# Semantic annotation of publication entities

David Shotton, Department of Zoology, University of Oxford, UK

Silvio Peroni, Department of Computer Science, University of Bologna, Italy

A   The SPAR Ontologies   (*Topic 2.4 Annotation standards and ontologies*)

We bring to the table SPAR, the Semantic Publishing and Referencing Ontologies, a freshly-minted suite of ontologies written in OWL 2.0 that enable the creation of rich metadata to surround scientific publications. These eight ontologies are named in the following figure:
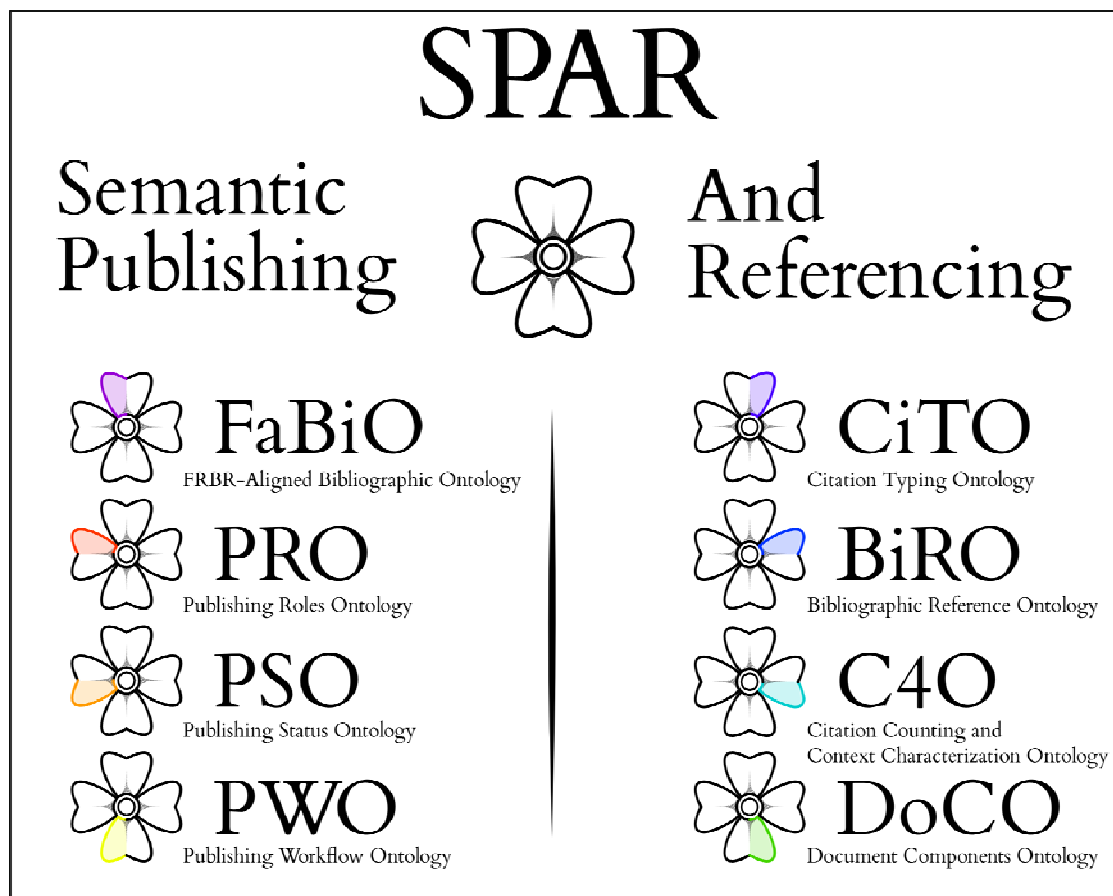
Two of these ontologies, CiTO, the Citation Typing Ontology [1], and FaBiO, the FRBR-aligned Bibliographic Ontology, have recently been harmonized with the SWAN ontologies developed by Tim Clark and Paolo Ciccarese, in a joint collaboration described in [2].

The SPAR Ontologies are all available through their PURLS: http://purl.org/spar/fabio/, http://purl.org/spar/cito, etc., with content negotiation being used to provide a human-readable version if accessed through a browser, or the OWL file itself if accessed through an ontology editor such as Protégé. (Note that for full compatibility with OWL 2.0 in which these ontologies are encoded, Build 200 of Protégé version 4.1 beta, or subsequent versions, should be used.)

