A review of ontologies for describing scholarly and scientific documents

Almudena Ruiz-Iniesta and Oscar Corcho

Ontology Engineering Group, Facultad de Informática
Universidad Politécnica de Madrid, Madrid, Spain
{almudenari, ocorcho}@fi.upm.es

Abstract. Several ontologies have been created in the last years for the semantic annotation of scholarly publications and scientific documents. This rich variety of ontologies makes it difficult for those willing to annotate their documents to know which ones they should select for such activity. This paper presents a classification and description of these state-of-the-art ontologies, together with the rationale behind the different approaches. Finally, we provide an example of how some of these ontologies can be used for the annotation of a scientific document.

Keywords: ontology, document semantics, semantic publishing

1 Introduction

Semantic publishing [1,2] can be defined as the activity of enhancing a document (e.g. a journal article) with semantic annotations, providing a way to understand the meaning of the published information and enabling the linking to related documents. Semantic publications offer a better access to both their content and metadata describing the entire documents, their structure, their rhetorical elements and related information. Several ontologies have been created to support this activity in different scholarly domains, e.g., EXPO [3] in the scientific experiments domain, OMDoc as a markup format and data model for Open Mathematical Documents [4], the SWAN ontology for modelling the scientific discourse, developed in the context of building a series of applications for biomedical research [5]. However, the variety of works that describe documents in different domains makes it difficult to choose the best ontology for the annotation of scientific papers, besides the obvious use of Dublin Core Terms[1]. Moreover, there are no general conventions or rules on how to use the existing ontologies in semantic publishing. In order to shed light on this variety of works, in this paper we review the most relevant ontologies for describing scholarly publications and we also present other vocabularies that allow embedding formal metadata in documents using markup languages.

[1] Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/2008/01/14/dces/

Therefore, the result of this work is a classification of the most important ontologies for describing scholarly documents. The proposed classification divides ontologies into three main groups: ontologies for describing the document structure (sections, paragraphs, etc.), ontologies for describing the rhetorical elements (introduction, results, etc.) and ontologies for describing bibliographies and citations. In what follows, we expand on these ontologies, and show how they can be employed to describe a scientific paper. We also illustrate how some of them could be applied to a published article from the journal *Future Generation Computer Systems*, which is already encapsulated as a Research Object [6][2].

The rest of the paper is organized as follows: Section 2 explains the main ontologies that describe documents; Section 3 focuses on describing ontologies for the scientific discourse; Section 4 presents the works that attempt to annotate the references of a document; Section 5 introduces other vocabularies that allow annotating documents. Finally, Section 6 concludes the paper and depicts some recommendations to annotate a scientific document.

2 Ontologies for describing documents

In this section we describe ontologies that allow describing the structure of a scholarly article or, more generally, of a document. Each ontology is presented with its main characteristics (classes and properties) and an example of use.

One of the earliest works in this direction was the no-longer maintained Document ontology[3], implemented in the SHOE language. This ontology focuses only on the document type. Some of the documents types defined in this ontology are: Abstract, Letter, Form, Lecture, etc.

The Ontology of Rhetorical Blocks (ORB)[4] captures a coarse-grained rhetorical structure of scientific publications, independently of their domain. The ontology models a publication by means of three artefacts: the header, the body and the tail. The header is the part of the publication that models meta-information about the publication, including title, authors, affiliations, publishing venue and abstract. The body is composed by four rhetorical blocks: introduction, methods, results and discussion, according to the IMRAD [7] structure. Finally the tail provides additional meta-information about the paper, related to external references. The tail is represented by two ontological entities: acknowledgments and references.

A recent work that attempts to annotate the entire characteristics of a document is the Semantic Publishing and Referencing Ontologies[5] a set of ontologies that allow describing books and journal articles, citations, bibliographic records, the component parts of documents, and various aspects of the scholarly publication process. This set of ontologies is composed by:

[2] `http://rohub.linkeddata.es/motifs_bundle_page-FGCS/`
[3] `http://www.cs.umd.edu/projects/plus/SHOE/onts/docmnt1.0.html`
[4] `http://www.w3.org/2001/sw/hcls/notes/orb/#ontology`
[5] SPAR, namespace `http://purl.org/spar`

