CiTO + SWAN: The Web Semantics of Bibliographic Records, Citations, Evidence and Discourse Relationships

Paolo Ciccarese[a,b *], David Shotton[c *], Silvio Peroni[c,d] and Tim Clark[a,b §]

[a] Massachusetts General Hospital, Department of Neurology, Mindinformatics, 65 Landsdowne Street, Cambridge, 02139 MA, USA.
[b] Harvard Medical School, Boston, MA, USA.
[c] Image Bioinformatics Research Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.
[d] Department of Computer Science, University of Bologna, Mura Anteo Zamboni 7, 40127 Bologna (BO), Italy
[*] These authors contributed equally to this work.
[§] Corresponding author. Email: tim_clark@harvard.edu

Abstract. Most literature searching in biomedicine is now conducted via PubMed, Google Scholar or other web-based bibliographic search mechanisms. Yet until now a public, open, interoperable and complete web-adapted information schema for bibliographic citations, bibliographic references and scientific discourse has not been available. Such a schema, expressed in the form of a description logic compatible with current web semantics approaches, would provide the ability to treat bibliographic references and citations, and rhetorical discourse in scientific publications, as semantic metadata on the web, with all the benefits that implies for organization, search and mash-up of web-based scientific information.

In this paper we present CiTO + SWAN, a set of fully harmonized ontology modules resulting from the harmonization of CiTO (the Citation Typing Ontology) with SWAN (Semantic Web Applications in Neuromedicine), which we have developed by jointly adapting and evolving version 1.6 of CiTO, the Citation Typing Ontology, and version 1.2 of the SWAN Scientific Discourse Ontology (v1.2). The CiTO + SWAN model is specified in OWL 2 DL, is fully modular, and inherently supports agent-based searching and mash-ups.

Through the harmonization activity presented here, and previous work that harmonized SWAN with the SIOC (Semantically-Interlinked Online Communities) Ontology for describing blogs, wikis and discussion groups, we have construct the basis of a powerful new web framework for scientific communications.

Keywords: Bibliographic ontology, FRBR, scientific discourse, OWL

1 Introduction

Motivation

The web is now the primary platform by which biomedical scientists find, retrieve and share textual information and, in certain cases, research data. Most literature searching in biomedicine is now conducted via PubMed, Google Scholar or some other web-based bibliographic search mechanism. Such approaches have substantially replaced physical searching in library stacks for periodicals, and PubMed is now processing in the range of 60 million web searches per month [1].

Each web-based (or other electronic) search method constructs and uses an information schema to represent the cataloguing metadata for publications, and has methods for interrogating and displaying these metadata. However, so long as the schemas for these metadata are either (a) not made publicly available, (b) encoded in *sui generis* representations, and/or (c) made available programmatically only in application-specific APIs (if at all), we all have a problem. The problem is that each repository of metadata and publication contents – and each application for searching and processing them – must stand alone as an information island.

While stand-alone programs and databases were reasonable in the pre-web world, this is no longer the case. Modern web programming depends on reuse and extensive cross-linking of information. For scientific information in particular, we need the ability to mash-up (integrate) and query data and metadata across multiple repositories, using computer programs to undertake this work automatically, as in e.g. Miles *et al*. [2]. The style of programming current on the web which best supports this involves (a) surfacing data and metadata in machine readable standard form such as REF and OWL [3-6], and (b) making data and metadata queries available to programs via standard RESTful APIs [7] or as SPARQL endpoints (query interfaces) [8]. A RESTful API for SPARQL endpoints, suitable for Linked Data applications, has recently been published [9].

Because of the centrality of scientific documents to the social processes and practice of science, it is of fundamental importance to have available robust information schemas for bibliographic citations, bibliographic references and scientific discourse in the form of proper OWL ontologies. Such ontologies would enable inference about the structure and provenance of the collective scientific discourse presented in academic journals and conference proceedings, which are the media through which new discoveries in science have been presented since the mid-seventeenth century. Conversely, lack of such schemas naturally inhibits development of agent-based search and mash-up capabilities for scientific discourse in Web 3.0 style [10]. Scientists and software developers who wish to construct web applications supporting machine-readable metadata about science publications must have appropriate ontologies to support these activities.

Ontologies

An ontology encapsulates formal specifications of concepts within a particular domain of knowledge, and the relationships between them, in a machine-readable manner. Ontologies designed for the semantic web have two principal functions. First, they enable the collective development of controlled terminology systems with natural language definitions of terms and properties, enabling communities to be secure in the knowledge that they are talking about the same things when using them to engage in structured conversations about particular domains of knowledge, thereby extending the advantages of controlled vocabularies for the creation of metadata. Such terminological systems are both objects of collaboration in their own right, and enable further collaboration. Additionally, because the DL

logic upon which ontologies are frequently based permits a greater logical structure than that possessed by simple controlled vocabularies or taxonomies, and because the metadata encoded using them can be processed by computer, such ontologies permit automation of logical inferencing about the items under discussion.

Within the Semantic Web community, ontologies are customarily recorded using OWL, the Web Ontology Language defined by the World Wide Web Consortium [5, 6]. If ontological terms are defined by unique International Resource Identifiers (IRIs) [11], metadata created using them consistently contribute to what has become known as the Web of Linked Data [12], and may be integrated from disparate sources while preserving consistency of meaning.

Ontology modularization and ontology normalization

A *suite of ontologies* may be defined as a number of ontologies created to complement one another and work together in their coverage of different aspects of a particular domain of knowledge. Such a suite can be described as an ontology ecosystem.

One ontology within such a suite of ontologies, which exists as a complete, internally consistent and independent ontology object (saved, for example, as a unique OWL file), may be described as an *ontology module*. For example, GO, the Gene Ontology [13], comprises three ontology modules: the GO Cellular Component Ontology, the GO Biological Process Ontology, and the GO Molecular Function Ontology. It is in this sense of being a component ontology within a suite of ontologies that the term *ontology module* is used in this paper, in contrast to the alternative use of the term to mean a sub-section within a single ontology.

