Dec 13, 2024 Version 4

Protocol PROCI: A methodology for creating the Peer Review OpenCitations Index V.4

DOI
https://dx.doi.org/10.17504/protocols.io.261ge56qwg47/v4

Chiara Parravicini[1], Daniele Spedicati[1], Matteo Guenci[1], Nicole Liggeri[1]

[1]University of Bologna

Nicole Liggeri
University of Bologna

DOI: https://dx.doi.org/10.17504/protocols.io.261ge56qwg47/v4

Protocol Citation: Chiara Parravicini, Daniele Spedicati, Matteo Guenci, Nicole Liggeri 2024. Protocol PROCI: A methodology for creating the Peer Review OpenCitations Index. **protocols.io** https://dx.doi.org/10.17504/protocols.io.261ge56qwg47/v4Version created by **Nicole Liggeri**

Protocol status: Working
We use this protocol and it's working

Created: December 05, 2024

Last Modified: December 13, 2024

Protocol Integer ID: 114331

Keywords: peer review opencitations index, peer review data, analysis of peer review data, many peer reviews in crossref, peer review, number of peer review, new index of citation, many peer review, citation, opencitations index, opencitations meta, step methodology for the systematic extraction, top publication venue, systematic extraction, top publication venues in term, cited entity, specific citation function, compliant with the opencitations data model, following research question, opencitations data model, typed citation, research, research question, data analysis phase, publication, citing entity, data extraction, first two research question, new index, crossref

Abstract

We present a step-by-step methodology for the systematic extraction, alignment, and analysis of **peer review data** from **Crossref** to enhance the **OpenCitations Index**.

This process aims to address the following research questions:

1. Is it possible to create a **new index of citations** which contains typed citations where a peer review (citing entity) reviews (specific citation function) a publication (cited entity)?
2. What is the transformation of the Crossref dump necessary to create such an index to be **compliant with the OpenCitations Data Model** ?
3. What are the **top publication venues** in terms of the number of peer reviews received?
4. How many **peer reviews** in Crossref are included **in OpenCitations Meta**?
5. How many **articles** that have been reviewed by a peer review are included **in OpenCitations Meta**?

The protocol delineates four key phases: (1) **data extraction**, (2) **data processing**, (3) **post-processing**, and (4) **data analysis**. The first two research questions were addressed during the data processing and post-processing phases, while the answers to the third, fourth, and fifth questions were found during the data analysis phase.

Guidelines

The code developed for this research, along with detailed guidelines on how to run it, can be found in the **github repository**.

The required Python version for running the current software is Python 3.10.8.
Other libraries needed are:

- pytz version: 2022.7.1
- python-dateutil version: 2.8.2
- polars version: 0.20.27
- tqdm version: 4.64.1

Safety warnings

It is important to note that the capabilities of the computers at our disposal did not allow us to download the entire Crossref dump. Therefore, we leveraged more powerful servers provided by the University of Bologna to download the dataset, which subsequently has been divided into 18 chunks, each approximately 10 GB in size.

It should also be noted that expected results are just examples and depend on the specific dumps used.

Before start

To execute this protocol efficiently, ensure the system has at least **8 GB of RAM** (though 16GB would be recommended) and sufficient storage (at least **200 GB of free space**) to accommodate large datasets and processing. Additionally, the system should be equipped with Python and the required libraries (Polars, Matplotlib, and RDFLib) installed.
Details on how to install the dependencies and how to run each step of the process are included in the **readme file of the software repository**. We recommend creating a virtual environment to install libraries and dependencies, ensuring no conflicts arise with the existing setup on the computer.

Data extraction: acquiring data from Crossref and Meta

1 **Downloading data dumps**
In order to answer to our research questions, we gathered data from **Crossref** (dump released on April 2023, zip file of 185.88GB) and **OpenCitations Meta** (dump released on April 2024, zip file of 11GB).

