (Blümel & Schniedermann, 2020). While both types are considered the backbone of the main research discourse, editorials, letters, and news items typically do not communicate novel insights. Rather they represent an editorial discourse or meta-debate on research practices and the field as a whole. In practice, the boundaries between different types are less well-defined. Dissertations can include extensive and systematic literature reviews, and some research letters or comments report the results of small experiments performed to validate or question the results of a research article. To systematise the problem of classifying document types, three dimensions are helpful.

- *Publication or source types* represent the different venues in which document types are published. Such venues have been source types in traditional databases and can include periodic journals, monographs, edited volumes, and conference proceedings. These venues have been the subject to the curating procedures of databases such as WoS, leading to the inclusion of all items from included sources. Indexing practices can vary, for example, Scopus only indexes serial publications assigned to an International Standard Serial Number (ISSN), as well as one-off conferences and one-off books⁴. In contrast to WoS and Scopus, OpenAlex includes additional venues, such as preprint servers like the arXiv. This may lead to higher number of versions of the very same research contribution. OpenAlex makes use of a fingerprinting algorithm that matches and chronologically sorts different versions of a text across venues. This ensures that the latest version of a document (i.e., a preprint) is identifiable (Priem, Piwowar, & Orr, 2022).
- *Document types* represent the types or genres of texts rather than methods or research practices, and are usually included as metadata in databases. Some document types, such as editorials or letters, are similar in their textual characteristics and scholarly functions. However, the majority of document types are classified as common journal articles and are not further specified in the major databases. As such, they can represent the results of different study types, theoretical discussions, and opinion pieces.
- The *study type* usually represents a set of research methods or project designs and describes what researchers actually did in their labs. Study types are especially apparent in standardised research fields, such as biomedicine,

⁴https://web.archive.org/web/20240527182403/https://assets.ctfassets.net/o78em1y1w4i4/EX1iy8VxBEQKf8aN2Xz0p/c36f79db25484cb38a5972ad9a5472ec/Scopus_ContentCoverage_Guide_WEB.pdf

while they can be less accentuated in other fields. Therefore, they are not commonly represented as metadata in databases such as WoS or Scopus. However, they can be even more important in specialist databases where they are used for indexing and information retrieval. For example, medical experts need to retrieve particular types of information, such as randomised-controlled trials for high-quality evidence synthesis and informing health policies (Glanville, Lefebvre, Miles, & Camosso-Stefinovic, 2006). Illustratively, the Medical Subject Headings (MeSH) which serve as a standardised set of tags for information retrieval in medicine provide their own category, “V03 - Study Characteristics” for classifying study types ⁵.

These dimensions reflect the different epistemic, textual, or social characteristics of texts and can provide complex classification problems that lead to confusion and errors in bibliometric research. Study types can be used synonymously to document types in specialist databases based on field-specific definitions. For example, medical experts have standardised reporting practices in their domain which lead to the emergence of new text genres. Consequently, the systematic review became a separate type in MEDLINE/PubMed in 2019 that can be used to filter results and inform information retrieval (Collins, 2019). Vice versa, genre types and their textual characteristics can also shape study types in a field, for example by requiring certain types of visualisations or suggesting a certain sequence of events in the research process (Bazerman, 1988).

Document types can be confused with publication types or venues in bibliometric studies, or used synonymously (Scheidsteger & Haunschild, 2023). This is further complicated as document-type classifications can be derived from the publication venue. Preprints may be classified solely by their repository as a source venue, as suggested below. Likewise, journal articles published in review journals may be classified as reviews. However, in such cases, classifications are derived from disciplinary definitions that make them less comparable in cross-disciplinary analyses (Sigogneau, 2000). As an illustrative example, encyclopedia entries in the humanities may be closer to the idea of a review article than other publications. Bibliometricians that use such information on the basis of academic databases such as WoS, Scopus or OpenAlex must be aware that there is always a trade-off between the internal and external validity of document type classifications.

In addition, the role of multiple venues for the same or almost similar text provides a legitimate yet problematic source for multiple and conflicting document type classifications. For example, the addition or removal of references during the peer review process may turn a preprinted review into a published article if the classification system is solely based on the number of references. Similarly, systematic reviews in medicine can be (re-)published as updated or abbreviated versions (Bashir, Surian, & Dunn, 2018), and method papers or guidelines are translated into other languages. Thus, even with more sophisticated classification algorithms that consider a palette of metadata, the same

⁵<http://id.nlm.nih.gov/mesh/D052182>

act of writing can result in differently classified texts. For this reason, bibliometric studies must differentiate between the publication of a text on the one hand and the publication of an idea or a result on the other when they conclude the social dynamics of science. Similarly, database providers should consider methods such as the fingerprinting approach in OpenAlex (Priem et al., 2022) which allows keeping both - a record of all published items and also pointers that reconnect the different faces and versions of the same text.

3 Data and Methods

3.1 Data

In this work, publication and document types from Openalex (August 2023 snapshot), Web of Science (July 2023 snapshot), Scopus (July 2023 snapshot), PubMed (December 2022 snapshot) and Semantic Scholar (September 2023 snapshot) were analysed. The WoS data includes the collections: Science Citation Index Expanded (SCI), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI), Conference Proceedings Citation Index-Science (ISTP) and Conference Proceedings Citation Index-Social Sciences & Humanities (ISSHP). Because PubMed features mostly biomedical research, it is not used as commonly as other databases in bibliometric research, although it has been used to compare and cross-validate metadata (Rotolo & Leydesdorff, 2015). With MeSH, PubMed features a multi-purpose classification system that was originally based on human indexing and only recently shifted to automated indexing (Mork, Yepes, & Aronson, 2013). Its document-type assignment provides a valuable reference point for the comparison of different data sources. However, the MeSH categories that represent document types, publication types, and study types were provided as a single datatype in PubMed's XML files. For example, according to the logic of the indexing rules, all "Reviews" are also co-assigned "Journal Article" because the latter is used as a general tag for scholarly content in academic journals. To make the data comparable, we reduced the multi-assignment by employing a hierarchy that ranks the individual document types. For example, research-related types ranked highest, followed by editorials, letters, news items, review articles, and journal articles (see Table 11 for the complete scheme). After ranking, only the highest-ranked type was retained for each item. We restricted the data to the publication years 2012 to 2022 and to publications with a Digital Object Identifier (DOI). All data were analysed within a custom SQL database environment maintained by the German Competence Network of Bibliometrics⁶.

