The OpenCitations Data Model

Version 2.0.1, February 23, 2020

Publication date for this document: February 23, 2020

Version number of this document: 2.0.1

Previous version v2.0, published November 8, 2019
https://doi.org/10.6084/m9.figshare.3443876.v6

Authors

Marilena Daquino University of Bologna, Italy
marilena.daquino2@unibo.it
marilena.daquino@opencitations.net
https://orcid.org/0000-0002-1113-7550

Silvio Peroni University of Bologna, Italy
silvio.peroni@unibo.it
silvio.peroni@opencitaitons.net
http://orcid.org/0000-0003-0530-4305

David Shotton University of Oxford, UK
david.shotton@oerc.ox.ac.uk
david.shotton@opencitations.net
http://orcid.org/0000-0001-5506-523X

License

Citation

Marilena Daquino, Silvio Peroni, David Shotton (2020). The OpenCitations Data Model. Version 2.0.1. Figshare. https://doi.org/10.6084/m9.figshare.3443876

Main changes since the previous versions

version 2.0.1

Corrected the DOI URL in section "Citation" above to comply with the current guidelines of Figshare.

version 2.0

Rationale

Initially, we developed the OpenCitations Data Model (OCDM) to describe explicitly data in the OpenCitations Corpus (OCC). However, in recent years OpenCitations has also been developing other datasets, and the OCDM has also been adopted by external projects. This new version of OCDM is designed to describe a generic bibliographic dataset, making easier its adoption and independent use by third parties for their own data, while still using OCC as an example.

[1] https://creativecommons.org/licenses/by/4.0/legalcode

We have achieved this by removing all the particular constraints related to our implementation of the OpenCitations Corpus, to make OCDM more flexible and easier to adopt by other. Retained references to the OpenCitations Corpus are now used as examples instead of being mandatory requirements. Open Citation Identifiers (OCIs) have been introduced since the previous version of the OCDM, and, for this reason, the local identifiers used for citations are now based on their OCIs. Further, we have extended such local identifiers for citations to permit the identification of individual in-text reference pointers (aka "in-text citations") to the same cited bibliographic resource, to which annotations can now be added. In addition, we found a way to simplify the number of provenance statements while expressing the same amount of information, facilitating implementation. Overall, the approach is lighter than before, with simplification of the way we handle virtual entities, where their provenance is now handled as for other entities.

Specific additions

Added definitions of four new bibliographic entities: reference pointer, pointer list, discourse element, and reference annotation. Added a new identifier, i.e. the XPath expression or function of discourse elements and reference pointers. Added four new ontology classes for describing reference pointers, i.e. *c4o:InTextReferencePointer*, pointer lists, i.e. *c4o:SingleLocationPointerList*, discourse elements, i.e. *deo:DiscourseElement* (and its subclasses), and reference annotations, i.e. *oa:Annotation*. Added the mapping to OWL of properties linking the aforementioned entities. Added a description of the metadata associated with the aforementioned bibliographic entities. Extended the JSON-LD excerpts showing how to linearize new data on citation annotations and in-text reference pointers.

The OpenCitations datasets

OpenCitations is an independent infrastructure organization for open scholarship directed by Silvio Peroni (Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy) and David Shotton (Oxford e-Research Centre, University of Oxford, Oxford, UK). It is dedicated to the publication of open bibliographic and citation data using Semantic Web (Linked Data) technologies. It is also engaged in advocacy for open citations, particularly in its role as a key founding member of the Initiative for Open Citations (I4OC, https://i4oc.org).

OpenCitations has a persistent URL at w3id.org, https://w3id.org/oc, which resolves to our server at http://opencitations.net. In terms of data, OpenCitations first developed the OpenCitations Corpus (OCC, https://w3id.org/oc/corpus), a database of open downloadable bibliographic and citation data, and is currently developing a number of Open Citation Indexes (https://w3id.org/oc/index) using the data openly available in third-party bibliographic databases. The first and largest of these is COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations (https://w3id.org/oc/index/coci). All OpenCitations published data are recorded in RDF and released under a Creative Commons CC0 public domain waiver[2].

This document describes the OpenCitations Data Model (OCDM), the data model used to model all such bibliographic and citation data exposed in any OpenCitation dataset in RDF, the 'language' of the Semantic Web, in particular by employing the OpenCitations SPAR (Semantic Publishing and Referencing) Ontologies[3]. Such usage permits the publication of bibliographic and citation data as Linked Open Data, thereby conferring machine readability and interoperability of the data on the Web. This OpenCitations Data Model may also be employed by third parties, either for their own use or to structure their data for submission to and publication by OpenCitations.

RDF resources in the OpenCitations datasets

Kinds of metadata

The OpenCitations Data Model makes available five levels of metadata:

- Dataset metadata
- Bibliographic entity metadata
- Identifiers
- Provenance metadata
- Virtual entities

Within the datasets, different classes of information (different types of entity) are identified and described using unique names and, usually, are accompanied with two-letter abbreviations ("short names"), for example **Bibliographic resource** (short: **br**).

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

[3] Peroni, S., Shotton, D. (2018). The SPAR Ontologies. In Proceedings of the 17th International Semantic Web Conference (ISWC 2018): 119-136. DOI: https://doi.org/10.1007/978-3-030-00668-6_8

Dataset metadata

All the datasets exposed by OpenCitations are described appropriately by means of standard vocabularies, such as the *Data Catalog Vocabulary*[4] and the *VoID Vocabulary*[5]. Such datasets can have particular distributions.

- **Dataset** (short: not applicable and strictly dependent on the implementation of the dataset infrastructure): a set of collected information about something.
- **Distribution** (short: **di**): an accessible form of a dataset, for example a downloadable file.

Bibliographic entity metadata

The following bibliographic entities are handled as RDF resources:

- **Bibliographic resource** (short: **br**): a published bibliographic resource that cites/is cited by another published bibliographic resource. Subclasses (extracted from CrossRef[6] Types[7], and extended to meet the specific needs of collaborating projects and institutions) include:
  - *Archival document*
  - *Book*
  - Book chapter
  - *Book part*
  - *Book section*
  - *Book series*
  - *Book set*
  - Book track (audio tracks, usually related to a particular article or book)
  - Component (figures, tables, and the like, contained in an article)
  - Dataset
  - Dissertation
  - *Edited book*
  - Journal article
  - *Journal issue*
  - *Journal volume*
  - *Journal*
  - *Monograph*
  - Proceedings article
  - *Proceedings*
  - *Reference book*
  - Reference entry
  - *Report series*
  - Report
  - *Standard series*
  - Standard (a specification document)

[4] http://www.w3.org/TR/vocab-dcat/
[5] http://www.w3.org/TR/void/
[6] http://crossref.org/
[7] http://api.crossref.org/types

Those in *italics* refers to resources that can also be treated as container resources, i.e. those that may contain another cited resource (e.g. a journal containing a cited article, a book containing a cited chapter).

- **Resource embodiment** (short: **re**): the particular physical or digital format in which a bibliographic resource was made available by its publisher. Subclasses include:
  - Digital embodiment
  - Print embodiment
- **Discourse element** (short: **de**): a document component, either structural (e.g. paragraph, section, chapter, table, caption, footnote, title) or rhetorical (e.g. introduction, discussion, acknowledgements, reference list, figure, appendix), in which the content of a bibliographic resource can be organized. Subclasses include, but are not limited to:
  - Section
  - Paragraph
  - Sentence
  - Text chunk
  - Table
  - Footnote
  - Caption
- **Reference pointer** (long: in-text reference pointer; short: **rp**): a textual device (e.g. "[1]"), denoting a single bibliographic reference, that is embedded in the text of a document within the context of a particular sentence or text chunk. A bibliographic reference can be denoted in the text by one or more in-text reference pointers.
- **Pointer list** (short: **pl**): a textual device (e.g. "[1, 2, 3]" or "[4-9]") which includes a number of reference pointers denoting the specific bibliographic references to which the list pertains.
- **Bibliographic reference** (short: **be**)[8]: the particular textual bibliographic reference, usually occurring in the reference list (and denoted by one or more in-text reference pointers within the text of a citing bibliographic resource), that references another bibliographic resource.
- **Responsible agent** (short: **ra**): the agent (usually a person or an organisation) having a certain role with respect to a bibliographic resource (e.g. an author of a paper or book, or an editor of a journal).
- **Agent role** (short: **ar**): a particular role held by an agent with respect to a bibliographic resource.
- **Citation** (short: **ci**): a permanent conceptual directional link from the citing bibliographic resource to a cited bibliographic resource. A citation is created by the performative act of an author citing a published work that is relevant to the current work by using a particular textual device. Typically, citations are made by including a bibliographic reference in the reference list of the citing work and by denoting such a bibliographic reference using one or more in-text reference pointers (e.g. "[1]"), or by the inclusion within the citing work of a link, in the form of an HTTP Uniform Resource Locator (URL), to the cited bibliographic resource on the World Wide Web. The class

[8] The short name "be" refers to the old naming convention we used in the previous versions of the OpenCitations Data Model, which specified this kind of entities as *bibliographic entry*. While we have updated the name of the entity with the (more appropriate) wording of *bibliographic resource*, we decided to keep the original short name as "be" for back-compatibility.

