The SWAN biomedical discourse ontology

The SWAN Scientific Discourse Ontology

Paolo Ciccarese(1,4)*, Elizabeth Wu(2)*, June Kinoshita(2), Gwendolyn T. Wong(2), Marco Ocana(1), Alan Ruttenberg(3), and Tim Clark(1,4)**

1 Massachusetts General Hospital, Boston MA 02129 USA

2 Alzheimer Research Forum Foundation, Boston MA 02109 USA

3 Science Commons, Cambridge MA 02139 USA

4 Harvard Medical School, Boston MA 02115 USA

*These authors made equal contributions to the work.

** Corresponding Author tim_clark@harvard.edu

ABSTRACT

SWAN (Semantic Web Application in Neuromedicine) is a project to construct a semantically-organized, community-curated, distributed knowledge base of Theory, Evidence, and Discussion in biomedicine. Unlike Wikipedia and similar approaches, SWAN's ontology is designed to represent and foreground *both harmonizing and contradictory assertions* within the total community discourse.

Releases of the software, content and ontology will be initially by and for the Alzheimer Disease (AD) research community, with the obvious potential for extension into other disease research areas. The Alzheimer Research Forum, a 4,000-member web community for AD researchers, will host SWAN's initial public release, currently scheduled for late 2007.

This paper presents the current version of SWAN's ontology of scientific discourse and presents our current thinking about its evolution including extensions and alignment with related communities, projects and ontologies.

Keywords

Ontology, semantic web, discourse, biomedicine, SWAN

1. INTRODUCTION

The SWAN project (Semantic Web Applications in Neuromedicine) aims to develop a practical, common, semantically-structured, framework for scientific discourse initially applied, but not limited, to significant problems in Alzheimer Disease (AD) research. SWAN is a collaboration between the Alzheimer Research Forum (Alzforum) and informaticians and clinicians at Harvard Medical School and the Massachusetts General Hospital. The initial concept was proposed in a talk at the W3C Semantic Web in Life Sciences workshop, October 2004 [1]. SWAN has since been developed through a pilot application and is currently in the limited beta release of its first production-quality software and content [2,3,4]. The ability to use SWAN as an integrator of other semantic web ontologies for life science has begun to be shown in several collaborative demonstrator projects [5,6,7] and is an element of current use-case development work in the W3C Health Care and Life Science Task Force [8].

SWAN has built on Alzforum's successful ten-year history as a scientific web community and strong social network [9,10]. The Alzforum web site reports on the latest scientific findings, from basic research to clinical trials; creates and maintains public databases of essential research data and reagents, and produces discussion forums to promote debate, speed the dissemination of new ideas, and break down barriers across the numerous disciplines that can contribute to the global effort to cure Alzheimer's disease. Alzforum currently has over 4,000 registered members, with many members actively contributing to the site by serving as scientific advisors, partnering in creation of databases such as AlzGene, commenting on published papers, and participating in discussion forums.

Alzforum in the past ten years not only has amassed a rich array of scientific contents related to AD, but has also captured vast knowledge from scientists in the field. The SWAN project aims to construct a semantically-structured network of hypotheses, claims, dialogue, publications and digital repositories, incorporating and extending this knowledge. Rather than attempting to construct a logically coherent model of the known facts about AD, SWAN sets itself the goal to model the scientific discourse about AD and its supporting evidence in a rich way that is

compatible with functioning of the current social network as a technology-mediated ecosystem.

SWAN applications currently include an annotator's workbench and a public browser. Both are in limited beta release. The workbench is currently being used by Alzforum annotators to create an initial knowledge base of major hypotheses, claims and evidence in AD research. These applications are discussed more fully in Section 5.

In many formal models of knowledge acquisition in science, research proceeds in a cycle – from hypothesis development; through experiment and data collection; to interpretation and drawing of conclusions; to communication of results to other scientists; to assimilating, criticizing and synthesizing the communications of colleagues. These practice-theory-practice cycles are socially interconnected in an extremely rich and complex way in what has been termed the "knowledge ecosystem" of science. More and more this ecosystem is mediated by the technology of the Web.

Theoretically this "ecosystemic" approach derives from work in industrial knowledge management [11,12] and is also inspired by third-generation activity-theory approaches to human-computer interaction [13]. Practically, it is based on many experiences in constructing information systems to support rapidly-evolving science, in which social factors and the social frame of the system were seen to strongly interact with the technology and content, critically influencing its ultimate success [14]. This approach is naturalistic and materialistic, in that it emphasizes social practice, that is, what scientists actually do, in synthesizing and communicating knowledge of science.

2. SCIENTIFIC DISCOURSE AND TRUTH ON THE WEB

Philosophers of science have defined knowledge as "warranted true belief" [15]. The classical knowledge management (KM) definition of knowledge is "information in context" [16]. The KM definition, while suggestive of how interconnection of information enriches understanding, omits the explicit notions of evidence and truth required for doing science.