Dataset	
META	NAME
https://opencitations.net/download#meta	LINK

Dataset	
Crossref	NA
https://academictorrents.com/details/d9e554f4f0c3047d9f49e448a7004f7aa1701b69	LI

The Crossref dump includes extensive metadata for scholarly publications. The dataset is typically organized with each record representing a single publication, containing various fields such as DOI, title, authors, publication date, journal name, and, notably, information about citations and peer reviews. Here is an example of the JSON data structure for peer reviews:

```
{
        "URL": "http://dx.doi.org/10.7554/elife.69960.sa2",
        "resource": {
            "primary": {
                "URL":
"https://elifesciences.org/articles/69960#sa2"
            }
        },
        "member": "4374",
        "score": 0.0,
        "created": {
            "date-parts": [
                [
                    2022,
                    9,
                    6
                ]
            ],
            "date-time": "2022-09-06T13:30:35Z",
            "timestamp": 1662471035000
        },
        "license": [
            {
                "start": {
                    "date-parts": [
                        [
                            2021,
                            10,
                            13
                        ]
                    ],
                    "date-time": "2021-10-13T00:00:00Z",
                    "timestamp": 1634083200000
                },
                "content-version": "unspecified",
                "delay-in-days": 0,
                "URL":
"http://creativecommons.org/licenses/by/4.0/"
            }
        ],
        "issued": {
            "date-parts": [
                [
                    2021,
                    10,
                    13
                ]
            ]
        },
        "review": {
            "type": "author-comment",
            "stage": "pre-publication"
        },
        "prefix": "10.7554",
        "reference-count": 0,
        "indexed": {
            "date-parts": [
                [
                    2022,
                    9,
                    6
                ]
            ],
            "date-time": "2022-09-06T14:14:29Z",
            "timestamp": 1662473669310
```

```
        },
        "author":
            {
                "ORCID": "http://orcid.org/0000-0002-6672-5202",
                "authenticated-orcid": true,
                "given": "Allison",
                "family": "Schad",
                "sequence": "first",
                "affiliation": [
                    {
                        "id": [
                            {
                                "id": "https://ror.org/0130frc33",
                                "id-type": "ROR",
                                "asserted-by": "publisher"
                            }
                        ],
                        "name": "Office of Medical Education,
University of North Carolina at Chapel Hill School of Medicine",
                        "place": [
                            "Chapel Hill, United States"
                        ]
                    }
                ]
            },
       "DOI": "10.7554/elife.69960.sa2",
        "is-referenced-by-count": 0,
        "published": {
            "date-parts": [
                [
                    2021,
                    10,
                    13
                ]
            ]
        },
        "published-print": {
            "date-parts": [
                [
                    2021,
                    10,
                    13
                ]
            ]
        },
        "content-domain": {
            "domain": []
        },
        "title": [
            "Author response: Mental health in medical and
biomedical doctoral students during the 2020 COVID-19 pandemic and
racial protests"
        ],
        "source": "Crossref",
        "type": "peer-review",
        "publisher": "eLife Sciences Publications, Ltd",
        "references-count": 0,
        "deposited": {
            "date-parts": [
                [
                    2022,
                    9,
                    6
                ]
            ],
            "date-time": "2022-09-06T13:30:36Z",
```

```
            "timestamp": 1662471036000
        },
        "relation": {
            "is-review-of": [
                {
                    "id-type": "doi",
                    "id": "10.7554/eLife.69960",
                    "asserted-by": "subject"
                }
            ]
        }
    }
```