Table 1 compares the number of joint publications indexed in OpenAlex. Because PubMed contains the fewest publications, the number of publications in all five databases is correspondingly lower. Following Culbert et al. (2024), this intersection is referred to as a shared corpus. For comparison, we used the DOI instead of the PubMed identifier (PMID) to match records from the

⁶<https://www.bibliometrie.info>

respective databases. Since the inception of versioning for PMIDs, they are no longer unique-related items, that is subsequent stages of the same preprint, can obtain the same PMID but differ only by version (Torre, 2012). As version information is currently not indexed by WoS or Scopus, matching by PMID would cause inflated numbers of matches. The DOIs for items in each database were normalised beforehand. Because the selected PubMed snapshot is from December 2022, not all publications in PubMed from the publication year 2022 were fully covered.

Data source	Number of publications
OpenAlex	69,456,021
Scopus	31,922,514 (96% included in OpenAlex)
Web of Science	23,540,852 (99% included in OpenAlex)
Semantic Scholar	56,279,413 (97% included in OpenAlex)
PubMed	12,681,219 (99% included in OpenAlex)
Shared Corpus	9,575,603

Table 1: Data Sources. The number of publications is limited to those records with publishing year between 2012 and 2022 inclusively.

Each database has its own classification of items according to the publication and document type. In PubMed, publication types are not clearly separable from document types because the database only provides a single metadata field that can cover publication types, document types, study types and even funding information. In OpenAlex and PubMed, one publication and one document type were assigned to each item. In the case of OpenAlex, the venue type specification in the primary location was considered. The primary location is where the best copy of a work can be found (closest to the version of record)⁷. In Semantic Scholar, Scopus and WoS, documents are covered multiple times without being linked or subordinate to a primary location, resulting in multiple records of the same article. These were all taken into account in this study and were not reduced to one type, which means that the calculated numbers in the Tables 3 and 4 can differ between the data sources when the shared corpus is analysed. Scopus, WoS, and Semantic Scholar specified one publication type for each publication. Venue types were compared pairwise. For example, the intersection between OpenAlex and Scopus was first calculated and then the corresponding venue types were compared. The publication year from OpenAlex was used, unless otherwise stated. The publication year of a publication may differ between the investigated databases, depending on which version of a work is indexed in the database (Delgado-Quirós & Ortega, 2024; Ortega, 2022).

Additional data from Crossref was used for Table 5 (snapshot from October 2023). The publication data from Crossref was linked to the shared corpus and supplemented. All the characteristics analysed in Table 5 were compiled from

⁷https://docs.openalex.org/api-entities/works/work-object#primary_location

information from Crossref (such as whether funding information or licensing information is available or the number of references.)

3.2 Methodology

First, publication types in the various databases were analysed (Section 4.2). Then, the document types of items in the specified databases that were assigned to the publication type *journal* were investigated (Section 4.3). To allow for database comparisons, document types were categorised into three groups (see Table 4):

Research Discourse: This includes articles and reviews that focus on presenting and analysing research findings.

Editorial Discourse: This covers letters, editorials, and comments that express opinions and engage in discussions.

Not Assigned: All works that cannot be assigned to either research or editorial discourse were categorised here.

In July 2023, OurResearch implemented a new item classification, replacing the previously used item classification of Crossref in OpenAlex⁸. In this process, the types *journal-article*, *proceedings-article*, and *posted-content* were combined into the type *article*, which makes it difficult to differentiate between conference articles, preprints, and journal articles. As this transition has not yet been fully completed in the used snapshot, we have manually reclassified the remaining old types using the OpenAlex procedure⁹.

Within the OpenAlex database, retractions were not assigned distinct document types. Instead, OpenAlex provided a designated field to label retractions (*is_retracted*)(Hauschke & Nazarovets, 2024). Because retractions aren't separated into a unique category within OpenAlex, they are likely included in most analyses by default. For this analysis, we ignored the information from OpenAlex in the fields *is_paratext* and *is_retracted*.

4 Results and Discussion

4.1 Overview

First, we identified the different typologies in our data sources and assess how comprehensively each category is covered (see Table 2). We then calculated the coverage ratio separately for publication and document types.

We found that, compared to the proprietary providers, the open databases have a lower coverage of publication and document types. Scopus and WoS achieve full coverage of document types. In WoS, each item is also assigned a publication type. This value is similarly high in Scopus (99.97%). However, both databases differ in terms of typology. WoS offers a much more granular

⁸<https://docs.openalex.org/api-entities/works/work-object#type>

⁹<https://github.com/ourresearch/openalex-guts/>

document type classification system compared to Scopus, with 87 categories versus 18.

Focusing on the open databases, it can be seen that OpenAlex and PubMed also nearly have 100% of the items assigned with a document type. However, the proportion of items with a publication type was lower than that of Scopus and WoS (86% of items in OpenAlex have a publication type). No publication types were separately specified in PubMed. The number of document types in PubMed is similar to Web of Science (79 document types).

Overall, Semantic Scholar had the lowest coverage, providing two publication and 12 document types. Less than half of the items indexed in Semantic Scholar were assigned to a publication type, and 37% to a document type.

Data Source	Publication Type Typologies	Publication Type Coverage	Document Type Typologies	Document Type Coverage
OpenAlex	5	85.85%	16	100%
Scopus	7	99.97%	18	100%
Web of Science	3	100%	87	100%
Semantic Scholar	2	43.60%	12	37.04%
Pubmed	/	/	79	99.99%

Table 2: Size of publication type and document type typologies per data source. The coverage ratio is calculated separately for publication and document types.

4.2 Publication types

The analysis of publication types illustrated that the various database providers adopt different approaches to classifying venues (see Figure 1, 2, 3). In addition, providers have different focuses on certain types of publication. For Figures 1, 2 and 3, we have considered the intersection of a selected database with OpenAlex for the publication years 2012 to 2022. This means that publications and the number of publication types that are not included in one of the databases are not shown. The comparison is a 1:1 allocation, because only one publication type was present to each of the databases examined.

Journals comprised the largest share in all databases. This was followed by conference proceedings and books. A large proportion of items was not assigned to a venue, in particular in Semantic Scholar.

