# Coverage of DOAJ journals' citations through OpenCitations - Protocol V.5

COMMENTS 0

DOI
dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

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

[1]University of Bologna

VERSION 5
DEC 14, 2022

WORKS FOR ME 1

Open Science 2021/2022

Chiara Manca

DISCLAIMER

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

ABSTRACT

This is the protocol for the research of the coverage of DOAJ journals' citations through OpenCitations.

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.

After careful considerations on the best way to retrieve data from DOAJ and OpenCitations, we opted for downloading the public data dumps. Using the API resulted in a way too long running time, and the same problem arose for using the SPARQL endpoint of OpenCitations.

**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

Pandita, R., & Singh, S. "A Study of Distribution and Growth of Open Access Research Journals Across the World. Publishing Research Quarterly" (2022), 38(1), 131–149. https://doi.org/10.1007/s12109-022-09860-x

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/v5

PROTOCOL CITATION

Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk 2022. Coverage of DOAJ journals' citations through OpenCitations - Protocol. **protocols.io** https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5
Version created by Constance Dami

KEYWORDS

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



CREATED

Jul 13, 2022

LAST MODIFIED

Dec 14, 2022

OWNERSHIP HISTORY

Dec 14, 2022 Chiara Manca

PROTOCOL INTEGER ID

66635

MATERIALS TEXT

This protocol uses the following Python libraries: tarfile, pandas, JSON, pickle, DateTime, zip file, and plotly.

The GitHub repository for our research software, including all python code mentioned in the protocol, is available here.

We used the data dump of DOAJ articles of May 01, 2022 and the data dump of DOAJ journals of May 07, 2022. The most recent ones can be found on the DOAJ website.

For Open Citations data, we used the COCI dump of March 2022. This dump, as well as the most recent one, is available on the OpenCitations website.

DISCLAIMER

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

BEFORE STARTING

Make sure to have Python 3.9 installed on your device.
All the dependecies of the script can be installed using the requirements.txt file stored into the github repository.

Computer technical specifications:
CPU: Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz
RAM: 20,0 GB (19,9 GB usable) 2666 mhz

## Data Gathering: DOAJ



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

1 **Collecting data from DOAJ**: we download data about journals and articles from the DOAJ website, and then refine it excluding all information that we are not interested in.

Expected result

1.1 We download the data dumps from DOAJ in .tar.gz. format.

Dataset

**DOAJ articles public data dump** NAME

https://doaj.org/public-data-dump/article LINK

Dataset

**DOAJ journals public data dump** NAME

https://doaj.org/public-data-dump/journal LINK

Both datasets contain metadata that is not useful for our research, so we need to filter only the necessary data.

1.2 From the **DOAJ dump**, we create a unique key for each journal by concatenating the issn and the eissn, having as values: the issn (if it is present), eissn (if it is present), the title of the journal, the subject of the journal and the list of all the articles' DOIs.

After opening the tarfile containing the data, for every journal, we extract only the information about **issn** and **eissn**, first making sure that there is always at least one of the two for each record in the dump:



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

We then add to the set of journals our unique identifier "**issn+eissn**"

1.3 We extract the data of the articles: we open the file with tarfile, then for each article, we collect the information about **issn** and **eissn** of the journal publishing it, as well as the **DOI** of the article:

If the article doesn't have any DOI registered, we add it to a list that we will store separately.

Otherwise we handle cases where the issn and eissn have been wrongly registered in the articles dump by aligning data with the journals set previously created.



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

We collect the subject of the journal.

Once all of the information are collected, we add them to our final json, adding a new key if it doesn't exist or adding it to the list of dois for the journal.

An example of an element in the final file:



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

Expected result

1.4 We create a file containing a dictionary with all DOAJ articles' DOIs from DOAJ as keys and the "issn+eissn" identifier of the journal who published it as value, to simplify the next steps.

Expected result

## Data Gathering: OpenCitations

2 **Collecting and filtering data from OpenCitations**: we take the data from the download section, on the OpenCitations website, and then refine them using the files obtained from the previous step.