B   Research Objects   (*Topic 2.1 New formats for scientific papers*)

Sean Bechhofer *et al.* [3] have recently proposed the notion of Research Objects that will encapsulate data, models and textual descriptions constituting a complete scientific investigation,

providing digital knowledge containers that enable exchange and re-use. In a new collaboration with Sean Bechhofer, Carole Goble and Katy Wolstencroft in Manchester, and with Graham Klyne and Jun Zhao in Oxford, we will instantiate and exemplify Research Objects to encapsulate datasets arising from the ADMIRAL Project, a small Oxford-based project that is developing a digital management infrastructure for research activities in the life sciences, and the SysMO Project, a large European trans-national funding and research initiative concerned with the systems biology of microorganisms. Both have the advantage of being able to trial the Research Objects publication model with real data from active research groups, and will employ the SPAR ontologies, together with the Open Provenance Model Vocabulary (OPMV), the Annotation Ontology AO and the SWAN ontologies for semantic annotation.

C   Structural and rhetorical annotation of document components   *(Topic 2.2 Models of annotation: what parts of the document are marked up)*

DoCO, the Document Component Ontology (http://purl.org/spar/doco/), explained diagrammatically at in the following diagram
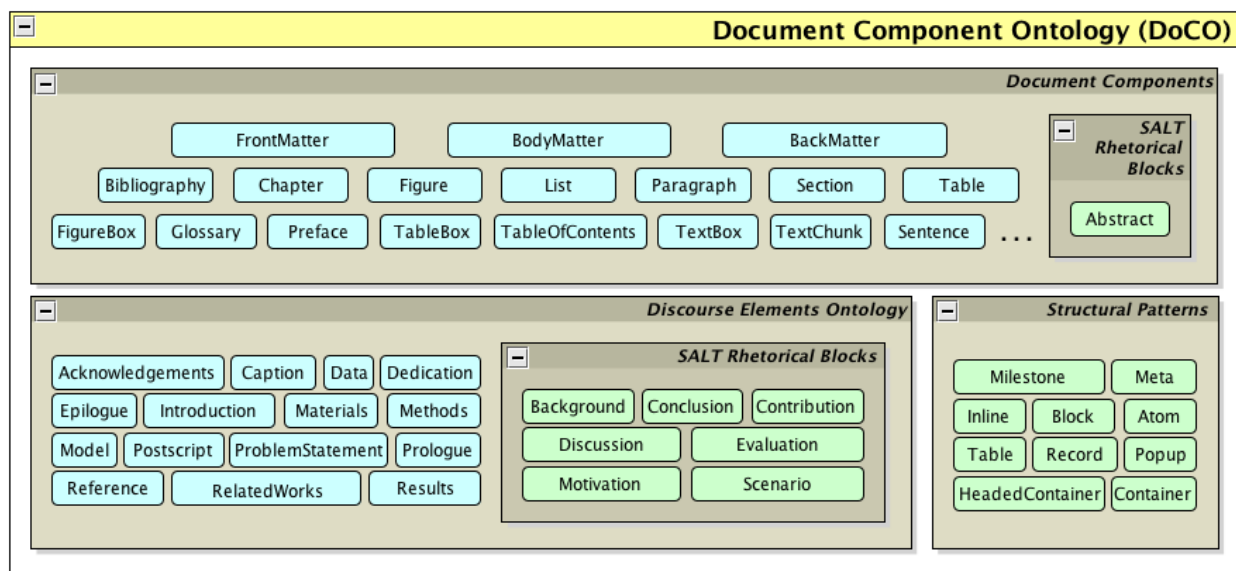
provides the semantic tools for detailed structural and rhetorical annotation of document components, as exemplified in the following diagram and others to be found at https://sempublishing.svn.sourceforge.net/svnroot/sempublishing/DoCO/Examples.pdf.   DoCO provides the possibility for real semantic markup, and can unify the present plethora of distinct syntactic markup schemas provided by the various DTDs employed by different publishers.

The DoCO, FaBiO and CiTO ontologies have now been adopted by the developers of **Utopia**, a PDF reading and annotation environment that provides semantic enrichment to the articles being read (see Workshop Paper #1), and that already employs SWAN and AO, the Annotation Ontology. DoCO, FaBiO and CiTO will be used for describing document components, PDF documents and citations on the Utopia server. Utopia is employed by Portland Press to prove semantic enrichment to the Semantic *Biochemical Journal* (http://www.biochemj.org/bj/semantic_faq.htm).