The activity of reorganizing a complex ontology into a suite of simpler complementary ontology modules is described as *ontology modularization*. *Ontology normalization* is a related activity that involves ensuring that each such ontology module is represented as a single subsumption (*is_a*) hierarchy. The advantages of ontology modularization and normalization are the following:

- Small and cohesive ontologies are easier to create, verify, maintain and understand. An iterative development method can be applied, and those interested in one single aspect of the domain can focus on a specific module without having to understand the architecture and details of the entire suite of ontologies.
- Modularization allows the separate independent reuse of individual components of the ontology suite.
- Modularization allows ontology module swapping when needed. Given their high cohesion and low coupling, it is easy to remove a module and substitute a third-party domain ontology covering the same topic.

For further discussion of the advantages of ontology normalization and modularization, see [14] and [15].

The need for ontology harmonization

As Semantic Web activities accelerate, new ontologies are being independently created to cover an expanding range of knowledge domains. The ideal is that separate ontologies should be *orthogonal* to one another, covering complementary domains and fitting together without overlap like pieces of a jigsaw puzzle or patchwork quilt. However, in reality this is not always the case, since inevitably some of these new ontologies overlap in scope.

In such situations, 'harmonization' between pairs of individual ontologies or suites of related ontologies may be required, to remove overlap and enable these ontologies to be used in conjunction without logical ambiguity or conflict. Such harmonization activities are usually undertaken collaboratively by the groups responsible for authoring the respective ontologies.

Best practice guidelines

Best practice guidelines for creating ontologies are given in the OBO (Open Biological Ontologies) Foundry Principles [16]. In summary, each ontology:

- should be open for use by all;
- should be expressed in a common shared syntax such as OWL;
- should possess a unique identifier space (namespace);
- should be published in distinct successive versions;
- should have clearly specified and delineated content;
- should be orthogonal to other ontologies;
- should include textual definitions for all terms;
- should use relationships (object and data properties) that are unambiguously defined;
- should be well documented;
- should serve a plurality of independent users; and
- should be developed collaboratively.

While our ontologies are not themselves currently housed within the OBO Foundry, we have taken these principles as our guide for ontology development. The first eight technical points have all been implemented, the description of the ontologies in peer-reviewed journal papers is an ongoing process of which this paper is part, and we are working to establish open communities of developers for the ongoing support and development of our ontologies, and of users for their widespread application.

The ontology harmonization activity described in this paper, and its purpose

This paper reports the processes and results of ontology harmonization activity between two suites of ontologies. The first of these, the SWAN (Semantic Web Applications in Neuromedicine) Ontologies [17, 18, 19], covers the domain of scientific discourse in general, with particular application to neuromedicine, while the second, the SPAR (Semantic Publishing and Referencing) Ontologies [20], which have been developed from CiTO (the Citation Typing Ontology) [21], describe the domain of scientific publishing and referencing.

The purpose of these ontologies is to provide controlled vocabularies and logical structures for the items of discourse surrounding bibliographic entities, references and citations, and the entities and processes involved in scientific discourse more generally, in which researchers use experimental evidence to support or refute hypotheses and to develop the arguments that are embedded within the text of research papers.

Areas of ontology overlap

Prior to our harmonization activity, there were two area of overlap between these ontologies, concerning (a) terms for referring to and citing others' work, and (b) terms for describing the bibliographic objects of such citations (i.e. books, journal articles, etc.) Two of the authors (Clark and Ciccarese) developed terms for referring to and citing others' work within the SWAN Relations Ontology version 1.2, and terms for describing bibliographic entities within the SWAN Citations Ontology version 1.2, while a third author (Shotton), starting with a focus on semantic annotation of scientific documents [22], independently developed a somewhat more detailed ontology for bibliographic citations and entities, CiTO version 1.6, that covered both areas within the single ontology.

It is the harmonization of these independent developments which is described in this paper. By jointly discussing, criticizing, adapting and evolving modules of the SWAN and SPAR ontology suites, we have developed a cluster of fully harmonized ontology modules for describing citations, bibliographic reference and biomedical

discourse.

Re-use of pre-existing vocabularies

We have made use of pre-existing vocabulary specifications wherever possible, including Dublin Core, SKOS, FOAF and PRISM (Publishing Requirements for Industry Standard Metadata) [23], a metadata specification widely used in scientific publishing. These ontologies are specifically intended to fill the technology gap identified above, and we will show that they do so in a robust and fully evolvable way.

Our revised ontologies are specified in OWL 2 DL [6], are fully modular, and inherently supports agent-based searching and mash-ups. We believe they can be further extended and mapped to other related ontologies, to build the nucleus of an extended information ecosystem for scientific communications.

In this paper, Section 2 (Materials) describes the ontologies to be harmonized, Section 3 (Methods) describes our approaches to the harmonization task, Section 4 (Results) details the changes introduced into the new versions of CiTO (version 2.0) and SWAN (version 2.0), and Section 5 (Discussion) presents a discussion of these changes and the lessons learned.

2 Materials – the Ontologies to be Harmonized

SWAN

The SWAN (Semantic Web Applications in Neuromedicine) ontology ecosystem is a set of ontological modules to represent scientific discourse in biomedical research [17, 18, 19]. Thus, when we refer to "the SWAN Ontology", we are actually referring to this suite of ontology modules. SWAN was initially developed to represent scientific discourse in neuromedicine. However, the current architecture allows interested parties to adopt significant components of the SWAN Ontology for representing scientific discourse, quite broadly, in many other domains of science, while assuring an important level of integration with all SWAN ontology-based applications. The SWAN Ontology has been published as a W3C HCLS Working Group Note [19], and it has been the topic of a preceding process of integration with the SIOC (Semantically-Interlinked Online Communities) Ontology for describing blogs, wikis and discussion groups [24]. Within SWAN v1.2, the eight modules of particular relevance to this paper are shown in Table 1.