– FaBiO[6], the FRBR-aligned Bibliographic Ontology, which allows recording and publishing bibliographic records of scholarly documents.
– CiTO[7][8], the Citation Typing Ontology, which allows characterising citations, both factually and rhetorically.
– BiRO[8], the Bibliographic Reference Ontology, which allows describing bibliographic records and references, and their compilation into bibliographic collections and reference lists.
– C4O[9], the Citation Counting and Context Characterization Ontology, which allows the characterization of bibliographic citations in terms of their number and their context.
– DoCO[10], Document Components Ontology, which allows describing the component parts of a document. DoCO imports the Discourse Elements Ontology[11] and the Document Structural Patterns Ontology[12].
– PRO[13], the Publishing Roles Ontology, which allows characterising the roles of agents in the publication process.
– PSO[14], the Publishing Status Ontology, which allows characterising the publication status of a document at each of the various stages in the publishing process.
– PWO[15], the Publishing Workflow Ontology, which allows describing the steps in the workflow associated with the publication of a document.

In this work we analyse those focused on describing the document content. Hence we will describe in detail DoCO (see Section 2.1) and CiTO (see Section 4).

2.1 DoCO, Documents Components Ontology

The DoCO ontology provides a broad number of classes and relationships that allow describing a document based on its structure and content. DoCO imports two ontologies: Deo and the Document Structural Patterns Ontology. Deo is an OWL2 ontology that describes the major rhetorical elements of a document. It also provides a structured vocabulary for rhetorical elements within documents and it uses all the rhetorical block elements from the SALT Rhetorical Ontology [9]. The pattern ontology defines formally patterns for segmenting a document into atomic components, in order to be manipulated independently and re-flowed in different contexts.

DoCO describes the vast majority of document components such as chapter, preface, glossary, etc. Table 1 shows some of the classes from this ontology.

[6] Namespace `http://purl.org/spar/fabio/`
[7] Namespace `http://purl.org/spar/cito`
[8] Namespace `http://purl.org/spar/biro`
[9] Namespace `http://purl.org/spar/c4o`
[10] Namespace `http://purl.org/spar/doco`
[11] Namespace, `http://purl.org/spar/deo`
[12] Namespace, `http://www.essepuntato.it/2008/12/pattern`
[13] Namespace `http://purl.org/spar/pro`
[14] Namespace `http://purl.org/spar/pso`
[15] Namespace `http://purl.org/spar/pwo`

Table 1. List of some of the classes in DoCO

table	section	list
chapter	figure	glossary
front matter	body matter	preface

As far as Deo is concerned, Deo supports the main rhetorical elements in a document, e.g., Introduction, Methods, Results and Conclusions. These elements give a defined rhetorical structure to the paper, which assists readers to identify the important aspects of the paper. Notice that the rhetorical organization of a paper does not necessarily correspond neatly to its structural components (sections, paragraphs, etc.). In this sense, Deo and DoCO complement one another. Table 2 shows some of the most relevant classes from Deo.

Table 2. List of some of classes in Deo

acknowledgements	background	conclusion
introduction	future work	methods
related work	results	discussion
motivation	problem statement	biography

3 Scholarly and scientific discourse ontologies

The scientific discourse has particular characteristics that are not covered by the aforementioned ontologies. Particularly, a scientific discourse has goals, claims, experiments, evaluations and so on. Indeed, the reasoning of the assertion of the scientific document is crucial for scholarly and scientific publishing, in proposing hypotheses and advancing evidence in their support. Several works have been proposed to model the discourse argumentation normally present in scientific articles.

One of the first works to address the modelling of the scholarly discourse was ScholOnto[10]. The ScholOnto ontology provided a small set of conceptual and relational types. The main class of the ontology is the *Claim*. All claims are owned by an agent, and have some form of justification. Claims assert new relationships with other claims, or between concepts.

The work proposed by [11,12] identifies the main components of scientific investigations and construct the Core Information about Scientific Papers (CISP) metadata about the content of papers. The main classes proposed in CISP are: Goal of investigation, Motivation, Object of investigation, Research method, Experiment, Observation, Result and Conclusion. CISP metadata makes use of the ontology of experiments EXPO[16][3] as a core ontology. CISP includes eight key

[16] `http://expo.sourceforge.net/`

classes that are presented in Table 3. Many of these key classes have additional subclasses and properties.

