Journal of Open
Humanities Data

The Integration of the Japan Link Center's Bibliographic Data into OpenCitations

RESEARCH PAPER

]u[ubiquity press

The production of bibliographic and citation data structured according to the OpenCitations Data Model, originating from an Anglo-Japanese dataset

ARIANNA MORETTI
MARTA SORICETTI
IVAN HEIBI
ARCANGELO MASSARI
SILVIO PERONI
ELIA RIZZETTO

*Author affiliations can be found in the back matter of this article

ABSTRACT

In this article, we present OpenCitations' main data collections: the unified index of citation data (OpenCitations Index), and the bibliographic data corpus (OpenCitations Meta) in view of the integration of a new dataset provided by the Japan Link Center (JaLC). Based on a computational analysis of the titles of the publications performed in October 2023, 8.6% of the bibliographic metadata stored in OpenCitations Meta are not in English. Nevertheless, the ingestion of an Anglo-Japanese dataset represents the first opportunity to test the soundness of a language-agnostic metadata crosswalk process for collecting data from multilingual sources, aiming to preserve bibliodiversity and to minimize information loss considering the constraints imposed by the OpenCitations data model, which does not allow the acceptance of multiple values in different translations for the same metadata field. The JaLC dataset is set to join OpenCitations' collections in November 2023, and it will be made available in RDF, CSV, and SCHOLIX formats. Data will be produced using open-source software and provided under a CC0 license via API services, web browsing interfaces, Figshare data dumps, and SPARQL endpoints, ensuring high interoperability, reuse, and semantic exploitation.

CORRESPONDING AUTHOR:
Arianna Moretti
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
arianna.moretti4@unibo.it

KEYWORDS:
Japan Link Center; OpenCitations; bibliographic data; citations; workflow; multilingualism

TO CITE THIS ARTICLE:
Moretti, A., Soricetti, M., Heibi, I., Massari, A., Peroni, S., & Rizzetto, E. (2024). The Integration of the Japan Link Center's Bibliographic Data into OpenCitations. *Journal of Open Humanities Data*, 10: 21, pp. 1–16. DOI: https://doi.org/10.5334/johd.178

1. CONTEXT AND MOTIVATION

In the Social Sciences and Humanities (SSH) disciplines, interdisciplinary teams often need to integrate data from various formats and models, which can be complex, particularly in terms of data modeling. This intricacy is evident in the integration of differently structured bibliographic and citation data.

OpenCitations[1] (Peroni & Shotton 2020) is a non-profit independent infrastructure organization dedicated to open scholarship for the collection, curation, management, and publication of citation and bibliographic data. Some of the infrastructure's distinctive traits include treating scholarly citations as first-class data entities (each with its persistent identifier) and offering comprehensive, freely accessible global citation data with a focus on semantic interoperability implemented via the adoption of Semantic Web technologies (Shotton 2013). Furthermore, OpenCitations provides inclusive, transparent, and interoperable scholarly citation services, driven by open principles under academia-based governance (Peroni & Shotton 2022b). The data it provides are collected from different sources, reshaped according to the OpenCitations Data Model (OCDM) (Daquino et al. 2020a; Daquino et al. 2020b), and exposed using Linked Data technologies (Berners-Lee & Kagal 2008), with the ultimate goal of providing free access to data in highly interoperable formats, such as RDF, SCHOLIX (Burton et al. 2017), and CSV.

In this paper, we illustrate the process of producing and updating OpenCitations data collections following the ingestion of an Anglo-Japanese dataset provided by the Japan Link Center (JaLC). The aim is to test the flexibility of a language-agnostic methodology for the management of multilingual or non-English datasets, arguing that this approach is functional for the safeguarding of the bibliodiversity of the acquired data, which represents a crucial aspect in cultural heritage preservation and humanities-oriented studies (Horvath 2021), as also stressed in the Recommendation on Open Science by UNESCO (2021).

1.1 OPENCITATIONS DATASETS

OpenCitations currently maintains two primary datasets. The first dataset is OpenCitations Index, the unified collection of open citations ingested from different sources created starting from the experience done with its first index, i.e. COCI (Heibi et al. 2019), that gathered citations from Crossref. The other dataset is OpenCitations Meta (Massari et al. 2024a), a collection of bibliographic metadata of all the resources included in the OpenCitations Index as citing or cited entities.

To date, the information integrated into the OpenCitations infrastructure comes from Crossref[2] (Hendricks et al. 2020), DataCite[3] (DataCite Metadata Working Group, 2021), PubMed (Hutchins et al. 2019; Canese & Weis, 2013), OpenAIRE (La Bruzzo et al. 2023; Alexiou et al. 2016), and – since November 2023 – the Japan Link Center (JaLC) (斉史 et al. 2012).

OpenCitations Meta data are available as CSV and RDF dumps. The stored metadata for the bibliographic resources includes document IDs, titles, authors, publication dates, venue information, volume and issue identifiers, page ranges, resource types, publishers, and editors. Instead, OpenCitations Index data is made available in CSV, N-Triples, and Scholix (JSON) formats, and it exposes the Open Citation Identifier (OCI) (Peroni & Shotton 2019) for the citation, the OpenCitations Meta Identifier (OMID, a persistent identifier for the entities included in OpenCitations Meta) of both entities involved in the citation, the citation creation date, the timespan between the publication dates of the cited and citing entities, and fields indicating whether both entities are published in the same journal and if they share at least one common author.

Both OpenCitations Index and Meta data are accompanied by provenance information (Massari et al. 2023), which includes the responsible agent, the source URL, and the creation and modification date of the record, as well as tracking the changes of the data associated with an entity.

[1] https://opencitations.net.

[2] https://api.crossref.org/snapshots.

[3] https://archive.org/details/datacite_dump_20230713.

1.2 JaLC DATA AS A PROTOTYPE TESTING FACILITY FOR A LANGUAGE-AGNOSTIC APPROACH

This paper focuses on the process of metadata crosswalk (Chen 2015) of multilingual data provided by JaLC to the OCDM. To this end, we implemented a particular workflow (Moretti & Heibi 2023) to produce citation and bibliographic data.

The introduction of an Anglo-Japanese dataset marks OpenCitations' first formal effort to ingest extensively non-English sources, emphasizing the importance of handling multilingual data and promoting inclusivity and global knowledge dissemination.

We avail ourselves of this opportunity to expose a methodology that covers data acquisition, curation, and the production of citation and bibliographic data, highlighting the benefits of a language-agnostic approach to data integration. Indeed, we claim that this strategy fosters the preservation of bibliodiversity, enhances scholarly research, and facilitates knowledge access to data provided by any data source, including multilingual ones.

Coherently, the integration of the JaLC collection tests the management of multilingual datasets following a language-agnostic approach that prioritizes displaying data in the original language when available, to facilitate access and reuse for an international academic and research community.

2. DATASET DESCRIPTION

OBJECTS NAMES, FORMAT NAMES, VERSIONS, AND REPOSITORY LOCATION

JaLC citation data are integrated into OpenCitations Index, and the bibliographic metadata provided by JaLC is published in OpenCitations Meta. The JaLC data are included in the last dumps of OpenCitations Index – in CSV (OpenCitations 2023a), RDF (OpenCitations 2023b), and Scholix (OpenCitations 2023c) formats – and OpenCitations Meta, in CSV format (OpenCitations 2023d).

CREATION DATES

JaLC's ingestion process took six months, from June 1, 2023, to the dataset production on November 29, 2023.

DATASET CREATORS

The input data was supplied by the Japan Science and Technology Agency[4] and then analyzed and processed using custom software components. Below is the list of students and researchers affiliated with the Research Centre for Open Scholarly Metadata[5] and the Digital Humanities Advanced Research Center of the University of Bologna[6] who took part in the OpenCitations' datasets creation, together with their roles:

- Marta Soricetti: Software;
- Arianna Moretti: Data curation, Software;
- Ivan Heibi: Data curation, Software;
- Arcangelo Massari: Data curation, Software;
- Silvio Peroni: Supervision, Project administration, Conceptualization;
- Elia Rizzetto: Software.

