# How to structure citations data and bibliographic metadata in the OpenCitations accepted format

Arcangelo Massari<sup>1</sup>, Ivan Heibi<sup>1</sup>

<sup>1</sup>Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

#### Abstract

The OpenCitations organization is working on ingesting citation data and bibliographic metadata directly provided by the community (e.g., scholars and publishers). The aim is to improve the general coverage of open citations, which is still far from being complete, and use the provided metadata to enrich the characterization of the citing and cited entities. This paper illustrates how the citation data and bibliographic metadata should be structured to comply with the OpenCitations accepted format.

OpenCitations, Bibliographic metadata, Open citations, CSV

### 1. Introduction

The Declaration on Research Assessment [1], the Leiden Manifesto for Research Metrics [2], and the Initiative for Open Citations (I4OC, https://i4oc.org/) have successfully convinced almost all major academic publishers to release their publication reference lists. To date, more than 1.2 billion citations are available through the Crossref REST API [3] and distributed by OpenCitations [4] as structured, separated from the original bibliographic source and under the CC0 license [5].

Nevertheless, the coverage of open citations is still far from complete [6]. On the one hand, some publishers have not yet made their citations public. On the other hand, many citations are lost because they are only present in unstructured format within PDF files, especially in social sciences.

OpenCitations is working on ingesting citations and bibliographic metadata directly coming from the community (e.g., scholars and publishers). In this way, projects like EXCITE [7] - aimed at extracting citations from PDFs - could significantly contribute to increasing the data cov-

The following section illustrates how to structure the citation data and bibliographic metadata in the accepted format of OpenCitations. We conclude this paper with a description of the upcoming future related works.

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License CEUR Workshop Proceedings (CEUR-WS.org)

# 2. Metadata and citations

OpenCitations manages and processes two different CSV files to separately characterize the ingested documents, one containing their metadata (META-CSV), and a second one holding their citations (CITS-CSV). On this section we discuss how these files should be structured and defined before providing them to OpenCitations. The discussion presented in this section is based on a more exhaustive documentation [8].

CSV files are logically structured as tables. In META-CSV each document (row), is characterised by 11 attributes (columns):

• id. the ID(s) of the corresponding document. A document can have more than one ID, each ID is defined by its type (using an acronym) and value. Multiple IDs must be separated using single white space, as follow:

For example "doi:10.3233/ds-170012" indicates a DOI identifier having the value "10.3233/ds-170012".

- title. a textual value to express the title of the document.
- author and editor. Data regarding the authors and the editors of the document. Each character (author/editor) is defined by several attributes, e.g., his family name or ID. Multiple characters are separated by a semicolon followed by a white space character. Generally, the definition of an actor follows this structure:

The IDs of the authors/editors are specified in square brackets and follow the format used for the "id" attribute.

JCDL'22: ULITE-ws, Understanding LIterature references in academic full TExt. June 24-06, 2022, Cologne, Germany

arcangelo.massari@unibo.it (A. Massari); ivan.heibi2@unibo.it

D 0000-0002-8420-0696 (A. Massari); 0000-0001-5366-5194 (I. Heibi)

e.g. "Peroni, Silvio [orcid:0000-0003-0530-4305]"

In case of no IDs, the square brackets are omitted from the character description either. The given name is not mandatory, however, the description of the character should still contain a comma to indicate such absence (e.g. "Peroni, [orcid:0000-0003-0530- 4305]")

pub\_date. the date of publication of the document. The date is defined according to ISO 86014[9], the ISO standard for "Representation of dates and times":

### YYYY-MM-DD

It is mandatory to specify at least the publication year. The values of the month and day are not required. However, if the day is specified, the month must be specified as well.

venue. data regarding the venue of the document.
For example, if the document is a journal article, the venue defines the journal where the document has been published. Each venue is described as follows:

The IDs of a venue are described using the same format used previously. In case of no identifiers, the square brackets are omitted.

- volume and issue. these values are required only if the document is contained in a journal volume or a journal issue.
- page. the page range of the corresponding document, defined through the specification of the first and the last page, divided by a hyphen "-".
- type. a textual value to identify the document. This value is taken from the list of the currently supported bibliographic resource types: book, book chapter, book part, book section, book series, book set, book track, component, dataset (or data file), dissertation, edited book, journal, journal article, journal issue, journal volume, monograph, other, peer review, posted content (or web content), proceedings, proceedings article, proceedings series, reference book, reference entry, report, report series, standard, and standard series.
- publisher. the publisher name of the corresponding document. To define a publisher we apply the same format used in the definition of the venue.