[Table 1 here]

These SWAN ontology modules are orthogonal: each module covers one single topic and was developed to have the highest cohesion and the lowest coupling possible. Of the SWAN Ontologies, the SWAN Scientific Discourse Relationships Ontology and the SWAN Citations Ontology have been the objects of the harmonization activity described in this paper.

CiTO

CiTO, the Citation Typing Ontology, was first developed as an ontology for describing the nature of reference citations in scientific research articles and other scholarly works, both to other such publications and also to web information resources, and for publishing these descriptions on the semantic web [21]. Using it, citations could be described in terms of both the factual and rhetorical relationships between citing publication and cited publication, in terms of the in-text and global citation frequencies of each cited work, and in terms of the nature of the cited work itself, including its publication and peer review status.

Distinguishing between citation as an act, and as the thing being cited

In the context of the Citation Typing Ontology, a bibliographic citation is a reference within a particular citing work to another publication (e.g. to a journal article, a book chapter or a web page) termed the cited work. As first emphasized in [21], this use of the word 'citation' should be clearly distinguished from the common related use of this word to indicate the cited work itself, for

which the term 'bibliographic record' is to be preferred. Within CiTO, 'cite' and 'citation' denote the performative act of citation, not the target of that citation.

CiTO modularization

In part stimulated by our harmonization activity, two of the authors (DS and SP) recently modularized CiTO v1.6 into a suite of orthogonal and complementary ontologies to describe citations, bibliographic entities, citation counts and publication status, to which they added four further ontologies to create the SPAR (Semantic Publishing and Referencing) Ontologies [20, Peroni and Shotton (in preparation)], a suite of complementary and orthogonal ontologies that can be used individually or in combination, listed in Table 2. Of the SPAR Ontologies, CiTO, the Citation Typing Ontology v2.0 and FaBiO, the FRBR-aligned Bibliographic Ontology v1.0, are the two of relevance to the harmonization activity described in this paper.

[Table 2 here]

The current version of CiTO (v2.0) fulfills the original role of CiTO to characterize bibliographic citations, both factually and rhetorically. To enrich expressivity, several new sub-properties have been added to *cito:cites*, of which *cito:agreesWith* and *cito:disputes* are of particular relevance to rhetoric. Their meanings are subtly different from those of the pre-existing object properties *cito:confirms* and *cito:disagreesWith*. Additionally, for convenience of use, the inverse properties of all the sub-properties of *cito:cites* have been added as sub-properties of its inverse property *cito:isCitedBy*. To permit the relationships described in CiTO to be used widely, we removed the original domain and range restrictions on the object properties *cito:cites* and *cito:isCitedBy* and their sub-properties, following established principles for ontology modularization and development [14, 15].

The core of FaBiO, the FRBR-aligned Bibliographic Ontology, are the classes originally within CiTO v1.6 for describing bibliographic entities, which have been extended by the addition of some new classes, object properties and data properties (see below).

FRBR

Harmonization between our ontologies was readily achievable both because the original ontologies were specified in OWL, and also because they used the same fundamental conceptual model for bibliographic entities, namely FRBR. As a result, FaBiO and the SWAN Citations module were essentially pre-aligned, to an extent that made them highly compatible.

The FRBR (Functional Requirements for Bibliographic Records) classification model is a conceptual entity-relationship model, developed by the International Federation of Library Associations and Institutions (IFLA) as a "generalized view of the bibliographic universe, intended to be independent of any cataloging code or implementation" [25, 26].

Importance of distinctions in FRBR: Works, Expressions, Manifestations and Items

FRBR makes important distinctions between *Works*, *Expressions*, *Manifestations* and *Items, as bibliographic objects*.

A *Work* is a distinct intellectual or artistic creation, an abstract concept recognized through its various expressions; an *Expression* is the specific form that a *Work* takes each time it is 'realized' in physical or electronic form; while a *Manifestation* of an expression of a work defines its particular physical or electronic embodiment, e.g. online, in print, or in PDF format. For example, a research paper (a *Work*) may be realized as a journal article (an *Expression* of that *Work*) and embodied in a print object (a *Manifestation* of that *Expression*). An *Item* is an individual copy of a manifestation that someone can own, for example a print issue of a journal or a PDF file on a computer hard drive.

Recognition and advantages of FRBR

FRBR is widely recognized as a sound fundamental model for bibliographic records,

and was previously used independently by both CiTO and the SWAN Citations module as a basis for ontology design, since its hierarchical structure permits greater expressivity and descriptional accuracy than other 'flat' ontologies and vocabularies for dealing with citations and bibliographic records, such as BIBO, the Bibliographic Ontology [27], PRISM [23], the MeSH tags used in MEDLINE and PubMed [28-30], and various reference management software systems such as Endnote [31] and BibTEX [32].

Building on earlier work that represented the core FRBR concepts in RDF [33], we have recently represented these essential FRBR concepts in OWL 2 DL [34], and have used them in our ontologies.

3 Methods Employed for Ontology Harmonization

Possible harmonization activities

Harmonization activities may involve (a) renaming classes (concepts) or properties (relationships) in one or both ontologies to avoid apparent overlap, (b) more carefully defining classes or properties to resolve actual overlap, and (c) deprecating elements of individual ontologies, or even whole ontologies, in favour of others that more effectively serve the domain of knowledge under consideration, perhaps by having greater granularity or a more effective structure. All three activities were employed to achieve CiTO + SWAN harmonization.

Communication methodologies

The work described in this paper was undertaken collaboratively between the SWAN authors (PC and TC) in Boston and the SPAR authors (DS and SP) in Oxford, without face-to-face meetings. Instead we used a combination of Skype and phone discussions, e-mail exchanges, a collaborative wiki page to record discussions and decisions, and joint participation in Scientific Discourse teleconferences of the W3C Health Care and Life Sciences Interest Group [35], convened and chaired by one of the authors (TC).