LANGUAGE

For communication and dissemination purposes, the datasets' structure follows the OpenCitations Data Model, an English-based data model. However, since the OCDM has no restrictions concerning the language choice, metadata values are exposed in the original

[4] https://www.jst.go.jp/EN/.

[5] https://openscholarlymetadata.org/.

[6] https://centri.unibo.it/dharc/en.

formulation when possible. For JaLC, the data include a combination of information in English, Japanese, and, occasionally, other languages.

LICENSE

In full compliance with FAIR principles (Wilkinson et al. 2016) and the spirit of Open Science, the citation and bibliographic data produced from JaLC are released under a CC0 waiver.[7]

REPOSITORY NAME

Dumps are released on OpenCitations' Figshare page,[8] and the related links are published at https://opencitations.net/download. In addition, the data are available through other services for programmatic access, listed on the OpenCitations' website.[9]

PUBLICATION DATE

The latest versions of the dumps of OpenCitations Index and OpenCitations Meta, which include JaLC data, were published on Figshare on 11 December 2023 and 30 November 2023 respectively.

3. METHOD

In this section, we introduce the methodology for generating citation and bibliographic data compliant with OCDM given an external source's dataset, focusing on the extension of OpenCitations software infrastructure for integrating data from JaLC. This case study offers the opportunity to introduce an updated version of the workflow – adopted for the first time for JaLC data ingestion – and elaborate on the measures taken for handling multilingual aspects.

3.1 DATA INGESTION IN OPENCITATIONS

Originally, the data in Meta was derived exclusively from Crossref. However, between December 2022 and July 2023, data from PubMed, DataCite, and OpenAIRE were introduced, and simultaneously the software extension activities for JaLC dataset integration commenced. To facilitate software maintenance and define a systematic approach at the time of such expansions, a structural re-engineering of the ingestion and production workflow became necessary (Figure 1).

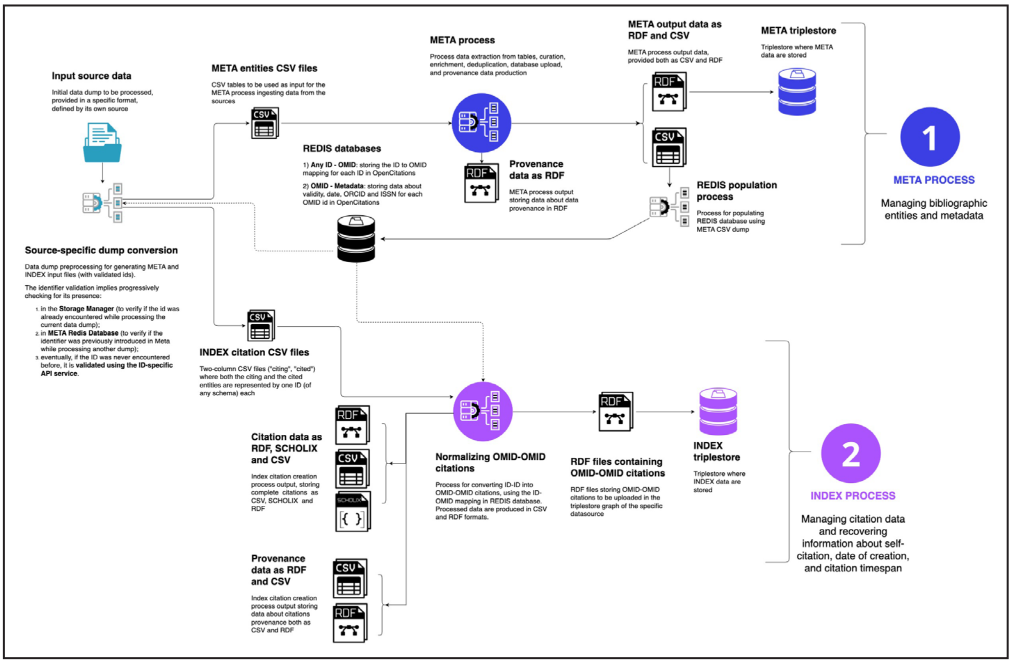
Input source data
Initial data dump to be processed, provided in a specific format, defined by its own source
META entities CSV files
CSV tables to be used as input for the META process ingesting data from the sources
CSV
META process
Process data extraction from tables, curation, enrichment, deduplication, database upload, and provenance data production
META output data as RDF and CSV
META process output data, provided both as CSV and RDF
RDF
CSV
META triplestore
Triplestore where META data are stored
REDIS databases
1) Any ID - OMID: storing the ID to OMID mapping for each ID in OpenCitations
2) OMID - Metadata: storing data about validity, date, ORCID and ISSN for each OMID id in OpenCitations
Provenance data as RDF
META process output storing data about data provenance in RDF
REDIS population process
Process for populating REDIS database using META CSV dump
1
META PROCESS
Managing bibliographic entities and metadata
Source-specific dump conversion
Data dump preprocessing for generating META and INDEX input files (with validated ids).
The identifier validation implies progressively checking for its presence:
1. in the Storage Manager (to verify if the id was already encountered while processing the current data dump);
2. in META Redis Database (to verify if the identifier was previously introduced in Meta while processing another dump);
3. eventually, if the ID was never encountered before, it is validated using the ID-specific API service.
INDEX citation CSV files
Two-column CSV files ("citing", "cited") where both the citing and the cited entities are represented by one ID (of any schema) each
Citation data as RDF, SCHOLIX and CSV
Index citation creation process output, storing complete citations as CSV, SCHOLIX and RDF
SCHOLIX
Normalizing OMID-OMID citations
Process for converting ID-ID into OMID-OMID citations, using the ID-OMID mapping in REDIS database. Processed data are produced in CSV and RDF formats.
RDF files containing OMID-OMID citations
RDF files storing OMID-OMID citations to be uploaded in the triplestore graph of the specific datasource
INDEX triplestore
Triplestore where INDEX data are stored
2
INDEX PROCESS
Managing citation data and recovering information about self-citation, date of creation, and citation timespan
Provenance data as RDF and CSV
Index citation creation process output storing data about citations provenance both as CSV and RDF

Figure 1 Workflow for the ingestion of citation data and bibliographic metadata into the OpenCitations datasets.

[7] https://creativecommons.org/public-domain/cc0/.

[8] https://figshare.com/authors/OpenCitations_/3068259.

[9] http://opencitations.net/querying.

As anticipated, a section of the methodology was formalized in a workflow (Moretti and Heibi 2023) published on the Social Sciences & Humanities Open Marketplace[10] (Barbot et al. 2019; Ilijašić Veršić & Ausserhofer 2019). However, while this workflow primarily focuses on citation data production, here we provide an overview of its application in the JaLC case study by exposing it in conjunction with the procedure for producing bibliographic metadata. Indeed, OpenCitations' data integration methodology relies on the interconnection between bibliographic and citation data creation processes.

3.1.1 Data Source Selection and Documentation Analysis

The Japan Link Center is a Japanese DOI registration agency, and the data ingestion in OpenCitations was planned as an agreement between the parties. Accordingly, the JaLC technical team organized the data exposed via their API into a dataset in JSON format, to allow OpenCitations to reuse its previously developed software components.

The resulting dataset is organized as follows. A main ZIP archive contains the dataset directory, which stores a JSON file about all the DOI prefixes handled and a series of ZIP archives. Each of these archives is named after a DOI prefix and contains a directory storing all the JSON files of the DOI entities having that prefix – thus, the number of files per directory is variable. Note that each file represents a single bibliographic entity and citation data are provided as its bibliographic metadata.

To gain a deep understanding of the source data model, we conducted an e-mail correspondence to clarify any doubts and studied in detail the documentation,[11] available both in Japanese and in English translation.

