

CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks

David Shotton*

Image Bioinformatics Research Group, Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

ABSTRACT

CiTO, the Citation Typing Ontology, is an ontology for describing the nature of reference citations in scientific research articles and other scholarly works, and for publishing these descriptions on the Semantic Web. Citations are described in terms of the factual and rhetorical relationships between citing publication and cited publication, the in-text and global citation frequencies of each cited work, and the nature of the cited work itself, including its peer review status. This paper describes CiTO and illustrates its usefulness both for the annotation of bibliographic reference lists and for the visualization of citation networks. CiTO Version 1.3, published on 5 May 2009, is written in the Web Ontology Language OWL, uses the namespace <http://purl.org/net/cito/>, and is available from <http://purl.org/net/cito/>, which uses content negotiation to deliver to the user an OWLDoc Web version of the ontology if accessed via a Web browser, or the OWL ontology itself if accessed from an ontology management tool such as Protégé (<http://protege.stanford.edu/>).

1 CITO SCOPE AND USAGE

1.1 What is meant by a citation

In the context of the Citation Typing Ontology, a bibliographic citation is a reference within a particular citing work of another publication (e.g. a journal article, a book chapter or a web page) termed the cited work. This use of the word 'citation' should be distinguished from the common related use of this word to indicate the cited work itself. Within CiTO, 'cite' and 'citation' denote the performative act of citation itself, not the target of the citation.

1.2 Citation networks

The first purpose of CiTO is to enable the citations within a citing work to be recorded and published in machine-readable form as RDF, thus (in Notation3 format):

```
<http://example1.com/citingwork> cito:cites  
<http://example2.com/citedwork> .
```

While the advent of on-line publishing and bibliographic search engines has made the problem of finding individual research articles considerably easier, the present scholarly citation system inadequately exposes the knowledge networks that exist within the scientific literature, linking papers, authors and research projects.

Much of the problem stems from the lack of freely available citation data. In this Open Access age, it is a scandal that reference lists from journal articles, the core elements of the academic data cycle, are not freely available for use by scholars. If CiTO-enabled machine-readable citation data were to be associated with all scholarly publications and published freely on the Web, the construction and interrogation of citation networks would become trivially simple, with enormous advantages to scholarship. Figure shows a simple citation network of papers directly or indirectly cited by Reis *et al.* (2008), the target paper for our recent semantic enhancement demonstration (<http://dx.doi.org/10.1371/journal.pntd.0000228.x001>), described by Shotton *et al.* (2009). This diagram was created automatically by using an RDF graph of CiTO citations as input to the RDF graph visualization tool Welkin (<http://simile.mit.edu/wiki/Welkin>), with the nodes arranged along a vertical temporal axis.

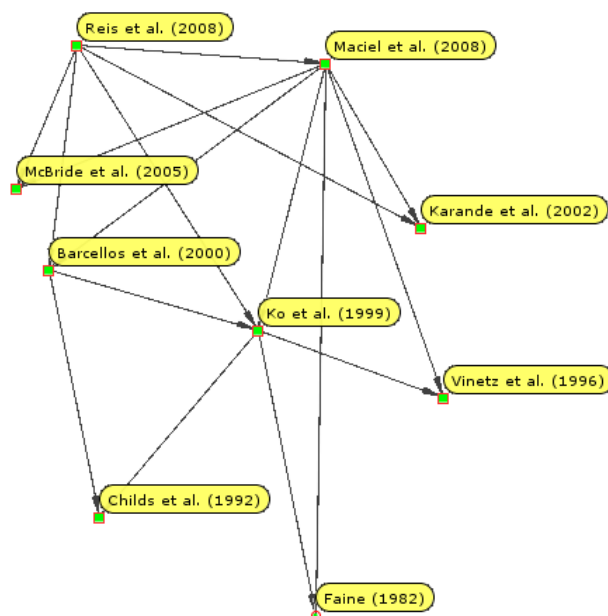


Figure 1. A citation network of selected articles directly or indirectly cited by Reis *et al.* (2008), automatically displayed using Welkin from an input RDF graph of *cito:cites* relationships.

