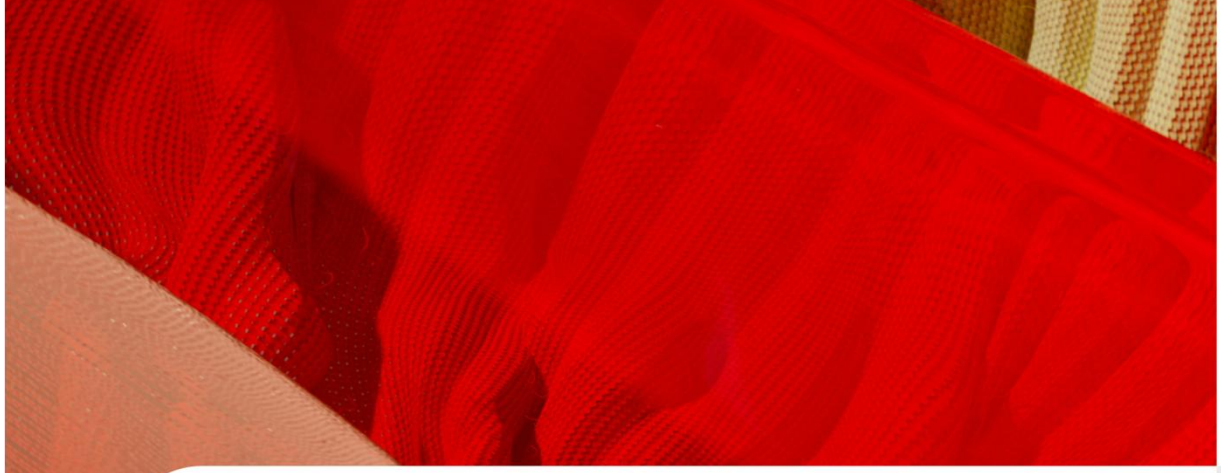

A Comparative Analysis of Open and Commercial Bibliographic Infrastructures: Scale, Metadata Standardization, and Implications for Bibliometric Evaluation

SCImago

scimagolab.com

Recommended citation:

De-Moya-Anegón, Félix; Sánchez-Jiménez, Rodrigo; Halevi, Gali; Guerrero-Bote, Vicente P.; Guerrero-Castillo, Pablo; Rivadeneyra, Federico (2026). *A Comparative Analysis of Open and Commercial Bibliographic Infrastructures: Scale, Metadata Standardization, and Implications for Bibliometric Evaluation*. Granada: Ediciones Profesionales de la Información, 48 pp. ISBN: 978-84-125757-8-1

Ediciones Profesionales de la Información SL
Av. Doctor Olóriz, 15, 4º
18012 Granada, España
Tel.: +34-639 878 489

First published: May 2026
ISBN: 978-84-125757-8-1
https://doi.org/10.3145/aca
https://www.scimagoepi.com/producto/a-comparative-analysis

Retail price: € 20

Team Composition

Team Director: Félix de Moya Anegón
https://orcid.org/0000-0002-0255-8628

Team Coordinator: Rodrigo Sánchez Jiménez
https://orcid.org/0000-0002-3685-7060

Research Team:

- **Gali Halevi**, Senior Scientific Advisor
  https://orcid.org/0000-0003-3478-8804
- **Vicente P. Guerrero Bote**, Senior Scientific Advisor
  https://orcid.org/0000-0003-4821-9768
- **Pablo Guerrero Castillo**, Computer Scientist
  https://orcid.org/0009-0000-9495-6553
- **Federico Rivadeneyra**, Computer Scientist

Proofreading:

- **Tomàs Baiget**, *Ediciones Profesionales de la Información SL*
  https://orcid.org/0000-0003-0041-2665

Contents

List of Figures and Tables

Figures

Tables

Table 17: Distribution of source types, difference between SJR sources and matching sources in OpenAIRE

Table 18: Top 20 countries by editorial output, comparison of sources and documents indexed in the free databases in the SJR source list.

Table 19: World regions by editorial output, comparison of sources and documents indexed in the free databases in the SJR source list.

Table 20: Top 20 publishers by published sources, comparison of sources and documents indexed in the free databases in the SJR source list.

Executive summary

This report evaluates the structural viability of open bibliographic infrastructures for research assessment purposes, with a particular focus on how leading open databases compare with Scopus in terms of coverage, metadata quality, transparency, interoperability, and suitability for research evaluation workflows.

While recent policy frameworks such as the Coalition for Advancing Research Assessment (CoARA) and the Barcelona Declaration mandate a transition toward open research data, an empirical analysis reveals a critical bottleneck: a structural trade-off between scale and metadata standardization. Platforms such as OpenAIRE, which aggregates more than 150 million records, and open bibliographic platforms including OpenAlex and The Lens, each with over 200 million records, significantly surpass the publication volume covered by commercial curated databases, most notably Scopus, across the analyzed 1996–2024 period.

However, this aggregation model prioritizes recall over structural consistency, which can lead to metadata gaps that compromise direct bibliometric application. The massive ingestion capabilities of open platforms are counterbalanced by substantial limitations in key metadata fields. Affiliation data are absent in more than 55% of records, severely constraining the feasibility of institutional evaluations, and key identifiers such as ISSNs and DOIs exhibit significantly lower levels of completeness than in Scopus. Document type classification also frequently lacks editorial rigor, relying heavily on algorithmic labeling that does not consistently standardize the categorization of scholarly outputs.

Furthermore, the analysis of citation flows reveals a markedly asymmetric dynamic: the expansive long tail of open databases functions primarily as a reference feeder that reinforces the impact indicators of the already established commercial core, rather than substantially redistributing measured impact across the broader scholarly corpus. In this way, the additional literature that open sources seek to incorporate ultimately serves to strengthen the prominence of the publications already represented in commercial databases. This finding points to a structural paradox in open scholarly infrastructures and raises important questions that warrant further reflection and investigation.

Geographic and editorial analyses reveal persistent asymmetries. Within the Global South, representation trajectories diverge: while regions such as Africa and Latin America have improved their visibility, significant coverage gaps, reaching up to 25%, remain in Asia and the Middle East. Additionally, deficits persist in specialized humanities monographs and complex publication structures like conference proceedings. Consequently, the theoretical advantage of the open "long tail" cannot currently be leveraged to offset these geographic and editorial biases, as its source-level metadata remains structurally incomplete or absent.

This operational friction stems from a fundamentally bifurcated data reality. Within the core literature that overlaps with Scopus, open infrastructures achieve high metadata completeness in fields essential for research evaluation. However, the extended literature outside this overlapping core suffers from profound structural deficiencies, including empty essential fields, duplication, and incomplete source data.

The corpus derived from Scopus's editorial processes exhibits a structural consistency without a direct equivalent in open platforms. While all databases utilize normalization methods, open

infrastructures depend intensively on algorithmic procedures which are notably prominent in OpenAlex. Conversely, Scopus integrates automated processes with author and institutional feedback to refine data disambiguation. Although the data indicates that Scopus captures a higher number of affiliations per document, this study does not include an empirical comparison regarding the effectiveness of their respective disambiguation systems.

Conversely, open platforms face significant structural trade-offs: The Lens struggles with global metadata standardization, reporting the lowest global rates of ISSN and DOI presence and a 71.67% deficit in capturing conference proceedings. OpenAlex relies heavily on unstructured source data, with 41.5% of its records (having a source) lacking an ISSN, and faces potential analytical bias due to algorithmic over-labeling of documents as "articles". Finally, OpenAIRE presents important technical anomalies, including over one million duplicated DOIs and the highest rate of unclassified documents (23.1%) within the curated core, resulting in the lowest overall citation impact ratio of the group.

Despite the structural limitations observed in their extended corpora, open bibliographic infrastructures present advantages when applied to targeted use cases. The Lens, with over 215 million records, integrates scholarly outputs with patent data, making it highly effective for mapping technology transfer while maintaining a 96.1% citable document density within its core overlap. OpenAlex demonstrates the highest absolute alignment with commercial standards by capturing 63.8 million Scopus-indexed records and highest citation density in that core among the three open databases. Finally, OpenAIRE offers the highest coverage of persistent identifiers (73.2% for DOIs and 59.7% for ISSNs) and the lowest rate of missing institutional affiliations (40.55%) among the open platforms.

The high structural availability of open data must not be uniformly equated with evaluative viability. Uncritical adoption of the full open dataset in its raw state risks introducing new, systemic biases into the global science policy landscape, imposing significant methodological compromises. Nevertheless, these infrastructures have evolved considerably. While direct aggregation currently complicates standard institutional evaluation, these platforms can deliver highly functional solutions for specialized bibliometric analyses, provided that institutions commit to investing in rigorous data normalization and disambiguation processes. Consequently, the transition toward open research assessment requires a technical shift from mere data accessibility to active data validation.

Introduction

In recent years, the emergence of large-scale open databases has been driven by a convergence of technological and structural factors. Primarily, advances in data infrastructure have made the massive algorithmic processing of bibliographic records feasible. This capacity has been fueled by the open provision of extensive datasets from initiatives such as PubMed and Crossref, alongside the foundational legacy of the Microsoft Academic Graph (**Priem** *et al.*, 2022; **Delgado-Quirós** & **Ortega**, 2024a**)**. Furthermore, the maturation of critical interoperability standards in scholarly communication—such as DOI, ORCID, and ROR, and the existence of established identifiers like the ISSN—has provided the necessary structural backbone for such databases (**Delgado-Quirós** & **Ortega**, 2024b). While the broader open science paradigm and the demand for alternative models of information access have acted as policy catalysts, it is fundamentally this technical integration that explains the current operational viability of open sources and the strategic expectations stakeholders place on their future trajectory.

The push for responsible research evaluation, institutionalized through frameworks such as the San Francisco Declaration on Research Assessment (*DORA*, 2012) and the Coalition for Advancing Research Assessment (*CoARA*, 2022), requires transparent and verifiable indicators. In parallel, the *Barcelona Declaration on Open Research Information* (2024) explicitly requires that the data underlying these assessments must be open by default. Consequently, the curation and utilization of open data sources have become instrumental to facilitate this policy shift (**Torres-Salinas** & **Arroyo-Machado**, 2026). Although several open infrastructure initiatives seem structurally consolidated and highly promising, significant operational shortcomings and technical limitations have also been documented (**Céspedes** *et al.*, 2025; **Haupka** *et al.*, 2024; **Zhang** *et al.*, 2024). Specifically, studies on databases like Dimensions have highlighted the high effort required for data processing and the risks of information loss during source integration (**Guerrero-Bote** *et al.*, 2021). Therefore, for institutions and policymakers seeking to implement these reforms, it is essential to systematically audit these open datasets to determine their fitness for purpose across different use scenarios (**Abramo** *et al.*, 2026; **Scheidsteger** *et al.*, 2025).

Traditional controlled databases such as Scopus continue to play a central role in scholarly research by providing curated datasets with high levels of standardization and quality control (**Birkle** *et al.*, 2020). Their selectivity means they capture only a fraction of global scientific output, which can introduce coverage gaps with respect to emerging fields, regional scholarship, and non-traditional formats (**Khanna** *et al.*, 2022; **Priem** *et al.*, 2022; **Visser** *et al.*, 2021). However, this restricted scope is also a consequence of editorial oversight, an operational mechanism designed to preserve metadata integrity and maintain consistent indexing standards across the collection (**Baas** *et al.*, 2020).

In contrast, open data sources are designed to be more inclusive, aiming to provide broader coverage across disciplines, geographies, and publication types **(Khanna** *et al.*, 2022; **Maddi** *et al.*, 2025). This approach supports a more comprehensive and democratic representation of scholarly communication, making a wider range of research outputs accessible to the global academic community. At the same time, inclusivity introduces methodological challenges related to metadata consistency, entity disambiguation, and citation graph integrity (**Céspedes** *et al.*, 2025; **Delgado-Quirós** & **Ortega**, 2024b; **Zhang** *et al.*, 2024).

The divergence between curated commercial databases and large-scale open infrastructures reflects a classic information retrieval trade-off between precision and recall. Traditional bibliographic databases prioritize precision through editorial control, standardization, and selective coverage, reducing noise but limiting exhaustiveness. In contrast, large open infrastructures prioritize recall by aggregating content at scale, increasing coverage but also introducing ambiguity and heterogeneity in metadata. This tension has direct implications for scientometric analysis, where results may vary significantly depending on the underlying data source and its architecture.

Commercial databases also impose access restrictions and often charge substantial subscription fees, limiting availability to well-resourced institutions. These constraints create barriers for researchers and organizations seeking to conduct transparent and reproducible analyses. In contrast, open infrastructures increasingly support data accessibility, reproducibility, and independent validation, allowing analysts to audit data provenance, reconstruct indicators, and perform large-scale analyses without dependency on proprietary systems (**Culbert** *et al.*, 2025).

Despite these advantages, important methodological challenges remain. Open infrastructures rely heavily on automated disambiguation and machine learning models to resolve authors' names, institutional affiliations, and citations. When these models fail, statistical noise can affect micro and meso-level indicators, particularly at the author and institutional levels **(Abramo** *et al.*, 2026; **Scheidsteger** *et al.*, 2025). At the same time, curated systems, while precise, may omit relevant outputs, leading to incomplete assessments.

The analysis presented in this report examines essential data fields required for bibliometric studies and research evaluation, alongside coverage and record overlap. The main conclusions are derived from a comparative study of four databases, encompassing over 650 million records published between 1996 and 2024 (inclusive). Furthermore, the report provides a comparative analysis based on selected sources restricted to the 2021–2024 period. This second part of the analysis highlights recent developments in coverage and citation data, ensuring the findings accurately reflect the current state of open scholarly infrastructure.

By identifying strengths, limitations, and areas for improvement, this report aims to provide actionable insights for teams working with open scholarly data, while also informing the decisions of research evaluation policymakers from a strictly technical standpoint. Understanding the trade-offs among coverage, precision, and reproducibility is essential for optimizing data selection, strengthening analytical workflows, and enhancing the overall effectiveness of research evaluation processes.

Open Data Sources Used in this Study

This study compares three prominent open datasets, OpenAIRE, OpenAlex, and The Lens against Scopus data. These platforms were selected because they provide large-scale scholarly metadata, support programmatic access, and enable citation-based analysis across multiple disciplines.

The Lens (https://www.lens.org) is an open platform that integrates scholarly literature with patent and innovation data. It aggregates content from other macro sources such as Microsoft Academic, PubMed, Crossref and OpenAlex. It also includes academic repositories, and other open data infrastructures, as well as patent offices, providing access to both research

publications and intellectual property records. This study only considers the former. By linking scholarly outputs with patent citations, The Lens enables analysis of the relationship between research and innovation. The platform includes hundreds of millions of scholarly works and a large corpus of patent records, offering a unique perspective on knowledge transfer and technological development. Its open access model and analytical tools support trend analysis, citation tracking, and exploration of research-to-innovation pathways.

OpenAlex (https://openalex.org) is a large-scale, open-access database of scholarly information developed by the nonprofit organization OurResearch. It aggregates metadata for hundreds of millions of scholarly works, including journal articles, books, datasets, and other research outputs. While originally built upon the legacy records of the Microsoft Academic Graph (MAG), OpenAlex currently derives its core data from Crossref and PubMed, structurally enriching these records with persistent identifiers such as ORCID and ROR. Furthermore, as of November 2025, the platform has deployed its own automated full-text parsing infrastructure to natively extract citations and metadata directly from open-access PDFs across the web. By combining this native extraction with automated entity linking, OpenAlex constructs a comprehensive global citation graph. Its architecture emphasizes openness, transparency, and reproducibility, making all structured data—spanning authors, institutions, and research topics—fully accessible via APIs and complete data dumps to support advanced filtering and analytics. The data snapshot utilized for this report was extracted immediately prior to the deployment of the new architecture; consequently, it does not reflect the aforementioned changes.