For the reproducibility of the process, it is crucial to declare that the JaLC input dataset was not made publicly available as it was ingested by OpenCitations. Despite this, the contained data can be freely retrieved by querying the API[12] with the help of the provided documentation[13] and used to restructure the bibliographic information as exemplified in the sample data used for OpenCitations' software tests.[14] As a result of possible API source data updates, the obtained dataset might contain more recent information than OpenCitations'.

3.1.2 Development of a Software Plug-in for Data Conversion

Although the current OpenCitations software infrastructure allows for the reuse of many general components common to the processing procedure of all the input datasets, two specific extensions had to be developed for each new data source. The first is aimed at reading the structure of the source dataset, extracting the bibliographic entities data and citations, and producing output files; the second performs the metadata crosswalk between the data model used by the source[15] – in this case, JaLC – and OCDM.[16] These components are developed as plugins of the OpenCitations software `oc_ds_converter` (Moretti et al. 2024), released under an ISC license.

Since JaLC is the DOI registration agency responsible for assigning DOIs to the citing entities, these identifiers were accepted as valid without further checks. Nevertheless, no information was provided about the registration agencies that assigned DOIs to the cited entities. Thus, to avoid redundant API checks, we adopted a two-step validation process relying on an ad-hoc data storage system.

During the first iteration of the dataset, all citing DOIs are accepted as valid and stored in memory as such. Metadata CSV tables are produced, concerning the citing entities only (Figure 2). In the second iteration, the cited identifiers are analyzed to verify their validity. During this phase, we produce metadata and citations CSV tables concerning the cited entities whose DOIs proved to be valid (Figure 3).

[10] https://www.sshopencloud.eu/ssh-open-marketplace.

[11] https://japanlinkcenter.org/top/doc/REST_API_Functional_Description.pdf.

[12] https://api.japanlinkcenter.org/.

[13] https://api.japanlinkcenter.org/api-docs/index.html.

[14] https://github.com/opencitations/oc_ds_converter/tree/4382eea1fdce83945c88e6da76da2f2dbd49a2f7/test/jalc_process/sample_dump.

[15] https://github.com/opencitations/oc_ds_converter/blob/a5316f09a6b7fdefc42a6b327fcdf8374114fa26/oc_ds_converter/jalc/jalc_processing.py.

[16] https://github.com/opencitations/oc_ds_converter/blob/a5316f09a6b7fdefc42a6b327fcdf8374114fa26/oc_ds_converter/run/jalc_process.py.

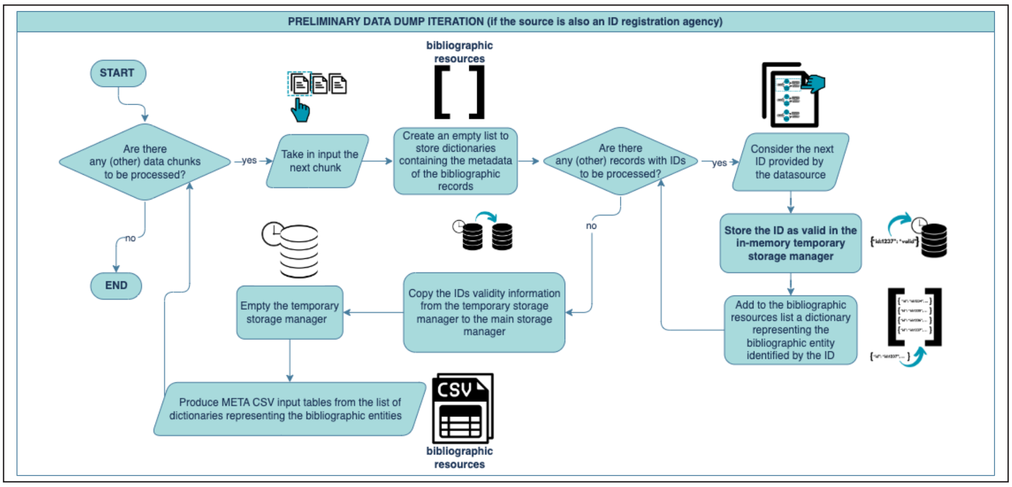
PRELIMINARY DATA DUMP ITERATION (if the source is also an ID registration agency)
START
Are there any (other) data chunks to be processed?
yes
no
END
Take in input the next chunk
bibliographic resources
Create an empty list to store dictionaries containing the metadata of the bibliographic records
Are there any (other) records with IDs to be processed?
yes
no
Consider the next ID provided by the datasource
Store the ID as valid in the in-memory temporary storage manager
Add to the bibliographic resources list a dictionary representing the bibliographic entity identified by the ID
Copy the IDs validity information from the temporary storage manager to the main storage manager
Empty the temporary storage manager
Produce META CSV input tables from the list of dictionaries representing the bibliographic entities
CSV
bibliographic resources

Figure 2 Flowchart describing the preliminary processing of citing bibliographic entities.

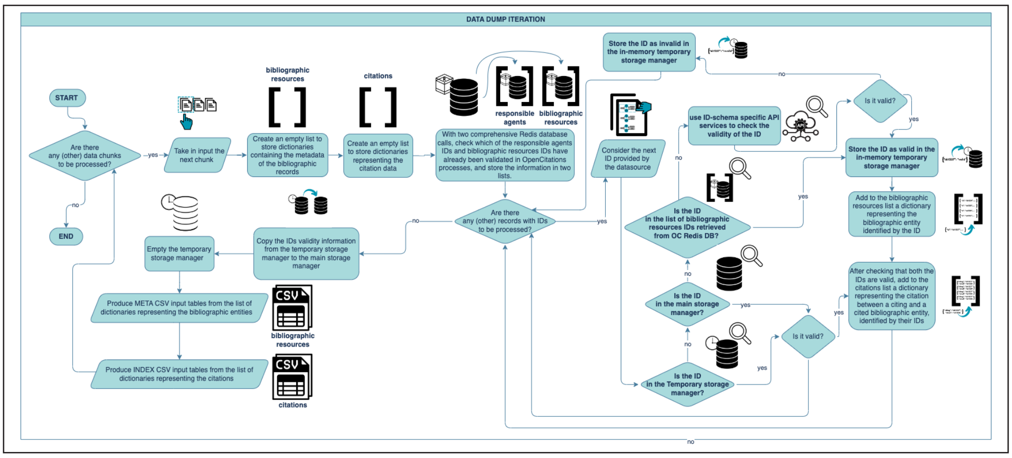
DATA DUMP ITERATION
START
Are there any (other) data chunks to be processed?
yes
no
END
Take in input the next chunk
bibliographic resources
Create an empty list to store dictionaries containing the metadata of the bibliographic records
citations
Create an empty list to store dictionaries representing the citation data
responsible agents
bibliographic resources
With two comprehensive Redis database calls, check which of the responsible agents IDs and bibliographic resources IDs have already been validated in OpenCitations processes, and store the information in two lists.
Are there any (other) records with IDs to be processed?
yes
no
Consider the next ID provided by the datasource
Is the ID in the Temporary storage manager?
Is the ID in the main storage manager?
Is the ID in the list of bibliographic resources IDs retrieved from OC Redis DB?
use ID-schema specific API services to check the validity of the ID
API
Is it valid?
Store the ID as invalid in the in-memory temporary storage manager
Store the ID as valid in the in-memory temporary storage manager
Add to the bibliographic resources list a dictionary representing the bibliographic entity identified by the ID
After checking that both the IDs are valid, add to the citations list a dictionary representing the citation between a citing and a cited bibliographic entity, identified by their IDs
Copy the IDs validity information from the temporary storage manager to the main storage manager
Empty the temporary storage manager
Produce META CSV input tables from the list of dictionaries representing the bibliographic entities
Produce INDEX CSV input tables from the list of dictionaries representing the citations
CSV
bibliographic resources
citations

Figure 3 Flowchart describing the processing of cited bibliographic entities, their validation, and the production of metadata and citation tables.

