Using annotations to model discourse: an extension to the Annotation Ontology

Leyla Jael García-Castro[1], Olga Giraldo[2], Alexander García[3]

[1] Universität der Bundeswehr München, Werner-Heisenberg-Weg 39,
85779 Neubiberg, Germany
w31blega@unibw.de
[2] Universidad Politécnica de Madrid, Ontology Engineering Group,
Madrid, Spain
oxgiraldo@gmail.com
[3] Florida State University,
Tallahassee, Florida, USA
alexgarciac@gmail.com

Abstract. The Annotation Ontology (AO) has proven to be a valuable resource for structuring annotations in scientific documents. We are representing elements of discourse with the AO; by using our proposed extension it is possible to mark up specific rhetorical structures and build a network of interconnected documents. The extension presented in this paper also makes it possible to represent more expressive associations across nanopublications.

Keywords: Scientific publications, social tagging systems, social and semantic web, knowledge discovery

1 Introduction

Digital Libraries such as Elsevier Science Direct[1] or PubMed[2] store electronic versions of scientific publications. Although these resources provide some information retrieval mechanisms, it is still difficult to extract facts buried in the text [1]; for instance, retrieving definitions and claims from literature is usually a manual process. Making the content explicitly identifiable by means of Semantic Web (SW) technology has been proposed as a feasible solution for improving information retrieval across digital libraries; enriching the metadata should make it possible to identify and extract facts buried in documents [1, 2]. Documents should be self-descriptive and fully immersed in the web of data [3].

Heading towards a self-descriptive document requires a well-organized annotation structure consistent with the underlying rhetorical structure. Annotation should support not only marking segments but also making the relationships across these portions explicit. Furthermore, annotations should scaffold relations across documents. It is not enough to know the concepts in a document; it is also necessary

[1] http://www.sciencedirect.com/

[2] http://www.ncbi.nlm.nih.gov/pubmed/

to know how are they related [4] -within the document, across documents and to the web of data.

The Annotation Ontology (AO) [5] facilitates modeling annotations on static resources; an ongoing project will extend the AO in order to facilitate the annotation on mutable objects as well [6]. The AO is built upon the Annotea Project[3] and supports both free and semantic annotations: free annotations are expressed by plain text attached to resources whilst semantic annotations should also include a relation *ao:hasTopic* to an ontological entity. Annotations can be attached to the whole resource but also to portions of it, *e.g.* sentences, paragraphs, sections, images, tables, etc. Annotations on any fragment within a document should be modeled by using selectors; a selector identifies the fragment depending on its nature: *aos:TextSelector* identifies a exact match to a piece of text, *aos:StartEndSelector* identifies the initial and final position that the annotation refers to, *aos:InitEndCornerSelection* identifies the initial and final (x,y) coordinates within an image, etc. The AO offers several types of annotations such as notes, comments, erratum, etc. Qualifiers are a particular type of annotations mapped to the Simple Knowledge Organization System[4] (SKOS) properties and particularly useful for semantic annotations: *ao:Qualifer* maps to *skos:RelatedMatch*, *ao:ExactQualifer* to *skos:exactMatch*, *ao:CloseQualifier* to *skos:closeMatch*, *ao:BroadQualifer* to *skos:broadMatch*, and *ao:NarrowQualifier* to *skos:narrowMatch*. The provenance within AO is supported by the Provenance Authoring and Versioning ontology[5] that provides features on provenance to support scientific content and its curation. Scientific discourses are modeled by integrating the AO and SWAN [7].