Publications that were not assigned to a venue in OpenAlex were largely classified as conference proceedings in the other investigated databases. Conference proceedings have been included in WoS, however, the publication type field lacks an explicit designation for this type. Conference proceedings contained in WoS are indexed as publication type *book*, *book in series* or *journal*.

In Semantic Scholar books are not included as a publication type so that books from OpenAlex cannot be matched to any venue. Items that are assigned

to the venue type repository in OpenAlex are generally less covered in Web of Science and Scopus where they are categorised as a journal. Compared to Semantic Scholar, more repositories from OpenAlex are covered and match either to the type journal (410,444 items) or no venue at all (653,592 items).

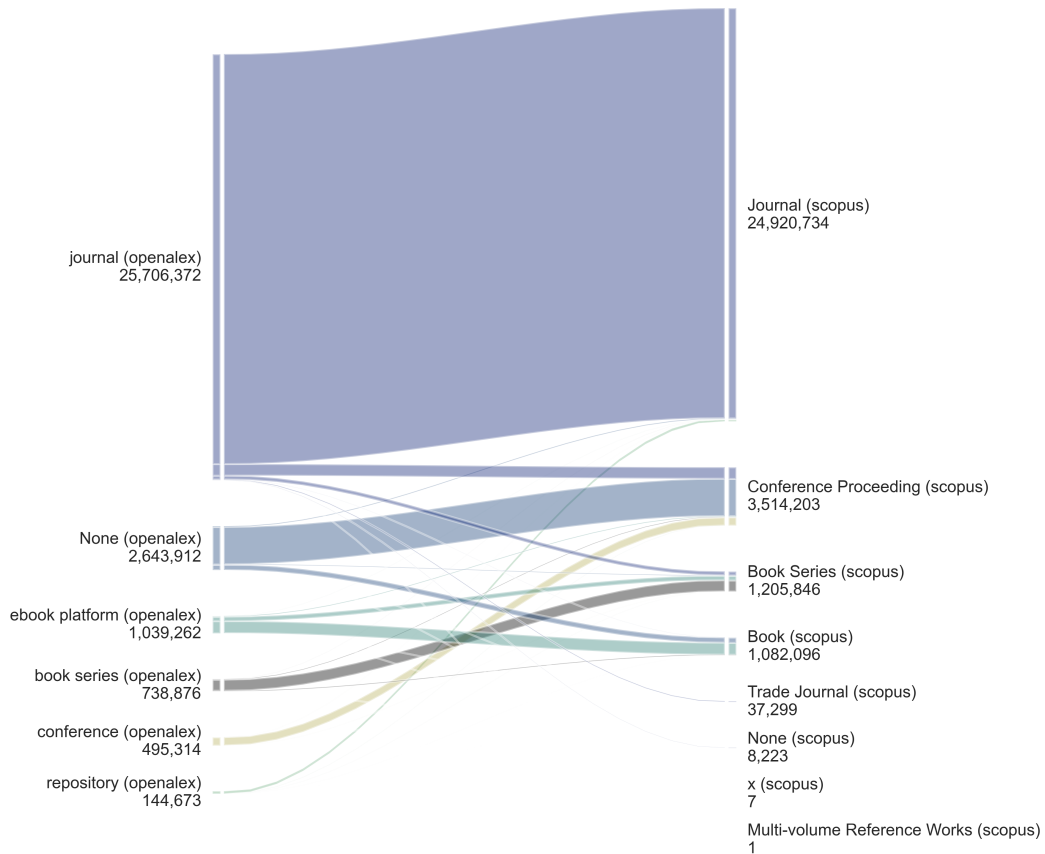


Figure 1: Publication type classification between OpenAlex and Scopus. The intersection between Scopus and OpenAlex is limited to those records with publishing year between 2012 and 2022 inclusively.

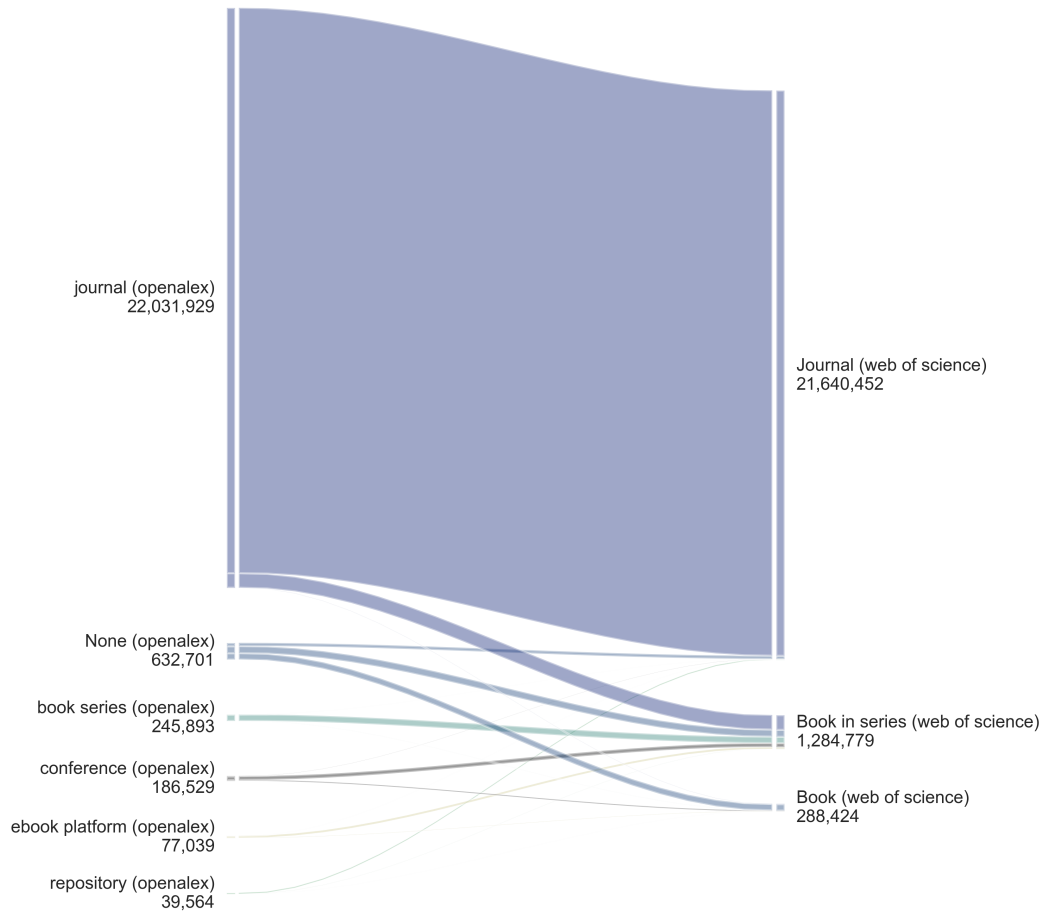


Figure 2: Publication type classification between OpenAlex and WoS. The intersection between WoS and OpenAlex is limited to those records with publishing year between 2012 and 2022 inclusively.