OpenAIRE (Open Access Infrastructure for Research in Europe, https://www.openaire.eu) is a major European initiative supporting open access to research outputs through a comprehensive and interoperable infrastructure. It aggregates metadata from a wide variety of sources, including institutional repositories, open-access journals, and data archives. It also systematically ingests data from global registries such as Crossref and Europe PMC (PubMed) to enrich, deduplicate, and structurally link repository records with definitive publication metadata and funding information. OpenAIRE connects publications, datasets, software, and research projects with funding information, creating a rich network of research outputs and contextual metadata. Its coverage spans thousands of repositories worldwide, ensuring broad geographic and disciplinary representation, although its analytical strength lies in funding traceability and its ability to monitor European open access policies. The platform is continuously updated, emphasizing discoverability, reuse, and transparency. It intends to play a key role in advancing open science and research accessibility.

These platforms were selected based on their ability to support bibliometric analysis. Key selection criteria included API availability and data extraction capabilities, citation graph integrity, entity disambiguation approaches, and longitudinal stability. These requirements are essential for constructing co-authorship networks, citation-based indicators, and comparative impact metrics.

Together, the selected platforms represent different approaches to open scholarly infrastructure, ranging from repository aggregation to large-scale citation graphs and research–innovation integration. Comparing these systems against curated data from a prestigious source provides insight into their readiness for research evaluation and their suitability for building reproducible, large-scale scientometric indicators.

Several alternative sources were excluded for methodological reasons. Disciplinary databases such as PubMed or IEEE Xplore lack multidisciplinary coverage, preventing field-normalized

comparisons. Academic social networks rely on self-archived content and unvalidated metrics, limiting their reliability for scientometric analysis. Microsoft Academic Graph was excluded due to its discontinuation in 2021, which prevents longitudinal analysis. Regional databases were also excluded as primary sources, since their geographic focus can bias global comparisons.

Methodology

Although all datasets were retrieved during the 2025 calendar year, the data snapshots were not obtained simultaneously. The Scopus dataset was finalized in March, OpenAIRE's full version dates from September, and both The Lens and OpenAlex were extracted in November. While this variance of up to eight months may introduce discrepancies due to retroactive indexing, the analysis maintains a consistent chronological scope across all sources, covering the period from 1996 to 2024. This ensures a uniform longitudinal framework for the examined bibliographic records.

This study relies in an exhaustive matching methodology, executed at scale across three massive open databases against the Scopus corpus. While reliance on the Digital Object Identifier (DOI) is the standard heuristic in bibliometric integration, the systemic absence of this identifier and the inherent structural imperfections within the source databases necessitated a more ambitious, multi-tiered approach. Consequently, a phased matching pipeline was designed, strictly prioritizing the most reliable metadata fields to systematically minimize false positives.

The protocol initiated with DOI matching. To prevent systemic misattributions, any records presenting duplicated DOIs within their source database are immediately excluded from the active pipeline and tallied separately. For records failing to match by DOI and provided they do not contain a duplicated DOI, the system attempted an exact title match. At this juncture, any unlinked documents sharing identical titles were deemed structurally ambiguous, flagged as un-linkable, and permanently discarded from the matching process. The final phase applied a rigorous normalization algorithm to the titles of the remaining un-linked records. This normalization process standardized the text by converting all characters to lowercase, replacing diacritics and accented letters with their base Latin equivalents, stripping all non-alphabetic characters, and reducing consecutive whitespaces to a single space. As in the previous phase, any records yielding duplicated normalized titles were classified as un-linkable and excluded.

This hierarchical methodology guaranteed a very high degree of confidence that the consolidated links genuinely correspond to the exact same document. However, the necessary tradeoff for this strict disambiguation standard is the generation of false negatives, which inherently inflates the volume of records classified as residing "outside the overlap" when a theoretical match might exist. A deliberate decision was made to abstain from applying secondary, independent deduplication algorithms to resolve these ambiguous cases. Remediating these structural conflicts falls strictly under the operational purview of the data providers. Consequently, the sheer volume of these discarded, ambiguous records acted as an implicit proxy for the maturity and quality of the databases' internal curation pipelines, even if not explicitly formulated as an evaluation metric in this study.

Results of the matching process

The analysis of the cross-referencing between these open databases and Scopus revealed distinct patterns regarding integration viability, highlighting both linking efficiencies and structural metadata deficiencies. The overwhelming majority of the validated overlap across all three systems relies heavily on the unique identifier. Specifically, 91% of the overlap in both OpenAlex (58.3 million out of 63.8 million) and The Lens (57.4 million out of 62.9 million) is driven by exact DOI matches. OpenAIRE exhibits a slightly lower, yet still dominant, dependency at 82% (48.6 million out of 59.3 million). Consequently, heuristic matching methods based on exact or normalized titles yield only residual performance for OpenAlex and The Lens. OpenAIRE stands as the exception, extracting substantial value from title normalization, with 10.6 million documents linked through this method.

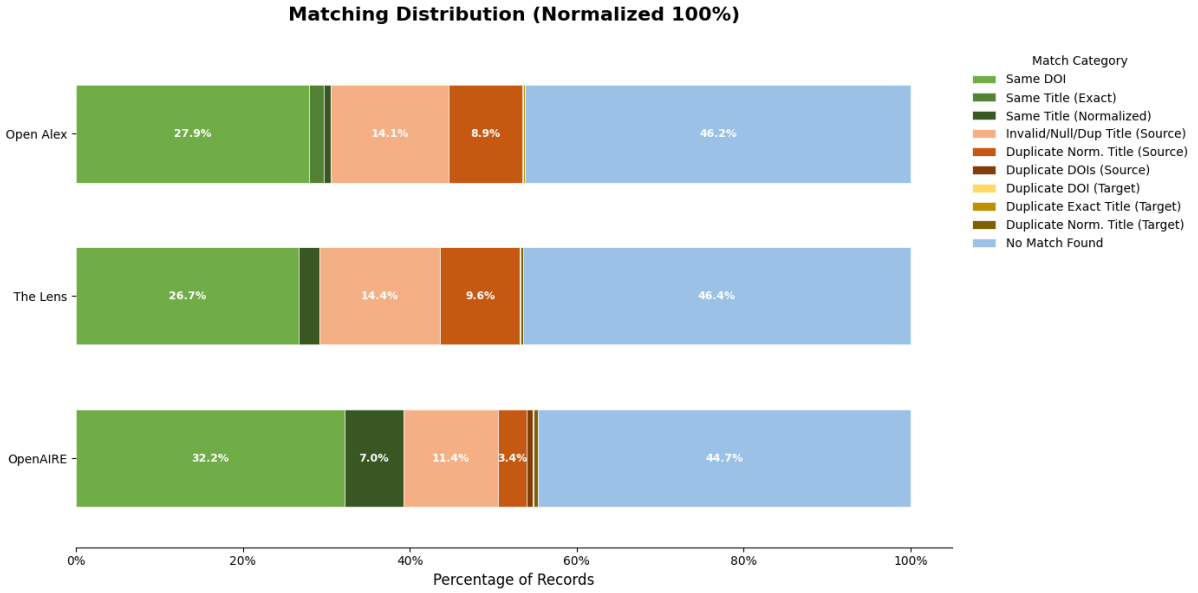

Figure 1: Matching Distribution Percentages

A closer review of the unlinked records reveals significant data quality issues, particularly within text fields such as titles, where a substantial number of documents are excluded due to ambiguity.

OpenAlex includes over 48 million records, and Lens over 51 million records, categorized under duplicated, null, or empty titles. This substantiates the premise that attempting large-scale bibliometric database cross-referencing relying on plain text fields is highly inefficient, leading to systemic rejections and rendering title-based heuristics fundamentally unreliable without extensive prior data cleaning.

Despite significant variations in the total volume of processed documents, the final overlap volume remains remarkably consistent across the three platforms. OpenAlex processes approximately 209 million documents to achieve the highest absolute overlap of 63.8 million. Lens, despite processing a larger total volume of 215 million records, yields a slightly lower overlap of 62.9 million, consequently generating the highest number of non-matching records, which

exceed 100 million. In contrast, OpenAIRE demonstrates higher proportional efficiency. From a notably smaller processing base of 151 million documents, it achieves a highly competitive overlap of 59.3 million. This suggests that OpenAIRE's coverage is either more strictly aligned with Scopus's indexing parameters, or its underlying metadata facilitates a more effective pairing process.

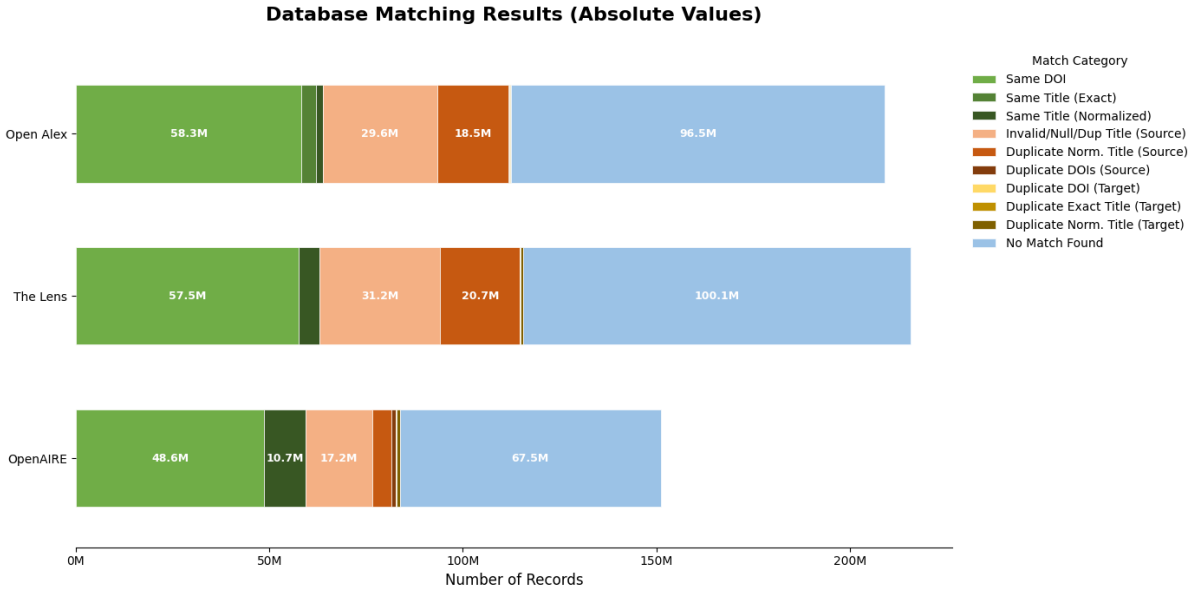

Figure 2: Databases matching results in absolute numbers

Finally, a structural anomaly within the OpenAIRE dataset necessitated specific scrutiny since the instances of duplicated DOIs reaches over one million documents. When juxtaposed with the negligible 18 instances in Lens and the roughly 79,000 in OpenAlex, this magnitude of duplicated identifiers in OpenAIRE could point to a flaw in their data ingestion pipelines or persistent identifier resolution mechanisms. This could be a problematic issue for any rigorous downstream bibliometric analysis and warrants further attention.

Matching the sources

The second part of the study used SCImago Journal & Country Rank (SJR) as a base line to enable a comparison between the four databases. To do so, there was a need to match the sources from each database to SJR's source list. The procedure uses ISSN as the first resource to establish a match between each source. This process is not perfect, though, as ISSN metadata was not always available. To expand the overlap between the databases, source titles were also used for matching, increasing the number of available sources by several hundred.

Size and overlap of the databases

There is a well-known and significant divergence in the scale of contemporary bibliographic databases, highlighting the structural differences between proprietary and open-source models. Traditional, commercial databases like Scopus operate on a strict curation model, which is

reflected in its more conservative volume of approximately 74.1 million documents. Its growth is constrained by editorial selection criteria and rigorous indexing standards.

In stark contrast, open-source aggregators show a massive scale. The Lens and OpenAlex lead with over 215 million and 209 million records, respectively, while OpenAIRE follows with approximately 151 million. This unprecedented volume is achieved through aggressive data aggregation from thousands of disparate sources, including institutional repositories, pre-print servers, and cross-reference registries.

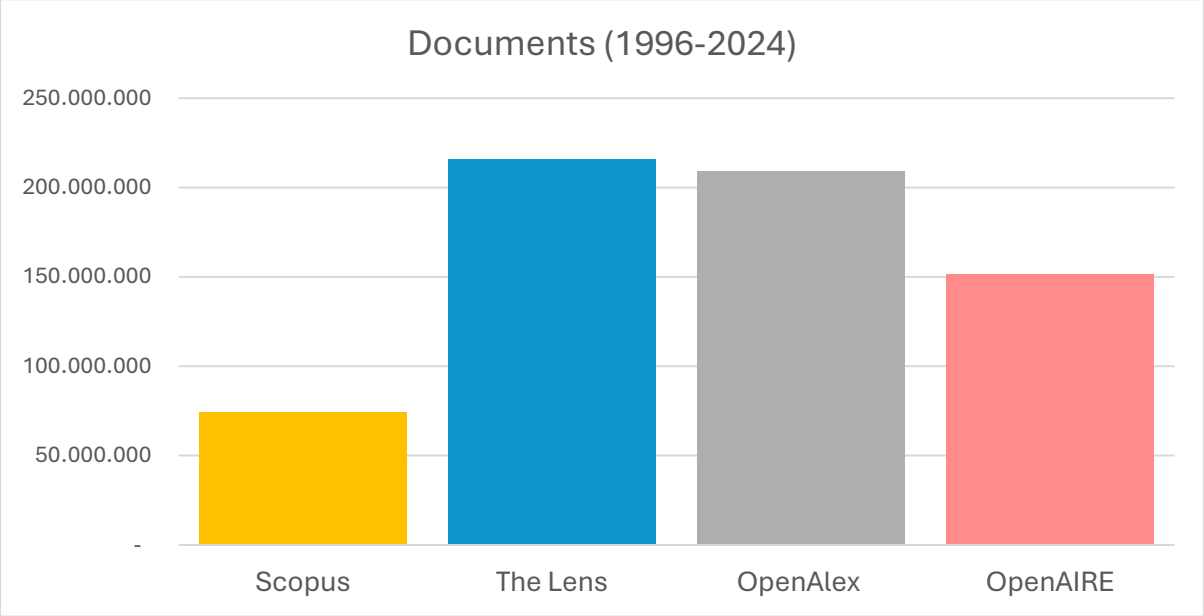

Figure 3: Number of documents indexed in the four databases from 1996-2024.

Using the procedure outlined in the methodology, we performed record matching. This operation yielded three distinct sets of records for each database. First, the overlap with Scopus, which enabled a record-by-record comparison between the open databases and the baseline of an established commercial database. Second, the exclusive portion of the open databases, comprising of content not indexed in Scopus. This segment reflects the potential to extend our view of scientific output beyond traditional database coverage, making it a valuable subject of study in its own right. Finally, a third, significantly smaller portion consists of records exclusively covered by Scopus, which the open platforms fail to capture for various reasons.

