Metadata for the OpenCitations Corpus

Version 1.5, June 19, 2016

Publication date for this document: June 19, 2016

Version number of this document: 1.5

Previous version v1.4, published June 18, 2016

Authors

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

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

License

The Open Citations Corpus

The Open Citations Corpus (herewithin abbreviated "the corpus" or "OCC") is an open access corpus of scholarly citation data, namely information about the author-created bibliographic references present in publications that cite other publications. The Open Citations project has a persistent URL at w3id.org, https://w3id.org/oc, which resolves to our OCC server at http://opencitations.net. The OCC stores metadata relevant to these citations in RDF, specifically BibJSON[2] encoded as JSON-LD[3] and makes them available through a SPARQL endpoint and as downloadable datasets.

RDF resources in the Open Citations Corpus

Kinds of metadata

The OCC stores three levels of metadata:

- Corpus metadata
- Bibliographic entity metadata
- Provenance metadata

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

Corpus metadata

The Open Citations Corpus is itself a dataset, as are the contents of the individual entity classes within it, for example all the entries within the class **Bibliographic resource** (short: **br**). These datasets are described appropriately by means of standard vocabularies, such as

[1] https://creativecommons.org/licenses/by/4.0/legalcode
[2] http://okfnlabs.org/bibjson/
[3] https://www.w3.org/TR/json-ld/

the *Data Catalog Vocabulary*[4] and the *VoID Vocabulary*[5]. Such datasets can have particular distributions.

- **Dataset** (short: the related *entity short name* if appropriate, e.g. **br** for all the bibliographic resources); *none* in case of the main OCC dataset, i.e. the corpus itself): a set of collected information about something.
- **Distribution** (short: **di**): an accessible form of an OCC dataset, for example a downloadable file.

Bibliographic entity metadata

The following OCC bibliographic entities (short: **en**) are handled as proper RDF resources:

- **Bibliographic resource** (short: **br**): a bibliographic resource that cites/is cited by another bibliographic resource. Subclasses (extracted from CrossRef[6] Types[7]) include:
  - *Book*
  - Book chapter
  - *Book part*
  - *Book section*
  - *Book series*
  - *Book set*
  - Book track
  - Component
  - 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

  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). Using the Functional Requirements for Bibliographic Records (FRBR)[8] distinction between works, expressions, manifestations and items, these bibliographic resources are expressions of works, that may be manifested in physical (e.g. printed paper) or electronic form.

---

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

- **Resource embodiment** (short: **re**): the particular physical or digital format in which a bibliographic resource was made available by its publisher. Subclasses:
    - Digital embodiment
    - Print embodiment
- **Bibliographic entry** (short: **be**): the particular textual bibliographic entry ("a reference") occurring in the reference list (or elsewhere) within a bibliographic resource, that references another bibliographic resource.
- **Responsible agent** (short: **ra**): the agent 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.

Identifiers for bibliographic entities

All the aforementioned bibliographic entities **must** have a corpus identifier:

- The **corpus identifier** assigned to the entity upon initial curation into the OCC is the two-letter short name for the class of items (e.g. **be** for a bibliographic entry) followed by an oblique slash ("/") and a number assigned to each resource, unique among resources of the same type, which increments for each new entry in that resource class (e.g. "be/537"). Note that this identifier is for internal OCC use only, and is distinct from any "public" Internationalized Resource Identifier (abbreviated IRI) that may be used to identify the entity.

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

- **Identifier** (short: **id**): an external identifier (e.g. DOI[9], ORCID[10], PubMedID[11]) associated with the bibliographic entity. Members of this class of OCC metadata are themselves given unique numbers, e.g. "id/129".