If the resource identifier is specified in the "id" field, all the other fields are optional. Conversely, if the "id" field is empty, there are mandatory fields that vary depending on the resource type:

- The fields "title", "pub\_date", and "author" (or "editor") are mandatory for the resources of type book, dataset (or data file), dissertation, edited book, journal article, monograph, other, peer review, posted content (or web content), proceedings article, report, and reference book. Moreover, this information is compulsory if the "type" field is empty.
- The "title" and "venue" fields are required for the resources of type book chapter, book part, book section, book track, component, and reference entry.
- Only the "title" field is required for the resources of type book series, book set, journal, proceedings, proceedings series, report series, standard, and standard series.
- Regarding the resources of journal volume type, the fields "venue" and "volume", or "venue" and "title", are mandatory. Conversely, as for resources of journal issue type, the fields "venue" and "issue", or "venue" and "title", are mandatory.

Table 1 shows an example of a well-formed META-CSV representation. The table contains a small sample of ten documents (rows) and their corresponding attributes (columns).

On the other hand, in the CITS-CSV each entity (row) represents a citation. A citation is characterised by 4 attributes (columns): ing\_id, citing\_publication\_date, cited\_id, The citing\_id and cited\_publication\_date.  $\verb|cited_id| values represent the identifiers of the citing|$ and cited document, respectively. These values are both mandatory, and they are structured following the same scheme used for id definition in META-CSV. The citing\_publication\_date and cited\_publication\_date represent the date of publication of the citing and cited document, respectively. Both these values are optional, and follow the same structural scheme used for pub\_date definition in META-CSV.

Table 2 shows an example of a well-formed CITS-CSV representation. The table contains a small sample of ten different citations (rows) and their corresponding attributes (columns).

# 3. Discussion and conclusion

This paper described how to define well-formed CSV files storing citations and metadata of bibliographic resources, ready to be provided and later processed by OpenCitations.

The ingestion of bibliographic metadata will be possible starting from the release of OpenCitations Meta (OC-Meta), expected by the end of 2022. OC-Meta will

store bibliographic metadata for the documents involved (as citing or cited entities) in OpenCitations citation indexes.

The ingestion of the citations is possible thanks to CROCI, the Crowdsourced Open Citations Index, which allows individuals identified by ORCIDs to deposit the citation data that they have legal right to submit [10]. Citation data are submitted to either Figshare (https://figshare.com) or Zenodo (https://zenodo.org), accompanied by the ORCID of the contributor. Aftwerwards, the submitter can inform OpenCitations using the GitHub issue tracker on the CROCI repository (https://github.com/opencitations/croci/issues).

Future works include implementing an interface that simplifies and automates the entire publication process via CROCI, also providing input data validation and modification suggestions.

Moreover, CROCI currently handles only DOI-to-DOI citations. The upcoming plan is to let CROCI manage also any-to-any citations.

# Acknowledgments

This work was funded from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017452 (OpenAIRE-Nexus Project). We want to thank Silvio Peroni for supervising the entire work on OpenCitations, Philipp Mayr-Schlegel and Ahsan Shahid for the feedback on the documentation from which this demo paper is drawn, and Davide Brambilla for the valuable insights about CROCI and its future developments.

## References

- [1] R. Cagan, San francisco declaration on research assessment, Disease Models & Mechanisms (2013) dmm.012955. URL: https://journals.biologists.com/ dmm/article/doi/10.1242/dmm.012955/261854/ San-Francisco-Declaration-on-Research-Assessment. doi:10.1242/dmm.012955.
- [2] D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, I. Rafols, Bibliometrics: The leiden manifesto for research metrics, Nature 520 (2015) 429–431. URL: https://www.nature.com/articles/520429a. doi:10. 1038/520429a.
- [3] G. Hendricks, D. Tkaczyk, J. Lin, P. Feeney, Crossref: The sustainable source of community-owned scholarly metadata, Quantitative Science Studies 1 (2020) 414–427. doi:10.1162/qss\_a\_00022.
- [4] S. Peroni, D. Shotton, OpenCitations, an infrastructure organization for open scholarship, Quantitative Science Studies 1 (2020) 428–444. doi:10.1162/qss\_a\_00023.

- [5] S. Peroni, D. Shotton, Open citation: Definition, 2018. doi:10.6084/M9.FIGSHARE.6683855.V1, artwork Size: 95436 Bytes Publisher: figshare.
- [6] A. Martín-Martín, Coverage of open citation data approaches parity with web of science and scopus, OpenCitations blog (2021).
- [7] A. Hosseini, B. Ghavimi, Z. Boukhers, P. Mayr, Excite–a toolchain to extract, match and publish open literature references, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2019, pp. 432–433.
- [8] A. Massari, How to produce well-formed CSV files for OpenCitations, 2022. URL: https://doi.org/10.5281/zenodo.6597141. doi:10.5281/zenodo.6597141.
- [9] M. Wolf, C. Wicksteed, Date and time formats, https://www.w3.org/TR/NOTE-datetime, 1997.
- [10] I. Heibi, S. Peroni, D. M. Shotton, Crowdsourcing open citations with CROCI an analysis of the current status of open citations, and a proposal, CoRR abs/1902.02534 (2019). URL: http://arxiv.org/abs/1902.02534. arxiv:1902.02534.