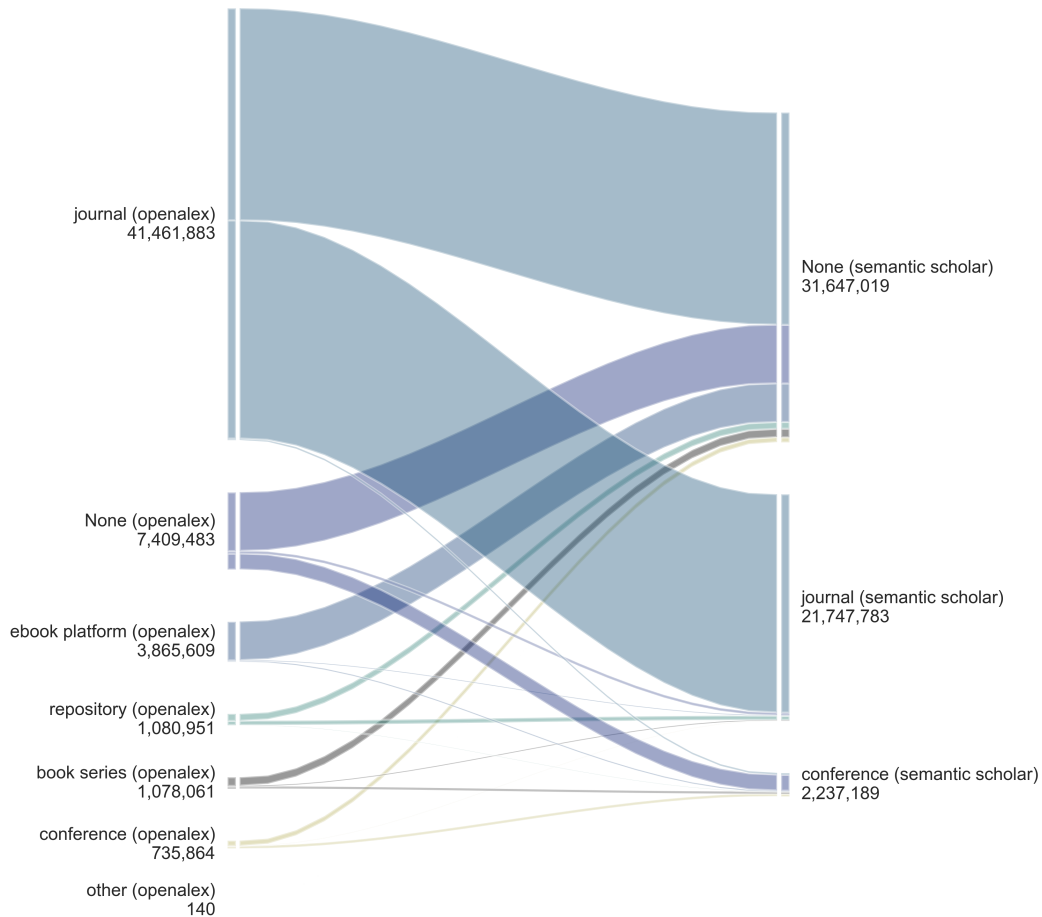


Figure 3: Publication type classification between OpenAlex and Semantic Scholar. The intersection between Semantic Scholar and OpenAlex is limited to those records with publishing year between 2012 and 2022 inclusively.

4.3 Document Types

Firstly we demonstrate how the different item types correspond to the venue types in OpenAlex in Figure 4. The majority of the items classified as article had the publication type journal. Approximately five million items classified as article were not assigned to a publication type. However, Figures 1 and 3 suggest that these could be conference articles, as a certain proportion of the items that are not assigned to a publication type in OpenAlex have been classified

as conference publications by Scopus and Semantic Scholar. Around two million articles have the venue type repository. According to OurResearch, these are preprints¹⁰, because Microsoft Academic Graph (MAG), as a precursor to OpenAlex, also covered preprint servers as venues (Xie, Shen, & Wang, 2021) and their items were labeled as journal articles in OpenAlex (Scheidsteger & Haunschild, 2023). In fact, when counting items from repositories per source title since 2020, most stem from arXiv, SSRN, or bioRxiv. Figures 1 and 3 demonstrate that these have the publication type journal in Scopus and Semantic Scholar. The publication type book and book series was represented by around 9 million items. The number of publication type classified as *conference* was determined to be almost 1 million for the period 2012 to 2022.