Dataset

| | |
|---|---|
| COCI March 2022 Dump | NAME |
| https://doi.org/10.6084/m9.figshare.6741422.v14 | LINK |



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

Expected result

Expected result

Expected result



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

Expected result

## Filter Open Citations

2.1 We iterate all the records from the **Open Citations dump**, which have at least one doi in either the ***citing*** or ***cited*** column. For each directory:

1. We unpack all the zip directory files in a temporary folder and iterate all over the unzip CSV files:

2. We split the CSV file in two dataframes. For each dataframe we delete all the records that have a null value on the citing or cited column:

3. For each dataframe, we filter all records which have a DOAJ doi either in the *citing* or the *cited* column:

4. We add the journal name that matches the doi in the citing or cited column. Additionally, we add a column



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

for both the cited (isDOAJ_cited) and the citing column (isDOAJ_citing), for identifying which doi belongs to DOAJ for each record (only the one in the cited column, the one in citing column, or in the dois in both columns):

5. We merge the two dataframe in a single one with an outer join.

Expected result

Expected result

Expected result

## Group By Open Citations results

2.2 We iterate on each file of the **filtered** directory and for each one:

1. We transform the *creation* column into a date format:



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

2. We save and discard all the records that don't have any creation dates or have a date bigger than 2024:

3. We split the main dataframe in two sub-dataframe: one for the groupBy with only the year (df_normal); another one for the groupBy with both year and journal (df_by_journal).

Expected result

Expected result

Expected result



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

Expected result

## Concatenate all results

2.3 We concatenate, using the Pandas library, all the files in the **normal** repository and in the **by_journal** repository, to summarize all values in two dataframes:

We add to the df_by_journal the group of fields extracted from DOAJ for each journal, which adds useful information about the journal:

Finally, we concatenate all error files into one single file:



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

Expected result

Expected result

Expected result

## Add Ratios to the final results



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

2.4 1. We add ratios to the **normal.json:**

2. We add ratios to the **by_journal.json**:

Expected result

Expected result

## Add useful metrics

2.5 We add the following metrics to a JSON file, in order to provide a summary of useful research information about dois processed from DOAJ.

Expected result



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

## Data Visualization

3 We visualize our results in *line*, *bar* and *scatter* graphs with the use of the **plotly Python library**.
We load our json data from the **queried** folder in DataFrames of the pandas library.

3.1 We query the **final_df_journal** data frame to find the biggest of DOAJ in terms of the most number of citations, references, citations to DOAJ journals and citations from DOAJ journals.

Expected result



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

We create the **final_df_journal_1** data frame with the result of the query.

3.2 To have a better understanding of our data, we examine the most recurring subjects among DOAJ journals.

We represent it with a *line* plot.

Expected result

3.3 In order to examine the citations made by journals overall, regardless of the year, we group the journals by title and sum the relevant columns.



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

We then use *bar* plots to visualize the citations data about DOAJ journals

Expected result

3.4 We repeat step 3.3 with *scatter* plots, including information about the number of articles per journal.

Expected result

3.5 We examine the journals doing the most self-citations by year, using a *line* plot.

Expected result



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

3.6 To have a better comparison between the ***citing*** and ***cited*** of DOAJ journals in the last 20 years, we do some bar plots that stack the two amounts in the same column.

Expected result

3.7 We use a bar plot to visualize the timeline, in the last 20 years, of the number of citations involving DOAJ journals as both citing and cited entities and the percentage of it inside the number of general citations.

Expected result

3.8 We use a bar plot to show the number of errors we encountered in the project divided by category.

Expected result

## Publishing data



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5

4 We publish the following JSON files in Zenodo and also in our Github repository (**queried** folder).



**Citation:** Constance Dami, Alessandro Bertozzi, Chiara Manca, Umut Kucuk Coverage of DOAJ journals' citations through OpenCitations - Protocol https://dx.doi.org/10.17504/protocols.io.n92ldz598v5b/v5