Faceted Documents: Describing Document Characteristics Using Semantic Lenses

Silvio Peroni
University of Bologna (Italy)
essepuntato@cs.unibo.it

David Shotton
University of Oxford (UK)
david.shotton@zoo.ox.ac.uk

Fabio Vitali
University of Bologna (Italy)
fabio@cs.unibo.it

ABSTRACT

The semantic enhancement of a traditional scientific paper is not a straightforward operation, since it involves many different aspects or facets. In this paper we propose eight different *semantic lenses* through which these facets may be viewed, and describe and exemplify the ontologies by which these lenses may be implemented.

Categories and Subject Descriptors

I.7.2 [**Document And Text Processing**]: Document Capture— *Document analysis*

Keywords

Semantic Web, document semantics, semantic publishing

1. INTRODUCTION

The enhancement of a traditional scientific paper with semantic annotations – one of the most important activities within the expanding field of *semantic publishing* [13] – is not a straightforward operation, since it involved much more that simply making semantically precise statements about named entities within the text. There are many additional aspects to a paper beyond the bare words it contains, that combine together to create an effective unit of scholarly communication. These include the context of the publication, contributing to the overall credibility and authoritativeness of the scientific activity, the structural components of the publication (e.g. author list, sections, tables, reference list, etc.) and in particular the rhetorically distinct sections of the publication (e.g. Introduction, Results, Discussion), the rhetorical devices used in the text, that contribute to its argumentative and persuasive power, and the citations that connect the publication with its wider context of scholarship.

These and other aspects coexist, and are usually so well integrated into the paper as a whole, and into the rhetorical flow of the natural language of the text, as to be scarcely discernible as separate entities by the reader. However, in order to create machine-readable semantic annotations over the paper, each of these aspects has to be clearly and separately identified and described, since each impacts and affects the semantic characterization of the content in different ways. Examining the semantic characterization of each of these aspects of a document can be envisaged as applying a set of *semantic lenses*, each of which magnifies or reveals one aspect or facet of the whole.

In this paper we propose a model for the semantic enhancement of scientific papers based on eight such *semantic lenses* that can be used to characterize its facets and in this way enhance its usefulness. These eight semantic lenses are:

- *Research context*: information about the background from which the paper emerged (the research reported, the institutions involved, the sources of funding, etc.).
- *Contributions and roles*: details about which individuals hold particular authorship roles for the paper and what specific contributions different people made.
- *Publication context*: information about related conferences, the journal in which the paper was published and the other papers with which it appeared.
- *Structure*: the explicit structural components (sections, paragraphs, etc.) into which the paper is organized.
- *Rhetoric*: the organization of the paper in terms of rhetorical sections having different purposes (Introduction, Results, Discussion, etc.).
- *Citation*: the purpose and target of each individual reference in the paper, and the manner in which the paper fits within the citation network.
- *Argumentation*: the structure and expression of each assertion within the paper, as a component of an argument to justify or invalidate a claim.
- *Semantics*: the actual meaning of each assertion, statement or named entity within the paper.

In what follows, we expand on these *semantic lenses*, and show how ontologies can be employed to describe each of the relevant facets of a paper, so that, when take together, these provide a complete semantic description of a scientific publication, its relationships with similar publications, and its role in the world of scholarship. The rest of this paper is structured as follows: in Section 2 we introduce our model in greater detail and present the various semantic technologies used to describe each facet of the paper; and in Section 3 we summarize and draw some conclusions.

DocEng'12, September 4–7, 2012, Paris, France.

2. SEMANTIC LENSES

The semantics of a scientific paper (or, more generally, of a document) is definable from different perspectives. Each perspective may be thought of as a *semantic lens* that can be *applied* to a document to reveal a particular semantic facet. In Section 1, we identified eight different semantic lenses that cover different perspectives. In the following sections we elaborate on the theories behind each of these semantic lenses, describe them, and discuss and exemplify the use of ontologies, developed either in previous works or specifically for this paper, that make possible the application of these lenses to documents by means of Semantic Web technologies.