Provenance metadata

All the aforementioned OCC bibliographic entities and identifiers **must** have metadata describing their provenance. These provenance metadata entities are:

- **Snapshot of entity metadata** (short: **se**): a particular snapshot recording the metadata associated with the entire corpus, a dataset within the corpus, or an individual entity (either a bibliographic entity or an identifier) at a particular time.
- **Curatorial activity** (short: **ca**): a curatorial activity relating to that entity. Possible activities in the are:
    - Creation: the activity of creating a new entity and of associating new metadata with it, within the corpus;
    - Modification: the activity of modifying (adding/removing) the metadata associated with an existing entity, or even of deprecating the entire entity;
    - Merging: the activity of unifying the metadata relating to two different OCC bibliographic entity descriptions, if they actually represent the same thing. This

[9] https://www.doi.org/
[10] http://orcid.org/
[11] http://www.ncbi.nlm.nih.gov/pubmed

can result in the deprecation of one of the corpus entries in favour of the other one.

- **Provenance agent** (short: **pa**): the agent, such as a person, organisation or process, that creates or modifies entity metadata, or that is used as source provider of those metadata (e.g. CrossRef).
- **Curatorial role** (short: **cr**): a particular role held by a provenance agent with respect to a curatorial activity (e.g. OCC curator, metadata source).

Naming convention for entities and provenance data

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

URLs for datasets and distributions

The URL identifying the corpus is the following:

[corpus URL] : [base URL]/corpus/

where the *base URL* has been chosen for guaranteeing persistency over time. The Open Citations project has a persistent URL at w3id.org, https://w3id.org/oc. Therefore, the URL of the Open Citations Corpus is:

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

The *corpus URL* identifies the main aggregated dataset, which is split in several sub-datasets, one for each kind of entity included in the corpus. The URLs of such sub-datasets follow the following schema:

[sub-dataset URL] : [corpus URL][entity short name]/

where the *entity short name* is that two-character abbreviation specified above for each of the entity classes within the corpus. For example, the URL of the dataset of all OCC bibliographic resources is:

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

The URL defining one or more distributions of the main dataset (i.e. the entire corpus) is:

[corpus distribution URL] : [corpus URL]di/[iterative number]

where the *iterative 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 corpus is:

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

Similarly, the URL defining a distribution of any of the corpus sub-datasets is:

[sub-dataset distribution URL]: [sub-dataset URL]di/[iterative number]

All the distributions of a dataset must be assigned to the relevant distribution dataset (e.g. within the OCC, "https://w3id.org/oc/corpus/di/1" is stored in the dataset graph "https://w3id.org/oc/corpus/di/").

URLs for bibliographic entities and their identifiers

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

[entity URL] : [corpus URL][entity short name]/[iterative number]

where corpus URL, entity short name, and iterative number are as previously defined. For example, the third entry within the OCC class of bibliographic resources, and the 129th entry within the OCC class of identifiers, have the following URLs respectively:

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

All these entities must be assigned to the dataset related to the entity class (e.g. within the OCC, "https://w3id.org/oc/corpus/br/3" is stored in the bibliographic resource dataset graph "https://w3id.org/oc/corpus/br/").

URLs for provenance metadata

Each of the OCC bibliographic entities and identifiers has associated with it a particular provenance RDF 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 only exception to the aforementioned graph URL construction concerns the graph containing provenance information about provenance agents (curators, metadata providers, etc.), since they can be involved in several curatorial activities of different bibliographic entities or identifiers within the OCC and, thus, are not necessary tied to one specific entity. For this reason, we store provenance agent metadata in the more appropriate general provenance graph, namely:

[corpus provenance URL]: [corpus URL]prov/

OCC provenance metadata entities (i.e. snapshots, curatorial activities, provenance agents and curatorial roles) relating to a particular OCC bibliographic entity or identifier use the following convention for their URLs:

[provenance agent URL] : [corpus provenance URL]pa/[iterative number]

[other provenance metadata entity URL] :
[entity provenance URL][provenance metadata entity short name]/[iterative number]

where *provenance metadata entity short name* and *iterative number* are assigned as explained for the other entities in the corpus.

For example, the second curatorial activity related to the fifteenth bibliographic resource and the third provenance agent involved in that curatorial activity have the following URLs:

- https://w3id.org/oc/corpus/br/15/prov/ca/2
- https://w3id.org/oc/corpus/prov/pa/3

Please note that all the provenance entities (except the information about the provenance agents, such as their names) are assigned to the provenance dataset graph associated with the entity of the corpus for which they provide provenance information (e.g. "https://w3id.org/oc/corpus/br/15/prov/ca/2" is stored in provenance graph "https://w3id.org/oc/corpus/br/15/prov/"). This has been done 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 graph.