doco:BodyMatter

doco:Section ∧ pattern:HeadedContainer ∧ doco:Introduction

Introduction doco:SectionTitle

doco:Paragraph

The human pancreas secretes liters of enzymes daily to aid in food digestion, and bovine mammary glands produce eight liters of milk each day, largely for human consumption. To do this, secretory organs must adapt to the increased need for protein secretion that occurs during development, differentiation, or changing physiological conditions. An important question is how changes in secretory capacity are coordinated to allow for efficient targeting, folding, modification, and delivery of secreted products. A few transcription factors have been discovered to up-regulate genes in the secretory pathway, including Xbp1, which is expressed and required in B cells as they differentiate into antibody secreting plasma cells (Shaffer et al., 2002), and which also regulates secretory function in a subset of specialized secretory organs (Shaffer et al., 2004; Lee et al., 2005). The bZip transcription factor ATF6 activates expression of chaperone proteins required for efficient protein folding (Adachi et al., 2008) as well as many of the lipid components of secretory organelles (Bommiasamy et al., 2009). Two other bZip transcription factors, Creb3L1/OASIS and Creb3L2/ BBF2H7 (herein referred to as Creb3L1 and Creb3L2), are required for efficient bone deposition and cartilage matrix secretion, respectively (Murakami et al., 2009; Saito et al., 2009). A major question is whether these transcription factors

function more broadly to up-regulate the entire secretory pathway in multiple specialized cell types or if their function is restricted to the up-regulation of only a subset of secretory genes in a few specialized cells.

doco:Paragraph

The *Drosophila* salivary gland (SG) provides an excellent model for identifying and studying the factors required for secretory function. The SG is the largest secretory organ in *Drosophila*, and the processes of morphogenesis and differentiation have been well characterized (Kerman et al., 2006). The SG comprises two large secretory tubes, each containing ∼100 polarized epithelial cells that are specialized for the production and delivery of secreted proteins. Consistent with the high-level secretory activity of the SG, at least 34 secretory pathway component genes (SPCGs) are highly expressed in the secretory cells (Abrams and Andrew, 2005), and this expression requires at least two transcription factor genes, *fork head (fkh)* and *CrebA* (Andrew et al., 1997; Myat et al., 2000).

SG expression of *fkh* and *CrebA* is activated in the most posterior head segment (parasegment two) by the homeotic gene *Sex combs reduced (Scr)* and two more generally expressed homeotic cofactor genes *extradenticle (exd)* and *homothorax (hth;* Henderson and Andrew, 2000). Dpp signaling in dorsal cells blocks expression of *fkh* and *CrebA*, limiting their activation to only the ventral cells of parasegment two (Henderson et al., 1999).

D   Publication workflows   (Topic 3: Use cases and examples)

Phil Bourne provided files in the *Workflow Materials* folder of the Beyond the PDF web site that form a linear workflow from research activity to publication and beyond. The Publication Workflow Ontology (http://purl.org/spar/pwo/) is an *extremely* simple ontology for describing workflows that we have applied to create an exemplar description of part of Phil's workflow, shown in the following diagram.
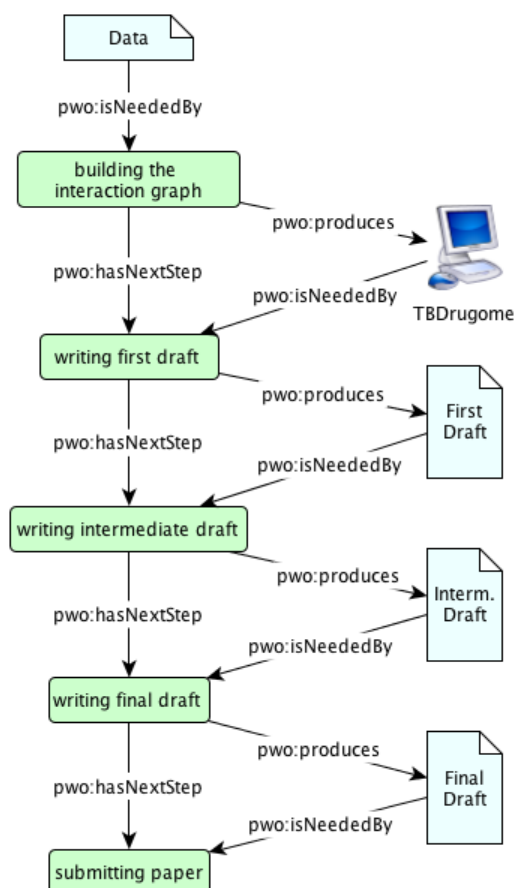
Despite the beguiling simplicity of this diagram, the point we wish to make is that using PWO enables such a workflow to be described in RDF.

References

[1]   Shotton D (2010) CiTO, the Citation Typing Ontology. *J. Biomedical Semantics* 1 (Suppl. 1): S6. http://dx.doi.org/10.1186/2041-1480-1-S1-S6.

[2]   Ciccarese P, Shotton D, Peroni S and Clark T (2010)   CiTO + SWAN: The Web Semantics of Bibliographic Records, Citations, Evidence and Discourse Relationships. (Submitted for publication.)

[3]   Sean Bechhofer, David De Roure, Matthew Gamble, Carole Goble and Iain Buchan (2010) Research Objects: Towards Exchange and Reuse of Digital   Knowledge.   *Nature   Preceedings*. doi:10.1038/npre.2010.4626.1.