ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

The OpenCitations Data Model

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Daquino, M., Peroni, S., Shotton, D., Colavizza, G., Ghavimi, B., Lauscher, A., et al. (2020). The OpenCitations Data Model. Cham : Springer [10.1007/978-3-030-62466-8_28].

Availability:
This version is available at: https://hdl.handle.net/11585/778712 since: 2021-02-25

Published:

DOI: http://doi.org/10.1007/978-3-030-62466-8_28

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

07 July 2026

This is the final peer-reviewed accepted manuscript of:

Daquino Marilena, Peroni Silvio, Shotto David, Colavizza Giovanni, Ghavimi Behnam, Lauscher Anne, Mayr Philipp, Romanello Matteo, Zumstein Philipp (2020) The OpenCitations Data Model, in The Semantic Web – ISWC 2020. Lecture Notes in Computer Science, vol. 12507. Springer, pp. 447-463.

The final published version is available online at:

https://dx.doi.org/10.1007/978-3-030-62466-8_28

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

arXiv:2005.11981v2 [cs.DL] 24 Aug 2020

The OpenCitations Data Model

Marilena Daquino[1,2][0000−0002−1113−7550], Silvio Peroni[1,2][0000−0003−0530−4305], David Shotton[2,3][0000−0001−5506−523X], Giovanni Colavizza[4][0000−0002−9806−084X], Behnam Ghavimi[5][0000−0002−4627−5371], Anne Lauscher[6][0000−0001−8590−9827], Philipp Mayr[5][0000−0002−6656−1658], Matteo Romanello[7][0000−0002−7406−6286], and Philipp Zumstein[8][0000−0002−6485−9434]⋆

[1] Digital Humanities Advanced research Centre (/DH.arc), Department of Classical Philology and Italian Studies, University of Bologna
{marilena.daquino2,silvio.peroni}@unibo.it
[2] Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna
[3] Oxford e-Research Centre, University of Oxford
david.shotton@opencitations.net
[4] Institute for Logic, Language and Computation (ILLC), University of Amsterdam
g.colavizza@uva.nl
[5] Department of Knowledge Technologies for the Social Sciences, GESIS - Leibniz-Institute for the Social Sciences
ghavimi.behnam@gmail.com, philipp.mayr@gesis.org
[6] Data and Web Science Group, University of Mannheim
anne@informatik.uni-mannheim.de
[7] cole Polytechnique Fdrale de Lausanne
matteo.romanello@epfl.ch
[8] Mannheim University Library, University of Mannheim
philipp.zumstein@bib.uni-mannheim.de

Abstract. A variety of schemas and ontologies are currently used for the machine-readable description of bibliographic entities and citations. This diversity, and the reuse of the same ontology terms with different nuances, generates inconsistencies in data. Adoption of a single data model would facilitate data integration tasks regardless of the data supplier or context application. In this paper we present the OpenCitations Data Model (OCDM), a generic data model for describing bibliographic entities and citations, developed using Semantic Web technologies. We also evaluate the effective reusability of OCDM according to ontology evaluation practices, mention existing users of OCDM, and discuss the use and impact of OCDM in the wider open science community.

Keywords: Open citations · Scholarly data · Data model

⋆ Authors' contributions specified according to the CrediT taxonomy. MD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing original draft, Writing review & editing. She is responsible for section "Background" and section "Analysis of OCDM reusability". SP and DS: Conceptualization, Investigation, Methodology, Software, Data curation, Supervision, Funding acquisition, Project administration, Writing original draft. GC, BG, AL, PM, MR and PZ: Investigation, Resources, Validation, Writing review & editing.

1 Introduction

In recent years, largely thanks to the Initiative for Open Citations (I4OC)[9], most major scholarly publishers have made their bibliographic reference data open, resulting, for example, in more than 700 million citations now being made openly available in the OpenCitations Index of Crossref open DOI-to-DOI citations (COCI) [17]. As a consequence, scholarly data providers and bibliometric analysis software have started to integrate open citation data into their services, thereby offering an alternative to the current reliance on proprietary citation indexes.

Open bibliographic and citation metadata are beneficial because they enable anyone to perform meta-research studies on the evolution of scholarly knowledge, and allows national and international research assessment exercises characterized by transparent and reproducible processes. Within this context, bibliographic citations are essential components of scholarly discourse, since they remain the dominant measurable unit of credit in science [12]. They carry evidence of scholarly networks and of the progress of theories and methods, and are fundamental aids in tenure evaluation and recommendation systems. To perform open bibliometric research and analysis, the publications upon which the work is based should be FAIR, namely Findable, Accessible, Interoperable, and Reusable [35]. Ideally, such data should be made available without any restrictions, licensed under a Creative Commons CC0 waiver[10], and the software for programmatically accessing and analysing them should be also released with open source licences.

However, data suppliers use a variety of licenses, technologies, and vocabularies for representing the same bibliographic information, or use ontology terms defined in the same ontologies with different nuances, thereby generating diversity in data representation. The adoption of a common, generic, open and documented data model that employs clearly defined ontological terms would ensure data consistency and facilitate integration tasks.

In this paper we present the OpenCitations Data Model (OCDM), a data model based on existing ontologies for describing information in the scholarly bibliographic domain with a particular focus on citations. OCDM has been developed by OpenCitations [29], an infrastructure organization for open scholarship dedicated to the publication of open bibliographic and citation data using Semantic Web technologies. Herein, we propose a holistic approach for evaluating the reusability of OCDM according to ontology evaluation methodologies, and we discuss its uptake, impact, and trustworthiness.