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 OCC dataset (graph: https://w3id.org/oc/corpus/[entity short name]/)

- 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 publication of a particular dataset by the OCC.
- has modification date: *date*
  The date describing when the dataset has been modified.
- has keyword: *literal*
  A keyword 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.

Metadata elements that may be associated with the main OCC dataset (graph: https://w3id.org/oc/corpus/)

All the attributes for datasets defined in the previous section, plus the following ones:

- has landing page: *document*
  An HTML page (indicated by its URL) representing a browseable page for the corpus.
- has sub-dataset: *dataset*
  A link to a subset of the whole corpus dataset.
- has SPARQL endpoint: *URL*
  The link to the SPARQL endpoint for querying the corpus.

Metadata elements that may be associated with a distribution (graph: https://w3id.org/oc/corpus/[entity short name or *none* for the main corpus]/di/)

- 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 first date of publication of the distribution.
- has license: *document*
  The resource describing the licence associated with the data in the distribution.
- has download URL: *document*
  The resource which is the representation of the distribution in a certain format.
- has file type: *media type*
  The file type of the downloadable representation of the distribution (according to IANA media types).
- has byte size: *literal*
  The size in bytes of the downloadable 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 OCC bibliographic entities.

Metadata elements that may be associated with any OCC bibliographic entity

- has identifier: *identifier*
  In addition to the internal **corpus identifier** assigned to the entity upon initial curation into the OCC (format: [entity short name]/[iterative number], 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 (graph: https://w3id.org/oc/corpus/br/)

- has type: *thing*
  The type of the bibliographic resource, conforming to those introduced above.
- has title: *literal*
  The title 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 entry (be)*
  The literal text of a reference within the bibliographic resource
- has publication year: *gYear*
  The year 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*
  The number identifying 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.

Metadata elements that may be associated with a responsible agent's role (graph: https://w3id.org/oc/corpus/ar/)

- 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 agents' roles of the same type associated with the same bibliographic resource (so as to define, for instance, its ordered list of authors).

Metadata elements that may be associated with a responsible agent (graph: https://w3id.org/oc/corpus/ra/)

- 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.

Metadata elements that may be associated with a resource embodiment (graph: https://w3id.org/oc/corpus/re/)

- 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 entry (graph: https://w3id.org/oc/corpus/be/)

- has bibliographic entry text: *literal*
  The literal text of a bibliographic entry (i.e. a 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 corpus identifier of the cited bibliographic resource to which this bibliographic entry relates.

Metadata elements associated with identifiers

In this section we introduce all the metadata elements with which an external third-party identifier of an OCC bibliographic entity may be associated.

Metadata elements associated with an identifier (graph: https://w3id.org/oc/corpus/id/)

- has literal value: *literal*
  The string representing the identifier (e.g. 10.1987/4567.98).
- has scheme: *thing*
  The particular identifier scheme to which the identifier belongs (e.g. DOI).

Provenance information

Each of the aforementioned bibliographic entities introduced into the corpus has associated provenance information that documents the curatorial processes that have led to the current OCC description of that resource. In this section we introduce all the provenance metadata elements that constitute the provenance information for a particular OCC 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) (graph: [entity provenance URL])

- has creation date: *date time*
  The date on which a particular snapshot of a bibliographic entity's metadata was created within the OCC.
- 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. the addition of some metadata that was not specified in the previous snapshot) or a merger with another one.
- is snapshot of: *bibliographic entity (en)*
  This property is used to link a snapshot of entity metadata to the bibliographic entity in the OCC 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. the result of querying the CrossRef API).
- is generated by: *curatorial activity (ca)*
  This property is used to specify the curatorial activity whereby the snapshot of entity metadata entity was generated.
- is invalidated by: *curatorial activity (ca)*
  This property is used to specify the curatorial activity whereby the snapshot of entity metadata entity was invalidated, i.e. the reason for the invalidation.

