3

May 09, 2022

Coverage of open citations in DOAJ journals - Protocol V.3

Constance Dami[1], Alessandro Bertozzi[1], Chiara Manca[1], Umut Kucuk[1]

[1]University of Bologna

dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3

Open Science 2021/2022

Chiara Manca

This protocol refers to a research done for the Open Science course 21/22 of the University of Bologna.

Citation: Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of open citations in DOAJ journals - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3

This is the protocol for the research of the coverage of open citations in DOAJ journals. Our goal is to find out:

- about the coverage of articles from open access journals in DOAJ journals as citing and cited articles,
- how many citations do DOAJ journals receive and do, and how many of these citations involve open access articles as both citing and cited entities,
- as well as the presence of trends over time of the availability of citations involving articles published in open access journals in DOAJ journals.

Our research focuses on DOAJ journals exclusively, using OpenCitations as a tool. Previous research has been made on open citations using COCI (Heibi, Peroni & Shotton 2019), and on DOAJ journals' citations (Saadat and Shabani 2012), paving the grounds for our present analysis.

Minimal Bibliography

Björk, B.-C.; Kanto-Karvonen, S.; Harviainen, J.T. "How Frequently Are Articles in Predatory Open Access Journals Cited." *Publications*, *8*, 17. (2020) https://doi.org/10.3390/publications8020017

Heibi, I.; Peroni, S.; Shotton, D. "Crowdsourcing open citations with CROCI -- An analysis of the current status of open citations, and a proposal" arXiv:1902.02534 (2019) https://doi.org/10.48550/arXiv.1902.02534

Saadat, R., A. Shabani. "Investigating the citations received by journals of Directory of Open Access Journals from ISI Web of Science's articles." *International Journal of Information Science and Management (IJISM)* 9.1 (2012): 57-74.

Solomon, D. J., Laakso, M., Björk, B.-C. "A longitudinal comparison of citation rates and growth among open access journals", *Journal of Informetrics*, 7, 3 (2013): 642-650. https://doi.org/10.1016/j.joi.2013.03.008.

DOI

dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3

Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk 2022. Coverage of open citations in DOAJ journals - Protocol. **protocols.io** https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3
Chiara Manca

citations, OpenCitations, DOAJ, open access, journals, open science

protocol ,

May 06, 2022

May 09, 2022

May 09, 2022 Chiara Manca

62117

:

This protocol refers to a research done for the Open Science course 21/22 of the University of Bologna.

Data Gathering

1 We download the public data dump of DOAJ articles in .tar.gz format and the public data dump of OpenCitations' COCI in zipped files containing several .csv files. The other index of OpenCitation, CROCI, is not available as public data dump, so for this research we consider only COCI.

Downloading from DOAJ was not problematic. A few issues had to be addressed with COCI. The first problem we faced is how to obtain these citations directly from the platform. The possibilities are essentially reduced to two: either through the API service or through the direct download of the entire dataset, made free and accessible by the platform itself.

The first one, after several attempts, has been discarded. The DOIs of our interest, coming from DOAJ, are approximately 5 million. Considering that each request containing a DOI addressed to the OC API takes an average of 1.2 seconds to return a response, it would be impossible to quickly conclude the collection of all the data needed for research purposes. OC implements beyond the API a SPARQL endpoint, immediately discarded for the same reasons just explained. In fact this system, more dated than the API, has response times much longer than the latter, despite allowing greater flexibility in research.

The only solution that could be reconciled with the search times is **the direct download of the entire dataset**. The main problem with this option is the amount of data to be preserved on the hard drive. Especially if you try to read the data by unzipping the content.
The choice to focus on the latter method, then, is mainly to be found in the timing to complete the data collection.

1.1 From the **DOAJ dump** we create a **DOAJ_journals.json.** Inside this JSON

Citation: Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of open citations in DOAJ journals - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3

file we create a key for each journal, selecting: the issn (if it is present), eissn (if it is present), the title of the journal, the list of all the articles' DOIs.

We also create another file containing all articles's DOIs from DOAJ (**articles_DOIs.json),** which has just the aim to semplify next working steps.

1.2 Then we manage the **OC dump**. We unzip the first level of files, obtain a series of repository. In this way, we avoid extending the required memory for storing files.

Then we iterate over all the CSV files inside the zip directory and we obtain from them a series of data frames.
On each dataframe, we select only rows that have a DOAJ DOI in the *cited* or *citing* column. To accomplish this work we use the DOIs stored inside the **articles_DOIs.json**.

Once we apply this filter on dataframe we saved for each iteration two new CSV: one with only the DOAJ DOIs in the cited (**references.json**) column and another one with citing (**citations.json**) ones.

1.3 Then for convenience, we transform these latter CSV into a JSON format. In this way, we can reorganize all the content divided as the **DOAJ_journals.json**, by journals.

1.4 We then count the number of elements in **references.json** and **citations.json** and update the **DOAJ_journals.json** with the fields: reference_count and citation_count.

Then, we check if the *citing* DOI in the citations of **references.json** and the *cited* DOI in the citations of **citations.json** are open by looking at the *oa_link.*

The result will be **open_citations.json** and **open_references.json** with all of the citations for each journal and **DOAJ_journals.json** will be updated with a counter of the open citations coming and done by the journals (**open_citations_count**, **open_references_count**).

1.5 We combine open_citations.json and open_references.json. We group the citations based on the *creation_date,* taking just the year.
The result will be **open_access_citations_by_year**.**json** of all of these citations for every year that was found.

Data Visualization

2 We visualize our results with python libraries.
We specifically visualize **open_access_citations_by_year**.**json** in a graph that specifies the

number of citations in the y axis and the list of years in the x axis.

Publishing data

3 We publish all the CSV and JSON data that we gathered on our Data Management Plan in Zenodo and also in our Github repository.

Citation: Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of open citations in DOAJ journals - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v3