We compared OCDM to similar existing solutions and found that, to the best of our knowledge, OCDM (a) has the broadest vocabulary coverage, (b) is the best documented data model in this area, and (c) has already a significant uptake in the scholarly community. The main advantages of OCDM, in addition to the consistency of data description that it facilitates, are that it was designed from the outset to enable use by those who are not Semantic Web practitioners, as well as by those that are, that it is properly documented, and it is provided

[9] https://i4oc.org

[10] https://creativecommons.org/publicdomain/zero/1.0/legalcode

with accompanying software for managing the entire life-cycle of data created according to OCDM.

The paper is organized as follows. In Section 2 we clarify the scope and motivations for this work. In Section 3 we present the data model and its documentation, software and current early adopters. In Section 4 we present the criteria we have used to evaluate OCDM reusability and we present results, including figures about OCDM views, downloads and citations according to Figshare and Altmetrics, which are further discussed in Section 5.

2 Background

The OpenCitations Data Model (OCDM) [9] was initially developed in 2016 to describe the data in the OpenCitations Corpus (OCC). In recent years OpenCitations has developed other datasets while OCDM has been adopted by external projects, and OCDM has been expanded to accommodate these changes. We have recently further expanded the OpenCitations Data Model to accommodate the extended metadata requirements of the Open Biomedical Citations in Context Corpus project (CCC). This project has developed an exemplar Linked Open Dataset that includes detailed information on citations, in-text reference pointers such as Berners-Lee et al. 2011, and identifiers of the citation contexts (e.g. sentences, paragraphs, sections) within which in-text reference pointers are located, to facilitate textual analysis of citation contexts. The citations are treated as first-class data entities [26], enriched with open bibliographic metadata released using a CC0 waiver that can be mined, stored and republished. This includes identifiers specifying the specific positions of the various in-text reference pointers within the text. However, the literal text of these contexts are not stored within the Open Biomedical Citations in Context Corpus, and regrettably in many cases the full text of the published entities cannot be mined from elsewhere in an open way, even for some (view only) Open Access articles, because of copyright, licensing and other Intellectual Property (IP) restrictions.

Table 1 shows the representational requirements (hereinafter, for the sake of simplicity, also called citation properties and numbered (P1-P8)) that we were interested in recording for each citation instantiated from within a single paper.

3 The OpenCitations Data Model

The OCDM permits one to record metadata about bibliographic references and their textual contexts, bibliographic entities (citing and cited publications) and the citations that link them, agents and their roles (e.g. author, editor), identifiers for the foregoing entities, provenance metadata and much more, as shown diagrammatically in Fig. 1. All terms described in the OCDM are brought together in the OpenCitations Ontology (OCO)[11]. OCO aggregates terms from

[11] https://w3id.org/oc/ontology

ID	Description
P1	A classification of the type of citation (e.g. self-citation).
P2	The bibliographic metadata of the citing and cited bibliographic entities (e.g. type of published entity, identifiers, authors, contributors, publication date, publication venues, publication formats).
P3	The bibliographic reference, typically found within the reference list of the citing bibliographic entity, that references a cited bibliographic entity.
P4	The separate identifiers of all the in-text reference pointers included in the text of the citing entity, that denote bibliographic references within the reference list.
P5	The co-occurrence of in-text reference pointers within each in-text reference pointer lists (e.g. [3,5,12]).
P6	The identifiers of structural elements (e.g. XPath of sentences, paragraphs, captions) that specify where, in the full text, an in-text reference pointer appears.
P7	The function or purpose of the citation (e.g. to cite as background, extend, or agree with the cited entity) to which each in-text reference pointer relates.
P8	Provenance information of the citation extraction process (e.g. responsible agents, data sources, extraction dates).

Table 1. Representational requirements of the OpenCitations Data Model

the SPAR (Semantic Publishing and Referencing) Ontologies [28] and other well-known ontologies, such as PROV-O [4] and Web Annotation Ontology [32].

Citations are instances of the class `cito:Citation` defined in CiTO, the Citation Typing Ontology[12]. Subclasses (not shown in Fig. 1), relevant for P1, include `cito:AuthorSelfCitation`, `cito:JournalSelfCitation`, `cito:FunderSelfCitation`, `cito:AffiliationSelfCitation`, and `cito:AuthorNetworkSelfCitation`. In addition, citations can be characterized with a purpose or function with respect to the related citation context, by means of the property `cito:hasCitationCharacterisation` and the use of one or more CiTO properties (e.g. `cito:usesMethodIn`) (P7).

Instances of the class `fabio:Expression`, defined in the FRBR-aligned Bibliographic Ontology (FaBiO)[13], can be linked to bibliographic metadata such as publication dates, authors, and venues. Instances of `fabio:Manifestation` aggregate information on specific editions and formats (P2).

Instances of `oa:Annotation`, defined in the Web Annotation Ontology (OA)[14], link instances of the class `cito:Citation` to instances of `biro:BibliographicReference` (P3), defined in BiRO, the Bibliographic Reference Ontology[15], and individuals of `c4o:InTextReferencePointer` (P4), defined in C4O, the Citation Counting and Context Characterisation Ontology[16]. Lists of in-text reference pointers are represented by the class `c4o:SingleLocationPointerList` (P5).

