Metadata model for the Open Citation Corpus, v5

Introduction

**Scope**

The Open Citations Corpus (OCC) records citations between citing and cited published scholarly entities. Such an entity 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. Where possible, the OCC also records other types of relationship between such entities (e.g. sharing authorship, having a common funding agency, sharing authors' institutions, or having some other relationship, for example a semantic relationship of being about the same thing in the sense of PubMed's "Related Citations"). This is made possible by carefully recording the funding sources for such entities, the authors, editors and contributors of such entities, and the institutional affiliations of these people, where such information is available.

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

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

[While no external authority is ever totally free from erroneous data, CrossRef, DataCite, PubMed, PubMed Central, and the arXiv metadata database are *assumed* to contain authoritative metadata records for the purposes of creating the OCC, until proved to the contrary, as are the published Versions of Record of entities with regard to metadata concerning *themselves*.]

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 six 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 beexpressed 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 each OCC record, we need to record the metadata items listed in this document, if available:

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

Normal text metadata items may only be available for citing entities, unless the information for cited entities can be drawn from elsewhere (e.g. a separate ingest of the cited entity into the OCC)

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

**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 version of an OCC Record:

**VersionNumber** The version number for the Record (An integer number, incrementing from 1)

**VersionSource** Source of new information about the entity included in this version of the Record (e.g. ArXiV, PubMed Central, Dryad)

**FormatReceived** The format in which the information was received and ingested by the OCC, e.g. xml, laTex, text.

**FormatStored** BibJSON by default.

**VersionCurator** Name of the person responsible for curating this version of the OCC Record

**VersionDate**  Date of creation of this version of the Record in OCC (format *yyyy-mm-dd*)

**Metadata records in the Open Citations Corpus**

The Open Citations Corpus hosts both potentially error-ridden citation information 'as received', and also authoritative metadata derived from trustworthy sources. There are thus two basic types of entity metadata in the Open Citations Corpus: on the one hand references and names as unverified textual strings obtained directly from reference lists in citing entities, recorded within the OCC as **ReferenceTextRecords** and **NameTextRecords**, and on the other hand authoritative metadata concerning published entities and people recorded within the OCC as **BibliographicRecords** and **PersonalRecords**.

Linking these are the citations themselves, recorded within the OCC as **UnmatchedCitationRecords** if no BibliographicRecord exists within the OCC for the cited entities, and as **MatchedCitationRecords** if the citations are made to cited entities for which the OCC holds BibliographicRecords.

All the metadata held by the OCC falls within one of these six classes.

**Entity metadata**

The following two entity metadata record types are 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 entity*. It consists of an identifier plus the reference 'payload' as a literal, and a TargetBibliographciRecordID linking to a BibliographicRecord providing authoritative metadata about that entity, if such a record exists in the OCC, plus provenance metadata as detailed above (not repeated here):

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

**ReferenceText** The reference should be recorded "as given" in the reference list of the citing entity, 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 entity, should never be changed or corrected.

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

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

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

When a new citing entity being ingested into the OCC contains a reference to an entity 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 entity, and a TargetBibliographicRecordID link should immediately be added to this, linking the new ReferenceTextRecord from the citing entity to the BibliographicRecord of the cited entity.

Since popular publications will be cited many times, there will exist many references, from different citing entities, all describing the same cited entity in potentially slightly different ways. Each should have a separate ReferenceTextRecord.

**BibliographicRecord**

A Bibliographic Record is an authoritative statement of the metadata about a published entity. For each entity for which a Bibliographic Record is held in the OCC, the following metadata items should be recorded, if available, plus provenance metadata as detailed above (not repeated here):

*Identifiers and links*

**BibliographicRecordIdentifier**: A unique internal OCC identifier for an authoritative bibliographic record describing a published entity, having the form OCC-BR-*nnnnnnnnnn* expressed as a resolvable URI.

