Setting our bibliographic references free: towards open citation data

Silvio Peroni
Department of Computer Science and Engineering, University of Bologna, Italy
ORCiD: 0000-0003-0530-4305; E-mail: silvio.peroni@unibo.it

Alexander Dutton
IT Services, University of Oxford, UK
ORCiD: 0000-0003-1448-3114; E-mail: alexander.dutton@it.ox.ac.uk

Tanya Gray
Bodleian Library, University of Oxford, UK
ORCiD: 0000-0003-1561-7364; E-mail: tanya.gray@bodleian.ox.ac.uk

David Shotton
Oxford e-Research Centre, University of Oxford, UK
ORCiD: 0000-0001-5506-523X; E-mail: david.shotton@oerc.ox.ac.uk

Abstract

Purpose. Citation data needs to be recognized as a part of the Commons – those works that are freely and legally available for sharing – and placed in an open repository.

Design/methodology/approach. The Open Citation Corpus is a new open repository of scholarly citation data, made available under a Creative Commons CC0 1.0 public domain dedication and encoded as Open Linked Data using the SPAR Ontologies.

Findings. The Open Citation Corpus presently provides open access to reference lists from 204,637 articles from the Open Access Subset of PubMed Central (OA-PMC), containing 6,325,178 individual references to 3,373,961 unique papers.

Research limitations/implications. We need tools, such as the CiTO Reference Annotation Tools and CiTalO, to facilitate the semantic enhancement of the references in scholarly papers according to Semantic Publishing models and technologies.

Originality/value. Scholars, publishers and institutions may freely build upon, enhance and reuse the open citation data for any purpose, without restriction under copyright or database law.

Keywords: Open Citation Corpus, SPAR Ontologies, CiTO, CiTalO, CiTO Reference Annotation Tool, citations, references, semantic publishing, open access

1 Introduction

We are living in the early part of the decade of open information. Following a spate of recent reports and government policy statements (Boulton, 2012; Finch, 2012; Research Councils UK, 2013; Wellcome Trust, 2013), it can be fairly stated that the policy debate on open access has been won. Interest is now focused on implementation of the open agenda.

Over the past decade, several studies have demonstrated the importance and benefits of releasing articles and data as open access (OA) material: (Lawrence, 2001; Harnad and Brody, 2004; Swan, 2009) gave empirical evidence of the advantages of OA in terms of better visibility, findability and accessibility for research articles. Following an initial study showing similar results (Piwowar *et al.*, 2007), a new larger study by Piwowar and Vision shows that making research data publicly available can increase the citation rates of articles between 9% and 30%, depending on the publication dates of the datasets (Piwowar and Vision, 2013).

But what of open access to the *citation data*, in other words to the reference lists within scholarly papers that cite other bibliographic resources, from which citation rates can be calculated? Heather Piwowar, a resident of Vancouver, Canada, never anticipated the difficulties in collecting such citation data for that study (Piwowar and Vision, 2013). She needed to analyse citation counts for thousands of articles, but three of the major sources of citation data, Thomson Reuter's Web of Science[1], Google Scholar[2] and Microsoft Academic Search[3], did not support PubMed ID queries. Scopus[4], Elsevier's database of scholarly citations, did, but because Piwowar lacked institutional access to that resource, and with direct appeals to Scopus staff falling on deaf ears, she had a problem. She eventually obtained access through a Research Worker agreement with Canada's National Science Library, but, because she had recently worked in the USA, this required her first to obtain a police clearance certificate and to have her fingerprints sent to the FBI. "It was just ridiculous – for Scopus data! I wasted days trying to access the citation data required for my study. I had 10K PubMed IDs to look up. Had there been open citation data, I could have written my own script!" (Piwowar, personal communication).

A similar story can be told concerning Steven Greenberg's striking analysis of citation distortion (Greenberg, 2009), revealing how hypotheses can be converted into 'facts' simply by repeating citation. His work involved the manual construction and analysis of a citation network contained 242 papers, 675 citations, and 220,553 distinct citation paths relevant to a particular hypothesis relating to Alzheimers Disease. Had those citation data been readily accessible online, he would have been saved considerable effort.

These two examples demonstrate how actual research practice suffers because access to citation data is currently so difficult.

In this open access decade, we think it is a scandal that reference lists from academic articles, core elements of scholarly communication that permit the attribution of credit and integrate our independent research endeavours, are not already freely available for use by scholars. To rectify this, *citation data now needs to be recognized as a part of the Commons* – those works that are freely and legally available for sharing – and placed in an *open repository*, where they should be stored in appropriate *machine-readable formats* so as to be easily reused by machines to assist people in producing novel services. So there is work to be done.

In this paper, we first introduce the issues affecting the currently available sources of citation data, and then describe our own contributions to this field which attempt to improve the current situation: the *Open Citations Corpus* (*OCC*)[5], the *Citations Typing Ontology* (*CiTO*)[6] (Peroni and Shotton, 2012), the *CiTO Reference Annotation Tools*[7] and *CiTalO*[8]. OCC is an open repository for citations data, available under a Creative Commons CC0 1.0 public domain

dedication and encoded as Open Linked Data. CiTO is an OWL2 DL ontology (Motik, Patel-Schneider and Parsia, 2012) that enables the assertion of citations in RDF, and their machine-readable characterization in terms of the reasons for such citations. The CiTO Reference Annotation Tools are JavaScript implementations that assist one to assign CiTO annotations to individual references in the reference list of a journal article, in one of two ways. Finally, CiTalO is a web tool that tries to infer an author's reasons for citing a particular paper by using the techniques of ontology learning and mapping, natural language processing, sentiment analysis, and word-sense disambiguation.

The rest of the paper is organised as follows. In Section 2 we introduce a precise definition of the "bibliographic citation", clarifying the components typically used within a text to build the citational device of a scholarly document, and we outline the need for citation typing. In Section 3 we list some of the main drawbacks of current sources of citation information. In Section 4 we describe the structure of the Open Citation Corpus and its use of ontological models, while in Section 5 we describe such models, including CiTO, which, if adopted, can bring advantages to the provision of citation services. In Section 6, we further describe citation typing, the act of assigning a particular rhetorical or factual characterisation to a citation using CiTO, and the CiTO Reference Annotation Tools that assist that process. After introducing a metaphor that captures the benefits of open citations and citation typing in Section 7, we conclude the paper in Section 8.

2 What exactly is a reference within a scholarly document?

The act of bibliographic citation – a scholar referencing the published work of others, that usually originates "from the interestingness of the phenomenon that has been addressed in the old article and considered again in the new one" (Liu and Rousseau, 2013) – is central to scholarly communication, permitting the attribution of credit and integrating our independent research endeavours. Citations knit together the whole world of scholarship into a gigantic citation network, a directed graph with publications as the nodes and citations as the links between them. Scholarly communication involves the flow of information and ideas through this citation network, and analysis of changes in the network over time can reveal patterns of communication between scholars and the development and demise of academic disciplines.

2.1 The word "reference" and its overlapping meanings

In bibliometrics parlance, a *reference* is made from a citing paper, while a *citation* is received by the cited paper. However, as shown in Figure 1, the word "reference" is colloquially used to mean many things: either the bibliographic reference itself, or the entry in the body text of an article that denotes such a reference, or the act of citing the target publication, or the actual target publication itself (as in "Have you read that reference yet?"). The word also has variety of other meanings, and in particular is widely used in academia to mean a statement about a person's achievements, qualifications, competence and character, supplied, for example, in support of a job application or an academic promotion.

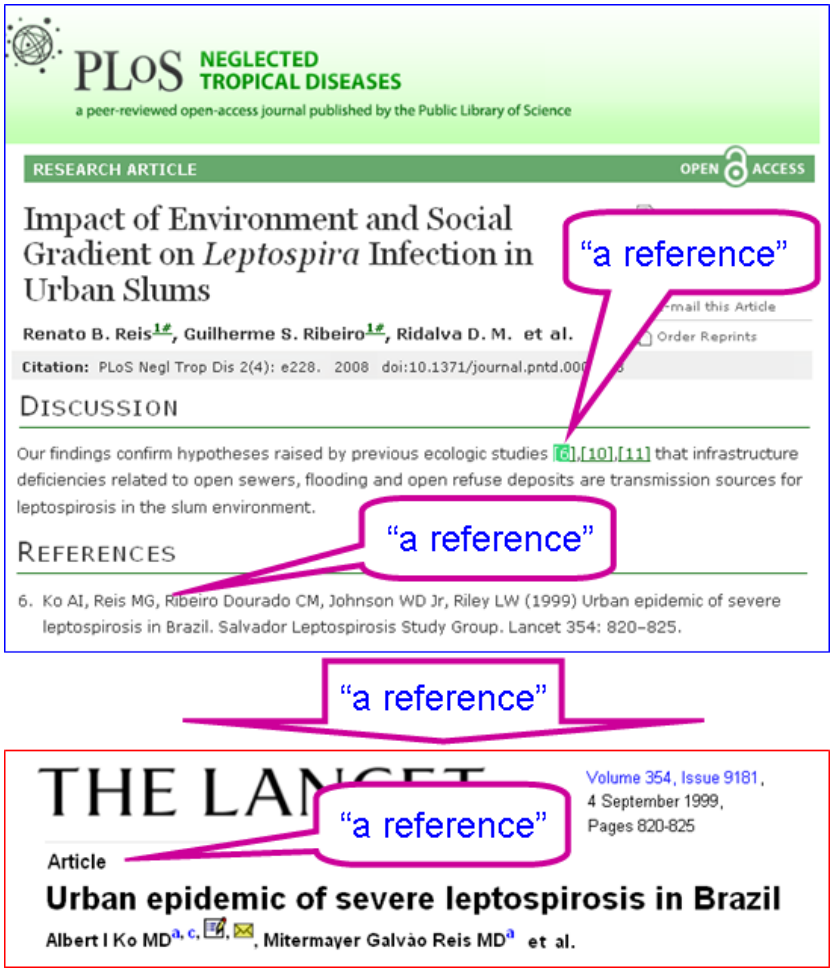

Figure 1. Lazy and ambiguous use of the word "reference" in the context of bibliographic citation.