* e-mail: david.shotton@zoo.ox.ac.uk.

1.3 Citation characterization

The second purpose of CiTO is to permit characterization of bibliographic citations. The reasons that one publication cites others are varied. Usually, it is because the more recently published citing work has gained assistance of some sort, perhaps in the form of background information, ideas, methods or data, from the older cited works. It is for this reason that Google Scholar has as its strapline “Stand on the shoulders of giants”, echoing Sir Isaac Newton’s famous remark to his rival Robert Hooke “If I have seen a little further, it is by standing on the shoulders of Giants”. However, more rarely, citations may also be made to critique or refute previous work. CiTO makes it possible to capture and publish such distinctions, i.e. the intent of the author when citing a particular publication, permitting authors (or others) to create metadata describing their citations, quite distinct from metadata describing the cited works themselves. The full list of possible citation typing relationships presently recordable using CiTO is as follows:

Relationships between citing and cited document: *cites*, *citesForInformation*, *confirms*, *corrects*, *credits*, *critiques*, *disagreesWith*, *discusses*, *extends*, *isCitedBy*, *obtainsBackgroundFrom*, *obtainsSupportFrom*, *refutes*, *reviews*, *sharesAuthorsWith*, *updates*, *usesDataFrom*, *usesMethodIn*.

A single citation can be characterized by several different relationships, both factual and rhetorical. In Notation3 format, such characterizations can be made as follows:

```
<http://example1.com/citingwork>
  cito:cites <http://example2.com/citedwork> ;
  cito:usesMethodIn <http://example2.com/citedwork> ;
  cito:extends <http://example2.com/citedwork> ;
  cito:sharesAuthorsWith
    <http://example2.com/citedwork> ; .
```

1.4 Citation frequency

The third purpose of CiTO is to permit two different sorts of citation frequency to be recorded. We are familiar with journal impact factors, based on the frequency of citation of the papers they contain by the scholarly community as a whole. Such global citation frequencies are also widely used to evaluate the academic merits of individuals and their institutions, on the crude premise that all citations are ‘votes of confidence’ in the cited papers. Another and lesser used aspect of citation frequency relates to the local importance of a cited publication to the citing publication. Put crudely, if Paper A cites Paper B once, but cites Paper C ten times at different points in the text, then, *from the point of view of the citing paper*, Paper B is more significant, irrespective of its global citation frequency. CiTO permits one to record both the in-text citation frequency from Paper A to each of the papers it cites, and also the global citation frequency of each cited papers, as determined by consulting third-party authorities such as Google Scholar (<http://scholar.google.com/>), the ISI Web of Knowledge (<http://www.isiwebofknowledge.com/>) and SCOPUS (<http://www.scopus.com/>). Such global citation counts providing proxy estimates of the importance of each cited paper

to the whole academic community. In CiTO, such information is recorded using the following classes and properties:

Citation frequency: *inTextCitationFrequency*, *InTextCitationCount*, *inTextCountValue*, *inTextCitationTarget*, *globalCitationFrequency*, *GlobalCitationCount*, *globalCountValue*, *globalCountSource* and *globalCountDate*.

In-text and global citation information for particular cited publications can be recorded in the following manner.

```
<http://example1.com/citingwork>
  cito:cites <http://example2.com/citedwork> ;
  cito:inTextCitationFrequency [
    a cito:InTextCitationCount ;
    cito:inTextCountValue "10"^^xsd:integer ;
    cito:inTextCitationTarget
      <http://example2.com/citedwork> ;
  ] ; .
<http://example2.com/citedwork>
  cito:isCitedBy <http://example1.com/citingwork> ;
  cito:globalCitationFrequency [
    a cito:GlobalCitationCount ;
    cito:globalCountValue "206"^^xsd:integer ;
    cito:globalCountSource
      <http://scholar.google.com>;
    cito:globalCountDate "2009-03-11"^^xsd:date ;
  ] ; .
```

There is intentional redundancy in these sets of triples, since ‘A cites B’ and ‘B is cited by A’ could both be deduced from the other statements. This level of redundancy has a practical usefulness, since only the direct citation statements can be used to provide clean input to citation network visualization programs such as Welkin (Figure 1), and since the explicit reciprocal statement in the second set of triples preserves the identity of the citing work if the ‘citing’ and ‘cited’ sets if triples were to be separated.