2.1 The research context lens

Writing a scientific paper is usually the final stage of an often complex collaborative and multi-domain activity of undertaking the research investigation from which the paper arises. The organizations involved, the people affiliated to these organizations, the grants provided by funding agencies, the research projects funded by such grants: all these provide the research *context* that leads, directly or indirectly, to the genesis of the paper, and awareness of these may have a strong impact on the credibility and authoritativeness of its scientific content. A number of vocabularies for the description of research projects and related entities have been developed, e.g. the VIVO Ontology[1] – developed for describing the social networks of academics, their research and teaching activities, their expertise, and their relationships to information resources – and DOAP, the *Description Of A Project*[2] – an ontology with multi-lingual definitions that contains terms specific for software development projects.

To permit description of this research context, we have developed FRAPO, the *Funding, Research Administration and Projects Ontology*[3], that can be used for applying the *research context* lens to a paper, as illustrated as follows[4]:

```
:jisc a frapo:FundingAgency ;
  frapo:awards [ a frapo:Grant ;
    frapo:funds :open-citations-project ] .
:open-citations-project a foaf:Project ;
  foaf:homepage <http://opencitations.org> ;
  frapo:enables :spar .
:spar a frapo:Endeavour ;
  foaf:homepage <http://purl.org/spar> ;
  foaf:page :lenses-paper .
:lenses-paper a foaf:Document . # This paper
```

2.2 The contributions and roles lens

People can have a variety of *roles* in research projects and in the authorship of articles, and additionally can make different *contributions* to these activities with varying degrees of effort. This aspect of semantic description is made possible by our development of SCoRO, the *Scholarly Contributions and Roles Ontology*[5], as shown as follows:

```
:adventures a fabio:ResearchPaper . # Ref. [13]
:shotton a foaf:Person ;
  foaf:name "David Shotton" ;
  scoro:hasAuthorshipRole [ a pro:RoleInTime ;
    pro:withRole scoro:corresponding-author ;
    pro:relatesToDocument :adventures ] ;
  scoro:makesContribution [
    a scoro:ContributionSituation ;
    scoro:withContribution scoro:writes-paper ;
    scoro:withContributionEffort
      scoro:major-effort ;
    scoro:relatesToEntity :adventures ] .
```

2.3 The publication context lens

When analysing the social context in which a scientific paper is written, it is important to understand how it is *grouped* with other documents. For instance, it is relevant to know the book, journal and/or conference proceedings within which a paper appears, and separately to be able to describe groupings of bibliographic records and references, e.g. in tables of contents, reference lists, reference management systems and library catalogues. One of the most widely used ontology for describing bibliographic entities and their aggregations is BIBO, the *Bibliographic Ontology* [4]. FRBR, *Functional Requirements for Bibliographic Records* [8], is yet another more structured model for describing documents and their evolution in time. One of the most important aspects of FRBR is the fact that it is not tied to a particular metadata schema or implementation.

For this purpose we have developed two ontologies, FaBiO, the *FRBR-aligned Bibliographic Ontology*[6] [12] to describe bibliographic entities (e.g. books and journal articles) and their grouping (e.g. into book series and journal issues), and BiRO, the *Bibliographic Reference Ontology*[7], that permits the description of collections, for example of references in a reference list. Their use is exemplified as follows:

```
:version-of-record a fabio:JournalArticle ;
  frbr:realisationOf :adventures ;
  frbr:partOf [ a fabio:JournalIssue ;
     prism:issueIdentifier "4" ;
  frbr:partOf [ a fabio:JournalVolume ;
    prism:volume "5" ;
  frbr:partOf [ a fabio:Journal ;
    dcterms:title "PloS Computational Biology"
       ] ] ] ;
  frbr:part [ a biro:ReferenceList ;
    co:element [ biro:references
      <http://dx.doi.org/10.1371/journal.pcbi
          .0010034> ] ... ] .
```