We are broadening the interoperability between SWAN and AO [5], going beyond the current integration[6]. We are including concepts from CoreSC [8], SWAN, the Sample Processing and Separation Techniques (SEP), Ontology for Biomedical Investigations (OBI), Micro Array Gene Expression Ontology (MGED), and National Cancer Institute Thesaurus (NCIt). On the one hand, we want to make explicit some discursive elements in scientific publications, for instance, the structural elements related to the arrangement and distribution of the document. We are also interested in identifying argumentative elements, *i.e.* elements of the discourse. On the other hand, we are adding new qualifiers and representing annotations on annotations to express relations across elements and the initiation of a topic thread, *i.e.* an argumentative line anchored in the document. We have initially focused our efforts in modeling literature reviews; although these papers summarize findings reported in other documents and offer insightful analysis of existing literature, extracting claims, definitions, data, and other data types is cumbersome –partly due to the lack of markers for these structures. Moreover, as literature reviews bring together information from existing documents by pulling out facts and structuring them in a new document, such a network is not explicit; we are providing the structure so that literature reviews can be seen as a collection of scaffolded annotations and/or nanopublications.

[3] http:// www.w3.org/2001/Annotea/

[4] http:// www.w3.org/2004/02/skos/

[5] PAV, swan.mindinformatics.org/spec/1.2/pav.html

[6] http://code.google.com/p/annotation-ontology/wiki/SWANDiscourse

2 Rhetoric and discourse from Annotations

2.1 AO extension to model rhetoric and discourse elements

We have extended the AO with new classes and properties that facilitate making explicit the rhetoric and discourse embedded in a scientific publication. Classes are meant to categorize the type of structures expressed in a publication -*e.g.* definitions, examples, claims, etc. From the AO, we have reused *ao:Definition* making it compliant to the Meaning-of-a-tag (MOAT) [9]. We also reused *ao:Example* as it was originally proposed. In addition, we integrated classes from SWAN vr. 1.2[7] as well as from CoreSC. Table 1 summarizes the proposed classes and presents a short description of the intended purpose. Whenever a class matches an entity from another vocabulary the description is taken from there, descriptions taken from the Cambridge Dictionaries Online[8] are identified as CDO, and no quoted descriptions are defined by the authors.

Table 1. Classes modeling concepts in a scientific publication

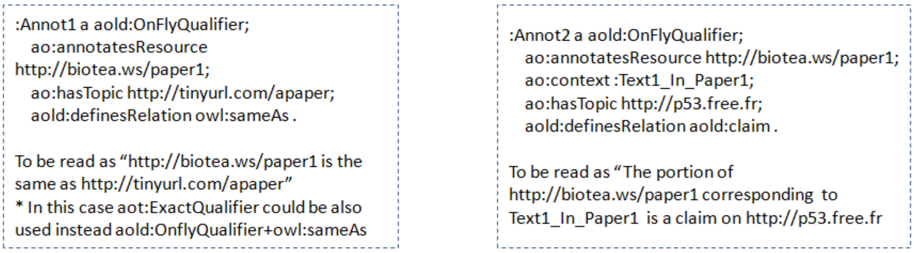
Class name	Description
aold:Introduction	Section used to broadly present the problem, existing solutions, and what the research work intends to achieve
aold:Motivation	coresc:Motivation, "The reasons behind an investigation"
aold:Aim	CDO "a result that your plans or actions are intended to achieve"
aold:Goal	coresc:Goal, "A target state of the investigation where intended discoveries are made"
aold:Hypothesis	coresc:Hypothesis, "A statement not yet confirmed rather than a factual statement"
ao:ResearchQuestion	swan:ResearchQuestion
ao:ResearchStatement	swan:ResearchStatement
aold:Reference	coresc:Background, "Generally accepted background knowledge and previous work"
aold:Report	obi:report "a document assembled by an author for the purpose of providing information for the audience. A report is the output of a documenting process and has the objective to be consumed by a specific audience."
aold:Counter-Example	Example that contradicts a statement or idea
aold:Opinion	CDO "a thought or belief about something or someone, a judgment about someone or something"
aold:Claim	CDO "to say that something is true or is a fact, although you cannot prove it and other people might