The process leverages a system involving multiple storage solutions to prevent duplications and minimize external API calls. The validation pipeline includes the checks listed below:

1. Search for the identifier in the temporary in-memory storage containing data concerning the current chunk of data being processed. If the identifier is among these data, it can be considered valid, since it was encountered and validated while processing the current data chunk.
2. If the identifier was not found in (1), search for it in the main storage containing data concerning the whole dataset. Finding the DOI here implies its validity since it was detected and validated previously.
3. If neither (1) nor (2) was successful, search for the identifier in the OpenCitations databases, containing a mapping between each identifier ever encountered while ingesting any dataset and its assigned OMID. If the DOI is found in there, it is considered valid.
4. Use ID-schema-specific API services to check the validity of the ID if none of the previous attempts was successful.

After having processed each chunk of the dataset, the data tables for bibliographic and citation information are generated. These tables (CSV) serve as input for the two tools performing the data production tasks for Meta and Index. Simultaneously, the ID validation information in the temporary memory storage is transferred to the permanent storage, collecting all the valid identifiers encountered in the dataset up to the current moment.

3.1.3 Production of Metadata and Citation Data Collections

Using the software extensions on the dataset provided by JaLC as input, tables of bibliographic and citation data are obtained. These tables are used as input for the subsequent steps of the process.

ID	TITLE	AUTHOR	PUB_DATE	VENUE	VOLUME	ISSUE	PAGE	TYPE	PUBLISHER	EDITOR
DOI: 10.14825/kaseki.68.0_14	本邦産白亜紀アンモナイトデータベースおよび種多様性について	利光, 誠一; 平野, 弘道; 松本, 崇; 高橋, 一晴	2000	化石 [issn:0022-9202 issn:2424-2632 jid:kaseki]	68	0	14–16	journal article	日本古生物学会	
DOI: 10.1126/science.235.4793.1156	Chronology of fluctuating sea levels since the Triassic		1987	Science	235		1156–1167			

Table 1 Sample of Meta input tables produced by `oc_ds_converter`, storing bibliographic entities' metadata.

The bibliographic entities CSV tables (Table 1) are used as input for the Meta software (Massari et al. 2024b), which curates the provided information and generates new data compliant with OCDM.

The DOI-to-DOI citation CSV tables (Table 2) serve as input for the Index software (Heibi et al. 2024), a tool used for producing collections of references between bibliographic entities identified by OMIDs.

CITING	CITED
DOI: 10.14825/kaseki.68.0_14	DOI: 10.1126/science.235.4793.1156

Table 2 Sample of Index input tables, produced by `oc_ds_converter`, storing citation data.

3.1.4 Ingestion of Metadata Collection into OpenCitations Meta

We use the Meta software to perform curation tasks and produce the collection of bibliographic metadata, where JaLC bibliographic entities will be included together with the ones derived from all the other sources. The software assigns a new OMID to each new entity and propagates an existing OMID to those records that have already been included in OpenCitations Meta. This latter operation is performed by deduplicating identical entities ingested from different data sources, potentially with multiple identifiers.

3.1.5 Production of Citation Data

After the integration of the new bibliographic records in Meta, the Index software, which is responsible for producing citation data (RDF, SCHOLIX, and CSV) compliant with OCDM, generates the OMID-to-OMID citations from JaLC data. In this step, OpenCitations Meta is exploited to retrieve OMIDs.

3.1.6 Input Dataset Analysis and Multilingualism Information Loss Forecast

OCDM allows only a single value for each metadata field (title, authors, venue, etc.). Therefore, we prioritized metadata in the original language when both the original and English versions were provided. However, an analysis of the initial JaLC dataset revealed that, in a few instances, the declared original language is not Japanese, and linguistic information is not always provided. Adding to the complexity, the dataset permits the specification of multiple values for publishers related to the same entity. This goes beyond allowing different translations of the same publisher name; it extends to cases where a single entity may have associations with multiple distinct publishers. Therefore, to avoid attributing wrong metadata information to bibliographic entities when the linguistic information is not formally specified and when the declared language differs from Japanese or English, our approach was to prioritize the first encountered value. We assumed the first available value to be the most commonly used, thereby mitigating the risk of inaccurate metadata attribution. This choice was motivated by the need for a pragmatic solution following OCDM without introducing data inconsistencies.

We sought to assess the impact of the single-language constraint by analyzing specific metadata elements in the JaLC dataset, such as the bibliographic entity title, publication venue title, and author names. In the case of JaLC, we found that the metadata attributed with the highest forecasted loss of information compared to the input dataset (since only one language can be accepted) is the citing entities' journal title (41.44%) (Figure 5, Table 3).

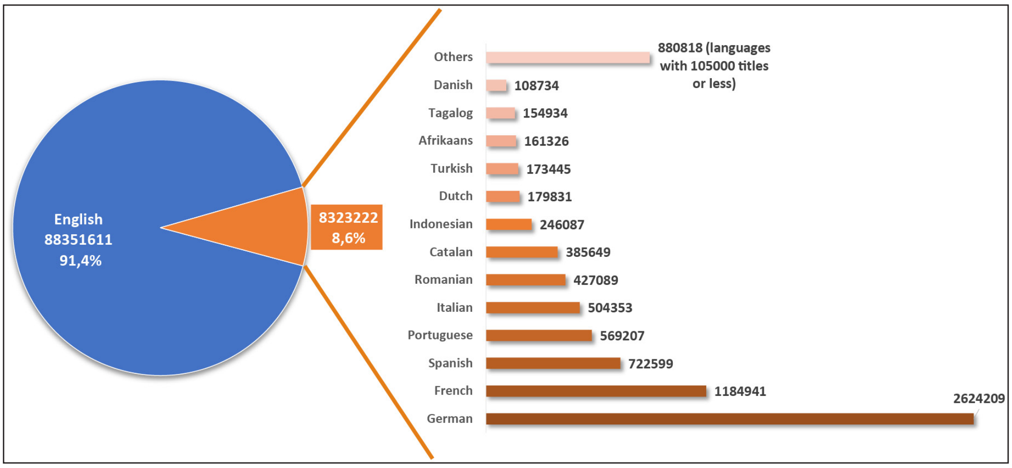

Figure 4 Language distribution in Meta bibliographic entities, calculated on Meta dump, version 5 (https://doi.org/10.6084/m9.figshare.21747461.v5). The analysis was performed on bibliographic entities with a declared title.