Citation has subclasses defining particular types of citation, enabling an individual citation to be qualified by the reason an author has cited this other resource (e.g. agrees with, uses method in, cites as authority), or by type, (e.g. self-citation).

  - Self-citation: a citation in which the citing and the cited entities have something significant in common with one another. Subclasses include:
    - Affiliation self-citation: a citation in which at least one author from each of the citing and the cited entities is affiliated with the same academic institution.
    - Author self-citation: a citation in which the citing and the cited entities have at least one author in common.
    - Author network self-citation: a citation in which at least one author of the citing entity has direct or indirect co-authorship links with one of the authors of the cited entity.
    - Funder self-citation: a citation in which the works reported in the citing and the cited bibliographic entities were funded by the same funding agency.
    - Journal self-citation: a citation in which the citing and the cited entities are published in the same journal.
  - Journal cartel citation: a citation from one journal to another journal which forms one of an abnormally large number of citations from the citing journal to recent articles in the cited journal, for the purpose of promoting the cited journal.
  - Distant citation: a citation in which the citing and the cited entities have nothing significant in common with one another over and beyond their subject matter.

- **Reference annotation** (short: **an**): an annotation, attached either to an in-text reference pointer or to a bibliographic reference, describing the related citation. If an in-text reference pointer is annotated, the related citation may be characterized with a citation function (the reason for that citation) specific to the textual location of that in-text reference pointer within the citing entity. If a bibliographic reference is annotated, the related citation may be similarly characterized in a more general way with a citation function (the reason for that citation).

Identifiers

All the aforementioned bibliographic entities stored within the dataset[9] **must** have a dataset identifier:

- The **dataset identifier** assigned to the entity is composed by the *two-letter short name* for the class of items (e.g. **be** for a bibliographic reference) followed by an oblique slash ("/") and a *local identifier* (defined below). Note that the dataset identifier is for

[9] This property of being 'stored within the dataset' does not apply to the entities belonging to an external dataset that are simply referenced from within the dataset by using their IRIs. For instance, the citing and cited entities of each citation defined in COCI (the OpenCitations Index of Crossref open DOI-to-DOI citations; http://opencitations.net/index/coci) are defined by means of their DOI URLs (e.g. http://dx.doi.org/10.1108/jd-12-2013-0166). See, for example, the COCI citation https://w3id.org/oc/index/coci/ci/020070505043614211815143703020802020200101000836191363010263020001036300010606. In this case, the dataset (i.e. COCI), while storing metadata about the citation itself, does not store any metadata about the citing or cited entities other than their HTTP URLs. The metadata for the citing and cited entities can usually be retrieved on-the-fly by accessing such an external dataset by means of the entities' IRIs and the HTTP protocol.

internal use only within the dataset, and is distinct from any "public" Internationalized Resource Identifier (abbreviated IRI) that may be used to identify the entity.

The structure of the **local identifier** depends on the type of bibliographic entity under consideration:

- For all the bibliographic entities included within OpenCitations datasets *except citations*: the **local identifier** is composed of a prefix[10] and a body. The prefix consists of a positive number (following the pattern "*nnn*", where "*nnn*" is a string of numerals of variable length which includes no zeros), enclosed between two zeros (e.g. "0420"). The body, which is present in all local identifiers, is a positive integer (e.g. "23"). The prefix and the body together form the local identifier (e.g. "042023")[11].

  The prefix may be used to identify a supplier of the entity information stored within the dataset (for example, when bibliographic entity metadata compliant with the OpenCitations Data Model, provided by a third party, the Excite Project[12], is ingested into the OpenCitations Corpus, it uses the Excite Project prefix "0110" to identify its provenance). A list of all currently assigned supplier prefixes is available at https://github.com/opencitations/oci/blob/master/suppliers.csv.

  Third parties wishing to adopt the OpenCitations Data Model for their own use are encouraged to use unique positive integers as local bibliographic entity identifiers. Should they wish to submit their date for inclusion within a dataset published by OpenCitations, a unique OpenCitations local identifier prefix will at that stage be assigned to the third party, denoting the provenance of their data within the OpenCitations dataset.

- For citations: the numerical part of the citation's **Open Citation Identifier** (OCI), prefixed by "ci/", is used as the **local identifier** for a citation. The OCI is defined in the Open Citation Identifier Definition available at https://doi.org/10.6084/m9.figshare.7127816. Each OCI is composed of an initial "oci:" followed by two sequences of numerals separated by a dash, where the first sequence of numerals identifies the citing entity while the second sequence of numerals identifies the cited entity. For instance, if a bibliographic resource with local identifier "br/042023" cites another bibliographic resource with local identifier "br/042027", the OCI for that citation would be oci:042023-042027, and the local identifier for the citation would be "ci/042023-042027".

  To permit the identification of a citation involving a particular in-text reference pointer (a.k.a. "an in-text citation"), and the recording of its local textual context, in cases where multiple in-text reference pointers denote the same bibliographic reference within the text of the citing bibliographic resource, the citation local identifier can optionally be followed by "/" and a positive number. For instance, if we have a bibliographic reference in the reference list of a citing entity, and two in-text reference pointers within the body text denoting such a bibliographic reference, we may wish to describe three citations: "ci/042023-042027" related to the bibliographic reference itself, and "ci/042023-042027/1" and "ci/042023-042027/2" respectively referring to the citations related to the first and second in-text reference pointers encountered in

[10] The prefix is omitted only in the case of local identifiers for entities that were ingested into the OpenCitations Corpus prior to February 2018.
[11] "0420" is a hypothetical prefix, not (currently) assigned to any particular data supplier, and used here solely as an example.
[12] https://west.uni-koblenz.de/en/research/excite

the body text of the citing entity, counted using the conventional beginning-to-end reading order.

In addition, a bibliographic entity may have one or more other public identifiers assigned to it by external third parties:

- **Identifier** (short: **id**): an external identifier (e.g. DOI[13], ORCID[14], PubMedID[15], Open Citation Identifier[16]) associated with the bibliographic entity. Members of this class of metadata are themselves given unique corpus identifiers, as described above, e.g. "id/0420129".

Provenance metadata

All the aforementioned bibliographic entities and their identifiers **must** have metadata describing their provenance (except in the case of *virtual* entities, described below). This provenance metadata entity is:

- **Snapshot of entity metadata** (short: **se**): a particular snapshot recording the metadata associated with an individual entity (either a bibliographic entity or an identifier) at a particular date and time, including the agent, such as a person, organisation or automated process that created or modified the entity metadata.

Virtual entities

Bibliographic entities can be made available within the Corpus as *virtual* RDF resources, by which we mean entities that are created on-the-fly, and only when they are requested (i.e. by accessing their URLs). These are defined either by using data relating to non-virtual bibliographic entities that are already available within the dataset, or by using data that are themselves obtained on-the-fly from an external supplier. Note that this approach of using virtual RDF resources is optional, and is simply employed for storage efficiency, to avoid duplication of information within the dataset[17].

---

[13] https://www.doi.org/

[14] http://orcid.org/

[15] http://www.ncbi.nlm.nih.gov/pubmed

[16] https://w3id.org/oc/oci

[17] For instance, as of September 2019, the OpenCitations Corpus (OCC, https://w3id.org/oc/corpus) defines as virtual entities only one type of bibliographic entity, namely citations (i.e. members of the class Citation). The local identifiers for members of this class and for their public identifiers are defined as follows:

- **Citation** (short: **ci**): the local identifier for a citation is the string obtained by combining the local identifiers for the citing and cited bibliographic resources relating to that citation, both prefixed by "030" (i.e. the prefix assigned to OCC) and separated with a dash ("-"). For instance, the citation from citing resource "br/1" to cited resource "br/18", both resources being within the OCC, is given an OCC local identifier "0301-03018". The corresponding OCI for this citation between two bibliographic resources within the OpenCitations Corpus is oci:0301-03018.
- **Identifier** (short: **id**): the local identifier for the recorded public Identifier of a citation is the string obtained by taking the corpus identifier of the citation it identifies (e.g. "ci/0301-03018") and then substituting the "/" with a dash "-" (e.g. to become "ci-0301-03018").