Metadata elements that may be associated with a curatorial activity (ca) (graph: [entity provenance URL])

- has type: *thing*
  The type of OCC curatorial activity, conforming to one of those defined above (creation, modification or merging).
- has description: *literal*
  A textual description of the activity and its consequence.
- has update action: *thing*
  The UPDATE SPARQL query that keeps track of which metadata have been modified as the result of a modification of some of the metadata or the merging of the metadata relating to a particular bibliographic entity.
- involves agent with role: *curatorial role (cr)*
  The curatorial role of the provenance agent involved in this curatorial activity.

Metadata elements that may be associated with a curatorial role (cr) (graph: [entity provenance URL])

- has role type: *thing*
  The specific type of role under consideration (e.g. the merging activity of an OCC curator, or an external authority acting as a metadata source).
- held by agent: *provenance agent (pa)*
  The provenance agent (OCC curator or external authority) holding that curatorial role.

Metadata elements that may be associated with a provenance agent (pa) (graph: [entity provenance URL])

- has name string: *literal*
  The name of a provenance 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 a provenance agent, if a person.
- has family name: *literal*
  The family name of a provenance agent, if a person.

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 used in the Open Citations Corpus using OWL ontologies, in particular the Semantic Publishing and Referencing (SPAR) Ontologies[12], the well-known Web, library and publishing vocabularies Dublin Core[13], FRBR[14],

[12] http://www.sparontologies.net
[13] http://dublincore.org/documents/dcmi-terms/
[14] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

PRISM[15] and RDF[16], and the following additional models: DCAT[17], FOAF[18], Literal Reification[19], OCO[20], PROV-O[21], PROV-DC[22], and VOID[23].

The following prefixes are employed:

```
biro:       http://purl.org/spar/biro/
cito:       http://purl.org/spar/cito/
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/
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/
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 entry: biro:BibliographicReference
- Responsible agent: foaf:Agent
- Agent role: pro:RoleInTime
- Bibliographic resource: fabio:Expression
  Subclasses:
  - 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

---

15 http://www.idealliance.org/specifications/prism-metadata-initiative
16 https://www.w3.org/TR/rdf11-concepts/
17 http://www.w3.org/TR/vocab-dcat
18 http://xmlns.com/foaf/spec/
19 http://ontologydesignpatterns.org/wiki/Submissions:Literal_Reification
20 https://w3id.org/oc/ontology
21 http://www.w3.org/TR/prov-o
22 http://www.w3.org/TR/prov-dc
23 http://www.w3.org/TR/void

    - Book track fabio:Expression (part of a fabio:ExpressionCollection)
    - 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
- Resource embodiment: fabio:Manifestation
  Subclasses:
    - Digital embodiment fabio:DigitalManifestation
    - Print embodiment fabio:PrintObject

Identifier

- Identifier: datacite:Identifier

Provenance data

- Snapshot of entity metadata: prov:Entity
- Curatorial activity: prov:Activity
  Subclasses:
    - Creation: prov:Create
    - Modification: prov:Modify
    - Merging: prov:Replace
- Provenance agent: prov:Agent
- Curatorial role: prov:Association

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 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):

- 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 entry

- has bibliographic entry text: c4o:hasContent
- references: biro:references

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

Bibliographic resource

- has type: rdf:type
- has title: dcterms:title
- is part of: frbr:partOf
- cites: cito:cites
- has publication year: fabio:hasPublicationYear
- is embodied as: frbr:embodiment
- has number: fabio:hasSequenceIdentifier
- has edition: prism:edition
- has part: frbr:part
- has contributor: pro:isDocumentContextFor

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 generated by: prov:wasGeneratedBy
- is invalidated by: prov:wasInvalidatedBy

Curatorial activity

- has type: rdf:type
- involves agent with role: prov:qualifiedAssociation
- has description: dcterms:description
- has update action oco:hasUpdateQuery

Curatorial role

- has role type: prov:hadRole
- held by agent: prov:agent

Provenance agent

- has name: foaf:name
- has given name: foaf:givenName
- has family name: foaf:familyName

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

The *OCC Context* (http://w3id.org/oc/corpus/context.json) is a mapping document that formally maps terms used in the OCC's JSON-LD files to the entities defined in the various ontologies used for describing OCC data in RDF. The OCC Context is defined as follows.