For scientific knowledge management systems, the context provided for information will be its "warrant for belief", that is, the evidence of experiment and its theoretical interpretation. Scientific discourse refers to, and validates itself by, a body of theory and experimental practice, which supplies the criterion of truth. What we must know about a context for any scientific assertion is, (a) what warrant-for-belief is provided by the author for the assertion's content, (b) how the content relates logically to other assertions within the body of theory for the domain, and (c) how we may validate this context for ourselves through similar or other experiments. The "ecosystem" of these contextualized assertions is home to a continuous evolutionary process whereby scientific understanding and practice evolve.

Current approaches in providing warrant-for-belief as context are poorly adapted to the reality evolved in the scientific knowledge ecosystem over the past decade - most scientific discourse now takes place mediated by digital artifacts on the Web. The information content of these artifacts is not bound with its context – the forms in which *context* is provided are historically inhomogeneous with the forms of the *content*:

- Scientific information is currently only exchanged digitally as individual documents and data files;
- Knowledge annotation and organization is performed independently by websites and researchers;
- Knowledge schemas are therefore idiosyncratic, incompatible and not easily transferable.

The aim of the SWAN project is to enable a social-technical ecosystem in which semantic context of scientific discourse can be created, stored, accessed, integrated and exchanged along with unstructured or semi-structured digital scientific information. The SWAN ontology is presented here in overview. It is freely accessible on the web [17] and provides a formal basis in OWL [18] for organizing a very rich context for scientific information and discussion. We intend it to evolve to incorporate a large part of the biomedical research life cycle including support for personal data organization, hypothesis generation, and digital pre-publication collaboration. Potentially, community, laboratory, and personal digital resources may all be organized, interconnected and shared using SWAN's common semantic framework. Later this year, we plan to extend this ontology to cover the most common forms of experimental activities and laboratory data.

3. THE SWAN ONTOLOGY AND THE SOCIAL CREATION PROCESS

The SWAN ontology presented here is the knowledge schema for the current limited-beta version of SWAN, with some discussion of changes we plan to adopt prior to full public production release later this year (2007). In preparing for this release, we have not only addressed issues revealed in the beta, but have expanded our effort in

aligning with other existing ontologies. The SWAN applications utilizing this ontology have been in limited but increasing use by scientific annotators, in an iterative development process, since March 2007. While application use has identified various modifications of the initial concept, we have also been in discussion with colleagues working in related areas of biomedical informatics about the ontology. With several of them we have developed approaches to knowledge integration resulting in demonstration prototypes [5,6,7,8], further contributing to refining the ontology. The SWAN ontology creation is a social process.

The SWAN ontologies are freely accessible on the web [17] and provide a formal basis in OWL [18] for organizing a very rich context for scientific information and discussion. We intend it to evolve to incorporate a large part of the biomedical research life cycle including support for personal data organization, hypothesis generation, and digital pre-publication collaboration. Potentially, community, laboratory, and personal digital resources may all be organized, interconnected and shared using SWAN's common semantic framework. We also plan to extend this ontology to cover the most common forms of experimental activities and laboratory data.

4. SWAN CLASSES

4.1 Provenance of data and the root class 'SWANThing'

Every conceptual entity in SWAN is a sub class of SWANThing. SWANThing has been designed to keep track of the provenance of data. Besides the creation date, it records the persons who curate, enter, and author the SWANThing. The curator is a person who performs the process of structuring the knowledge coming from a resource. One or more persons could provide different roles for the same SWANThing. For example, when structuring one of the hypotheses that has been published in Alzforum (http://www.alzforum.org/res/adh/cur/default.asp), the curator can be the same or a different person than the one who authored the hypothesis. The curator can also assign the role of entering data to another person. Initially, a core set of Alzheimer disease hypotheses will be curated in the SWAN knowledge base by a curator other than the original authors of published journal articles. Having a core set of information in the knowledge base will jump-start the community curation process.

4.2 'People', Groups and Organizations in SWAN

In scientific discourse we encounter persons, groups and organizations as participants in the ecosystem. Persons in SWAN are currently defined as an extension of the 'Person' class of the FOAF (Friend Of A Friend) [19] ontology. In the same way, the SWAN ontology defines an extension of the FOAF Organization and Group classes.

When referencing the authors of an external source such as a journal article, most of the time we only have available parts of the person's name, first name or initial letter, last name and sometimes title - basically a set of literals or syntactic objects. However, it is not always possible to uniquely identify the person given this limited set of information. We want to avoid creating an instance of the person class for two reasons. First, we might end up creating more than one instance for the same person if we come across multiple ways that the person's name is represented (e.g. Doe J, Doe JJ, John Doe, etc.). Second we might improperly resolve multiple real persons to the same literal name. Although the prior situation can be managed by the use of owl:sameAs relationships, this puts an additional burden on the implementation.

In our initial implementation we used two classes to represent people, the class KnownPerson and a superclass Person for representing unknown. Properties representing relationships to people had a range of Person, or KnownPerson, taking advantage of the fact that using a range of Person allowed us to also specify a KnownPerson. This was considered desirable as we wished to be able to update such relations if we later identified the person, by creating an instance of KnownPerson, and discarding the Person instance. When we aligned to FOAF, we changed our naming to map to FOAF usage, without changing the general "known/unknown Person" model.