Because we do not separately store these virtual entities within the Corpus triplestore, they cannot be directly queried by means of the OCC SPARQL end-point. In addition, they are not stored within its data dumps. However, the data associated with a virtual entity within the OCC can be obtained by accessing its URL (defined below).

In particular, the following rules hold for each virtual entity:

- Its local identifier may not follow the usual structure provided for other bibliographic entities, but may be defined according to specific and *ad hoc* rules;
- Its URL is clearly distinguishable from those used for the other (non-virtual) bibliographic entities (see the following section defining URLs).

Naming convention for entities and provenance data

In the corpus, we distinguish four different kinds of URLs: URL for datasets and distributions, URLs for bibliographic entities, URLs for provenance data, and URL for virtual entities.

URLs for datasets and distributions

The URL identifying the corpus is the following:

[dataset URL] : [base URL]/[dataset name]/

where the *base URL* must be chosen for guaranteeing persistence over time. OpenCitations has a persistent URL at w3id.org: https://w3id.org/oc, and uses it for defining its dataset URLs, for instance:

- the URL of the OpenCitations Corpus is https://w3id.org/oc/corpus/ (dataset name: "corpus");
- the URL of the sub-dataset containing bibliographic resources within the OpenCitations Corpus is https://w3id.org/oc/corpus/br/ (dataset name: "corpus/br");
- the URL of COCI is https://w3id.org/oc/index/coci/ (dataset name: "index/coci").

The URL defining one or more distributions of a dataset is:

[dataset distribution URL] : [dataset URL]di/[iterative positive number]

where the *iterative positive number* is a number assigned to each distribution, unique among distributions of resources of the same type. For example, the first distribution of the entire dataset of the OpenCitations Corpus is:

- https://w3id.org/oc/corpus/di/1

URLs for bibliographic entities and their identifiers

The URL of each of the bibliographic entities in the dataset is constructed according to a particular naming convention scheme, introduced as follows:

[entity URL] : [dataset URL][entity dataset identifier]

where *dataset URL* and the *entity dataset identifier* are as previously defined. For example, the third entry within the OpenCitations Corpus class of bibliographic resources, and the 129th entry within the OpenCitations Corpus class of identifiers, have the following URLs respectively:

- https://w3id.org/oc/corpus/br/3
- https://w3id.org/oc/corpus/id/129

URLs for provenance metadata

Each of the dataset bibliographic entities and identifiers has associated with it a particular RDF provenance metadata graph that record information about its creation, modification and/or merging. The URL for such an entity provenance graph has the following structure:

[entity provenance URL] : [entity URL]/prov/

Such a graph contains all the provenance information related to the bibliographic entity/identifier under consideration, except that relating to provenance agents. For example, URL for the provenance graph for the 15th bibliographic resource in the corpus is:

- https://w3id.org/oc/corpus/br/15/prov/

The dataset provenance metadata entities (i.e. snapshots) relating to a particular dataset bibliographic entity or identifier use the following convention for their URLs:

[snapshot entity URL] : [entity provenance URL]se/[iterative number]

where *iterative number* is assigned as explained for the other bibliographic entities (except citations).

For example, the second snapshot related to "br/1423490" in the OpenCitations Corpus has the following URLs:

- https://w3id.org/oc/corpus/br/1423490/prov/se/2

Please note that all the provenance entities must be assigned to the provenance metadata graph associated with the entity of the dataset for which they provide provenance information (e.g. "https://w3id.org/oc/corpus/br/1423490/prov/se/2" is stored in provenance graph "https://w3id.org/oc/corpus/br/1423490/prov/"). This has been prescribed so as to make it easy to retrieve all the provenance information related to a particular entity simply by accessing all the statements in the relevant provenance metadata graph.

URLs for virtual entities

The URL of each of the virtual entities in the corpus is constructed according to a particular naming convention scheme, introduced as follows:

[virtual entity URL] : [base URL]/virtual/[entity dataset identifier]

where *base URL* and the *entity dataset identifier* are as previously defined. For example, the citation between the 1st and the 18th bibliographic resources in the OpenCitations Corpus, and

the OpenCitations Corpus identifier associated with this citation, have the following URLs respectively:

- https://w3id.org/oc/virtual/ci/0301-03018
- https://w3id.org/oc/virtual/id/ci-0301-03018

All such virtual entities **are not** assigned to any dataset graph, since they are derivative bibliographic entities. Instead, all the provenance data must be recorded within the appropriate graph – https://w3id.org/oc/virtual/ci/0301-03018/prov/ for the entity https://w3id.org/oc/virtual/ci/0301-03018.

Metadata elements associated with OCC datasets and distributions

In this section, we introduce all the metadata elements that may be associated with each dataset or distribution.

Metadata elements that may be associated with any non-virtual dataset

- has title: *literal*
  The title of the dataset.
- has description: *literal*
  A short textual description of the content of the dataset.
- has release date: *date*
  The date of first publication of the dataset.
- has modification date: *date*
  The date on which the dataset has been modified.
- has keyword: *literal*
  A keyword or phrase describing the content of the dataset.
- has subject: *concept*
  A concept describing the primary subject of the dataset.
- has distribution: *distribution*
  A distribution of the dataset.
- has landing page: *document*
  An HTML page (indicated by its URL) representing a browsable page for the dataset.
- has sub-dataset: *dataset*
  A link to a subset of the present dataset.
- has SPARQL endpoint: *URL*
  The link to the SPARQL endpoint for querying the dataset.

Metadata elements that may be associated with a distribution

- has title: *literal*
  The title of the distribution.
- has description: *literal*
  A short textual description of the content of the distribution.
- has release date: *date*
  The date of first publication of the distribution.

- has license: *document*
  The resource describing the license associated with the data in the distribution.
- has download URL: *document*
  The URL of the document where the distribution is stored.
- has file type: *media type*
  The file type of the representation of the distribution (according to IANA media types).
- has byte size: *literal*
  The size in bytes of the distribution.

Metadata elements associated with an individual bibliographic entity

In this section, we introduce all the metadata elements that may be associated with each of the following bibliographic entities.

Metadata elements that may be associated with any bibliographic entity

- has identifier: *identifier*
  In addition to the internal **dataset identifier** assigned to the entity upon initial curation (format: [entity short name]/[local identifier], as specified above), other external third-party identifiers can be specified through this attribute (e.g. DOI, ORCID, PubMedID).

Metadata elements that may be associated with a bibliographic resource

- has type: *thing*
  The type of the bibliographic resource, conforming to those introduced above.
- has title: *literal*
  The title of the bibliographic resource.
- has subtitle: *literal*
  The subtitle of the bibliographic resource.
- is part of: *bibliographic resource (br)*
  The corpus identifier of the bibliographic resource (e.g. issue, volume, journal, conference proceedings) that is a container for the subject bibliographic resource.
- cites: *bibliographic resource (br)*
  The corpus identifier of the bibliographic resource cited by the subject bibliographic resource.
- has part: *bibliographic reference (be)* or *discourse element (de)*
  A bibliographic reference within the bibliographic resource, or a discourse element wherein the text of the bibliographic resources can be organized.
- has publication date: *gYear* or *gYearMonth* or *date*
  The date of publication of the bibliographic resource.
- is embodied as: *resource embodiment (re)*
  The corpus identifier of the resource embodiment defining the format in which the bibliographic resource has been embodied, which can be either print or digital.
- has number: *literal*
  A literal (for example a number or a letter) that identifies the sequence position of the bibliographic resource as a particular item within a larger collection (e.g. an article

number within a journal issue, a volume number of a journal, a chapter number within a book).

- has edition: *literal*
  An identifier for one of several alternative editions of a particular bibliographic resource.
- has contributor: *agent role (ar)*
  The role (e.g. author, editor, or publisher) of one of the contributors of this bibliographic resource.
- has related document: *thing*
  A document external to the Corpus, that is related to the bibliographic resource (such as a version of the bibliographic resource – for example a preprint – recorded in an external database).

Due to the precise specification of these metadata, the names of the authors and the information about the publication (journal name, volume number, etc.) are not directly accessible as literal values associated with the bibliographic resource under consideration. However, they are accessible by following other metadata elements that *are* directly associated with the bibliographic resource – for instance, the authors can discovered via the agent role entities specified by the *has contributor* element, while the name of the journal in which the bibliographic resource has been published (as well as the volume number and the issue) can be obtained by looking at the entities specified by the *is part of* element.