¹⁰<https://github.com/ourresearch/openalex-guts/>

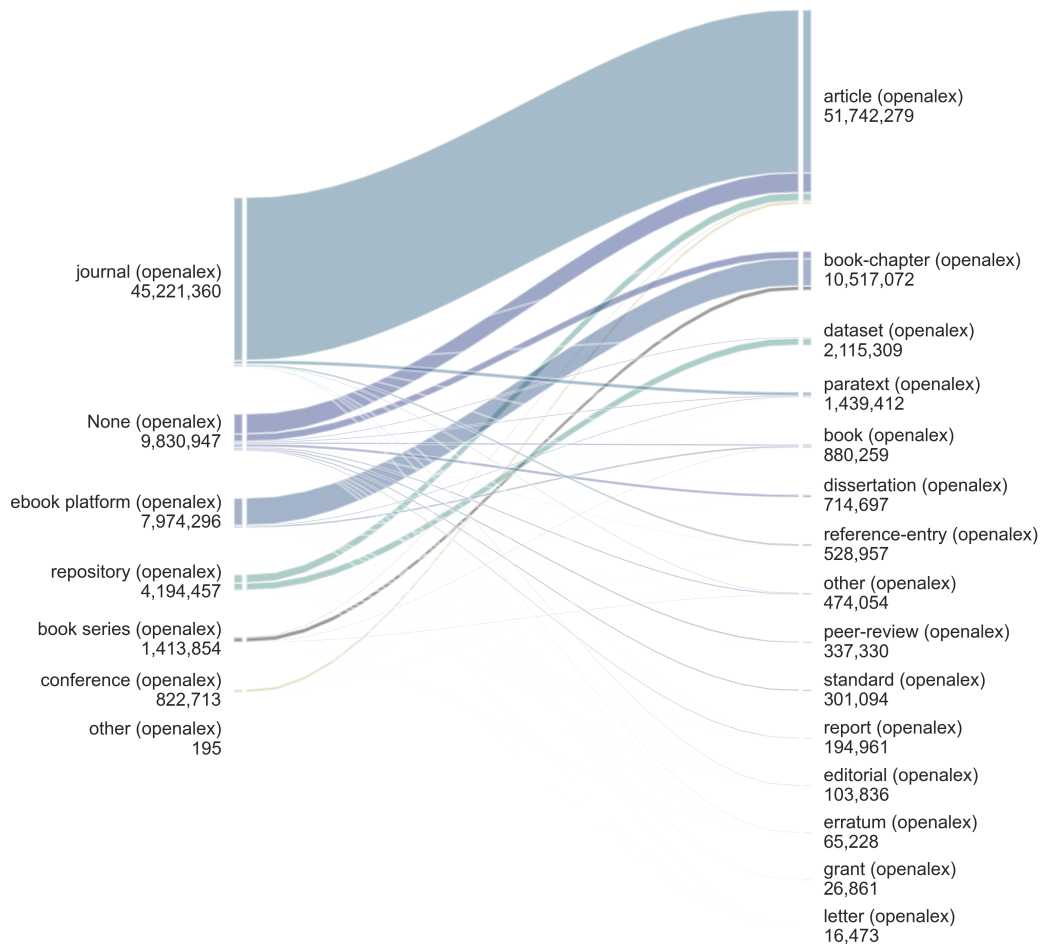


Figure 4: Publication types and their corresponding document types in OpenAlex. Data is limited to those records with publishing year between 2012 and 2022 inclusively.

Table 3 presents the document types of items published in journals between 2012 and 2022 in our shared corpus. Journal articles constituted the largest share. OpenAlex classified over 99% of the items as journal articles. This proportion was lower in Scopus (80%), WoS (78%), Semantic Scholar (73%) and PubMed (75%). Except for OpenAlex, all databases have classified reviews. With around 9%, reviews constituted the second largest share. Only Semantic Scholar had a larger share (15%). Editorial materials such as editorials, errata or

letters are classified in all databases. However, this proportion was also higher in the other databases. In OpenAlex, for example, only around 0.3% of the items were editorial material. In Scopus, around 4% of items were letters, 2% note and 2% editorials. In WoS, the proportion of letters was similar (4%), but the proportion of editorials was slightly higher (around 7%). The proportion of letters in PubMed and Semantic Scholar were also similar to those in WoS and Scopus at 4%.

OpenAlex		Scopus		WoS		Semantic Scholar		PubMed	
Type	n	Type	n	Type	n	Type	n	Type	n
article	9,528,567 (99.51%)	Article	7,618,472 (79.51%)	Article	7,535,712 (77.77%)	Journal	8,836,640 (72.95%)	Journal	7,178,813 (74.96%)
editorial	20,316 (0.21%)	Review	923,486 (9.64%)	Review	878,361 (9.07%)	Review	1,765,638 (14.58%)	Review	878,361 (9.19%)
erratum	15,201 (0.15%)	Letter	358,420 (3.74%)	Editorial	637,569 (6.58%)	Letters	455,431 (3.76%)	Letter	352,031 (3.68%)
letter	9,712 (0.1%)	Note	251,305 (2.62%)	Letter	375,866 (3.88%)	Com-ments	410,735 (3.40%)	Clinical	293,069 (3.06%)
paratext	636 (0.01%)	Editorial	217,034 (2.27%)	Correction	93,702 (0.97%)	Study	263,532 (2.18%)	Trial	279,377 (2.92%)
reference- entry	520 (0.01%)	Erratum	99,979 (1.04%)	Proceedings	70,846 (0.73%)	Report	215,889 (1.78%)	Report	229,707 (2.4%)
report	1 (0%)	Conference	53,858 (0.56%)	Paper	37,479 (0.39%)	Editorial	71,035 (0.59%)	Editorial	198,037 (2.07%)
Other	654 (0.01%)	Paper	59,403 (0.62%)	News	60,029 (0.62%)	Clinical	93,663 (0.77%)	Meta	165,334 (1.73%)
		Other		Item		Trial		Analysis	
				Other		Other		Other	

Table 3: Item classification per data source (shared corpus)

It is noticeable that the classification of items is similar between Semantic Scholar and PubMed. Similar percentages were calculated except for reviews. Note that PubMed types were re-classified to obtain a more compact overview.

To determine the relationship between scientific and editorial texts in the respective data sets, the types were reclassified into the categories of research discourse and editorial discourse. For example, articles and reviews are classified as research discourse and letters and editorials as editorial discourse (see Table 4). Types that cannot be assigned to either class (such as books or dataset) are not assigned to any class (not assigned). Only items that were published in journals between 2012 and 2022 and can be found in all five databases were considered being part of the analysed shared corpus.

Table 4 shows that the proportion of editorial discourse is low in OpenAlex (0.5%). The proportion of editorial material is higher in Scopus (10%), WoS

(12%), Semantic Scholar (6%) and PubMed (8%). The proportion of research discourse is correspondingly lower (between 87% and 94%), although this proportion is still highest in Semantic Scholar.