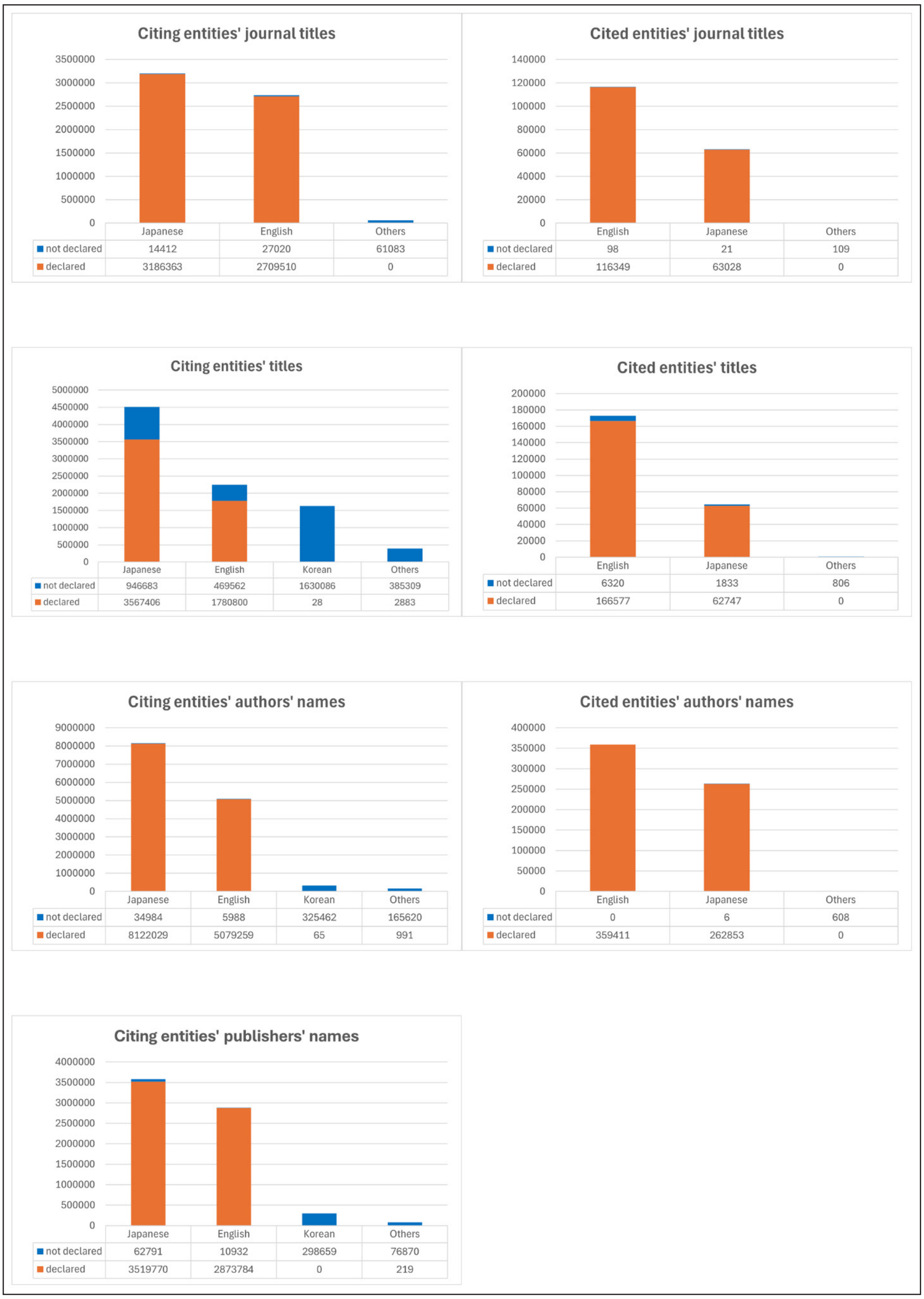

Figure 5 Bar charts illustrating the analysis of multilingualism within the input dataset, categorized by bibliographic metadata fields.

	1 LANGUAGE	2 LANGUAGES	3+ LANGUAGES	TOTAL VALUES PROVIDED	INFORMATION LOSS WRT. THE ORIGINAL DATASET
title citing	5,701,285	1,641,895	39 (3 languages)	8,985,192	1,641,973; 18.27%
title cited	217,316	12,616	0	242,548	12,616; 5.2%
authors citing	9,892,522	4,556,812	39 (3 languages)	19,006,263	4,556,890; 23,98%
authors cited	308,079	157,556	0	623,191	157,556; 25.28%
journal title citing	1,137,368	2,658,678	21,213 (20,572 3 languages; 641 4 languages)	6,519,004	2,701,745; 41.44%
journal title cited	180,515	0	0	180,515	0

Table 3 Table showing the metadata languages in the original dataset and the linguistic information loss due to OCDM constraints. The total amount of metadata provided for a field is the sum of the number of values provided solely in one language, twice the number of values supplied in two languages, and the product between the number of values provided in more than two languages and the precise number of furnished languages. The information loss is calculated as the sum of values provided in more languages out of the total calculated. The publisher's name field has not been included in the table since it does not necessarily concern the loss of linguistic information but might involve cases where the information loss derives from having multi-publisher values.

3.2 TOOLS AND SOFTWARE

Three OpenCitations software tools are involved in the process:

1. **`oc_ds_converter`** (Moretti et al. 2024). This tool handles metadata crosswalks from the JaLC data model to OCDM, the identifier validation, and the production of citations and bibliographic data in CSV format, meant to serve as input for subsequent processes. This software includes two nested modules, i.e. the `oc_idmanager`, for the validation of persistent identifiers, and the `oc_data_storage`, for the storage system management. On the occasion of codebase extension for JaLC integration, the Identifier Manager module was extended to handle JIDs, the identifiers assigned by J-STAGE[17] to publication venues (e.g. journals). To facilitate the reuse of the software components, we released a Python package on PyPI. The latest version is 1.0.0,[18] distributed on 27 October 2023.
2. **`oc_meta`** (Massari et al. 2024b). This software curates bibliographic data, deduplicates entities, assigns OMIDs, and generates a metadata dataset related to the bibliographic entities involved in OpenCitations Index citations. As output, it produces bibliographic data in RDF and CSV formats, and a dataset for provenance and change tracking in RDF. For this software, a PyPI package was released to maximize reuse potential: the latest available version is 1.2.4,[19] published on 16 February 2023.
3. **`index`** (Heibi et al. 2024). The Index software produces RDF, CSV, and SCHOLIX data formats of OMID-to-OMID citations and a corresponding collection of provenance data in CSV and RDF formats. The current version can be consulted in the "meta-index" branch,[20] bound to be merged into the master branch once a stable version is assessed.

Each software is released under an ISC license and hosted on a public GitHub repository, developed following a Test-Driven Development approach (Tilley 2004) and monitoring the percentage of tested code features with the Python "coverage" library.[21] Currently, 83% of both `oc_ds_converter` and `oc_meta` software code is covered by tests. All the abovementioned repositories come with README documentation and use Poetry[22] as a dependency management system to facilitate maintenance and foster workflow reproduction.

3.3 LANGUAGE AGNOSTIC APPROACH

Despite the current dominance of English-language content on the web, language-agnostic architectures are needed to address the challenges posed by globalization and the rising demand for multilingual web accessibility. Nonetheless, a universal solution for representing,

[17] https://www.jstage.jst.go.jp/browse/-char/ja.

[18] https://pypi.org/project/oc-ds-converter/1.0.0.

[19] https://pypi.org/project/oc-meta.

[20] https://github.com/opencitations/index/tree/meta-index.

[21] https://pypi.org/project/coverage/.

[22] https://python-poetry.org/.

storing, and processing multilingual data (Jain & Kysliak 2022) and, in particular, bibliographic data (Bergamin & Guerrini 2022) is still lacking. Over time, OpenCitations has developed general solutions aimed at meeting the needs of a wide range of data to be incorporated into a comprehensive and constantly expanding database capable of accepting information from diverse sources irrespective of the language in which it is generated (Cameron 1997). By not allowing the specification of multiple translations for the same value, the OCDM imposes the selection of a single language for each bibliographic metadata entry, even in cases where the data source provides more options, leading to an inevitable loss of information.

Recognizing the implications of this endeavor, we emphasize the need to strike a delicate balance between preserving linguistic elements and respecting infrastructural and data model constraints, leading us to the choice of adopting a language-agnostic approach. For this reason, we decided to store metadata in the original language only, where it is possible. This choice is the most suitable in our case since it allows the preservation of bibliodiversity over global uniformity in a dominant language.

3.3.1 The Authors' Names Management

In the context of adopting a language-agnostic approach, in addition to the project's design goal of preserving linguistic diversity, additional considerations have emerged during the consolidation phase of the methodology. Particularly, a noteworthy case came out during the process of cleaning and standardizing the names of authors in the Meta dataset.

Initially, there was a proposition to eliminate all characters from names except for letters, numbers, periods preceded by a letter, and the ampersand. However, such a decision assumes an exhaustive knowledge of all permissible characters in personal names worldwide. This would require an understanding of all global alphabets and discerning which letters from these alphabets are genuinely used in personal names. For instance, we would need to consider scripts ranging from Basic Latin, Latin-1 Supplement, and Latin Extended-A, to Cyrillic Supplement, Armenian, Hebrew, and more. A pertinent question that arises is whether African click letters (ǀ, ǁ, ǂ, and !) can be used in personal names. While it is feasible to craft a regular expression capturing all these alphabets broadly (e.g., '[A-Za-zÀ-ÖØ-öø-ÿĀ-ňƀ-ǌΑ-ω…]'), it does not seem to be a robust solution. Furthermore, other characters, not necessarily letters, are permissible in personal names. For instance, characters resembling an apostrophe (e.g., O'Connel) and a hyphen (e.g., Mun, Ji-Hye) are valid. Given these complexities, we decided to remove only the reserved characters used in the syntax of the CSV files with which Meta is populated, specifically ";", "[", ",", and "]". We chose to remain agnostic regarding all other characters. Our observations indicated that creating whitelists introduced far more errors than those resolved. This is primarily because the diversity in names is vast, and it is impractical to verify them all.

