https://doi.org/10.51408/issi2025_112

What Are We Missing? A Systematic Approach to Overlap Analyses of Local and Global Repositories

Simon Willemin[1], Gaël Bernard[2], Julian Dederke[3], Mahmoud Hemila[4], Michelle Koch[5]

[1] simon.willemin@library.ethz.ch, [3] julian.dederke@library.ethz.ch, [4] mahmoud.hemila@library.ethz.ch,
[5] koch_michelle@outlook.com
ETH Zurich, ETH Library, Rämistrasse 101, CH-8092 Zurich (Switzerland)

[2] gael.bernard@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL), Institutional Data and Analytics Unit, CH-1015 Lausanne (Switzerland)

Abstract

Local repositories, managed by institutions, often differ in coverage and metadata from the research output affiliated with the concerned institutions in global repositories such as OpenAlex, which aggregate records from numerous sources for broader visibility. This paper introduces a DOI Screening System that systematically identifies and explains mismatches between local and global repositories by classifying publications as local-only, matched, or global-only. The system applies predefined rules and allows identifying patterns such as misattributed affiliations, unrecognized DOI prefixes, incomplete metadata, or underrepresented publication types. Based on these patterns, one can derive 'curative' actions. We demonstrate the system's utility by comparing the repositories of EPFL and of ETH Zurich to OpenAlex, showing how subtle inconsistencies in identifiers and affiliations can account for many discrepancies. The system provides insights into how targeted interventions addressing the root causes of these discrepancies can be used to enhance coverage and reliability in both local and global repositories.

Introduction

The landscape of bibliometric data has expanded considerably in recent years, with numerous openly accessible repositories complementing established, subscription-based platforms. Several global repositories (i.e., bibliometric databases) such as OpenAlex and OpenAIRE now coexist alongside institutional or national repositories, each serving distinct yet complementary purposes. *Local repositories* for scholarly outputs give institutions control over their data, provide archival continuity, and capture the full breadth of their scientific production—features critical for accurate record-keeping and institutional sovereignty. Conversely, *global repositories* of scholarly metadata bolster discoverability, expand the global reach of publications, facilitate benchmarking across institutions, and influence university rankings or decision-making, and therefore command significant attention from research administrators. Both types of repositories play a major role in international and national initiatives such as Plan S, the European Open Science Cloud (EOSC) and the Swiss National Open Access Strategy to ensure that research output is findable and accessible.

However, there are still discrepancies between the research output available locally and globally. Moreover, there is no systematic method to clarify why certain publications appear in one repository but not the other, and there are few tools to quickly perform an "overlap analysis" that compares the extent of coverage between several data sources. Similar to the biblioverlap package (Vieira & Leta, 2024), we aim to offer a semi-automated approach that allows anyone to perform such analyses. While our current system focuses on the overlap between a local and a global repository, our approach not only aims to identify the gaps, but also seeks to uncover the underlying reasons for these discrepancies, enabling a better understanding of the factors contributing to the mismatches. Systematically identifying the reasons behind overlaps (or lack thereof) among repositories can lead to concrete curation actions—such as updating metadata or affiliations. We refer to this approach to overlap analysis as *curative* in the sense that it aims to identify gaps to ultimately improve the coverage and metadata quality in both local and global repositories. It therefore goes beyond *descriptive* approaches that merely aim to understand the logic behind the selection and indexing process of a repository.

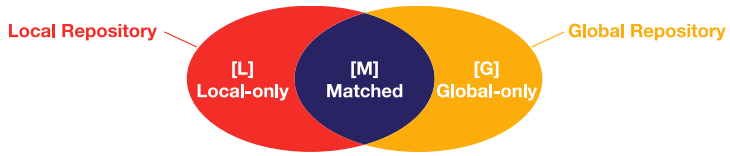

Figure 1. Sets representation for the overlap analysis.

As illustrated in Figure 1, a basic overlap analysis comparing local and global repositories typically categorizes publications into three groups: local-only, matched, and global-only. However, institutions need more than just these counts—they require practical explanations of why a publication is missing from one repository, whether due to issues like unregistered prefixes, incomplete metadata, or incorrect institutional attributions. Hence, we present a system that we call *DOI Screening System* and which is designed to help quickly gauge the overlap and the gap between a local and a global repository, as well as to deliver a list of indicators that can be used for curative purposes. The system takes a minimum set of information as input and automatically queries a global repository, classifies each publication into local-only, matched, or global-only, and applies automated rules to pinpoint the likely reasons for any discrepancies. This enables institutions to curate the related metadata by improving them or correcting institutional attributions where needed. Designed to be easily extended to additional repositories and new explanatory rules, the tool is available on GitHub, allowing for community-driven enhancements over time.
In the next section, we introduce the DOI Screening System and demonstrate its utility by comparing two local repositories to one global repository.