2.4 The structure lens

The *structure* of a textual document is often expressed by means of markup languages such as XML and LaTeX, that have constructs for describing content hierarchically. We have been investigating patterns in XML vocabularies to understand how the structure of digital documents can be segmented into atomic components which can then be addressed and understood independently. Instead of defining a large number of complex and different structures, we identified eleven *structural patterns* [5] that have proved to be sufficient to express the structure of most documents, including scientific papers. This model, implemented in the *Pattern Ontology* (*PO*)[8], can be used in combination with

[1]VIVO Ontology: http://vivoweb.org/ontology/core
[2]DOAP: http://usefulinc.com/ns/doap
[3]FRAPO: http://purl.org/cerif/frapo
[4]This and the following RDF examples are written in Turtle (http://www.w3.org/TeamSubmission/turtle/), with namespace definitions defined at http://www.essepuntato.it/lenses-paper/prefixes.
[5]SCoRO: http://purl.org/spar/scoro
[6]FaBiO: http://purl.org/spar/fabio
[7]BiRO: http://purl.org/spar/biro
[8]PO: http://www.essepuntato.it/2008/12/pattern

EARMARK [6], an ontology[9] describing a markup metalanguage, to describe the structure of the document as a set of OWL assertions, and then to associate formal and explicit semantics with these descriptions. Thus we can associate a particular structural semantics to elements (e.g. an element *h1 expresses* the concept of being a block of text, while the element *div* containing it is a container). For instance, the first section of this paper[10] can be described as follows:

```
:div1 a earmark:Element # Sec. Introduction
  la:expresses pattern:HeadedContainer ;
  earmark:hasGeneralIdentifier "div" ;
  c:firstItem [ c:itemContent :h1 ... ] .
:h1 a earmark:Element # Title of the sec.
  la:expresses pattern:Block ;
  earmark:hasGeneralIdentifier "h1" ;
  c:firstItem [ c:itemContent :r1 ] .
:r1 a earmark:PointerRange ... # Text content
```

2.5 The rhetoric lens

Often, scientific communities require their papers to follow a particular rhetorical organization of sections, in order to identify meaningful aspects of the scientific discourse explicitly. These rhetorical components, for example Introduction, Methods, Results and Conclusions, give a defined rhetorical structure to the paper, which assists readers.

Previous works that introduced models to characterise such *rhetorical* aspect of a scientific publication included the rhetorical blocks within SALT, the *Semantical Annotated LaTeX Ontology* [7] and ORB, the *Ontology of Rhetorical Blocks* [2], as well as DEO, the *Discourse Elements Ontology*[11]. However, the rhetoric organization of a paper does not necessarily correspond neatly to its structural components (sections, paragraphs, etc.). Thus, in order to enable description both of the purely structural components of a document (introduced in Section 2.4) and its rhetorical components, we have developed DoCO, the *Document Components Ontology*[12]. This ontology provides the means to describe the organization of a document from both the structural and the rhetorical perspectives, as shown in the following brief example:

```
:div1 la:expresses doco:Section ,
  deo:Introduction .
:h1 la:expresses doco:SectionTitle . # etc.
```

2.6 The citation lens

Measuring how papers *cite* each other is often undertaken to generate metrics for the productivity of scientists and the impact of journals. Although citation metrics presently register the simple fact that one paper cites another, improved measures of the impact of the research and the productivity of authors might wish to take into account the *reasons* for particular citations, e.g. to express qualification of or disagreement with the ideas presented in the cited paper, which may significantly effect the evaluation of a citation network. In fact, it would seem sensible to weight differently citations that criticises a cited work from those used to acknowledge the benefit gained from its authoritative content.