To avoid such ambiguity in the context of bibliographic citation, we use the term "bibliographic reference" to refer to the reference itself, and employ other terms for other purposes, as defined in the following list:

- ***Bibliographic record***: a data record held in some authoritative information system or library catalogue that fully describes a particular publication. Each bibliographic record comprises a set of entities defined by the publisher describing aspects of the publication, which for a journal article include but are not restricted to: names of all authors, title of article, journal name, volume number, issue number, first and last page numbers, full publication date, publisher's name, copyright information, peer reviewed status, open access status, Digital Object Identifier (DOI) for the article, and International Standard Serial Number (ISSN) for the journal.
- ***Bibliographic reference***: the textual entity within a citing work that identifies a cited work. A bibliographic reference contains some of the elements of the full bibliographic record for the cited work, arranged in a specific format determined by the house style of the citing publication. Some journal house styles require omission of particular elements of the bibliographic record regarded as essential by others, such as the names of all the authors, the title of the article, and the Digital Object Identifier. For a journal article, the bibliographic reference *minimally* comprises: first author's surname and initials, publication year, abbreviated journal name, volume number, and first and last

page numbers. Typically, when the citing work is a scientific journal article, each bibliographic reference is complete in itself and forms a reference list item in the "References" section at the end of the article. In other types of publication, particularly in the humanities, bibliographic references may be contained within footnotes, may be mixed with comments, and may contain pointers, such as "ibid." (abbreviation of the Latin *ibidem*, meaning "the same place") and "op. cit." (abbreviation of the Latin phrase *opere citato*, meaning "in the work cited"), that refer the reader to a previous bibliographic reference from which information needs to be extracted and duplicated to complete the current incomplete bibliographic reference. For this reason, automated parsing of bibliographic references within humanities publications is particularly difficult.

Because errors can be introduced when an author creates a bibliographic reference, a published bibliographic reference should not be trusted to be a fully accurate expression of the information contained within the authoritative bibliographic record for that cited work.

- ***Citing work***: the article that contains a bibliographic reference to another work.
- ***Cited work***: the article that is being referred to by such a bibliographic reference.
- ***Citation***: the attribution link between a citing work and the cited work that is created when the author of the citing work makes a bibliographic reference to the cited work.
- ***To cite*** (transitive verb): the *performative act of citing a published work*, typically instantiated (for textual works) by means of a bibliographic reference included in the citing work, which should be explicitly denoted somewhere within the text of the citing work by an in-text reference pointer. The passive condition of "having been cited" exists for a published work when a bibliographic reference to it exists within another published work.
- ***In-text reference pointer***: the entity present in the body text of a citing work that denotes a particular bibliographic reference in the reference list or a footnote. In scientific literature, this in-text reference pointer can be presented in different forms – as a square-bracketed or superscripted number (e.g. "[3]" or "[3]"); as a square-bracketed text string comprising the first letter of each author's surname (to a maximum of three) plus the last two digits of the publication year (e.g. "[RDS02]"); or as a parenthesised text string containing, for a single-author publication the author's surname and the publication year (e.g. "(Renear, 2002)"), for a two-author publication both authors' surnames and the publication year (e.g. "(Renear & Jones, 2002)"), or for a multi-author publication the first author's surname followed by "*et al*." and the publication year (e.g. "(Renear *et al.*, 2002)").
- ***Citation context***: the textual content of that component of the published paper (e.g. sentence, paragraph, section or chapter) within which an in-text reference pointer appears, which provides the rhetorical rationale for the existence of that citation.

These terms are defined in machine-readable form within four of the SPAR (Semantic Publishing and Referencing) Ontologies: *CiTO*, the *Citation Typing Ontology*[6]; *BiRO*, the *Bibliographic Reference Ontology*[9], *C4O*, the *Citation Counting and Context Characterization Ontology*[10], and *DoCO*, the *Document Components Ontology*[11], that are introduced more fully in Section 5.2.

To summarize: in the context of scholarly citations between scientific articles, a citation from a citing work to a cited work is typically made by including a bibliographic reference to the

cited work in the reference list of the citing work, and by denoting this bibliographic reference at one or more appropriate points within the body text of that citing work by inclusion of one or more in-text reference pointers, as shown in Figure 2, which clarifies the ambiguous situation shown in Figure 1.

Figure 2. SPAR nomenclature provides more precise terms relating to bibliographic citation.

2.2 Classifying citing and cited entities

As introduced above, a citation can be expressed as the simple assertion "Paper A cites Paper B". However, to be useful, full bibliographic details of the citing and cited entities need to be included with each citation record, or to be automatically retrievable from web services using the Digital Object Identifiers (DOIs)[12], PubMed IDs[13], or equivalent globally unique identifiers[14] of these publications.

Traditionally, bibliographic references have been to monographs, book chapters, peer-reviewed journal articles, papers in conference proceedings, etc. Today, citation targets additionally include patents, web pages, blog posts, datasets, presentations and other forms of digital content. Reciprocally, citations are sometimes made from these entities to related research articles.

In addition to this classification, referring to the *nature* of each citing and cited work, one might also be interested in retrieving a more abstract characterisation of such entities. For instance, whether a particular entity contains experimental research results, the conclusions from a survey or questionnaire, a system description, etc. Such descriptions characterise the content of the publication rather than its publication type.

To our knowledge, there exists no shared taxonomy describing such characterisations. The Journal of Documentation[15] allows one to classify an article in one of seven categories – *research paper*, *viewpoint*, *technical paper*, *conceptual paper*, *case study*, *literature review* or *general review* – while the Journal of Web Semantics[16] provides authors with a choice of four categories – *research paper*, *ontology paper*, *survey* and *system paper*. Other journals adopt different classifications depending on their particular scopes.

Any open citation system should be able to include descriptions of this diversity of types of bibliographic entities, since the clear and unambiguous declaration of the types of citing and cited entities is relevant to task of creating alternative metrics to assess the quality of research (Priem *et al.*, 2010; Roemer and Borchardt, 2012; Priem, 2013).

2.3 Classifying the citations themselves

As introduced in (Ciancarini *et al.*, 2013a), the mere existence of a citation, i.e. "Paper A cites Paper B", might not be enough to capture the actual relevance of the cited article. It may be necessary to determine why the citation has been made, to permit one to act appropriately. For example:

- should negative and positive citations be counted in the same way?
- should self-citations have the same weight as those from other scholars?
- is it appropriate to give a review the same weight as a research paper, by giving equal importance to the number of times each has been cited?

Having an effective means of characterising citations opens interesting perspectives that go beyond the quantitative evaluation of research products. For instance, it should be possible automatically to analyse the pertinence of documents to particular research areas, to discover emerging research trends, to observe how new research methods and ideas are propagated, to build sophisticated recommenders, and so on.

In any source of citation information, it would thus be highly desirable if all citations were annotated with their proper functions, which are "the author's reasons for citing a given paper" (Teufel *et al.*, 2009).

In the past, several models have been proposed to describe citation functions. For instance, (Teufel *et al.*, 2006; Teufel *et al.*, 2009) provide a categorisation of citation functions by introducing twelve different classes clustered in three sets: those that convey a negative (CoCo-, Weak), positive (PMot, PUse, PBas, PModi, PSim, PSup) and neutral connotation (CoCoGM, CoCoR0, CoCoXY, Neut) respectively. Starting from a large dataset of research papers, (Jörg, 2008) found one hundred and fifty cue verbs – i.e. verbs usually used to carry important information about the nature of citations, e.g. "based on", "outperform", "focus on", "extend", etc. – that could be mapped to classes of citation functions according to the classification provided in (Moravcsik and Murugesan, 1975) – conceptual/operational, organic/perfunctory, evolutionary/juxtapositional, and confirmative/negational). In the biological field, (Agarwal *et al.*, 2010) introduces eight different top-level classes describing different kinds of citations: background, contemporary, contrast, evaluation, explanation of results, material and methods, modality, and similarity. And the list could go on.