Metadata elements that may be associated with a discourse element

- has type: *thing*
  The type of discourse element – such as "paragraph", "section", "sentence", "acknowledgements", "reference list" or "figure".
- has title: *literal*
  The title of the discourse element, such as the title of a figure or a section in an article.
- has part: *discourse element (de)*
  The discourse element hierarchically nested within the parent element, such as a sentence within a paragraph, or a paragraph within a section.
- has next: *discourse element (de)*
  The following discourse element that includes at least one in-text reference pointer.
- is context of: *reference pointer (rp)* or *pointer list (pl)*
  Provides the textual and semantic context of the in-text reference pointer or list of in-text reference pointers that appears within the discourse element.
- has content: *literal*
  The literal document text contained by the discourse element.

Metadata elements that may be associated with a reference pointer

- has next: *reference pointer (rp)*
  The following in-text reference pointer, when included within a single in-text reference pointer list.
- denotes: *bibliographic reference (be)*
  The bibliographic reference included in the list of bibliographic references, denoted by the in-text reference pointer.

- has annotation: *reference annotation (an)*
  An annotation characterizing the citation to which the in-text reference pointer relates in terms of its citation function (the reason for that citation) specific to the textual location of that in-text reference pointer within the citing entity.
- has reference pointer text: *literal*
  The literal text of the textual device forming an in-text reference pointer and denoting a single bibliographic reference (e.g. "[1]").

Metadata elements that may be associated with a pointer list

- has element: *reference pointer (rp)*
  The in-text reference pointer that is part of the in-text reference pointer list present at a particular location within the body of the citing work.

Metadata elements that may be associated with a responsible agent's role

- has role type: *thing*
  The specific type of role under consideration (e.g. author, editor or publisher).
- is held by: *responsible agent (ra)*
  The agent holding this role with respect to a particular bibliographic resource.
- has next: *agent role (ar)*
  The following role in a sequence of agent roles of the same type associated with the same bibliographic resource (so as to define, for instance, an ordered list of authors).

Metadata elements that may be associated with a responsible agent

- has name string: *literal*
  The name of an agent (for people, usually in the format: given name followed by family name, separated by a space).
- has given name: *literal*
  The given name of an agent, if a person.
- has family name: *literal*
  The family name of an agent, if a person.
- has related agent: *thing*
  An external agent that/who is related in some relevant way with this responsible agent (e.g. for inter-linking purposes).

Metadata elements that may be associated with a resource embodiment

- has type: *thing*
  It identifies the particular type of the embodiment, either digital or print.
- has format: *media type*
  It allows one to specify the IANA media type of the embodiment.
- has first page: *literal*
  The first page of the bibliographic resource according to the current embodiment.
- has last page: *literal*
  The last page of the bibliographic resource according to the current embodiment.

- has url: *document*
  The URL at which the embodiment of the bibliographic resource is available.

Metadata elements that may be associated with a bibliographic reference

- has bibliographic reference text: *literal*
  The literal text of a bibliographic reference occurring in the reference list (or elsewhere) within a bibliographic resource, that references another bibliographic resource. The reference text should be recorded "as given" in the citing bibliographic resource, including any errors (e.g. mis-spellings of authors' names, or changes from "*β*" in the original published title to "beta" in the reference text) or omissions (e.g. omission of the title of the referenced bibliographic resource, or omission of sixth and subsequent authors' names, as required by certain publishers), and in whatever format it has been made available. For instance, the reference text can be either as plain text or as a block of XML.
- references: *bibliographic resource (br)*
  The cited bibliographic resource to which this bibliographic reference relates.
- has annotation: *reference annotation (an)*
  An annotation characterizing the related citation, in terms of its citation function (the reason for that citation).

Metadata elements that may be associated with a citation

- has citing document: *bibliographic resource (br)*
  The bibliographic resource which acts as the source for the citation.
- has cited document: *bibliographic resource (br)*
  The bibliographic resource which acts as the target for the citation.
- has citation creation date: *gYear* or *gYearMonth* or *date*
  The date on which the citation was created[18].
- has citation time span: *duration*
  The date interval between the publication date of the cited bibliographic resource and the publication date of the citing bibliographic resource.
- has citation characterization: *thing*
  The citation function characterizing the purpose of the citation.

Metadata elements that may be associated with a reference annotation

- has citation: *citation (ci)*
  The citation to which the annotation relates, that is relevant *either* to a bibliographic reference *or* to an in-text reference pointer that denotes such a bibliographic reference.

Provenance information

Each of the aforementioned bibliographic entities introduced into the dataset has associated provenance information that documents the processes that have led to the current description of that resource. In this section, we introduce all the provenance metadata

[18] This has the same numerical value as the publication date of the citing bibliographic resource, but is a property of the citation itself. When combined with the citation time span, it permits that citation to be located in history.

elements that constitute the provenance information for a particular bibliographic entity, all of which elements are stored within the entity's single provenance graph.

Metadata elements that may be associated with a snapshot of entity metadata (se)

- has creation date: *date time*
  The date on which a particular snapshot of a bibliographic entity's metadata was created.
- has invalidation date: *date time*
  The date on which a snapshot of a bibliographic entity's metadata was invalidated due to an update (e.g. a correction, or the addition of some metadata that was not specified in the previous snapshot), or due to a merger of the entity with another one.
- is snapshot of: *bibliographic entity (en)*
  This property is used to link a snapshot of entity metadata to the bibliographic entity to which the snapshot refers.
- is derived from: *snapshot of entity metadata (se)*
  This property is used to identify the immediately previous snapshot of entity metadata associated with the same bibliographic entity.
- has primary source: *thing*
  This property is used to identify the primary source from which the metadata described in the snapshot are derived (e.g. Crossref, as the result of querying the CrossRef API).
- has update action: *thing*
  The UPDATE SPARQL query that specifies which data, associated to the bibliographic entity in consideration, have been modified (e.g. for correcting a mistake) in the current snapshot starting from those associated to the previous snapshot of the entity.
- has description: *literal*
  A textual description of the events that have resulted in the current snapshot (e.g. the creation of the initial snapshot, the creation of a new snapshot following the modification of the entity to which the metadata relate, or the creation of a new snapshot following the merger with another entity of the entity to which the previous snapshot related).
- is attributed to: *thing*
  The agent responsible for the creation of the current entity snapshot.

Mapping with OWL

This section introduces all the mapping of the entities mentioned in the previous section with OWL ontology definitions.

Mapping entities types

We provide a mapping to RDF of the bibliographic entities described in this OpenCitations Data Model using OWL ontologies, in particular the OpenCitations SPAR (Semantic Publishing and Referencing) Ontologies[19]; the well-known library, publishing and Web vocabularies

[19] http://www.sparontologies.net

Dublin Core[20], FRBR[21], PRISM[22] and RDF[23]; and the following additional models: DCAT[24], FOAF[25], Literal Reification[26], OCO[27], OA[28], PROV-O[29], and VOID[30].

The following prefixes are employed:

```
biro:       http://purl.org/spar/biro/
cito:       http://purl.org/spar/cito/
co:         http://purl.org/co/
c4o:        http://purl.org/spar/c4o/
datacite:   http://purl.org/spar/datacite/
dcat:       http://www.w3.org/ns/dcat#
dcterms:    http://purl.org/dc/terms/
deo:        http://purl.org/spar/deo/
doco:       http://purl.org/spar/doco/
fabio:      http://purl.org/spar/fabio/
foaf:       http://xmlns.com/foaf/0.1/
frbr:       http://purl.org/vocab/frbr/core#
literal:    http://www.essepuntato.it/2010/06/literalreification/
oa:         http://www.w3.org/ns/oa#
oco:        https://w3id.org/oc/ontology/
prism:      http://prismstandard.org/namespaces/basic/2.0/
pro:        http://purl.org/spar/pro/
prov:       http://www.w3.org/ns/prov#
rdf:        http://www.w3.org/1999/02/22-rdf-syntax-ns#
void:       http://rdfs.org/ns/void#
```

Datasets and distributions

- Dataset: dcat:Dataset
- Distribution: dcat:Distribution

Bibliographic entities

- Bibliographic reference: biro:BibliographicReference
- Responsible agent: foaf:Agent
- Agent role: pro:RoleInTime
- Bibliographic resource: fabio:Expression
  Subclasses:
  - Archival document fabio:ArchivalDocument