ExtPubID External identifier for the entity

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

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

*VoRURI* For a preprint, a link resolving to the landing page for the separate Version of Record journal article at the publisher's web site

*RelatedURI* A link to any related entity, for example from an article to a supplementary information file, or from an article in PubMed to a semantically related article (described by PubMed as a "Related Citation")

*Nature of publication*

PubType Nature of the entity, e.g. preprint, journal article, conference paper, book, Wikipedia article, dataset, model, computer software

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

Format The format(s) in which the entity 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 entity

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

*Bibliographic details of the entity*

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

**Title** Title of the entity

**Journal** Journal name or standardized abbreviation (for a journal article)

**Volume** Volume number (for a journal article)

*Issue*Issue number (for a journal article)

**FirstPage** (for a journal article)

**LastPage** (for a journal article)

N.B.This should not be truncated! Thus if the first page number is 567 and the last page number is 579, this last page should be recorded as "579", not as "79".

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

*Database or repository Name* (for a dataset, a model or software)

*Version* (for a dataset, a model or software)

*ResourceType* (for a dataset, a model or software)

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

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

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

*Publisher* Name of the organization that has published the entity (e.g. Elsevier, arXiv, EBI, Dryad)

*Funding agency* Name of the funding agency that has provided financial support to the project of which the entity is a product, if known

*Contributors* Names of any named contributors to the entity

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

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

*ContributorList* The names of the contributors to each entity (if any) must be separately recorded in an **ordered numbered list of contributors**, i.e. FirstContributor, SecondContributor, etc., *each* identified by an NameTextIdentifier (see below for details)

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

One OCC BibliographicRecord should be created for each citing entity recorded in the Open Citation Corpus, and for each cited entity 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* entity obtained from the ingest source, e.g PubMed Central, the arXiv metadata database, DataCite or CrossRef;
2. bibliographic metadata about a *cited* entity obtained from PubMed in response to a PubMed ID query;
3. bibliographic metadata about a *cited* entity obtained from CrossRef in response to a DOI query; or
4. (as a last resort, if the cited entity lacks both DOI and PubMed ID) bibliographic metadata about a *cited* entity obtained by a voting algorithm comparing at least four OCC ReferenceTextRecords referring to the same cited entity, where the proposed OCC BibliographicRecord is specified by at least three OCC ReferenceTextRecords are identical and contain all the necessary metadata elements.

[For such a voting algorithm, bibliographic metadata containing 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 entity for which no BibliographicRecord already exists **when, and only when**, accurate bibliographic metadata for the entity have been obtained from an external authority, or (failing that) when accurate metadata for the entity have been generated by comparison between at least four references to that entity, as detailed above.

An OCC BibliographicRecord can be incrementally supplemented with additional metadata from reliable sources whenever such additional metadata is obtained. When this happens, a new version of that BibliographicRecord should be created, with provenance metadata.

**Personal metadata**

The following two name record types are held in the OCC, each instance bearing a unique identifier. The first (NameTextRecord) is for names "as given" in a reference within the citing entity, while the PersonalNameRecord contains authoritative metadata about a person:

**NameTextRecord**

A NameTextRecord is made up of a NameTextIdentifier and the NameText 'payload', plus the SourceReferenceID identifying the reference from which the name was extracted, the Role attributable to the name in the reference (i.e. Author, Consortium Author or Editor), and a TargetPersonalNameID linking to a PersonalNameRecord providing authoritative metadata about that person, if such a record exists in the OCC, plus provenace metadata as detailed above (not repeated here)

**NameTextIdentifier**: A unique internal OCC identifier, having the form OCC-NT-*nnnnnnnnnn* expressed as a resolvable URI, for a name extracted from a ReferenceTextRecord that must be separately identified using the SourceReferenceID.

**NameText**: This should be recorded the name "as given" in the ReferenceTextRecord from which it was extracted, either as text or as a block of XML. This NameText, which may contain errors, should never changed or corrected. This NameText may be the name of a group or an organization (e.g. "The MalariaGen Consortium") listed in the author list within the reference, specified as a Consortium Author.

**SourceReferenceID**: The ReferenceTextIdentifier for the reference from which the NameText was extracted.

**Role**: Author, Consortium Author, or Editor, as specified in the reference from which the name was extracted.

*TargetPersonalNameID*The PersonalNameIdentifier to which this NameTextRecord relates, if such a PersonalNameRecord exists in the OCC, and if the correspondence between the NameTextRecord and the person represented by the PersonalNameRecord can be reliably established.

One NameTextRecord should be created for each name in every reference from a citing entity imported into the OCC, *whether or not* that person has a PersonalNameRecord.