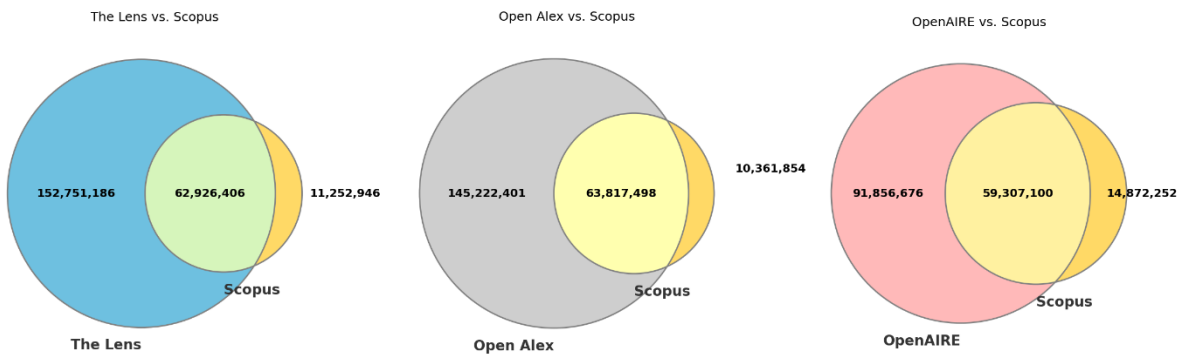

Figure 4: Coverage overlap between the four databases

Coverage overlap analysis showed that open aggregators largely encompass the content indexed in Scopus. Both The Lens and OpenAlex include approximately 85% of Scopus's total collection. However, Scopus still retains a substantial exclusive corpus of approximately 10–11 million documents that is not covered by these open platforms. The most striking feature of the coverage landscape is the vast volume of literature that is exclusive to the open platforms. For example, The Lens contains more than 150 million records outside its overlap with Scopus. Put differently, Scopus accounts for only about 30% of the total content indexed in The Lens and OpenAlex, and just under 40% of the content available through OpenAIRE. The composition and data quality of this additional content are of particular interest. However, our analysis takes a broader perspective by examining the three databases as complete collections, while also using their overlap with Scopus and the records extending beyond Scopus coverage to better understand how their coverage is structured.

Key metadata

Titles

Titles are a fundamental element of bibliographic records, central to information retrieval and the effective use of databases. They support both exploratory discovery (as tools for information seeking) and structured discovery (as a basis for systematic bibliographic analysis). Their presence is therefore essential, particularly for enabling document identification. This function becomes critical in supporting deduplication and record disambiguation, especially when standardized identifiers such as DOIs are absent.

Table 1: Percentage of records with a title and percentage with a unique title

Database	Total records	Records with title		Records with unique titles	
		n	%	n	%
Lens	215,677,592	214,441,361	99.43	184,504,018	85.55
OpenAlex	209,039,899	207,409,448	99.22	179,296,598	85.77
OpenAIRE	151,161,885	150,997,883	99.89	133,466,112	88.29
Scopus	74,179,352	74,179,352	100	72,376,782	97.57

The analysis of title data revealed a clear dichotomy in the integrity of textual metadata across open databases, heavily conditioned by their concurrent indexing in Scopus. While Scopus contains titles for 100% of its records and open databases also exhibit very high completion rates, a more stringent examination reveals a concerning pattern whereas the three open databases contain a significant number of records with duplicated titles. These duplicate titles typically arise from overly generic naming or from titles that have lost their subtitles. This is a common phenomenon in the open databases that Scopus, to a great extent, does not share, as it maintains a 97.57% overall uniqueness rate compared to the 85.5%, 85.7%, and 88.3% rates of the other databases.

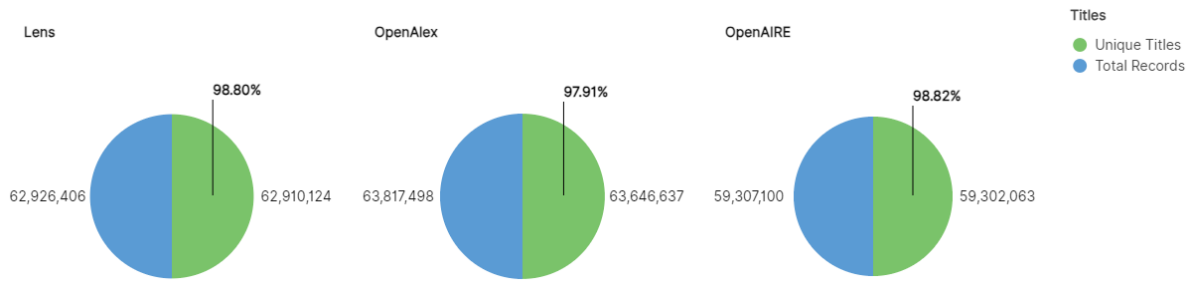

Figure 5: Unique titles of databases in their overlap with Scopus

Isolating the intersection with Scopus clarifies this pattern. Within the overlap region, the title uniqueness indicator increases significantly, achieving near-perfect disambiguation: The Lens reaches 98.8%, OpenAlex 97.3%, and OpenAIRE 98.8%. This indicates that the subset of open-source data overlapping with the commercial corpus consists of highly structured and curated literature.

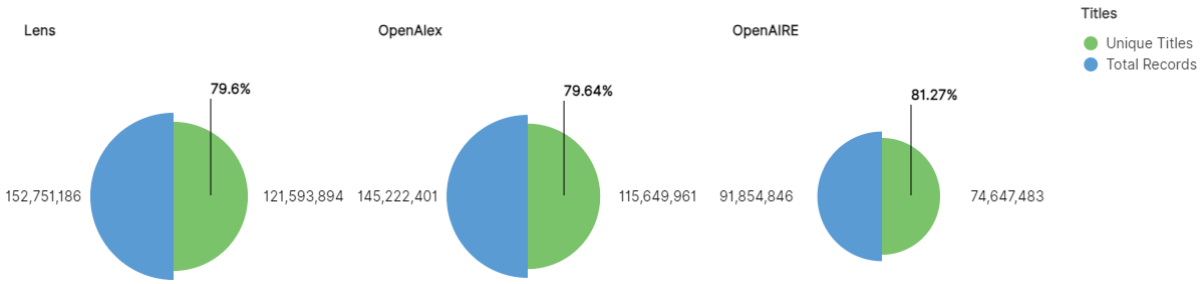

Figure 6: Unique titles of databases outside of their overlap with Scopus

Conversely, the data located outside the overlap exposes a problematic feature of open metadata infrastructures. Within the records exclusive to these platforms, the rate of unique titles drops significantly: The Lens falls to 79.60%, OpenAlex to 79.64%, and OpenAIRE to 81.27%. This metric indicates that approximately 20% of titles are non-unique, affecting around 30 million records in The Lens and OpenAlex, and roughly 17 million in OpenAIRE. While this pattern may point to unresolved internal deduplication issues, it could also result from generic titles or truncated metadata. In any case, the discrepancy highlights significant structural differences across the various data regions within open databases.

DOIs comparison

In the architecture of modern bibliometric databases, the Digital Object Identifier (DOI) functions as the main mechanism for unambiguous entity resolution. The presence of this identifier is widely recognized for imparting critical merits to the scholarly record, such as guaranteeing the long-term discoverability of documents, enabling the accurate traceability of citation networks, and establishing a reliable foundation for impact metrics. Unlike text-based metadata fields, which are inherently susceptible to typographic variations and formatting inconsistencies, the DOI provides a persistent, standardized anchor. Furthermore, DOIs are a crucial tool for database integration and interoperability; its absence is therefore problematic, rendering these processes a more complex and error-prone endeavor.

Table 2: Percentage of records with a DOI in the four databases

Database	Records with DOI	Records without DOI	Total records	Records with DOI %
Lens	86,705,991	128,971,601	215,677,592	59.8
OpenAlex	72,618,862	136,421,035	209,039,897	65.3
OpenAIRE	40,503,436	110,660,340	151,163,776	73.2
Scopus	13,492,911	60,686,441	74,179,352	81.8

The global overview shows an inverse relationship between total corpus size and the %age of records containing a DOI. Scopus shows the highest DOI coverage at 81.8%. Among the open databases, OpenAIRE achieves 73.2% coverage across its 151 million records. In contrast, OpenAlex and The Lens, each containing more than 200 million documents, report lower DOI coverage rates of 65.3% and 59.8%, respectively.

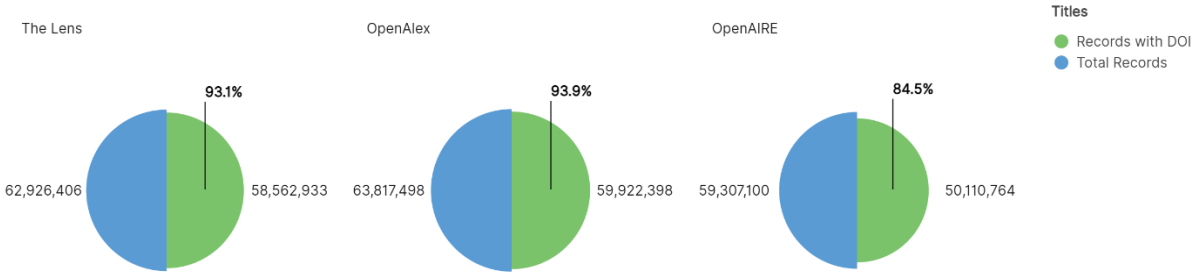

Figure 7: Percentage of records with a DOI for the databases in their overlap with Scopus

Analyzing the data portion that intersects with Scopus reveals a higher concentration of DOIs. Within this segment, The Lens and OpenAlex reach coverages of 93.1% and 93.9%, respectively. This increase of roughly 30 %age points relative to their global averages suggests that the subset of open data overlapping with the commercial index is composed primarily of standardized literature.%. OpenAIRE records a DOI presence of 84.5% in this region; while higher than its overall average, this figure aligns with the DOI metadata inconsistencies noted in previous cross-referencing data.

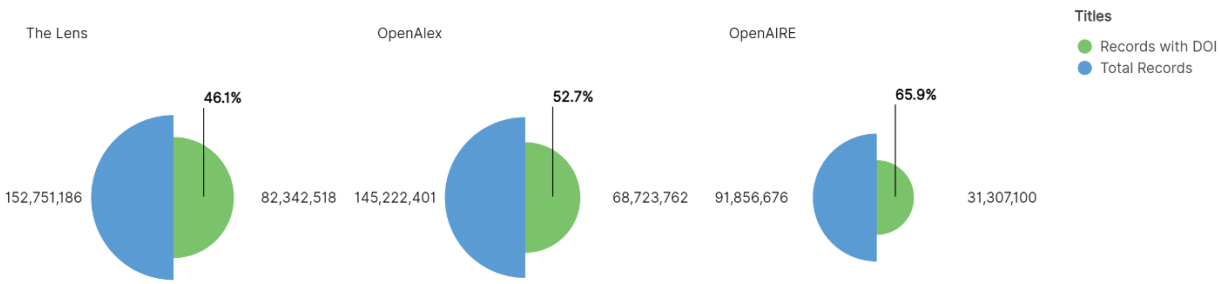

Figure 8: Percentage of records with a DOI for the databases outside of their overlap with Scopus

The portion of data exclusive to the open databases shows a significant decline in DOI coverage. For records outside the Scopus overlap, DOI presence in The Lens and OpenAlex drops to 46.1% and 52.7%, respectively. This means that a substantial portion of their exclusive volume, amounting to approximately 82 and 68 million records, lacks this identifier. OpenAIRE, conversely, exhibits a more consistent DOI distribution across its dataset, maintaining a 65.9% coverage in its exclusive segment. Nevertheless, the general reduction of DOI presence in this non curated literature, combined with the title duplication rates identified earlier, poses a structural challenge for the algorithmic deduplication and analysis of this specific subset.

ISSN

In bibliometric databases, the International Standard Serial Number (ISSN) serves as the primary mechanism for unambiguously identifying serials and publication venues. It enables accurate source aggregation, journal-level metric calculations, and longitudinal tracking, providing a stable alternative to text-based journal titles, which are susceptible to inconsistent abbreviations and variations. In the absence of an ISSN, data integration relies on string-matching algorithms applied to unstructured names. This approach introduces systemic noise due to homonymy and unrecorded title changes, compromising source-level aggregation. Therefore, the ISSN is an infrastructural prerequisite for reliable cross-referencing and rigorous macro-level bibliometric analysis.

Table 3: Percentage of records with an ISSN in the four databases

Database	Records no source		Records no ISSN		Records with ISSN		Total records
	n	%	n	%	n	%	
The Lens	72,946,773	33.8	41,768,689	19.4	100,962,130	46.8	215,677,592
OpenAlex	42,400,107	20.3	86,752,733	41.5	79,887,057	38.2	209,039,897
OpenAIRE	52,501,716	34.7	8,461,854	5.6	90,198,315	59.7	151,161,885
Scopus	0	0.0	9,425,096	12.7	64,754,256	87.3	74,179,352

While the absence of a source or an ISSN is perfectly legitimate for certain typologies, such as datasets, standalone reports, or books, standardized serial venues are the basic means for scientific knowledge dissemination. In this context, Scopus establishes the baseline for a highly curated scholarly corpus, exhibiting complete source attribution and an 87.3% ISSN coverage. The open platforms, however, struggle to balance their expansive scale with structural standardization. OpenAIRE achieves the highest ISSN presence among the open alternatives (59.7%) but carries a substantial 34.7% of records without any source affiliation, indicating a massive influx of isolated scholarly artifacts. OpenAlex displays an opposing imbalance: it holds the lowest rate of missing sources (20.3%) but relies heavily on unstructured data, with 41.5% of its records containing a source string that lacks an ISSN. Lens exhibits a similarly fragmented landscape, with only 46.8% of its total corpus linked to an ISSN.

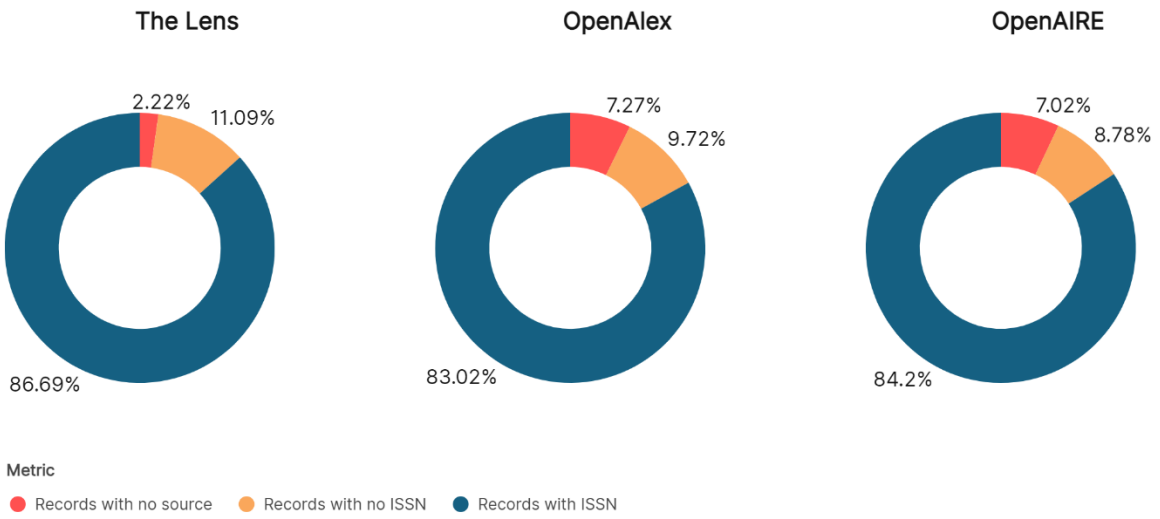

Figure 9: Percentage of records with any source information and ISSN information for the databases in their overlap with Scopus

Analyzing the data intersecting with Scopus reveals a substantial increase in metadata completeness. Within this overlap region, ISSN presence rises across the three platforms: The Lens reaches 86.7%, OpenAIRE 84.2%, and OpenAlex 83.0%. Concurrently, the proportion of records lacking source information decreases, ranging from 2.2% in The Lens to 7.3% in OpenAlex. These figures indicate that the subset of open data overlapping with Scopus corresponds primarily to structured serial literature. Consequently, the overlap isolates a segment of the open databases characterized by higher metadata standardization compared to their broader collections.