We have decided to modify this aspect of our ontology before full public release of the SWAN knowledgebase. In dealing with documents and related digital resources, obviously the resource may become known to us long before we can uniquely map real Persons to the names in an author list. Our current approach does not adequately support real world operations on documents, names and Persons because it prematurely characterizes mere strings as real Persons (even if "unknown").

To get over this and related problems, and especially to avoid losing monotonicity, we identified the following

approach. We define a class PersonName and a class Person. If somebody named "John Doe" has authored an article, we instantiate a PersonName *for the*

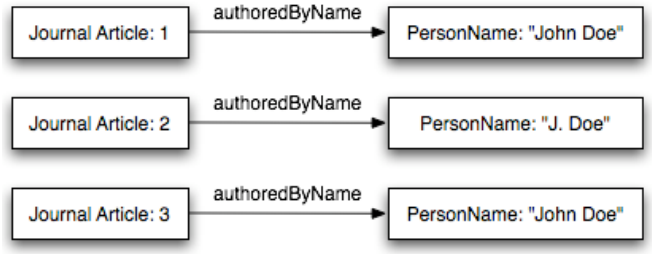

Figure 1 - Three examples of journal articles with the author names.

string "John Doe" – without assuming the string corresponds to any living being. Later on, when the person "John Doe" is uniquely identified, we instantiate a corresponding Person object, and connect the Person instance to one or more instances of PersonName. If we can connect any journal articles written by "John Doe" (the PersonName), to "John Doe" (the Person), we do so. We could define a different instance of PersonName for every name even if literally identical to another one already present.

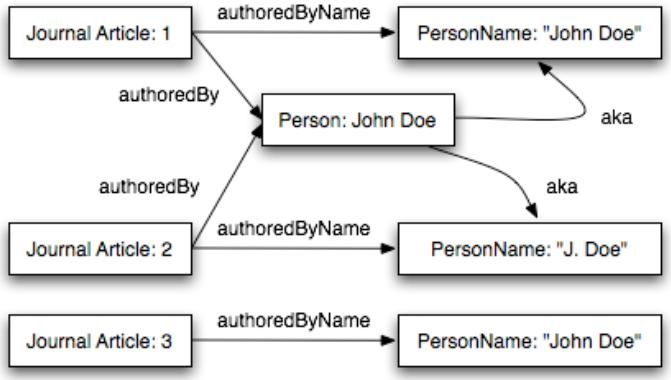

Figure 2 - Three examples of journal articles with the author names and the related person. The person names are redundant.

Or to avoid name proliferation we could instantiate adopting the approach in Figure 3, in which duplicate PersonNames are eliminated.

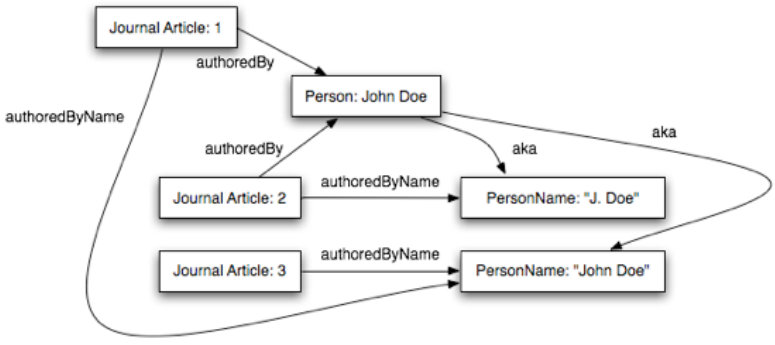

Figure 3 - Three examples of journal articles with the author names and the related person entity.

With this configuration we can keep track of the original format of the authors of a journal article and have the

connection to the actual Person who authored that article. In fact **Figure 3** shows the approach we plan to implement in our next revision to the ontology.

4.3 Digital Resources: Intellectual Products in SWAN

Scientific discourse within SWAN includes references to intellectual products that exist as digital resources outside the SWAN environment. These intellectual products, such as published articles and internet-mediated communications may be noted as evidence for or against Research Statements, and so must be 'citable'. SWAN provides for users to create bibliographic records to identify these resources. We have designed this part of the ontology to represent the citations or bibliographic records of these products and not the products themselves. Records identifying intellectual products are currently modeled in SWAN as the class DigitalResource. These can be records identifying journal articles, published comments, news stories, web pages, simple images or data files. Nowadays the majority of such resources can be found through websites like PubMed or simply through a Google search. In this first iteration of the SWAN ontology we focused on all those resources fundamental for representing the scientific discourse mediated by the Alzforum website.

The SWAN-managed Digital Resources are:

- From journals: articles, news, comments, and images
- From magazines: articles, news, comments, and images
- On web pages: articles, news, comments, and images

All these classes are bibliographic representations of the original sources and do not include the content which is available external to SWAN. Each class defines a set of attributes and relationships useful to uniquely define the resource and to provide a sufficiently useful set of information to the users who deal with it.

