The Journal of Open Source Software

ORC: The Open Research Converter

Jack H. Culbert [1¶], **Muhammad Ahsan Shahid** [1], **and Philipp Mayr** [1]

1 GESIS – Leibniz Institute for the Social Sciences ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Tristan Miller

Reviewers:

- @evamaxfield
- @sarahsid10
- @dmnapolitano

Submitted: 07 November 2024
Published: unpublished

Summary

The Open Research Converter (ORC) is a tool designed to allow researchers, developers and others using bibliographic data to bulk convert their data to a shareable format utilising OpenAlex.

Statement of need

Bibliometrics and in particular Scientometrics suffers from a lack of reproducibility, wherein the databases used to perform bibliometrics are often proprietary and therefore bound by copyright and access agreements which forbid sharing the underlying data used to create the scientific insights shared in papers.

OpenAlex (Priem et al., 2022), released in 2022, is a open-source bibliometric database compiled by Our Research which releases its data with a maximally permissive copyright (specifically under the CC0 1.0 Universal deed), allowing free sharing of all data. This has allowed bibliometric researchers to download and interrogate the data as they see fit, and enables sharing of data.

However, dealing with OpenAlex data can be cumbersome. The methods of access are currently via the website, API, or a data dump, each of which have challenges for researchers associated with it. Namely, to use the website limits the amount of information available to be displayed and may require downloading and then processing the data further to achieve the desired insights, to use the API requires a level of technical knowledge and is rate limited by OpenAlex, and the data dumps are very large (approximately 300GB at time of writing) and also require technical knowledge in the processing and interrogation of the data.

Easing the barrier of access to OpenAlex is a current theme of work in the bibliometrics community, for example Massimo et al. (2024) have created a tool in the R programming language, openalexR, capable of bulk collection of OpenAlex data and processing this data from OpenAlex's JSON based data format to a tabular format. Similarly OpenAlex Networks (Silva, 2023) is a Python library for generation of OpenAlex datasets and processing of citation and coauthorship networks. OpenAlexNet is a C# wrapper for OpenAlex enabling searching of OpenAlex.

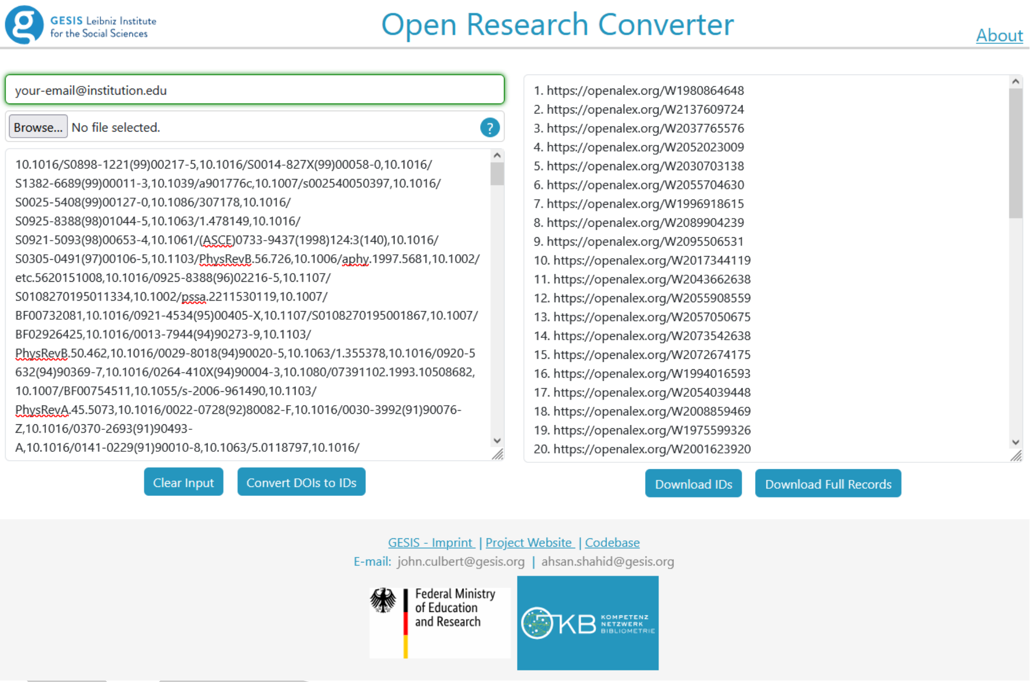
Currently OpenAlex has no easy method for researchers to convert their datasets from proprietary formats to OpenAlex. While it is possible to manually convert smaller datasets using OpenAlex's website, or download the OpenAlex data dump and process this to enable matching.

We provide here in the Open Research Coverter a tool utilising the OpenAlex API enabling simple bulk conversion of bibliometric data (DOIs) to a shareable format.

Functionality

The Open Research Converter is a containerised Python and Javascript based tool which when run serves a webpage allowing a user to enter either a string of DOIs via copy and paste, or by uploading a correctly formatted CSV file. The user can then convert these to OpenAlex WorkIDs or retrieve the full bibliographic record from OpenAlex.

The tool has been tested on datasets of 100,000 DOIs and was stable. At time of writing, a running version of the ORC can be found at orc-demo.gesis.org, and the code is released here on Github under a GPL-3.0 license.