[20] http://dublincore.org/documents/dcmi-terms/
[21] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
[22] http://www.idealliance.org/specifications/prism-metadata-initiative
[23] https://www.w3.org/TR/rdf11-concepts/
[24] http://www.w3.org/TR/vocab-dcat
[25] http://xmlns.com/foaf/spec/
[26] http://ontologydesignpatterns.org/wiki/Submissions:Literal_Reification
[27] https://w3id.org/oc/ontology
[28] https://www.w3.org/TR/annotation-model/
[29] http://www.w3.org/TR/prov-o
[30] http://www.w3.org/TR/void

    - Book — fabio:Book
    - Book chapter — fabio:BookChapter
    - Book part — doco:Part, part of a fabio:Book
    - Book section — fabio:ExpressionCollection, part of a fabio:Book
    - Book series — fabio:BookSeries
    - Book set — fabio:BookSet
    - Book track — fabio:Expression
    - Component — fabio:Expression
    - Dataset — fabio:DataFile
    - Dissertation — fabio:Thesis
    - Edited book — fabio:Book
    - Journal article — fabio:JournalArticle
    - Journal Issue — fabio:JournalIssue
    - Journal Volume — fabio:JournalVolume
    - Journal — fabio:Journal
    - Monograph — fabio:Book
    - Proceedings article — fabio:ProceedingsPaper
    - Proceedings — fabio:AcademicProceedings
    - Reference book — fabio:ReferenceBook
    - Reference entry — fabio:ReferenceEntry
    - Report series — fabio:Series (of some fabio:ReportDocument)
    - Report — fabio:ReportDocument
    - Standard series — fabio:Series (of some fabio:SpecificationDocument)
    - Standard — fabio:SpecificationDocument
- Discourse element — deo:DiscourseElement

  Subclasses:
    - Caption — deo:Caption
    - Footnote — doco:Footnote
    - Paragraph — doco:Paragraph
    - Section — doco:Section
    - Section title — doco:SectionTitle
    - Sentence — doco:Sentence
    - Table — doco:Table
    - Text chunk — doco:TextChunk
- Reference pointer — c4o:InTextReferencePointer
- Pointer list — c4o:SingleLocationPointerList
- Resource embodiment: — fabio:Manifestation

  Subclasses:
    - Digital embodiment — fabio:DigitalManifestation
    - Print embodiment — fabio:PrintObject
- Citation: — cito:Citation

  Subclasses:
    - Self-citation — cito:SelfCitation

      Subclasses:
        - Affiliation self-citation — cito:AffiliationSelfCitation

        - Author network self-citation — cito:AuthorNetworkSelfCitation
        - Author self-citation — cito:AuthorSelfCitation
        - Funder self-citation — cito:FunderSelfCitation
        - Journal self-citation — cito:JournalSelfCitation
    - Journal cartel citation — cito:JournalCartelCitation
    - Distant citation — cito:DistantCitation
- Reference annotation — oa:Annotation

Several of the aforementioned bibliographic entities have been mapped to entities defined in FaBiO, the FRBR-aligned Bibliographic Ontology (http://purl.org/spar/fabio), which is based on the Functional Requirements for Bibliographic Records (FRBR)[31]. While FRBR distinguishes between works, expressions, manifestations and items, all the bibliographic resources discussed here are defined as **expressions** of works, that may be manifested in physical (e.g. printed paper) or electronic form.

Identifier

- Identifier: datacite:Identifier

Provenance data

- Snapshot of entity metadata: prov:Entity

Mapping entities attributes and properties

In this section, we introduce the mapping between all the attributes and properties with OWL-related entities.

Datasets and distributions

Any dataset:

- has title: dcterms:title
- has subtitle: fabio:hasSubtitle
- has description: dcterms:description
- has publication date: dcterms:issued
- has modification date: dcterms:modified
- has keyword: dcat:keyword
- has subject: dcat:theme
- has distribution: dcat:distribution

Main dataset (all the above, plus the following ones):

[31] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

- has landing page: dcat:landingPage
- has sub-dataset: void:subset
- has SPARQL endpoint: void:sparqlEndpoint

Distribution:

- has title: dcterms:title
- has description: dcterms:description
- has publication date: dcterms:issued
- has license: dcterms:license
- has download URL: dcat:downloadURL
- has file type: dcat:mediaType
- has byte size: dcat:byteSize

Bibliographic entities

Any of the following resources

- has identifier: datacite:hasIdentifier

Bibliographic reference

- has bibliographic reference text: c4o:hasContent
- references: biro:references
- has annotation: oco:hasAnnotation

Citation

- has citing document: cito:hasCitingEntity
- has cited document: cito:hasCitedEntity
- has citation creation date: cito:hasCitationCreationDate
- has citation time span: cito:hasCitationTimeSpan
- has citation characterization: cito:hasCitationCharacterization

Reference annotation:

- has citation: oa:hasBody

Agent role

- has role type: pro:withRole
- is held by: pro:isHeldBy
- has next: oco:hasNext

Responsible agent

- has name: foaf:name
- has given name: foaf:givenName

- has family name: foaf:familyName
- has related agent: dcterms:relation

Bibliographic resource

- has type: rdf:type
- has title: dcterms:title
- is part of: frbr:partOf
- cites: cito:cites
- has publication date: prism:publicationDate
- is embodied as: frbr:embodiment
- has number: fabio:hasSequenceIdentifier
- has edition: prism:edition
- has part: frbr:part
- has contributor: pro:isDocumentContextFor
- has related document: dcterms:relation

Discourse element:

- has type: rdf:type
- has title: dcterms:title
- has part: frbr:part
- has next: oco:hasNext
- is context of: c4o:isContextOf
- has content: c4o:hasContent

Reference pointer:

- has reference pointer text: c4o:hasContent
- has next: oco:hasNext
- denotes: c4o:denotes
- has annotation: oco:hasAnnotation

Single location pointer list:

- has pointer list text: c4o:hasContent
- has element: co:element

Resource embodiment:

- has type: rdf:type
- has format: dcterms:format
- has first page: prism:startingPage
- has last page: prism:endingPage
- has url: frbr:exemplar

Identifier

Identifier

- has literal value: literal:hasLiteralValue
- has scheme: datacite:usesIdentifierScheme

Provenance data

Snapshot of entity metadata

- has creation date: prov:generatedAtTime
- has invalidation date: prov:invalidatedAtTime
- is snapshot of: prov:specializationOf
- is derived from: prov:wasDerivedFrom
- has primary source: prov:hadPrimarySource
- is attributed to: prov:wasAttributedTo
- has description: dcterms:description
- has update action oco:hasUpdateQuery

Linearization in BibJSON + JSON-LD

The RDF data included in the OCC is available in a triplestore, accompanied by a SPARQL endpoint, and is stored in JSON-LD format. The BibJSON specification (http://okfnlabs.org/bibjson/) has been adopted, since it provides JSON labels for the description of bibliographic entities. In the following subsections, we introduce alignment between OCC terms and the IRIs of the ontological entities described in the previous section, and give examples of linearization of some of the aforementioned entities.