[7] http://swan.mindinformatics.org/ontologies/1.2/discourseelements.owl

[8] http://dictionary.cambridge.org/

	not believe it”
aold:Method	NCIt “A means, manner of procedure, or systematic course of actions that have to be performed in order to accomplish a particular goal.”
aold:Sample	SEP “A sample is a substance role played by a biological substance as an input substance to a protocol.”
aold:Protocol	OBI “a protocol is a plan specification which has sufficient level of detail and quantitative information to communicate it between domain experts, so that different domain experts will reliably be able to independently reproduce the process.”
aold:Model	coresc:Model, “A statement about a theoretical model or framework”
aold:Experiment	MGED “The complete set of assays and their descriptions performed as an experiment for a common purpose.”
aold:ObservationInResearch	NCIt “Watching something and taking note of what happens.”
aold:Result	coresc:Result, “Factual statements about the outputs of an investigation”
aold:Discussion	The annotation identifies a fragment corresponding to the discussion of the document
aold:Conclusion	coresc:Conclusion, “Statements inferred from observations & results relating to research hypothesis”

We are also proposing classes that facilitate the definition of relations between entities; this makes it possible to relate fragments within the same document or across multiple documents. Qualifiers in the AO are mapped to SKOS properties; here we interpret them as expressing a subjacent relationship between the annotated fragment/document and the topic. It is recommended to use ao:Annotation+*ao:hasTopic* whenever it is needed to point to examples, definitions and external links related to URIs. The ao:Qualifier/***aold:OnFlyQualifier***+*ao:hasTopic* should be used to relate the annotated fragment/document to an entity/resource, *e.g.* to relate “mouse” to an ontological term “ncbitaxon:10090”. Qualifiers, as proposed by AO, express only five relationships, we are proposing ***aold:OnFlyQualifier*** that extends the original *ao:Annotation* to model any relationship that has been defined somewhere else, typically an ontology. The property ***aold:definesRelation*** maps to the subjacent relation, *e.g. owl:sameAs*, between the annotated fragment/document and the topic. Relations used by a specific annotation project can be narrowed down to a set of predefined relations; they may also be open to represent any relation expressed as free text. The last scenario would probably require curation mechanisms to see whether the relation is new or can be mapped to an existing one. Fig. 1 shows two possible uses of ***aold:OnFlyQualifier****:* on the left, an annotator has identified the paper with URI “http://biotea.ws/paper1” as the same at “http://tinyurl.com/apaper”; on the right, an annotator has identified a claim and its corresponding source.



```
:Annot1 a aold:OnFlyQualifier;
  ao:annotatesResource
http://biotea.ws/paper1;
  ao:hasTopic http://tinyurl.com/apaper;
  aold:definesRelation owl:sameAs .
```

To be read as "http://biotea.ws/paper1 is the same as http://tinyurl.com/apaper"
* In this case aot:ExactQualifier could be also used instead aold:OnflyQualifier+owl:sameAs

```
:Annot2 a aold:OnFlyQualifier;
  ao:annotatesResource http://biotea.ws/paper1;
  ao:context :Text1_In_Paper1;
  ao:hasTopic http://p53.free.fr;
  aold:definesRelation aold:claim .
```

To be read as "The portion of http://biotea.ws/paper1 corresponding to Text1_In_Paper1 is a claim on http://p53.free.fr

Fig. 1. aold:OnFlyQualifier in use

The ***aold:Relation*** extends the ***aold:OnFlyQualifier***, it represents an annotation that brings together a pair of annotations; the *ao:body* of the annotation establishes the intended name for the annotation. In this way, it is possible to use ***aold:OnFlyQualifier*** for known relations while using the ***aold:Relation*** may be reserved for new ones. Two subclasses have been proposed, ***aold:UnidirectionalRelation*** for those relations where the subject and object cannot be interchanged, *e.g. is_a*, and ***aold:BidirectionalRelation*** for all other relations, e.g. synonyms. If ***aold:definesRelation*** is used, it defines a relation, e.*g. sameAs*, between the *ao:body* and the topic. Fig. 2 shows a graphical representation of these classes.