# A. Appendix

Table 1.: A sample of ten documents characterized by their corresponding metadata attributes

| pi  | title  | author  | pub_date venue | venue   | volume | volume issue page |         | type                    | publisher  | editor               |
|---|--|---|----------------|---|--------|-------------------|---------|-------------------------|--|----------------------|
| doi:10.1007/978-3-030-00668-6_8 The SPAR Ontologies                       | The SPAR Ontologies                                      | Peroni, Silvio [orcid:0000-0003-0530-4305];<br>Shotton, David [orcid:0000-0001-5506-523X] | 2018           | 17th ISWC<br>[doi:10.1007/978-3-030-00668-6]        |        |                   | 119-136 | 119-136 book chapter    | Springer International<br>Publishing<br>[crossref:297]       |                      |
| doi:10.3233/DS-170012   | Automating semantic publishing                           | Peroni, Silvio [orcid:0000-0003-0530-4305]  | 2017           | Data Science<br>[issn:2451-8484 issn:2451-8492]     | 1      | 1-2               | 155-173 | 155-173 journal article | IOS Press<br>[crossref:7437]                                 |                      |
| doi:10.1007/978-3-476-00160-3<br>isbn:9783476021144<br>isbn:9783476001603 | Literatur  |   | 2005           |   |        |                   |         | book                    | Springer Science and<br>Business Media LLC<br>[crossref:297] | Gfrereis, Heike      |
| doi:10.1057/9780230316645<br>isbn:9780230276604<br>isbn:9780230316645     | New Waves in Philosophy of Law                           |   | 2011           |   |        |                   |         | book                    | Springer Science and<br>Business Media LLC<br>[crossref:297] | Mar, Maksymilian Del |
| doi:10.4324/9781003115830<br>isbn:9781003115830                           | Governing Savages  | Markus, Andrew  | 2020-7-31      |   |        |                   |         | book                    | Informa UK Limited<br>[crossref:301]                         |                      |
| doi:10.1515/9781503600836<br>isbn:9781503600836                           | Newsworthy   | Barbas, Samantha  | 2020-6-24      |   |        |                   |         | book                    | Walter de Gruyter GmbH<br>[crossref:374]                     |                      |
| doi:10.1134/s0018151x17020055   | On the theory of convection of electrons in metals       | Gladkov, S. O.  | 2017-5         | High Temperature<br>[issn:0018-151X issn:1608-3156] | 55     | 3                 | 321-325 | 321-325 journal article | Pleiades Publishing Ltd<br>[crossref:137]                    |                      |
| doi:10.1134/s0018151x17050029 Stability of boiling shock                  | Stability of boiling shock                               | Avdeev, A. A.   | 2017-9         | High Temperature<br>[issn:0018-151X issn:1608-3156] | 55     | 5                 | 753-760 | 753-760 journal article | Pleiades Publishing Ltd<br>[crossref:137]                    |                      |
| doi:10.1134/s0018151x17050224   | The high-temperature<br>and radiative effect on concrete | Zhakin, A. I.   | 2017-9         | High Temperature<br>[issn:0018-151X issn:1608-3156] | 55     | 5                 | 767-776 | 767-776 journal article | Pleiades Publishing Ltd<br>[crossref:137]                    |                      |
| doi:10.1134/s0018151v18010169   | Relaxation of Rayleigh                                   | Skrabkov O V  | 2018-1         | High Temperature                                    | 26     | -                 | 77-83   | 77-83 ionanoi article   | Pleiades Publishing Ltd                                      |                      |

Table 2.: A sample of ten citations characterized by their related attributes

| citing_id                                | citing_publication_date cited_id | cited_id                              | cited_publication_date |
|--|----------------------------------|---------------------------------------|------------------------|
| doi:10.1016/j.websem.2012.08.001 2012-12 | 2012-12                          | doi:10.1087/2009202                   | 2009-04-01             |
| doi:10.1016/j.websem.2012.08.001 2012-12 | 2012-12                          | doi:10.1371/journal.pcbi.1000361      |                        |
| doi:10.1016/j.websem.2012.08.001 2012-12 | 2012-12                          | doi:10.1007/978-3-642-33876-2_35 2012 | 2012                   |
| doi:10.1016/j.websem.2012.08.001 2012-12 | 2012-12                          | doi:10.1186/2041-1480-1-S1-S6         | 2010-06-22             |
| doi:10.1016/j.websem.2012.08.001 2012-12 | 2012-12                          | doi:10.1145/945645.945664             | 2003-10-23             |
| pmid:23636598                            | 2013                             | pmid:19151427                         | 2005                   |
| pmid:23636598                            | 2013                             | pmid:19782561                         | 2008-10                |
| pmid:23636598                            |                                  | pmid:18686754                         | 2012-09-05             |
| pmid:23636598                            | 2013                             | pmid:15890079                         | 2009-07-15             |
| pmid:23636598                            | 2013                             | pmid:18191757                         |                        |