Accuracy of affiliation information in open bibliographic data sources: A comparative study of OpenAlex and OpenAIRE for Leiden University

Nees Jan van Eck [1], Bram van den Boomen [1], Martijn Visser [1]

[1] *Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands*

Introduction

OpenAlex[1] and OpenAIRE[2] are prominent open bibliographic data sources that index a large collection of scholarly works, sources, authors, and institutions. They offer public access to their data with no or only very limited usage restrictions. OpenAlex data is available under a CC0 public domain declaration for unrestricted use and distribution, and OpenAIRE data is available under a CC-BY license, with certain elements also reusable under CC0[3]. The fact that OpenAlex and OpenAIRE data can be used without significant restrictions is important for several reasons. It allows anyone to freely share, reuse, and build on the data, fostering open science practices. It promotes transparency and scrutiny, allowing the community to verify the accuracy of the data, and to contribute to its improvement. It reduces the dependence on proprietary data providers, contributing to more equitable access to data. This is particularly advantageous for researchers in resource-limited settings who may lack access to expensive proprietary data sources. By openly providing their data without significant restrictions, OpenAlex and OpenAIRE contribute to more transparent, collaborative, and equitable access to research information.

Despite their shared commitment to open and comprehensive indexing of research outputs, OpenAlex and OpenAIRE differ in coverage and in the level of attention they have received from the research community. OpenAlex has been the focus of several studies examining its features, coverage, and data quality (e.g., Alperin et al., 2024; Culbert et al., 2024, Gusenbauer, 2024). In contrast, OpenAIRE has attracted less community attention.

Little is known about the completeness and accuracy of the affiliation information in OpenAIRE. Kramer (2024) demonstrated that OpenAlex and OpenAIRE complement each other in covering research outputs from Dutch research-performing organizations (RPOs), but her study focused primarily on the number of records attributed to an RPO in each data source, without assessing the accuracy of these affiliations. For instance, when a research output was found in both OpenAlex and OpenAIRE but the affiliation with a particular RPO was identified in only one data source, the study did not investigate whether this data source had correctly identified the affiliation. The present paper aims to address this gap by evaluating the accuracy

[1] https://openalex.org
[2] https://www.openaire.eu
[3] https://graph.openaire.eu/docs/next/license/

of affiliation information in OpenAlex and OpenAIRE for research outputs attributed to Leiden University.

Data and method

We used the OpenAlex snapshot from August 2024, available in the Google BigQuery instance of CWTS[4], and the OpenAIRE Graph Dataset version 9.0.0 (Manghi et al., 2025), also accessible via Google BigQuery (Mannocci & Mazoni, 2024).

Using an SQL script, we selected all publications indexed in both data sources by matching on DOI. Publications without a DOI in one of the data sources were excluded. We included only publications with a publication year in OpenAlex between 2020 and 2023.

Next, for each of the selected publications, we determined both for OpenAlex and for OpenAIRE whether they are attributed to Leiden University (LU). The following ROR IDs were used: https://ror.org/027bh9e22 (Leiden University), https://ror.org/05xvt9f17 (Leiden University Medical Center), and https://ror.org/03es66g06 (Leiden Observatory). These ROR IDs were selected because they represent the primary institutions that are commonly associated with LU's research activities. In OpenAlex these three ROR IDs correspond to three institutions, whereas in OpenAIRE they are linked to six organizations. We collected all selected publications linked to the three institutions in OpenAlex, as well as those linked to the six organizations in OpenAIRE.

Results

Our analysis of publications indexed in OpenAlex and OpenAIRE and attributed to LU revealed differences between the two data sources. Of the 42,215 publications attributed to LU across both data sources, 30,374 publications (72%) are attributed to LU in both OpenAlex and OpenAIRE, 4,548 (11%) are attributed to LU exclusively in OpenAlex, and 7,293 (17%) are attributed to LU exclusively in OpenAIRE. These discrepancies are visually depicted in Figure 1. A further examination of the three subsets of publications is presented in the following subsections.