The OpenCitations Meta dump, similar to Crossref, includes detailed metadata on scholarly works. The dataset is structured to facilitate the exploration of citation networks, including peer reviews. The data typically includes fields such as the citing DOI, cited DOI, publication dates, and reviewer information. Here is an example of how data is structured in OpenCitations Meta:

	id	title	author	issue	volume	venue	page	pub_date	type	publisher	editor
	omid/0690163767doi:10.3390/inorganics7030032openalex	Computational Characterization Of Single-Electron Transfer Steps In Water Oxidation	De Aguirre, Adiran [omid/0624025336 orcid:0000-0001-7991-6406]; Funes-Ardoiz, Ignacio [omid/0670126897171 orcid:0000-0002-5843-9660]; Maseras, Feliu [omid/0690494082 8]	3	7	Inorganics [omid/0690166425 5 issn:2304-6740 openalex]	32-32	2019-03-01	journal article	Mdpi Ag [omid/0610116165 crossref:1968]	

This phase takes the Crossref data as input and selects the values from the JSON files that are necessary for alignment with the OpenCitations Index data model. The software extracts and categorizes the data from the input dataset, distinguishing between peer review and non-peer review items. Only the information useful according to the OC data model is extracted.

2 **Peer-review extraction**

To extract peer review items, we employ a targeted approach using compressed JSON files stored in a ZIP archive. The extraction process operates in parallel to handle large datasets efficiently. Files are processed in manageable batches, where each file is read, decompressed, and parsed to identify items of type "peer-review". These entries are saved with relevant metadata: citing DOI, cited DOI, creation date, and author information. Authors' details, including names and ORCID identifiers (if available), are consolidated for additional context.

2.1 **OCI Generation**

Open Citation Identifiers (OCIs) are generated for peer review relationships to uniquely link citing and cited entities. Characters in each DOI are mapped to unique codes using a pre-defined lookup table. OCIs are formed as

```
oci:<citing_entity_local_id>-<cited_entity_local_id>
```

Expected result

Peer review CSV containing fields for OCIs, citing and cited DOIs, citation dates, URLs, and author details.

	oci	citing_doi	cited_doi	citing_date	citing_url	author_info
	0200100000236312228033709020136310236271428252423281401-020010000023699999991699999991799999991803370902021	10.1002/vms3.921/v2/response1	10.1002/VMS3.921	2022-09-15	http://dx.doi.org/10.1002/vms3.921/v2/response1	Ye Sun; Xufeng Hou; Lingjie Li; Yanqing Tang; Mingyue Zheng; Weisen Zeng; XiaoLong Lei

3 **Non-peer review data extraction**

Similarly, the process for identifying non-peer review items focuses on filtering out any records tagged as "peer-review". The remaining entries, which include items like journal articles, conference papers, or other publication types, are analyzed to capture details such as the cited DOI, URL, ISSN, venue title, and publication date.

Expected result

Non-peer review CSV containing cited DOIs, citation dates, URLs, issn and venue details.

	cited_doi	cited_date	cited_url	cited_issn	cited_venue
	10.3133/ofr20051113a	2018-08-15	http://dx.doi.org/10.3133/ofr20051113a	2331-1258	Open-File Report

Data processing: combining and integrating peer review and non-peer review data

4 This step consolidates the data extracted in the previous step, joining the peer review and non-peer review datasets into a unified structure and calculating the temporal difference between the citing and cited publications. The goal is to obtain a combined

CSV with all the information needed to create an index of citations compliant with the OpenCitations Data Model and to enable subsequent analyses that will be illustrated later.

4.1 **Data integration**

The integration process begins by loading and combining CSV files from the directories containing peer review and non-peer review datasets. To ensure consistency, the following actions are needed:

1. **Normalizing** DOI fields, which are cleaned to remove stray characters and standardized to lowercase.
2. **Validating** the required column for joining (e.g., cited_doi), making sure it exists in both datasets.
3. **Joining** the datasets, using an inner join based on the shared cited_doi column to create a consolidated dataset containing both citing and cited metadata.

4.2 **Provenance addition**

To align data with OpenCitations Data Model and to ensure reproducibility and traceability, provenance metadata is appended to the joined dataset:

- **Agent**: The source of the data (e.g., https://academictorrents.com).
- **Source**: DOI of the dataset.
- **Timestamp**: A UTC timestamp marking the integration's execution.