```
{
  "@context": {
    "@base": "https://w3id.org/oc/",

    "gocc": "https://w3id.org/oc/corpus/",
    "gar": "https://w3id.org/oc/corpus/ar/",
    "gbe": "https://w3id.org/oc/corpus/be/",
    "gbr": "https://w3id.org/oc/corpus/br/",
    "gcr": "https://w3id.org/oc/corpus/cr/",
    "gid": "https://w3id.org/oc/corpus/id/",
    "gra": "https://w3id.org/oc/corpus/ra/",
    "gre": "https://w3id.org/oc/corpus/re/",
    "gdi": "https://w3id.org/oc/corpus/di/",

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

```
"agent": "foaf:Agent",
"article": "fabio:JournalArticle",
"book": "fabio:Book",
"book_part": "doco:Part",
"book_section": "fabio:ExpressionCollection",
"book_series": "fabio:BookSeries",
"book_set": "fabio:BookSet",
"creation": "prov:Create",
"curatorial_activity": "prov:Activity",
"curatorial_role": "prov:Association",
"dataset": "fabio:DataFile",
"digital_format": "fabio:DigitalManifestation",
"generic_format": "fabio:Manifestation",
"entry": "biro:BibliographicReference",
"inbook": "fabio:BookChapter",
"inproceedings": "fabio:ProceedingsPaper",
"merging": "prov:Replace",
"metadata_snapshot": "prov:Entity",
"document": "fabio:Expression",
"monograph": "fabio:Book",
"occ_dataset": "dcat:Dataset",
"occ_distribution": "dcat:Distribution",
"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",
"role": "pro:RoleInTime",
"series": "fabio:Series",
"standard": "fabio:SpecificationDocument",
"techreport": "fabio:ReportDocument",
"thesis": "fabio:Thesis",
"web": "fabio:WebContent",
"unpublished": "fabio:Preprint",
"unique_identifier": "datacite:Identifier",
"modification": "prov:Modify",

"citation": { "@id": "cito:cites", "@type": "@vocab" },
"contributor": { "@id": "pro:isDocumentContextFor", "@type": "@vocab" },
"crossref": { "@id": "biro:references", "@type": "@vocab"},
"curatorial_role_type": { "@id": "prov:hadRole" , "@type": "@vocab" },
"derived_from": { "@id": "prov:wasDerivedFrom", "@type": "@vocab" },
"distribution": { "@id": "dcat:distribution", "@type": "@vocab" },
"download": { "@id": "dcat:downloadURL", "@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" },
"role_of": { "@id": "pro:isHeldBy", "@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_of": { "@id": "frbr:partOf", "@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" },
"document_url": { "@id": "frbr:exemplar", "@type": "@vocab" },
"webpage": { "@id": "dcat:landingPage", "@type": "@vocab" },

"byte": { "@id": "dcat:byteSize", "@type": "xsd:decimal" },
"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",
    "title": "dcterms:title",
    "year": { "@id": "fabio:hasPublicationYear", "@type": "xsd:gYear" },
    "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",
    "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",
    "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",
    "occ": "datacite:occ",
    "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"
  }
}
```