Table 3. List of the CISP key classes.

goal of investigation	motivation	object of investigation	research method
experiment	observation	result	conclusion

As aformentioned, CISP makes use of EXPO, a very complete ontology about scientific experiments. The aim of this ontology is to provide a controlled vocabulary of scientific experiments. For this purpose EXPO defines over 200 concepts to allow providing a formal description of experiments for efficient analysis, annotation and sharing of results. EXPO is able to describe computational and physical experiments, experiments with explicit and implicit hypothesis. EXPO defines general classes including ScientificExperiment, ExperimentGoal, ExperimentTechnology, ExperimentResult, etc. (see Table 4).

Table 4. Some of EXPO classes.

experimental design strategy	fact	field of study
procedure	variable	experimental technology
scientific activity	hypothesis forming	interpreting results

Inspired by EXPO and CISP metadata, the work described in [13] proposes the *Core Scientific Concepts, CoreSCs*. The CoreSCs is a scheme built upon eleven categories at the sentence level that allows the automatic recognition of each one of the categories in scientific articles. The *CoreSC* includes: hypothesis, motivation, goal, object, background, method, experiment, model, observation, result and conclusion. The authors argued that these categories describe the main components of a scientific investigation. The first application of this scheme has been used to automatically annotate papers in Biochemistry and Chemistry.

Other works, such as The Argument Model Ontology[17], use the 'Toulmin Model of Argument' [14]. Toulmin proposed a layout containing six interrelated components for analysing arguments: claim, evidence, warrant, backing, qualifier and rebuttal. The Argument Model Ontology models this components through a set of 8 classes and 21 properties (see Table 5). The following snippet shows an example of use of this ontology:

[17] `http://www.essepuntato.it/2011/02/argumentmodel`

```
:sentence1 dcterms :description " We propose a catalog of
    domain independent conceptual abstractions for workflow
    steps that we call scientific workflow motifs " .
:sentence2 dcterms :description " We present an empirical
    analysis performed over 260 scientific workflow
    descriptions.  " .
:argument1 a amo:Argument ;
  amo:hasClaim :sentence1 ;
  amo:hasEvidence :sentence2 .
```

Table 5. List of the Argument Model Ontology classes, together with some properties.

Classes			
argument	argumentation entity	backing	claim
evidence	qualifier	rebuttal	warrant

Properties			
backs	forces	has evidence	has claim
involves	relates to	support	proves

A very recent work is the one proposed in [15] MicroPublications[18]. In this work the authors employ the Toulmin's model updated by Bart Verheij [16] and then they propose a semantic model of scientific argument and evidence designed for representing the key arguments and evidence in scientific articles. MicroPublications proposes a model to construct an argumentation network linking textual statements and data as evidence for claims.

Beyond the works that employ a linguistic model there are other works that are focused on describing the scientific discourse itself and the relations among the claims and hypotheses made by the author of the document. That is the case of the last two works that we present here.

The SWAN[19] ontology [5] models the scientific discourse. The SWAN project is part of the Annotation Ontology [17] and it has evolved into the Domeo annotation toolkit[20] (a web application enabling users to create and share ontology-based annotations on HTML and XML documents). The core of the SWAN ontology models the discourse elements providing a model of assertions, questions and hypotheses. The SWAN discourse elements are:

- Research statements: a claim or an hypothesis.
- Research questions: topics under investigation.
- Structured comments: the structured representation of a comment published in a digital resource.

On the other hand the SWAN ontology also provides discourse relationships, which are a set of relationships that can be used to build scientific discourse. Some of the discourse relationships are in Table 6.

[18] `http://purl.org/mp`
[19] Semantic Web Applications in Neuromedicine `http://www.w3.org/TR/hcls-swan/`
[20] `http://swan.mindinformatics.org/`

Table 6. The discourse relationships properties proposed by the SWAN ontology.

refers to	inconsistent with	alternative to
relevant to	arises from	motivates

4 Ontologies for describing bibliography and citations

The references of a document play an important role in the paper. One of the most widely used ontologies for describing bibliographic entities is BIBO, The Bibliographic Ontology Specification [18]. BIBO defines a set of classes to identify the type of document based on its origin (journal, book, webpage, etc.), where *bibo:Document* is the key class of this model. BIBO includes Dublin Core terms to cover common needs, uses FOAF (Friend of a Friend)[21] to describe authors, and adds other classes and properties, as shown in Table 7.

Table 7. List of some of the classes from BIBO, together with some properties.

