Metadata requirements for the Open Citation Corpus

Introduction

**Scope**

The Open Citations Corpus (OCC) 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. The OCC also records the authors and editors of such works.

**Golden rules for working on the Open Citations Corpus**

1. Never trust a reference in a reference list until it has been independently verified.
2. Never trust an identifier (e.g. DOI, PubMed ID) in a reference list until it has been independently verified.

[While all external authorities will contain erroneous data, CrossRef, PubMed, PubMed Central and the arXiv metadata database are *assumed* to contain authoritative data for the purposes of creating the OCC, until proved to the contrary.]

1. Never trust that a software data processing algorithm is working correctly until its output has been manually verified on multiple inputs.
2. Never process thousands of articles until the processing pipeline's algorithms have been verified.
3. Never work alone, and never trust that your colleague has done something correctly – work as a team and check one another's work. Ideally use agile pair programming techniques.

**Golden rules for Open Citations Corpus records**

1. Each OCC record, of which there are seven distinct types detailed below, must have a unique internal identifier having the form "OCC-XX-*nnnnnnnnnn*" (where XX specifies the record type, and *nnnnnnnnnn* is a 10-digit number unique for that record type). Each identifier should beexpressable as a URI, to permit each record to be described in RDF.
2. Provenance information must be explicitly recorded for every OCC record.

**Explanation of metadata items**

For OCC record, 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 ingest of the cited work into the OCC)

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

**Encoding metadata items**

The internal format for marking up OCC metadata entities is BibJSON.

Imported data held in its literal form should be held as plain text or as a block of XML.

OCC entities have been mapped to RDF, enabling their export as Linked Open Data.

**Provenance data for all OCC records**

The following provenance information must be explicitly recorded for each OCC record.

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

**RecordCreationDate** Date on which the information was entered from the source into OCC

**RecordCreator** Name of person responsible for the initial BibliographicRecord creation in OCC

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

**UpdateCurator** Name of person responsible for updating the work's BibliographicRecord in OCC

**Bibliographic records**

The following two bibliographic record types shall be held in the OCC, each instance bearing a unique identifier:

**ReferenceTextRecord**

A ReferenceTextRecord records a bibliographic reference *as it appeared in the reference list of the citing work*. It consists of an identifier plus the reference 'payload' as a literal, plus provenance metadata:

**ReferenceTextIdentifier** (OCC-RT-*nnnnnnnnnn*): A unique internal OCC identifier, expressable as a resolvable URI, for a reference, i.e. an item in the reference list of a citing paper.

**ReferenceText** The reference should be recorded "as given" in the reference list of the citing work, either as plain text or as a block of XML. This RT record, reflecting how the reference appeared in the Version of Record of the citing work, should never changed or corrected.

*RelatedBibliographicRecord*The BibliographicRecordIdentifier to which this ReferenceTextRecord relates, if such a BibliographicRecord exists in the OCC.

One ReferenceTextRecord should be created for each reference in the reference list of every citing work imported into the OCC, *whether or not* the cited work has a BibliographicRecord.

When a new BibliographicRecord is created for a cited work, each original ReferenceTextRecord to the cited work should be left unchanged, and a new RelatedBibliographicRecord link should be added to it, linking it to the new authoritative BibliographicRecord for that cited work.

When a new citing work being ingested into the OCC contains a reference to a work for which the OCC already has a BibliographicRecord, a new ReferenceTextRecord should still be created from the reference in the reference list of the citing work, and a RelatedBibliographicRecord link should immediately be added to this, linking the new ReferenceTextRecord from the citing paper to the BibliographicRecord of the cited paper.

Since popular publications will be cited many times, there will exist many references, from different citing papers, all describing the same cited work in potentially slightly different ways.

**BibliographicRecord**

For each work for which a Bibliographic Record is held in the OCC, the following metadata items should be recorded, if available:

*Identifiers and links*

**BibliographicRecordIdentifier** (OCC-BR-*nnnnnnnnnn*): A unique internal OCC identifier, expressable as a resolvable URI, for an authoritative bibliographic record describing a publication.