	OpenAlex	Scopus	WoS	Semantic Scholar	Pubmed
Research Discourse	9,528,567 (99.51%)	8,541,958 (89.15%)	8,414,073 (86.84%)	11,408,341 (94.19%)	11,362,371 (88.58%)
Editorial Discourse	45,865 (0.48%)	926,738 (9.67%)	1,153,900 (11.91%)	703,573 (5.81%)	985,505 (7.68%)
Not assigned	1,175 (0.01%)	113,261 (1.18%)	121,591 (1.25%)	649 (0.01%)	479,936 (3.74%)

Table 4: Item reclassification (shared corpus)

Table 5 provides information on the characteristics of research and editorial texts, based on the shared corpus. For comparison, additional data from Crossref were used and added to the shared corpus. PubMed document types were reclassified to allow for comparison. The data shows, that in contrast to editorial articles, research publications on average have a higher number of authors, citations and references. The average research article has almost twice as many authors as the average editorial text. The same applies to the number of pages and word length of titles and abstracts of scientific articles. For example, the abstracts of scientific articles have an average count of 207 words and editorial articles have 106 words. In addition, it is more likely that funding information and abstract information are provided for research articles.

	Research Discourse			Editorial Discourse			Not assigned		
	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std
Number of authors	5.0	6.5	14.3	3.0	3.8	10.6	1.0	2.7	4.9
Number of citations	11.0	25.1	89.2	1.0	5.2	32.4	1.0	6.6	37.7
Number of references	38.0	44.2	38.5	5.0	7.7	12.0	1.0	11.9	30.3
Number of pages	5.0	37.3	6.6	1.0	14.0	247.5	1.0	7.4	141.8
Abstract word length	207.0	206.6	81.1	44.0	94.8	106.4	129.0	141.5	84.6
Title word length	14.0	14.0	5.0	11.0	11.7	6.1	8.0	9.0	5.8
Abstract	36.51%			6.65%			12.41%		
Funding information	37.11%			7.52%			3.71%		
License information	74.50%			72.89%			64.40%		

Table 5: Characteristics (shared corpus)

4.4 Discussion

This study compared the publication and document types in OpenAlex with those in Scopus, WoS, Semantic Scholar, and PubMed. The results show, that the open databases OpenAlex and Semantic Scholar seem to comprise a broader range of materials as research publications compared to Scopus, WoS and PubMed (see Table 4). This difference might be due to the reliance of OpenAlex and Semantic Scholar on Crossref, which has a less precise classification than Pubmed, Scopus and WoS (Delgado-Quirós & Ortega, 2024). In contrast, Scopus, WoS and PubMed utilise independent and more detailed classification systems. In addition, a classification system that assigns multiple types to each document (e.g. WoS and Scopus) has the potential to capture a more complex picture of scientific genres, or is even able to distinguish between study types, document types and source types. However, independent of their granularity or complexity, such categorisations can differ substantially from those of other databases or journal-based assignments (Donner, 2017). Therefore, database users must consider that there are no correct or incorrect categorisation schemes. Rather, categorisations represent the different curation approaches.

Nevertheless, document type classifications are critical for robust and reproducible bibliometric research, as well as bibliometrics-based research evaluation. A finer-grained palette of document types in OpenAlex, i.e. classifying review articles, as well as an improved type accuracy would be highly beneficial. However, achieving a universally standardised classification across bibliometric databases is a substantial hurdle. In addition, generalist databases such as OpenAlex must find a classification scheme that balances the definitions of document types across a diverse set of research fields.

We propose multiple strategies to enrich the document type assignment practices in bibliometric databases. For instance, the abstract of a publication can be interpreted semantically to identify the conceptual or textual markers of a particular document type. For example, the phrase “we systematically review the literature” can be interpreted as *review*. This approach however, is limited by the availability of abstracts, as not all items are provided with an abstract in bibliometric databases (Culbert et al., 2024). Consequently, the accuracy of such a classification system would differ based on data availability, which introduces its own systematic data biases.

In addition, descriptions of studies in abstracts can differ from those of studies themselves. For example, research shows that authors may exaggerate the research results in their abstracts, so that abstracts might be better considered as a “promotional genre” (Bordignon, Ermakova, & Noel, 2021). Similarly, authors might spin the wording within their abstracts towards document genres that sound more systematic or to be considered as of higher value. Hence, abstracts might not provide sufficient text to classify article.

In future, full-text analyses are promising to detect publication and document types. For example, using semi-automatic methods to identify the structure of a document can help to improve the classification of a document (Caragea, Wu, Gollapalli, & Giles, 2016).

Research texts can also be distinguished from non-research texts based on various characteristics (see Table 5). However, this approach is problematic; for example, reviews and traditional research reports typically have different reference counts and page lengths. By contrast, editorials often have unique structures that are typically limited to one or two pages.

For future analyses, it is advisable to curate the document types. For example, by authors themselves selecting a document type for submission (Yeung, 2019). In addition, a stricter distinction made by editors and peer reviewers between research and non-research publications can lead to an improvement in data quality (Di Girolamo & Meursinghe Reynders, 2020). Some repositories, such as the French open archive Hyper Articles en Ligne (HAL), made author-selected document types available ¹¹. Also, OurResearch is actively working to improve categorisation of types in OpenAlex. Since May 2024, reviews and preprints will now be recognised as separate document types ¹².

5 Conclusion

This paper demonstrates inconsistencies in publication and document type classification across bibliometric databases. This problem stems from the lack of a standardised system for classifying publications. Each database utilises its own typologies and approaches which makes overarching comparisons difficult (see Table 5). Furthermore, the results indicate that the proprietary databases Web of Science and Scopus classify more publications as editorial content compared to the open databases (OpenAlex, PubMed and Semantic Scholar). However, when comparing only publication types, it shows that, with the exception of Semantic Scholar, all databases use a similar classification of venues, with a large overlap between OpenAlex, Scopus and Web of Science. Also, the results shows that research and non-research texts have different metadata characteristics, such as different numbers of references and availability of abstracts and funding information. Future studies should be aware of these typification differences when interpreting results from bibliometric databases.

6 Limitations

This study has some limitations. First, the categorisation of documents into research and editorial discourse in this study provides a simplified method for examining research and non-research content in the databases analysed. Second, for the analysis of the publication types, the selection was limited to a specific publication source, which was identified by the databases as the closest version to the original. However, it is common for publications to have multiple publication locations. The inclusion of PubMed data in the shared corpus has led to an over-representation of publications from the life sciences, as PubMed restricts

¹¹<https://hal.science>

¹²<https://groups.google.com/g/openalex-users/c/YujaIIjY02A>

its data to certain types of topics. Therefore, the analysed data might be biased in this context. Finally, the lack of ground truth makes it difficult to analyse publication and document types to determine their reliability in bibliometric databases.

