Open Citation Identifier: Definition

Version 1.0, 5 October 2018

Publication date for this document: 5 October 2018
Version number of this document: 1.0
DOI of this version: https://doi.org/10.6084/m9.figshare.7127816.v1
DOI of the last version: https://doi.org/10.6084/m9.figshare.7127816

Authors

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

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

Cite this as

Silvio Peroni, David Shotton (2018). Open Citation Identifier: Definition. Figshare.
DOI: https://doi.org/10.6084/m9.figshare.7127816

Preamble

The key word “MUST” in this definition is to be interpreted as described in Request For Comments 2119, meaning that the item or condition referred to is an absolute requirement of the specification.

Preliminaries

A *bibliographic citation* is a conceptual directional link from a citing entity to a cited entity, for the purpose of acknowledging or ascribing credit for the contribution made by the author(s) of the cited entity. The citing and cited entities may be scholarly publications, on-line documents, blog posts, datasets, or any other authored entities capable of giving or receiving citations.

A citation is created by the performative act of making a citation, typically by the inclusion of a bibliographic reference in the reference list of the citing entity or in one of its footnotes, or by the inclusion within the citing entity of a database accession number or of link, in the form of an HTTP Uniform Resource Locator (URL), to a resource on the World Wide Web. While the act of citation by the author may be the work of a moment, the citation itself, once the citing work is published, becomes an enduring component of the academic ecosystem.

Citations play a major part in knitting together independent works of scholarship into a global endeavour. Analyses of citations can both reveal how scholarly knowledge develops over time and also be used to assess scholars' influence and make wise decisions about research investment.

Metadata describing a citation can be stored in a database (or in other kinds of storages) containing bibliographic information about or links to the citing and/or cited entity.

While citations are normally treated simply as the links between published entities, a recent proposal by OpenCitations is to provide an alternative richer view: to regard each citation as a data entity in its own right. The main advantages of treating citations as first-class data entities are:

1. all the information regarding each citation can be stored in one place;
2. citations become easier to describe, distinguish, count and process;
3. if available in aggregate, citations described in this manner are easier to analyze using bibliometric methods, for example to visualize citation networks or to determine how citation time spans vary by discipline;
4. citations can be identified by means of globally unique and persistent citation identifiers. Such citation identifiers can be used to reference the citations contained within a specific bibliographic database.

The scope of this document is the definition of one particular identifier for citations, the Open Citation Identifier (OCI), which can be used to provide identifiers to *open citations*[1].

[1] For a detailed description of what constitutes an open citation, please read "*Silvio Peroni, David Shotton (2018). Open Citation: Definition. Figshare. DOI: https://doi.org/10.6084/m9.figshare.6683855*".

Definition of an Open Citation Identifier

The **Open Citation Identifier** (**OCI**) is a globally unique persistent identifier (PID) for the identification of **open** bibliographic citations stored in a specific database or in other kinds of storages. What is meant by an **open** bibliographic citation is defined at https://doi.org/10.6084/m9.figshare.6683855. The Open Citation Identifier system, and a resolution service for OCIs that returns metadata about the identified citations, have been developed by OpenCitations, a scholarly infrastructure organization dedicated to open scholarship and the publication of open bibliographic and citation data.

Each OCI has a simple structure: the lower-case letters “oci” followed by a colon, followed by two sequences of numerals separated by a dash. Precisely, an OCI is defined as follows (in Backus-Naur form):

```
<oci>        ::= "oci:" <identifier> "-" <identifier>
<identifier> ::= <pos_number> | <pos_number> <any_number> |
                 <prefix> <any_number>
<pos_number> ::= "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"
|
                 <pos_number> <pos_number>
<any_number> ::= "0" | <pos_number> | <any_number> <any_number>
<prefix>     ::= "0" <pos_number> "0"
```

If an `<identifier>` starts with “0” then a `<prefix>` is always defined. In addition, if `<prefix>` is used, it MUST be exactly the same in both `<identifier>`s.