ExtPubID External identifier for the work

ExtPubIdScheme DOI, ISBN, ArXiV, PubMedCentral, PubMed, GenBank

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

*VoRURI* A link, recorded for a preprint, resolving to the separate 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

*Nature of publication*

PubType 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(s) in which the work is published, e.g. txt, pdf, html, latex, xml, ps

*Rights* Details of copyright and other asserted rights

*License* Details of publication license

*Description of publication*

*Keywords* Keywords and phrases relevant to the work

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

*Bibliographic details of the work*

**BibCitation** Full bibliographic citation of the work, as a text string

**Title** Title of the work

**Journal** Journal name or standardized abbreviation

**Volume** Volume number

*Issue*Issue number

**FirstPage**

**LastPage** (This should not be truncated! i.e. if the first page number is 567 and the last page number is 579, this should be recorded as "579", not as "79".**)**

**PubYear** Publication year of the work (format *yyyy)*

*Version* (for a dataset in a data repository)

*ResourceType* (for a dataset in a data repository)

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

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

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

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

*Contributors* Names of any named contributors to the work

*List of authors*

The names of the authors of each work must be separately recorded in an **ordered numbered list of authors**, i.e. FirstAuthor, SecondAuthor, etc., *each* identified by an AuthorNameIdentifier (see below for details)

*List of editors*

The names of the editors of each work (if any) must be separately recorded in an **ordered numbered list of editors**, i.e. FirstEditor, SecondEditor, etc., *each* identified by an EditorNameIdentifier (see below for details)

*List of references*

The references in the reference list of each work must be separately recorded in an **ordered numbered list of references**, i.e. FirstReference, SecondReference, etc., *each* identified by a ReferenceTextIdentifier (see above for details)

One OCC BibliographicRecord should be created for each citing work recorded in the Open Citation Corpus, and for each cited work for which authoritative metadata can be obtained or created.

An OCC BibliographicRecord does not have to be complete, but it must be accurate. For this reason, a BibliographicRecord must be derived from an authoritative source, e.g.

1. bibliographic metadata about a *citing* work obtained from the ingest source, e.g PubMed Central, the arXiv metadata database or CrossRef;
2. bibliographic metadata about a *cited* work obtained from PubMed in response to a PubMed ID query;
3. bibliographic metadata about a *cited* work obtained from CrossRef in response to a DOI query; or
4. (as a last resort, if the cited work lacks both DOI and PubMed ID) bibliographic metadata about a *cited* work obtained by a voting algorithm comparing at least three OCC ReferenceTextRecords referring to the same cited work.

[For such a voting algorithm, accented words (e.g. café), symbols (e.g. β, π), and words devoid of HTML or XML escape characters are to be selected in preference to non-accented words (e.g. cafe), English transliterations of symbols (e.g. beta, pi), and words including escape characters.]

A separate BibliographicRecord should be created for each cited work for which no BibliographicRecord already exists **when, and only when**, accurate bibliographic metadata for the work have been obtained from an external authority, or (failing that) when accurate metadata for the work have been generated by comparison between at least three references to that work, as described above.

An OCC BibliographicRecord can be incrementally supplemented with additional metadata from reliable sources whenever such additional metadata is obtained. When this happens, the provenance metadata for this BibliographicRecord should be updated.

**Citation records**

A citation is a performative act linking two works, taking the simple form "*WorkA cites WorkB*". In the OCC:

1. the citing work, *WorkA*, is identified by its OCC BibliographicRecordIdentifier (OCC-BR-*nnnnnnnnnn*), and
2. the cited work, *WorkB*, is identified by its OCC BibliographicRecordIdentifier (OCC-BR-*nnnnnnnnnn*) if available, or else by an UnverifiedReferenceIdentifer for the reference to the cited work in the reference list of *WorkA* (OCC-UR-*nnnnnnnnnn*).

The following two citation record types shall be held in the OCC, each instance bearing a unique identifier:

**UnmatchedCitationRecord**

An UnmatchedCitationRecord exists where the citation is to a cited work for which the OCC has as yet no independent BibliographicRecord, but only an ReferenceTextRecord to the cited work.

Each UnmatchedCitationRecord must have identifiers for the citing and cited works as detailed above, provenance metadata for the citation record, and a unique identifier for the citation record itself:

**UnmatchedCitationIdentifier** (OCC-UC-*nnnnnnnnnn***)** A unique internal OCC identifier, expressable as a resolvable URI, for an UnmatchedCitation.

**Status** Either Active or Deprecated.

**MatchedCitationRecord**

A MatchedCitationRecord exists where the citation is made to a cited work for which the OCC has a BibliographicRecord.

Each MatchedCitation Record must have identifiers for the citing and cited works as detailed above, provenance metadata for the citation record, and a unique identifier for the citation record itself:

**MatchedCitationIdentifier** (OCC-MC-*nnnnnnnnnn*) A unique internal OCC identifier, expressable as a resolvable URI, for a MatchedCitation.

When a BibliographicRecord is created for a cited work that had previously been described in the OCC only by one or more ReferenceTextRecords, a new MatchedCitationRecord must also be created for each work in the OCC that cites the cited work, linking its BibliographicRecord to the new BibliographicRecord of that cited work.