Context

A JSON-LD context document is a mapping document that formally maps terms used in a dataset following the OpenCitations Data Model to the entities defined in the various ontologies used for describing such data in RDF. An implementation of such JSON-LD context is defined as follows – where "[dataset URL]" should be replaced with the particular URL of the dataset in consideration.

```
{
  "@context": {
    "gan": "[dataset URL]an/",
    "gar": "[dataset URL]ar/",
    "gbe": "[dataset URL]be/",
    "gbr": "[dataset URL]br/",
    "gci": "[dataset URL]ci/",
    "gde": "[dataset URL]de/",
    "gdi": "[dataset URL]di/",
    "gid": "[dataset URL]id/",
    "gpl": "[dataset URL]pl/",
    "gra": "[dataset URL]ra/",
    "gre": "[dataset URL]re/",
    "grp": "[dataset URL]rp/",

    "application": "https://w3id.org/spar/mediatype/application/",
    "biro": "http://purl.org/spar/biro/",
    "co": "http://purl.org/co/",
```

```
"c4o": "http://purl.org/spar/c4o/",
"cito": "http://purl.org/spar/cito/",
"datacite": "http://purl.org/spar/datacite/",
"dbr": "http://dbpedia.org/resource/",
"dcat": "http://www.w3.org/ns/dcat#",
"dcterms": "http://purl.org/dc/terms/",
"deo": "http://purl.org/spar/deo/",
"doco": "http://purl.org/spar/doco/",
"fabio": "http://purl.org/spar/fabio/",
"foaf": "http://xmlns.com/foaf/0.1/",
"frbr": "http://purl.org/vocab/frbr/core#",
"literal": "http://www.essepuntato.it/2010/06/literalreification/",
"oa": "http://www.w3.org/ns/oa#",
"oco": "https://w3id.org/oc/ontology/",
"prism": "http://prismstandard.org/namespaces/basic/2.0/",
"pro": "http://purl.org/spar/pro/",
"prov": "http://www.w3.org/ns/prov#",
"rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
"rdfs": "http://www.w3.org/2000/01/rdf-schema#",
"text": "https://w3id.org/spar/mediatype/text/",
"void": "http://rdfs.org/ns/void#",
"xsd": "http://www.w3.org/2001/XMLSchema#",

"iri": "@id",
"a": "@type",
"value": "@value",

"affiliation_self_citation": "cito:AffiliationSelfCitation",
"agent": "foaf:Agent",
"author_network_self_citation": "cito:AuthorNetworkSelfCitation",
"author_self_citation": "cito:AuthorSelfCitation",
"article": "fabio:JournalArticle",
"book": "fabio:Book",
"book_part": "doco:Part",
"collection": "fabio:ExpressionCollection",
"book_series": "fabio:BookSeries",
"book_set": "fabio:BookSet",
"caption": "deo:Caption",
"citation_relationship": "cito:Citation",
"creation": "prov:Create",
"curatorial_activity": "prov:Activity",
"curatorial_role": "prov:Association",
"dataset": "fabio:DataFile",
"digital_format": "fabio:DigitalManifestation",
"discourse_element": "deo:DiscourseElement",
"distant_citation": "cito:DistantCitation",
"entry": "biro:BibliographicReference",
"generic_format": "fabio:Manifestation",
"footnote": "doco:Footnote",
"funder_self_citation": "cito:FunderSelfCitation",
"inbook": "fabio:BookChapter",
"journal_self_citation": "cito:JournalSelfCitation",
"journal_cartel_citation": "cito:JournalCartelCitation",
"inproceedings": "fabio:ProceedingsPaper",
"merging": "prov:Replace",
"metadata_snapshot": "prov:Entity",
"document": "fabio:Expression",
"ocdm_dataset": "dcat:Dataset",
"ocdm_distribution": "dcat:Distribution",
"paragraph": "doco:Paragraph",
"patent": "fabio:PatentDocument",
"periodical_issue": "fabio:JournalIssue",
"periodical_volume": "fabio:JournalVolume",
"periodical_journal": "fabio:Journal",
"print_format": "fabio:PrintObject",
"proceedings": "fabio:AcademicProceedings",
"provenance_agent": "prov:Agent",
"reference_book": "fabio:ReferenceBook",
"reference_entry": "fabio:ReferenceEntry",
"reference_pointer": "c4o:InTextReferencePointer",
"role": "pro:RoleInTime",
"self_citation": "cito:SelfCitation",
"section": "doco:Section",
```

```
"section_title": "doco:SectionTitle",
"sentence": "doco:Sentence",
"series": "fabio:Series",
"pointer_list": "c4o:SingleLocationPointerList",
"note": "oa:Annotation",
"standard": "fabio:SpecificationDocument",
"techreport": "fabio:ReportDocument",
"thesis": "fabio:Thesis",
"table": "doco:Table",
"text_chunk": "doco:TextChunk",
"web": "fabio:WebContent",
"unpublished": "fabio:Preprint",
"unique_identifier": "datacite:Identifier",
"modification": "prov:Modify",

"note_content": { "@id": "oa:hasBody", "@type": "@vocab" },
"annotation": { "@id": "oco:hasAnnotation", "@type": "@vocab" },
"attributed_to": { "@id": "prov:wasAttributedTo", "@type": "@vocab" },
"citation": { "@id": "cito:cites", "@type": "@vocab" },
"characterization": { "@id": "cito:hasCitationCharacterization", "@type": "@vocab" },
"citing_document": { "@id": "cito:hasCitingEntity", "@type": "@vocab" },
"cited_document": { "@id": "cito:hasCitedEntity", "@type": "@vocab" },
"context_of": { "@id": "c4o:isContextOf", "@type": "@vocab" },
"contributor": { "@id": "pro:isDocumentContextFor", "@type": "@vocab" },
"crossref": { "@id": "biro:references", "@type": "@vocab"},
"curatorial_role_type": { "@id": "prov:hadRole" , "@type": "@vocab" },
"denoted_entry": { "@id": "c4o:denotes", "@type": "@vocab" },
"derived_from": { "@id": "prov:wasDerivedFrom", "@type": "@vocab" },
"distribution": { "@id": "dcat:distribution", "@type": "@vocab" },
"document_url": { "@id": "frbr:exemplar", "@type": "@vocab" },
"download": { "@id": "dcat:downloadURL", "@type": "@vocab" },
"pointer": { "@id": "co:element", "@type": "@vocab" },
"endpoint": { "@id": "void:sparqlEndpoint", "@type": "@vocab" },
"file_type": { "@id": "dcat:mediaType", "@type": "@vocab" },
"format": { "@id": "frbr:embodiment", "@type": "@vocab" },
"generated_by": { "@id": "prov:wasGeneratedBy", "@type": "@vocab" },
"held_by": { "@id": "prov:agent" , "@type": "@vocab" },
"identifier": { "@id": "datacite:hasIdentifier", "@type": "@vocab" },
"invalidated_by": { "@id": "prov:wasInvalidatedBy", "@type": "@vocab" },
"involved": { "@id": "prov:qualifiedAssociation", "@type": "@vocab" },
"license": { "@id": "dcterms:license", "@type": "@vocab" },
"mime_type": { "@id": "dcterms:format", "@type": "@vocab" },
"next": { "@id": "oco:hasNext", "@type": "@vocab" },
"reference": { "@id": "frbr:part", "@type": "@vocab" },
"part": { "@id": "frbr:part", "@type": "@vocab" },
"part_of": { "@id": "frbr:partOf", "@type": "@vocab" },
"related": { "@id": "dcterms:relation", "@type": "@vocab" },
"role_of": { "@id": "pro:isHeldBy", "@type": "@vocab" },
"role_type": { "@id": "pro:withRole", "@type": "@vocab" },
"snapshot_of": { "@id": "prov:specializationOf", "@type": "@vocab" },
"source": { "@id": "prov:hadPrimarySource", "@type": "@vocab" },
"subject": { "@id": "dcat:theme", "@type": "@vocab" },
"subset": { "@id": "void:subset", "@type": "@vocab" },
"type": { "@id": "datacite:usesIdentifierScheme", "@type": "@vocab" },
"webpage": { "@id": "dcat:landingPage", "@type": "@vocab" },

"byte": { "@id": "dcat:byteSize", "@type": "xsd:decimal" },
"citation_creation_date": "cito:hasCitationCreationDate",
"citation_time_span": { "@id": "cito:hasCitationTimeSpan", "@type": "xsd:duration" },
"date": "prism:publicationDate",
"description": "dcterms:description",
"edition": "prism:edition",
"fname": "foaf:familyName",
"fpage": "prism:startingPage",
"generated": { "@id": "prov:generatedAtTime", "@type": "xsd:dateTime" },
"gname": "foaf:givenName",
"id": "literal:hasLiteralValue",
"invalidated": { "@id": "prov:invalidatedAtTime", "@type": "xsd:dateTime" },
"keyword": "dcat:keyword",
"label": "rdfs:label",
"lpage": "prism:endingPage",
"mod_date": { "@id": "dcterms:modified", "@type": "xsd:dateTime" },
"name": "foaf:name",
```

```
"number": "fabio:hasSequenceIdentifier",
"pub_date": { "@id": "dcterms:issued", "@type": "xsd:dateTime" },
"content": "c4o:hasContent",
"subtitle": "fabio:hasSubtitle",
"title": "dcterms:title",
"update_action": "oco:hasUpdateQuery",

"ark": "datacite:ark",
"arxiv": "datacite:arxiv",
"author": "pro:author",
"bibliographic_database": "dbr:Bibliographic_database",
"cc0": "https://creativecommons.org/publicdomain/zero/1.0/legalcode",
"ccby": "https://creativecommons.org/licenses/by/4.0/legalcode",
"citations": "dbr:Citation",
"curator": "oco:occ-curator",
"dia": "datacite:dia",
"docx": "application:vnd.openxmlformats-officedocument.wordprocessingml.document",
"doi": "datacite:doi",
"ean13": "datacite:ean13",
"editor": "pro:editor",
"eissn": "datacite:eissn",

"f_agrees_with": "cito:agreesWith",
"f_cites_as_authority": "cito:citesAsAuthority",
"f_cites_as_data_source": "cito:citesAsDataSource",
"f_cites_as_evidence": "cito:citesAsEvidence",
"f_cites_as_metadata_document": "cito:citesAsMetadataDocument",
"f_cites_as_potential_solution": "cito:citesAsPotentialSolution",
"f_cites_as_recommended_reading": "cito:citesAsRecommendedReading",
"f_cites_as_related": "cito:citesAsRelated",
"f_cites_as_source_document": "cito:citesAsSourceDocument",
"f_cites_for_information": "cito:citesForInformation",
"f_compiles": "cito:compiles",
"f_confirms": "cito:confirms",
"f_contains_assertion_from": "cito:containsAssertionFrom",
"f_corrects": "cito:corrects",
"f_credits": "cito:credits",
"f_critiques": "cito:critiques",
"f_derides": "cito:derides",
"f_describes": "cito:describes",
"f_disagrees_with": "cito:disagreesWith",
"f_discusses": "cito:discusses",
"f_disputes": "cito:disputes",
"f_documents": "cito:documents",
"f_extends": "cito:extends",
"f_includes_excerpt_from": "cito:includesExcerptFrom",
"f_includes_quotation_from": "cito:includesQuotationFrom",
"f_links_to": "cito:linksTo",
"f_obtains_background_from": "cito:obtainsBackgroundFrom",
"f_obtains_support_from": "cito:obtainsSupportFrom",
"f_parodies": "cito:parodies",
"f_plagiarizes": "cito:plagiarizes",
"f_qualifies": "cito:qualifies",
"f_refutes": "cito:refutes",
"f_replies_to": "cito:repliesTo",
"f_retracts": "cito:retracts",
"f_reviews": "cito:reviews",
"f_ridicules": "cito:ridicules",
"f_speculates_on": "cito:speculatesOn",
"f_supports": "cito:supports",
"f_updates": "cito:updates",
"f_uses_conclusion_from": "cito:usesConclusionFrom",
"f_uses_data_from": "cito:usesDataFrom",
"f_uses_method_in": "cito:usesMethodIn",

"fundref": "datacite:fundref",
"handle": "datacite:handle",
"html": "text:html",
"infouri": "datacite:infouri",
"isbn": "datacite:isbn",
"isni": "datacite:isni",
"issn": "datacite:issn",
"lissn": "datacite:lissn",
```

```
    "istc": "datacite:istc",
    "json": "application:json",
    "jsonld": "application:ld+json",
    "jst": "datacite:jst",
    "localfunder": "datacite:local-funder-identifier-scheme",
    "localpersonal": "datacite:local-personal-identifier-scheme",
    "localresource": "datacite:local-resource-identifier-scheme",
    "lsid": "datacite:lsid",
    "nii": "datacite:nii",
    "nationalinsurancenumber": "datacite:national-insurance-number",
    "nihmsid": "datacite:nihmsid",
    "oci": "datacite:oci",
    "odt": "application:vnd.oasis.opendocument.text",
    "open_access": "dbr:Open_access",
    "openid": "datacite:openid",
    "orcid": "datacite:orcid",
    "pdf": "application:pdf",
    "pii": "datacite:pii",
    "plain": "text:plain",
    "pmcid": "datacite:pmcid",
    "pmid": "datacite:pmid",
    "metadata_provider": "oco:source-metadata-provider",
    "publisher": "pro:publisher",
    "purl": "datacite:purl",
    "rdfxml": "application:rdf+xml",
    "researcherid": "datacite:researcherid",
    "scholarly_communication": "dbr:Scholarly_communication",
    "sici": "datacite:sici",
    "social_security_number": "datacite:social-security-number",
    "turtle": "text:turtle",
    "upc": "datacite:upc",
    "uri": "datacite:uri",
    "url": "datacite:url",
    "urn": "datacite:urn",
    "viaf": "datacite:viaf",
    "xhtml": "application:xhtml+xml",
    "xpath": "datacite:local-resource-identifier-scheme",

    "year": "xsd:gYear",
    "year_month": "xsd:gYearMonth",
    "year_month_day": "xsd:date"
  }
}
```