For example, `oci:1-18`, `oci:2544384-7295288`, `oci:01027931310-01022252312`, and `oci:020010108063601070506630807020263066305090-020010108063601070506630807020263056303010` are all valid OCIs, while `oci:0123456-123456` is not.

Meaning

In each OCI, the first `<identifier>` is the identifier for the citing bibliographic resource, while the second `<identifier>` is the identifier for the cited bibliographic resource. Considering a particular database containing citations, each OCI referring to one of the citations within that database MUST carry the following two pieces of information:

1. specification of the supplier's database in which the bibliographic citation is recorded, and
2. the encoded numerical equivalent of (part of) the identifier used in the database that uniquely identifies the citing or the cited entity.

The first point is encoded by the `<prefix>` part of each half of the OCI, which MUST be the same for both halves of the OCI, and which it is assigned by OpenCitations. Please see the following section for more information about existing prefixes.

The second point is more nuanced. As a mandatory requirement for the assignment of OCIs to its content, any supplier's database containing citations MUST clearly identify the citing and cited entities of each citation by assigning them a particular internal identifier compliant with **one and only one** identifier scheme. Examples of such identifier schemes are given by the following bibliographic databases:

- the OpenCitations Corpus uses URLs following the pattern "https://w3id.org/oc/corpus/br/<unique sequence of numbers>" to identify bibliographic entities;
- Wikidata uses URLs following the pattern "http://www.wikidata.org/entity/Q<unique sequence of numbers>";
- Crossref uses full Digital Object Identifiers (DOIs).

The sequence of number in each `<identifier>` part of an OCI is a numerical encoding of the original supplier's identifier for the citing/cited entity. For instance:

- The OpenCitations Corpus has no prefix. The sequence of numbers (i.e. the *local identifiers* for the entities within the Corpus between which the citation exists) are used to create the OCI – e.g. `oci:1-18` is the OCI for the citation within the OpenCitations Corpus from https://w3id.org/oc/corpus/br/1 to https://w3id.org/oc/corpus/br/18.
- The prefix "010" is assigned to Wikidata. As in the OpenCitations Corpus, the unique numbers in the URL of the citing and cited entities are used to create the OCI – e.g. `oci:01027931310-01022252312` is the OCI for the citation from http://www.wikidata.org/entity/Q27931310 to http://www.wikidata.org/entity/Q22252312.
- The prefix "020" is assigned to Crossref. In this case, the sub-DOIs (i.e. the full DOI alphanumeric string omitting the initial "10.") of the citing and cited entities are translated into the two numerical sequences used in the OCI, according to this simple lookup table – e.g. `oci:020010108063601070506630807020263066305 09-02001010806360107`

`0506630807020263056303 01` represents a citation between the two bibliographic resources identified in the Crossref database by the DOIs 10.1186/1756-8722-6-59 and 10.1186/1756-8722-5-31.

Thus, for every prefix identifying an individual supplier of bibliographic and citation information, a particular conversion rule is applied to move from the unique identifiers used within the supplier's database to the numerical identifiers of the citing and cited entities used in the OCI to identify that citation. Reciprocally, this conversion rule MUST also allow one to recover the original identifier employed within the supplier's database when starting from its encoded version available in an OCI.

Prefixes

The prefix used for building an OCI defines the supplier of the original citation data, which is typically a database containing bibliographic information. The list of currently supported suppliers is maintained by OpenCitations and is available at https://github.com/opencitations/oci/blob/master/suppliers.csv. This list will be expanded to include other bibliographic databases in due time.

Software

OpenCitations provides an Open Citation Identifier Resolution Service that takes an OCI and returns information about the identified citation. The service is available at https://w3id.org/oc/oci. The prefixes currently handled by the Open Citation Identifier Resolution Service are mentioned in its webpage, this prefix list being updated every time the Resolution Service is expanded to handle the open citations within a new bibliographic database.

In addition to that, a script has been developed for validating a given OCI and for retrieving the citation data related to the citation identified such OCI. The script, called `oci.py`, is available on the OCI repository on GitHub.