4.3 **Temporal delta calculation**

The temporal difference between the publication dates of citing and cited items is calculated in compliance with ISO 8601 duration format. This involves:

1. **Parsing** the citing_date and cited_date to datetime objects.
2. **Calculating the difference** in years, months, and days using *relativedelta* from the *dateutil* library.
3. **Formatting** the result as an ISO 8601 duration string, prefixed with a minus symbol ("-") if the citing date is earlier than the cited date.

Expected result

Combined CSV file with all the information regarding both peer reviews (citing entities) and non-peer reviews (cited entities).

	oci	citing_doi	cited_doi	citing_date	citing_url	author_info	cited_url	cited_issn	cited_venue	cited_date	prov_agent	source	prov_date	time_span
	oci:020070505043614211815143700000003013700000808-020070505043614999999905181514370000000301	10.7554/elife.00031.008	10.7554/elife.00031	2012-10-26	http://dx.doi.org/10.7554/elife.00031.008	Pretto, Paolo; Bresciani, Jean-Pierre; Rainer, Gregor; Bülthoff, Heinrich H	http://dx.doi.org/10.7554/elife.00031	2050-084X	eLife	2012-10-26	https://academictorrents.com/details/d9e554f4f0c3047d9f49e448a7004f7aa1701b69	https://doi.org/10.13003/8wx5k	2024-11-22T15:44:10.680404Z	P0D

5 **Dataset compartmentalization**

In this step, the unified dataset obtained after temporal delta calculation is further divided into specialized compartments for targeted analysis. This separation ensures that each subset focuses on specific data attributes while retaining structural consistency.

This part of the process involves:

1. **Cloning** the full dataset twice to maintain the original data and allow independent column manipulation for each subset.
2. **Dropping** columns irrelevant to each compartment's purpose as specified during initialization.

Each compartment is saved to a separate CSV file.

Expected result

Three CSV files each containing information regarding:

- **Citations**, with fields for OCI, citing DOI, cited DOI, citing date, citing url, author information , cited url, cited date, and time span.
- **Provenance**, with fields for OCI, author information, provenance agent, source, and provenance date.
- **Venues**, with fields for cited DOI, author information, cited ISSN, and cited venue.

Post-processing: RDF Graph construction

6 This phase involves generating RDF graphs from the processed data to represent citations in a structured format. RDF allows the creation of machine-readable datasets with explicit semantics, which are essential for interoperability and querying using SPARQL.

Note

Input:
- The combined CSV file obtained as output of step 4.3.
- A base url for creating each element of the triples.

6.1 **RDF representation of peer review data**

This phase proceeds following these steps:

1. **Mapping** each row from the processed CSV dataset RDF triples.
2. **Emplyoing** RDF predicates from the CiTO (Citation Typing Ontology) to describe citation relationships. Key relationships include linking citing entities (*hasCitingEntity*) to cited entities (*hasCitedEntity*) and recording temporal spans (*hasCitationTimeSpan*) or creation dates (*hasCitationCreationDate*).

6.2 **Provenance integration**

Provenance metadata, based on the PROV ontology, is included for documenting the source, agent responsible for data generation, and the timestamp of generation.

6.3 **Data processing**

1. **Parsing** the input CSV file line by line.
2. **Creating the RDF graph** taking from each row an instance of the PeerReview class.

The attributes (e.g., citing url, cited url, time span, and citing date) are included based on the mode of operation:

- Data-Only Mode: Generates RDF triples representing the citation relationships and attributes.
- Provenance-Only Mode: Focuses on documenting provenance information.