The public access version of SWAN only includes information not covered by copyrights. Thus, the abstract of the articles and the full text are not included. On the other hand, directly, and through annotation to be introduced later on, we collect a variety of attributes useful for improving search and data mining capabilities.

Although copyright may prevent inclusion of copies of resources our approach is, in any case, to duplicate the least amount of information possible. We represent information that is required to guarantee the ability to implement necessary functionality and to enable proper data integration. Thus, for a web page the content will not be duplicated in SWAN – it is outside the scope of SWAN to manage the risk of losing the resource if the web page is not maintained over time.

In the current version of the ontology we are still not considering such resources as manuscripts-in-process because these would, typically, not be published in journals, magazines, or on web pages. We plan to allow for management of such resources in the future. In the next iteration of our ontology a manuscript could represent, for instance, an idea that is under development for a journal article. Such an artifact is different from the resources currently referred to by SWAN, as it is not the result of a publishing process, but rather an embryonic form of something that might eventually be published. A manuscript-in-progress will, typically, be an entity managed in the private space of the user. In such cases we would provide a means to include abstract and a full text, as a way of having the author make it public if it cannot be found in a digital format elsewhere. Other Digital Resources that will be integrated in the future versions of the SWAN ontology will identify files of data, images, and database entries from the user's personal workspace.

4.4 'DiscourseElements': the Core of SWAN

DiscourseElement classes represent the core of the SWAN Ontology. Through such classes it is possible to use self-annotated discourse as a bridge connecting the many specialized research sub-domains contributing to AD research and to research in general. One is not required to make "value judgments" about the entities referred to or the propositions that the bridging statements make. The bridging level consists of communicative acts that we record, representing what is said as well as its evidential status and where possible, its relationship as argumentation relative to other statements. The classes in SWAN are, in Hausser's terminology, a "+constructive" ontology [20], that is, an ontology about what is said, rather than about agreed-upon objective facts. The DiscourseElements in the ontology

characterize digital resources which themselves contain statements expressed in natural language. Each DiscourseElement may also be linked to terms or statements in other domain ontologies and folksonomies, which classify or describe it in terms of relatively undisputed facts or objective categories (in Hausser's framework, "-constructive" ontologies). SWAN thus captures a middle, transitional ground between the more inventive, nuanced, contentious, and inherently ambiguous flow of natural language – in which scientific discourse is conducted – and the far more controlled, formal, unambiguous, rigorous, and fixed nature of formal ontologies "about" the science. In SWAN's ontology the connecting point that links to externally defined entities in biology is the LifeScienceEntity (LSE), which functions as a kind of "adapter".

The SWAN DiscourseElements 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

DiscourseElements only attributes, aside from those inherited from SWANThing, are title and description, but the elements participate in a variety of important relationships. In order to give an idea of the relations involving discourse elements, consider the following example of creating a research statement, specifically a hypothesis.

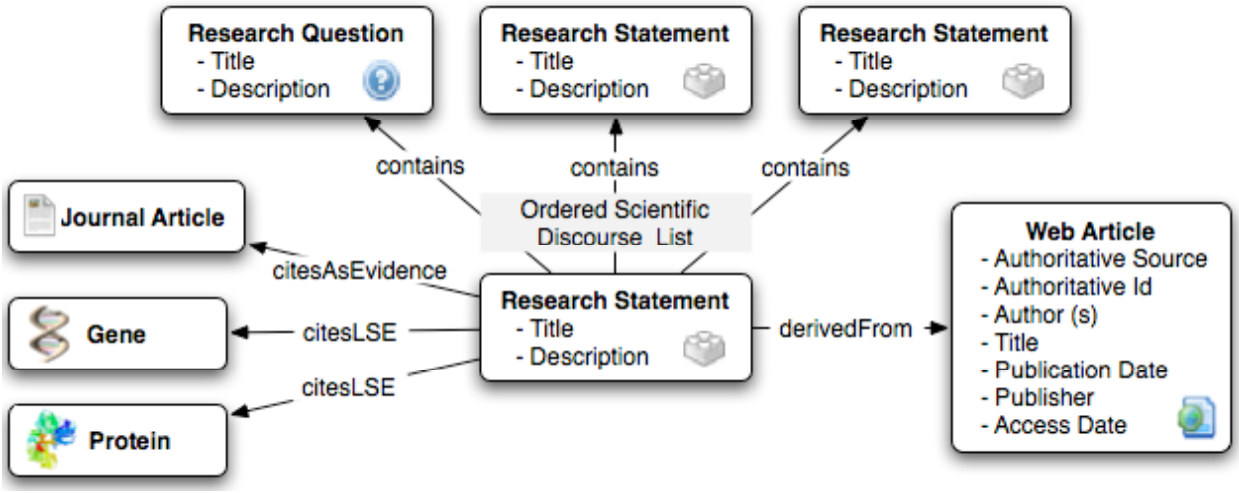

Figure 4 – Some example relationships of a DiscourseElement.