When a new PersonalNameRecord is created for a person, each original NameTextRecord for that person should be left unchanged, and a new TargetPersonalNameID link should be added to each NameTextRecord, linking it to the new authoritative PersonalNameRecord for that person, if this relationship can be reliably established.

When a new reference being ingested into the OCC contains a personal name for which the OCC already has a PersonalNameRecord, a new NameTextRecord should still be created from the name in the reference from the citing entity, and a TargetPersonalNameID link should immediately be added to this, linking the new NameTextRecord from the citing entity to the PersonalNameRecord of the named individual, provided that other evidence exists to show that the NameText really does relate to that particular person.

Since productive authors will be named many times, there will exist many name texts, from different references, all naming the same person in potentially slightly different ways. Each should have a separate NameTextRecord.

**PersonalNameRecord**

A PersonalNameRecord is made up of a PersonalNameIdentifier, plus an external name identifier and external identifier scheme, if available, plus the Personal Name metadata components 'payload', plus provenance metadata as detailed above (not repeated here):

**PersonalNameIdentifier** An internal OCC identifier for a unique individual person, having the form OCC-PN-*nnnnnnnnnn* expressed as a resolvable URI.

*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.

For each PersonNameRecord, the role of the person (e.g. Author, Editor, Contributor), the e-mail address and the institutional affiliation of that person in the context of a particular entity should also be recorded, if that information is available:

**Role** Author, Editor or Contributor

*Email* Email address for the person (ideally the institutional email address), if available. N.B. THIS FIELD MUST BE ENCRYPTED USING

*InstitutionalAff* Person's institutional affiliation, if available

**Context**: The BibliographicRecord of the entity to which these metadata relate.

The role of Author or Editor can be determined from the position of the name in a ReferenceText, but the role of contributor, and individuals' e-mails and institutional affiliations, can normally only be determined from examination of the Front Matter and Acknowledgements section of the Version of Record of a published entity. Since these details may change between entities and over time (an individual can be the editor of one paper and the author of several others, and can move institutions and thus change academic email addresses), it is essential to record the context to which these metadata relate:

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* entity obtained from the Version of Record of the citing entity, ingested from e.g PubMed Central, the arXiv metadata database or CrossRef;
2. metadata about the author or editor of a *cited* entity obtained from PubMed in response to a PubMed ID query;
3. metadata about the author or editor of a *cited* entity obtained from CrossRef in response to a DOI query; or
4. metadata about the author or editor of a *cited* entity 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 entity lacks both DOI and PubMed ID), a canonical version of the name obtained by a voting algorithm comparing at least four OCC NameTextRecords relating to the same person derived from four separate ReferenceTextRecords to the same cited entity, within which at least three NameTextRecords are identical.

[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.]

**Citation records**

A citation is a performative act linking two entities, taking the simple form "*EntityA cites EntityB*". The following two citation record types are held in the OCC, each instance bearing a unique identifier:

**UnmatchedCitationRecord**

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

Each UnmatchedCitationRecord requires a unique identifier for the citation record itself, identifiers for the citing and cited entities, status information, and provenance metadata for the citation record as detailed above (not repeated here):

**UnmatchedCitationIdentifier** A unique internal OCC identifier for an UnmatchedCitation, having the form OCC-UC-*nnnnnnnnnn* expressed as a resolvable URI.

**CitingEntityID** The citing entity is identified by its BibliographicRecordIdentifier

**CitedEntityID** The cited entity is identified the ReferenceTextRecord created for the text identifying the cited entity within the reference present in the citing entity

**Status** Either Active or Deprecated

**MatchedCitationRecord**

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

Each MatchedCitation Record must have a unique identifier for the citation record itself, identifiers for the citing and cited entities, and provenance metadata for the citation record as detailed above (not repeated here):

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

**CitingEntityID** The citing entity is identified by its BibliographicRecordIdentifier

**CitedEntityID** The cited entity is identified by its BibliographicRecordIdentifier

When a BibliographicRecord is created for a cited entity that had previously been described in the OCC only by one or more ReferenceTextRecords, a new MatchedCitationRecord must also be created for each entity in the OCC that cites the cited entity, forming a citation link between its BibliographicRecord and the new BibliographicRecord of that cited entity.

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

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

When the reference list from a new citing entity that is being ingested into the OCC contains a reference to an entity 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 entity and the new ReferenceTextRecord to the cited entity created from its reference list entry.