Bibliographic resources and their metadata

The following excerpt shows how to linearize the information about a bibliographic resource into JSON-LD according to the aforementioned JSON-LD context document.

```
{
  "@context": "[JSON-LD context URL]",
  "iri": "gbr:04201",
  "a": [ "document", "article" ],
  "identifier": [
    {
      "iri": "gid:04201",
      "a": "unique_identifier",
      "id": "10.1108/jd-12-2013-0166",
      "type": "doi"
    },
    {
      "iri": "gid:04202",
      "a": "unique_identifier",
      "id": "http://www.emeraldinsight.com/doi/abs/10.1108/jd-12-2013-0166",
      "type": "url"
    },
    {
      "iri": "gid:04203",
      "a": "unique_identifier",
      "id": "http://dx.doi.org/10.1108/JD-12-2013-0166",
```

```
    "type": "url"
  }
],
"title": "Setting our bibliographic references free: towards open citation data",
"date": { "value": "2015", "a": "year" },
"related": "http://dx.doi.org/10.1108/jd-12-2013-0166",
"contributor": [
  {
    "iri": "gar:04201",
    "a": "role",
    "role_type": "author",
    "role_of": {
      "iri": "gra:04201",
      "a": "agent",
      "gname": "Silvio",
      "fname": "Peroni",
      "identifier": {
        "iri": "gid:04204",
        "a": "unique_identifier",
        "type": "orcid",
        "id": "0000-0003-0530-4305"
      },
      "related": "http://orcid.org/0000-0003-0530-4305"
    },
    "next": "gar:04202"
  },
  {
    "iri": "gar:04202",
    "a": "role",
    "role_type": "author",
    "role_of": {
      "iri": "gra:04202",
      "a": "agent",
      "gname": "Alexander",
      "fname": "Dutton"
    },
    "next": "gar:04203"
  },
  {
    "iri": "gar:04203",
    "a": "role",
    "role_type": "author",
    "role_of": {
      "iri": "gra:04203",
      "a": "agent",
      "gname": "Tanya",
      "fname": "Grey"
    },
    "next": "gar:04204"
  },
  {
    "iri": "gar:04204",
    "a": "role",
    "role_type": "author",
    "role_of": {
      "iri": "gra:04204",
      "a": "agent",
      "gname": "David",
      "fname": "Shotton",
      "related": "http://orcid.org/0000-0001-5506-523X"
    }
  },
  {
    "a": "role",
    "iri": "gar:04205",
    "role_type": "publisher",
    "role_of": {
      "iri": "gra:04205",
      "a": "agent",
      "name": "Emerald"
    }
  }
],
```

```
  "format": [
    {
      "iri": "gre:04201",
      "a": [ "generic_format", "digital_format"],
      "mime_type": "pdf",
      "fpage": "253",
      "lpage": "277",
      "document_url": "http://www.emeraldinsight.com/doi/pdfplus/10.1108/jd-12-2013-0166"
    },
    {
      "iri": "gre:04202",
      "a": [ "generic_format", "digital_format"],
      "mime_type": "html",
      "document_url": "http://www.emeraldinsight.com/doi/full/10.1108/jd-12-2013-0166"
    }
  ],
  "reference": [{
    "iri": "gbe:04201",
    "a": "entry",
    "content": "Agarwal, S., Choubey, L. and Yu, H. (2010), “Automatically classifying the role of
citations in biomedical articles”, Proceedings of the 2010 AMIA Annual Symposium, pp. 11-15.",
    "crossref": "gbr:04205"
  },
  {
    "iri": "gbe:04202",
    "a": "entry",
    "content": "Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D.
(2010). “Utopia documents: linking scholarly literature with research data”. Bioinformatics, 26(18):
i568- i574.",
    "crossref": "gbr:04206"
  },
  {
    "iri": "gbe:04203",
    "a": "entry",
    "content": "Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D.
(2009). “Calling International Rescue: knowledge lost in literature and data landslide!”.
Biochemical Journal, 424(3): 317–333.",
    "crossref": "gbr:04207"
  }],
  "part": [{
    "iri": "gde:04201",
    "a": ["discourse_element", "section"],
    "identifier": {
      "iri": "gid:04206",
      "a": "unique_identifier",
      "id": "/article/body/sec[2]",
      "type": "xpath"
    },
    "part": [{
      "iri": "gde:04202",
      "a": ["discourse_element", "sentence"],
      "identifier": {
        "iri": "gid:04207",
        "a": "unique_identifier",
        "id": "substring(string(/article/body/sec[2]/p[15]),223,259)",
        "type": "xpath"
      },
      "content": "In the biological field, Agarwal et al. (2010) introduces eight different
top-level classes describing different kinds of citations: background, contemporary, contrast,
evaluation, explanation of results, material and methods, modality and similarity [10,12].",
      "context_of": [
        {
          "iri": "grp:04207",
          "a": "reference_pointer",
          "identifier": {
            "iri": "gid:04208",
            "a": "unique_identifier",
            "id": "substring(string(/article/body/sec[2]/p[15]),248,21)",
            "type": "xpath"
          },
          "content": "Agarwal et al. (2010)",
          "denoted_entry": "gbe:04201"
        },
```

```
        {
          "iri": "gpl:04201",
          "a": "pointer_list",
          "identifier": {
            "iri": "gid:04209",
            "a": "unique_identifier",
            "id": "substring(string(/article/body/sec[2]/p[15]),473,7)",
            "type": "xpath"
          },
          "content": "[10,12]",
          "pointer": [
            {
              "iri": "grp:04208",
              "a": "reference_pointer",
              "denoted_entry": "gbe:04202",
              "next": "grp:04209"
            },
            {
              "iri": "grp:04209",
              "a": "reference_pointer",
              "denoted_entry": "gbe:04203"
            }
          ]
        }
      ]
    }]
  }],
  "part_of": {
    "iri": "gbr:04202",
    "a":  [ "document", "periodical_issue" ],
    "number": "2",
    "part_of": {
      "iri": "gbr:04203",
      "a":  [ "document", "periodical_volume" ],
      "number": "71",
      "part_of": {
        "iri": "gbr:04204",
        "a":  [ "document", "periodical_journal" ],
        "identifier": [
          {
            "iri": "gid:04205",
            "a": "unique_identifier",
            "id": "0022-0418",
            "type": "issn"
          }
        ],
        "title": "Journal of Documentation"
      }
    }
  },
  "citation": [
    {
      "iri": "gbr:04205",
      "a":  [ "document", "inproceedings" ],
      "title": "Automatically classifying the role of citations in biomedical articles",
      "date": { "value": "2010", "a": "year" },
      "format": [
        {
          "iri": "gre:04203",
          "a": "generic_format",
          "fpage": "11",
          "lpage": "15"
        }
      ],
      "part_of": {
        "iri": "gbr:04206",
        "a":  [ "document", "proceedings" ],
        "title": "Proceedings of the 2010 AMIA Annual Symposium"
      }
    }
  ]
}
```