In **Figure 4** we depict a possible instantiation showing some relationships between the primary research statement and other SWAN entities, as well as attributes of each type.

In this case, the research statement (an hypothesis) is "derivedFrom" a Web Article. The relationship "derivedFrom" is used to assert that the research statement is mirroring a digital resource, in this case, an article published on the web. This distinguishes a derived resource from one created, *de novo*, by an author in the SWAN environment.

The "content" of a ResearchStatement is then composed of an ordered list of other DiscourseElements This whole-part relationship is named "contains". The proper order of the contained entities establishes the logical flow of discourse expressed by the resource. At the same time it is possible that the article cites as evidence other digital resources ("citesAsEvidence") or LifeScienceEntities and Reagents through "citesLifeScienceEntity" or "citesReagent", respectively.

After the original hypothesis has been detailed using nested DiscourseElements in the proper order, it is possible to relate each DiscourseElement to others. This is done with the set of relationships "discusses", "refutes", "supports" and "alternativeTo". The contained entities can be newly defined or reused if already present.

Therefore, it is possible to have three cases:

1. A research statement not derived from any resource. This is shown in **Figure 5** by the upper rightmost statement, which is evolved from a pre-existing one in the knowledgebase. The research statement can be detailed in a title and description and it is possible to relate it to other DiscourseElements through the already mentioned relationships. In particular the relationship "alternativeTo" is used to refer the new research statement to already existing ones. In this case the research statement's provenance will be defined by the curator - and could correspond to the original author or to a knowledge base editor.

2. Full reuse of an existing research statement. In this case it is possible to create a research statement (e.g. a hypothesis) to include another existing research statement (e.g a claim). The provenance of the existing claim is maintained. But the new connection between the hypothesis and the claim will have a different curator.

3. Partial reuse of an existing research statement. It is possible to partially reuse another research statement through the relationship "evolvedFrom" in which we connect a newer version of a research statement to a previously existing one, upon which it was based.

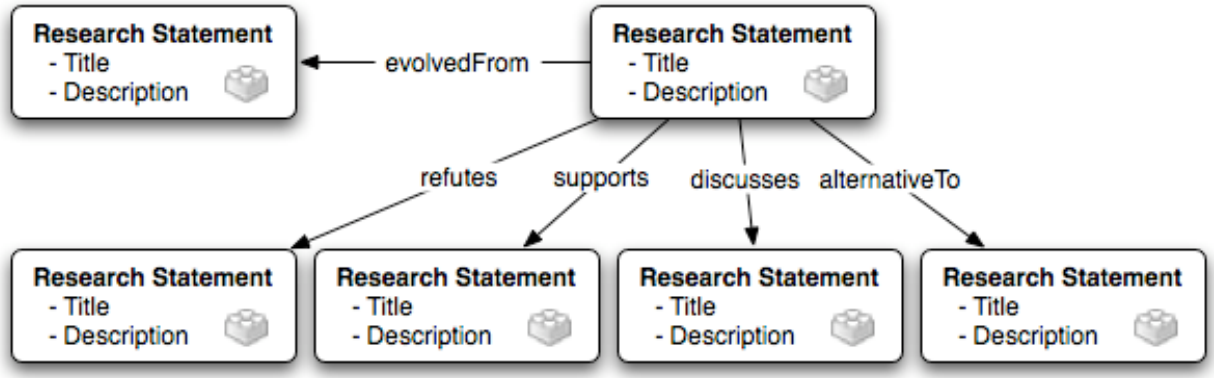

Figure 5 – Examples of logical relationships among DiscourseElements

Other interesting use cases come from the practice of commenting, which in SWAN are captured as Comment entities. A Comment upon a ResearchStatement can mirror a comment that has been published in a journal or on a web site, or can be written by the user. In the first case the relationship *derivedFrom* is applied to connect the Comment and the original digital resource. In the second case the comment is defined within the SWAN workbench directly. The comment is always *inResponseTo* some other DiscourseElement (because we are modeling dialogue) and such relationship can be characterized more fully through a *supports*, *discusses*, or *refutes* relationship.

A Comment has a form similar, in certain respects, to a ResearchStatement. It can be composed of an ordered list of DiscourseElements, it can refute, support or discuss other DiscourseElements and it can be alternative to some other DiscourseElement. It can cite life science entities or reagents as well as digital resources. As with all the other