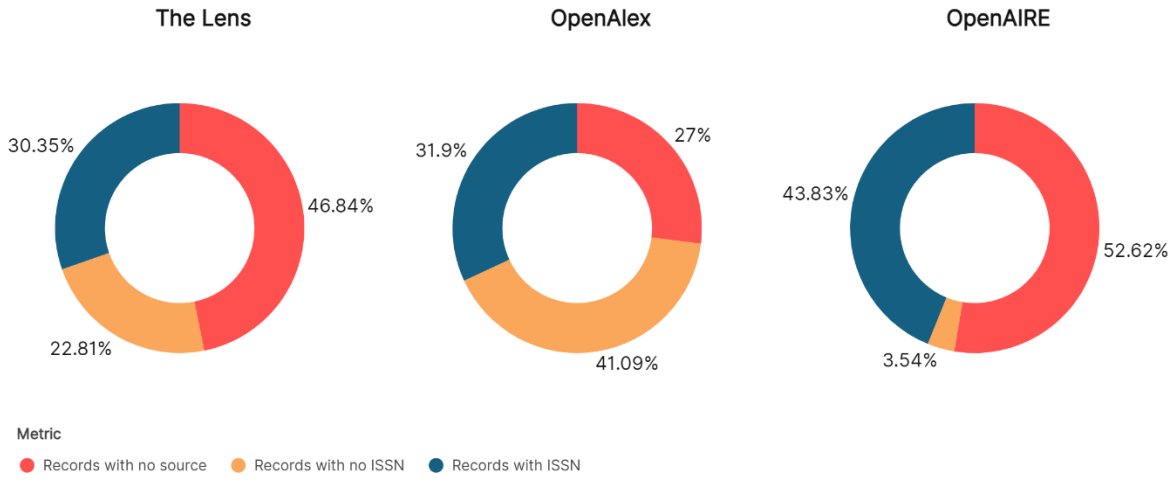

Figure 10: Percentage of records with any source information and ISSN information for the databases outside of their overlap with Scopus

The portion of data exclusive to the open databases exhibits lower rates of source attribution. Outside the Scopus overlap, ISSN presence decreases to 30.3% in The Lens, 31.9% in

OpenAlex, and 43.8% in OpenAIRE. Concurrently, the proportion of records lacking any source information increases, reaching 46.8% in The Lens and 52.6% in OpenAIRE. This indicates that tens of millions of documents within this segment, such as pre-prints, working papers, or repository deposits, lack both a defined publication venue and a standardized serial identifier. This absence of source-level metadata limits the feasibility of conducting venue-based bibliometric aggregation and evaluation within the non curated domains of these platforms.

Document types

Correctly classifying document types, including distinguishing citable research such as articles and reviews from non-citable material like editorials or abstracts, is essential for reliable bibliometric evaluation. Missing or incorrect metadata compromises citation metrics by distorting the denominators in impact calculations. In large open infrastructures, this poses methodological risks. While documents indexed in established commercial databases usually retain standardized categorization, the gray literature and repository deposits common in open databases often lack reliable document type tags. Consequently, conducting comparative assessments across these open corpora without filtering by document type can lead to skewed conclusions, as inadequate categorization limits the utility of the raw data for research assessment and policy formulation.

Comparing document type distributions reveals a lack of standardization across open infrastructures, which hinders comparative analysis. While Scopus uses a taxonomy focused on traditional scholarly outputs, with articles at 65.9% and conference papers at 15.4%, open databases show varied classification frameworks. OpenAIRE includes granular subsets of gray literature, separately indexing master's, bachelor's, and doctoral theses, which complicates macro-level aggregation. Meanwhile, The Lens and OpenAlex include categories like datasets, preprints, paratexts, and libguides. This inconsistency means harmonizing these datasets requires subjective mapping, introducing methodological biases as definitions of document types vary between platforms.

Citable documents can be mapped using a set of document types that have been marked in bold type in the table below. Scopus citable documents would include articles, conference papers, reviews, book chapters and short surveys. The Lens citable documents could be mapped to journal articles, book chapters, conference proceedings articles and reviews. OpenAlex citable types include articles (that subsume conference proceedings articles), book chapters and reviews. OpenAIRE citable types include articles, part of book or chapter of book, conference objects and reviews.

Table 4: Main distribution of the main document types in the four databases. Citable types in bold.

The Lens		OpenAlex		OpenAIRE		Scopus	
Doctype	**%**	**Doctype**	**%**	**Doctype**	**%**	**Doctype**	**%**
Journal article	47.1	**Article**	72.6	**Article**	55.9	**Article**	65.9
Book chapter	8.7	**Book chapter**	9.2	**Part of book or chapter of book**	10.8	**Conference paper**	15.4
Dataset	3.4	Dataset	3.6	Other literature type	3.4	**Review**	5.9
Conference proceedings article	3.4	Preprint	3.1	**Conference object**	2.5	**Book chapter**	4.7
Dissertation	2.6	Dissertation	2.7	Book	1.8	Note	2.2
Book	2.2	Book	1.7	Master thesis	1.5	Editorial	2.1
Preprint	2.0	**Review**	1.7	Doctoral thesis	1.1	Letter	1.6
Libguide	0.8	Paratext	1.5	**Review**	0.9	**Short survey**	0.8
Other	0.8	Libguides	0.9	Bachelor thesis	0.7	Book	0.8
Journal issue	0.5	Other	0.8	Thesis	0.6	Conference review	0.5
Reference entry	0.4	Letter	0.6	Report	0.6	Data paper	0.2
Conference proceedings	0.4	Reference-entry	0.4	Research	0.3	Report	0.0
Report	0.3	Peer-review	0.3	Preprint	0.2	Article in press	0.0
Review	0.3	Editorial	0.3	Presentation	0.1	Abstract Report	0.0
Other	0.6	Other	0.7	Other	0.3	Business article	0.0
Null	83.4	Null	0	Null	19.1	Null	0.0
Total	100	Total	100	Total	100	Total	100

The distribution of document types across the four databases highlights differences in metadata curation between Scopus and open infrastructures. Scopus shows a high level of editorial control, with an unclassified rate of 0.001% and 92.6% of its documents categorized as citable. In contrast, The Lens and OpenAIRE lack document type labels for 26.6% (57.2 million records) and 19.1% (28.9 million records) of their data, respectively. This volume of missing metadata complicates macro-level filtering and impact normalization, making it difficult to accurately

isolate citable outputs. To address this issue, OpenAlex employs an algorithmic approach, assigning document types automatically.

Table 5: Global number and percentages of records with citable types, other types and no type assigned of the four databases

Database	Citable documents		Other documents		Unclassified	
	n	%	n	%	n	%
Lens	128,343,972	59.5	30,056,711	13.9	57,276,909	26.6
OpenAlex	174,480,972	83.5	34,558,925	16.5	0	0
OpenAIRE	105,989,818	70.1	16,245,860	10.7	28,926,207	19.1
Scopus	68,697,709	92.6	5,480,798	7.4	845	0.001

Table 6: Global percentages of records with citable types and other types, total number of records with a document type assigned in the four databases

Database	% Citable documents	% Other documents	Total (not null)
The Lens	81.0	19.0	158,400,683
OpenAlex	83.5	16.5	209,039,897
OpenAIRE	86.7	13.3	122,235,678
Scopus	92.6	7.4	74,178,507

By excluding unclassified records from the document type distribution analysis, the comparison focuses only on the curated portions of each database. Within these subsets, Scopus retains the highest density of citable literature (92.6%), while open infrastructures show slightly lower proportions, ranging from 81.0% in The Lens to 86.7% in OpenAIRE. It is noteworthy that OpenAlex maintains a similar proportion despite its heavy reliance on automated classification. It is plausible that missing data is not random, and that unclassified records are more likely to be non curated gray literature or repository deposits rather than standardized, peer-reviewed articles. This creates a difficult trade-off, as many impact indicators depend on the distinction between citable and non-citable documents. Relying on algorithmic classification remains problematic; recent research indicates that OpenAlex over-labels its records as 'articles', leading to important discrepancy rates when compared to manual verification (Haupka *et al.*, 2026; Mongeon *et al.*, 2025).

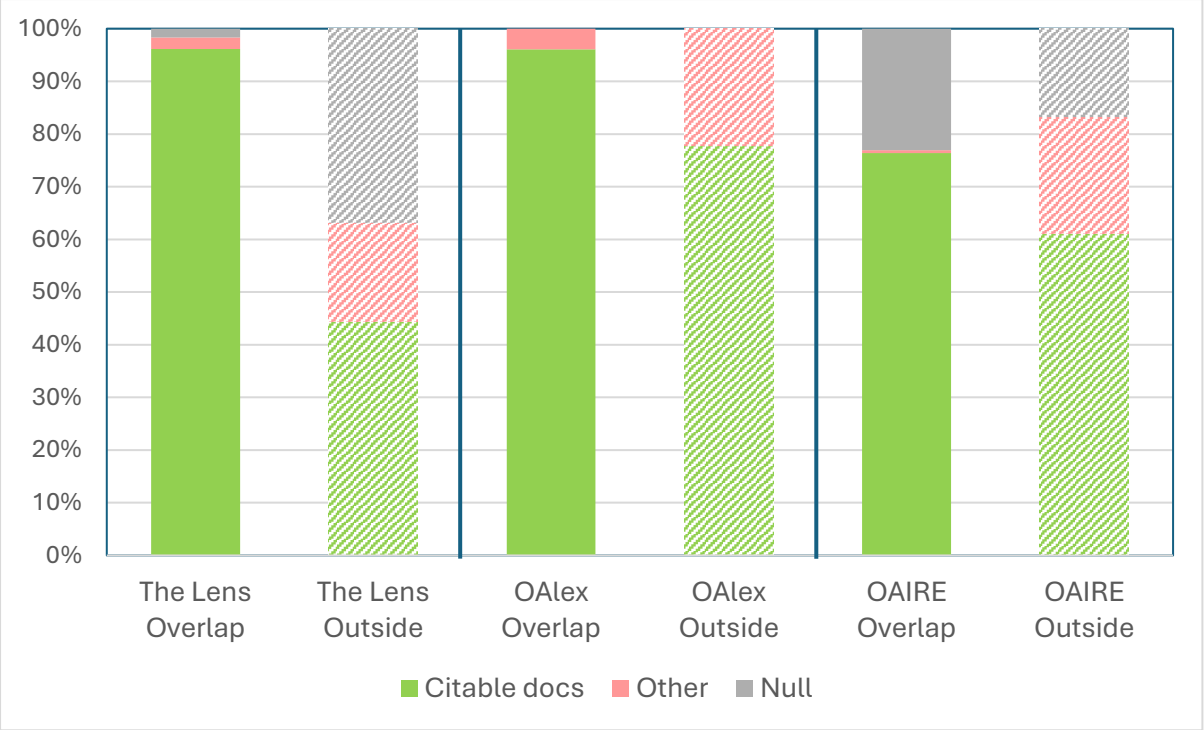

Figure 11: Percentage of citable documents, other types and null document types for the three databases, inside and outside of the overlaps

Cross-tabulating document types by overlap status highlights the differences in metadata curation between commercial proxies and open infrastructures, revealing specific anomalies within each database. The Lens illustrates this curation gradient: its overlapping subset shows a 96.1% density of citable documents and a 1.7% null rate, whereas its exclusive, non-overlapping literature presents a 36.8% null rate and only 44.4% citable outputs. Conversely, OpenAlex maintains a 0% null rate across both overlapping and exclusive regions, resulting in a 77.9% citable density in its exclusive corpus. OpenAIRE shows a 23.1% null rate for records shared with Scopus, which exceeds the 16.6% null rate found in its exclusive region. This high null rate within the overlapping subset might indicate internal processing or ingestion issues, affecting OpenAIRE's metadata even for core scholarly literature.

Document types related to source information

Analyzing the intersection between document typologies and source metadata provides valuable insights into maturity of these databases for large-scale bibliometric applications. Because evaluative metrics traditionally rely on a consistent and comparable delineation of "citable" literature, exploring how these typological categories interact with the presence or absence of standard identifiers helps contextualize structural discrepancies across platforms.

Table 7: Global distribution of document typologies relative to source attribution and ISSN presence in the four databases

Doctype The Lens	% of Total	% Source ISSN	% Source no ISSN	% No source
Citable documents	59,5	97,5	68,6	1,8
Other	13,9	2,4	28,6	21,5
Null	26,6	0,1	2,9	76,7
Doctype OpenAlex	**% of Total**	**% Source ISSN**	**% Source no ISSN**	**% No source**
Citable documents	83,5	95,3	76,1	67,9
Other	16,5	4,7	23,9	32,1
Null	0,0	0,0	0,0	0,0
Doctype OpenAIRE	**% of Total**	**% Source ISSN**	**% Source no ISSN**	**% No source**
Citable documents	70,1	79,2	72,7	54,1
Other	10,7	0,2	0,1	30,6
Null	19,1	20,6	27,2	15,3
Doctype Scopus	**% of Total**	**% Source ISSN**	**% Source no ISSN**	**% No source**
Citable documents	92,6	93,5	92,1	0,9
Other	7,4	6,5	7,9	99,1
Null	0,0	0,0	0,0	0,0

The cross-tabulation of document types with source metadata reveals a highly consistent structural alignment at the top tier of data quality. Across all platforms, the presence of a standardized serial identifier (ISSN) acts as a proxy for core scholarly literature. In Scopus, Lens, and OpenAlex, over 93% of the records possessing an ISSN fall into the "citable documents" category. OpenAIRE demonstrates a lower concentration at 79.2%, primarily due to an internal mapping discrepancy where 20.6% of its ISSN-bearing records are left with a 'Null' document type.

Conversely, the distribution of records entirely lacking source attribution ("No source") exposes divergences in how the open platforms handle classification for incomplete metadata. The Lens seems to adopt a cautious heuristic: the absence of a source heavily correlates with an unclassified status, with 76.7% of its source-less documents assigned a 'Null' document type. In contrast, OpenAlex and OpenAIRE use more inclusive classification models, assigning 67.9% and 54.1% of their source-less records, respectively, to citable types. This highlights a significant analytical caveat: within these two open platforms, document types that are typically considered as citable are broadly applied to tens of millions of orphaned artifacts regardless of their complete isolation from any identifiable publication venue.

Affiliations

Metadata on institutional affiliations is essential for valid bibliometric assessment, providing the necessary spatial and institutional context to map knowledge production. This data is the primary mechanism for measuring research and funding impact, conducting geopolitical analyses of science, and establishing global institutional rankings. When affiliation variables contain significant gaps, structural biases can be introduced into the database. Without comprehensive institutional linking, datasets lose utility for rigorous science policy evaluation or institutional benchmarking.