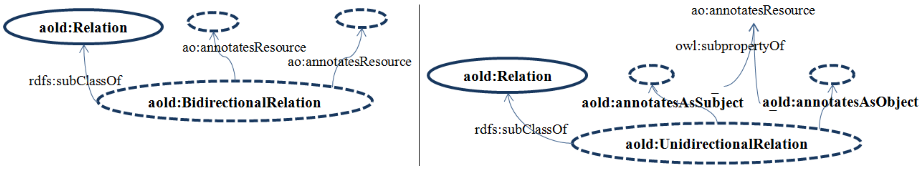
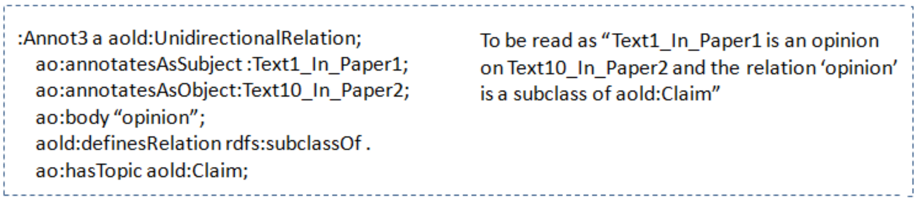


Fig. 2. aold:Relation, unidirectional and bidirectional

aold:Relation is useful in cases such as definitions in vocabularies as well as when making explicit a process described by a document. Additional information can be found at http://biotea.ws. Fig. 3 illustrates how could ***aold:Relation*** be used when expressing new ways to relate documents. An annotated fragment in a document is categorized as an "opinion" on a fragment of a second document.



```
:Annot3 a aold:UnidirectionalRelation;
  ao:annotatesAsSubject :Text1_In_Paper1;
  ao:annotatesAsObject:Text10_In_Paper2;
  ao:body "opinion";
  aold:definesRelation rdfs:subclassOf .
  ao:hasTopic aold:Claim;
```

To be read as "Text1_In_Paper1 is an opinion on Text10_In_Paper2 and the relation 'opinion' is a subclass of aold:Claim"

Fig. 3. aold:UnidirectionalRelation

We have also added two new selectors that work on RDF documents. When working with these selectors it is assumed that only one ***rdfs:comment*** will be present in the RDF element being annotated. The ***aold:ElementSelector*** extends the ***aos:textSelector,*** it identifies an exact chunk of text in the ***rdfs:comment*** while the ***aold:StarEndElementSelector*** extends the **ao:StartEndSelector** by identifying the *start* and *end* positions of the text being annotated in the ***rdfs:comment***. Fig. 4 illustrates an example using the latter selector: A text from position 27 to position 38 in the "*Introduction*" section of the corresponding RDF representation for the paper with DOI 10.1016/SO014-5793(03)00051-6 has been annotated; the annotation body is "*mus musculus*". The ***aold:OnFlyQualifier*** is here used to indicate that the annotated text corresponds to *ncbitaxon:10090*.

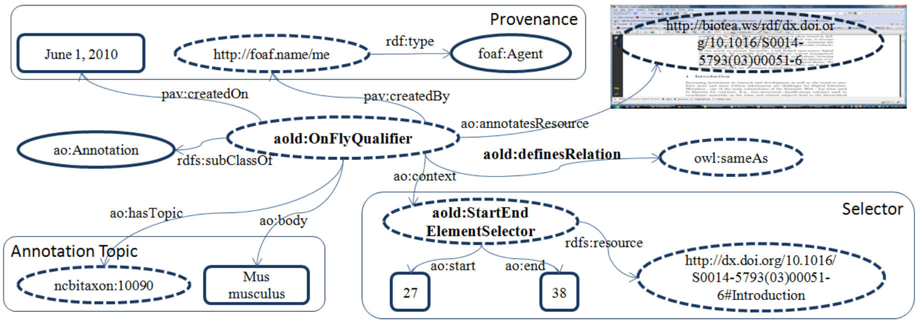


Fig. 4. aold:OnFlyQualifier and aold:StartEndElement selector in use

In order to support the argumentative process hidden in the text, we are reusing the properties proposed by the SWAN [7]; we are also adding new properties, some of them based on [4]. Table 2 summarizes these properties.