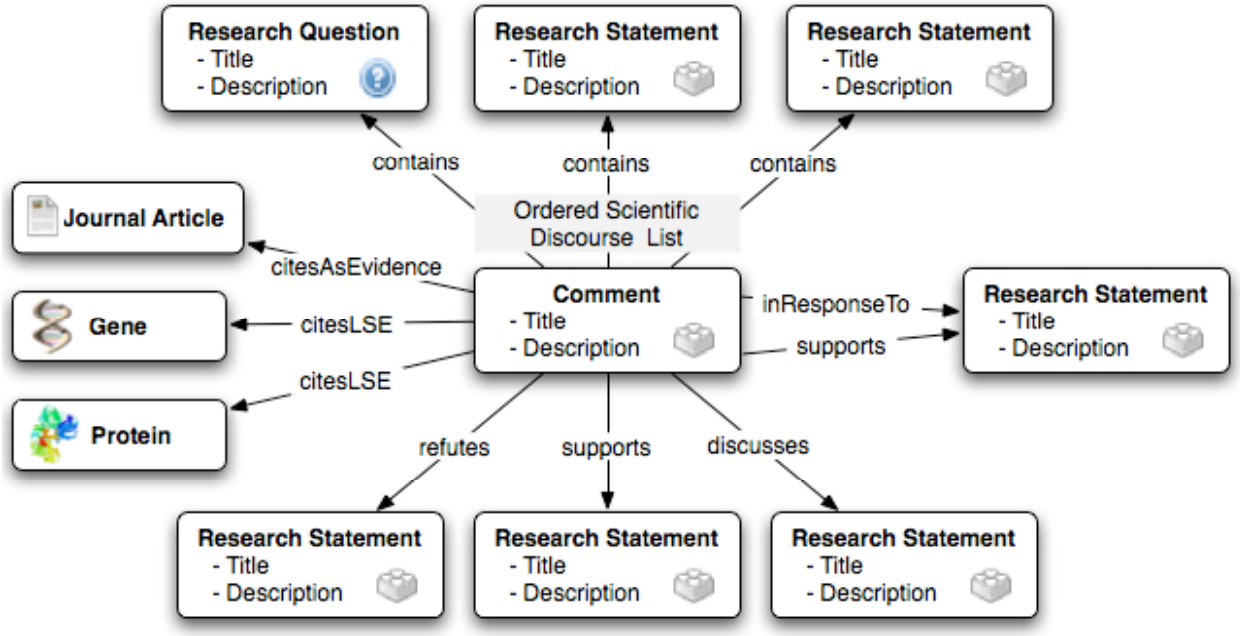

Figure 6 – Some possible relationships between Comment and other classes.

DiscourseElements, Comments can present an ordered list of authors. An example showing relationships which Comments can participate in are shown in Figure 6.

The last DiscourseElement in this version of SWAN is the ResearchQuestion. A ResearchQuestion can be contained in another DiscourseElement or it can be "motivatedBy" another DiscourseElement. ResearchQuestions are open topics of investigation where dialogue is initiated and, based on them, experiments performed.

4.5 Life Science Entities (LSEs) and Reagents

An important element of scientific discourse in Alzheimer Disease (AD) research, and in many other biomedical contexts, is inclusion of scientific statements that refer to entities defined by external resources. Examples of such external resources include the Gene Ontology (http://www.geneontology.org/), Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene), and the Alzforum antibody database (http://www.alzforum.org/res/com/ant/default.asp).

Using the same approach that informs our treatment of digital resources, we represent only enough information about such external entities to enable simple search, to present the entity in a way a user can recognize, and to provide a way to link to the external information about the resource. Analogous to Digital Resources, an instance of a Life Science Entity is created for each referred-to external resources. For genes, links are provided to web pages in Entrez Gene or to the HUGO Nomenclature Committee Database [21]. We anticipate that in the near future we will link antibodies to an RDF translation of the Alzforum antibody database.

The current version of SWAN provides the ability to refer to the following types of entities, which were priorities of our target community.

- Life Science Entity (LSE)
  - Gene
  - Protein
- Reagent
  - Antibody
  - Transgenic Model

4.6 Tags

SWAN includes support for "tagging" entities, in a manner popularized by sites such as social collaboration sites such as Flickr and del.icio.us. One is able to create user-defined "custom" tags by providing a textual label as the sole attribute, or to collect a set of tags as a tag "class". In this way, one can use and distinguish identifiers from external databases or taxonomies as tags. For example, MeSH (http://www.nlm.nih.gov/mesh/) terms can serve as tags by being members of the class of MeSH tags, and then having their textual labels be the names or identifiers from MeSH.

The use of MeSH for tags is an interesting case. Together with the bibliographic record, PubMed provides keywords encoded using the MeSH terminology. When SWAN imports reference information from Pubmed, an instance of Comment, authored by the PubMed organization, is created and related to Digital Resource that refers to the article. We consider this to be a comment in the sense that it is an elaboration of the original work. The Comment instance is tagged with the MeSH terms that have been associated with the Pubmed record.

4.7 Qualifiers

Qualifiers are predefined tags that can be assigned to specific types of entities. Currently we only allow qualification of the research statements. Research statements can be tagged as *Claim* or *Hypothesis*, with a Pathogenic Narrative* term describing the AD disease process, or by an evidence type that describes the sort of scientific evidence supporting the statement. A curator does initial assignment of qualifiers, however users may subsequently tag or qualify these entities in order to organize knowledge according to their own assessment.

4.8 Versioning and Evolution

The SWAN ontology has been designed to support the scientific knowledge life cycle. This includes evolution of research statements as the underlying knowledge evolves, or as new researchers start work inspired by the ideas of others. Each time a researcher edits a research statement a new version is created. Users may also edit statements created by others, in which case a new research statement is created and a new relation, *evolvedFrom,* is established between the new statement and the original. Previous versions of research statements remain available for review.

