Metadata requirements for the Open Citation Corpus

The Open Citations Corpus records citations between citing and cited scholarly works. A 'work' may be a research article appearing as a preprint or a journal article, a book, a Wikipedia entry, a dataset in a data repository, a software application or computational model, or any other type of scholarly output.

We envisage one publication record for each citing work recorded in the Open Citation Corpus, and one publication record for each of the works cited in the reference lists of the citing publications.

For each work we need to record the metadata items listed below, if available:

**Bold text metadata items** should be available for both citing and cited works.

Normal text metadata items may only be available for citing works, unless the information for cited works can be drawn from elsewhere (e.g. a separate submission of that paper to the OCC)

*Italic text metadata items* may not be available even for the citing work.

*Identifiers and links*

**IntPubID** Each work will have an internal Open Citation Corpus database identifier

ExtPubID External identifiers for the work – as many as are available – e.g. DOI, ArXiV ID, PubMedCentral ID, PubMed ID, GenBank database accession number

SourceURI A link resolving to the work itself (or its landing page)

*VoRURI* A link, recorded for a preprint, resolving to the Version of Record journal article at the publisher's web site

*RelatedURI* A link to any related work, for example from an article to a supplementary information file, or from a dataset to a related journal article

*Provenance information*

Source Source of the metadata about the citing work, e.g. ArXiV, PubMed Central, Dryad

DepositDate Date on which the information was entered from the source into OCC

DepositCurator Name of person responsible for the initial deposit in OCC

*UpdateDate* Date on which the information initially entered into OCC was subsequently updated

*UpdateCurator* Name of person responsible for updating the publication's metadata in OCC

*Nature of publication*

Type Nature of the work, e.g. preprint, journal article, conference paper, dataset, model

*Language* The language in which the work is written (e.g. English, C++)

Format The format in which the work is published, e.g. txt, pdf, html, latex, xml, ps

Rights Details of copyright and publication license

*Description of publication*

**Title** Title of the work

*Keywords* Keywords of relevance to the subject matter of the work

*Abstract* The Abstract from the work, as a text string

*Bibliographic details* (there is intentional overlap in this list)

**BibCitation** Full bibliographic citation of the work, as a text string (see below)

**AuthorList** An ordered list of authors of the work

*EditorList* An ordered list of editors of the work, or of the larger work (e.g. conference proceedings) in which the particular work was published, if relevant

**PubYear** Publication year for the work

PubDate Publication date for the work, as yyyy-mm-dd or yyyy-mm, with sub-records:

*PrintPubDate* Publication date for the print version of the work, if known and if relevant

*OnlinePubDate* Publication date for the online version of the work, if known

*Publisher* The organization that has published the work (e.g. Elsevier, ArXiv, Dryad)

*Ordered lists*

Each BibCitation text string must be parsed into the following ordered lists of structured metadata:

**BibRef** e.g. AuthorList, PubDate, Title, Journal, VolumeNumber, FirstPage, LastPage, Identifier; or AuthorList, PubDate, Title, Version, Publisher, ResourceType, Identifier (datasets).

**AuthorList** First Name, second Name, etc.

*EditorList* First Name, second Name, etc. (if relevant and available)

*ContributorList* First Name, second Name, etc. (if relevant and available)

The complete reference list in each citing work must be parsed into a structured ordered list of references:

RefList An ordered structured list (First cited BibRef, second cited BibRef, etc.) of all the references in the citing work

*People*

**IntPersID** Each person (author, editor) must have a unique internal OCC personal identifier

**Name** The text string giving the name of an author

**Role** Author, Editor, Contributor, etc.

*Each name text string must be parsed into*:

*GivenName(s)* (If available from the name text string)

**GivenNameInitial(s)** Either extracted from the name text string or created from GivenName(s)

**FamilyName**

*In addition*:

*ExtPersID* If an external personal identifiers can be found – e.g. ORCID – they should be recorded

InstitutionalAff If one or more institutional affiliation for a person can be found, these should be recorded.

**For possible future inclusion: citation contexts**

While **not** required for the core functionality of the Open Citations Corpus, we might at a later data return to the source material and extract the following information for each individual reference in the reference list:

InTextReferencePointerCount A count of the number of InTextReferencePointers in a textual work denoting the same bibliographic reference.

InTextReferencePointerList An ordered list of the InTextReferencePointers in a textual work denoting the same bibliographic reference.

InTextReferencePointerContext The context of each InTextReferencePointer, i.e. the text of the sentence or paragraph in which it is embedded

ContextLocation The location of that sentence or paragraph as part of one of the rhetorical blocks of the textual work (Introduction, Methods, Results, etc.)

GlobalCitationCount A count of global citations from a source on a date

GlobalCountSource e.g. Google Scholar, Scopus

GlobalCountDate Date on which count recorded.