Classes			Properties		
academic article	journal	collection	homepage	publisher	rights
book	chapter	issue	status	subject	time

Let us suppose that we want to annotate our journal paper with BIBO. In this case we would include the following information:

```
@prefix bibo: <http://purl.org/ontology/bibo> .
@prefix dc: <http://purl.org/dc/terms> .
[ a bibo:Article ;
  dc:title "Common motifs in scientific workflows..." ;
  dc:date "2013-09-21"
  bibo:volume "In press" ;
  dc:creator "Daniel Garijo"
  bibo:authorList ("Daniel Garijo" "..." "...")
 ...
]
```

Although the BIBO ontology identifies in a unique way each paper, BIBO is unable to express the history of a paper. In this sense, the work proposed by [19] extends the BIBO ontology in order to include the internal workflow of journals, conferences, and so forth. The result of this work allows tracking the history of a scientific paper.

As mentioned in Section 2, there are other ontologies focused on describing bibliographic entities, such as FaBiO[22] [8]. FaBiO describes bibliographic entities (e.g. books and journal articles) and their grouping (e.g. into book series and

[21] http://www.foaf-project.org/
[22] FaBiO http://purl.org/spar/fabio

journal issues). FaBiO classes are structured according to the FRBR schema of Works, Expressions, Manifestations and Items [20]. FaBiO has additional properties to extend the FRBR data model by linking the different parts of the FRBR schema. The FaBiO classes are divided into four main groups: Works with 69 subclasses, Expressions with 92 subclasses, Manifestations with 10 subclasses and Items with 4 subclasses.

Finally, the Citation Typing Ontology[23] enables the characterization of the nature or type of citations, both factually and rhetorically. CiTO contains 41 object properties that add more information to the cite (e.g., agrees with, corrects, likes, uses method in). CiTO allows characterising citations in three ways: explicit citations (e.g. the reference list of a journal article), indirect citations (e.g. a citation to a more recent paper by the same research group on the same topic), or implicit citations (e.g. as in artistic quotations or parodies, or in cases of plagiarism). Some of the CiTO properties are enumerated in Table 8.

Table 8. List of some of CiTO properties

agrees with	cites as authority	cites as evidence	confirms	corrects
describes	disagrees with	extends	includes excerpt from	discusses
supports	updates	uses conclusions from	uses data from	uses method in

Let us see an example of using CiTO for defining the citations of the journal paper.

```
@prefix cito: <http://purl.org/spar/cito/> .
:journalPaper cito:agrees_with <http://dx.doi.org/10.1016/j.
    future.2008.06.012> ;
         cito:extends <http://dx.doi.org/10.1109/eScience
             .2012.6404427> ;
```

5 Other vocabularies for describing documents

There are also other general ontologies that have been used for the annotation of any type of documents, not just scientific ones.

The most obvious and extended one is Dublin Core Metadata Terms (DCT), which contains fifteen properties to specify the characteristics of electronic documents (*creator, date, contributor, description, format*, etc). The terms in DCT are intended to be used in combination with terms from other vocabularies, as we saw above in the ORB vocabulary.

Friend of a Friend (FOAF) is a stable ontology that contains some classes such as Agent, Person, Organization, Group, Project, Document, Image, etc., and some properties to describe the instances of these classes. This vocabulary

[23] CiTO, namespace `http://purl.org/spar/cito`

allows describing the authors of documents, their affiliations and other relevant information about them.

Finally, the Semantic Web Conference Ontology[24] is an ontology for describing academic conferences. This ontology establishes how to use classes and properties from other ontologies (FOAF, Dublin Core, SIOC[25] and iCal/RDF Calendar[26]) and provides some classes for things relative to conferences. Some of these classes are: *AcademicEvent, ConferenceVenuePlace, Proceedings*, etc.

6 Conclusions and recommendations

We conclude this review paper by doing some recommendations about how to annotate scholarly documents with the aforementioned ontologies. These recommendations are part of our approach to describe scientific documents, which we are applying in the context of the DrInventor project in the domain of computer graphics.

First, we propose to use the DoCO ontology for describing the document structure. In Figure 1 (annotations with doco prefix) we can see how the *title, section, table* and *figure* classes are employed to annotate the different parts of the document. Due to space restrictions in the figure, all the classes do not appear. The annotation of the document structure allows comparing two or more papers according to their structure. It will also be possible to detect frequent document structures according to the domain or the kind of document (e.g. journal paper, book).