CiTO, the *Citation Typing Ontology*[13] [12] allows one not only to assert in RDF that citations exists, but also to define the factual or rhetorical nature of the citations, as shown in the following example:

```
:lenses-paper cito:usesMethodIn
  <http://www.cambridge.org/0521092302> ;
  cito:citesForInformation :adventures .
```

2.7 The argumentation lens

The *argumentation* of the claims of the paper is crucial for scholarly and scientific publishing, in proposing hypotheses and advancing evidence in their support. Several works have been proposed in the past to model the argumentation of papers. For instance, the SALT application [7] permits someone such as the author "to enrich the document with formal descriptions of claims, supports and rhetorical relation as part of their writing process". There are other works, based on [14], that offer an application of Toulmin's model within specific scholarly domains, for instance the legal and legislative domain [9]. In [14], Toulmin proposed that arguments (including scientific arguments) are composed of statements having specific argumentative roles, of which three are essential:

- **The claim.** A fact that must be asserted – e.g. "This text is a scientific paper".
- **The evidence.** Another fact that represents a foundation for the claim – e.g. "This paper has been accepted to a scientific conference".
- **The warrant.** A statement bridging from the evidence to the claim – e.g. "A paper accepted to a scientific conference is a scientific paper".

In Toulmin's model, each instance that has a certain role in an argument (e.g. a warrant) may very well be the claim of another sub-argument. And, each statement of the sub-argument could be the claim of yet other sub-sub-arguments.

In order to use this argumentation theory, we have developed AMO, the *Argument Model Ontology*[14]. It allows one to express an argument using Toulmin's argumentation theory, as shown in the following excerpt:

```
:sentence1 dcterms:description "This is a
    scientific paper" .
:sentence2 dcterms:description "This paper has
    been accepted to a scientific conference" .
:sentence3 dcterms:description "A paper
    accepted to a scientific conference is a
    scientific paper" .
:argument1 a amo:Argument ;
  amo:hasClaim :sentence1 ;
  amo:hasEvidence :sentence2 ;
  amo:hasWarrant :sentence3 .
:argument2 a amo:Argument ;
  amo:hasClaim :sentence3 ; # etc.
```

2.8 The semantics lens

The main goal of a scientific paper is to express (and cite) findings that have specific scientific value. These finding are expressed through text, tables, and figures. Usually, they

[9]EARMARK: http://www.essepuntato.it/2008/12/earmark
[10]The XML version of this paper is available at http://www.essepuntato.it/lenses-paper/xml.
[11]DEO: http://purl.org/spar/deo
[12]DoCO: http://purl.org/spar/doco
[13]CiTO: http://purl.org/spar/cito
[14]AMO: http://www.essepuntato.it/2011/02/argumentmodel

are meant for human interpretation only and are not directly suitable for machines. This because the *semantics* of a piece of text, such as "EARMARK is more expressive than XML", is not explicitly defined in any formal way: it is just text that requires human interpretation. The *semantics lens* is employed to encode the meaning of the original scientific message contained in the paper using Semantic Web technologies such as RDF, as shown in the following statement:

```
<http://www.essepuntato.it/2008/12/earmark>
:isMoreExpressiveThan <http://www.w3.org/XML> .
```

2.9 An authorial activity

The application of a particular semantic lens to a paper involves adding information about the particular facet of semantics described by that lens, and constitutes an *authorial activity*, i.e. an action of a person (who may be the original author of the human-readable text, or someone else) who takes responsible for the specification of the semantic interpretations given to the document. Therefore, the tracking of semantic lens applications involves a requirement to record *data provenance*, i.e. the identification of the tools and processes that were involved in the creation of an artefact or resource, and the people involved in that creation. To encode provenance information, OPM, the *Open Provenance Model* [11], is a well-known model whose main requirements concerns the exchangeability of provenance data between systems, the digital representation of provenance for any resource, and the definition of a set of rules identifying the valid inferences that can be made on provenance graphs. Another implemented model for provenance is included in the SWAN ontology ecosystem [3], and aims to describe resources in terms of their accessing, authoring and versioning.