Table 8: Global affiliations per record and percentage of records without affiliations

Database	Number of affiliations	Avg. affiliations per record	Records without affiliation	% records without affiliation	Total records
The Lens	190,165,109	0.88	121,855,485	56.50	215,677,592
OpenAlex	181,323,018	0.87	116,799,521	55.87	209,039,899
OpenAIRE	223,237,420	1.48	61,303,477	40.55	151,163,776
Scopus	156,046,558	2.10	8,666,731	11.68	74,179,352

Examining the global datasets reveals a significant gap in affiliation metadata across open infrastructures compared to the commercial index. Scopus establishes a curated baseline, averaging 2.1 affiliations per record, with only 11.68% of its 74.1 million records lacking institutional links. In contrast, the larger open databases show substantial amount of missing data. The Lens and OpenAlex, despite containing over 200 million records each, lack affiliation data for 56.50% and 55.87% of their respective corpora, with averages of 0.88 and 0.87 affiliations per record. While OpenAIRE shows a lower rate of missing affiliations (40.55%) and a higher average (1.48), it remains below the Scopus baseline. This suggests that the large-scale ingestion processes characteristic of open platforms often result in lower completeness of structural metadata.

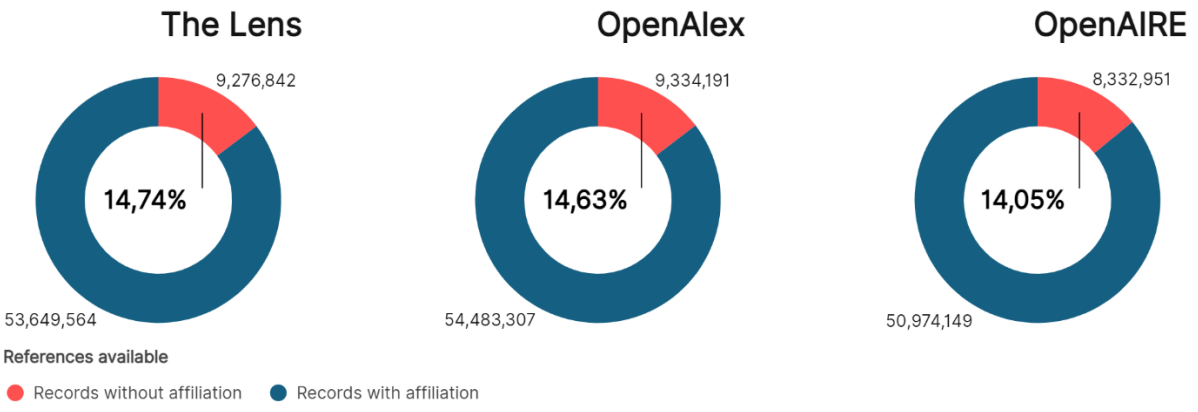

Figure 12: Percentage of records without affiliation in the overlap of the three databases with Scopus

When focusing only on the literature shared with Scopus, a clear shift becomes apparent. In this overlapping subset, many of the structural limitations found in the open databases are substantially reduced. The proportion of records entirely lacking affiliations drops drastically to 14.74% in Lens, 14.63% in OpenAlex, and 14.05% in OpenAIRE, aligning closely with the Scopus global baseline. This suggests that the overlapping data segment approaches commercial levels in terms of metadata availability.

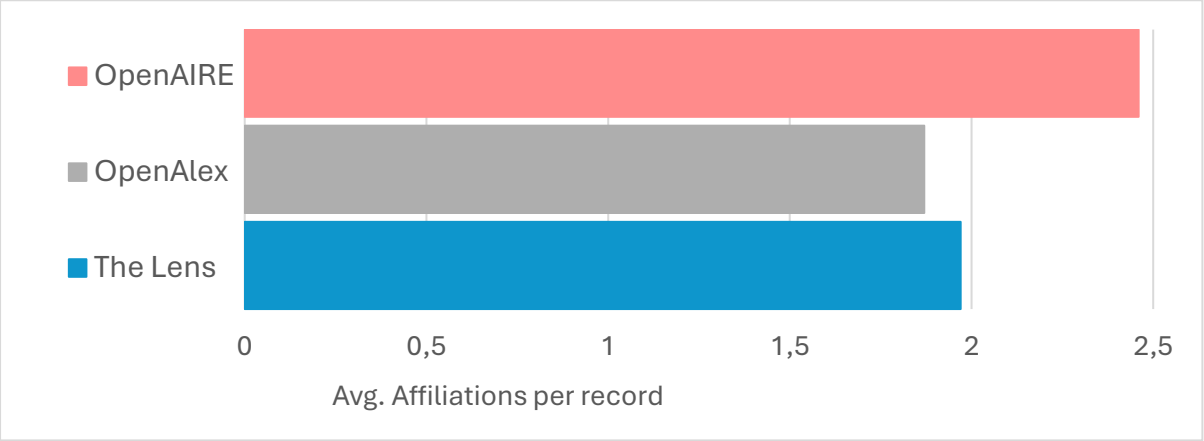

Figure 13: Average number of affiliations per record in the overlap of the three databases with Scopus

Furthermore, the average number of affiliations per record is higher in this subset: 1.97 in The Lens, 1.87 in OpenAlex, and 2.46 in OpenAIRE. These results suggest that, within the segment aligned with Scopus coverage, the available metadata may be sufficient to support analysis and evaluation tasks. However, since this study does not examine the quality of normalization and disambiguation across platforms, this potential analytical usefulness should be validated using additional methods.

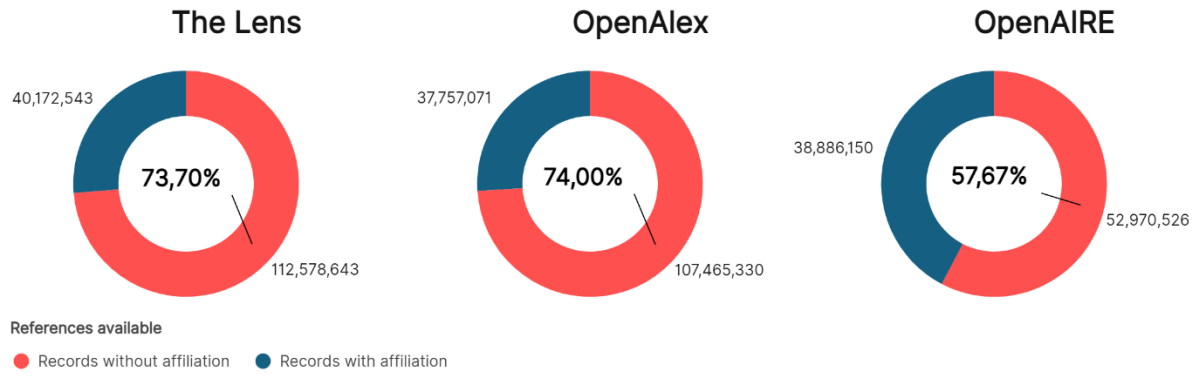

Figure 14: Percentage of records without affiliation outside of the overlap of the three databases with Scopus

An analysis of the exclusive long tail (the literature residing entirely outside the overlap) shows a significant decrease in metadata completeness within open infrastructures. In this domain, missing data is highly prevalent. The Lens and OpenAlex lack affiliation data for 73.70% and 74.00% of their exclusive records, respectively, with the average number of affiliations falling to 0.43 per record.

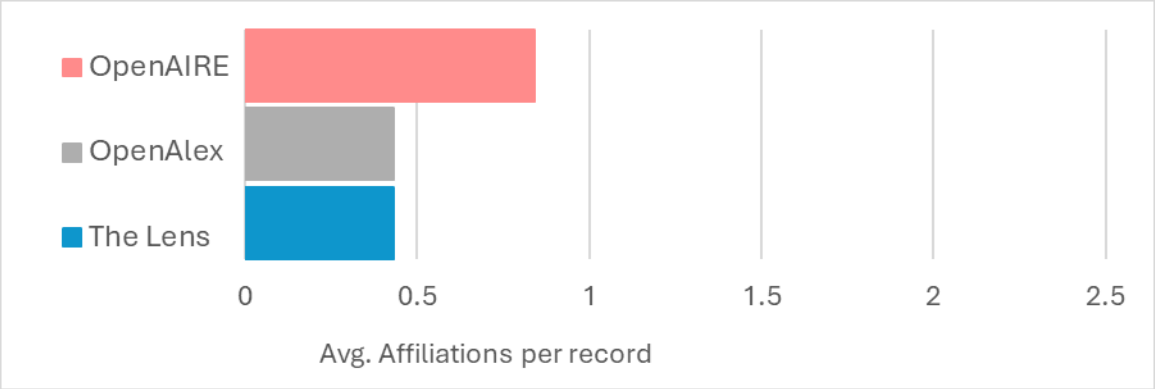

Figure 15: Average number of affiliations per record outside of the overlap of the three databases with Scopus

OpenAIRE shows a similar pattern, although it exhibits slightly higher completeness, with 57.67% of records lacking affiliations and an average of 0.84 per record. Since the matching protocol deliberately excludes structurally ambiguous records, this lower completeness reflects the limitations of open ecosystems in curating non-commercial content. Consequently, the high proportion of incomplete metadata in this non-overlapping region limits its utility for institutional tracking without extensive external remediation.

Citations

Citation metadata serves as a core metric in evaluative bibliometrics, providing the empirical basis for measuring research impact, tracking knowledge diffusion, and generating performance indicators such as the h-index or field-weighted citation metrics. When a bibliographic database lacks comprehensive citation linkages, its reliability for research assessment decreases. Missing or poorly extracted citation data underrepresents scientific impact, introducing biases that can affect comparative evaluations and funding allocation models.

Table 9: Global number of references, percentages of records without references and average number of references per record of the four databases

Database	Number of citations	Avg. citations per record	Records without citations		Total records
			n	%	
The Lens	1,843,656,358	8.55	139,628,271	64.74	215,677,592
OpenAlex	2,037,287,209	9.75	126,795,392	60.66	209,039,899
OpenAIRE	1,572,159,104	10.40	89,503,964	59.21	151,163,776
Scopus	1,465,727,109	19.76	19,288,076	26.00	74,179,352

Examination of the datasets reveals lower citation density across open infrastructures compared to the commercial index. Scopus establishes a baseline, averaging 19.76 citations per record, with 26.00% of its corpus lacking citation data. In contrast, open databases show a higher proportion of missing citation data. The Lens, OpenAlex, and OpenAIRE lack citations in 64.74%, 60.66%, and 59.21% of their respective global corpora. Consequently, their global average citations per record are lower, ranging between 8.55 and 10.40. This suggests that the large-scale

ingestion processes of open platforms include a high volume of isolated or poorly parsed documents, making their global aggregates less suitable for macroscopic impact assessment without prior filtering.

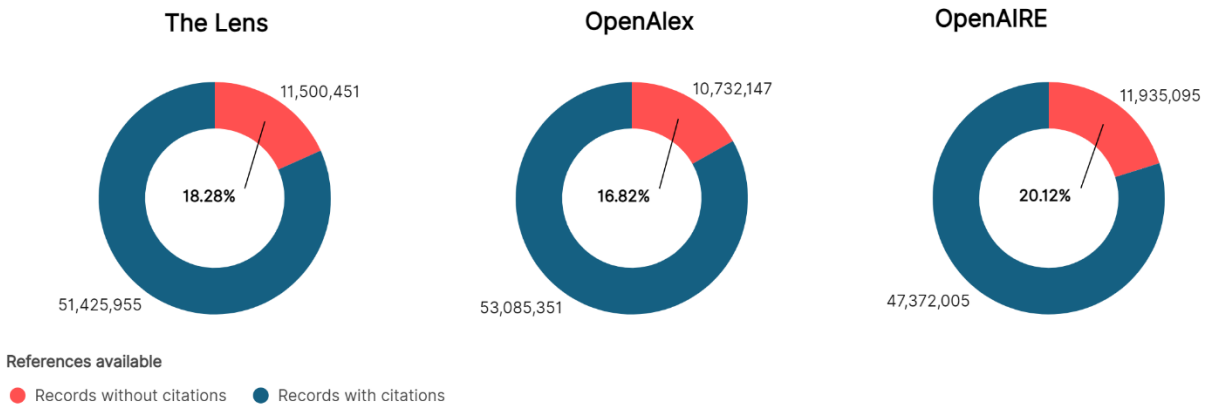

Figure 16: Number of records with citations and percentage of records without citations in the overlap with Scopus

Isolating the literature that overlaps with Scopus shows increased citation completeness, suggesting this subset provides sufficient density for reliable citation analysis. Within this intersection, the proportion of records lacking citations drops to 18.28% in The Lens, 16.82% in OpenAlex, and 20.12% in OpenAIRE. Furthermore, the average citations per record in this overlapping subset are substantially higher, reaching 25.97 in The Lens, 27.31 in OpenAlex, and 23.32 in OpenAIRE. These values exceed both their respective global averages (8.55, 9.75, and 10.40) and the global average reported by Scopus. The rates of missing data in the overlapping portions are also much lower than in the global figures provided before, ranging somewhere between 18% and 20%. These figures highlight a significant structural dynamic: open databases accumulate a disproportionate share of citations directed toward the core literature they share with Scopus. While the literature exclusive to these open platforms provides references that nourish this central core, it receives a minimal share of citations in return. Consequently, the expanded coverage of open databases primarily serves to amplify the impact indicators of the research already indexed in commercial databases.

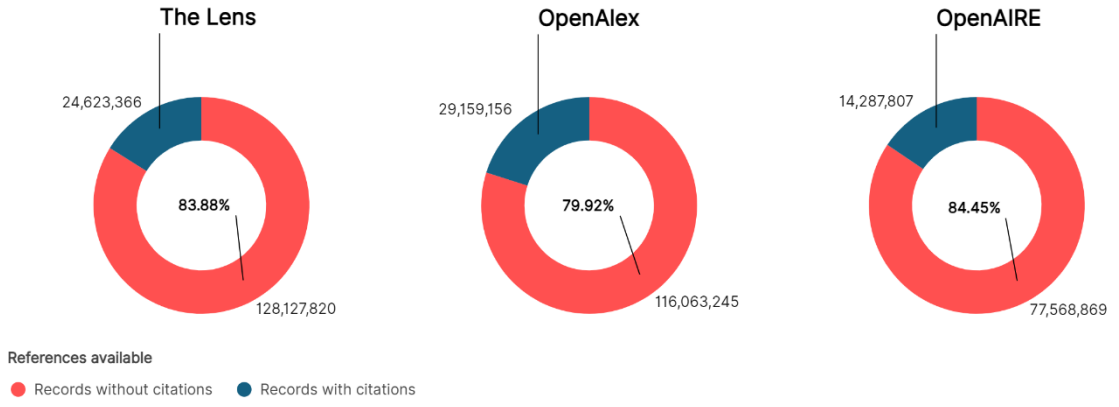

Figure 17: Number of records with citations and percentage of records without citations outside of the overlap with Scopus

Conversely, analyzing the exclusive long tail covering the literature residing entirely outside the overlap content, reveals a significant drop in citation connectivity within open infrastructures. In this region, a high proportion of documents remain uncited. Specifically, 83.88% of exclusive records in The Lens, 79.92% in OpenAlex, and 84.45% in OpenAIRE register zero citations. Because the matching protocol excludes structurally ambiguous records, this high rate of un-citedness reflects the open ecosystem's ingestion of fragmented or predominantly archival material. Consequently, this non-overlapping tail remains largely disconnected from the active citation network, limiting its utility for impact-driven institutional or author-level evaluations.