Structural elements wherein in-text reference pointers appear are represented as individuals of `deo:DiscourseElement`, defined in DEO, the Discourse Ele-

[12] http://purl.org/spar/cito
[13] http://purl.org/spar/fabio
[14] https://www.w3.org/ns/oa
[15] http://purl.org/spar/biro
[16] http://purl.org/spar/c4o

ment Ontology[17]. Elements are uniquely identified (P6) by means of instances of `datacite:Identifier`, defined in the DataCite Ontology[18].

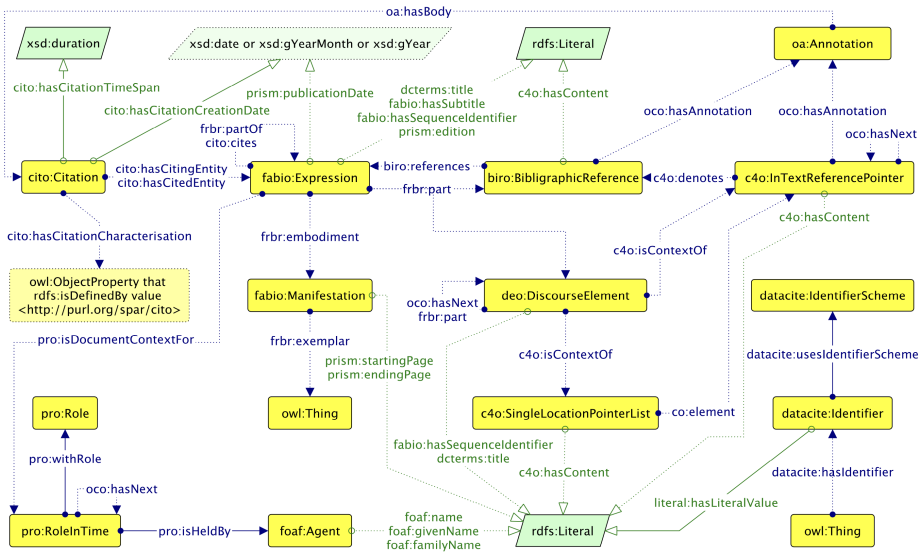

Fig. 1. Main classes and properties of the OpenCitations Ontology

Finally, as summarized in Figure 2, OCDM provides guidance for describing the provenance and versioning of each entity under consideration, and also enables the specification of the main metadata related to the datasets containing such entities (P8). To this end, the OCDM reuses terms from PROV-O, the Provenance Ontology[19], VoID, the Vocabulary of Interlinked Datasets[20] [2], and DCAT, the Data Catalog Vocabulary[21] [24].

Each bibliographic entity described by the OCDM is annotated with one or more provenance snapshots (i.e. instances of `prov:Entity`, each snapshot intended as a specialisation of the bibliographic entity via `prov:specializationOf`) as defined in [30]. In particular, each snapshot records the set of statements having the bibliographic entity as its subject at a fixed point in time, validity dates, responsible agents for either the creation or the modification of the metadata, primary data sources, and a SPARQL query summarising changes with respect to any prior snapshot.

Lastly, a dataset (`dcat:Dataset`) containing information about the bibliographic entities is described with cataloguing information (e.g. title, description, publication and change dates, subjects, webpage, SPARQL endpoint) and distri-

[17] http://purl.org/spar/deo
[18] http://purl.org/spar/datacite
[19] http://www.w3.org/ns/prov
[20] http://rdfs.org/ns/void
[21] http://www.w3.org/ns/dcat

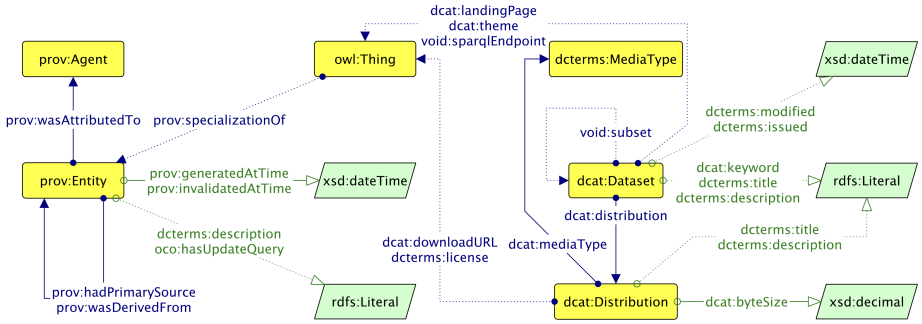

Fig. 2. Provenance, versioning, and dataset description in the OCDM

bution information (`dcat:Distribution`) which also includes the specification of licenses, dumps, media types, and data volumes.

3.1 OCDM documentation and resources

In order to make the OCDM understandable and reusable by both the Semantic Web community and communities with no expertise in Semantic Web technologies, support material has been produced. All materials are available at `http://opencitations.net/model` and include the following resources.