Definition of ontology scope and overlap

Our first activity was to carefully analyze the scope of the original ontologies, and discuss their purposes and use cases. From this, it became clear that there was much to be gained simply by using CiTO for its intended limited purpose of specifying and characterizing literature citations, while using the SWAN Ontology for describing hypotheses, relationships to evidence, and scientific discourse more generally, rather than attempting to cover both tasks by creating a single super-ontology. This simple approach is exemplified in Fig. 1.

[Figure 1 here.]

Having made that decision, our harmonization task was simplified to that of inspecting the two ontologies to determine common or related classes, relationships (object properties) and data properties, and then of modifying these as required to clarify their intended purposes and permit their smoother and more coherent integration, or where necessary to rename, redefine or deprecate certain terms in favor of existing or new terms in the other ontology.

In the following description, the past tense is used to describe aspects of the ontologies as they were before our harmonization activity, and the present tense is used to describe the situation that now exists with the publication of the harmonized versions of the ontologies.

4 Results – Harmonization Outcomes

Describing citations

CiTO was developed around the relationship *cito:cites*, encoded as an object property within the ontology to connect citing and cited bibliographic entities. In CiTO v1.6, there were 21 sub-properties of *cito:cites*,

including both factual relationships (e.g. *cito:citesForInformation*, *cito:sharesAuthorsWith*, *cito:usesMethodIn*) and rhetorical relationships (e.g. *cito:supports*, *cito:discusses*, *cito:critiques*). Full details are given in [21].

The SWAN Discourse Relationships Ontology included a relationship *swanrel:cites* (Fig. 2), but here the scope was intended to be more general than in CiTO, for instance, to relate a *swande:ResearchStatement* with a gene or protein.

[Figure 2 here.]

Although the "*cites*" relationships in the two original ontologies had different namespaces and definitions, they shared a common name, which was thought likely to induce confusion in users' minds. We therefore decided to deprecate *swanrel:cites*, and in future to use *cito:cites* and its sub-properties when referring specifically to citations between publications that are the source or target of bibliographic citations, leaving use of the pre-existing more general relationship *swanrel:refersTo* to permit entities such as a *swande:DiscourseElement* to refer to scientific entities such as genes and proteins.

In the context of SWAN, the relationship *cito:cites* is declared to be a sub-property of the SWAN relationship *swanrel:refersTo*. Since *swanrel:refersTo* had previously been defined as a sub-property of *sioc:relatedTo* [24], *cito:cites* thereby becomes a sub-sub-property of *sioc:relatedTo*.

The SWAN relationships hierarchy was then further revised to accommodate these changes. The original subclasses of *swanrel:cites* were renamed and moved to become subclasses of *swanrel:refersTo*. In addition, the subclasses of *swanrel:inResponseTo* were renamed to avoid term collision with CiTO, and other relationship names were modified to harmonize the use of English tenses across the SWAN Relationships hierarchy, as shown in Table 3.

[Table 3 here.]

Fig. 3 shows the resulting revised Relationships hierarchy in SWAN v2.0. Comparison with Fig. 2 will reveal the changes detailed in Table 3.

[Figure 3 here.]

In this manner, we eliminated clashes and redundancy by conforming the SWAN evidence relationships to fit those in CiTO.

Directionality of citation

It is important to note that the directionality of CiTO object properties is always *from* the citing work *to* the cited work. Thus *cito:supports* mean that the citing entity provides intellectual or factual support *for* the cited entity. Conversely, *swanrel:referencesAsSupportiveEvidence* is used to identify a cited item that provides supporting evidence for the argument in the citing document from which the reference is made. Similarly, *cito:discusses* and *cito:refutes* are used, respectively, when the citing entity discusses or refutes the cited entity. These usages are quite different from *swanrel:referencesAsRelevantEvidence* and *swanrel:referencesAsInconsistentEvidence*, which involve bringing relevant or inconsistent evidence *from* the cited work into the argument under consideration.

Describing bibliographic entities

While the primary purposes for which CiTO and SWAN were originally developed were those of describing citations and elements of scientific discourse, respectively, both needed to describe the targets of citations within the FRBR framework, and thus both included classes such as *Book*, *Journal* and *Journal Article* (for an example, see Fig. 4).

[Figure 4 here]

Because of variations in interpretation and application of the FRBR data model, the SWAN Citations Ontology v1.2 lacked the class *swancitations:Work*. However, it had the classes *swancitations:Citation,*

swancitations:Expression and *swancitations:PublicationEnvironment*, the sub-classes of which are very similar to the subclasses of *Work*, *Expression* and *Manifestation* originally in CiTO v1.6 and now part of v1.0. (Note that the class *swancitations:Citation* was used to define a bibliographic record designating the *target* of a citation, *not* the citation itself in the CiTO sense of "A *cites* B".)

It can be seen in Table 4 that, following the inclusion in FaBiO v1.0 of the classes from CiTO v1.6 describing bibliographic entities (book chapters, journal articles, etc. - the *objects* of citations), and the enrichment of FaBiO by the creation of seven new classes, FaBiO v1.0 now provides almost perfect coverage of the classes in the original SWAN Citations Ontology v1.2 for describing bibliographic entities.

[Table 4 here.]

During the development of the SWAN ontology ecosystem, it had always been its creators' intention to leave open the possibility of later 'retiring' one or more SWAN modules, and substituting better or more complete third-party ontologies or ontology fragments as they appeared. The recently created FaBiO Ontology is the very first candidate for such a substitution. Since FaBiO provides more complete coverage of bibliographic records than did the SWAN Citations Ontology, the decision was taken to deprecate the SWAN Citations Ontology in favour of using this alternative ontology, rather than to attempt their integration.