The next step is to annotate the rhetorical elements of the document. In this sense, Deo seems to be the most appropriate ontology. In Figure 1(annotations with deo prefix) we use the *background* class to identify what is the essential knowledge for understanding the problem. The class *contribution* annotates a description of the part that this publication plays in the overall field. Notice that Deo covers the most relevant rhetorical elements, but we should extend Deo with some specific classes according to the concrete domain. Particularly we are interested in the scientific domain, and we can see in Figure 1(annotations with scientific prefix) two new classes that are necessary for this domain: hypothesis and goal. We have defined a hypothesis as an "educated guess about how things work". A hypothesis should be "something that you can test". On the other hand, a goal of an investigation, according to the definition provided by [21], is the "target state of the investigation where intended discoveries are made, approaches are tested, problems are demonstrated, tasks formulated etc."

For describing the bibliography and citations we have employed BIBO and CiTO. BIBO describes in a detailed way each one of the references. In this sense we can establish a document classification. The second ontology, CiTO, has been employed for characterising the nature of the bibliographic citation linking the citing paper to the cited paper. Figure 1 shows that the citing paper extends

[24] `http://data.semanticweb.org/ns/swc/ontology`
[25] `http://sioc-project.org/`
[26] `http://www.w3.org/TR/rdfcal/`

the cited paper, because the author of the citing paper says that this document extends their previous work.

Finally, we point out how to use the FOAF ontology to annotate each one of the authors of the document. The use of this ontology allows describing people and their relations.

In brief, we propose to use the DoCO ontology for describing the document structure, the Deo ontology gives way to describe the vast majority of rhetorical elements, but we believe that it is necessary to extend this ontology in order to cover the concrete elements of the application domain. The BIBO ontology describes references according the kind of document and all the characteristics involved in the publication process. We propose to use the CiTO ontology for describing the rhetoric of the citations (in this way we can establish a network with other works). In order to describe the scientific discourse it is necessary to describe the most important elements of the domain (e.g. the biological domain has probably different discourse elements than the computer graphics domain). An important part in the scientific discourse are the claims done by the author of the paper and how they are contextualized. For this purpose we propose to use the discourse model proposed by [11]. This model can be annotated with most of the CiTO properties. In this sense, our nearly future work is to provide an ontology for describing the scientific discourse according to the aforementioned model by doing an extension of the proposed ontologies.

In this paper we have described the most relevant ontologies used for describing scientific documents. We have classified these ontologies into three main groups, those that describe the discourse of a document (either scientific or generic), those that allow describing the document structure and those dedicated to describe references and citations. At last, we have presented other well-known vocabularies that provide generic information about any type of document. Moreover, we have also sketched out a model for the semantic enhancement of documents based on some of these ontologies, mainly those that describe the document structure, their rhetorical elements and their references.

Acknowledgments

This work is supported by the FP7 European project Dr Inventor FP7-611383.

References

1. Shotton, D.: Semantic publishing: the coming revolution in scientific journal publishing. Learned Publishing **22**(2) (April 2009) 85–94
2. Waard, A.d.: From proteins to fairytales: Directions in semantic publishing. IEEE Intelligent Systems **25**(2) (March 2010) 83–88
3. Soldatova, L.N., King, R.D.: An ontology of scientific experiments. Journal of The Royal Society Interface **3**(11) (December 2006) 795–803
4. Kohlhase, M.: OMDoc: Open mathematical documents. In: OMDoc An Open Markup Format for Mathematical Documents. Number 4180 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (January 2006) 25–32