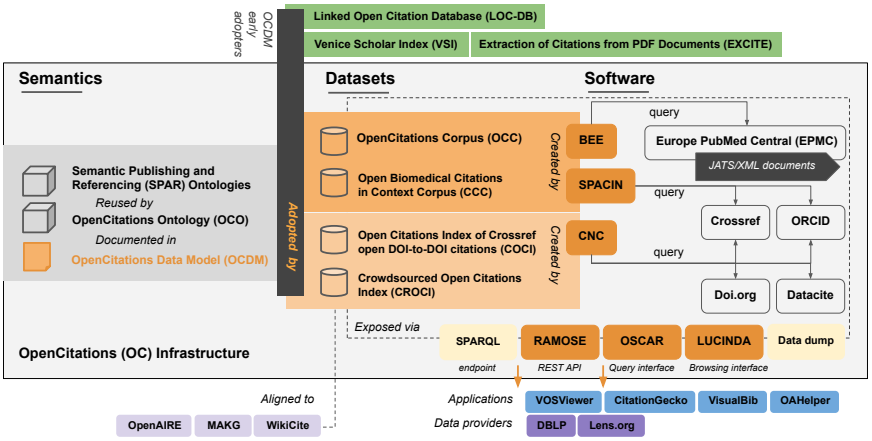

Fig. 3. Overview of OpenCitations ecosystem and acronyms used in this paper

Human-readable documentation. The OCDM documentation [9] provides (1) detailed definitions of terms characterising open citation data and open bibliographic metadata, (2) naming conventions and URI patterns, and (3) real-world examples. OCDM is supplemented by two additional specifications, i.e.

the definition of the Open Citation Identifier (OCI) [26] and the definition of the In-Text Reference Pointer Identifier (InTRePID) [33].

OCDM-compliant data examples. All the data introduced in the OCDM documentation are expressed and provided in JSON-LD to make it easily understandable both to RDF experts and other Web users. In addition, CSV templates have been adopted so as to express and share parts of the OCDM e.g. to store the citation data in COCI [17].

Ontology development documentation. The first version of the OCDM, released in 2016, addressed citation properties P1-P3 and P8, by directly reusing the SPAR Ontologies and other vocabularies [28]. Within the context of the CCC project described above, we used SAMOD [27], an agile data-driven methodology for ontology development, to extend OCO with terms relevant to P4-P7. Motivating scenarios, competency questions, and a glossary of terms of all the new entities included in the OCDM, are available for reproducibility purposes.

Open source software leveraging the data model. The source code of the knowledge extraction and data re-engineering pipeline for managing data according to OCDM is available at `http://opencitations.net/tools`. The pipeline includes software originally developed for creating the OpenCitations Corpus (BEE and SPACIN) and the OpenCitations Indexes (Create New Citations CNC), and a user-friendly web application (BCite)[10] for creating OCDM-compliant RDF data from lists of bibliographic references. In addition, we have released tools to support the development of applications leveraging data organized according to OCDM: RAMOSE (to create RESTful APIs over SPARQL endpoints), OSCAR (to create user-friendly search interfaces for querying SPARQL endpoints [16]) and LUCINDA (a configurable browser for RDF data). Configuration files for setting up these tools are available in their GitHub repositories.

Licenses for reuse. OCDM (both the documentation and OCO) is released under a CC-BY license. Software solutions are released under the ISC license. The OCDM-compliant data served by OpenCitations are made open under CC0.

3.2 OCDM early adopters

To date, OCDM is central to the work of OpenCitations. The OpenCitations datasets modelled using OCDM include: the OpenCitations Corpus (OCC), including about 13 million citation links and the OpenCitations Indexes, which include more than 721 million citations. Forthcoming datasets, that will be released later in 2020, include OpenCitations Meta, which stores metadata of the citing and cited entities involved in the citations included in the Indexes, and the Open Biomedical Citations in Context Corpus (CCC), mainly derived from the Open Access corpus of biomedical articles provided by PubMed Central, that will include detailed information on in-text reference pointers denoting each reference in the reference list, and their textual contexts.

Moreover, OCDM has three external acknowledged early adopters. The Extraction of Citations from PDF Documents (EXCITE) project [20] is run by GESIS and the University of Koblenz. The aim of EXCITE is to extract and

match citations from social science publications. To date, EXCITE has extracted around 1 million citations, has converted the data to RDF according to OCDM, and has then published it by ingestion into the OCC.

The Linked Open Citation Database (LOC-DB) [21] is a project which aims to demonstrate that it is possible for academic libraries to catalogue citation relations sustainably, accurately, and cooperatively. So far, the project has stored bibliographic and citation data for about 7000 published entities. LOC-DB has used a customisation of the OCDM as the data model for defining its data, and exports data in OCDM/JSON-LD so as to be ingested into the OCC.

The Venice Scholar Index (VSI)[22] is an instance of the Scholar Index, originated from the Linked Books project [8] founded by the Swiss National Science Foundation. The citation index includes about 4 million references to publications cited in the historiography of Venice. VSI exports data into RDF formats according to OCDM so as to be integrated into the OCC.

4 Analysis of OCDM reusability

A holistic approach has been used to evaluate the OCO ontology and to infer properties relevant to OCDM. We adopted seminal definitions and classifications of ontology evaluation approaches [6,14] and we selected the following dimensions and approaches that are representative with respect to OCDM reusability.

[E1] Lexical keyword similarity. This addresses the similarity of definitions (labels of terms) in OCO with respect to the real-world knowledge to be mapped. We adopted a data-driven evaluation [7] to map OCO definitions with terms included in a corpus of documents encoded in the Journal Article Tag Suite (JATS) XML schema[23]. JATS is used by Europe PubMed Central (EPMC)[24] to encode scholarly documents, that are in turn harvested by OpenCitations.