Citations and reference annotations

The following excerpts show how to linearize the information about a citation between two papers into JSON-LD according to the aforementioned JSON-LD context document. In particular, there are two scenarios, namely: (1) when a citation is extracted from a list of bibliographic references, and there is no other information available on the in-text reference pointers denoting that bibliographic reference, and (2) when information on the in-text reference pointers denoting that bibliographic reference and their textual contexts within the citing entity is available.

Scenario 1

The following excerpt shows how to linearize the information about a citation when it is extracted from the list of bibliographic references, and there is no information on the reference pointers denoting that bibliographic reference.

```
{
  "@context": "[JSON-LD context URL]",
  "iri": "gci:04201-04205",
  "a": "citation_relationship",
  "citing_document": "gbr:04201",
  "cited_document": "gbr:04205",
  "citation_creation_date": { "value": "2015", "a": "year" },
  "citation_time_span": "P5Y",
  "identifier": {
    "iri": "gid:042012",
    "a": "unique_identifier",
    "id": "04201-04205",
    "type": "oci"
  }
}
```

Scenario 2

The following excerpt shows how to linearize the information about a reference annotation when information on both the related bibliographic reference and a reference pointer denoting that bibliographic reference is available.

```
{
  "@context": "[JSON-LD context URL]",
  "iri": "grp:04207",
  "a": "reference_pointer",
  "content": "Agarwal et al. 2010",
  "denoted_entry": "gbe:04201",
  "annotation": {
      "iri": "gan:04201",
      "a": "note",
      "note_content": {
        "iri": "gci:04201-04025/1”,
        "a": "citation_relationship",
        "citing_document": "gbr:04201",
        "cited_document": "gbr:04205",
        "citation_creation_date": { "value": "2015", "a": "year" },
        "citation_time_span": "P5Y",
        "characterization": "f_extends"
      }
  }
}
```

Datasets and distributions

The following excerpt shows how to linearize the information about a dataset. In the example below, the distributions and its related sub-datasets of a fictional dataset – i.e. Dataset following OCDM (DFO) – are linearised into JSON-LD according to the aforementioned JSON-LD context document.


```
{
  "@context": "[JSON-LD context URL]",
  "iri": "[dataset URL]",
  "a": "ocdm_dataset",
  "title": "The Dataset following OCDM (DFO)",
  "description": "The Dataset following OCDM (DFO) is an open repository of scholarly citation data
made available under a Creative Commons public domain dedication, which provides in RDF accurate
citation information (bibliographic references) harvested from the scholarly literature (described
using the SPAR Ontologies) that others may freely build upon, enhance and reuse for any purpose,
without restriction under copyright or database law.",
  "pub_date": "2016-02-01T00:00:00",
  "mod_date": "2016-04-01T00:00:00",
  "keyword": [
    "DFO",
    "Dataset following OCDM (DFO)",
    "SPAR Ontologies",
    "bibliographic references",
    "citations"
  ],
  "subject": [
    "scholarly_communication",
    "bibliographic_database",
    "open_access",
    "citations"
  ],
  "distribution": [
    {
      "iri": "gdi:04201",
      "a": "ocdm_distribution",
      "title": "The Dataset following OCDM: distribution in Turtle dated 3rd April 2016",
      "description": "The 3rd April 2016 distribution of the Dataset following OCDM (DFO) stored in
Turtle.",
      "pub_date": "2016-04-03T12:00:00",
      "license": "cc0",
      "download": "http://www.dfo.example/distribution/dfo-2016-04-03.ttl.zip",
      "file_type": "turtle",
      "byte": "14098371"
    }
  ],
  "webpage": "[Landing Page URL]",
  "subset": [
    {
      "iri": "gbr:",
      "a": "ocdm_dataset",
      "title": "Dataset following OCDM (DFO): Bibliographic Resource dataset",
      "description": "The Dataset following OCDM is an open repository of scholarly citation data
made available under a Creative Commons public domain dedication, which provides in RDF accurate
citation information (bibliographic references) harvested from the scholarly literature (described
using the SPAR Ontologies) that others may freely build upon, enhance and reuse for any purpose,
without restriction under copyright or database law. This sub-dataset contains all the
'bibliographic resource' resources.",
      "pub_date": "2016-02-01T00:00:00",
      "mod_date": "2016-03-29T00:00:00",
      "keyword": [
        "DFO",
        "Dataset following OCDM",
        "SPAR Ontologies",
        "bibliographic references",
        "citations",
        "bibliographic resource"
      ],
      "subject": [
        "scholarly_communication",
```

```
        "bibliographic_database",
        "open_access",
        "citations"
      ]
    }
  ],
  "endpoint": "https://w3id.org/dfo/sparql"
}
```

Provenance data

The following excerpt shows how to linearize the information about the provenance of a bibliographic resource into JSON-LD according to the aforementioned JSON-LD context document.

```
{
  "@context": "[JSON-LD context URL]",
  "iri": "gbr:04201/prov/se/2",
  "a": "metadata_snapshot",
  "generated": "2016-04-01T00:00:00",
  "snapshot_of": "gbr:04201",
  "derived_from": {
    "iri": "gbr:04201/prov/se/1",
    "a": "metadata_snapshot",
    "generated": "2016-02-01T00:00:00",
    "invalidated": "2016-04-01T00:00:00",
    "snapshot_of": "gbr:04201",
    "source": "http://api.crossref.org/works/10.1108/jd-12-2013-0166",
    "attributed_to": "https://orcid.org/0000-0003-0530-4305",
    "description": "The entity has been created."
  },
  "source": "https://doi.org/10.1108/jd-12-2013-0166",
  "attributed_to": "https://orcid.org/0000-0003-0530-4305",
  "description": "The field 'title' of the entity has been modified.",
  "update_action": "PREFIX gbr: <[dataset URL]br/> PREFIX dcterms: <http://purl.org/dc/terms/>
DELETE DATA { gbr:1 dcterms:title 'Setting our bibliographic references free: towards open citation
data' }; INSERT DATA { gbr:1 dcterms:title 'Setting Our Bibliographic References Free: Towards Open
Citation Data' }"
}
```