Funding Information

This work was funded by the Federal Ministry of Education and Research (grant funding number: 16WIK2301B / 16WIK2301E, The OpenBib project; grant funding number: 01PU17017, The FuReWiRev project). We acknowledge the support from the Federal Ministry of Education and Research, Germany under grant number 01PQ17001, the Competence Network for Bibliometrics.

Code Availability

The Jupyter notebook containing the source code analysis can be found on GitHub: https://github.com/naustica/openalex_doctypes.

Author Contributions

Nick Haupka: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—Original Draft, Writing—Review and Editing.

Jack Culbert: Data Curation, Investigation, Validation, Writing—Review and Editing.

Alexander Schniedermann: Data Curation, Writing—Review and Editing.

Najko Jahn: Funding acquisition, Project administration, Supervision, Writing—Review and Editing.

Philipp Mayr: Funding acquisition, Project administration, Supervision, Writing—Review and Editing.

References

- Alperin, J. P., Portenoy, J., Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses*. arXiv. Retrieved 2024-05-06, from <http://arxiv.org/abs/2404.17663> doi: 10.48550/arXiv.2404.17663
- Bashir, R., Surian, D., & Dunn, A. G. (2018). Time-to-update of systematic reviews relative to the availability of new evidence. *Systematic Reviews*, 7(1), 195. doi: 10.1186/s13643-018-0856-9
- Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison, Wisconsin: University of Wisconsin Press.
- Blümel, C., & Schniedermann, A. (2020). Studying review articles in scientometrics and beyond: a research agenda. *Scientometrics*, 124(1), 711–728. Retrieved 2023-10-16, from <https://doi.org/10.1007/s11192-020-03431-7> doi: 10.1007/s11192-020-03431-7
- Bordignon, F., Ermakova, L., & Noel, M. (2021). Over-promotion and caution in abstracts of preprints during the COVID-19 crisis. *Learned Publishing*, 34(4), 622–636. Retrieved 2024-04-30, from <https://onlinelibrary.wiley.com/doi/10.1002/leap.1411> doi: 10.1002/leap.1411
- Caragea, C., Wu, J., Gollapalli, S., & Giles, C. (2016). Document Type Classification in Online Digital Libraries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(2), 3997–4002. Retrieved from <https://ojs.aaai.org/index.php/AAAI/article/view/19075> doi: 10.1609/aaai.v30i2.19075
- Collins, M. (2019). Pubmed updates february 2019. *NLM Tech Bull*, 426, e4. Retrieved 2024-04-24, from https://www.nlm.nih.gov/pubs/techbull/jf19/jf19_february_pubmed_updates.html
- Culbert, J., 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*. Retrieved 2024-02-13, from <http://arxiv.org/abs/2401.16359> doi: 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*, 1–19. Retrieved from https://direct.mit.edu/qss/article/doi/10.1162/qss_a.00286/119466/Completeness-degree-of-publication-metadata-in doi: 10.1162/qss_a.00286
- Di Girolamo, N., & Meursinghe Reynders, R. (2020). Characteristics of scientific articles on COVID-19 published during the initial 3 months of the pandemic. *Scientometrics*, 125(1), 795–812. Retrieved from <https://doi.org/10.1007/s11192-020-03632-0> doi: 10.1007/s11192-020-03632-0
- Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, 113(1), 219–236. Retrieved from <https://doi.org/10.1007/s11192-017-2483-y> doi: 10.1007/s11192-017-2483-y

- Glanville, J. M., Lefebvre, C., Miles, J. N. V., & Camosso-Stefinovic, J. (2006). How to identify randomized controlled trials in MEDLINE: ten years on. *Journal of the Medical Library Association: JMLA*, *94*(2), 130–136.
- Hauschke, C., & Nazarovets, S. (2024). *(Non-)retracted academic papers in OpenAlex*. Retrieved 2024-03-21, from <http://arxiv.org/abs/2403.13339> doi: 10.48550/arXiv.2403.13339
- Moed, H. F., & Van Leeuwen, T. N. (1995). Improving the Accuracy of the Institute for Scientific Information's Journal Impact Factors. *Journal of the American Society for Information Science*, *46*(6), 461–67.
- Mokhnacheva, Y. V. (2023). Document Types Indexed in WoS and Scopus: Similarities, Differences, and Their Significance in the Analysis of Publication Activity. *Scientific and Technical Information Processing*, *50*(1), 40–46. Retrieved from <https://doi.org/10.3103/S0147688223010033> doi: 10.3103/S0147688223010033
- Mork, J. G., Yepes, A. J. J., & Aronson, A. R. (2013). *The NLM Medical Text Indexer System for Indexing Biomedical Literature*. Lister Hill National Center for Biomedical Communications. Retrieved from <https://lhncbc.nlm.nih.gov/ii/information/Papers/MTI.System.Description.Expanded.2013.Accessible.pdf>
- Ortega, J. L. (2022). *When is a paper published?* Retrieved 2024-03-28, from <https://researchwhisperer.org/2022/02/08/when-is-a-paper-published/>
- Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. arXiv. Retrieved 2024-04-10, from <http://arxiv.org/abs/2205.01833> doi: 10.48550/arXiv.2205.01833
- Rotolo, D., & Leydesdorff, L. (2015). Matching Medline/PubMed data with Web of Science: A routine in R language: Matching Medline/PubMed Data With Web of Science: A Routine in R Language. *Journal of the Association for Information Science and Technology*, *66*(10), 2155–2159. Retrieved 2020-07-01, from <http://doi.wiley.com/10.1002/asi.23385> doi: 10.1002/asi.23385
- Scheidsteger, T., & Haunschild, R. (2023). Comparison of metadata with relevance for bibliometrics between Microsoft Academic Graph and OpenAlex until 2020. *El Profesional de la información*. Retrieved from <http://arxiv.org/abs/2206.14168> doi: 10.3145/epi.2023.mar.09
- Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the science citation index database. *Scientometrics*, *47*(3), 589–604. Retrieved from <https://link.springer.com/article/10.1023/A:1005628218890> doi: <https://doi.org/10.1023/A:1005628218890>
- Torre, S. (2012). Versioning in PubMed. *NLM Tech Bull*(384), e6. Retrieved from https://www.nlm.nih.gov/pubs/techbull/jf12/jf12_pm_versioning.html
- van Buskirk, N. E. (1984). The Review Article in MEDLINE: Ambiguity of Definition and Implications for Online Searchers. *Bull. Med. Libr. Assoc.*,

72(4), 349–352.