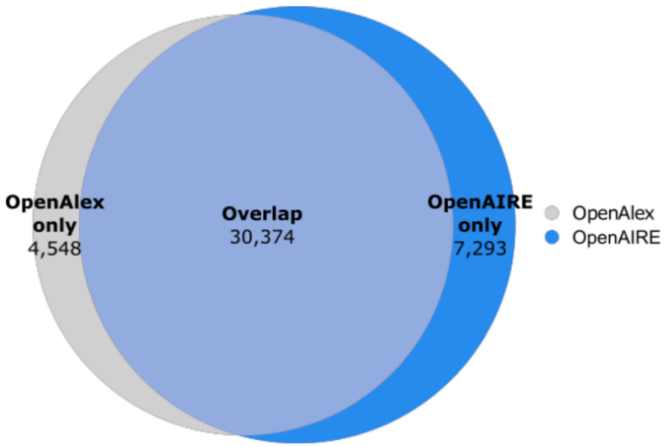

Figure 1. Comparison of publications in OpenAlex and OpenAIRE attributed to LU (2020 2023).

[4] https://console.cloud.google.com/bigquery?ws=!1m4!1m3!3m2!1scwts-leiden!2sopenalex_2024aug

Publications attributed to LU in both OpenAlex and OpenAIRE

Of the 30,374 publications in the overlapping subset, the majority (88%) are journal publications, including 75% articles, 9% reviews, 2% letters, and 1% errata. Additionally, 5% of the publications are preprints hosted in repositories like arXiv, bioRxiv, and medRxiv.

Given that publications in the overlapping subset are attributed to LU in both OpenAlex and OpenAIRE, it is reasonable to assume they include a LU affiliation. To confirm this, we examined a random sample of 30 publications from the overlapping set. For each publication, we inspected the landing page and the PDF. In this way, we confirmed that LU is listed as an affiliation in all 30 publications.

Publications attributed to LU exclusively in OpenAlex

Of the 4,548 publications attributed to LU exclusively in OpenAlex, we also examined a random sample of 30 publications. As before, we manually inspected the landing pages and PDFs of the publications and verified whether LU is listed as an affiliation. Table 1 summarizes the results of our manual inspection.

Our inspection confirmed that all 30 publications contain a LU affiliation, indicating that OpenAlex correctly attributes these publications to LU, whereas OpenAIRE incorrectly does not attribute them to LU. The sample of publications includes abstracts and conference-related contributions published in special or supplementary journal issues (11), journal articles (7), preprints (7), and book chapters (5).

Since OpenAIRE does not expose raw affiliation strings, we were unable to determine whether the discrepancies resulted from missing affiliation data or from the algorithm[5] OpenAIRE uses to link organizations to publications. However, in most cases (26 of the 30 publications), the publications are not linked to any organization in OpenAIRE.

Table 1. Results of the manual validation of a sample of 30 publications attributed to LU exclusively in OpenAlex.

OpenAlex LU attribution	**No. of pub.**	**Observation**
Correct	26	LU affiliation is included in the publication. No linked organizations in OpenAIRE.
Correct	4	LU affiliation is included in the publication. Not linked to LU in OpenAIRE, but there are other linked organizations in OpenAIRE.
Incorrect	0	-

Publications attributed to LU exclusively in OpenAIRE

Similarly, for the 7,293 publications attributed to LU exclusively in OpenAIRE, we examined a random sample of 30 publications. Inspection of the landing pages and the PDFs revealed that 19 publications list a LU affiliation, while 11 do not. This means that OpenAIRE correctly

[5] https://graph.openaire.eu/docs/graph-production-workflow/enrichment-by-mining/affiliation_matching

attributes about two-third of the sampled publications to LU, while OpenAlex incorrectly does not attribute them to LU. Table 2 summarizes the results of our manual inspection.

Of the 19 publications correctly attributed to LU by OpenAIRE but not by OpenAlex, six are preprints (primarily from arXiv) for which no raw affiliations strings are available in OpenAlex. Nine are journal articles suffering from the following issues: in one case, no raw affiliation strings are available in OpenAlex; in five cases, the LU affiliation is missing among the raw affiliation strings; and in two cases, the correct raw affiliation string is available but OpenAlex did not link it to LU.

The 11 publications incorrectly attributed to LU by OpenAIRE and correctly not attributed to LU by OpenAlex are all journal articles. It is unclear why OpenAIRE attributes these publications to LU.

Table 2. Results of the manual validation of a sample of 30 publications attributed to LU exclusively in OpenAIRE.

OpenAIRE LU attribution	**No. of pub.**	**Observation**
Correct	15	LU affiliation is included in the publication. Missing raw affiliation string in OpenAlex.
Correct	4	LU affiliation is included in the publication. Correct raw affiliation string in OpenAlex, but linked to LU.
Incorrect	11	LU affiliation is not included in the publication.