[E2] Vocabulary coverage. This addresses the coverage of concepts, instances, and facts of OCO with respect to the domain to be covered. **[E2.1]** We validated OCO coverage by comparing it with competing ontologies [25]. **[E2.2]** Secondly, we adopted an application-based approach [31] to address OCO coverage in four sources that leverage it: OpenCitations, EXCITE, LOC-DB, and ScholarIndex.

Also, we addressed aspects peculiar to OCDM reusability, namely:

[E3] Usability-profiling. This encompasses the communication context of OCDM, i.e. its pragmatics. We evaluated OCDM recognition level [13], i.e. the efficiency of access to OCDM ontologies, documentation, and software, by comparing it with competing ontologies [25].

[22] https://scholarindex.eu/
[23] https://jats.nlm.nih.gov/
[24] https://europepmc.org/downloads/openaccess

Lastly, we addressed current uptake, potential impact, and trustworthiness of OCDM, including metrics about OCDM views, downloads and citations according to Figshare and Altmetrics.[25]

4.1 E1: Lexical keyword similarity

We created a randomized corpus of 2800 JATS documents taken from the Open Access Subset of biomedical literature hosted by Europe PubMed Central. We extracted the list of XML elements used in the documents within this corpus (117 elements), and we expanded element names with definitions scraped from the online XML schema guidelines (e.g. `<p>` became Paragraph). We manually pruned non-relevant elements such as MathML markup, text style elements (e.g. `<italic>`), redundant wrapping elements (`<keywordGroup>`) and elements that are out of scope (e.g. `<biography>`), resulting in a refined list of 45 terms.

Secondly, we extracted definitions from OCO (118). We manually pruned terms that were not relevant (e.g. annotation properties, provenance, and distribution related terms), terms that represent hierarchy, sequences, and linguistic aspects not available in XML (e.g. partOf, hasNext, Sentence), and terms dependent on post-processing activities (e.g. self-citation, hasCitationCharacterisation), resulting in a refined list of 77 OCO definitions.

We then used Wordnet[26] to automatically expand both XML and ontology definitions with synonyms, and we matched synsets similarities. We used a symmetric similarity score to find best matches between the synsets. We considered two thresholds for the similarity match, 0.7 and 0.5, and we manually computed precision and recall. Table 2 shows the results.

The coverage of JATS terms in OCO was 55.5% when the threshold was greater than 0.7, with high precision (96%) and average recall (53.3%). The coverage was 73.3% when the threshold was greater than 0.5, with still high precision (93.3%) and average recall (68.8%). False negative results included acronyms (e.g. issn) that did not have a match in Wordnet, and terms of the taxonomy that were underrepresented in the corpus (e.g. book). Likewise, false positive results were due to acronyms used in XML definitions that were not correctly parsed (e.g. URI for This Same Article Online was incorrectly matched with fabio:JournalArticle).

Threshold	Matches	Precision	Recall
0.7	(25/45) 55.5%	(24/25) 96%	(24/45) 53.3%
0.5	(33/45) 73.3%	(31/33) 93.9%	(31/45) 68.8%

Table 2. Lexical similarity between JATS/XML elements and OCO terms

[25] Source code and results of this analysis are available at https://github.com/opencitations/metadata

[26] https://wordnet.princeton.edu/

4.2 E2: Vocabulary coverage

[E2.1] Vocabulary coverage in existing vocabularies. Since gold standard ontologies are not available, we referred to existing data models and relevant ontologies used by citation data providers. For the sake of completeness, we addressed both open and non-open citation data providers[27], and both graph data providers and others. We reviewed the vocabulary coverage with respect to P1-P8. We did not take into account discipline coverage or citation counting. The complete list of data models and references is available at https://github.com/opencitations/metadata. Table 3 summarizes the comparison of vocabularies coverage, an x indicating that the source had metadata of relevance to the citations properties P1-P8 (Table 1).

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
Google Scholar		x						x
Scopus		x						x
Web of Science	x	x	x					x
CiteseerX	x	x	x			x		x
Dimensions		x	x					x
Crossref		x	x					x
EPMC		x	x					x
Datacite		x	x				x	x
DBLP		x						x
MAKG		x	x			x		
ORC		x						x
GORC		x	x	x	x	x		x
SciGraph		x						x
WikiCite		x						x
OpenCitations	x	x	x	x	x	x	x	x

Table 3. Vocabulary coverage in existing vocabularies according to P1-P8

Non-open citation data providers include Google Scholar, Scopus [1], Web of Science (WoS) [5], CiteSeerX [22] and Dimensions [19]. Their data models cover a few aspects of bibliographic metadata (P2) and provenance data (P8). WoS, CiteSeerX, and Dimension also includes bibliographic references (P3). In addition, Wos and CiteSeerX also cover types of citations (P1), and only CiteSeerX includes citation context sentences (P6).