DOI Screening System

Description of the system

We present a *DOI Screening System* that automatically compares DOIs from a local and a global repository to identify which DOIs are missing from each source, as well as the reasons for these gaps. By classifying publications into local-only, matched, and global-only and then applying a set of predefined rules, the system provides insights that allow deriving actionable, "curative" steps to improve metadata accuracy and institutional coverage. The code is publicly available on GitHub (https://github.com/gaelbernard/DOI-screener), and can be extended to any repository or adapted with new rules as needed. Figure 2 shows an overview of the system.

Figure 2. Overview of the DOI Screening System.

Input. The system requires three minimal inputs: the ROR ID (Research Organization Registry Identifier), capturing the institution of interest, a year range specifying the temporal scope of the analysis, and a list of DOIs from the local repository, structured as a list of lists to accommodate multiple DOIs per publication. These inputs minimize setup complexity while still enabling a robust overlap analysis. A common source of error is incorrectly formatted DOIs, which may fail to be matched even after basic normalization. In addition, a limitation of the system is that it currently does not work for publications that do not have any DOIs.

Steps. Figure 2 illustrates the overall workflow, which consists of four main steps. First, the system queries the global repository using the specified ROR ID and chosen year range to collect all corresponding DOIs. Our system uses a deterministic approach to match DOIs, applying minimal text normalization (e.g., converting to lowercase, removing the "https://doi.org/" prefix). In its current iteration, the system uses only OpenAlex as a global repository (Priem et al., 2022), but it can readily be adapted to incorporate other data sources. Second, based on these global DOIs and the local repository's DOI list, it categorizes each publication into one of three sets: local-only (present locally but not globally within the expected time range and affiliation), matched (present locally and globally within the expected time range and affiliation), and global-only (present globally within the expected time range and affiliation, but not locally). Third, the system queries the global repository again—this time querying each unmatched local DOI instead of filtering by institutional affiliation or publication year. The DOIs retrieved during this third step are placed in the local-only category, that hence contains DOIs present locally but not globally or present globally without the expected time range or affiliation. Finally, the system applies a predefined set of rules (see Table 1) to diagnose why a DOI may not appear in both sources. Such a diagnose can be used to identify underrepresented output or inaccurate metadata and to target curation actions.

Table 1. List of rules implemented in the system.

Rule name	*Result from screening*	*DOI is present in Local list (L) or Global repo (G)*	*Description*
L-other	Unmatched	L	DOI from L that does not satisfy any other rule
L-prefix	Unmatched	L	DOI has a DOI-prefix that is significantly more frequently unmatched than matched (odds ratio)
L-time	Unmatched	L & G	DOI is outside the time range in G
L-inst	Unmatched	L & G	DOI is not affiliated with the institution in G
Matched	Matched	L & G	DOI is affiliated with institution and is within time range in G
G-prefix	Unmatched	G	DOI has a DOI-prefix that is significantly more frequently unmatched than matched (odds ratio)
G-type	Unmatched	G	DOI has a public. type that is significantly more frequently unmatched than matched (odds ratio)
G-authors	Unmatched	G	DOI has an author (identified through ORCID) that is affiliated with the institution in at least one of the matched DOIs in the same year
G-other	Unmatched	G	DOI from G that does not satisfy any other rule

These rules focus on issues such as misattributed affiliation (L-inst), out-of-range publication years (L-time), potential underrepresentation of certain sources identified through DOI prefixes (L-prefix / G-prefix), or of certain publication types in the local repository (G-type). Users can tailor or expand these rules to distinguish specific prefixes, to address unique metadata fields, or institution-specific patterns.

Output. The DOI Screening System provides two main outputs. The first is an *Overlap Bar Chart* visible in Figure 3 that shows how many publications fall under each rule, offering an immediate snapshot of the most common coverage or metadata issues. The second output is a detailed report that not only identifies which DOIs match each rule, but also provides additional information specific to each rule, such as the list of problematic prefixes. This enables librarians and research administrators to pinpoint and correct specific issues, such as updating metadata fields or resolving institutional attribution errors. Overall, this semi-automated approach offers both descriptive insights (quantifying the degree of overlap) and actionable items (pinpointing causes for mismatches) that enable curative measures, helping

institutions to curate data sources and to maintain robust, accurate bibliometric records. Although all rules are tested on each publication and appear in the detailed report, the order in which these rules are applied affects the distribution of publications in the bar chart output.

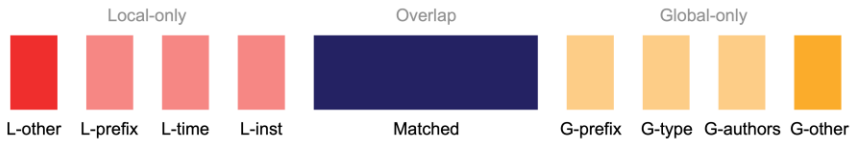

Figure 3. Bar chart representation for the overlap analysis with DOIs categorisation.