We have furthered this work by developing *CiTO*, the *Citation Typing Ontology*[6], an OWL 2 DL ontology that permits citations to be characterised, typed and described in RDF (Peroni and Shotton, 2012). CiTO permits the motivations of an author when referencing another document to be captured and described according to forty-one different citation properties, each with its inverse property. These properties are categorised as factual (e.g. cito:includesQuotationFrom) and/or rhetorical properties[17], with the latter being sub-grouped into positive properties (e.g. cito:supports), neutral properties (e.g. cito:reviews) and

negative properties (e.g. cito:disputes), these groupings being themselves described in a new extension ontology called *Functions of Citations*[18].

3 Drawbacks of currently-available sources of citation data

3.1 Cost

The most authoritative sources of scholarly citation data are the Thomson Reuters Web of Science (WoS)[1], which grew from the Science Citation Index created by Eugene Garfield in 1964, originally published by the Institute for Scientific Information (ISI); and its main commercial rival, Elsevier's Scopus[4], released in 2004. Both have wide coverage of the leading literature, but because neither is complete, they are widely regarded as complementary (Chadegani *et al.*, 2013).

For access to these two resources, UK research universities each pay tens of thousands of pounds annually (Chadegani et al., 2013), with equivalent sums being charged to academic institutions in other developed countries. The exact values of these subscriptions are closely guarded industrial secrets, and the university librarians who pay these fees are bound by confidentiality agreements from disclosing them to their academic colleagues.

This high cost severely disadvantages all who work outside such rich institutions, including most small and medium sized businesses, and members of the general public.

The other significant sources of citation information, also run by commercial companies but accessible without subscriptions, are Google Scholar[2] and Microsoft Academic Search[3], released in 2004 and 2009, respectively. Google Scholar's coverage is wider than that of the others, because it includes books, theses, preprints, conference papers, technical reports and other non-peer-reviewed 'grey' literature.

All these sources have licence restrictions that prevent the free re-use of their citation data. As a result, bibliometrics papers are rarely permitted to publish the citation data upon which their conclusions are based, so that the benefits of open data (for validation of findings, reuse, etc.) cannot be embraced.

3.2 Citation count accuracy

Available citation data are not necessarily accurate. For instance, David Shotton's citation record differs considerably between Web of Science[1], Scopus[4], Google Scholar[2] and Microsoft Academic Search[3]. For example, on 11th October 2013, the "Adventures in Semantic Publishing" paper that he co-authored (Shotton *et al.*, 2009) had citation counts of 21, 38, 90 and 16, respectively, in these four databases. Which to trust? More worryingly, an earlier paper (Shotton *et al.*, 1972) had two separate entries in Web of Science, with citation counts of 59 and 19, respectively, for this single publication, as shown in Figure 3. The bibliographic information in the second entry with 19 citations, reported by Web of Science as having been obtained from Medline, is correct, but that in the first entry with 59 citations, reported to have been obtained from Web of Science itself, has the wrong author order, the wrong publication year, the omission of one word ("the") and the addition of two erroneous hyphens in the title, and an incorrect last page number.

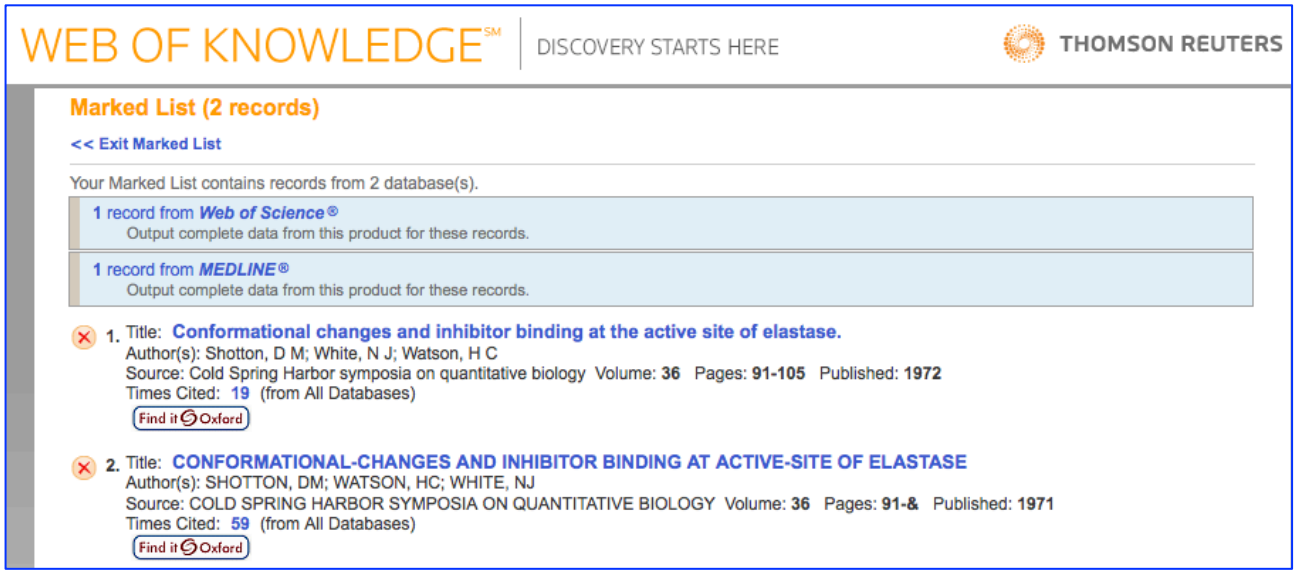

Figure 3. Duplicate entries for the same paper in Web of Science with different citation counts.

What do these differing citation counts for the same paper say about the intrinsic reliability of the Thomson-Reuters impact factor, which is based on such counts?

Additionally, Web of Science is totally misleading when it comes to assessing research within the Computer Science domain, a field in which most important works are published in conference proceedings that it does not fully index (Meyer *et al.*, 2009). For this reason, computer scientists prefer to use Google Scholar to look for papers and citations, since that source does index conference papers – even if its consistency and accuracy is lower compared to the other commercial citation services (Franceschet, 2010). For instance, within the Computer Science domain, a paper that Silvio Peroni co-authored five years ago (Peroni, Motta and d'Aquin, 2008), which was awarded Best Research Paper at the 3rd Asian Semantic Web Conference, has citation counts of 0 (Web of Science), 2 (Scopus), 53 (Google Scholar) and 9 (Microsoft Academic Search). However, although it is richer than other sources for the Computer Science domain, Google Scholar does not permit one to query its data and return them in a defined format – most probably for political reasons rather than technical ones – reducing its value for constructing citation networks.

3.3 Citation-based metrics

Citation analysis has for too long been dominated by 'local' thinking. The Impact Factor (Garfield, 2006), computed to measure the relative importance of journals, is based simply on the number of direct citations received by articles within the journal, without consideration of the wider network. But a citation from a paper that is itself highly cited should count for more than a citation from a paper that is not. For this reason, other measures such as the EigenFactor (Bergstrom *et al.*, 2008) use the structure of the entire network to evaluate the importance of each journal.

The Impact Factor for a journal, flawed as it is, is then further used out of context as a proxy for the importance of individual articles and their authors. These short-falls, and the delays intrinsic to accumulating a significant number of citations for one's papers, have led to the development of alternative metrics of impact and esteem (Piwowar, 2013; Priem, 2013; Hahnel, 2013).

Several studies have highlighted how classification of citations into different types (e.g. supportive, critical, neutral) would lead to changes in the values of the Impact Factor and similar metrics. For instance, knowledge of the fact that a citation has been made for rhetorically negative reasons (MacRoberts and MacRoberts, 1989), or that the citing and the cited works share at least one author (i.e. self-citations) (Aksnes, 2003), can change radically the perceived impact of such a work within the scholarly community.

Nevertheless, direct bibliographic citation of another's work will remain a keystone indicator of that cited work's significance and, indirectly, of the impact of the journal in which it appears. The free availability of citation data is therefore important for scholarship.

4 The Open Citation Corpus

The *Open Citations Corpus* (*OCC*), a new open repository of scholarly citation data, is attempting to improve the current situation. Its objective is to provide accurate bibliographic citation data that others may freely build upon, enhance and reuse for any purpose. It does this by making the citation data available under a Creative Commons CC0 1.0 public domain dedication [19], without restriction under copyright or database law.

4.1 Aims of OCC

To achieve this objective, we had four initial aims:

- To create a semantic infrastructure that makes possible the description of citations, references and bibliographic entities in RDF, since we found existing ontologies inadequate for our purpose.
- To extend that semantic infrastructure to handle data citations and data entities, as well as bibliographic citations and bibliographic entities, mindful of Philip Bourne's prediction that soon there will be no meaningful difference between a journal article and a database entry (Bourne, 2005).
- To provide exemplars of how these ontologies can be applied to real-world data, by creating mappings from existing encodings to RDF, and by creating RDF metadata relating to bibliographic and data entities and their citations.
- To encode in RDF the reference lists within all the articles in the Open Access Subset of PubMed Central[20], and to publish them as Open Linked Data in the Open Citations Corpus, so that third parties can be free to use them in novel ways.

Originally developed with funding from Jisc[21], a UK information technology research and development funding organization, the prototype OCC was published in the summer of 2011, providing open access to reference lists from the 204,637 articles that comprised the Open Access Subset of PubMed Central (OA-PMC) on 24 January 2011, which contained 6,325,178 individual references to 3,373,961 unique papers. This original corpus was encoded as Linked Open Data using the SPAR Ontologies[22], and is available at http://opencitations.net/. Figure 4 shows part of the outgoing citation network of papers documented in the OCC that are cited by one paper in the corpus, the one that formed the subject of the semantic enhancements described in (Shotton *et al.*, 2009), namely the paper by (Reis *et al.*, 2008).