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 mapping document (i.e. the OCC Context).

```
{
    "@context": "https://w3id.org/oc/corpus/context.json",
    "iri": "gbr:1",
    "a": "article",
    "identifier": [
        {
            "iri": "gid:1",
            "a": "unique_identifier",
            "id": "br/1",
            "type": "occ"
        },
        {
            "iri": "gid:2",
            "a": "unique_identifier",
            "id": "10.1108/JD-12-2013-0166",
            "type": "doi"
        },
        {
            "iri": "gid:3",
            "a": "unique_identifier",
            "id": "http://www.emeraldinsight.com/doi/abs/10.1108/JD-12-2013-0166",
            "type": "url"
        },
        {
            "iri": "gid:4",
            "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",
    "year": "2015",
    "format": [
        {
            "iri": "gre:1",
            "a": "digital_format",
            "identifier": {
                "iri": "gid:5",
                "a": "unique_identifier",
                "id": "re/1",
                "type": "occ"
            },
            "mime_type": "pdf",
            "fpage": "253",
            "lpage": "277",
            "document_url": "http://www.emeraldinsight.com/doi/pdfplus/10.1108/JD-12-2013-0166"
        },
        {
            "iri": "gre:2",
            "a": "digital_format",
            "identifier": {
                "iri": "gid:6",
                "a": "unique_identifier",
                "id": "re/2",
                "type": "occ"
            },
            "mime_type": "html",
            "document_url": "http://www.emeraldinsight.com/doi/full/10.1108/JD-12-2013-0166"
        }
    ],
    "reference": {
        "iri": "gbe:1",
        "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.",
        "cross_reference": "gbr:5"
    },
    "part_of": {
        "iri": "gbr:2",
        "a": "periodical_issue",
        "identifier": {
            "iri": "gid:7",
            "a": "unique_identifier",
            "id": "br/2",
            "type": "occ"
        },
        "number": "2",
```

```
        "part_of": {
            "iri": "gbr:3",
            "a": "periodical_volume",
            "identifier": {
                "iri": "gid:8",
                "a": "unique_identifier",
                "id": "br/3",
                "type": "occ"
            },
            "number": "71",
            "part_of": {
                "iri": "gbr:4",
                "a": "periodical_journal",
                "identifier": [
                    {
                        "iri": "gid:9",
                        "a": "unique_identifier",
                        "id": "br/4",
                        "type": "occ"
                    },
                    {
                        "iri": "gid:10",
                        "a": "unique_identifier",
                        "id": "0022-0418",
                        "type": "issn"
                    }
                ],
                "title": "Journal of Documentation"
            }
        }
    },
"citation": [
        {
            "iri": "gbr:5",
            "a": "inproceedings",
            "identifier": [
                {
                    "iri": "gid:11",
                    "a": "unique_identifier",
                    "id": "br/5",
                    "type": "occ"
                }
            ],
            "title": "Automatically classifying the role of citations in biomedical articles",
            "year": "2010",
            "format": [
                {
                    "iri": "gre:3",
                    "a": "generic_format",
                    "identifier": {
                        "iri": "gid:12",
                        "a": "unique_identifier",
                        "id": "re/3",
                        "type": "occ"
                    },
                    "fpage": "11",
                    "lpage": "15"
                }
            ],
            "part_of": {
                "iri": "gbr:10",
                "a": "proceedings",
                "identifier": {
                    "iri": "gid:13",
                    "a": "unique_identifier",
                    "id": "br/10",
                    "type": "occ"
                },
                "title": "Proceedings of the 2010 AMIA Annual Symposium"
            }
        }
    ]
}
```

Datasets and distributions

The following excerpt shows how to linearize the information about the OCC, its distributions and its related sub-datasets into JSON-LD according to the aforementioned mapping document (i.e. the OCC Context).


```
{
    "@context": "https://w3id.org/oc/corpus/context.json",
    "iri": "gocc:",
    "a": "occ_dataset",
    "label": "OCC",
    "title": "The OpenCitations Corpus",
    "description": "The OpenCitations Corpus 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": [
        "OCC",
        "OpenCitations",
        "OpenCitations Corpus",
        "SPAR Ontologies",
        "bibliographic references",
        "citations"
    ],
    "subject": [
        "scholarly_communication",
        "bibliographic_database",
        "open_access",
        "citations"
    ],
    "distribution": [
        {
            "iri": "gdi:1",
            "a": "occ_distribution",
            "label": "distribution 1 of OCC [di/1 - OCC]",
            "title": "The Open Citations Corpus: distribution in Turtle dated 3rd April 2016",
            "description": "The 3rd April 2016 distribution of the Open Citations Corpus (OCC)
stored in Turtle.",
            "pub_date": "2016-04-03T12:00:00",
            "license": "cc0",
            "download": "http://www.opencitations.net/static/distribution/occ-2016-04-03.ttl.zip",
            "file_type": "turtle",
            "byte": "14098371"
        }
    ],
    "webpage": "http://opencitations.net/",
    "subset": [
        {
            "iri": "gbr:",
            "a": "occ_dataset",
            "label": "OCC / br",
            "title": "The Open Citations Corpus: Bibliographic Resource dataset",
            "description": "The OpenCitations Corpus 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": [
                "OCC",
                "OpenCitations Corpus",
                "OpenCitations",
                "SPAR Ontologies",
                "bibliographic references",
                "citations",
                "bibliographic resource"
            ],
            "subject": [
                "scholarly_communication",
                "bibliographic_database",
                "open_access",
                "citations"
            ]
```

```
        }
    ],
    "endpoint": "https://w3id.org/oc/corpus/sparql"
}
```