Describing bibliographic records

The definition of the class *swancitations:Citation* was:

> "Information which fully identifies a publication. A complete citation usually includes author, title, name of journal (if the citation is to an article) or publisher (if to a book), and date. Often pages, volumes and other information will be included in a citation."

The SWAN Citations Ontology had a number of data properties that could be applied to sub-classes of the class *swancitations:Citation*, such as *swancitations:JournalArticle*, permitting the details of the bibliographic reference to a particular published work to be specified.

Applications of FRBR distinctions between a Work and its Manifestation in practice

Despite sharing a common DOI, different manifestations of a particular published resource may differ in several details, such as the rendering of figures, and there may be occasions when it is important to distinguish between them, and to refer to a particular manifestation specifically. Additionally, the bibliographic records for journal articles with different manifestations differ. Articles in journals that have print manifestations are identified by the first and last page numbers (e.g. reference [17] in this paper), while those in online-only journals, which are presented in a single unbroken web page, are often identified by the article number rather than page numbers (e.g. reference [21] in this paper).

One of the use cases in building the AlzSWAN Knowledge Base [36] – the application of SWAN for Alzheimer's Disease, undertaken in collaboration with the Alzheimer Research Forum [37] – was to be able to declare that the referenced Journal Article appeared in the printed version of the journal, or alternatively in the version of the journal published electronically.

Thus, in SWAN v1.2, every bibliographic reference to a journal article was made to its manifestation, either in printed form or in electronic format, these being distinguished through the relationship *swancitations:contributionPublicationEnvironment*, an object property that connects the manifestation to a publication environment – in the case of a journal article to a printed journal or to a journal in electronic format, identified respectively by a print ISSN (*swancitations:printISSN*) or an electronic

ISSN (*swancitations:electronicISSN*) – as shown in Fig. 5.

[Figure 5 here.]

In contrast, CiTO v1.6 was intentionally limited to identifying the nature of the citing or cited Work, and of its expression or manifestation, in holistic terms, e.g. *cito:ResearchPaper*, *cito:BookChapter*, *cito:JournalArticle*, *cito:WebPage*, and did not cover the task of specifying the complete bibliographic record of such cited entities.

For instance, CiTO v1.6 contained the following textual definition of the class *cito:PeriodicalIssue*:

> "A particular issue of a periodical, identified and distinguished from other issues of the same publication by date and/or issue number and/or volume number, and comprising separate editorials, articles, news items and/or other writings."

In this definition, the concepts of date, issue number and volume number are present, but CiTO v1.6 intentionally contained no corresponding classes or data properties for defining these elements.

Incorporation of PRISM terms in the FaBiO bibliographic ontology

As a consequence of the decision to deprecate the SWAN Citations Ontology module in favour of FaBiO, the need arose to enable the specification of such elements of the bibliographic record within FaBiO. This was achieved by inclusion of terms from the RDF specification of PRISM, the Publishing Requirements for Industry Standard Metadata [23], to permit full specification of bibliographic records and references. To these PRISM terms, additional useful data properties were added to FaBiO, including *fabio:hasArticleIdentifier*, *fabio:hasCopyrightDate*, *fabio:hasPageCount*, *fabio:hasPublicationYear*, *fabio:hasPubMedID*, *fabio:hasSubtitle* and *fabio:hasURL*. The degree to which these properties accurately cover the properties in the SWAN Citations Ontology previously used for describing bibliographic records is shown in Table 5.

[Table 5 here.]

Using CiTO v2.0 and FaBiO v1.0 together

Used together, CiTO v2.0 and FaBiO v1.0 possess all of CiTO's original capabilities to characterize the nature, sources and targets of bibliographic citations, and now also have the additional ability to fully characterize the bibliographic references themselves, as did the SWAN Citations Ontology that they replace. FaBiO can be employed to describe *Works*, *Expressions* or their *Manifestations*, provided that these entities can be identified by a unique resolvable IRI. FaBiO can thus be employed to describe particular manifestations of a publication, as required by the AlzForum use case. It does so by directly describing each manifestation (e.g. *fabio:WebPage*; *fabio:Paperback*), rather than through use of the object property *swancitations:contributionPublicationEnvironment* as was the case when using the SWAN Citations Ontology.

Web availability and characteristics of the harmonized ontologies

The most recent version of the SWAN ontology ecosystem, SWAN v2.0, published at http://purl.org/swan/2.0/swan.owl, includes the revised modules that have resulted from the harmonization activity and decisions described in this paper:

- **Scientific Discourse Relationships Module:** This has been revised to better integrate with CiTO and to provide a more consistent set of relationship names. As explained above, the main changes involved deprecating the SWAN relationship swanrel:cites in favour of cito:cites, and modifying the names and sub-classing of other SWAN Relationships object properties as summarized in Table 3.
- **SWAN Citations Module:** This has been deprecated in favor of using FaBiO.

- **SWAN Commons:** The purpose of the SWAN Commons Ontology is to import and integrate all the ontological building blocks considered helpful for managing the scientific discourse of online scientific communities. This has been updated to import FaBiO v1.0 in place of the deprecated SWAN Citations module. As consequence, a certain number of integration constraints - such has OWL disjoints and property restrictions - defined in this module have been updated accordingly.

CiTO version 2.0 was published on 4 November 2010 at http://purl.org/spar/cito/, to which the original URL http://purl.org/net/cito/ now redirects, while FaBiO version 1.0 was published on 10 November 2010 at http://purl.org/spar/fabio/. These sites use content negotiation to deliver to the user a human-readable 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 4 [38, 39]. (For full compatibility with OWL 2 in which these ontologies are encoded, please use Build 200 or later of Protégé version 4.1 beta, or subsequent versions.) The principle revisions to these ontologies brought about by the harmonization activity are:

- **CiTO**: Addition of a small number of new object properties relating citing entity to cited entity. Addition of inverse classes of all subclasses of *cito:cites*, as subclasses of *cito:isCitedBy*. Removal of domain and range restrictions on *cito:cites* and *cito:isCitedBy*.
- **FaBiO**: Creation of new classes to cover classes in the deprecated SWAN Citations Ontology required for describing bibliographic entities. Inclusion of PRISM data properties and creation of additional FaBiO data properties to permit the full description of the elements of a bibliographic reference to a published entity, a role previously fulfilled by the deprecated SWAN Citations Ontology. Where appropriate, these data properties have been made functional, to ensure that the entities they describe can be assigned only one publication date and only one identifier of a particular type (e.g. DOI).