Table 2. Classes modeling concepts in a scientific publication

Property name	Description
aold:supportedBy	To specify where the support for the annotated fragment/document can be found (inverse *aold:supports*)
aold:contradicts	When the annotated fragment/document expresses an opposite idea (inverse aold:contradictedBy)
aold:takenFrom	When a fragment has been taken from another text not mentioned as a reference (inverse *aold:takenIn*)
aold:introducedBy	To specify a term, concept, definition introduced by a document
aold:proves	When the annotated fragment/document offers proof (inverse *aold:provedBy*)
aold:rebuts	When the annotated fragment/document offers a rebuttal (inverse *aold:rebutedBy*)
aold:useDataFrom	To specify a data source used in a document (inverse *aold:dataUsedAt*)
aold:cites	similar to bibo:cites but without restrictions on domain and range (inverse *aold:citedBy*, similar to bibo:citedBy)

Figures 5 and 6 present a hypothesis in a document using the definition given in a different document; Fig. 5 uses the ***aold:OnFlyQualifier*** to establish the relation "*cites to*" whereas Fig. 6 uses ***aold:UnidirectionalRelation***. Both figures illustrate how a fragment in the document with DOI 10.1016/SO014-5793(03)00051-6 has been annotated as a hypothesis; this hypothesis cites a fragment in the document PMC1435992 that corresponds to a definition that has been identified as the same concept defined by the entity CHEBI_16113.