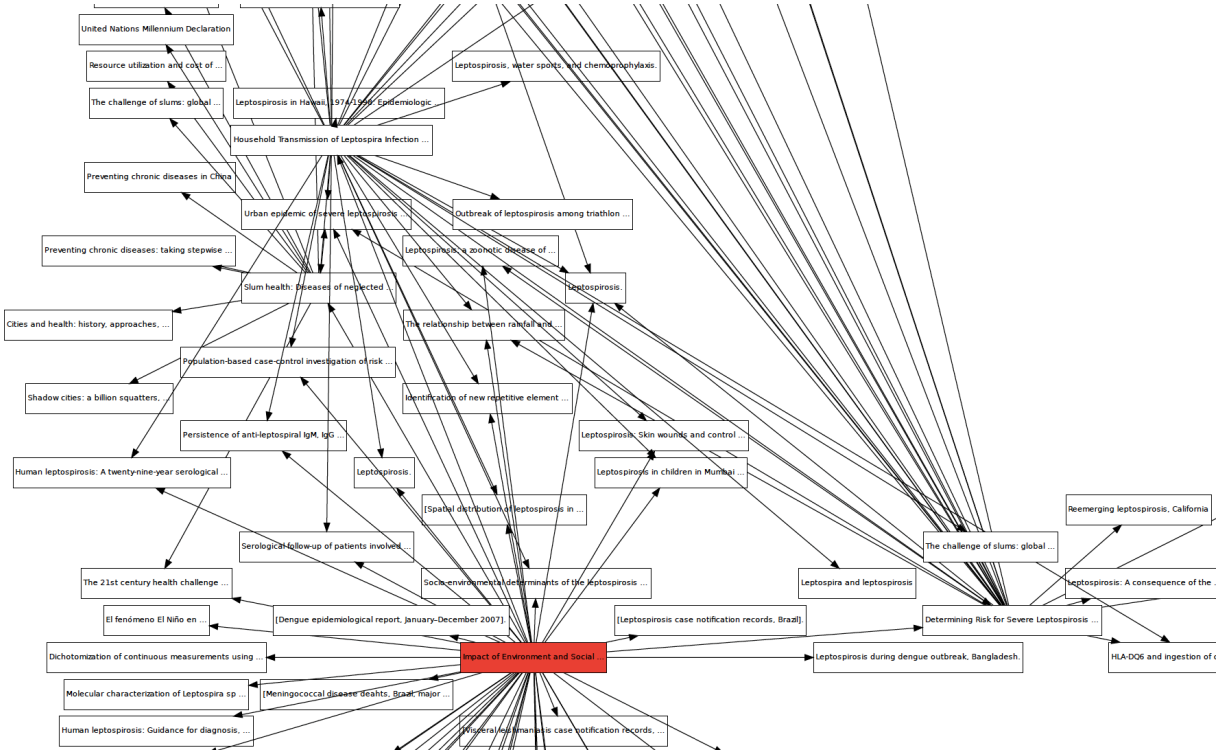

Figure 4. Part of the network of outgoing citations from (Reis *et al.*, 2008) (shown in red) to other publications documented in the Open Citations Corpus, with a depth of two links.

Despite its small size, this corpus contains references to about 20% of all the biomedical literature indexed in PubMed that had been published between 1950 and 2010, including all highly cited papers in every biomedical field.

Other open citations resources exist. The two main ones are *CiteSeerX*[23], which at the time of writing contained around 13,500,000 references from 1,242,041 articles, primarily in computer science; and *CitEc* (*Citations in Economics*)[24], which contained 13,544,970 references from 545,641 documents. Together, these resources and the OCC contain the references from some 1,980,000 articles – a mere 4% of the estimated 50 million scholarly articles that have ever been published.

4.2 From where do, will and should references come

We are currently revising the OCC data model, improving its hosting infrastructure, and expanding its coverage, both by updating the PubMed Central Open Access holdings, which have more than doubled since the initial ingest, and by ingesting citation data from the arXiv preprint server[25], thus adding citations in mathematics and the 'hard' sciences to augment the initial biomedical coverage. Future inclusion targets for citation data include CiteSeerX[23] and the Dryad Data Repository[26].

But references should ideally come from publishers directly. Most publishers are sympathetic to the idea of putting article reference lists outside the journal subscription pay wall, as they do for copyright abstracts.

We already have agreements with several major journal publishers for the future routine harvesting of their journals' reference data. As well as the 'pure' open access publishers, the references from which are open by definition, the publishers of subscription-access journals include Nature Publishing Group[27], Oxford University Press[28], the American Association for

the Advancement of Science[29] (which publishes *Science*), Royal Society Publishing[30], Portland Press[31], MIT Press[32] and Taylor & Francis[33], all of which will make references available either from some or from all of their journals. This represents a small but growing proportion of all the journal articles published in a year.

The references from current journal issues will be harvested from CrossRef[34], to which these publishers already submit article references as part of its free Cited-By Linking service[35]. However, since by default references are not exposed, these data can only be obtained from *CrossRef Metadata Search* **if publishers provide specific consent for this, which they can easily do within the article metadata**, as detailed in the new CrossRef metadata best practice document[36]. Additionally, to provide consent for CrossRef to release references from journal back numbers, each publisher needs to e-mail CrossRef (support@crossref.org), as detailed in David Shotton's "Open Letter to Publishers"[37], stating their willingness for their references to be open. No other action is required, making this extremely straightforward and cost free.

Our long-term aim is to host citation information for the majority of the world's scholarly literature, in the arts and humanities as well as the sciences. This will require major curatorial effort and underpinning technical innovation, on the scale of the National Library of Medicine's PubMed[13].

4.3 Services

In an ideal world, publishers would host their own bibliographic and citation data, following the example of Nature Publishing Group, which is exemplary in freely exposing the bibliographic and citation metadata for all its Nature titles, including all back numbers, in RDF on the NPG Linked Data Platform[38] (Carpenter, 2012).

But there are separate benefits to be gained from the aggregation of all such data into a single searchable corpus. The OCC will provide such integrated access to citation data from a variety of sources, both from within and from outside traditional scholarly publishing, with clear provenance data. It will expose entity relationships, including article-to-article, article-to-database and database-to-article citations, and, where the data are available, will reveal shared authorship and institutional membership, common funding, and semantic relationships between articles. The whole corpus is queryable through SPARQL (Harris and Seaborne, 2013), the standard W3C language to query RDF data stores.

Once citation data are openly available, useful analytical services can be built, including faceted search-and-browse tools, recommendation and trend identification services, and timeline visualization. Some of these we have already developed in prototype. The OCC's usefulness for calculating citation metrics will, of course, increase in proportion to its expanding coverage.

There is one additional service we envisage that could be of particular benefit to authors and editors – an erroneous reference correction service. About 1% of references in published papers contain errors of varying severity, from the trivial – for example, substitution of 'beta amylase' for 'β-amylase' in the reference title, or the omission of accents in author names – to the more serious – such as errors in the year, volume, page numbers or DOI. The OCC already uses citation correction methods internally for reference targets that are multiply cited, or for which authoritative bibliographic records can be obtained externally. A similar Web service provided by the OCC that could detect errors in up-loaded reference lists might significantly reduce the number of erroneous references in published papers.

4.4 The benefits of the OCC

An open and well-supported digital research infrastructure is crucial for facilitating research and innovation, and for promoting international growth and prosperity. By disrupting the present commercial stranglehold on scholarly bibliographic citation information, the Open Citations Corpus, a radical new element in the digital research infrastructure, provides a means of solving present problems, and will bring enormous benefits for scholarship.

Created by scholars for scholars using scholarly data, and with no profit motive constraining free publication, the OCC will directly benefit several distinct stakeholder communities. Clearly, it will benefit all scholars, particularly those who are not members of the elite club of research universities whose libraries can afford to purchase commercial citation data. These scholars will now be able to pursue their studies with greater freedom, following reference trails through the citation network without hindrance, and having their own publications more easily found, discussed and, hopefully, cited.

Publishers will also benefit considerably, since more readers will be readily guided to their online journal articles. Few people these days scan the Table of Contents of each issue of a journal as it is published. Rather, a researcher will typically come to an article by following a citation link, thinking it might be relevant to his or her line of enquiry. Thus, the more readily accessible such links, the greater the traffic to the articles. In addition, we anticipate that journals that are more readily discoverable will benefit by attracting additional article submissions.

Institutions and funding agencies, that wish to track the scholarly productivity and influence of their members/grantees, will also benefit by being able to do so more readily, once the bibliographic and citation data for these individuals are openly available in machine-readable form. It will also benefit research administrators, because the OCC data, semantically described in RDF, will be available for integration with other similarly described data resources, including research information encoded using *CERIF*, the *Common European Research Information Framework*[39]. Finally, it will benefit developers, who can exploit the freely available citation data to build new applications and visualizations that we cannot even begin to imagine.