1.5 Characterization of cited works

The fourth purpose of CiTO is to enable the cited works themselves to be characterized, so that someone reading a reference list marked up using CiTO can better appreciate their nature. In making this characterization, CiTO has adopted the classification developed by FRBR (Functional Requirements for Bibliographic Records; <http://www.ifla.org/VII/s13/frbr/frbr1.htm>) for characterizing different aspects of a publication:

1. **Work** A *Work* is a distinct intellectual or artistic creation, recognised through its various expressions. An example of a *Work* is your latest research paper.

Sub-classes of Work in CiTO: *BookReview*, *Catalogue*, *Dataset*, *Discussion*, *Editorial*, *Explanation*, *GrantApplication*, *Image*, *Message*, *Model*, *MovingImage*, *NewsItem*, *Ontology*, *Opinion*, *Patent*, *Protocol*, *ReferenceWork*, *Report*, *ResearchPaper*, *Review*, *ScholarlyText*, *Software*, *Specification*, *StillImage*, *Taxonomy*, *WorkingPaper*.

2. **Expression** An *Expression* is the specific form that a *Work* takes each time it is ‘realized’ in physical or electronic form. For your latest research paper, Draft 5, the preprint, and the published version to which the publisher assigned a DOI, are all expressions of the same work.

Since the number of citations in your research paper probably changed as it was developed through various drafts, the citations that matter for CiTO are those to be found in the final published expression, known to publishers as the ‘**version of record**’. The targets of the citations within a citing work are similarly the ‘version of record’ expressions of the cited works, since it is only these of which citing authors are normally aware. It is thus the *Expressions* of scholarly works that form the domain and range of CiTO object properties.

Sub-classes of *Expression* in CiTO: *Blog, Book, BookChapter, BookSection, ConferencePaper, ConferencePoster, Database, Email, Figure, JournalArticle, JournalItem, PatentDocument, Preprint, Presentation, ReportDocument, Spreadsheet, Table, TextFile, Thesis.*

If an author wishes to add citation typing to references in his or her citing work prior to publication and assignment of a unique identifier to its ‘version of record’ expression, the blank node `_:ThisWork` may be employed when using CiTO to annotate the work’s reference list, and subsequently replaced by the DOI of the published ‘version of record’.

The peer-review status of an expression of a work can also optionally be recorded:

Peer review status: *peerReviewed, notPeerReviewed.*

3. **Manifestation** A *Manifestation* of an expression of a scholarly work defines its particular physical or electronic embodiment. If your latest research paper appeared as an article in a print journal, and also in the on-line version of that journal as an HTML page, and as a downloadable PDF file, these are three separate manifestations of the same ‘version of record’ expression of your work, all bearing the same DOI. When annotating a reference list using CiTO, the nature of the manifestation, in terms of four broad categories, may optionally be recorded.

Sub-classes of *Manifestation* in CiTO: *DigitalMediaObject, OnlineDocument, PrintDocument, WebPage.*

CiTO thus has a number of subclasses of *Work*, *Expression* and *Manifestation* that enable accurate characterization of cited publications. Publications should be characterized using a single disjoint subclass of both *Work* and *Expression*. Each expression can optionally be given a *Manifestations* type and a *peerReviewed* status, thus:

```
<http://example2.com/citedpaper>
  dcterms:bibliographicCitation "Details" ;
  rdfs:label "FirstAuthor et al. (Year)"; #label
  rdf:type cito:ResearchPaper ; # work type
  rdf:type cito:JournalArticle; #expression type
  rdf:type cito:WebPage; #one manifestation type
  cito:peerReviewed "true"^^xsd:Boolean ; .
```

2 THE RELATIONSHIP OF CITO WITH OTHER METADATA SCHEMAS AND ONTOLOGIES

2.1 CiTO and FRBR

While CiTO follows the *Work*, *Expression*, *Manifestation* classification of FRBR, as explained above, the scope of

CiTO is more limited than FRBR, since CiTO covers scholarly works that contain bibliographic references, rather than artistic works such as plays or photographs that do not. Of course, a citation may occasionally be to an artistic work, such as a reference to *Macbeth*.