Provenance data

The following excerpt shows how to linearize the information about the provenance of a bibliographic entity contained in the OCC into JSON-LD according to the aforementioned mapping document (i.e. the OCC Context).

```
{
    "@context": "https://w3id.org/oc/corpus/context.json",
    "iri": "gbr:1/prov/se/2",
    "a": "metadata_snapshot",
    "label": "snapshot of entity metadata 2 related to bibliographic resource 1 [se/2 -> br/1]",
    "snapshot_of": "gbr:1",
    "generated": "2016-04-01T00:00:00",
    "generated_by": {
        "iri": "gbr:1/prov/ca/2",
        "a": ["curatorial_activity", "modification"],
        "label": "curatorial activity 2 related to bibliographic resource 1 [ca/2 -> br/1]",
        "involved": {
            "iri": "gbr:1/prov/cr/3",
            "a": "curatorial_role",
            "label": "curatorial role 3 related to bibliographic resource 1 [cr/3 -> br/1]",
            "curatorial_role_type": "curator",
            "held_by": {
                "iri": "gpa:3",
                "a": "provenance_agent",
                "name": "Silvio Peroni"
            }
        },
        "description": "The field 'title' of the entity 'https://w3id.org/oc/corpus/br/1' has been
modified.",
        "update_action": "DELETE DATA { GRAPH <https://w3id.org/oc/corpus/br/> {
<https://w3id.org/oc/corpus/br/1> <http://purl.org/dc/terms/title> 'Setting our bibliographic
references free: towards open citation data' } }; INSERT DATA { GRAPH
<https://w3id.org/oc/corpus/br/> { <https://w3id.org/oc/corpus/br/1>
<http://purl.org/dc/terms/title> 'Setting Our Bibliographic References Free: Towards Open Citation
Data' } }"
    },
    "derived_from": [
        {
            "iri": "gbr:1/prov/se/1",
            "a": "metadata_snapshot",
            "label": "snapshot of entity metadata 1 related to bibliographic resource 1 [se/1 ->
br/1]",
            "snapshot_of": "gbr:1",
            "generated": "2016-02-01T00:00:00",
            "generated_by": {
                "iri": "gbr:1/prov/ca/1",
                "a": ["curatorial_activity", "creation"],
                "label": "curatorial activity 1 related to bibliographic resource 1 [ca/1 -> br/1]",
                "involved": [
                    {
                        "iri": "gbr:1/prov/cr/1",
                        "a": "curatorial_role",
                        "label": "curatorial role 1 related to bibliographic resource 1 [cr/1 ->
br/1]",
                        "curatorial_role_type": "metadata_provider",
                        "held_by": {
                            "iri": "gpa:1",
                            "a": "provenance_agent",
                            "name": "CrossRef"
                        }
                    },
                    {
                        "iri": "gbr:1/prov/cr/2",
                        "a": "curatorial_role",
                        "label": "curatorial role 2 related to bibliographic resource 1 [cr/2 ->
br/1]",
                        "curatorial_role_type": "curator",
                        "held_by": {
                            "iri": "gpa:2",
                            "a": "provenance_agent",
                            "name": "SPACIN CrossrefProcessor"
```

```
                        }
                    }
                ],
                "description": "The entity 'https://w3id.org/oc/corpus/br/1' has been created."
            },
            "source": "http://api.crossref.org/works/10.1108/JD-12-2013-0166",
            "invalidated": "2016-04-01T00:00:00",
            "invalidated_by": "gbr:1/prov/ca/2"
        }
    ]
}
```