Open citation data providers include Crossref [18], Europe PubMed Central (EPMC), DataCite, DBLP, Microsoft Academic Knowledge Graph (MAKG) [11] (which is based on Microsoft Academic Graph [34] and which reuses the SPAR Ontologies and links to resources in Wikidata and OpenCitations), the Semantic Scholar Open Research Corpus (ORC) [3], the Semantic Scholars Graph of References in Context (GORC) [23], Springer Natures SciGraph [15] (which is based on Schema.org), WikiCite (which includes terms aligned to SPAR Ontologies and interlinks with the OpenCitations Corpus), and the OpenCitations datasets [29]. All data models cover P2, and all except MAKG also cover P8. Only OpenCitations covers P1. In addition, Crossref, Europe PMC, DataCite, MAKG, GORC, and OpenCitations cover P3. MAKG, GORC, and OpenCitations cover P6, while the latter two also includes in-text reference pointers (P4)

[27] See the definition of "open" at https://opendefinition.org/licenses/.

and related lists (P5). DataCite and OpenCitations allow the tracking of citation functions (P7).

[E2.2] Vocabulary coverage in early adopters. We separately analysed the vocabulary coverage in acknowledged adopters of OCDM (Table 4).

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
EXCITE		x	x					x
LOC-DB		x	x					x
VSI		x	x	x		x		x

Table 4. Vocabulary coverage in OCDM early adopters according to P1-P8

EXCITE data fully covers P2, P3 and P8. Its local data model also includes information about the data quality of extracted references, which is not currently mapped to OCDM. LOC-DB data fully covers P2, P3, and P8. The OCDM was extended in its local data model so as to cover information about its OCR activities performed on PDF scans. Venice Scholar Index (VSI) aligned data to OCDM terms so as to fully cover P2, P3, P4, P6, and P8. In order to cover peculiar needs of the project relevant to P2, the classes `fabio:Work` and `fabio:Expression` defined in the SPAR Ontologies (and reused in OCO) were specialized so as to include the following sub-classes: `fabio:ArchivalRecord`, `fabio:ArchivalRecordSet`, `fabio:ArchivalDocument`, and `fabio:ArchivalDocumentSet`[28].

4.3 E3: Usability profiling

We compared the documentation available for existing graph data providers, namely: MAKG, OC and GORC (Semantic Scholar), SciGraph, and WikiCite. We considered the same dimensions used to address OCDM documentation, namely: human-readable documentation, machine-readable data model and examples, ontology development documentation, open source software leveraging the model, and licenses for reuse (see Table 5).

	HR docum.	**MR data model**	**ontology dev. docum.**	**software**	**licenses**
MAKG	x				
ORC and GORC	x	x			
SciGraph	x	x			x
WikiCite	x			x	x
OCDM	x	x	x	x	x

Table 5. Usability of existing ontologies and data models

The MAKG data model is graphically represented in [11]. Software for creating RDF data is available, but no machine-readable data model and examples are provided. Likewise, the development of the data model is not described. Moreover, according to Frber [11], the property `c4o:hasContext` is used to annotate instances of `cito:Citation`, rather than `c4o:InTextReferencePointer` as prescribed in C4O, preventing it from representing consistently P3, P4, and P7 in

[28] As documented at https://github.com/SPAROntologies/fabio/issues/1.

future works, and from merging third-party data with OpenCitations. Lastly, no license is specified for the data model.

The Semantic Scholar Open Research Corpus data model is described in [3]. A machine-readable example of the data model is presented in a dedicated web page[29]. No further documentation is available. Similarly, GORC is described in [23], where an example of JSON data is presented. Both datasets are released under OCD-BY (i.e. an open license), although programmatically accessing data through their APIs requires one to subscribe to a more restrictive and non-open license (comparable to CC-BY-NC-ND). No license associated with the data model is stated.

The Schema.org main classes reused in SciGraph are described in a dedicated web page[30]. While the ontology is reused as-is, the SciGraph data model[31] is released as a JSON-LD file and machine-readable examples are available under a CC-BY license. Development documentation of the data model is not available.

Sources addressing the Wikidata model used by WikiCite include templates[32] and examples[33]. However, no dedicated documentation nor a machine-readable version of the model having citations as a scope is separately available. Data, software, and the general data model are all released under the CC0 license.

Lastly, OCDM [9] is described in dedicated human-readable documentation, including machine-readable data model and examples, available under a CC-BY license. The ontology development documentation and the open source software leveraging the model are available on github (ISC licence). All materials are gathered in the official page of the OCDM data model[34].

4.4 OCDM uptake, potential impact, and trustworthiness

We can quantify current uptake of the OCDM documentation by using statistics provided by Figshare and Altmetrics, and the number of users views of the model description page in the OpenCitations website. As of 18 August 2020, the Figshare document [9] has been viewed 10852 times, downloaded 1508 times, and cited 5 times. 100 tweets from 65 users include links to the document. The web page (`http://opencitations.net/model`) dedicated to the model has received 13,844 views from 8,202 unique users since 2018.

We can estimate the potential impact of OCDM by considering (a) different types of possible reuse of the model, (b) the number of current reusers of the data model, (c) projects and applications leveraging data created according to OCDM, and (d) the kind of users of data created according to OCDM.