Precision and recall analysis

Using the number of publications attributed to LU in OpenAlex and OpenAIRE, combined with the results of the manual inspection of the 90 sampled publications, we estimated the precision and recall for both data sources. Table 3 presents the precision and recall estimates for LU publications in each data source.

Table 3. Precision and recall estimates for LU publications in OpenAlex and OpenAIRE.

	OpenAlex	**OpenAIRE**
True positives	$\frac{30}{30} \times 4{,}548 + \frac{30}{30} \times 30{,}374 = 34{,}922$	$\frac{19}{30} \times 7{,}293 + \frac{30}{30} \times 30{,}374 = 34{,}993$
False positives	$\frac{0}{30} \times\ 4{,}548\ =\ 0$	$\frac{11}{30} \times\ 7{,}293\ =\ 2{,}674$
False negatives	$\frac{19}{30} \times\ 7{,}293\ =\ 4{,}619$	$\frac{30}{30} \times\ 4{,}548\ =\ 4{,}548$
Precision	$\frac{34{,}922}{34{,}922 + 0} =\ 100\%$	$\frac{34{,}993}{34{,}993 +\ 2{,}674}\ =\ 93\%$

Recall	$\frac{34{,}922}{34{,}922 + 4{,}619} = 88\%$	$\frac{34{,}993}{34{,}993 + 4{,}548} = 88\%$

Table 3 shows that OpenAlex and OpenAIRE have the same recall (88%) for LU affiliated publications, while OpenAIRE has a lower precision (93%) compared to OpenAlex (100%), largely due to its higher rate of false positives.

Conclusions

We evaluated the accuracy of affiliation information in two prominent open bibliographic data sources, OpenAlex and OpenAIRE, using publications attributed to LU as a case study. Our analysis highlighted differences in how these data sources attribute publications to organizations, with a specific focus on LU.

OpenAlex and OpenAIRE demonstrated a similar recall (88%) for LU affiliated publications. Precision was higher for OpenAlex (100%) than for OpenAIRE (93%). These precision and recall estimates are based on the manual inspection of a sample of 90 publications. This inspection showed that OpenAIRE exhibited a higher rate of false positives, with one-third of the manually inspected publications incorrectly attributed to LU. The cause of OpenAIRE's false positives was hard to determine due to the lack of raw affiliation strings in OpenAIRE. OpenAIRE could increase transparency by including raw affiliation strings in its data and by providing provenance data indicating on what basis a publication has been attributed to a particular organization.

This study examined the accuracy of the attribution of publications to LU in OpenAlex and OpenAIRE, focusing on publications indexed in both data sources. Further research could investigate the accuracy for other organizations, varying by type and region. Additionally, it could include publications uniquely indexed in either OpenAlex or OpenAIRE. Such research could provide valuable insights for improving the accuracy and usefulness of these data sources.

In conclusion, continuous evaluation and improvement of open bibliographic data sources like OpenAlex and OpenAIRE is essential to ensure not only equitable and unrestricted access to research information but also the accuracy of such information.

Acknowledgements

We would like to thank Marc Luwel and Paolo Manghi for their helpful comments on earlier versions of this paper.

Data and code availability

The Google BigQuery SQL code used in this study and the manual validation data are available in the following GitHub repository: https://github.com/neesjanvaneck/OpenAlexOpenAIRE-LU-comparison.

References

Alperin, J., 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

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. *arXiv*. https://doi.org/10.48550/arXiv.2401.16359

Gusenbauer, M. (2024). Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases' citation indices. *Research Synthesis Methods*, *15*(5), 802-817. https://doi.org/10.1002/jrsm.1729

Kramer, B. (2024). Coverage and quality of open metadata for Dutch research output. *Zenodo*. https://doi.org/10.5281/zenodo.10629457

Manghi, P., Atzori, C., Bardi, A., Baglioni, M., Dimitropoulos, H., La Bruzzo, S., Foufoulas, I., Mannocci, A., Horst, M., Iatropoulou, K., Kokogiannaki, A., De Bonis, M., Artini, M., Lempesis, A., Ioannidis, A., Manola, N., Principe, P., Vergoulis, T., & Chatzopoulos, S. (2025). OpenAIRE Graph Dataset (9.0.0) [Data set]. *OpenAIRE*. https://doi.org/10.5281/zenodo.14582029

Mannocci, A., & Mazoni, A. (2024). OpenAIRE Graph Training for Scientometrics Research. *Zenodo*. https://doi.org/10.5281/zenodo.13981535