Applications of the Semantic Publishing and Referencing (SPAR) Ontologies – The Open Citations Project

David Shotton, Silvio Peroni, Ben O’Steen and Alex Dutton

Image Bioinformatics Research Group
Department of Zoology
University of Oxford, UK

"It is a very sad thing that nowadays there is so little useless information"

Oscar Wilde
SPAR (Semantic Publishing and Referencing) Ontologies

These SPAR ontologies are described at http://purl.org/spar/ and in my blog Open Citations and Semantic Publishing at http://opencitations.wordpress.com

CiTO, the Citation Typing Ontology http://purl.org/spar/cito enable characterization of the nature or type of citations, both factually and rhetorically

FaBiO, the FRBR-aligned Bibliographic Ontology http://purl.org/spar/fabio is an ontology for describing bibliographic entities (books, articles, etc.)

BiRO, the Bibliographic Reference Ontology http://purl.org/spar/biro is an ontology to define bibliographic records and references, and their compilation into bibliographic collections and reference lists, respectively

(FaBiO and BiRO classes are structured according to the FRBR schema of Works, Expressions, Manifestations and Items)
The SPAR Ontologies, continued

**C4O, the Citation Counting and Context Characterization Ontology**  
http://purl.org/spar/c4o allows the characterization of bibliographic citations in terms of their number (both locally and globally), and their textual context.

**DoCO, the Document Components Ontology**  
http://purl.org/spar/doco provides a structured vocabulary of document components, both structural (e.g. heading, paragraph) and rhetorical (e.g. Abstract, Introduction).

**PRO, the Publishing Roles Ontology**  
http://purl.org/spar/pro is an ontology for the roles of agents (e.g., author, editor, publisher, librarian) in the publication process, and the times during which those roles are held.

**PSO, the Publishing Status Ontology**  
http://purl.org/spar/pso is an ontology for the temporal status of a document (e.g. draft, under review, published, Version of Record) during the publication process.

**PWO, the Publishing Workflow Ontology**  
http://purl.org/spar/pwo describing the steps in the workflow associated with the publication of a document or other publication entity.
An example of citation metadata in RDF

<http://dx.doi.org/10.1371/journal.pntd.0000228>


a fabio:JournalArticle ; # expression
frbr:realizationOf [ a fabio:ResearchPaper ] ; # work
cito:cites <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> # Reference [6];
cito:obtainsBackgroundFrom <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> ;
cito:usesDataFrom <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> ;
cito:confirms <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> ;
cito:extends <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> ;
cito:sharesAuthorsWith <http://dx.doi.org/10.1016/S0140-6736(99)80012-9> ;
frbr:part [a biro:BibliographicReference;
   c4o:hasInTextCitationFrequency "10"^^xsd:nonNegativeInteger;
An example of citation metadata in RDF  (continued)

# Reference [6]
<http://dx.doi.org/10.1016/S0140-6736(99)80012-9>


a fabio:JournalArticle ; # expression
frbr:realizationOf [ a fabio:ResearchPaper ] ; # work
prism:publicationDate "1999-09-04"^^xsd:date ;
cito:isCitedBy <http://dx.doi.org/10.1371/journal.pntd.0000228> ;

# Google Scholar on 11 March 2009: Cited 206 times.
c4o:hasGlobalCitationFrequency [ a c4o:GlobalCitationCount ;
c4o:hasGlobalCountValue "206"^^xsd:integer ;
c4o:hasGlobalCountSource <http://scholar.google.com> ;
c4o:hasGlobalCountDate "2009-03-11"^^xsd:date ] .
The Open Citations Corpus

- The reference lists from all **204,637 articles** in the Open Access Subset of PMC on 24 January 2011, encoded in RDF using the SPAR ontologies
- These lists contain **6,325,178 individual references**, some unique, but many from different citing articles to highly cited papers
- These references cite **3,373,961 papers** outside the Open Access Subset
  - ~ 20% of all PubMed Central papers (approx. 3,200,000 papers)
  - includes **ALL** the highly cited papers in every biomedical field
- Each reference list is maintained as a distinct unit, by encoding it as a Named RDF Graph with a unique URI
- Encoded these bibliographic records and the citations between them in RDF, creating **236,499,781 quads** occupying 2.1 gigabytes of compressed storage
- Freely available under a CC0 waiver from [http://opencitations.net/data/](http://opencitations.net/data/)
- The complete corpus can be downloaded, or can be queried via a SPARQL endpoint
Viewing citation networks at http://opencitations.net

Impact of Environment and Social Gradient on Leptospira Infection in Urban Slums

http://dx.doi.org/10.1371/journal.pntd.0000228

Application and Validation of PFGE ...

Isolation and Characterization of New ...

Word add-in for ontology recognition: ...

LINNAEUS: A species name identification ...

CITO ...

Adventures in Semantic Publishing: Exemplar ...

Impact of Environment and Social ...
The outward citation network of Reis et al. (2008)
Expanding the Open Citations Corpus

- In partnership with CrossRef, we now wish to
  - expand the Open Citations Corpus to include references from articles in subscription access journals
  - create an automated workflow to harvest references as they are published
- Many publishers already submit the reference lists of their articles to CrossRef as part of the CrossRef ‘CitedBy’ service
- No additional work is required of these publishers, just granting permission for the reference lists to be published as open data
- We would then harvest them from the XML feeds to CrossRef and parse them into RDF using the SPAR ontologies
- CrossRef would host the Open Citations Corpus for long-term security
Reasons for opening reference lists

- Open data is increasingly regarded as a common good
- Citation data (i.e. bibliographic references from one article to another: A \texttt{cito:cites} B.) are ‘just data’ – they contain no creative authorship
- With the exception of Elsevier (via Scopus), publishers do not make money from their reference information, and thus it would cost nothing to open it
- In true Web 2.0 style, it would increase in value the more publishers joined
- Give increased exposure of one’s own journal articles – a form of advertising
- Access to new services built over the open citation data:
  - automated citation correction (approx 1% of references contain errors)
  - citation network visualization
- Good will from funding agencies, government and researchers themselves, who would feel the publishers were at last giving back some of their own data

- I am pleased to say that a number of major academic publishers are now considering whether to open the reference data in this manner
- My message to publishers: “Go for it - what have you got to lose?”
Wellcome Trust joins 'academic spring' to open up science

Wellcome backs campaign to break stranglehold of academic journals and allow all research papers to be shared free online

Alok Jha, science correspondent
guardian.co.uk, Monday 9 April 2012 20.44 BST
The Force11 White Paper

- It summarizes key problems facing scholarly publishing today, and presents a vision that addresses these problems, proposing concrete steps that key stakeholders can take to improve the state of scholarly publishing.

Force11 White Paper:
Improving The Future of Research Communications and e-Scholarship

Editors: Phil Bourne, UCSD; Tim Clark, Harvard/MGH; Robert Dale, Macquarie University; Anita de Waard, Elsevier Labs; Ivan Herman, W3C; Eduard Hovy, ISI/USC; David Shotton, Oxford University

October 28, 2011