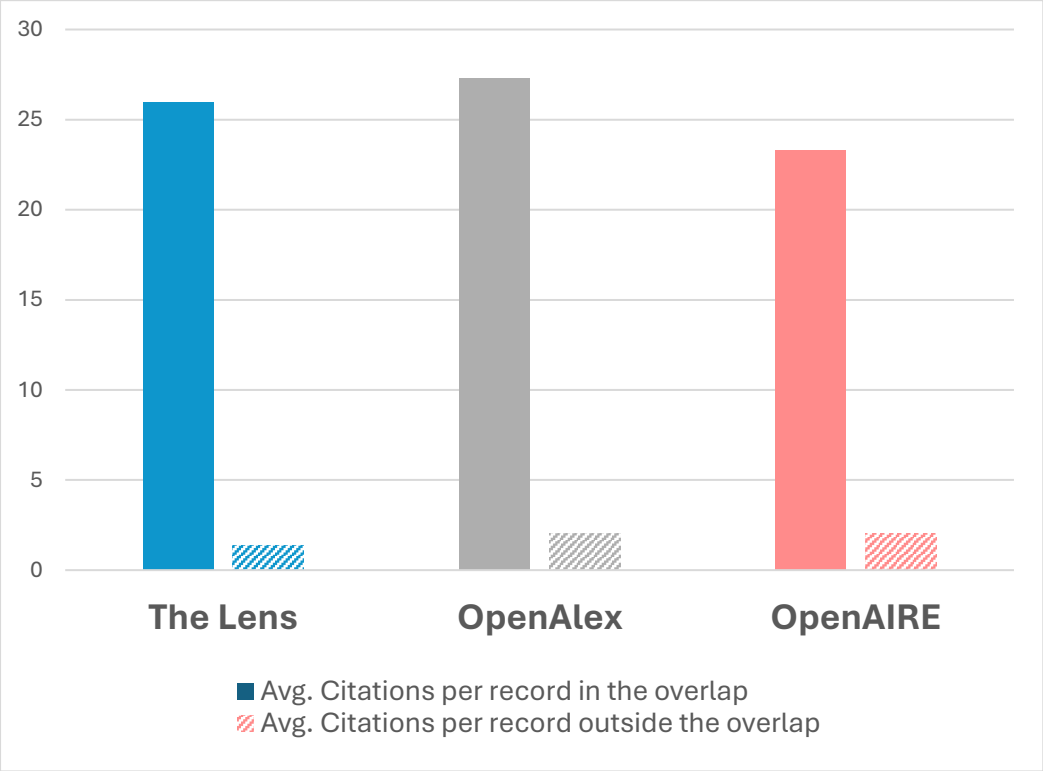

Figure 18: Average citations per record for the three databases, inside and outside the overlap with Scopus

The contrast between curated and un-curated data is clearly illustrated by comparing the average citations per document inside and outside the overlap between Scopus and the open databases. Within the overlapping segment, open databases show high citation averages, reaching 25.97 in The Lens, 27.31 in OpenAlex, and 23.32 in OpenAIRE. However, outside the overlap, these averages drop significantly to 1.37, 2.03, and 2.06, respectively. Consequently, the data suggests that the expanded volume of open infrastructures primarily consists of a long tail of infrequently cited literature, whereas the highly cited, networked corpus aligns almost entirely with the commercial overlap.

References

Comprehensive reference metadata is essential for advanced bibliometric analysis, providing the necessary linkages to track knowledge flows, calculate impact metrics, and map scientific

networks. Unlike descriptive metadata, reference lists connect individual publications, enabling the derivation of citation-based indicators used in research evaluation. When this citation infrastructure is incomplete, the capacity to perform accurate impact assessments or longitudinal network analyses is significantly hindered. Consequently, a database lacking robust reference data becomes less reliable for evaluative bibliometrics, as it cannot accurately establish the influence or intellectual context of the research it indexes.

Table 10: Global number of references, percentage of records without references and average number of references per record of the four databases

Database	Number of References	Avg. References per record	Records without references		Total records
			n	%	
The Lens	2,152,223,594	9.98	141,429,774	65.57	215,677,592
OpenAlex	2,381,281,796	11.39	130,069,270	62.22	209,039,899
OpenAIRE	1,492,384,750	9.87	86,954,153	57.52	151,163,776
Scopus	2,405,173,781	32.42	7,507,341	10.12	74,179,352

An examination of the global datasets exposes a profound systemic deficit in reference completeness across the open infrastructures when benchmarked against the Scopus index. Scopus establishes a highly curated baseline, maintaining an average of 32.42 references per record with only 10.12% of its 74.1 million records lacking cited literature. In contrast, the significantly larger open databases exhibit massive structural gaps that preclude large-scale citation analysis. The Lens and OpenAlex fail to provide reference data for 65.57% and 62.22% of their respective corpora, yielding depressed global averages of 9.98 and 11.39 references per record. OpenAIRE mirrors this deficiency with a 57.52% absence rate and an average of 9.87 references. These findings suggest that the broad ingestion models used by open platforms can weaken the integrity of citation networks by emphasizing high document volumes over the accurate linking needed for reliable citation tracking.

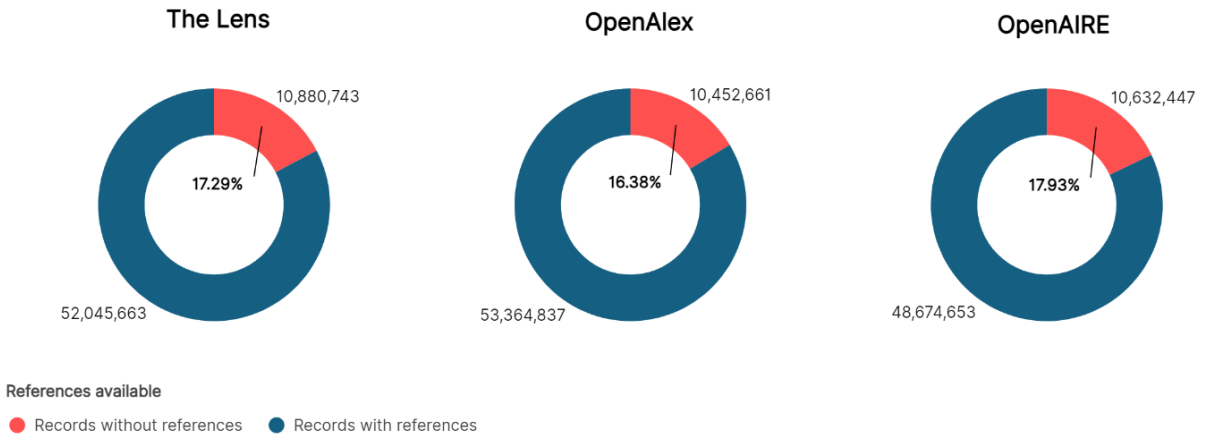

Figure 19: Number of records with references and percentage of records without references in the overlap with Scopus

Applying the cascade matching methodology to isolate the literature intersecting with Scopus reveals a significant improvement in reference completeness. Within this overlapping segment, the gaps in reference data observed globally in open databases are notably reduced. The proportion of records lacking references drops to 17.29% in The Lens, 16.38% in OpenAlex, and 17.93% in OpenAIRE, aligning more closely with the commercial baseline.

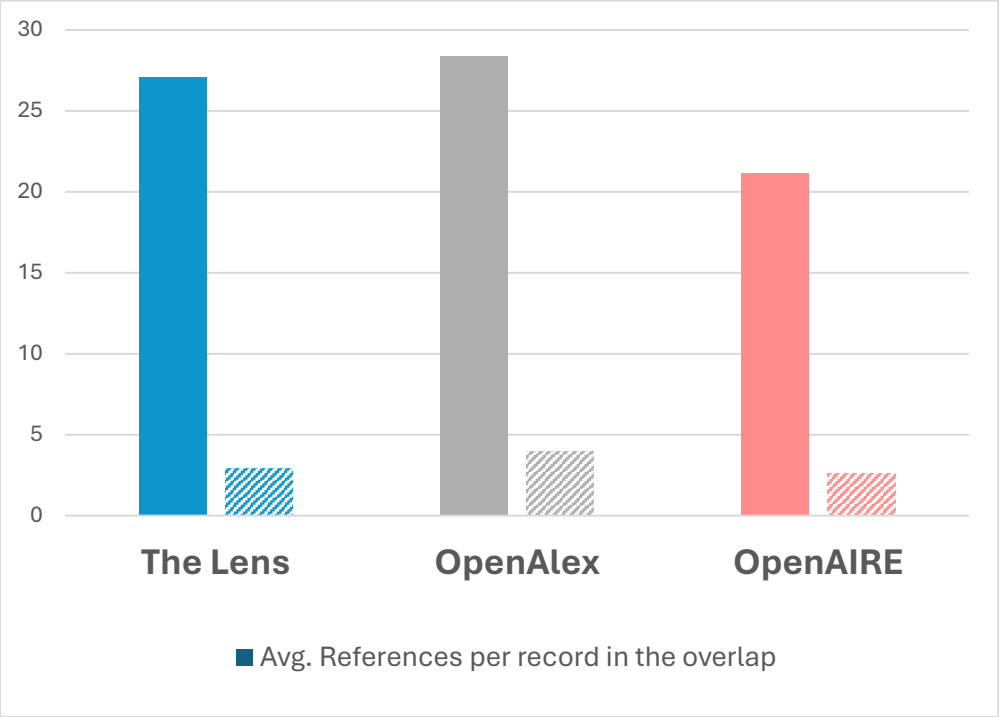

Figure 20: Average references per record for the three databases, inside and outside the overlap with Scopus

Correspondingly, the average number of references per record increases, reaching 27.07 in The Lens, 28.36 in OpenAlex, and 21.13 in OpenAIRE. This convergence suggests that the core scientific literature shared between the Scopus and open ecosystems exhibits higher metadata completeness, making this overlapping segment viable for supporting citation-based evaluative metrics.

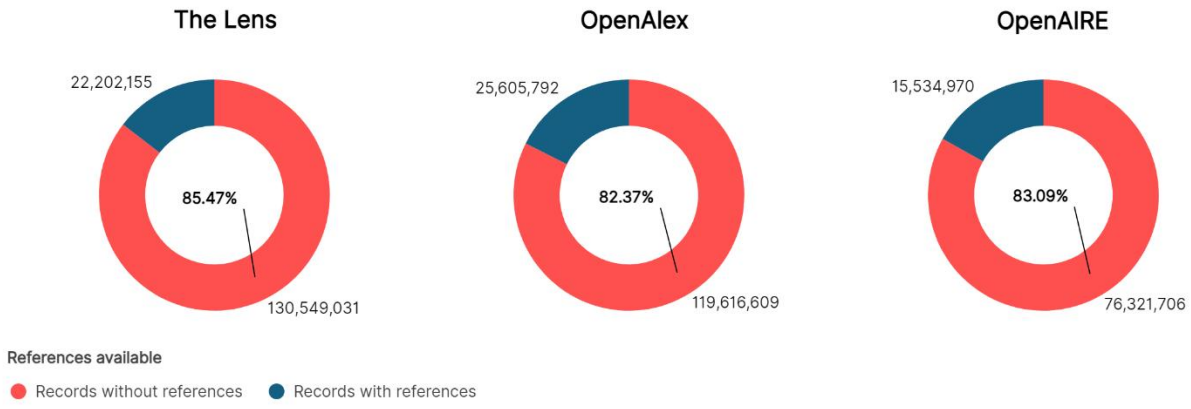

Figure 21: Number of records with references and percentage of records without references outside the overlap with Scopus

Ultimately, the analytical value of reference metadata depends on the ability of the different databases to convert references into citations. This conversion process requires accurately matching outgoing reference strings to existing target records within the corpus to build a cohesive citation network. While open platforms demonstrate higher reference completeness within the Scopus overlap, their broader global deficits suggest limitations in their internal linkage architectures. When references are extracted but fail to resolve into active citation links, whether due to parsing errors, inadequate disambiguation, or missing target documents, the underlying network remains fragmented. Therefore, a database's reliability for evaluative bibliometrics relies not only on the volume of references it ingests, but on the structural efficacy of its reference-to-citation resolution.

Source coverage – Overlap with SJR Sources

Accurate characterization of publication sources, and in particular academic journals, is foundational for bibliometric databases. It supports thematic classification, search precision, and quality control. Furthermore, reliable metadata enables macro-level analyses, informetric mapping, and the application of journal-level metrics. Ultimately, the analytical utility of these databases relies heavily on their ability to link documents to clearly identified sources.

However, source aggregation in open bibliographic infrastructures often lacks consistent descriptive normalization. In un-curated segments of the data, irregular ISSN adoption and lower document-type precision introduce descriptive limitations. Metadata reliability generally decreases outside the core overlap shared with commercial indexes. In these broader sections, fragmentary records and non-traditional venues frequently lack the standardized identifiers needed for bibliometric tracking, complicating journal-level analysis.

To evaluate source coverage, this phase focused on the overlap between Scopus and the open databases, moving away from the initial document-level matching protocol. This adjustment accounts for differing indexing policies: while open databases generally use comprehensive ingestion, Scopus applies selective editorial criteria. A strict document-level intersection could obscure these structural differences in indexing behavior.

To establish a standardized reference for this comparison, the analysis used the SCImago Journal & Country Rank (SJR) inventory. The SJR provides a curated, widely recognized benchmark of global publication venues. Assessing the coverage of open databases against this set ensures comparative consistency. The study uses the 2024 SJR edition, rather than the available 2025 version, to maintain chronological alignment with the temporal boundaries of the preceding empirical evaluations.

Coverage of SJR sources, fundamental metrics

Evaluating aggregate citation metrics and their derived impact ratios is necessary to determine the analytical viability of open bibliographic infrastructures. While volumetric discrepancies indicate ingestion biases, measuring total citations, citations per document, and documents per

source assesses the performance of the platforms' citation extraction processes and the practical effects of their indexing policies. Because these ratios are often used in global university rankings and research evaluations, structural distortions caused by incomplete citation coverage or differences in document counts can reduce the reliability of these databases for comparative assessment. Documenting these macro-level impact profiles illustrates how metadata limitations affect the metrics used to evaluate scientific performance.

Table 11: Comparative overview of aggregate document and citation metrics (2021-2023 publi-cation window) across SJR, The Lens, OpenAlex, and OpenAIRE.

Metric	SJR	The Lens	OpenALEX	OpenAIRE
Sources	31.136	27.960	28.000	27.706
Cites 2024 to 2021-2023 output	39.203.422	34.142.692	39.042.356	34.746.115
Documents 2021-2023	10.783.930	10.069.330	11.119.326	10.434.255
Cites/Doc (3 years)	3,6	3,4	3,51	3,3
Docs/Source	346	360	397	377
Cites/Source	1.259	1.221	1.394	1.254

An analysis of The Lens shows lower citation extraction rates alongside higher document counts per source. The platform records approximately five million fewer total citations than the Scopus baseline. Concurrently, its broader indexing approach results in a higher document count per source (360 compared to the baseline's 346). However, the disproportionate drop in absolute citations relative to its overall document volume reduces the citations per document ratio to 3.4, introducing a slight downward variation in aggregate impact metrics rather than a structural limitation.