5 Towards Semantic Publishing

5.1 Machine-readable metadata

In the modern world, in which academics are overwhelmed by the increasing number of relevant papers in their field, systems that undertake the automated processing of bibliographic and citation data, for example by filtering for relevance, become increasingly essential to support the scholarly endeavour. These requires machine-processable metadata relating to publications and citations, and these in turn require appropriate ontologies – structured descriptions of the scholarly communication, publishing and referencing domain – to permit these metadata to be encoded in RDF (Beckett, 2004), the *lingua franca* for interoperable linked data on the Web.

We are a very long way from such metadata being routinely available. As we already introduced in (Peroni and Shotton, 2012), there is a need for additional work in several fields to make Semantic Publishing (Shotton, 2009) a reality. In particular, to guarantee the advent of an open citation data era, we need to invest effort in:

- the development and adoption of semantic models (vocabularies, ontologies) that meet the requirements of scholarly authoring and publishing, and in particular that permit

the publication of bibliographic and citation data in a machine-readable form, e.g. RDF; and

- the development of annotation tools that allow authors to use these models to enhance documents with appropriate semantic assertions.

In the following sections we will introduce some of our recent contributions in this direction.

5.2 The SPAR (Semantic Publishing and Referencing) Ontologies

Machine-readable encoding of bibliographic and citation data, of citation contexts, of document components, and of the nature or type of individual citations, are now all made possible by a new set of OWL 2 ontologies (Motik, Patel-Schneider and Parsia, 2012), the *SPAR* (*Semantic Publishing and Referencing*) *Ontologies*[22], which we have created to describe different aspect of the scholarly publishing domain. The principal SPAR ontologies are summarized below.

CiTO*, the *Citation Typing Ontology[6] (Peroni and Shotton, 2012), previously introduced in Section 2, is an ontology written to enable the existence of bibliographic reference citations to be asserted, and their factual and rhetorical nature to be characterised. The citations characterised may be direct and explicit (as in the reference list of a journal article), indirect (e.g. a citation to a more recent paper by the same research group on the same topic), or implicit (e.g. as in artistic quotations or parodies, or in cases of plagiarism).

BiRO*, the *Bibliographic Reference Ontology[9], is an ontology structured according to the FRBR model (IFLA Study Group on the FRBR, 2009) that provides a logical system for describing an individual bibliographic reference, such as appears in the reference list of a published article, and its relationship to the complete bibliographic record for that cited article, which in addition to having metadata components (e.g. title, DOI) missing from the reference, may also include the name of the publisher, and the ISSN or ISBN of the publication. BiRO also permits one to describe the collection of individual bibliographic references into ordered reference lists, and the collection of bibliographic records into bibliographic record collections such as library catalogues.

C4O*, the *Citation Counting and Context Characterisation Ontology[10], provides the ontological structures required to permit one to record both the number of citations to a cited paper that exist in the citing paper (i.e. the number of in-text reference pointers to a single reference in the citing article's reference list), and also the number of citations that the cited paper has received globally up to a particular date, as determined by consulting an external citation information resource such as Microsoft Academic Search[3], Google Scholar[2], Scopus[4] or Web of Science[1]. Additionally, C4O can be used to define the context of a citation, i.e. the text within which an in-text reference pointer is embedded in the citing paper.

FaBiO*, the *FRBR-aligned Bibliographic Ontology[40] (Peroni and Shotton, 2012), is an ontology for recording and publishing on the Semantic Web descriptions of entities that are published or potentially publishable, particularly those that contain or are referred by bibliographic references. FaBiO entities are primarily textual publications such as books, magazines, newspapers and journals, and items of their content such as poems, conference papers and editorials. However, they also include blogs, web pages, datasets, computer algorithms, experimental protocols, formal specifications and vocabularies, legal records, governmental papers, technical and commercial reports and similar publications, and also anthologies, catalogues and similar collections. FaBiO uses terms from the RDF and OWL versions of FRBR[41] (IFLA Study Group on the FRBR, 2009), PRISM (Hammond, 2008), Dublin Core Metadata Elements (DCMI Usage Board, 2012b), Dublin Core Terms (DCMI Usage Board, 2012a), and SKOS (Miles and Bechhofer, 2009; Baker *et al.*, 2013).

DoCO*, the *Document Component Ontology[11], provides a structured vocabulary for describing document components, both structural (e.g. block, chapter, heading, inline, paragraph, section, text chunk) and rhetorical (e.g. Abstract, Introduction, Results, Discussion, Conclusions, Acknowledgements, Bibliography, Figure, Appendix), enabling these components, and documents composed of them, to be described in RDF.

PRO*, the *Publishing Roles Ontology[42] (Peroni *et al.*, 2012), is an ontology for the characterisation of the roles of agents (i.e. people, groups or organisations) in the publication process (e.g. author, editor, librarian, review panel, publisher), and for specifying the times during which those roles are held, and the contexts in which those roles are relevant.

PSO*, the *Publishing Status Ontology[43] (Peroni *et al.*, 2012), is an ontology for characterising the publication status of a document or other publication entity at each of the various stages in the publishing process (e.g. draft, submitted, under review, rejected, accepted for publication, proof, published, Version of Record, peer reviewed, *libre* open access, catalogued, archived). It can also be used for specifying the times during which those statuses are held, the events that trigger a transition from one status to the next, and the people involved in those events.

Finally, ***PWO*, the *Publishing Workflow Ontology***[44], is an ontology for characterising the main stages in the workflow associated with the publication of a document (e.g. being written, under review, XML capture, page design, publication to the Web), and for specifying the input and output necessary for each step, and the timings and events with which each step is associated. It is designed specifically to describe workflows that have happened, rather than to provide decision trees for future workflows.

These eight core SPAR ontologies, and additional ontological modules not described here, are discussed in more detail in the *Semantic Publishing Blog*[45].

6 New tools and interfaces to handle citations better

Scholarly communication, at this mid-point in the digital revolution, is in an ill-defined transitional state that lies somewhere between the world of print and paper and the world of the Web and computers, with the former still exercising significantly more influence than the latter. We now publish online, but five years after our own semantic publishing exemplar was put online[46] (Shotton *et al.*, 2009), the majority of papers are still published in the form of static PDF documents. We need *tools* to facilitate the transition of the online paper from its present 'horseless carriage' state into the scholarly communication equivalent of a Ferrari.

Of course, the communities working on Semantic Publishing technologies have already produced some excellent results in this direction, which we review in this section.

6.1 Utopia documents, ALC and the Pensoft Writing Tool

A valuable example of a semantic enhancement tool is *Utopia Documents*[47] (Attwood *et al.*, 2010), a novel 'smart' PDF reader that semantically integrates visualization and data-analysis tools with static published PDF research articles. It brings such PDF documents to life by linking to live resources on the web and by turning the static numerical data they contain into live interactive content. It is now regularly used by the editors of the Biochemical Journal[48] (Attwood *et al.*, 2009) to transform static document features into objects that can be linked, annotated, visualised and analysed interactively.

Utopia Documents includes a mechanism for deconstructing a PDF document into its constituent parts, which are then annotated using DoCO[11]. This is useful for a number of

things: generating bibliometric metadata within an article; improving "mouse selection" in multi-column documents; and identifying the correct flow of text in the document, thus allowing annoying intruding text such as running headers, footers and captions to be excluded. This in turn is useful for text and data mining algorithms, which can now be targeted, for example, at "all the main text excluding intruders" or "just the text in the figure captions". Recently, the Utopia Documents team released a free web service called PDFX[49] (Constantin *et al.*, 2013) that takes a PDF document, deconstructs it, and returns a JATS-based[50] XML document annotated with DoCO terms.

Another important work, called the ACL Anthology Network[51] (Radev *et al.*, 2013), has been developed by Mark Thomas Joseph, Amjad Abu-Jbara, Dragomir Radev and other members of the University of Michigan's CLAIR Group[52]. They have collected (and continue to collect) bibliographic entity descriptions and related citations from 19,647 papers (as of October 2013) from the Association of Computational Linguistics Anthology dataset[53]. The collected data can be browsed, and used to explore citation networks, authors' citations and authors' collaboration networks.

Additional to these researcher-led developments, some publishers also are making significant progress in the semantic annotation of scholarly documents. Of these, the most progressive is Pensoft[54], a small publisher in the area of biological taxonomy, whose open access online journals carry a significant number of semantic enhancements. Its most recently launched journal, the Biodiversity Data Journal (BDJ)[55], whose Editor-in-Chief is Vincent Smith, Cybertaxonomist at The Natural History Museum in London, uses a new collaborative authoring environment, the Pensoft Writing Tool[56], that is fully integrated into the journal's reviewing and publishing pipeline. This permits automated semantic markup of text and data during the writing process, with no additional effort for the authors, as described at http://www.pensoft.net/services-for-journals.

The Pensoft Writing Tool also permits references to be added automatically to the reference list in the appropriate format, once the cited papers are identified by their DOI or PubMed ID, for references to be parsed automatically into XML from pasted text, or for references to be entered manually. Furthermore, the associated in-text reference pointers can be inserted automatically into the text at the cursor position in the required format using the "Cite a reference" command.