In detail, OCDM can be reused as is, via alignment for interchange purposes, and as a JSON data model for non-Semantic Web users. Currently OCDM

[29] `http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/`
[30] `https://scigraph.springernature.com/explorer/datasets/ontology/`
[31] `https://github.com/springernature/scigraph`
[32] `https://www.wikidata.org/wiki/Template:Bibliographic_properties`
[33] `https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_MetaData`
[34] `http://opencitations.net/model`

is used by OpenCitations for all its datasets, and by the three acknowledged early adopters, namely: EXCITE and LOC-DB, which reuse OCDM as is, and VSI, which aligned terms to OCDM. EXCITE data have been ingested in the OpenCitations Corpus, while LOC-DB and VSI data are going to be ingested soon. VOSViewer[35], CitationGecko[36], VisualBib[37], and OAHelper[38] are applications that leverage OpenCitations data conforming to OCDM retrieved via the OpenCitations REST APIs or directly through its SPARQL endpoints. Moreover, OpenAIRE[39], MAKG, and WikiCite align data to OpenCitations. Both DBLP and Lens.org[40] use citation data from OpenCitations to enrich their bibliographic metadata records.

Users of OpenCitations data include scholars in scientometrics, life sciences, biomedicine, the physical sciences, and the information technology domain. OpenCitations is currently expanding its coverage to include the social science and the arts and humanities disciplines. The main users of EXCITE data are researchers in the social sciences, while those of the data held by LOC-DB and the Venice Scholar Index include librarians and researchers in the humanities.

Lastly, we address trustworthiness of OCDM. Long-term availability of ontologies is crucial for the development of the Semantic Web, and the trustworthiness of the ontology creators is important. OCDM, OCO, and the SPAR Ontologies are all maintained by OpenCitations, which has been recently selected by the Global Sustainability Coalition for Open Science Services (SCOSS)[41] as an open infrastructure deserving of crowdfunding support from the scholarly community, thereby helping to ensure its long-term sustainability.

Along with trustworthiness, another important factor is the general interest in the community towards research topics and outputs that can leverage OCDM. So far, two OpenCitations projects dedicated to the enhancement of the OpenCitations Corpus and the creation of the Open Biomedical Citations in Context Corpus have been funded by the Alfred P. Sloan Foundation[42] and the Wellcome Trust respectively, as mentioned above in Section Background. Moreover, the Internet Archive and Figshare have both offered to archive backup copies of the OpenCitations datasets without charge.

5 Discussion and conclusions

First, we evaluated lexical similarity of OCO definitions over the knowledge included in data sources encoded in JATS/XML, a gold standard for academic publications [E1]. While the recall is only average, mainly due to mistakes in

[35] https://www.vosviewer.com/
[36] https://citationgecko.com/
[37] https://visualbib.uniud.it/en/project/
[38] https://www.otzberg.net/oahelper/
[39] https://www.openaire.eu/
[40] https://lens.org
[41] https://scoss.org/
[42] See https://sloan.org/grant-detail/8017

parsing of acronyms, for those terms that were correctly matched the lexical similarity precision is high, showing that OCO is appropriate for representing data sources organized according to the gold standard. One of the known limits of data-driven evaluation methodologies is that these do not address possible changes in the domain knowledge over time. To date, early adopters of OCO continuously contribute with new scenarios to be represented in the model, which is correspondingly expanded. As a result, OCO will remain a comprehensive reference point for future developments. Other statistical semantic approaches will be evaluated in the future.

Secondly, we evaluated OCO vocabulary coverage as compared with competing data models [E2.1] and in the context of early adopters [E2.2]. Only OCDM fully covers P1-P8. In particular, only one other provider covers P4 and P5 (identifiers for in-text references and groups of these), three providers cover property P6 (although they only store full-text sentences, and lack identifiers for in-text reference pointers), and only one other provider covers property P7 (citation function). Two graph-data providers reuse terms from SPAR Ontologies (either directly or by alignment) in different ways, generating heterogeneity in data.

Among early adopters, LOC-DB required extensions in order to represent special information related to the cataloguing of digital objects, and VSI required us to expand the FaBiO ontology to permit description of unpublished archival entities. While such changes can be deemed marginal, these are relevant hints for future developments in the humanities domain and will require further analysis. Nonetheless, the OCDM vocabulary coverage is satisfying and strengthens its reusability across domains and applications.

We showed how alternative citation data providers ensure access to their data models [E3]. Peer-reviewed articles are the main access point to descriptions of those data models, with additional information scattered across various web pages. While machine-readable data models and examples are mostly available, none of the other providers referenced detailed development documentation. Moreover, the licenses for reusing the data models are not always defined. In summary, OCDM appears to be the most documented and findable data model.

Again, no comparison was possible of the uptake of the alternative models in the community. We showed that OCDM has been relatively popular in community social networks, and that the documentation has been downloaded and read by many people. At the moment we cannot measure for what purpose the OCDM documentation has been reused, with the exception of the three early-adopter projects of which we are aware listed in this paper.

We have shown that OCDM is potentially of significant usefulness to several communities, and fosters reuse in combination with legacy technologies, and we have highlighted ongoing interest from several parties in the maintenance and ongoing development of OCDM in support of several projects.

In future works, we will (a) create SHeX shapes to facilitate reusers in mapping their data to OCDM, and (b) trace OCDM usage scenarios by asking users to fill in a form for statistical purposes.

Acknowledgements

This work was funded by the Wellcome Trust (Wellcome-214471_Z_18_Z). We thank Ludo Waltman (Centre for Science and Technology Studies - CWTS, Leiden University) and Vincent Larivire (cole de bibliothconomie et des sciences de l'information, lUniversit de Montral) for supervising aspects of this work, and Ivan Heibi (University of Bologna) for contributing with suggestions.