As part of these revisions, the textual definitions (annotation comments) of all the CiTO and FaBiO classes and properties were individually checked and where necessary amended, primarily to bring these descriptions into line with the logical changes that had been introduced into the ontologies. During this process, to enhance readability, class and property labels, and occurrences of class and property names within the textual definitions, were uniformly changed to appear as separate lower case words (e.g. "work" and "patent application"), rather than being capitalized (e.g. "Work") or presented in CamelBack notation (e.g. "PatentApplication"). The exceptions to this are where, for clarity of meaning within the textual definitions, they are preceded by their namespace abbreviations (e.g. "*fabio:Work*", to distinguish it from *frbr:Work*), in which case the standard CamelBack notation of the class or property name is used where necessary (e.g. "See also *fabio:GrantApplication*").

5 Discussion

Why our harmonization activities succeeded

The ontology harmonization effort described in this paper succeeded because of the following factors:

- The fact that the original ontologies were devised for distinct, although related purposes.
- The decision to limit the usage of CiTO to bibliographic citations, clearly distinguishing its purpose from the broader purposes of the SWAN Scientific Discourse Relationships Ontology to describe scientific discourse relationships.
- The decision to ensure that there were no classes or properties with identical names

between the two ontologies, renaming and refining definitions where appropriate to avoid name collisions.

- The willingness of the authors of each ontology suite to suggest, and at times insist, that the authors of the other suite make particular changes, either for reasons of ontological correctness or to meet specific use case requirements. Such recommendations have been acknowledged by the inclusion of the names of those individuals as contributors to the others' ontologies.
- The willingness of the participating parties to seek the best outcome, rather than to 'defend' their prior work. This was particularly evident when it came to the decision to deprecate the SWAN Citations Ontology module in favour of using FaBiO.
- The adoption of a modular strategy in developing the SWAN Ontologies, and the extension of this principle to the SPAR ontology suite. This has been demonstrated to be a winning approach, since it allowed integration through the very limited set of changes, apart from the deprecation of the SWAN Citations module. This is a very important point, since modularization limits the number of cross-constraints that have to be applied or modified when the various SWAN ontology components are re-integrated after making a change to one of them.

The focus of attention on the structure of the ontologies at the start of this harmonization activity had a further benefit of providing additional incentive for the authors of the SPAR ontology suite (DS and SP) to undertake the modularization of CiTO v 1.6 discussed above.

The architecture of the harmonized ontology system resulting from our work is shown in Fig. 6.

[Figure 6 here]

Some comments on social process

The social process we used began with a pre-existing mutual understanding that ontology development in scientific domains is an inherently collaborative process. This arises from the nature of scientific work itself. The reason we develop ontologies is to make better use of, and to better understand, one another's research results.

Our social process can best be described as consensus-driven "give-and-take". We defined no rules of engagement at the outset, but we did define a goal to which we all subscribed, and an understanding that none of us had a monopoly on good sense. All participants realized that this goal would best be furthered if we could achieve interoperability and delegation of concerns. This helped us to be patient and flexible with one another when conflicts arose.

Ultimately the authors found that one of the keys to successful collaboration in this field, as in many others, was a dose of humility from time to time. It was essential to be willing to learn from each other, and to abandon previous approaches when better ones arose from another source. This was possible because we understood that the whole would be greater than the sum of its parts, and enabled the consensus driven approach to succeed.

Examples of usage of these harmonized ontologies

To demonstrate the manner in which our revised and harmonized ontologies can be used to encode bibliographic references, we provide two examples of bibliographic information encoded as RDF in Turtle notation [40], both before and after the harmonization activity described in this paper. These appear in **Supplementary Information File S1**.

In this supplementary information file, Text Box 1 shows the bibliographic record for the SWAN paper by Ciccarese *et al*., 2008 [17] and for the journal in which it was published.

In Text Box 1A, this is encoded in Turtle using the SWAN Ontology v1.2, while Text Box 1B it is re-coded using SWAN v2.0, FaBiO v1.0 and CiTO v2.0.

Similarly, Text Box 2 shows an excerpt from the document [41] published to provide machine-readable metadata about the paper by Shotton describing CiTO v1.6 [21], both as originally encoded using CiTO v1.6 and after re-coding using FaBiO v1.0 and CiTO v2.0.

Another example of how these new information models can be used is the SWAN Annotation Framework (AF), now in alpha release at a collaborating major pharmaceutical company and soon to be released more widely as part of the Neuroscience Information Framework [42]. SWAN AF provides a means of running and supervising text mining applications over full text scientific articles, as well as doing manual annotation. The annotation is represented as fully provenanced stand-off metadata in OWL/RDF using the Annotation Ontology (AO) [43]. Among the key metadata linked to any publication annotated with AF/AO is its bibliographic record, expressed in FaBiO, and its citations, expressed in CiTO.

A third example is that of Utopia, a PDF reading and annotation application environment that provides semantic enrichment to the articles being read [44, 45]. Utopia has decided to use SWAN, FaBiO and CiTO, in addition to DoCO, the Document Components Ontology [46] and AO, the Annotation Ontology, to describe PDF documents and citations on the Utopia server. Utopia is employed by Portland Press to prove semantic enrichment to the Semantic *Biochemical Journal* [47].

CiTO is also being used by the bibliographic reference service CiteULike, an activity of the Springer publishing group – for example see [48].