6.2 The CiTO Reference Annotation Tools

It is naïve to imagine that authors will undertake the additional work required to specify the typing of citations in their journal articles using CiTO's citation typing properties, unless the appropriate authoring tools are provided. A step towards this goal has been the development of the CiTO Reference Annotation Tools[7], which permits CiTO properties to be selected from a drop-down list when viewing the reference in a reference list.

Implementation of CiTO Reference Annotation can be achieved in two ways. For online articles encoded in XML using the National Library of Medicine Journal Publishing DTD[57], for example those in PubMed Central[58], PLoS Currents[59], eLife[60] and ZooKeys[61], the reader need only install a simple CiTO Chrome Extension[62], which then provides this functionality, as shown in Figure 5.

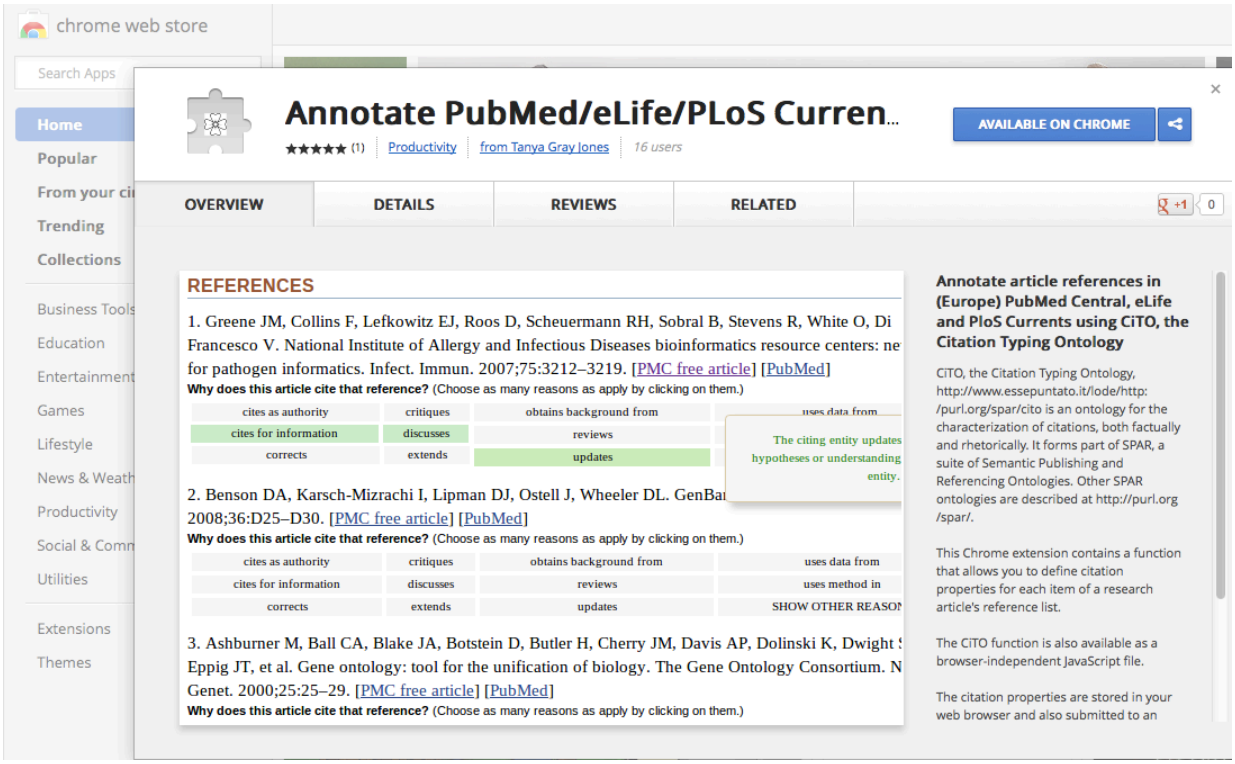

Figure 5. The homepage of the CiTO Chrome Exension, showing how to add CiTO annotations to references in the reference list of (Brinkac *et al*., 2009).

Alternatively, a publisher or user can insert three lines of XML code into an original published article from PubMed Central, PLoS Currents, eLife or ZooKeys, and view it in combination with the Javascript library provided by the second CiTO Reference Annotation Tool.

In either case, as shown in Figure 5, selected CiTO properties are highlighted in green. To ensure the user fully understands the meaning of each property, its textual definition is displayed in a pop-up when the mouse hovers over the property label in the table.

Clearly, this open source software needs to be developed further, so that CiTO properties can be associated not just with a reference in a reference list, but more specifically with each individual in-text reference pointer in the body text, since an article may be cited for different purposes at different places within the citing work, for example for general background information in the Introduction of the citing work, and to provide details of an experimental procedure in the Methods section.

It would be desirable to integrate such CiTO Reference Annotation functionality into an online authoring tool such as the Pensoft Writing Tool, but that has yet to be achieved.

6.3 Recommending citation types using CiTalO

Even with such tools, typing citations is an additional task that at present needs to be undertaken manually, ideally by an author, or alternatively by other interested parties such as editors, publishers or readers. In the latter case, for example when attempting to annotate previously published articles, "the citation function is hard to annotate because it in principle requires interpretation of author intentions (what could the author's intention have been in choosing a certain citation?)" (Teufel *et al.*, 2006). In addition, the person annotating a citation according to a particular ontology such as CiTO, has "to associate a particular meaning to each of the various functions defined" in it, and "has to map [...] the personal interpretation

of author's intention emerging from a written text containing a citation to one of the functions of the citation model" (Ciancarini *et al.*, 2013a).

It is thus of great significance that the first steps have been taken to automate this citation typing task, by employing computational natural language processing to determine the rhetorical meaning of the citation contexts, so as to recommend to the user (either the author or someone else) the most appropriate citation description(s) that should be associated with a particular citation.

We have accomplished this by developing *CiTalO* (Di Iorio *et al.*, 2013a; Di Iorio *et al.*, 2013b). (This name was obtained by merging the English words "**CiTO**" and "**al**gorithm"; "citalo" also happens to be an Italian word meaning "cite it"). CiTalO is an algorithmic tool that tries to infer an author's reasons for citing a particular paper by using the techniques of ontology learning and mapping, natural language processing, sentiment analysis, and word-sense disambiguation. The tool, which is available online[8], takes as input a sentence containing a citation, and returns a set of CiTO properties that best characterise that citation. In addition, CiTalO has been recently extended[63] to perform the automatic extraction and semantic characterisation of all the citations contained within an entire scientific paper stored as a PDF file (Ciancarini *et al.*, 2013b).

7 Building Venice: a vision

Imagine that we are trying to create the bustling city of Venice – the city of scholarly communication – from a collection of islands that represent individual scholarly publications. Citations are the bridges that enable people to pass from one island (e.g. a conference paper) to others (e.g. journal articles and book chapters). At present, while local travel to the next island is permitted, unrestricted travel over the entire network of bridges requires an expensive season ticket, affordable only by rich professionals. The general populace is excluded, and as a consequence the social and commercial growth of the city is stunted. In contrast, were the bridges to be opened without fee to the general populace, people would be able to travel freely through the whole city of scholarly knowledge, and the community would thrive.

The Open Citations Corpus is an attempt to provide these open bridges. Computers operating over the open machine-readable citation data within the OCC will enable us to build a flourishing Venice of knowledge services from the complex archipelago of separate publication islands, adding value and creating a whole that is greater than the sum of its parts.

But this is only part of the story. Anyone who has visited Venice knows how easy it is for a stranger to get lost, particularly if the road signs are not helpful, as is frequently the case (see Figure 6).

Figure 6. All Roads Lead to Piazza San Marco. © Philip Lakin 2009. CC BY-NC-ND 2.0. http://www.flickr.com/photos/dereklakin/3701211442/.

There is a parallel in the world of citations. When the citation links exist but are not described, the traveller through the city of scholarly publishing travels without specific directions, and will get lost in the maze of bridges that form the citation network. CiTO can be used to add informative citation typing "street signs" that facilitate travel in the desired direction.

8 Conclusions

In this article, we have presented the Open Citation Corpus (OCC), born from the need to make scholarly citation data open access, the SPAR Ontologies (including CiTO, the Citation Typing Ontology) that support that work, and some prototype tools that facilitate the annotation of references with CiTO properties, thus enabling the description of the purpose of a citation. Our hope is that scholars, publishers and institutions will build upon, enhance and reuse the citation data contained within the OCC. We also seek their assistance to increase the content of the OCC; to further develop user interfaces for searching, browsing and visualizing the network of scholarly citations; and to enhance tools such as the CiTO Reference Annotation tools that facilitate the annotation of citations by humans, aided by an automated recommendation system, CiTalO, that assists that process by proposing the most appropriate CiTO properties to describe the nature and purpose of citations from analysis of their textual contexts.

Acknowledgements

This article has been developed from the same textual source material from which was distilled a short Comment piece entitled "Open Citations" recently published by David Shotton in *Nature* (Shotton, 2013). It thus has substantial textual elements in common with that publication.