OpenAlex achieves high citation counts largely through extensive document ingestion. Although it comes close to matching Scopus in total citations, capturing more than 39 million citations despite relying on a smaller source base, this performance is associated with a substantially larger overall document collection. OpenAlex records the highest document-per-source ratio at 397, suggesting the inclusion of a wider range of document types, including potentially non-citable items. Consequently, the larger denominator reduces the citations per document ratio to 3.51, remaining below the commercial baseline. This indicates how comprehensive ingestion strategies can affect normalized impact metrics.

OpenAIRE shows lower values across the evaluated impact indicators, reflecting limitations in citation linking within this comparative context. The platform records approximately 4.5 million fewer citations than the Scopus baseline. At the same time, it has a higher document-per-source ratio of 377, indicating that it captures a broader range of document types within its source pool. However, the disproportionate drop in absolute citations relative to its overall document volume reduces its citations-per-document ratio to 3.3, the lowest in the group. This suggests its aggregator model would yield lower normalized impact metrics than commercial benchmarks if applied to the same source set.

Source coverage by SJR quartile

Stratifying source and document overlap by journal impact quartiles provides insight into the distribution patterns of open bibliographic infrastructures. Impact quartiles are often associated with editorial models, institutional visibility, and regional representation in global science. Evaluating coverage across these strata helps determine whether open databases capture a broad scholarly spectrum or primarily align with higher-impact journals, potentially underrepresenting regionally focused or lower-impact venues. Documenting this distribution is relevant to identify whether metadata limitations are uniform across the databases or concentrated in lower impact tiers, a factor that can influence macro-level evaluations of emerging research environments.

Table 12: Distribution of sources by SJR quartile, difference between SJR sources and matching sources in The Lens.

SJR best quartile	The Lens Sources	SJR Sources	Diff. %	The Lens Documents	SJR Documents	Diff. %
Q1	9.076	9.160	-0,92	6.113.880	5.193.175	17,73
Q2	7.395	7.707	-4,05	2.683.370	2.433.788	10,25
Q3	6.221	6.928	-10,20	1.337.235	1.319.770	1,32
Q4	4.900	6.406	-23,51	1.002.764	1.156.247	-13,27
-	368	935	-60,64	367.238	680.950	-46,07

The Lens shows variations in coverage across impact quartiles, with higher indexing rates in the upper tiers that decrease in the lower quartiles. In the first quartile, source coverage closely matches the baseline (with a minor difference of 0.92%), alongside a 17.73% increase in indexed documents. This higher document count likely reflects a broader ingestion of additional editorial content. However, this coverage declines in the lower strata. By the fourth quartile, The Lens does not capture 23.51% of the baseline sources, resulting in a 13.27% decrease in document coverage. This pattern indicates an indexing approach that captures high-impact literature effectively but is less comprehensive for lower-tier venues.

Table 13: Distribution of sources by SJR quartile, difference between SJR sources and matching sources in OpenAlex.

SJR best quartile	OpenAlex Sources	SJR Sources	Diff. %	OpenAlex Documents	SJR Documents	Diff. %
Q1	9.067	9.160	-1,02	5.994.604	801.429	15,43
Q2	7.400	7.707	-3,98	2.612.338	178.550	7,34
Q3	6.272	6.928	-9,47	1.271.777	-47.993	-3,64
Q4	4.938	6.406	-22,92	936.682	-219.565	-18,99
-	323	935	-65,45	358.517	-322.433	-47,35

OpenAlex shows a similar distribution pattern, with coverage decreasing from higher to lower impact tiers. The platform captures roughly 99 % of first-quartile sources, accompanied by a 15.43 % higher document volume, likely reflecting comprehensive cover-to-cover indexing

practices. Conversely, coverage declines in the lower impact tiers, resulting in a 22.92 % source deficit in the fourth quartile and a 65.45 % gap among unranked sources. This corresponds to an 18.99 % lower document count in the fourth quartile, indicating that OpenAlex's automated ingestion processes are more effective at capturing prominent, well-standardized journals than integrating the metadata of lower-impact scholarly outputs.

Table 14: Distribution of sources by SJR quartile, difference between SJR sources and matching sources in OpenAIRE.

SJR best quartile	OpenAIRE Sources	SJR Sources	Diff. %	OpenAIRE Documents	SJR Documents	Diff. %
Q1	9.035	9.160	-1,36	5.974.292	5.193.175	15,04
Q2	7.382	7.707	-4,22	2.536.938	2.433.788	4,24
Q3	6.236	6.928	-9,99	1.286.237	1.319.770	-2,54
Q4	4.762	6.406	-25,66	782.918	1.156.247	-32,29
-	291	935	-68,88	346.959	680.950	-49,05

OpenAIRE shows the most pronounced decrease in coverage across the impact quartiles among the evaluated platforms. While it maintains comparable coverage in the first quartile with a 1.36 % lower source count and a 15.04 % higher document volume, its coverage drops significantly in the lower tiers. In the fourth quartile, OpenAIRE records a 25.66 % source gap, the largest among the cohort, which corresponds to a 32.29 % lower document volume. This reduced capture rate in the lowest impact tier, combined with an almost 69 % source gap for unranked venues, indicates that OpenAIRE's aggregator model is primarily aligned with high-impact channels and presents limitations for comprehensively evaluating research outside these central tiers.

Source coverage – Source type

Disaggregating source and document overlap by publication typologies helps assess the capabilities and limitations of open indexing infrastructures. While macro-level aggregations can obscure specific variations, evaluating distinct formats such as journals versus conference proceedings or book series, shows how these databases handle diverse metadata structures. Documenting these differences indicates that the coverage of open platforms often varies by format, pointing to limitations in capturing non-journal scholarly outputs. This variation is a relevant factor that can affect cross-disciplinary evaluations, particularly in fields where conferences or books are the primary modes of scientific communication.

An analysis of The Lens indicates higher coverage rates for traditional journal articles. In this format, source coverage shows a 7.6 % gap compared to the Scopus baseline, alongside a 10.89 % higher document count, likely reflecting comprehensive cover-to-cover indexing practices. However, coverage decreases notably outside standard journal formats. The platform shows a 71.67 % source gap in conference proceedings and a 56.06 % gap for trade journals. This lower source capture corresponds to a 51.23 % lower document count for conference outputs,

suggesting current limitations in processing non-serial metadata and mapping event-based publication structures.

Table 15: Distribution of source types, difference between SJR sources and matching sources in The Lens

Type	The Lens Sources	SJR Sources	Diff. %	The Lens Documents	SJR Documents	Diff. %
Journal	27.306	29.553	-7,60	10.859.097	9.792.922	10,89
Book Series	380	725	-47,59	311.971	320.034	-2,52
Conferences and Proceedings	187	660	-71,67	310.290	636.259	-51,23
Trade Journal	87	198	-56,06	23.129	34.715	-33,37
Total	27.960	31.136	-10,20	11.504.487	10.783.930	6,68

Table 16: Distribution of source types, difference between SJR sources and matching sources in OpenAlex

Type	Open-Alex Sources	SJR Sources	Diff %	OpenAlex Docs.	SJR Docs.	Diff %
Journal	27.393	29.553	-7,31	10.537.450	9.792.922	7,60
Book Series	378	725	-47,86	306.529	320.034	-4,22
Conference and Proceedings	147	660	-77,73	307.678	636.259	-51,64
Trade Journal	82	198	-58,59	22.261	34.715	-35,87
Total	28.000	31.136	-10,07	11.173.918	10.783.930	3,62

The data for OpenAlex shows a similar pattern regarding standard journal formats, recording a 7.31 % source gap and a 7.6 % higher document count for traditional journals. However, coverage decreases significantly for other publication types. OpenAlex records the lowest source coverage for conference proceedings among the evaluated platforms, showing a 77.73 % gap compared to the baseline. This lower source capture corresponds to a 51.64 % lower document volume for this format. These results suggest limitations in OpenAlex's capacity to identify and link non-journal formats, such as conference events, trade literature, and serialized book collections.

Table 17: Distribution of source types, difference between SJR sources and matching sources in OpenAIRE

Type	OpenAIRE Sources	SJR Sources	Diff. %	OpenAIRE Documents	SJR Documents	Diff. %
Journal	27.287	29.553	-7,67	10.578.812	9.792.922	8,03
Book Series	180	725	-75,17	24.440	320.034	-92,36
Conference and Proceedings	172	660	-73,94	302.710	636.259	-52,42
Trade Journal	67	198	-66,16	21.382	34.715	-38,41
Total	27.706	31.136	-11,02	10.927.344	10.783.930	1,33

OpenAIRE shows the lowest coverage rates for non-journal typologies among the evaluated platforms, suggesting limitations in ingestion and classification within its aggregator model. While its journal metrics align with the other open platforms, recording a 7.67 % source gap and an 8.03 % higher document count, its coverage of book series is markedly lower. OpenAIRE registers a 75.17 % source gap for book series, corresponding to a 92.36 % lower document retrieval rate. Combined with reduced coverage in conference proceedings (a 73.94 % source gap) and trade journals, the data indicates that OpenAIRE presents significant limitations for comprehensively evaluating scholarly outputs outside standard journal articles.

Source coverage – Geographic distribution of publishers

Mapping source and document distributions across geopolitical regions is necessary to assess the global representation within open bibliographic infrastructures. Bibliometric databases often exhibit varied geographic coverage, frequently showing lower representation of research outputs from the Global South and non-Western regions. This is largely due to operational challenges in capturing and integrating regional publication venues. Integrating this geospatial analysis into the evaluation demonstrates that macro-level metrics derived from these platforms may reflect an overrepresentation of Western outputs. Documenting this distribution provides essential context, indicating that these databases require methodological calibration for comprehensive global research assessments.

Table 18: Top 20 countries by editorial output, comparison of sources and documents indexed in the free databases in the SJR source list.

Country	The Lens Sources	The Lens Documents	OAlex Sources	OAlex Documents	OAIRE Sources	OAIRE Documents
United States	-8,68	12,5	-9,80	9,4	-11,83	7,19
United Kingdom	-1,94	18,0	-2,02	15,2	-2,54	13,15
Netherlands	-12,45	11,2	-12,54	9,1	-15,04	7,48
Germany	-9,54	2,8	-9,90	-0,1	-9,18	-2,78
China	-43,39	-67,6	-42,12	-73,3	-48,73	-64,68
Switzerland	-3,13	1,6	-3,23	1,2	-5,57	-5,2
Spain	-9,31	10,1	-7,84	4,6	-9,19	9,6
Italy	-46,76	-33,4	-37,84	-41,5	-29,06	-28,65
Russian Federation	-2,24	8,0	-1,60	6,5	-3,37	3,21
Poland	-6,04	5,0	-7,01	2,7	-4,89	5,3
France	-11,85	23,8	-12,37	21,9	-12,2	23,18
India	-8,63	6,6	-9,19	-0,1	-11,26	-0,63
Brazil	-3,20	22,3	-2,56	21,5	-3,84	23,65
Japan	-17,84	-0,1	-16,20	-6,6	-20,66	-2,69
Iran	-30,85	-34,8	-21,81	-50,2	-18,35	-25,88
South Korea	-7,76	-7,7	-4,31	-9,2	-10,92	-12,11
Turkey	-6,59	-5,2	-6,89	-7,6	-7,78	-5,82
Canada	-7,43	9,3	-7,43	5,7	-11,46	3,8
Singapore	-0,39	10,4	-0,77	8,9	-8,11	2,24
Indonesia	-1,58	15,7	-3,16	6,0	-1,98	10,19

The overlap analysis between the open databases and the SCImago baseline indicates geographic variations in coverage, reflecting differences in indexing methodologies. While the capture of sources from the United States and the United Kingdom aligns closely with the baseline, open infrastructures show a 30 to nearly 50 % lower source coverage for journals from China, Italy, and Iran.

Document-level data further illustrates the effects of divergent ingestion policies. For example, The Lens and OpenAlex record 18.0 % and 15.2 % higher document counts for the United Kingdom, respectively. This reflects a comprehensive indexing approach that incorporates document types often excluded by Scopus's selective editorial criteria.

Conversely, in regions with lower source coverage, such as China, document counts drop significantly, showing a 64.6 % gap in OpenAIRE and a 73.3 % gap in OpenAlex. Furthermore, OpenAIRE presents variations in document retrieval even in regions with stable source coverage; for instance, it records lower document counts for Germany, whereas the other platforms maintain parity or show higher volumes. Consequently, comparative macro-level assessments using these databases require methodological adjustments to account for how varying source coverage and distinct indexing policies affect the total volume of records per country.

Table 19: World regions by editorial output, comparison of sources and documents indexed in the free databases in the SJR source list.

Country	The Lens		OpenAlex		OpenAIRE	
	Sources %	Documents %	Sources %	Documents %	Sources %	Documents %
Africa	-5,8	3,7	-10,3	-2,1	-6,6	5,1
Asiatic Region	-21,7	-34,0	-21,1	-39,7	-25,3	-35,2
Eastern Europe	-8,4	15,8	-8,5	12,0	-6,6	16,8
Latin America	-6,3	15,8	-4,2	13,7	-7,7	15,9
Middle East	-16,1	-9,6	-13,2	-19,5	-13,3	-9,9
Northern America	-8,6	12,4	-9,7	9,3	-11,8	7,1
Pacific Region	-16,2	-1,7	-16,5	-5,2	-18,6	1,7
Western Europe	-8,5	10,5	-8,0	8,1	-8,0	3,5

Aggregating the source and document overlap metrics to a macro-regional level highlights differences in geographic coverage across open bibliographic infrastructures. The data indicates a contrast between regions with established Western publishing networks and other global regions. While Northern America, Western Europe, and Eastern Europe show source-level gaps compared to the baseline, generally between 6 and 12 %, other regions exhibit lower coverage rates. The Asiatic Region shows larger source gaps, reaching 21.7 % in The Lens and 25.3 % in OpenAIRE. A similar pattern of lower coverage is observable in the Middle East and the Pacific Region. These macro-level variations suggest that open databases face challenges in

capturing diverse publication venues globally, introducing limitations for comprehensive inter-regional comparative assessments.

Document-level data across these macro-regions illustrates how comprehensive indexing practices interact with source availability to affect total document volumes. In regions with stable source coverage, such as Western Europe, Northern America, and Latin America, open databases often record 10 to 15 % higher document counts compared to the baseline. This higher document volume reflects broader indexing policies that capture additional document types within the covered sources. Conversely, in the Asiatic Region and the Middle East, lower source coverage corresponds to a decrease in document retrieval. Document gaps in the Asiatic Region, reaching 34.0 % in The Lens, 35.2 % in OpenAIRE, and 39.7 % in OpenAlex, indicate that comprehensive document ingestion within captured sources does not offset the initial lower source coverage. Consequently, using open databases for inter-regional comparative bibliometrics requires careful methodological consideration, as the combination of broader document indexing in high-coverage regions and lower source capture in others can affect the balance of global comparative assessments.

Source coverage – Publishers

A careful examination of source coverage and document ingestion is essential for reliable research evaluation. As open bibliographic infrastructures are increasingly used in science policy, institutional benchmarking, and funding decisions, it is important to understand their structural characteristics, including regional differences and variations in metadata ingestion. Rather than assuming that all platforms provide equivalent coverage, this analysis establishes an empirical baseline for identifying these differences. By documenting these patterns, the study highlights the limitations of raw data aggregation and shows that methodological adjustments are necessary to achieve accurate and balanced representations of global research activity in comparative bibliometric analyses.