2.2 CiTO and SWAP

The Scholarly Works Application Profile (SWAP; <http://www.ukoln.ac.uk/repositories/digirep/index/SWAP>) describes the metadata requirements for a scholarly work. SWAP also follows the FRBR model, but its scope is different from that of CiTO, in that SWAP concerns itself with items of metadata surrounding the scholarly work that fall outside the scope of a bibliographic citation, such as funding agency and copyright holder. Conversely, CiTO is concerned with the factual and rhetorical relationships between citing and cited works, something which cannot be captured within the metadata of a single work.

2.3 CiTO, BIBO and SWAN

Among many previous efforts to create metadata schemas and ontologies for characterizing bibliographic references, BIBO, the Bibliographic Ontology written in OWL (<http://bibliontology.com/>), provides the much-needed ability to describe the nature of the cited document in RDF to a high degree of granularity, in terms of *ISSN, Journal, Volume, Pages, Title, Abstract, DOI, dataCopyrighted, editor*, etc. BIBO also covers things outside conventional scholarly works, including broadcasts and legal entities. From the viewpoint of CiTO, BIBO is thus essentially orthogonal. Of greater overlap with CiTO is the fledgling SWAN Scientific Discourse Relationships Ontology (<http://swan.mindinformatics.org/spec/1.2/discourserelationships.html>), designed for characterization of rhetorical statements within text. Initially, CiTO was been constructed as an internally consistent self-contained entity. Further work is now required to establish and specify equivalent classes between CiTO, DCMI, SWAP, BIBO and SWAN.

2.4 CiTO vocabulary definitions

CiTO adopts the Dublin Core Metadata Initiative (DCMI) Type Vocabulary’s definitions for the terms *Dataset, Image, MovingImage, Software, StillImage* and *Text* (<http://dublincore.org/documents/dcmi-type-vocabulary/>).

Other CiTO class names and their definitions include all items in the vocabulary defined by SWAP for subclasses of the *dc:type* property *Text*, with only minor defined nomenclature variations. However, as explained above, CiTO makes a clear distinction between the *Work*, the *Expression* of the work, and the *Manifestation* of that expression, distinctions that are not made by BIBO and SWAN. Despite the clumsiness of this FRBR nomenclature, and the occasional seemingly redundant terminology that results from its use (e.g. *Work*: Report; *Expression*: ReportDocument), this level of granularity avoids ambiguities of meaning present in these other ontologies.

In summary, CiTO extends the vocabularies mentioned above by defining new relationships between citing work

and cited work, and by including a number of additional sub-classes of *Work*, *Expression* and *Manifestation*. In CiTO, all class name and properties are given full definitions, which may be found in the ontology itself at <http://purl.org/net/cito/>, and in textual form in a downloadable document at <http://purl.org/NET/cito/CiTO/terms.doc>.

3 GRANULARITY AND SCOPE

The commentary tradition of classical and biblical scholarship has well-developed methods for citing individual sections, paragraphs or verses of referenced works. In contrast, modern scientific references are typically made to the cited work as a complete entity. It was to enhance this latter practice that CiTO was initially developed. However, there are calls to permit a scientific article to be created compositionally from a set of pre-defined independent parts, and for individual rhetorical argument within the text to be referenced directly (de Waard et al. 2006, de Waard and Kircz 2008). Indeed, it is perfectly possible, using hidden XML code behind the displayed human-readable document, for the text of an on-line paper to be marked up to the level of the paragraph, the sentence or even the individual word, or to particular rhetorical elements (hypotheses, claims, supporting statements, refutations, etc.).