3.3.2 Handling a Multilingual Data Source

In this section, we delve into some additional methodological precautions of specific interest for handling multilingual or predominantly non-English sources.

- **Metadata Mapping and Data Selection.** Accurate metadata mapping is essential for a successful data ingestion process. Collaborative efforts with data providers facilitate mapping source metadata to the end data model, aligning language-specific terms and concepts.
- **Proper Encoding-Decoding Choices.** Appropriate encoding and decoding choices ensure the accurate transformation of data formats while preserving information integrity. When managing multilingual corpora, it is crucial to select broad encodings or develop ad-hoc solutions. For the JaLC Anglo-Japanese dataset, we adopted UTF-8 (Yergeau 2003), one of the most flexible encodings for representing Unicode[23] characters (Memon 2001), as it can represent any language. However, as a broad rule for all the cases when it is not possible to know the languages included in a multilingual dataset in advance, avoiding ASCII is a good practice, even if it is enough to represent English and Latin characters (Lide 2002).

[23] https://www.unicode.org/versions/Unicode15.1.0/.

- **Culture-Specific Considerations.** Each language has its peculiarities. In Asian languages, such as Japanese, homonymy is more common in surnames than in given names. For this reason, in the absence of a distinct and persistent identifier, we suggest facing the challenge with ad-hoc solutions, such as using an external DOI to ORCID mapping.

4. RESULTS AND DISCUSSION

4.1 CURRENT DATASET OVERVIEW

The infrastructure currently comprises 1,975,552,846 unique citations and 114,621,237 bibliographic entities. Based on a computational analysis of the titles of the publications performed in January 2024 with a script[24] exploiting the "langdetect" Python library for language detection,[25] 15% of the bibliographic metadata stored in OpenCitations Meta are not in English (Figure 6), representing a 6.4% increase compared to the previous version of the dataset (Figure 4) following the ingestion of JaLC data.

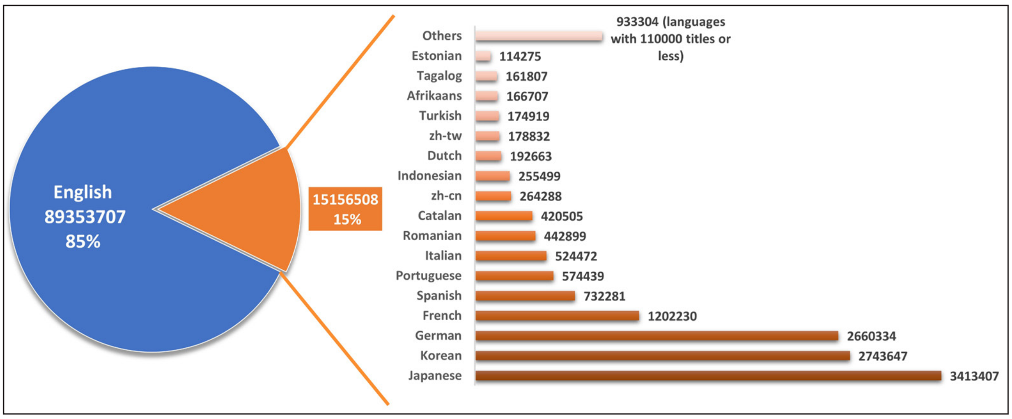

Figure 6 Language distribution in Meta bibliographic entities, calculated on Meta dump, version 6 (https://doi.org/10.6084/m9.figshare.21747461.v6). The analysis was performed on bibliographic entities with a declared title.

4.2 OPENCITATIONS DATASETS AFTER JaLC DATASET INGESTION

The JaLC input dataset counts 7,343,638 bibliographic entities and 416,125 citations. After processing the data, we eventually included in OpenCitations 7,333,238 bibliographic entities and 396,788 citations. Until the integration of this dataset, only bibliographic entities involved in Index citations were included in OpenCitations Meta. However, from a prior analysis of the JaLC dataset, we found that 6,908,305 entities had no citations, 329,078 entities were involved in citations to publications not identified with persistent identifiers managed by OpenCitations, and only 106,255 entities cited publications identified by DOI. For this reason, and given the small number of entities to be processed, we decided to integrate all bibliographic records in the dataset with DOIs assigned by JaLC into Meta. As can be seen in Table 4, it is not unusual for different sources to have overlapping information.

Table 4 Yellow cells represent the single contribution of each collection to OpenCitations Index, i.e., the number of citations uniquely derived by a given source. Pink cells represent the number of citations in the sources' intersection. The table is based on OpenCitations data at its latest update (29 November 2023).

	INDEX	Crossref	DataCite	PubMed	OpenAire	JaLC
INDEX	1,975,552,846	1,563,218,160	169,814,412	695,988,810	14,645,838	396,788
Crossref		1,100,963,346	27,051	458,309,297	3,917,329	1,137
DataCite			169,663,255	9,623	114,483	0
PubMed				237,208,867	9,711,789	125
OpenAire					1,067,712	0
JaLC						395,526

However, only 1,137 of the citations by JaLC are in common with Crossref, and the majority of its contribution to OpenCitations is unique.

[24] https://github.com/opencitations/oc_meta/blob/0e4bd2004592ab59c67bfb5a2fecffb5d3e843a7/oc_meta/plugins/language_count.py.

[25] https://pypi.org/project/langdetect/.

The most recent update of the datasets dates back to 29 November 2023. According to these data, the OpenCitations Index includes information collected from five sources and comprises 1.97 billion citations involving 72,848,673 citing entities and 71,805,805 cited entities, for a total of 89,920,081 unique entities involved in at least one citation link.

4.3 JaLC DATA RETRIEVAL FROM OPENCITATIONS DATASETS

JaLC data were included both in Meta and Index, available in all the formats mentioned above.

To retrieve JaLC citations from Index, there are two methods. The first approach implies retrieving the identifiers of the citations from the provenance dataset and then looking for the collected OCIs in the OpenCitations Index CSV dump. More in detail:

1. Access the OpenCitations Index on the “Download” page.[26]
2. Download the dataset *Citation data sources’ info (N-Triple): information regarding the data source collection*.
3. Identify OCI subjects containing the string “joci”, standing for JaLC OpenCitations Index, to locate citations from JaLC.
4. Download the citation data dataset in CSV format.
5. Match OCIs in the CSV dataset with those from N-Triples to obtain JaLC citation information.

The alternative approach is suggested for the Semantic Web experts and implies using the OpenCitations Index SPARQL Endpoint to execute a query to retrieve subjects of triples with `<http://www.w3.org/ns/prov#atLocation>` property and the IRI identifying the internal OpenCitations Index collection dedicated to JaLC derived data (https://w3id.org/oc/index/joci) as the object, as shown in the following SPARQL query:

```
SELECT ?s WHERE {
  ?s
      <http://www.w3.org/ns/prov#atLocation>
         <https://w3id.org/oc/index/joci/>.
}
```

To access JaLC records from OpenCitations Meta, it is key to note that primary source information is contained in the provenance data, as per the OCDM model. Each Meta entity is linked to a snapshot `(prov:Entity)` via `prov:specializationOf`. This snapshot includes a `prov:hadPrimarySource` property indicating the primary source. For JaLC records, the primary source is `https://api.japanlinkcenter.org/`. Since the provenance data are not in a triplestore, downloading the Meta Provenance dataset is necessary to identify JaLC records.