- van Raan, A. F. J. (2004). Measuring Science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 19–50). Dordrecht: Springer Netherlands. Retrieved 2020-07-07, from <http://link.springer.com/10.1007/1-4020-2755-9> doi: 10.1007/1-4020-2755-9
- 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. Retrieved from https://doi.org/10.1162/qss_a.00112 doi: 10.1162/qss_a.00112
- Waltman, L., & Larivière, V. (2020). Special issue on bibliographic data sources. *Quantitative Science Studies*, 1(1), 360–362. Retrieved from https://doi.org/10.1162/qss_e.00026 doi: 10.1162/qss_e.00026
- Xie, B., Shen, Z., & Wang, K. (2021). *Is preprint the future of science? A thirty year journey of online preprint services*. arXiv. Retrieved 2024-04-24, from <http://arxiv.org/abs/2102.09066> (arXiv:2102.09066 [cs])
- Yeung, A. W. K. (2019). Comparison between Scopus, Web of Science, PubMed and publishers for mislabelled review papers. *Current Science*, 116(11), 1909–1914. Retrieved from <https://www.jstor.org/stable/27138143>

7 Appendix

7.1 Mappings

The mappings used for matching document types to the categories *research discourse* and *editorial discourse* as mentioned in section 3.2 are listed here.

	Types
Research Discourse	article, journal-article
Editorial Discourse	erratum, editorial, letter, paratext
Not assigned	grant, book-chapter, dataset, book, other, reference-entry, dissertation, report, peer-review, standard, book-series

Table 6: OpenAlex Type Mapping

	Types
Research Discourse	Review, Article
Editorial Discourse	Erratum, Editorial, Letter, Note
Not assigned	Conference Paper, Chapter, Short Survey, Book, Tombstone, Data Paper, Article in Press, Conference Review, Abstract Report, Business Article, getItemType: unmatched: pp, Report

Table 7: Scopus Type Mapping

	Types
Research Discourse	Review, MetaAnalysis, JournalArticle, Study, CaseReport, Clinical-Trial
Editorial Discourse	Editorial, News, LettersAndComments
Not assigned	Conference, Book, Dataset

Table 8: Semantic Scholar Type Mapping

	Types
Research Discourse	Cochrane Systematic Review, Systematic Review, Meta-Analysis, Review, Case Reports, Randomized Controlled Trial, Clinical Trial, Clinical Trial, Phase II, Clinical Trial, Phase III, Clinical Trial, Phase I, Clinical Trial, Phase IV, Controlled Clinical Trial, Pragmatic Clinical Trial, Journal Article, Comparative Study, Multicenter Study, Observational Study, Evaluation Study, Historical Article, Validation Study, Clinical Study, Randomized Controlled Trial, Veterinary, Twin Study, Clinical Trial, Veterinary, Classical Article, Observational Study, Veterinary, Corrected and Republished Article, Adaptive Clinical Trial, Evaluation Studies, Validation Studies, "Research Support, Non-U.S. Govt", "Research Support, N.I.H., Extramural", "Research Support, U.S. Govt, Non-P.H.S.", "Research Support, U.S. Govt, P.H.S.", Preprint, "Research Support, N.I.H., Intramural", "Research Support, American Recovery and Reinvestment Act", Equivalence Trial
Editorial Discourse	Published Erratum, Retraction of Publication, Retracted Publication, Editorial, News, Letter, Comment, Introductory Journal Article, Newspaper Article
Not assigned	English Abstract, Video-Audio Media, Biography, Practice Guideline, Portrait, Congress, Clinical Trial Protocol, Interview, Personal Narrative, Consensus Development Conference, Overall, Patient Education Handout, Guideline, Dataset, Lecture, Address, Clinical Conference, Expression of Concern, Legal Case, Autobiography, Technical Report, Webcast, Bibliography, Festschrift, Consensus Development Conference, NIH, Interactive Tutorial, Scientific Integrity Review, Duplicate Publication, Directory, Periodical Index, Dictionary, Legislation

Table 9: PubMed Type Mapping

	Types
Research Discourse	Review, Article
Editorial Discourse	Correction, Retraction, Retracted Publication, Item Withdrawal, Editorial Material, News Item, Letter
Not assigned	Meeting Abstract, Book Review, Early Access, Biographical-Item, Book Chapter, Poetry, Reprint, Data Paper, Bibliography, Fiction, Creative Prose, Art Exhibit Review, Theater Review, Software Review, CC Meeting Heading, Record Review, Expression of Concern, Film Review, Music Performance Review, Music Score Review, TV Review, Radio Review, Excerpt, Database Review, Script, Hardware Review, Dance Performance Review, Book, Music Score, Chronology, Meeting Summary, Main Cite, Meeting-Abstract, Note, Proceedings Paper

Table 10: WoS Type Mapping

Rank	PubMed types
10 (Retractions)	Published Erratum; Retraction of Publication; Retracted Publication
11 (Editorials)	Editorial*
12 (Letters)	Letter*
13 (News)	News*
20 (Reviews)	Cochrane Systematic Review; Systematic Review; Meta-Analysis; Review
30 (Case reports)	Case Reports
31 (Clinical trial)	Clinical Trial, Phase IV; Pragmatic Clinical Trial; Controlled Clinical Trial; Randomized Controlled Trial; Clinical Trial; Clinical Trial, Phase II; Clinical Trial, Phase III; Clinical Trial, Phase I
50 (Articles)	Journal Article*
90 (Funding information)	Research Support, American Recovery and Reinvestment Act; Research Support, U.S. Gov't, Non-P.H.S.; Research Support, N.I.H., Intramural; Research Support, U.S. Gov't, P.H.S.; Research Support, Non-U.S. Gov't; Research Support, N.I.H., Extramural
99 (other)	...

*PubMed indexing rules define Editorials, Letters, News, and Articles as base types that can never co-occur

Table 11: PubMed Type Hierarchy