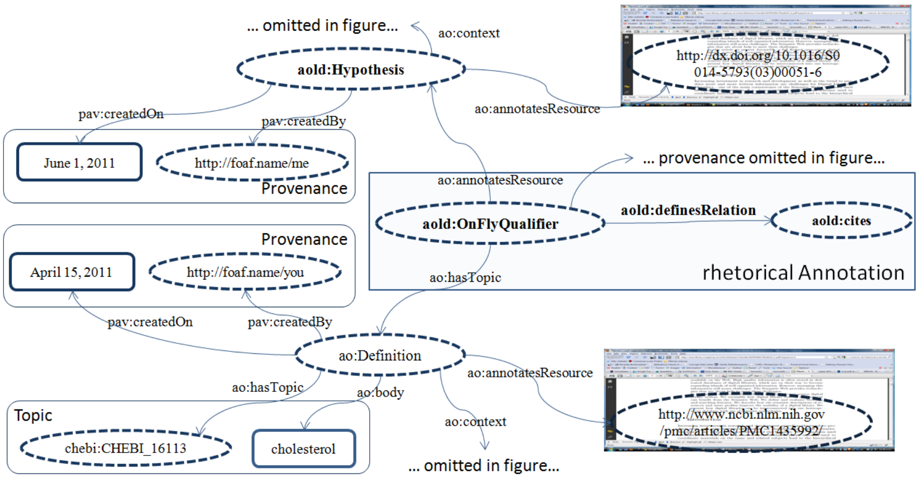


Fig. 5. aold:OnFlyQualifier - relating two documents using a known relation

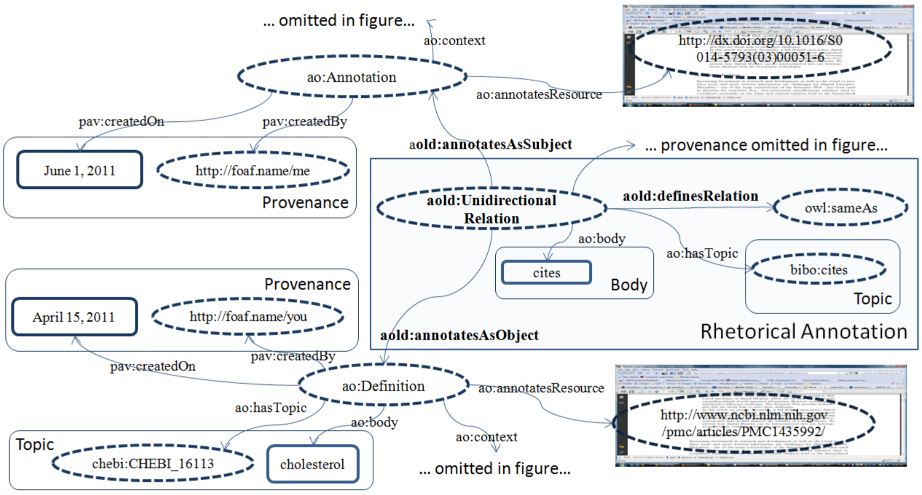


Fig. 6. aold:UnidirectionalRelation - relating two documents proposing a new relation

2.2 Nanopublications with the extended AO

Nanopublications are “a set of annotations that refers to the same statement and contains a minimum set of (community) agreed-upon annotations”; it is therefore feasible to use the proposed extension to represent nanopublications. Consider for instance the definitions for ontology; these vary depending on the field. One that is commonly used comes from Gruber, “*An ontology is a formal specification of a conceptualization*” [10]. The statement comprises three concepts: “an ontology”, “is a”, and “formal specification of a conceptualization”. Upon this statement, different annotations are possible; for instance: (i) Tom Gruber is the author of the statement, (ii) the statement is a definition introduced at http://tomgruber.org/writing/ontolingua-kaj-1993.pdf, (iii) this statement is cited by a particular paper, and (iv) this statement is extended by a particular person. Fig. 7 shows these annotations; provenance of the annotations has been omitted on the sake of readability.

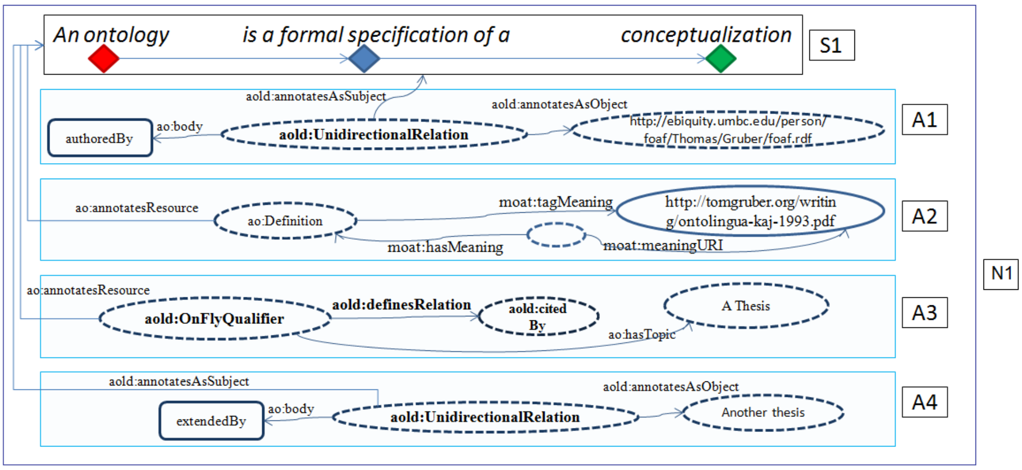


Fig. 7. A nanopublication

3 Discussion

The proposed extension to AO facilitates making explicit the rhetoric and discourse hidden in scientific publications in a way such that that machines can consume data. By using the extensions we can retrieve a list of publications related to a particular term, written by a specific author, or citing a specific gene or protein; for instance:

- All documents cited by document A that contains definitions coined in documents B, C, or D
- All materials used in documents cited by document A
- All documents from 2010 using method A but not method B
- All protocols used when materials A, B, and C have been also used

- All documents including some particular words (or entities) in a specific structural section, *e.g.* aim, thesis, discussion, results, etc.