We gratefully acknowledge the financial support of Jisc, which provided two small grants to David Shotton that, in addition to enabling the creation of the Open Citations Corpus of which he is the Director, in part also made possible his development of the SPAR ontologies in collaboration with Silvio Peroni, and of the CiTO Reference Annotation Tools in collaboration with Tanya Gray. The software development of the first public prototype of Open Citations

Corpus was primarily undertaken by Alexander Dutton during the first Jisc grant. Work currently in progress to revise the data model, infrastructure and ingest pipeline of the OCC was initiated during the second Jisc grant, in collaboration with Richard Jones, Mark Macgillivray and Martyn Whitwell of Cottage Labs, acting as development consultants, who are sincerely thanked for their excellent work.

Silvio Peroni would like to thank Angelo Di Iorio and Andrea Giovanni Nuzzolese, who co-authored CiTalO, and Paolo Ciancarini and Fabio Vitali for their help and for many fruitful and proactive discussions about citations, citation functions and citation metrics.

References

Aksnes, D. W. (2003). "A macro study of self-citation. Scientometrics", 56(2): 235–246. DOI: 10.1023/A:1021919228368.

Agarwal, S., Choubey, L., & Yu, H. (2010). "Automatically Classifying the Role of Citations in Biomedical Articles". In Proceedings of the 2010 AMIA Annual Symposium: 11–15.

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D. (2010). "Utopia documents: linking scholarly literature with research data". Bioinformatics, 26(18): i568–i574. DOI: 10.1093/bioinformatics/btq383.

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D. (2009). "Calling International Rescue: knowledge lost in literature and data landslide!". Biochemical Journal, 424(3): 317–333. DOI: 10.1042/BJ20091474.

Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., & Summers, E. (2013). "Key choices in the design of Simple Knowledge Organization System (SKOS)". Web Semantics: Science, Services and Agents on the World Wide Web, 20: 35–49. DOI: 10.1016/j.websem.2013.05.001.

Beckett, D. (2004). "RDF/XML Syntax Specification (Revised)". W3C Recommendation, 10 February 2004. World Wide Web Consortium. Retrieved October 7, 2013, from http://www.w3.org/TR/rdf-syntax-grammar/.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). "The Eigenfactor™ Metrics". Journal of Neuroscience, 28(45): 11433–11434. DOI: 10.1523/JNEUROSCI.0003-08.2008.

Boulton, G. (Chair) "Science as an Open Enterprise". Royal Society Report. Retrieved October 14, 2013, from http://royalsociety.org/policy/projects/science-public-enterprise/report/.

Bourne, P. (2005). "Will a Biological Database Be Different from a Biological Journal?". PLoS Computational Biology, 1(3): e34. DOI: 10.1371/journal.pcbi.0010034.

Brinkac, L. M., Davidsen, T., Beck, E., Ganapathy, A., Caler, E., Dodson, R. J., … Sutton, G. (2009). "Pathema: a clade-specific bioinformatics resource center for pathogen research". Nucleic Acids Research, 38(Database): D408–D414. DOI: 10.1093/nar/gkp850.

Carpenter, P. (2012). "Nature Publishing Group releases linked data platform". Retrieved October 7, 2013, from http://www.nature.com/press_releases/linkeddata.html.

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi M., & Ebrahim, M. A. (2013). "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases". Asian Social Science 9: 18-26. DOI: 10.5539/ass.v9n5p18.

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2013a). "Characterising citations in scholarly articles: an experiment". To appear in A. Lieto, & M. Cruciani (Eds.), Proceedings the 1st International Workshop on Artificial Intelligence and Cognition (AIC

2013). Postprint available at http://speroni.web.cs.unibo.it/publications/ciancarini-2013-characterising-citations-scholarly.pdf.

Ciancarini, P., Di Iorio, A., Nuzzolese, A. G., Peroni, S., & Vitali, F. (2013b). "Semantic Annotation of Scholarly Documents and Citations". To appear in M. Baldoni, C. Baroglio, & G. Boella (Eds.), Proceedings of 13th Conference of the Italian Association for Artificial Intelligence (AI*IA 2013), Lecture Notes in Computer Science. Berlin, Heidelberg, Germany: Springer. Postprint available at http://speroni.web.cs.unibo.it/publications/ciancarini-2013-semantic-annotation-scholarly.pdf.

Constantin, A., Pettifer, S., & Voronkov, A. (2013). "PDFX: fully-automated PDF-to-XML conversion of scientific literature". In Proceedings of the 2013 ACM symposium on Document Engineering (DocEng 2013): 177–180. New York, New York, US: ACM Press. DOI: 10.1145/2494266.2494271.

DCMI Usage Board. (2012a). DCMI Metadata Terms. DCMI Recommendation, 14 June 2012. Dublin Core Metadata Initiative. Retrieved October 7, 2013, from http://dublincore.org/documents/dcmi-terms/.

DCMI Usage Board. (2012b). Dublin Core Metadata Element Set, Version 1.1. DCMI Recommendation, 14 June 2012. Dublin Core Metadata Initiative. Retrieved October 7, 2013, from http://dublincore.org/documents/dces/.

Di Iorio, A., Nuzzolese, A. G., & Peroni, S. (2013a). "Characterising Citations in Scholarly Documents: The CiTalO Framework". In P. Cimiano, M. Fernández, V. Lopez, S. Schlobach, & J. Völker (Eds.), ESWC 2013 Satellite Events - Revised Selected Papers, Lecture Notes in Computer Science 7955: 66–77. Berlin, Heidelberg, Germany: Springer. DOI: 10.1007/978-3-642-41242-4_6.

Di Iorio, A., Nuzzolese, A. G., & Peroni, S. (2013b). "Towards the automatic identification of the nature of citations". In A. Garcia Castro, C. Lange, P. Lord, & R. Stevens (Eds.), Proceedings of 3rd Workshop on Semantic Publishing (SePublica 2013), CEUR Workshop Proceedings 994: 63–74. Aachen, Germany: CEUR-WS.org. Retrieved October 7, 2013, from http://ceur-ws.org/Vol-994/paper-06.pdf.

Finch, J. (Chair) (2012) "Expanding Access to Published Research – The Finch Report". Report of the UK Government Working Group on Expanding Access to Published Research. Retrieved October 14, 2013, from http://www.researchinfonet.org/publish/finch/.

Franceschet, M. (2009). "A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar". Scientometrics, 83(1): 243–258. DOI: 10.1007/s11192-009-0021-2.

Garfield, E. (2006). "The History and Meaning of the Journal Impact Factor". Journal of the American Medical Association, 295(1): 90. DOI: 10.1001/jama.295.1.90.

Greenberg, S. A. (2009). "How citation distortions create unfounded authority: analysis of a citation network". BMJ, 339(jul20 3): b2680–b2680. DOI: 10.1136/bmj.b2680.

Hahnel, M. (2013). "The reuse factor". Nature 502: 298-298. DOI: 10.1038/502298a.

Hammond, T. (2008). "RDF Site Summary 1.0 Modules: PRISM". Retrieved October 7, 2013, from http://nurture.nature.com/rss/modules/mod_prism.html.

Harnad, S., & Brody, T. (2004). "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals". D-Lib Magazine, 10(6). DOI: 10.1045/june2004-harnad.

Harris, S., & Seaborne, A. (2013). “SPARQL 1.1 Query Language”. W3C Recommendation, 21 March 2013. World Wide Web Consortium. Retrieved October 7, 2013, from DOI: http://www.w3.org/TR/sparql11-query/.

IFLA Study Group on the FRBR. (2009). “Functional Requirements for Bibliographic Records”. International Federation of Library Associations and Institutions. Retrieved October 7, 2013, from http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

Jörg, B. (2008). “Towards the Nature of Citations”. In Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems (FOIS 2008). Retrieved October 7, 2013, from http://www.dfki.de/~brigitte/publications/FOIS08_Poster_BrigitteJoerg.pdf.

Lawrence, D. (2001). “Free online availability substantially increases a paper’s impact”. Nature, 411(6837): 521. DOI: 10.1038/35079151.

Liu, Y., & Rousseau, R. (2013). Interestingness and the essence of citation. Journal of Documentation, 69(4): 580–589. DOI: 10.1108/JD-07-2012-0082.

MacRoberts, M. H., & MacRoberts, B. R. (1989). “Problems of citation analysis: A critical review”. Journal of the American Society for Information Science, 40(5): 342–349. DOI: 10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U.

Meyer, B., Choppy, C., Staunstrup, J., & van Leeuwen, J. (2009). “Research evaluation for computer science”. Communications of the ACM, 52(4): 31. DOI: 10.1145/1498765.1498780.

Miles, A., & Bechhofer, S. (2009). “SKOS Simple Knowledge Organization System, Reference”. W3C Recommendation, 18 August 2009. World Wide Web Consortium. Retrieved October 7, 2013, from http://www.w3.org/TR/skos-reference/.

Moravcsik, M. J., & Murugesan, P. (1975). “Some Results on the Function and Quality of Citations”. Social Studies of Science, 5(1): 86–92. DOI: 10.1177/030631277500500106.

Motik, B., Patel-Schneider, P. F., & Parsia, B. (2012). “OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax (Second Edition)”. W3C Recommendation, 11 December 2012. World Wide Web Consortium. Retrieved on October 7, 2013, from http://www.w3.org/TR/owl2-syntax/.