Disaggregating overlap metrics by publisher indicates that the coverage of open bibliographic infrastructures varies according to the metadata dissemination practices and editorial models of individual publishing houses. For major commercial publishers such as Elsevier, Springer Nature, Taylor & Francis, and Wiley, the open databases achieve near-complete source recognition, with gaps typically below two %. However, this source parity is accompanied by higher document counts, ranging between 7% and 22%. This pattern reflects the impact of comprehensive indexing, as open platforms capture a wider range of content including editorial material that Scopus excludes through selective criteria. This trend is also observed in several university and specialized presses, such as Oxford University Press, Cambridge University Press, and Wolters Kluwer Health, where document counts show larger increases despite minimal source differences.

In contrast, native Open Access publishers with digital, article-centric models notably MDPI and Frontiers Media, show high parity in both source and document counts across all platforms. This suggests that streamlined publication structures, which lack legacy print-based content, facilitate more synchronized and consistent data capture. The data also shows lower coverage rates correlated with specific publisher profiles. The source gaps recorded for IEEE, exceeding

54% across the three open databases, alongside lower document yields of over 28%, indicate challenges in mapping and integrating complex conference proceeding structures. Similarly, the source gaps observed for Brill (between 46% and 47%) highlight limitations in capturing humanities-oriented or book-heavy portfolios. Ultimately, these results show that the reliability of open metadata is influenced by the document typologies and distribution mechanisms of the underlying sources, requiring careful consideration of publisher-specific factors in bibliometric analysis.

Table 20: Top 20 publishers by published sources, comparison of sources and documents indexed in the free databases in the SJR source list.

Publisher	The Lens		OpenAlex		OpenAIRE	
	Sources %	Documents %	Sources %	Documents %	Sources %	Documents %
Elsevier	-0,41	21,86	-0,56	21,12	-2,67	21,99
Springer Nature	-0,44	6,95	-1,48	2,86	-3,64	-15,78
Taylor & Francis	-0,17	21,96	-0,25	8,35	-0,63	7,57
John Wiley & Sons	-0,19	22,79	-0,45	21,28	-0,51	14,99
SAGE	0,00	8,90	-0,09	7,02	-0,36	5,91
Null	-14,18	-7,08	-14,79	-10,42	-13,20	-14,24
IEEE	-54,30	-28,25	-59,86	-34,90	-59,50	-34,27
Brill	-46,34	9,29	-46,34	5,13	-47,65	0,35
Wolters Kluwer Health	-0,42	58,95	-0,42	54,73	-1,06	53,32
Oxford University Press	0,00	62,57	-0,22	57,83	-0,44	55,51
Walter de Gruyter	-2,09	19,11	-2,55	17,30	-2,55	16,05
Cambridge University Press	-0,26	62,53	-0,26	52,87	-0,78	50,35
Emerald Publishing	-0,55	16,57	-0,55	15,23	-4,99	10,62
MDPI	-0,32	1,11	-0,32	1,18	-0,32	1,21
Inderscience Publishers	0,00	111,11	0,00	111,73	0,00	22,86
Pleiades Publishing	0,00	6,25	0,00	2,43	0,00	-2,10
Bentham Science Publishers	0,00	27,31	0,00	12,35	-11,85	10,90
KeAi Communications Co.	-0,74	11,24	0,00	17,45	-0,74	9,04
World Scientific	0,00	9,92	0,00	8,80	-12,12	4,17
Frontiers Media S.A.	0,00	0,34	0,00	0,32	0,00	0,14

Conclusions

The Scale Paradox: Coverage *versus* Standardization

The evaluation of open bibliographic infrastructures reveals a clear trade-off between scale and metadata standardization. While platforms like The Lens and OpenAlex offer expansive coverage, exceeding 200 million records compared to the 74 million of Scopus, it is accompanied by variations in metadata completeness. The aggregation models of open platforms tend to prioritize recall, capturing a vast array of scholarly outputs. In many evaluative scenarios, the available information proves fundamentally insufficient to construct robust indicators, obliging analysts to either supplement the missing data through external sources or accept methodological compromises regarding the representativeness and validity of the findings.

The Curated Core *versus* Extended Literature

The analysis demonstrates that data within open infrastructures is not structurally homogeneous but rather divided into two distinct regions. The overlapping segment, which is the literature that open databases share with Scopus, maintains high metadata completeness, characterized by consistent DOI presence, unique titles, and citation densities comparable to commercial standards. This segment of the data provides a viable foundation for stable impact calculations. In contrast, the exclusive region (comprising over 150 million records in the largest platforms) shows lower metadata standardization. This extended segment frequently lacks persistent identifiers (DOIs and ISSNs) and affiliation data, exhibits higher title duplication, and presents lower connectivity to the active citation network.

The Dichotomy Between Structural Availability and Evaluative Viability

The empirical verification of high metadata availability across open infrastructures should not be equated with their operational viability for institutional evaluation. While the analysis demonstrates extensive data ingestion capabilities (including affiliations, persistent identifiers, and citation links) this scale does not guarantee qualitative integrity or standardization. Operating with these raw records requires an intermediate phase of preprocessing, disambiguation validation, and profile reconciliation. The evaluated databases do implement automated processes to address these challenges, but assessing the efficacy of their internal disambiguation models falls outside the scope of this study.

The Role of the "Long Tail" as an Impact Feeder

The analysis reveals an asymmetric dynamic in citation flows: the literature exclusive to open platforms (the 'long tail') seems to function primarily as a source of references directed toward the core literature already established in Scopus, generating a disproportionately low number of citations in return. Consequently, this expansion in coverage tends to reinforce the citation indicators of journals within commercial indices, rather than significantly redistributing measured impact across the broader corpus.

Limitations in Key Metadata Normalization

Variations in the normalization of key metadata fields introduce methodological challenges for micro- and meso-level indicators. First, document type classification lacks standardization; while Scopus maintains strict editorial categorizations, open platforms exhibit higher rates of unclassified records or rely on algorithmic labeling that may group diverse materials (including gray literature and preprints) under generic "article" tags.

Second, the absence of affiliation data, missing over 55% of the global records in The Lens and OpenAlex, presents challenges for conducting large-scale institutional rankings or science policy evaluations outside the region overlapping with Scopus.

Finaly, the uneven flow of citations already described, where the extended literature amplifies the impact of the core but remains largely uncited itself, coupled with source metadata limitations, suggests that developing robust, comparable journal-level metrics using the entirety of these open infrastructures will present significant methodological challenges.

Geographic and Editorial Variations

Evaluating editorial and geographic coverage is difficult given the problems in metadata completeness already described. This study uses the SJR source list, which requires acknowledging the inherent selectivity of this Scopus-derived baseline. By relying on SJR, the analysis focuses strictly on a curated, high-quality core, effectively masking the broader universe of sources that open databases may index outside the commercial scope.

Within this shared core, open infrastructures exhibit specific variations. Coverage of Western research outputs and major commercial or native Open Access publishers (such as MDPI or Frontiers Media) aligns closely with the baseline. However, notable gaps remain in the Global South, particularly in Asia and the Middle East, where source deficits reach up to 25%, as well as in complex publication structures like conference proceedings (e.g., IEEE) and specialized humanities monographs (e.g., Brill). While open platforms capture regional or specialized venues beyond the Scopus baseline, utilizing their exclusive "long tail" to offset these geographic and editorial biases seems currently unfeasible due to its defective or non-existent source-level metadata.

The Strategic Role of Empirical Audits in Research Assessment Reform

The shift toward responsible research evaluation, as advocated by the Coalition for Advancing Research Assessment (CoARA), and the mandate for default openness stipulated in the Barcelona Declaration on Open Research Information, represent a fundamental paradigm shift in scholarly communication. However, the successful implementation of these frameworks depends substantially on the technical reliability of the underlying data infrastructures.

Our findings indicate that while open databases offer the scale required to support more inclusive and transparent evaluation models, their operational maturity remains uneven. In the context of a research assessment reform, evaluation processes must move beyond a theoretical preference for openness and address the practical challenges of data integrity. Without a rigorous understanding of the trade-offs between scale, precision, and metadata quality, the move

toward open research information risks introducing new methodological biases into the global scholarly landscape.

Limitations

While all datasets were analyzed within the same chronological boundaries (covering literature from 1996 to 2024), the data snapshots were obtained at different points during 2025 (March for Scopus, September for OpenAIRE; November for The Lens and OpenAlex). Although this does not alter the historical corpus, the temporal gap may introduce minor volumetric discrepancies due to the ongoing retroactive indexing and continuous updating processes typical of these infrastructures.

To ensure the highest possible accuracy in document-to-document cross-referencing, the matching protocol strictly prioritized precision to minimize false positives. Records presenting duplicated DOIs or ambiguous, non-unique titles within their source databases were systematically excluded. While this approach guarantees a high-confidence overlap, it inherently generates false negatives. Consequently, the reported intersection volumes may slightly underestimate the theoretical overlap, prioritizing analytical reliability over maximum recall.

The evaluation of source-level coverage relies on the SCImago Journal & Country Rank (SJR) as the primary baseline. As a result, the analysis effectively measures the capacity of open platforms to replicate a highly curated, commercial core. This methodological choice precludes a comprehensive evaluation of the regional, non-traditional, or specialized venues that are exclusive to the open databases. As noted, the absence or fragmentation of standardized source identifiers in this exclusive "long tail" prevents reliable benchmarking against alternative frameworks.

This study primarily quantifies the structural availability of key metadata fields, such as affiliations, references, and document types. However, it does not evaluate the qualitative accuracy of the databases' internal entity disambiguation models. Issues such as the precision of algorithmic author clustering, institutional profile reconciliation, or the potential over-labeling of un-curated documents as standard "articles" (e.g., in automated classification systems) fall outside the scope of this analysis. Therefore, high metadata completeness should not be automatically equated with perfect metric precision.

References

Abramo, G.; **Cicero, T.**; **D'Angelo, C. A.** (2026). Data source effects in research performance assessment of individuals and institutions: Comparing OpenAlex with established bibliographic databases. *Scientometrics*.
https://doi.org/10.1007/s11192-026-05638-6

Baas, J.; **Schotten, M.**; **Plume, A.**; **Côté, G.**; **Karimi, R.** (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386.
https://doi.org/10.1162/qss_a_00019

Barcelona Declaration on Open Research Information (2024). *Barcelona Declaration on Open Research Information (Version 1.0)*. Zenodo.
https://doi.org/10.5281/zenodo.10958522

Birkle, C.; **Pendlebury, D. A.**; **Schnell, J.**; **Adams, J.** (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376.
https://doi.org/10.1162/qss_a_00018

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.** (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology*, *76*(6), 884–895.
https://doi.org/10.1002/asi.24979

Coalition for Research Assessment Reform (*CoARA*). (2022). *Agreement on reforming research assessment*.
https://coara.eu

Culbert, J. H.; **Hobert, A., Jahn, N.**; **Haupka, N.**; **Schmidt, M.**; **Donner, P.**; **Mayr, P.** (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics*, *130*(4), 2475–2492.
https://doi.org/10.1007/s11192-025-05293-3

Declaration on Research Assessment (*DORA*). (2012). *San Francisco Declaration on Research Assessment*.
https://sfdora.org/read

Delgado-Quirós, L.; **Ortega, J. L.** (2024a). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, *5*(1), 31–49.
https://doi.org/10.1162/qss_a_00293

Delgado-Quirós, L.; **Ortega, J. L.** (2024b). Research entity information and coverage in eight free access scholarly databases. *Online Information Review*.
https://doi.org/10.1108/OIR-07-2023-0348

Guerrero-Bote, V. P.; **Chinchilla-Rodríguez, Z.**; **Mendoza, A.**; **De-Moya-Anegón, F.** (2021). Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the country and institutional levels. *Frontiers in Research Metrics and Analytics*, *5*, 593494.
https://doi.org/10.3389/frma.2020.593494

Haupka, N.; **Culbert, J. H.**; **Schniedermann, A.**; **Jahn, N.**; **Mayr, P.** (2026). Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *Quantitative Science Studies*, *7*, 179-194.
https://doi.org/10.1162/QSS.a.406

Khanna, S.; **Ball, J.**; **Alperin, J. P.**; **Willinsky, J.** (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, *3*(4), 912–930.
https://doi.org/10.1162/qss_a_00228

Maddi, A.; **Maisonobe, M.**; **Boukacem-Zeghmouri, C.** (2025). Geographical and disciplinary coverage of open access journals: OpenAlex, Scopus, and WoS. *PLOS ONE*, *20*(4), e0320347.
https://doi.org/10.1371/journal.pone.0320347

Mongeon, P.; **Hare, M.**; **Riddle, P.**; **Wilson, S.**; **Krause, G.**; **Marjoram, R**; **Toupin, R.** (2025). Investigating document type, language, publication year, and author count discrepancies between OpenAlex and the Web of Science. *arXiv*.
https://doi.org/10.48550/arXiv.2508.18620

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

Scheidsteger, T.; **Haunschild, R.**; **Bornmann, L.** (2025). How similar are field-normalized citation impact scores obtained from OpenAlex and three popular commercial databases? An empirical comparison based on large German universities. *Scientometrics*, *130*(7), 3537–3569.
https://doi.org/10.1007/s11192-025-05338-7

Torres-Salinas, D.; **Arroyo-Machado, W.** (2026). The 'Big Three' of Scientific Information: A comparative bibliometric review of Web of Science, Scopus, and OpenAlex. Influscience Editions. https://doi.org/10.5281/zenodo.18411229

Visser, M.; **Van Eck, N. J.**; **Waltman, L.** (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112

Zhang, L.; **Cao, Z.**; **Shang, Y.**; **Sivertsen, G**; **Huang, Y.** (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics*, *129*, 5869–5891.
https://doi.org/10.1007/s11192-023-04923-y

About this report

In data infrastructure, size alone is not a sign of quality. In many cases, the larger the system becomes, the harder it is to maintain consistency, reliability, and control. This is ultimately a signal-to-noise problem. In this study, SCImago compares four databases —Scopus and three open platforms: The Lens, OpenAlex and OpenAIRE— regarding their coverage of scientific literature for the 1996–2024 interval. This comparison raises an important question: why doesn't access to 200 million records automatically produce better insights or better decisions? The answer lies not in what these databases include, but in what gets lost beneath the scale of automated aggregation.

About SCImago Lab

SCImago Lab is a technology-based company born as a spin-off of the SCImago Research Group, whose core purpose is specialized research in scientometrics, scientific publishing, and web visibility, designing and developing analytical solutions in Science, Technology, and Innovation. It is dedicated to the evaluation of science through techniques of analysis, visualization, and assessment of information contained in databases, and is involved in the development of various scientific information analysis tools. Among its most notable products are the *SCImago Journal & Country Rank* (SJR) and the *SCImago Institutions Rankings* (SIR), the latter capable of providing information on around 7,500 research-related entities, such as universities, research centers, hospitals, private companies, and government agencies.

About the publisher

Ediciones Profesionales de la Información SL (EPI SL) is a Spanish publishing house based in the province of Granada, Spain, though it operates primarily online. Currently is part of SCImago Research Group. EPI SL has shown a strong commitment to serving the scientific and professional community of Information specialists — including librarians, information scientists, and communicators — designing and producing a wide variety of products and services. EPI SL publishes the *Anuario ThinkEPI* and *Infonomy* journals, organizes the semi-annual CRECS (Congress on Scientific Journals) and CoDi (Congress on the Dissemination, Transfer, and Societal Impact of Science) congresses, and publishes books and reports on Social Communication and Information.

9 788412 575781