Two case studies

We applied the DOI screening system to the local repositories Infoscience at EPFL (ROR ID: https://ror.org/02s376052) and the Research Collection at ETH Zurich (ROR ID: https://ror.org/05a28rw58). OpenAlex served as the global repository, and the analysis covered the publication period from 2019 to 2023. For ETH Zurich, 46,579 publications were analyzed, with 9,518 (20.4%) not found in OpenAlex and 22,410 from OpenAlex not appearing in the local repository. At EPFL, 24,151 publications were analyzed, of which 5,369 (22.2%) did not appear in OpenAlex and 12,345 were found in OpenAlex but not in the local repository. The resulting Overlap Bar Charts are visible in Figure 4.

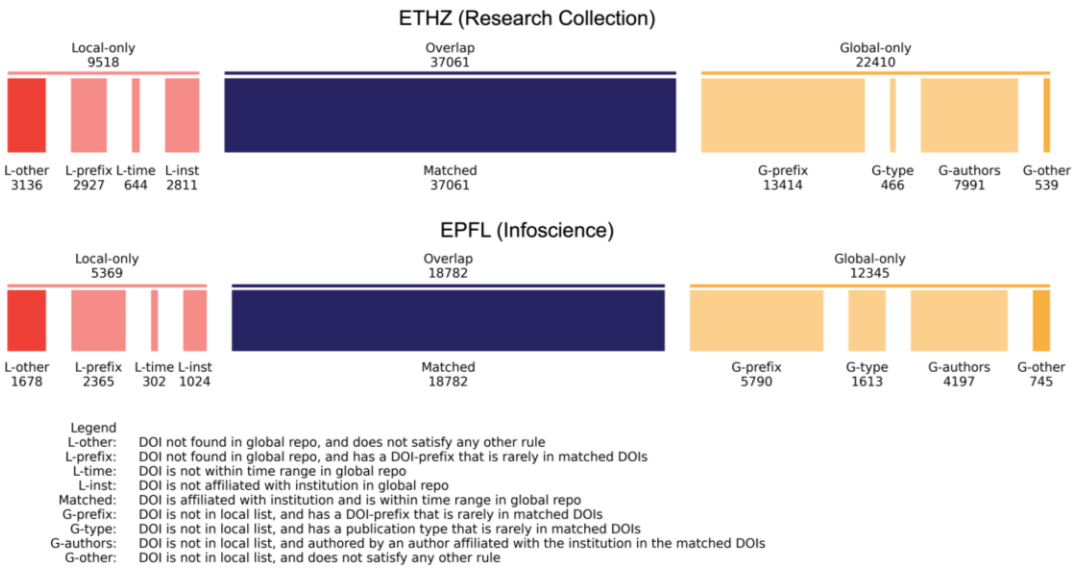

Figure 4. Visual output of the DOI Screening System based on OpenAlex, for DOI lists from local repositories of ETH Zurich and EPFL.

The system's predefined rules provide more detailed explanations for these mismatches that allow for informed decisions about the next curation steps. At EPFL, 44.0% (2,365) of local-only publications fell into the L-prefix category. Specifically, the system identified two prefixes, "10.5075" and "10.5281", that account for all

these mismatches: for example, “10.5075” appears in 2,280 local records but only 3 times in the global repository. Upon further investigation, we discovered that these prefixes are associated with EPFL theses, explaining why they are not automatically indexed in OpenAlex. Another 5.6% (302) were flagged as L-time, indicating that their publication dates fell outside the specified period and may require verification. In addition, 19.1% (1,024) were assigned L-inst, suggesting that these publications appear in OpenAlex but are not linked to EPFL, potentially requiring a curative action in the form of affiliation updates from the global repository. The remaining 31.3% (1,678) could not be explained by the current rules (L-other). Among global-only items, 46.9% (5,790) fell under G-prefix; for example, prefix “10.7910” appears 258 times in OpenAlex but never in the local repository, pointing to possible ingestion of new data sources locally. Another 13.1% (1,613) were categorized as G-type, as the system tagged peer-review, dataset, paratext, book-chapter, or preprint as underrepresented in the local repository. A further 34.0% (4,197) were flagged as G-authors, potentially indicating a missing research output in the local repository or a misattribution of affiliation in the global repository. The final 6.0% (745) were labeled G-other.

A parallel analysis at ETH Zurich revealed similar patterns, including publications not found in one source due to prefix or affiliation reasons, but with a notably lower percentage (2.0%) in the G-type category compared to 13.1% at EPFL. These results highlight how institutional policies or repository practices may influence coverage. Overall, the case studies demonstrate the value of the DOI Screening System in diagnosing coverage gaps, identifying metadata errors, and guiding targeted interventions to improve alignment between local and global repositories using a minimal set of input data.