GESIS Leibniz Institute for the Social Sciences
Open Research Converter
About
your-email@institution.edu
Browse... No file selected.
?
Clear Input
Convert DOIs to IDs
Download IDs
Download Full Records
GESIS - Imprint | Project Website | Codebase
E-mail: john.culbert@gesis.org | ahsan.shahid@gesis.org
Federal Ministry of Education and Research
KB KOMPETENZ NETZWERK BIBLIOMETRIE

```
10.1016/S0898-1221(99)00217-5,10.1016/S0014-827X(99)00058-0,10.1016/
S1382-6689(99)00011-3,10.1039/a901776c,10.1007/s002540050397,10.1016/
S0025-5408(99)00127-0,10.1086/307178,10.1016/
S0925-8388(98)01044-5,10.1063/1.478149,10.1016/
S0921-5093(98)00653-4,10.1061/(ASCE)0733-9437(1998)124:3(140),10.1016/
S0305-0491(97)00106-5,10.1103/PhysRevB.56.726,10.1006/aphy.1997.5681,10.1002/
etc.5620151008,10.1016/0925-8388(96)02216-5,10.1107/
S0108270195011334,10.1002/pssa.2211530119,10.1007/
BF00732081,10.1016/0921-4534(95)00405-X,10.1107/S0108270195001867,10.1007/
BF02926425,10.1016/0013-7944(94)90273-9,10.1103/
PhysRevB.50.462,10.1016/0029-8018(94)90020-5,10.1063/1.355378,10.1016/0920-5
632(94)90369-7,10.1016/0264-410X(94)90004-3,10.1080/07391102.1993.10508682,
10.1007/BF00754511,10.1055/s-2006-961490,10.1103/
PhysRevA.45.5073,10.1016/0022-0728(92)80082-F,10.1016/0030-3992(91)90076-
Z,10.1016/0370-2693(91)90493-
A,10.1016/0141-0229(91)90010-8,10.1063/5.0118797,10.1016/
```

```
1. https://openalex.org/W1980864648
2. https://openalex.org/W2137609724
3. https://openalex.org/W2037765576
4. https://openalex.org/W2052023009
5. https://openalex.org/W2030703138
6. https://openalex.org/W2055704630
7. https://openalex.org/W1996918615
8. https://openalex.org/W2089904239
9. https://openalex.org/W2095506531
10. https://openalex.org/W2017344119
11. https://openalex.org/W2043662638
12. https://openalex.org/W2055908559
13. https://openalex.org/W2057050675
14. https://openalex.org/W2073542638
15. https://openalex.org/W2072674175
16. https://openalex.org/W1994016593
17. https://openalex.org/W2054039448
18. https://openalex.org/W2008859469
19. https://openalex.org/W1975599326
20. https://openalex.org/W2001623920
```

Figure 1: Homepage for the ORC.

Research Projects

The Open Research Converter has been used in the release of two datasets: Culbert et al. (2024) which complements Gupta et al. (2024) and Smirnova et al. (2024) which complements Mir et al. (2024).

Acknowledgements

This work was funded by the Federal Ministry of Education and Research via funding numbers: 16WIK2301B / 16WIK2301E, The OpenBib Project (Schmidt et al., 2024). We acknowledge support by Federal Ministry of Education and Research, Germany under grant number 01PQ17001, the Competence Network for Bibliometrics.

Jack Culbert and Philipp Mayr received additional funding by the European Union under the Horizon Europe grant OMINO – Overcoming Multilevel INformation Overload under grant number 101086321 (Hołyst et al., 2024).

Our thanks go to Nina Smirnova for the initial inspiration for this project.

Contributor Role Taxonomy (CRediT)

Jack H. Culbert: Conceptualization (lead); Investigation (lead); Methodology (lead); Software (equal); Visualization (supporting) Writing - Original Draft Preparation (lead); Writing - Review and Editing (equal). **Muhammad Ahsan Shahid**: Software (equal); Visualization (lead); Writing - Review and Editing (equal). **Philipp Mayr**: Project Administration (lead); Supervision (lead); Writing - Review and Editing (equal).

References

Culbert, J. H., Gupta, S., Kanaujia, A., Lathabai, H. H., Kumar Singh, V., & Mayr, P. (2024). *Open AI literature 2010-2020 dataset* (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10997451

Gupta, S., Kanaujia, A., Lathabai, H. H., Singh, V. K., & Mayr, P. (2024). Patterns in the growth and thematic evolution of artificial intelligence research: A study using bradford distribution of productivity and path analysis. *International Journal of Intelligent Systems*, *2024*(1), 5511224. https://doi.org/10.1155/2024/5511224

Hołyst, J. A., Mayr, P., Thelwall, M., Frommholz, I., Havlin, S., Sela, A., Kenett, Y. N., Helic, D., Rehar, A., Maček, S. R., Kazienko, P., Kajdanowicz, T., Biecek, P., Szymanski, B. K., & Sienkiewicz, J. (2024). Protect our environment from information overload. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01833-8

Massimo, A., Le Trang, Corrado, C., Alessandra, B., & June, C. (2024). openalexR: An r-tool for collecting bibliometric data from OpenAlex. *The R Journal*, *15*, 167–180. https://doi.org/10.32614/RJ-2023-089

Mir, A. A., Smirnova, N., Jeyshankar, R., & Mayr, P. (2024). The rise of indo-german collaborative research: 1990–2022. *Global Knowledge, Memory and Communication*. https://doi.org/10.1108/GKMC-09-2023-0328

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (No. 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

Silva, F. N. (2023). OpenAlex networks (openalexnet). In *GitHub repository*. GitHub. https://github.com/filipinascimento/openalexnet

Smirnova, N., Culbert, J. H., & Mayr, P. (2024). *Indo-german literature dataset* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10607235