https://doi.org/10.51408/issi2025_173

Automating Reproducible Bibliometrics with the Open Research Converter

Jack H. Culbert[1], Philipp Mayr[2]

[1]*jack.culbert@gesis.org*, [2]*philipp.mayr@gesis.org*
GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667, Cologne (Germany)

Abstract

The Open Research Converter[1] (ORC) is an open-source tool that allows users to convert Digital Object Identifiers into OpenAlex[2] Work IDs and/or retrieve full bibliometric records from OpenAlex. In this poster paper, we introduce the ORC and show its main application: the generation of open and sharable bibliometric datasets, future development plans and a short analysis of usage patterns so far.

Introduction

Bibliometric and Scientometric studies which involve bibliometric data taken from proprietary databases (such as the Web of Science (WoS) or Scopus) suffer from a lack of openness, transparency, and reproducibility as researchers are not permitted to freely share and publish the underlying data from their analyses. Workarounds such as "we searched the [*query terms q*] and exported *n* records from WoS version *m*" have been utilised by the community but remain difficult a barrier to reproducibility as the underlying dataset from the study is unavailable.

Reproducibility in Bibliometrics and Scientometrics has been previously studied resulting in (Velden et al., 2018), and the current data sharing and publishing restrictions with the commercial providers are not likely to change in the short term. Consequently, bibliometric research based on WoS and Scopus data is likely to remain unreproducible and lacks the transparency which is required for Open Science research.

OpenAlex (Priem et al., 2022) was released in 2022 and is an open-source bibliometric database which releases its data under a maximally permissive license (CC0 1.0 Universal), which enables researchers to share their datasets. However, frictions for bibliometricians exist, including adapting to the website interface, the technical knowledge to utilise the API or raw data (provided by OurResearch[3] as a monthly approximately 300GB JSON snapshot), and a healthy suspicion of the quality of the bibliometric dataset.

The Open Research Converter (Culbert et al., 2024) was primarily designed to assist bibliometricians with the lattermost friction, allowing them convert DOIs from within their dataset to OpenAlex Work IDs, which can then be shared alongside their publications – increasing reproducibility and openness within the Scientometrics community and elsewhere. Since then, following community feedback at the Nordic Workshop on Bibliometrics & Research Policy 2024 (Culbert, 2024), we have been developing new features as detailed below.

[1] orc-demo.gesis.org
[2] openalex.org
[3] ourresearch.org

Figure SEQ Figure * ARABIC 1 – The Open Research Converter Interface, overlaid with a snippet from the full record output.

Open Research Converter

The ORC is a containerised Python application with a JavaScript frontend which allows researchers to input Digital Object Identifiers (DOIs) manually or upload a csv and returns either the OpenAlex Work ID or the full bibliographic record in csv format. (The codebase is available via Github.) [4]

The ORC provides bibliometrics researchers with the ability to use DOIs to identify the records in OpenAlex which match those in other databases. This approach has its limitations, as explored in (Vieira & Leta, 2024), such as missing or duplicated DOIs, and therefore we are working on a fuller approach which incorporates other publication metadata into the matching process.

DOIs accompany most bibliometric records in both proprietary academic databases such as the Web of Science (WoS), Scopus, and Dimensions and open databases such as PubMed, ArXiv, Semantic Scholar, OpenAIRE and OpenAlex. The degree of overlap and number of records without a DOI in WoS, Scopus and OpenAlex (and thereby the accuracy of this method) was explored in (Culbert et al., 2024).

The ORC backend is capable of processing over 300,000 records in a single request and is only limited by the size of the input CSV allowable in the frontend, to prevent abuse of the server.

Usage

We have been monitoring usage of the ORC and have found users accessing the ORC from around the globe, primarily from Europe and the US. So far, between August 31st 2024 and 9th April 2025 209,053 records have been processed from a total of 32 unique emails.

Future Development Goals

Fuzzy Matching

Instead of matching by DOI, we intend to implement a system which matches by Title, Author, Year and other identifying information, including a fuzzy matching step to allow for small differences in metadata, such as abbreviated names. This may be implemented alongside or directly as an optional BibTeX input.

Reference Lookups

Alongside the direct DOI to WorkID conversion, a feature allowing lookup of available references in OpenAlex for all papers identified is planned.

Reverse Lookups

Reversing the ORC to allow for libraries to identify which sources in OpenAlex are also in proprietary bibliometric databases via WorkID to DOI conversion has been requested and is in process of being implemented.

Network Visualisation

A planned extension of the ORC includes allowing for lightweight bibliometric analysis, transforming the ORC into an analysis platform. This includes incorporating a Neo4J instance into the codebase to allow for a visualisation of an OpenAlex dataset in the form of a graph.

Conclusion

The ORC enables bulk conversion of DOIs to OpenAlex WorkIDs, and allows for the generation of sharable research datasets,

[4] github.com/jhculb/Open-Research-Converter

increasing the reproducibility and openness of bibliometric research. It is being utilised by the Scientometrics community, and following user feedback is being expanded into an open-source, lightweight analysis platform for bibliometric analyses.

Acknowledgments

This work was funded by the Federal Ministry of Education and Research via funding numbers: 16WIK2301B / 16WIK2301E, The OpenBib project (Schmidt et al., 2024).
The authors thank Ahsan Shahid for building the web interface of the ORC and Nina Smirnova for providing the problem that inspired the ORC.

References

Culbert, J. H. (2024, November 26). *The Open Research Converter*. https://doi.org/10.5281/zenodo.14222479

Culbert, J. H., Shahid, M. A., & Mayr, P. (2024). *ORC: The Open Research Converter*. https://orc-demo.gesis.org/paper

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:2401.16359). arXiv. https://doi.org/10.48550/arXiv.2401.16359

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

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis*. Zenodo. https://doi.org/10.5281/zenodo.13932928

Velden, T., Hinze, S., Scharnhorst, A., Schneider, J. W., & Waltman, L. (2018). Exploration of reproducibility issues in scientometric research. *STI 2018 Conference Proceedings*, 612–624.

Vieira, G. A., & Leta, J. (2024). biblioverlap: An R package for document matching across bibliographic datasets. *Scientometrics*, *129*(7), 4513–4527. https://doi.org/10.1007/s11192-024-05065-5