5. IMPLICATIONS/APPLICATIONS

5.1 REUSE POTENTIAL

The user base of OpenCitations data includes Funders, Resource and Research Managers, Researchers, Policy Makers, Research Organisations, and Providers.[27] More in detail, the OpenCitations datasets benefit scholars in developing countries and professionals outside academic institutions without access to commercial citation indexes, as well as ordinary citizens seeking open data, open science partners, academic publishers, tool developers, bibliometricians, and librarians (Peroni & Shotton 2022a; Peroni & Shotton 2022b). OpenCitations’ citation collection is currently used in several projects, e.g. B!son,[28] Optimeta,[29] repositories, e.g. the Staatsbibliothek zu Berlin[30] and ORBi ULiege,[31] and search tools, e.g. PURE suggest.[32]

[26] https://opencitations.net/download.

[27] https://marketplace.eosc-portal.eu/services/opencitations/details.

[28] https://service.tib.eu/bison/.

[29] https://projects.tib.eu/optimeta/en/.

[30] https://blog.sbb.berlin/zitationsbasierte-recherche/.

[31] https://orbi.uliege.be/.

[32] https://fabian-beck.github.io/pure-suggest/.

JaLC data integration opens up interesting usage prospects, especially for bibliometric studies on bibliographic metadata and citations among non-English resources. This aspect is especially relevant for bibliodiversity preservation in cultural heritage and humanities-oriented studies, in which awareness toward reproducibility and replicability of research is still being established (Peels & Bouter 2018), and the introduction of open access tools that facilitate open science practices is a necessary starting point. In addition, the adopted ingestion workflow not only paves the way for addressing the intricacies of multilingual data ingestion but also opens doors for broader applications. Researchers and institutions dealing with non-English-based data sources can leverage such workflow, adapting and customizing it to suit their specific needs.

5.2 FUTURE DEVELOPMENTS

As a future development, since OCDM excludes multilingual storage but the language of cataloging can differ from the users' language preferences, data reuse limitations could be contrasted with software-driven data retrieval solutions for extemporaneous translation. In this perspective, scalable solutions to the lack of common practices for multilingual access include promoting the use of controlled terms and established value vocabularies for simplicity and cost-effectiveness (Riva 2022). However, the current state-of-the-art open-source solutions do not guarantee the level of accuracy we aim to achieve. Thus, since sophisticated tools are needed for cross-lingual information retrieval, we plan to develop software exploiting open-source technologies for data exposure only, specifically trained on translation tasks regarding bibliographic metadata. Of course, the tool would come with a clear statement of the nature of the data displayed, i.e., whether it is presented in its original form or has been translated. Recently, similar tasks were addressed with approaches to multilingual information retrieval based on using pre-trained multilingual language models (Hu et al. 2023). Such a solution would limit the storage issues caused by the maintenance of the English translations of the data, in addition to respecting the OCDM structure and maximizing the reuse potential of exposed data, therefore leveraging a balance between the need for information preservation and the constraints of physical space and data model nature.

DATA ACCESSIBILITY STATEMENT

The two main datasets are deposited on Figshare under a CC0 license and can be downloaded in different formats, alongside their provenance information.

Meta:

- OpenCitations Meta CSV dataset of all bibliographic metadata (https://doi.org/10.6084/m9.figshare.21747461.v6)
- OpenCitations Meta RDF dataset of all bibliographic metadata and its provenance information (https://doi.org/10.6084/m9.figshare.21747536.v5)

Index:

- OpenCitations Index CSV dataset of all the citation data (https://doi.org/10.6084/m9.figshare.24356626.v2)
- OpenCitations Index N-Triples dataset of all the citation data (https://doi.org/10.6084/m9.figshare.24369136.v2)
- OpenCitations Index Scholix dataset of all the citation data (https://doi.org/10.6084/m9.figshare.24416749.v2)
- OpenCitations Index CSV dataset of the provenance information of all the citation data (https://doi.org/10.6084/m9.figshare.24417733.v2)
- OpenCitations Index N-Triples dataset of the provenance information of all the citation data (https://doi.org/10.6084/m9.figshare.24417736.v2)

ACKNOWLEDGEMENTS

We want to thank the organizers of the Open Bibliodata Workflows Booksprint for introducing us to the use of the Social Sciences & Humanities Open Cloud Marketplace for the publication of formally structured workflows. Additionally, we express our gratitude to the team of researchers and students from the University of Bologna who contributed to the extension of the software infrastructure of OpenCitations for ingesting the JaLC dataset. We also thank the Japan Link Center for generously providing us with the input data, which significantly enhanced the quality and depth of the OpenCitations collections.

FUNDING INFORMATION

This work was partially funded by the European Union's Horizon Europe Framework Programme under grant agreement No 101095129 (GraspOS Project) and by Project PE 0000020 CHANGES – CUP B53C22003780006, NRP Mission 4 Component 2 Investment 1.3, funded by the European Union – NextGenerationEU.

COMPETING INTERESTS

The authors have no competing interest to declare.

AUTHOR CONTRIBUTIONS

Arianna Moretti: Conceptualization, Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing

Marta Soricetti: Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing

Ivan Heibi: Conceptualization, Formal analysis, Methodology, Software, Writing – review & editing

Arcangelo Massari: Conceptualization, Data curation, Formal analysis, Software, Writing – original draft

Silvio Peroni: Conceptualization, Funding acquisition, Methodology, Project administration, Writing – review & editing

Elia Rizzetto: Writing – original draft, Writing – review & editing, Software

AUTHOR AFFILIATIONS

Arianna Moretti orcid.org/0000-0001-5486-7070
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
Marta Soricetti orcid.org/0009-0008-1466-7742
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
Ivan Heibi orcid.org/0000-0001-5366-5194
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
Arcangelo Massari orcid.org/0000-0002-8420-0696
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
Silvio Peroni orcid.org/0000-0003-0530-4305
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
Elia Rizzetto orcid.org/0009-0003-7161-9310
Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

REFERENCES

Alexiou, G., Vahdati, S., Lange, C., Papastefanatos, G., & **Lohmann, S.** (2016). OpenAIRE LOD Services: Scholarly Communication Data as Linked Data. In A. González-Beltrán, F. Osborne, & S. Peroni (Eds.), *Semantics, Analytics, Visualization. Enhancing Scholarly Data* (Vol. 9792, pp. 45–50). Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-53637-8_6

Barbot, L., Moranville, Y., Fischer, F., Petitfils, C., Ďurčo, M., Illmayer, K., Parkoła, T., Wieder, P., & **Karampatakis, S.** (2019). *SSHOC D7.1 System Specification—SSH Open Marketplace*. DOI: https://doi.org/10.5281/ZENODO.3547649

Bergamin, G., & **Guerrini, M.** (Eds.). (2022). *Bibliographic control in the digital ecosystem*. Associazione italiana biblioteche. DOI: https://doi.org/10.36253/978-88-5518-544-8

Berners-Lee, T., & **Kagal, L.** (2008). The Fractal Nature of the Semantic Web. *AI Magazine*, *29*(3), 29–34. DOI: https://doi.org/10.1609/aimag.v29i3.2161

Burton, A., Fenner, M., Haak, W., & **Manghi, P.** (2017). *Scholix Metadata Schema For Exchange Of Scholarly Communication Links*. DOI: https://doi.org/10.5281/ZENODO.1120261

Cameron, R. D. (1997). A Universal Citation Database. *First Monday*, *2*(4). DOI: https://doi.org/10.5210/fm.v2i4.522

Canese, K., & **Weis, S.** (2013). PubMed: The Bibliographic Database. In: *The NCBI Handbook [Internet]*. 2nd edition. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK153385/

Chen, Y.-N. (2015). A RDF-based approach to metadata crosswalk for semantic interoperability at the data element level. *Library Hi Tech*, *33*(2), pp. 175–194. DOI: https://doi.org/10.1108/LHT-08-2014-0078

DataCite Metadata Working Group. (2021). *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4* [Application/pdf]. 82 pages. DOI: https://doi.org/10.14454/3W3Z-SA82

Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., Mayr, P., Romanello, M., & **Zumstein, P.** (2020a). The OpenCitations Data Model. In: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020*, pp. 447–463. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-62466-8_28