In our opinion, the *PROV Ontology* (*PROV-O*) [10] is one of the more appropriate ontologies for the definition of the authorial role of agents. Each set of RDF statements, produced as consequence of a lens application, should be enclosed within a *named graph* [1], in order to express all the provenance data as statements about the graph itself:

```
:citation-lens {
  :lenses-paper cito:citesForInformation
    :adventures ; ... }
:citation-lens a prov:Entity ,
  lens:CitationLensApplication ;
  prov:wasGeneratedBy [ a prov:Activity ;
    prov:wasAssociatedWith
      <http://www.essepuntato.it/me> ] . # etc.
```

3. CONCLUSIONS

In this paper we have sketched out a model for the enhancement of documents based on eight distinct *semantic lenses*, each of which can be used to identify a specific semantic facet of a document. Moreover, we have presented technologies, in the form of OWL ontologies, that can be used for the application of these lenses to actual documents, and for the specification of the authorial roles of the people responsible for these operations. Future works will mainly involve the improvement of the lens model, the development of automatic and semi-automatic tools for the application of the lenses to documents, the implementation of user interfaces for the dynamic interpretation and use of lens-related document semantics, and additional studies to analyse whether our model is enough generic or strictly depends on particular types of documents (e.g. scientific papers).

4. REFERENCES

[1] Carroll, J., Bizer, C., Hayes, P., Stickler, P. (2005). Named Graphs, Provenance and Trust. In Proceedings of the 14th International World Wide Web Conference: 613-622. DOI: 10.1145/1060745.1060835

[2] Ciccarese, P., Groza, T. (2011). Ontology of Rhetorical Blocks (ORB). W3C Editor's Draft. World Wide Web Consortium. http://www.w3.org/2001/sw/hcls/notes/orb/

[3] Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T. (2008). The SWAN biomedical discourse ontology. In Journal of Biomedical Informatics, 41 (5): 739-751. DOI: 10.1016/j.jbi.2008.04.010

[4] D'Arcus, B., Giasson, F. (2009). Bibliographic Ontology Specification. Specification Document, 4 November 2009. http://bibliontology.com/specification

[5] Di Iorio, A., Peroni, S., Poggi, F., Vitali, F. (2012). A first approach to the automatic recognition of structural patterns in XML documents. To appear in the Proceedings of the 2012 ACM symposium on Document Engineering.

[6] Di Iorio, A., Peroni, S., Vitali, F. (2011). A Semantic Web Approach To Everyday Overlapping Markup. In Journal of the American Society for Information Science and Technology, 62 (9): 1696-1716. DOI: 10.1002/asi.21591

[7] Groza, T., Moller, K., Handschuh, S., Trif, D., Decker, S. (2007). SALT: Weaving the claim web. In Proceedings of the 6th International Semantic Web Conference: 197-210. DOI:10.1007/978-3-540-76298-0_15

[8] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Functional Requirements for Bibliographic Records (FRBR). Final Report, http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm

[9] Lauritsen, M., Gordon, T. F. (2009). Toward a general theory of document modeling. In Proceedings of the 12th International Conference on Artificial Intelligence and Law: 202-211. DOI:10.1145/1568234.1568257

[10] Lebo, T., Sahoo, S., McGuinness, D. (2012). PROV-O: The PROV Ontology. W3C Working Draft 03 May 2012. World Wide Web Consortium. http://www.w3.org/TR/prov-o

[11] Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., Paulson, P. (2008). The Open Provenance Model: An Overview. In Proceedings of the 2nd International Provenance and Annotation Workshop: 323-326. DOI: 10.1007/978-3-540-89965-5_31

[12] Peroni, S., Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. In press, Journal of Web Semantics. http://imageweb.zoo.ox.ac.uk/pub/2012/publications/Peroni&Shotton_fabiocito_ontology_paper_JWSaccepted.pdf

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

[14] Toulmin, S. (1959). The uses of argument. Cambridge University Press. ISBN: 0521827485