5. APPLYING THE ONTOLOGY IN PRACTICE

The SWAN ontology is currently used in two software applications: the SWAN Workbench and the SWAN Browser.

The Workbench allows scientists and scientific editors to work with the ontology in a very friendly, AJAX-style, Web application. Instances created in the course of using the Workbench are written to a triple store - currently IBM's BOCA triple store [22]. The SWAN Workbench is designed to organize Hypotheses, Claims, and Evidence within a discourse; and to highlight relationships between them. The current version of the Workbench is designed to be used by "privileged" users who have curatorial responsibilities. Future work will be aimed at making the tool useful for a wider group of researchers. To this end, we will enhance the Workbench by providing private workspaces, and by adding a publication state model to support tracking and annotation of pre-publication papers.

A second Web application, the SWAN Browser, accesses and presents elements from the triple store. The Browser is designed to be used by ordinary scientists and supports access, manipulation and downloading of information from the SWAN knowledge base. We have adapted the Simile Exhibit faceted browser (http://simile.mit.edu/exhibit/) in this application to enable selection of sets of research statements and their evidence. The SWAN Browser supports navigation through the network of relationships stored in the knowledge base and provides tools that, for instance, enable the discovery of conflicting claims.

Both the Workbench and the Browser are implemented using the following technology: Java (http://java.sun.com/)

* The SWAN Pathogenic Narrative is an ordered list consisting of the following terms, applied as qualifiers to SWAN Claims: "Initial Condition", "Perturbation", "Pathogenic Event", "Pathologic Change". These terms describe general stages in the pathogenesis of neurodegenerative disorders.

is the programming language, and AspectJ (http://www.eclipse.org/aspectj/), Struts2 (http://struts.apache.org/2.x/), Spring (http://www.springframework.org/), and Hibernate (http://www.hibernate.org) form the object/relational persistency framework. DB2 (http://www-306.ibm.com/software/data/db2/9/) is the relational database upon which Boca implements the RDF (http://www.w3.org/RDF/) triple store. SPARQL (http://www.w3.org/TR/rdf-sparql-query/) is used to query RDF, and Jastor (http://jastor.sourceforge.net/) for generating Java Beans from OWL (http://www.w3.org/TR/owl-features/) ontologies. Dojo (http://dojotoolkit.org/) is used as a general Javascript framework and Exhibit of the Simile project (http://simile.mit.edu/exhibit/) is used for presentation within the Browser.

In **Figure 7** below we illustrate a small section of an example in which the SWAN ontology is used to structure scientific discourse where there is substantial uncertainty and conflict over the correctness of two competing models of AD pathology. Biologists and science curators on our team worked through details of numerous such examples in parallel with development of the ontology and the software in order to ensure that the software can support realistic use by scientists in their daily work. As an element of implementing the SWAN project, we are in the process of annotating in depth several dozen current hypotheses in AD research, which will be provided as an initial content store to our user community via the Alzforum. These annotations will be checked for correctness by scientific staff of the Massachusetts Alzheimer Disease Research Center, Massachusetts General Hospital (http://www.massgeneral.org/neurology/MADRC).

The example in **Figure 7** shows a small section of the discourse around two current hypotheses of AD etiology, one originating from Vincent Marchesi at the Yale Medical School (Marchesi, Intracellular A-Beta Dimer Hypothesis) [23] and the other from the group of Karen Hsiao Ashe at the University of Minnesota (Lesné et al, A-Beta*56 Hypothesis) [24]. Both scientists attempt to develop models that explain the thousands of research observations tying amyloid-beta protein to Alzheimer pathology. Among the questions at hand are (a) is amyloid-beta, or one of its derivatives or precursors, the toxic agent in AD? (b) if so, what is the mechanism of toxicity?

Note that Hypotheses are modeled as a nested set of Research Statements. Both Hypotheses and Claims are Research Statements; they are intended to be re-usable outside their original context. An Hypothesis in one context may be re-used as a Claim in another, broader context, and vice versa. There is no inherent limit to the nesting of Research Statements.

Research Statements are not regarded as valid in and of themselves. That is for the scientific community to determine. But clearly the authors of such statements intend them to be accepted. In modeling the discourse, therefore, we show the specific evidence cited by the authors in support of each claim, and in some cases the overall model or hypothesis. In the context our example illustrated in the Figure, the evidence cited is in the form of other publications but in the future SWAN will allow citations to supplemental data as well, including data on websites.

The example shows only a portion of the Marchesi and Lesné hypotheses. In our current content library, Marchesi hypothesis consists of 26 Claims. Claim 9 of Marchesi conflicts with Claim 3 of Lesné, as shown in the metadata by a symmetric "refutes" relationship between the research statements (shown in red). Use of the term "refutes" is not meant to imply objective refutation. We are simply modeling the conflict between these statements, one of which states that A-Beta exerts toxicity intra-membranously, while the other claims an extra-membranous mechanism of toxicity.