A final example is the use of these ontologies to encode bibliographic information in the SAO/NASA Astrophysics Data System hosted by the High Energy Astrophysics Division at the Harvard-Smithsonian Center for Astrophysics [49].

6 Conclusion

This ontology harmonization activity has improved the coverage, logical consistency and definitions of the ontologies under consideration, and their integration into an interoperable whole that is more powerful than the original ontologies alone. Our collection of ontologies extends the evolving ecosystem of ontology modules for scientific discourse on the web in a fundamental way. With CiTO, FaBiO and the SWAN ontologies, we can now offer an interoperable and complete ontology system in OWL 2 for describing bibliographic entities, bibliographic citations, bibliographic references, and the elements of scientific discourse more widely defined, as a coherent whole.

Extending from a core of the newly-aligned CiTO, FaBiO and SWAN ontologies, are several other harmonized ontologies of value in scientific discourse. These include the SIOC (Semantically-Interlinked Online Communities) Ontology for describing blogs, wikis and discussion groups, which had previously been aligned with the SWAN Ontologies; AO, the Annotation Ontology for annotation of documents; and the other SPAR Ontologies for describing other aspects of the publication domain, including reference collections and document components.

These ontologies represent the most important metadata for scientific discourse, because they provide key elements to underpin the scientific method as it embraces a web-based *modus operandi*. These ontologies allow us to create semantic metadata for web-based scientific publications, and can enable development of much more powerful facilities for organization, search and mash-up of web-based scientific discourse.

We commend these revised and integrated ontologies – CiTO, FaBiO and the SWAN ontology modules – to the publishing and

research communities for more widespread adoption and use, and welcome feedback on ways in which they may be further improved.

Acknowledgements

SWAN (PC and TC, Harvard): The development of SWAN was funded by generous grants from a philanthropic foundation that wishes to remain anonymous. We are grateful to Eric Prud'hommeaux of W3C for valuable technical support, and to Anita de Waard of Elsevier for many helpful comments and references and for her enthusiastic support.

CiTO and FaBiO (DS and SP, Oxford): The initial development of CiTO was undertaken as part of the work of the Ontogenesis Network, supported by EPSRC grant EP/E021352/1. The harmonization activity reported here, which paralleled the restructuring of CiTO and the creation of FaBiO, were initially undertaken without specific grant funding, and were then supported by the JISC Open Citations Project.

References

1. Databases and Tools, National Center for Biotechnology Information [http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html]
2. Miles A, Zhao J, Klyne G, White-Cooper H, Shotton D: OpenFlyData: An exemplar data web integrating gene expression data on the fruit fly Drosophila melanogaster. J Biomed Inform 43 (5). http://dx.doi.org/10.1016/j.jbi.2010.04.004.
3. Becket D (editor): RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. World Wide Web Consortium; 2004. http://www.w3.org/TR/REC-rdf-syntax/.
4. Brickley D, Guha RV (editors): RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. World Wide Web Consortium; 2004. http://www.w3.org/TR/rdf-schema/.
5. McGuinness D, van Harmelen F (editors): OWL Web Ontology Language. W3C Recommendation 10 February 2004. World Wide Web Consortium; 2004. http://www.w3.org/TR/owl-features/.
6. Bao J, Kendall, EF, McGuinness, DL, Patel-Schneider, PF (editors): OWL 2 Web Ontology Language Quick Reference Guide. W3C Recommendation 27 October 2009. World Wide Web Consortium; 2009. http://www.w3.org/TR/2009/REC-owl2-quick-reference-20091027/.
7. Fielding RT: Architectural styles and the design of network-based software architectures. Doctoral dissertation. University of California, Irvine, Information and Computer Science; 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm.
8. Prud'Hommeaux E, Seaborne A (editors): SPARQL Query Language for RDF: W3C Recommendation 15 January 2008. World Wide Web Consortium; 2008. http://www.w3.org/TR/rdf-sparql-query/.
9. Alexander, K: Linked Data APIs. Nodalities Magazine, Issue 10, p 21. http://www.talis.com/nodalities/pdf/nodalities_issue10.pdf.
10. Lassila O, Hendler J: Embracing "Web 3.0". IEEE Internet Computing 2007, 11:90-93. http://doi.ieeecomputersociety.org/10.1109/MIC.2007.52.
11. International Resource Identifiers. http://tools.ietf.org/html/rfc3987.
12. Web of Linked Data. http://linkeddata.org/.
13. Gene Ontology. http://www.geneontology.org/.
14. Rector A: Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. In K-CPA'03 Conference. pp. 121-128. Sanibel Island, Florida, USA; 2003:121-128.

http://portal.acm.org/citation.cfm?id=945664.