Peroni, S., Motta, E., & d’Aquin, M. (2008). “Identifying Key Concepts in an Ontology, through the Integration of Cognitive Principles with Statistical and Topological Measures”. In J. Domingue & C. Anutariya (Eds.), Proceedings of the 3rd Asian Semantic Web Conference (ASWC 2008), Lecture Notes in Computer Science 5367: 242–256. Berlin, Heidelberg, Germany: Springer. DOI: 10.1007/978-3-540-89704-0_17.

Peroni, S., & Shotton, D. (2012). “FaBiO and CiTO: Ontologies for describing bibliographic resources and citations”. Web Semantics: Science, Services and Agents on the World Wide Web, 17: 33–43. DOI: 10.1016/j.websem.2012.08.001.

Peroni, S., Shotton, D., & Vitali, F. (2012). “Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents”. In H. Sack & T. Pellegrini (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9–16. New York, New York, US: ACM Press. DOI: 10.1145/2362499.2362502.

Piwowar, H., Day, R., & Fridsma, D., (2007). “Sharing detailed research data is associated with increased citation rate”. PLoS ONE 2 (3), e308. DOI: 10.1371/journal.pone.0000308.

Piwowar, H., Vision, T.J., (2013). “Data reuse and the open data citation advantage”. PeerJ 1, e175. DOI: 10.7717/peerj.175.

Piwowar, H. (2013). "Altmetrics: value all research products". Nature 493, 159-159 (2013). DOI: 10.1038/493159a.

Priem, J., (2013). "Scholarship: Beyond the Paper". Nature 495: 437-440. DOI: 10.1038/495437a.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). "Altmetrics: a manifesto". Retrieved October 7, 2013, from http://altmetrics.org/manifesto.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). "The ACL anthology network corpus". To appear in Language Resources and Evaluation. DOI: 10.1007/s10579-012-9211-2.

Reis, R. B., Ribeiro, G. S., Felzemburgh, R. D. M., Santana, F. S., Mohr, S., Melendez, A. X. T. O., … Ko, A. I. (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.

Research Councils UK (2013). "RCUK Policy on Open Access". Retrieved October 14, 2013, from http://www.rcuk.ac.uk/documents/documents/RCUKOpenAccessPolicy.pdf.

Roemer, R. C., & Borchardt, R. (2012). "From bibliometrics to altmetrics: a changing scholarly landscape". College & Research Libraries News, 73(10): 596–600. Retrieved October 14, 2013, from http://crln.acrl.org/content/73/10/596.full.

Shotton, D. (2009). "Semantic publishing: the coming revolution in scientific journal publishing". Learned Publishing, 22(2): 85–94. DOI: 10.1087/2009202.

Shotton, D. (2013). "Open citations". Nature, 502: 295-297. DOI: 10.1038/502295a.

Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article". PLoS Computational Biology, 5(4): e1000361. DOI: 10.1371/journal.pcbi.1000361.

Shotton, D. M., White, N. J., & Watson, H. C. (1972). "Conformational Changes and Inhibitor Binding at the Active Site of Elastase". Cold Spring Harbor Symposia on Quantitative Biology, 36: 91–105. DOI: 10.1101/SQB.1972.036.01.015.

Swan, A. (2009). "The Open Access citation advantages: Studies and results to date". School of Electronics & Computer Science, University of Southampton. Retrieved October 7, 2013, from http://eprints.ecs.soton.ac.uk/18516/.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). "Automatic classification of citation function". In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 06): 103–110. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics.

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). "An annotation scheme for citation function". In Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue: 80–87. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics. Retrieved October 7, 2013, from http://acl.ldc.upenn.edu/W/W06/W06-1613.pdf.

Vision, T. J. (2010). "Open Data and the Social Contract of Scientific Publishing". BioScience 60(5): 330-330. DOI: 10.1525/bio.2010.60.5.2.

Wellcome Trust (2013). "Open Access Policy Statement". Retrieved October 14, 2013, from http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm.

Footnotes

[1] Thomas Reuter's Web of Science: http://thomsonreuters.com/web-of-science.
[2] Google Scholar: http://scholar.google.com.
[3] Microsoft Academic Search: http://academic.research.microsoft.com.
[4] Scopus: http://www.scopus.com.
[5] Open Citation Corpus (OCC): http://opencitations.org.
[6] CiTO, the Citation Typing Ontology: http://purl.org/spar/cito.
[7] CiTO Reference Annotation Tools: https://github.com/tgra/cito.
[8] CiTalO: http://wit.istc.cnr.it:8080/tools/citalo.
[9] BiRO, the Bibliographic Reference Ontology: http://purl.org/spar/biro.
[10] C4O, the Citation Counting and Context Characterisation Ontology: http://purl.org/spar/c4o.
[11] DoCO, the Document Components Ontology: http://purl.org/spar/doco.
[12] Digital Object Identifier System: http://www.doi.org.
[13] PubMed: http://www.ncbi.nlm.nih.gov/pubmed.
[14] An extended list of possible identifiers that can be used to refer to articles and users is defined by the DataCite ontology (http://purl.org/spar/datacite, class *IdentifierScheme*).
[15] Journal of Documentation: http://www.emeraldinsight.com/products/journals/journals.htm?id=jd.
[16] Journal of Web Semantics: http://www.journals.elsevier.com/journal-of-web-semantics.
[17] The CiTO properties *cito:citesAsEvidence* and *cito:obtainsSupportFrom* are two examples of properties which are both factual and rhetorical in character.
[18] Functions of Citations: http://www.essepuntato.it/2013/03/cito-functions.
[19] Creative Commons CC0 1.0 legal code: http://creativecommons.org/publicdomain/zero/1.0/legalcode.
[20] PubMed Central Open Access Subset: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.
[21] Jisc: www.jisc.ac.uk.
[22] Semantic Publishing and Referencing (SPAR) Ontologies: http://purl.org/spar.
[23] CiteSeerX: http://citeseerx.ist.psu.edu.
[24] CitEc, Citations in Economics: http://citec.repec.org.
[25] arXiv preprint server: http://arxiv.org.
[26] Dryad: http://datadryad.org.
[27] Nature Publishing Group: http://www.nature.com.
[28] Oxford University Press: http://global.oup.com.
[29] American Association for the Advancement of Science: http://www.aaas.org.
[30] Royal Society Publishing: http://royalsocietypublishing.org.
[31] Portland Press: http://www.portlandpress.com.
[32] MIT Press: http://mitpress.mit.edu.
[33] Taylor and Francis: http://www.taylorandfrancis.com.
[34] CrossRef: http://crossref.org.
[35] CrossRef cited-by linking service: http://www.crossref.org/citedby/index.html.
[36] CrossRef metadata best practice to support key performance indicators (KPIs) for funding agencies: http://fundref.crossref.org/docs/funder_kpi_metadata_best_practice.html.
[37] Open letter to publishers: http://opencitations.wordpress.com/2013/01/03/open-letter-to-publishers/.
[38] Nature.com Linked Data: http://data.nature.com.

[39] CERIF Introduction: http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1.
[40] FaBiO, the FRBR-aligned Bibliographic Ontology: http://purl.org/spar/fabio.
[41] Essential FRBR in OWL2 DL: http://purl.org/spar/frbr.
[42] PRO, the Publishing Roles Ontology: http://purl.org/spar/pro.
[43] PSO, the Publishing Status Ontology: http://purl.org/spar/pso.
[44] PWO, the Publishing Workflow Ontology: http://purl.org/spar/pwo.
[45] Semantic Publishing Blog: http://semanticpublishing.wordpress.com.
[46] Semantically enhanced version of (Reis *et al.*, 2008): http://dx.doi.org/10.1371/journal.pntd.0000228.x001.
[47] Utopia Documents: http://www.utopiadocs.com.
[48] Biochemical Journal: http://www.biochemj.org.
[49] PDFX: http://pdfx.cs.man.ac.uk.
[50] Journal Article Tag Suite: http://jats.nlm.nih.gov.
[51] ACL Anthology Network: http://clair.eecs.umich.edu/aan/index.php.
[52] Computational Linguistics And Information Retrieval (CLAIR) group at the University of Michigan: http://clair.eecs.umich.edu/aan/about.php.
[53] ACL Anthology: http://aclweb.org/anthology/.
[54] Pensoft Publishers: http://www.pensoft.net.
[55] Biodiversity Data Journal: http://biodiversitydatajournal.com.
[56] Pensoft Writing Tool: http://pwt.pensoft.net.
[57] National Library of Medicine Journal Publishing DTD: http://dtd.nlm.nih.gov/publishing/2.3/index.html.
[58] PubMedCentral: http://www.ncbi.nlm.nih.gov/pmc/.
[59] PLoS Currents: http://currents.plos.org/.
[60] eLife: http://www.elifesciences.org.
[61] ZooKeys: http://www.pensoft.net/journals/zookeys/.
[62] Annotate journal citations with CiTO (Google Chrome Extension): https://chrome.google.com/webstore/detail/annotate-journal-citation/geajighoohelnjnhfmhbcaddbcgcbphn.
[63] CiTalO^{PDF}: http://wit.istc.cnr.it:8080/tools/citalo/pdf.