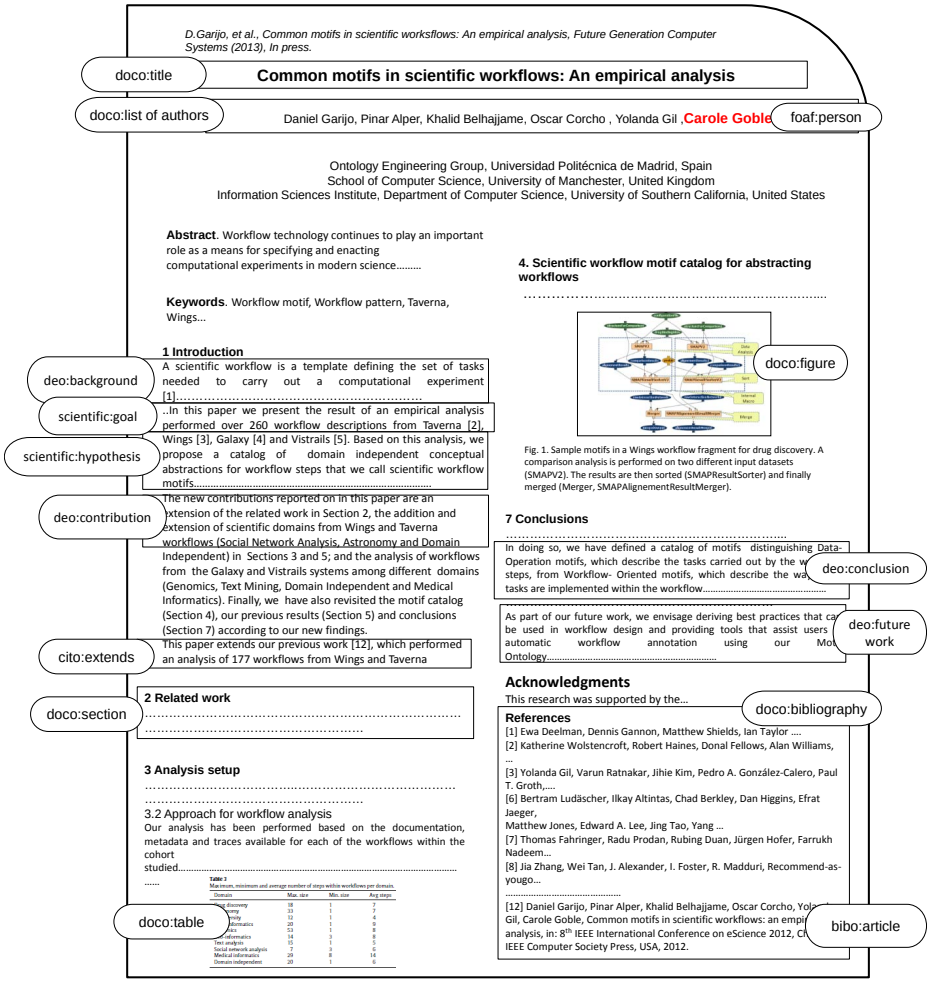

Fig. 1. The proposed model for describing a scientific document taking together some of the described ontologies.

5. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The SWAN biomedical discourse ontology. Journal of Biomedical Informatics **41**(5) (October 2008) 739–751
6. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems **29**(2) (February 2013) 599–611
7. Sollaci, L.B., Pereira, M.G.: The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. Journal of the Medical Library Association **92**(3) (July 2004) 364–371
8. Peroni, S., Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Web Semantics: Science, Services and Agents on the World Wide Web **17** (December 2012) 33–43
9. Groza, T., Handschuh, S., Möller, K., Decker, S.: SALT - semantically annotated LaTeX for scientific publications. In Franconi, E., Kifer, M., May, W., eds.: The Semantic Web: Research and Applications. Number 4519 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (January 2007) 518–532
10. Shum, S.B., Motta, E., Domingue, J.: ScholOnto: an ontology-based digital library server for research documents and discourse. International Journal on Digital Libraries **3**(3) (October 2000) 237–248
11. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R.: Corpora for the conceptualisation and zoning of scientific papers. In: Proceedings of the International Conference on Language Resources and Evaluation. (2010)
12. Soldatova, L., Liakata, M.: An ontology methodology and CISP - the proposed core information about scientific papers. Programme/Project deposit (December 2007)
13. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics **28**(7) (January 2012) 991–1000
14. Toulmin, S.E.: The Uses of Argument. Cambridge University Press (July 2003)
15. Clark, T., Ciccarese, P.N., Goble, C.A.: Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Submitted to the Journal of Biomedical Semantics (May 2013)
16. Verheij, B.: The toulmin argument model in artificial intelligence. In Simari, G., Rahwan, I., eds.: Argumentation in Artificial Intelligence. Springer US (January 2009) 219–238
17. Ciccarese, P., Ocana, M., Garcia Castro, L., Das, S., Clark, T.: An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics **2**(2) (July 2011) 1–24
18. D'Arcus, B., Giasson, F.: Bibliographic ontology specification (November 2009)
19. Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P.: A linked-data-driven and semantically-enabled journal portal for scientometrics. In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K., eds.: The Semantic Web. Number 8219 in Lecture Notes in Computer Science. Springer Berlin Heidelberg (January 2013) 114–129
20. Tillett, B.: What is FRBR? a conceptual model for the bibliographic universe. The Australian Library Journal **54**(1) (2005) 24–30
21. Liakata, M., Soldatova, L.: Guidelines for the annotation of general scientific concepts. Aberystwyth University, JISC Project Report (November 2008)