Additionally, each pre-existing UnmatchedCitationRecord between the BibliographicRecord of each citing work and its ReferenceTextRecord to the cited work shall be marked as **deprecated** in favour of the new MatchedCitationRecord that now exists between the BibliographicRecord of that citing work and the new BibliographicRecord of the cited work.

Deprecated UnmatchedCitationRecords should be ignored (i.e. not counted) when determining the numbers of unmatched citations in the OCC.

When a new citing work being ingested into the OCC contains a reference to a work for which the OCC already has a BibliographicRecord, a MatchedCitationRecord should be created immediately. However, no deprecated UnmatchedCitationRecord need be created between the BibliographicRecord of the citing work and the new ReferenceTextRecord to the cited work created from its reference list entry.

**Personal records**

The following four personal record types shall be held in the OCC, each instance bearing a unique identifier:

**AuthorNameRecord**

An AuthorNameRecord is made up of an AuthorNameIdentifier plus the AuthorName 'payload', plus the ReferenceTextIdentifer for the reference from which the name was extracted, plus provenance metadata:

**AuthorNameIdentifier** (OCC-AN-*nnnnnnnnnn*): A unique internal OCC identifier, expressable as a resolvable URI, for an author's name extracted from a ReferenceTextRecord that must be separately identified.

**AuthorName**: This should be recorded the author's name "as given" in the ReferenceTextRecord from which it was extracted, either as text or as a block of XML. This AN record should never changed or corrected.

One AuthorNameRecord should be created for each author name extracted from a ReferenceTextRecord

**EditorNameRecord**

An EditorNameRecord is made up of an EditorNameIdentifier plus the EditorName 'payload', plus the ReferenceTextIdentifer for the reference from which the name was extracted, plus provenance metadata:

**EditorNameIdentifier** (OCC-EN-*nnnnnnnnnn*): A unique internal OCC identifier, expressable as a resolvable URI, for an editor's name extracted from an ReferenceTextRecord, that must be separately identified.

**EditorName**: This should be recorded "as given", either as text or as a block of XML. An EN record should never changed or corrected.

One EditorNameRecord should be created for each editor name extracted from a ReferenceTextRecord.

**PersonalNameRecord**

A PersonalNameRecord is made up of an PersonalNameIdentifier, plus an external name identifier and external identifier scheme, if available, plus the Personal Name metadata components 'payload', plus provenance metadata:

**PersonalNameIdentifier** (OCC-PN-*nnnnnnnnnn*) An internal OCC identifier, expressable as a resolvable URI, for a unique individual person.

*ExternalPersonalIdentifier* An authoritative identifier for the person, from an external personal identification scheme (if available).

*ExternalPersonalIdentifierScheme* The name of the external personal identification scheme from which the ExternalPersonalIdentifier is taken, e.g. ORICD.

**FamilyName**

**GivenName(s)** and/or GivenNameInitials

*PersonalTitle*  (e.g. Dr, Professor, Revd.)

*DegreeSuffix* (e.g. Ph.D.)

*HonorificSuffix* (e.g. FRS)

*FamilialSuffix* (e.g. Sn, III) where helpful in distinguishing between individuals.

*Email* Email address for person, if available

*InstitutionalAff* Person's institutional affiliation (if available)

For each PersonNameRecord, the role of the person (e.g. Author, Editor, Contributor) in the context of a particular work must also be recorded:

**Role**: Author, Editor or Contributor.

**ContextForRole**: BibliographicRecord of the work for which the role applies. (Note that an individual can be the editor of one paper and the author of several others, so this information is essential to clarify in which contexts the person holds particular roles).

One PersonalNameRecord should be created in the Open Citation Corpus for each person for whom authoritative metadata can be obtained.

An OCC PersonalNameRecord must be derived from an authoritative source, e.g.

1. the name of the author of, editor of, or contributor to a *citing* work obtained from the Version of Record of the citing work, ingested from e.g PubMed Central, the arXiv metadata database or CrossRef;
2. metadata about the author or editor of a *cited* work obtained from PubMed in response to a PubMed ID query;
3. metadata about the author or editor of a cited work obtained from CrossRef in response to a DOI query; or
4. metadata about the author or editor of a cited work obtained from ORCID or a similar name authority in response to an ORCID or other personal identifier query,
5. (as a last resort, if the cited work lacks both DOI and PubMed ID), a canonical version of the name obtained by a voting algorithm comparing at least three OCC AuthorNameRecords or EditorNameRecords relating to the same person, derived from three separate ReferenceTextRecords to the same cited work.

[For such a voting algorithm, longer names, accented names, and names devoid of HTML or XML escape characters are to be selected in preference to shorter names, non-accented names and names including escape characters.]