References

1. Aagaard, K., Kladakis, A., Nielsen, M.W.: Concentration or dispersal of research funding? Quantitative Science Studies **1**(1), 117–149 (2020)
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Bizer, C., Heath, T., Berners-Lee, T., Idehen, K. (eds.) Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009. CEUR Workshop Proceedings, vol. 538. CEUR-WS.org (2009), http://ceur-ws.org/Vol-538/ldow2009_paper20.pdf
3. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al.: Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262 (2018)
4. Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C Recommendation (2013)
5. Birkle, C., Pendlebury, D.A., Schnell, J., Adams, J.: Web of science as a data source for research on scientific and scholarly activity. Quantitative Science Studies **1**(1), 363–376 (2020)
6. Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: Proceedings of the conference on data mining and data warehouses (SiKDD 2005). pp. 166–170. Ljubljana, Slovenia (2005)
7. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: 4th International Conference on Language Resources and Evaluation, LREC04 (2004)
8. Colavizza, G., Romanello, M.: Citation mining of humanities journals: The progress to date and the challenges ahead. Journal of European Periodical Studies **4**(1), 36–53 (2019)
9. Daquino, M., Peroni, S., Shotton, D.: The OpenCitations data model. Figshare (2018). https://doi.org/10.6084/m9.figshare.3443876.v7
10. Demidova, E., Zaveri, A., Simperl, E.: Creating open citation data with bcite. Emerging Topics in Semantic Technologies: ISWC 2018 Satellite Events **36**, 83 (2018)
11. Färber, M.: The microsoft academic knowledge graph: A linked data source with 8 billion triples of scholarly data. In: International Semantic Web Conference. pp. 113–129. Springer (2019)
12. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379), eaao0185 (2018)
13. Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In: European Semantic Web Conference. pp. 140–154. Springer (2006)

14. Gómez-Pérez, A.: Ontology evaluation. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 251–274. International Handbooks on Information Systems, Springer (2004)
15. Hammond, T., Pasin, M., Theodoridis, E.: Data integration and disintegration: Managing springer nature scigraph with shacl and owl. In: International semantic web conference (Posters, Demos & Industry Tracks) (2017)
16. Heibi, I., Peroni, S., Shotton, D.: Oscar: A customisable tool for free-text search over sparql endpoints. In: Semantics, Analytics, Visualization, pp. 121–137. Springer (2017)
17. Heibi, I., Peroni, S., Shotton, D.: Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. Scientometrics **121**(2), 1213–1228 (2019)
18. Hendricks, G., Tkaczyk, D., Lin, J., Feeney, P.: Crossref: The sustainable source of community-owned scholarly metadata. Quantitative Science Studies **1**(1), 414–427 (2020)
19. Herzog, C., Hook, D., Konkiel, S.: Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies **1**(1), 387–395 (2020)
20. Hosseini, A., Ghavimi, B., Boukhers, Z., Mayr, P.: Excite–a toolchain to extract, match and publish open literature references. In: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL). pp. 432–433. IEEE (2019)
21. Lauscher, A., Eckert, K., Galke, L., Scherp, A., Rizvi, S.T.R., Ahmed, S., Dengel, A., Zumstein, P., Klein, A.: Linked open citation database: Enabling libraries to contribute to an open and interconnected citation graph. In: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries. pp. 109–118 (2018)
22. Li, H., Councill, I., Lee, W.C., Giles, C.L.: Citeseerx: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th international conference on World Wide Web. pp. 883–884 (2006)
23. Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D.S.: Gorc: A large contextual citation graph of academic papers. arXiv preprint arXiv:1911.02782 (2019)
24. Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (dcat). W3C recommendation (2014)
25. Maedche, A., Staab, S.: Comparing ontologies-similarity measures and a comparison study. AIFB (2001)
26. Peroni, S., Shotton, D.: Open citation identifier: Definition. Figshare (2019). https://doi.org/10.6084/m9.figshare.7127816
27. Peroni, S.: A simplified agile methodology for ontology development. In: OWL: Experiences and Directions–Reasoner Evaluation, pp. 55–69. Springer (2016)
28. Peroni, S., Shotton, D.: The SPAR ontologies. In: International Semantic Web Conference. pp. 119–136. Springer (2018)
29. Peroni, S., Shotton, D.: Opencitations, an infrastructure organization for open scholarship. Quantitative Science Studies **1**(1), 428–444 (2020)
30. Peroni, S., Shotton, D.M., Vitali, F.: A document-inspired way for tracking changes of RDF data **1799**, 26–33 (2016)
31. Porzel, R., Malaka, R.: A task-based approach for ontology evaluation. In: ECAI Workshop on Ontology Learning and Population, Valencia, Spain. pp. 1–6 (2004)
32. Sanderson, R., Ciccarese, P., Van de Sompel, H.: Designing the w3c open annotation data model. In: Proceedings of the 5th Annual ACM Web Science Conference. pp. 366–375 (2013)
33. Shotton, D., Peroni, S., Daquino, M.: In-text reference pointer identifier: Definition. Figshare (2020). https://doi.org/10.6084/m9.figshare.11674032

34. Wang, K., Shen, Z., Huang, C., Wu, C.H., Dong, Y., Kanakia, A.: Microsoft academic graph: When experts are not enough. Quantitative Science Studies **1**(1), 396–413 (2020)
35. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)