6.4 **Temporal precision**

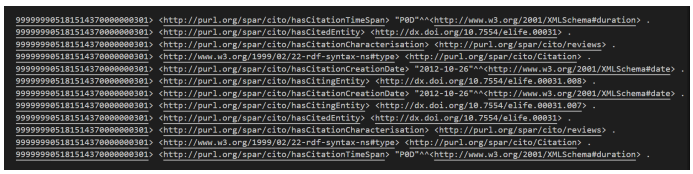
Dates from the input data are evaluated for precision using ISO 8601 standards (gYear, gYearMonth, or date). This ensures compatibility with RDF literals and accurate semantic representation.

6.5 **Batch RDF Writing**

1. **Serializing** the RDF graph in Turtle format (.ttl) and appending it to the specified output file.
2. **Minimizing** the writing process memory usage by streaming data directly to disk, facilitating scalability with large datasets.

Expected result

RDF graph containing triples encoding citation relationships and temporal attributes, and triples documenting the origin and generation context of the data.



A sample of the Turtle file generated as output of this phase of the workflow; information about peer reviews is formalised in semantic triples, which comprise a subject, predicate, and object.

Data analysis: checking coverage with OpenCitations Meta and calculating top publication venues

7 **Comparison with OpenCitations Meta**

This part of the process focuses on extracting, processing, and analyzing Digital Object Identifiers (DOIs) and their metadata from OpenCitations Meta. As mentioned before, OpenCitations Meta is a database that provides detailed bibliographic metadata for publications indexed in the OpenCitations Index, which focuses on openly sharing citation relationships between publications.

Note

Input:
- The Citation CSV file containing citation relationships (citing DOI and cited DOI).
- The OpenCitations Meta dump (a compressed ZIP archive of metadata files).

7.1 **DOI extraction**

1. **Reading** each CSV file within the ZIP archive is read and **extracting** the first column to obtain DOI-related information.
2. **Identifying** DOIs using a specific pattern (doi:) within the text.
3. **Saving** cleaned DOIs in a separate CSV file for subsequent processing.

7.2 **Classification into peer reviews and articles**

DOIs from the metadata are cross-referenced with the Citation dataset to classify them into two categories:
- **Peer Reviews**: DOIs that appear as citing entities in the Citation dataset.
- **Articles**: DOIs that appear as cited entities.

Two separate CSV files are generated, one for each category.

7.3 **Data deduplication and count calculation**

The final steps of this phase consist in:
1. **Cleaning** the DOI lists to remove duplicates using Pandas, ensuring the uniqueness of entries in both CSV files for peer reviews and articles.
2. **Calculating** the number of unique entries in the cleaned files to determine the total counts of peer reviews and articles.
3. **Saving** the counts to a summary CSV file, providing an overview of the dataset's composition.

Expected result

Four CSV files containing information regarding:
- **Meta DOIs**: A consolidated list of extracted DOIs from the ZIP file.
- **Meta peer reviews**: Unique DOIs categorized as peer reviews.
- **Meta articles**: Unique DOIs categorized as articles.
- **Summary CSV**: Contains the counts of peer reviews and articles for statistical reporting.

8 **Top publication venue analysis**

This part of the workflow focuses on analyzing the venues (e.g., journals or conferences) associated with citations.

8.1 **Data normalization**

This step involves:
1. **Splitting** the cited_issn column into separate entries for cases where multiple ISSNs are recorded (e.g., 1234-5678, 8765-4321).
2. **Separating** rows with empty ISSN fields and grouping them by venue name only.

8.2 **Top venues count**

The final count is achieved by:
1. **Grouping** venues by ISSN pairs (issn1, issn2) or by venue name alone when ISSNs are absent.
2. **Counting** of occurrences for each venue.
3. **Sorting** the dataset by the count column to identify the most frequently cited venues.

Expected result

Two CSV files containing the following information:
- **Venue Counts**: ISSNs, venue names, and their respective counts.
- **Top Venues Summary**: A filtered list of the most cited venues, useful for targeted analysis.

Protocol references

Heibi, I., Peroni, S. & Shotton, D. Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations.Scientometrics 121, 1213–1228 (2019). https://doi.org/10.1007/s11192-019-03217-6.