Scientific publications annotated in the proposed way benefit from the Semantic Web and Link Open Data initiatives. Annotations and information extraction based on these publications become easier, as do sharing them and enriching them with other information also available in RDF. In this way, we facilitate the integration between literature and databases making it easier to use data available in publications for additional analysis. Our ultimate goal is to increase the pace of available scientific data by helping researchers to find information relevant to their projects. As publications are annotated, their rhetoric becomes searchable, thus researchers can better focus on those publications meeting their needs depending on what they are looking for, *e.g.* similar experiments based on methods, materials and protocols, or rebuttals of a particular theory. Furthermore, as annotations can be linked to databases, this enriched content can also be used for more specialized queries.

We have introduced our own description for some existing concepts in CoreSC such as *Experiment* and *Observation,* as we wanted them to be more accurate from the workflow laboratory perspective. We have not used the relations proposed by AZ-II [8] as they use the same category to express a relation and its corresponding inverse. For instance, *Support* is described as “Other work supports current work or is supported by current work”.

Our approach is compatible with the principles of nanopublications. A concept would be a minimal *ao:Annotation* while relations on annotations could be used to define statements that are uniquely identified by a URI. Annotations, as they are understood in nanopublications, are also possible as relations, *i.e.* statement, are resources that can use as the subject of an annotation. The nanopublication itself becomes concrete by using the Annotation Set proposed by AO vr. 2.0; the Annotation Set is a container of annotations that is used to organize annotations that can be referred to as a whole. Furthermore, our approach makes it possible to relate nanopublications to any other type of publication. Our approach is also compatible with other annotation models, such as MOAT and Tag Ontology[9], making it easier to extend and integrate existing applications and tools.

4 Conclusions

We have presented an extension to AO that facilitates modeling rhetoric and discourse in scientific publications. Our approach entails using the common practice of annotating in order to identify specifics within the text; we are also gathering relationships between the annotated and referenced objects. Categories make it easier to identify whether it is about a claim, an example, a report, etc.; some of these categories come from the AO, others are new. In addition to the terms used during the exercise, we also worked with terms such as hypothesis, conclusion and research question; these are useful when analyzing the structure of scientific publications structure. Although not analyzed here, it could also be useful for commercial

[9] http://www.holygoat.co.uk/projects/tags/

documents, since they are also related to each other as well as to external resources. How to implement our model? This is a question beyond the scope of this paper that remains open; it is one of our top priorities.

Acknowledgments. We are grateful to Paolo Ciccarese for his fruitful discussions related to Hypertag migration to AO, *i.e.* using annotations to express relations.

References

1. Clare, A., Croset, S., Grabmüller, C., Kafkas, S., Liakata, M., Oellrich, A., Rebholz-Schuhmann, D.: Exploring the Generation and Integration of Publishable Scientific Facts Using the Concept of Nano-publications. SePublica - Workshop on Semantic Publishing, Vol. 721. CEUR, Heraklion, Crete, Greece (2011)
2. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use 30 (2010) 51-56
3. Garcia, A., Garcia, L.-J., Labarga, A., Giraldo, O., Montana, C., Bateman, J.: The Semantic Web and the Social Web heading towards a Living Document in life sciences. Journal of the Semantic Web. (2009)
4. de Waard, A., Breure, L., Kircz, J.G., Van Oosterndorp, H.: Modeling rhetoric in scientific publication. International Conference on Multidisciplinary Information Sciences and Technologies, Merida, Spain (2006) 1-5
5. 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 (2011) S4
6. Morris, R.A., Dou, L., Hanken, J., Kelly, M., Lowery, D., Ludaescher, B., Macklin, J.A., Morris, P.J.: Semantic Annotation of Mutable Data. To be submitted (2012)
7. Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T.: The SWAN biomedical discourse ontology. Journal of Biomedical Informatics 41 (2008) 739-751
8. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for the conceptualisation and zoning of scientific papers. International Conference on Language Resources and Evaluation, Malta (2010)
9. Passant, A., Laublet, P.: Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. International World Wide Web Conference - Linked Data on the Web Workshop, China (2008)
10. Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2) (1993) 199-220