Related Works

In this section we highlight some previous overlap analyses that are closest to our paper. Bologna et al. (2022) characterize studies on the coverage of global repositories including Web of Science, Scopus, Dimensions, Google Scholar and Microsoft Academic. Such analyses usually aim to determine the biases and provide an understanding of the selection processes at hand in global repositories (see also Delgado-Quirós, L. et al., 2024, Martín-Martín et al., 2021). They help researchers select the most appropriate data source and better evaluate the scope and limitations of the indicators they compute using such sources (Hug et al., 2017). Such analyses require strategies to match as many records as possible (see for instance Guerrero-Bote et al., 2021), in a context where global repositories are considered as relatively stable research objects.

Overlap analyses typically focus on comparing publications output or citations included in different repositories. In this study, we take a step further by categorizing unmatched publications according to a set of rules aiming to identify potential for improvements, an approach we characterize as *curative*. Descriptive and curative approaches should not be strictly contrasted, as recent papers focusing on open data sources illustrate. For example, Alperin et al. (2024) do not only explicitly compare OpenAlex and Scopus but also address critically the weaknesses of OpenAlex such

as accuracy and completeness of metadata. Hug and Brändle (2017) and Andreose et al. (2025) on the other hand explicitly compare a university's institutional bibliographic repository with Microsoft Academic or OpenCitations, respectively, as global repositories. This comes closest to our comparison of local and global repositories. With our curative approach, the goal is not primarily to describe the coverage and biases of the considered data sources regarding their suitability for research assessment. Instead, we explicitly aim to identify gaps and, ultimately, contribute to improve the coverage and metadata completeness of the considered data sources by enriching both local and global repositories. Such an approach emerges in a context where some open global repositories make their code for selection and indexation freely available, which allows not only for more transparency than commercial alternatives, but also allows to directly contribute to the improvement by identifying gaps. In contrast to strictly descriptive approaches, this curative approach has the side effect that it may render the results of an analysis obsolete shortly after being performed, but with the benefit that it can improve the considered data sources.

Conclusion and Future Directions

To identify what is missing when relying solely on a local or global repository, we propose a DOI Screening System that provides rapid overlap analysis and suggests "curative" actions to improve or interpret mismatches. The tool requires minimal input: an institutional identifier, a list of DOIs, and a time range. Since the system's code is publicly available, we expect community-driven enhancements to refine and expand the predefined rules, thereby increasing the portion of matched publications. Our immediate plans include extending the DOI Screening System, that currently only handles OpenAlex as a global repository, by incorporating OpenAIRE. We will also deepen the existing case studies, and explore how future iterations of the system could handle other units of analysis, such as a researcher's ORCID or a journal's ISSN. By embedding rules that specifically address gaps in the overlap analysis, the DOI Screening System serves as a catalyst for enhancing both the coverage and quality of local and global repositories, ultimately fostering more effective dissemination of scientific publications.

Acknowledgments

Contributors from the library of ETH Zurich worked on this paper as part of the TOBI project (Towards Open Bibliometric Indicators), co-funded by swissuniversities and the ETH Library.

References

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

Andreose, E., Di Marzo, S., Heibi, I., Peroni, S. & Zilli, L. (2025). Analysing the coverage of the University of Bologna's publication metadata in an existing source of open research information. *arXiv*. https://doi.org/10.48550/arXiv.2501.05821.

Bologna, F., Di Iorio, A., Peroni, S. & Poggi, F. (2022). Open bibliographic data and the Italian National Scientific Qualification: measuring coverage of academic fields. *Quantitative Science Studies*, 3 (3), 512–528. https://doi.org/10.1162/qss_a_00203.

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., López-Cózar, E. D., Orduña-Malea, E., & Ortega, J. L. (2024). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *Journal of the Association for Information Science and Technology*, 75(1), 43–58. https://doi.org/10.1002/asi.24839.

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. https://doi.org/10.3389/frma.2020.593494.

Hug, S.E., Ochsner, M. & Brändle, M.P (2017). Citation analysis with Microsoft Academic. *Scientometrics*, 111, 371–378. https://doi.org/10.1007/s11192-017-2247-8.

Hug, S.E. & Brändle, M.P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, 113, 1551–1571. https://doi.org/10.1007/s11192-017-2535-3.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126, 871–906. https://doi.org/10.1007/s11192-020-03690-4.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833.

Vieira, G.A. & Leta, J. (2024). *Biblioverlap*: an R package for document matching across bibliographic datasets. *Scientometrics*, 129, 4513–4527. https://doi.org/10.1007/s11192-024-05065-5.