Daquino, M., Peroni, S., Shotton, D., & **Massari, A.** (2020b). *The OpenCitations Data Model*. Version 2.0.1. Figshare. DOI: https://doi.org/10.6084/M9.FIGSHARE.3443876.V7

Heibi, I., Moretti, A., Grieco, G., & **Peroni, S.** (2024). *OpenCitations: Index* [Computer software]. https://archive.softwareheritage.org/swh:1:snp:edecc78aff97644d83e48f7373710554a2f55721;origin=https://github.com/opencitations/index

Heibi, I., Peroni, S., & **Shotton, D.** (2019). Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. *Scientometrics*, *121*(2), pp. 1213–1228. DOI: https://doi.org/10.1007/s11192-019-03217-6

Hendricks, G., Tkaczyk, D., Lin, J., & **Feeney, P.** (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), pp. 414–427. DOI: https://doi.org/10.1162/qss_a_00022

Horvath, A. (2021). *Enhancing Language Inclusivity in Digital Humanities: Towards Sensitivity and Multilingualism: Includes interviews with Erzsébet Tóth-Czifra and Cosima Wagner*, *1*(1), Article 1, pp. 1–21. DOI: https://doi.org/10.3828/mlo.v0i0.382

Hutchins, B. I., Baker, K. L., Davis, M. T., Diwersy, M. A., Haque, E., Harriman, R. M., Hoppe, T. A., Leicht, S. A., Meyer, P., & **Santangelo, G. M.** (2019). The NIH Open Citation Collection: A public access, broad coverage resource. *PLOS Biology*, *17*(10), e3000385. DOI: https://doi.org/10.1371/journal.pbio.3000385

Hu, X., Chen, X., Qi, P., Kong, D., Liu, K., Wang, W. Y., & **Huang, Z.** (2023). *Language Agnostic Multilingual Information Retrieval with Contrastive Learning* (arXiv:2210.06633). arXiv. DOI: https://doi.org/10.18653/v1/2023.findings-acl.581

Ilijašić Veršić, I., & **Ausserhofer, J.** (2019). Social sciences, humanities and their interoperability with the European Open Science Cloud: What is SSHOC? *Mitteilungen Der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare*, *72*(2), pp. 383–391. DOI: https://doi.org/10.31263/voebm.v72i2.3216

Jain, S., & **Kysliak, A.** (2022). Language-Agnostic Knowledge Representation for a Truly Multilingual Semantic Web. *International Journal of Information System Modeling and Design*, *13*(1), pp. 1–21. DOI: https://doi.org/10.4018/IJISMD.297045

La Bruzzo, S., Baglioni, M., Atzori, C., & **Manghi, P.** (2023). Scholix dump of the OpenAIRE inferred citations [dataset]. *Zenodo*. DOI: https://doi.org/10.5281/ZENODO.7845968

Lide, D. R. (Ed.). (2002). Code for Information Interchange—ASCII. In *A Century of Excellence in Measurements, Standards, and Technology*. CRC Press.

Massari, A., Mariani, F., Heibi, I., Peroni, S., & **Shotton, D.** (2024a). OpenCitations Meta. *Quantitative Science Studies* (to appear). DOI: https://doi.org/10.48550/arXiv.2306.16191

Massari, A., Persiani, S., Mariani, F., Peroni, S., Soricetti, M., & **Moretti, A.** (2024b). *OpenCitations Meta Software* [Computer software]. swh:1:snp:f4d58a91d35b2ffb7aa576cd8c4ef04005dd852b;origin=https://github.com/opencitations/oc_meta

Massari, A., Peroni, S., Tomasi, F., & **Heibi, I.** (2023). *Representing provenance and track changes of cultural heritage metadata in RDF: A survey of existing approaches*. DOI: https://doi.org/10.48550/ARXIV.2305.08477

Memon, A. P. (2001). Study of Unicode specifications and their implementation in Arabic script languages by designing a multilingual Unicode editor. *Proceedings. IEEE International Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century*, pp. 229–233. DOI: https://doi.org/10.1109/INMIC.2001.995342

Moretti, A., & **Heibi, I.** (2023). *Metadata crosswalk for citation data production in OpenCitations*. In: Social Sciences & Humanities Open Marketplace. https://marketplace.sshopencloud.eu/workflow/MHwO4l

Moretti, A., Heibi, I., Soricetti, M., Rizzetto, E., Massari, A., & **Peroni, S.** (2024). *OpenCitations Data Sources Converter* [Computer software]. https://archive.softwareheritage.org/swh:1:snp:f2bcfd68c9681a284c78fb64eb2fb0ac360ee566;origin=https://github.com/opencitations/oc_ds_converter

OpenCitations. (2023a). OpenCitations Index CSV dataset of all the citation data. figshare. *Dataset*. DOI: https://doi.org/10.6084/m9.figshare.24356626.v2

OpenCitations. (2023b). OpenCitations Index N-Triples dataset storing data source information about all the citation data. figshare. *Dataset*. DOI: https://doi.org/10.6084/m9.figshare.24427051.v2

OpenCitations. (2023c). OpenCitations Index Scholix dataset of all the citation data. figshare. *Dataset*. DOI: https://doi.org/10.6084/m9.figshare.24416749.v2

OpenCitations. (2023d). OpenCitations Meta CSV dataset of all bibliographic metadata. figshare. *Dataset*. DOI: https://doi.org/10.6084/m9.figshare.21747461.v6

Peels, R., & **Bouter, L.** (2018). The possibility and desirability of replication in the humanities. *Palgrave Communications*, 4(1), Article 1. DOI: https://doi.org/10.1057/s41599-018-0149-x

Peroni, S., & **Shotton, D.** (2019). *Open Citation Identifier: Definition*. DOI: https://doi.org/10.6084/m9.figshare.7127816.v2

Peroni, S., & **Shotton, D.** (2020). OpenCitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, *1*(1), pp. 428–444. DOI: https://doi.org/10.1162/qss_a_00023

Peroni, S., & **Shotton, D.** (2022a). *OpenCitations Mission Statement*. DOI: https://doi.org/10.5281/zenodo.6976670

Peroni, S., & **Shotton, D.** (2022b). *The Uniqueness of OpenCitations*. DOI: https://doi.org/10.5281/zenodo.6976696

Riva, P. (2022). The multilingual challenge in bibliographic description and access. *JLIS.It*, *13*(1), Article 1. DOI: https://doi.org/10.4403/jlis.it-12737

Shotton, D. (2013). Publishing: Open citations. *Nature*, *502*(7471), Article 7471. DOI: https://doi.org/10.1038/502295a

Tilley, S. (2004). Test-driven development and software maintenance. *20th IEEE International Conference on Software Maintenance, 2004. Proceedings*, pp. 488–491. DOI: https://doi.org/10.1109/ICSM.2004.1357840

UNESCO. (2021). UNESCO Recommendation on Open Science (Programme and Meeting Document SC-PCB-SPP/2021/OS/UROS; p. 36). https://unesdoc.unesco.org/ark:/48223/pf0000379949

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Yergeau, F. (2003). *UTF-8, a transformation format of ISO 10646* (RFC3629; p. RFC3629). RFC Editor. DOI: https://doi.org/10.17487/rfc3629

斉史加藤, 江里土屋, 壮一久保田, & **謹至宮川**. (2012). ジャパンリンクセンターによるリンク管理と日本語の電子的学術コンテンツへのDOI付与. 情報管理, 55(1), pp. 42–46. DOI: https://doi.org/10.1241/johokanri.55.42

TO CITE THIS ARTICLE:
Moretti, A., Soricetti, M., Heibi, I., Massari, A., Peroni, S., & Rizzetto, E. (2024). The Integration of the Japan Link Center's Bibliographic Data into OpenCitations. *Journal of Open Humanities Data*, 10: 21, pp. 1–16. DOI: https://doi.org/10.5334/johd.178

Submitted: 28 October 2023
Accepted: 30 January 2024
Published: 29 February 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.