15. Shotton D, Caton C, Klyne G: Ontologies for sharing, ontologies for use. In The Ontogenesis Knowledge Blog. 2010:Paper 3. http://ontogenesis.knowledgeblog.org/2010/01/22/ontologies-for-sharing/.
16. OBO Foundry Principles. http://www.obofoundry.org/crit.shtml.
17. Ciccarese P, Wu E, Wong G, Ocana M, Kinoshita J, Ruttenberg A, Clark T: The SWAN biomedical discourse ontology. J Biomed Inform 2008, 41:739-751. http://dx.doi.org/10.1016/j.jbi.2008.04.010.
18. The SWAN Ontology Ecosystem. http://swan.mindinformatics.org/ontology.html.
19. Ciccarese P (editor): Semantic Web Applications in Neuromedicine (SWAN) Ontology. W3C Interest Group Note 20 October 2009. http://www.w3.org/2001/sw/hcls/notes/swan/.
20. SPAR, the Semantic Publishing and Referencing Ontologies. http://opencitations.wordpress.com/2010/10/14/introducing-the-semantic-publishing-and-referencing-spar-ontologies/.
21. Shotton D: CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics 2010, 1 (Suppl. 1): S6. http://dx.doi.org/10.1186/2041-1480-1-S1-S6.
22. Shotton D, Portwin K, Klyne G, Miles A: Adventures in semantic publishing: exemplar semantic enhancements of a research article. PLoS Comput Biol 2009, 5:e1000361. http://dx.doi.org/10.1371/journal.pcbi.1000361.
23. PRISM (Publishing Requirements for Industry Standard Metadata) Specification: Version 2.1. http://www.prismstandard.org/specifications/2.1/PRISM_prism_namespace_2.1.pdf.
24. Passant A, Ciccarese P (editors): SWAN/SIOC: Alignment Between the SWAN and SIOC Ontologies. W3C Interest Group Note 20 October 2009. http://www.w3.org/TR/hcls-swansioc/.
25. Saur KG: FRBR (Functional Requirements for Bibliographic Records) Final Report. International Federation of Library Associations and Institutions; 1998. http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.
26. Tillett B: What is FRBR? A Conceptual Model for the Bibliographic Universe. Washington DC, USA: Library of Congress, Cataloguing Distribution Service; 2003. http://www.loc.gov/cds/downloads/FRBR.PDF.
27. BIBO, the Bibliographic Ontology. http://www.bibliontology.com.
28. MESH, Medical Subject Headings. http://www.nlm.nih.gov/mesh/.
29. Lipscomb CE: Medical Subject Headings (MeSH). Bull Med Libr Assoc 2000, 88:265-266. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/.
30. Sewell W: Medical Subject Headings in Medlars. Bull Med Libr Assoc 1964, 52:164-170. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC198088/.
31. EndNote bibliographic reference manager. http://www.endnote.com/.
32. BibTEX, a tool and a file format used to describe and process references. http://www.bibtex.org/.
33. Davis I, Newman R: Expression of Core FRBR Concepts in RDF. http://vocab.org/frbr/core.html.

34. Ciccarese P , Peroni S (2010) Essential FRBR in OWL 2 DL. http://purl.org/spar/frbr.

35. W3C Health Care and Life Sciences Interest Group. http://www.w3.org/2001/sw/hcls/.

36. The AlzSWAN Knowledge Base. http://www.alzforum.org/res/adh/swan/default.asp.

37. The Alzheimer Research Forum, AlzForum. http://www.alzforum.org.

38. Protégé, an open source ontology editor. http://protege.stanford.edu/.

39. Knublauch H, Fergerson RW, Noy NF, Musen MA: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Lecture Notes in Computer Science 2004, 3298:229-243. http://dx.doi.org/10.1007/978-3-540-30475-3_17.

40. Beckett D, Berners-Lee T: Turtle - Terse RDF Triple Language: W3C Team Submission 14 January 2008. http://www.w3.org/TeamSubmission/turtle/.

41. Shotton D (2010) Supplementary file S1 to Shotton D (2010). CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics 1(Suppl 1):S61 (Reference [21] in this paper), contains metadata descriptions of the article recorded in a structured machine-readable form, encoded as RDF and serialized in Notation3 format. http://dx.doi.org/10.1186/2041-1480-1-S1-S6/suppl/S1.

42. The Neuroscience Information Framework. http://www.neuinfo.org/.

43. Ciccarese P, Ocana M, Das S, Clark T: AO: An Open Annotation Ontology for Science on the Web. Bio-Ontologies 2010, July 9-10 2010, Boston MA. http://www.purl.org/ao/d/bo2010.

44. Attwood TK, Kell DB, McDermott P, Marsh J, Pettifer SR, Thorne D: Calling International Rescue: knowledge lost in literature and data landslide! Biochemical Journal 2009, 424:317–333. http://dx.doi.org/10.1042/BJ20091474 317.

45. Attwood TK, Kell DB, Mcdermott P, Marsh J, Pettifer SR, Thorne D: Utopia Documents: linking scholarly literature with research data. Bioinformatics 2010, 26:i540-i546. http://dx.doi.org/10.1093/bioinformatics/btq383.

46. DoCO, the Document Components Ontology. http://purl.org/spar/doco/.

47. The Semantic Biochemical Journal. http://www.biochemj.org/bj/semantic_faq.htm.

48. Example of CiTO being used in CiteULike. http://www.citeulike.org/user/egonw/article/1073448.

49. The SAO/NASA Astrophysics Data System. http://adswww.harvard.edu/.

FIGURE LEGENDS

Figure 1. The integrated use of CiTO to characterize citations (lower cloud) and of SWAN to describe scientific discourse (upper cloud), expressed as an RDF graph.

Figure 2. The original SWAN Relationships Ontology v1.2 relationships hierarchy, including the relationship *swanrel:cite*s. (Note: the sub-properties shown for *swanrel:cites* are found in the SWAN Commons Ontology module v1.2.)

Figure 3. The revised SWAN v2.0 Relationships hierarchy, that now includes *cito:cites* from CiTO v2.0 as a sub-property of *swanrel:refersTo*, in place of *swanrel:cites*. Other changes are as detailed in Table 3.

Figure 4. An example of a journal article representation in SWAN v1.2. This is the bibliographic record for the article (Reference [17]) "The SWAN biomedical discourse ontology" written by Paolo Ciccarese, Elizabeth Wu, June Kinoshita, Gwendolyn Wong, Marco Ocana, Alan Ruttenberg and Tim Clark, and published in Volume 41 of the *Journal for Biomedical Informatics* (PubMed id 18583197). Note that some of the authors are intentionally omitted from the diagram for clarity.

Figure 5. A diagram showing a journal article described using the SWAN Citations Ontology, manifested as part of an on-line journal with the electronic ISSN 15320464. The same journal also appears in printed form with the print ISSN 15320480. In this picture, the URIs of these different manifestations of the journal have been defined through their ISSNs.

Figure 6. A revision of the original SWAN architectural diagram showing the integration of CiTO and FaBiO, and the use of the OWL 2 DL version of the FRBR Core..