It is thus important for its future usefulness that CiTO is able to support citation of such items, which is possible provided that they have unique identifiers in the form of resolvable DOIs or URLs. However, an additional class vocabulary will need to be created or imported to characterize textual fragments in meaningful ways, for which the National Library of Medicine's Document Type Definition (NLM DTD; <http://dtd.nlm.nih.gov/publishing/>), as defined by the NLM Journal Publishing Tag Set Tag Library version 3.0 (<http://dtd.nlm.nih.gov/publishing/tag-library/3.0/index.html>), would seem a good starting place, since it is widely used as the *de facto* standard during journal production by many Scientific, Technical and Medical publishers.

CiTO has been developed with the scientific research community in mind. Its expansion to fulfill the citation needs of other disciplines will require engagement with appropriate domain experts. For example, classical scholarship in the commentary tradition requires comparison of textual variations between individual manuscripts (in the traditional meaning of the word as a hand-written documents) that are copies of a work of scholarship. Here, the FRBR concept of *Item* becomes important, but, for these unique creations, the distinction between *Manifestation* and *Item* becomes blurred.

4 AN EXAMPLE OF CITO IN USE

CiTO is new, and has not so far been used by third parties. While designed with biological citations in mind, it is potentially generic in usefulness, once expanded to meet the requirements of other disciplines. An example of the use of CiTO for annotation of a reference list in an on-line bio-

medical research article can be seen in the reference list of our enhanced version of Reis *et al.* (2008) at <http://dx.doi.org/10.1371/journal.pntd.0000228.x001>. Here, the human readable CiTO markup can be seen by first going to the References section of the paper (click the 'References' tab above the article's title), and then by turning on the optional citation typing display (click the 'Turn citation typing on' button just before the first reference). An exemplar downloadable file containing all the references from this article with their CiTO markup and their citation frequency information is available in RDF N3 format at <http://dx.doi.org/10.1371/journal.pntd.0000228.x004>.

In developing CiTO, we have created an ontology that should be sufficient in scope for the types of bibliographic citation encountered in biological research articles. Authors should be able to use it to type their own citations, although there is clearly scope for the development of an ontology-backed tool (e.g. a Word plug-in) that would assist that process during paper writing. Alternatively, citation typing can be made at the time of publication or later.

5 FEEDBACK

CiTO is published as open source under a Creative Commons attribution licence, and I invite community engagement and feedback concerning the usefulness of this ontology, and whether, and if so how, it should be extended. Suggestions for amendment or expansion should be sent by e-mail to david.shotton@zoo.ox.ac.uk using the subject 'CiTO Development'.

ACKNOWLEDGEMENTS

I am most grateful to Katie Portwin who participated in the development of the initial CiTO prototype, and to Alistair Miles for guidance in RDF modeling and syntax and for help in ontology coding and publication. The development of CiTO forms part of the work of the Ontogenesis Network, supported by EPSRC grant EP/E021352/1.

REFERENCES

- de Waard, A., L. Breure, J. G. Kircz and H. van Oostendorp (2006). Modeling rhetoric in scientific publications. *International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006*, Merida, Spain. <http://www.instac.es/inscit2006/papers/pdf/133.pdf>.
- de Waard, A. and J. Kircz (2008). Modeling scientific research articles – shifting perspectives and persistent issues. *ELPUB2008 Conf. Electronic Publishing*, Toronto. http://people.cs.uu.nl/anita/papers/ELPUB_2008_deWaardKircz.pdf.
- Reis, R. B., G. S. Ribeiro, R. D. M. Felzemburgh, et al. (2008). Impact of environment and social gradient on *Leptospira* infection in urban slums. *PLoS Neglected Tropical Diseases* 2 (4): e228. doi:10.1371/journal.pntd.0000228.
- Shotton, D., K. Portwin, G. Klyne and A. Miles (2009). Adventures in semantic publishing: exemplar semantic enhancement of a research article *PLoS Computational Biology*: (in press, publication data 17-04-2009). 10.1371/journal.pcbi.1000361.