By modeling the specific claims made by various models of AD pathogenesis and their argumentation relationships to one another, we hope to provide scientists in this highly multidisciplinary field with a tool for reasoning about the knowledge in their field, for thinking about what experiments need to be done to resolve conflicts and contradictions, and provide a framework for making serendipitous discoveries of research previously unknown to them.

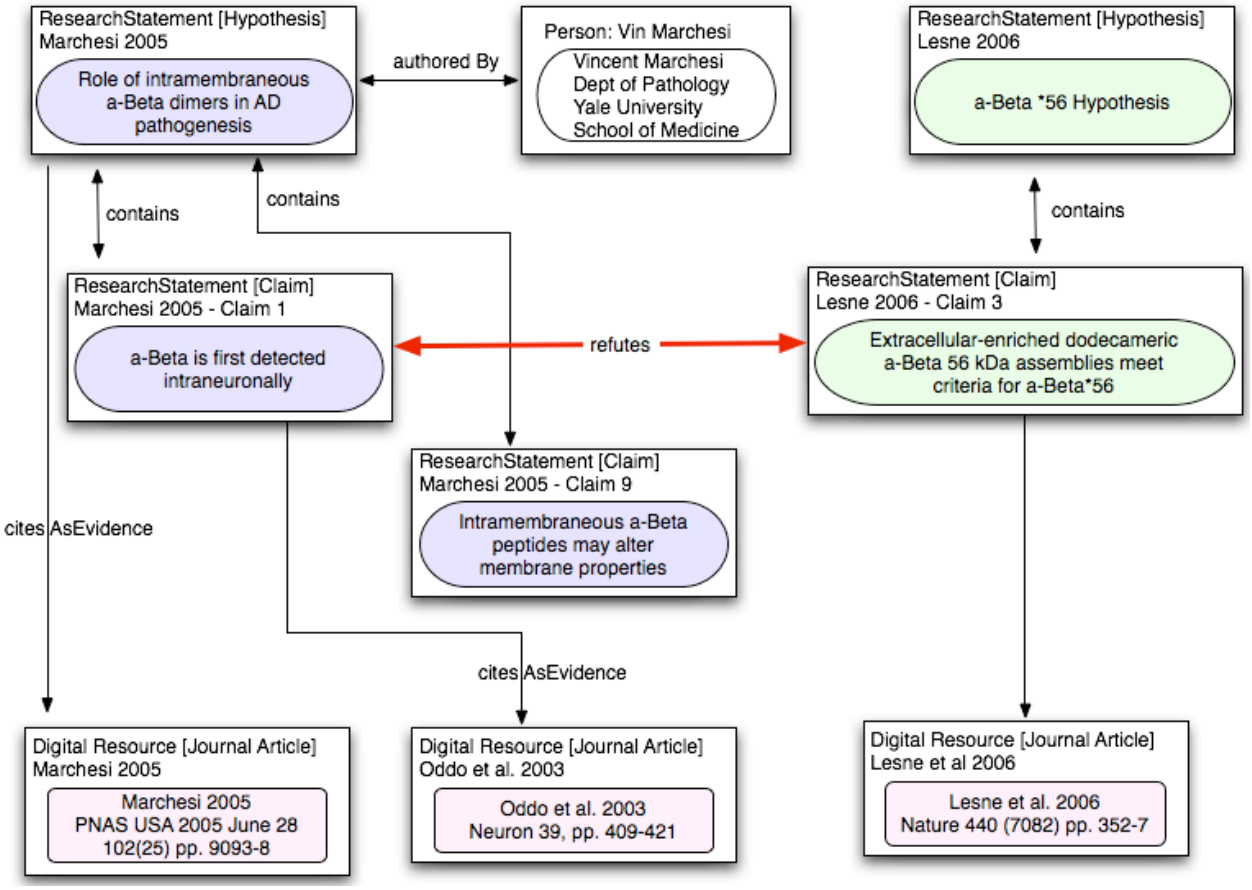

Figure 7 – Example of research statement instances representing conflicting hypotheses, claims and evidence. Note, in particular, the use of the refutation relation.

6. SWAN'S RELATIONSHIP TO OTHER EFFORTS IN THE BIOMEDICAL AND BIO-ONTOLOGY COMMUNITIES

6.1 Collaborating Groups and Centers

From the outset we have attempted to achieve as broad a set of collaborations and "friendly conversations" as possible between the SWAN project team, working AD researchers, bio-ontologists, and web technologists. We are currently collaborating with the Massachusetts Alzheimer Disease Research Center (http://www.madrc.org/) for quality control of hypothesis content; the W3C Health Care and Life Sciences Task Force (http://www.w3.org/2001/sw/hcls/) for development of AD and Parkinson's Disease research-based use cases and an interoperability demonstration; Science Commons (http://sciencecommons.org/); the SenseLab Group (http://senselab.med.yale.edu/) at Yale School of Medicine's Department of Medical Informatics; and the Open Biomedical Ontologies (http://obofoundry.org) effort. A number of other collaborations with groups developing ontologies of reagents, animal models, biological pathways, and so forth, are under active discussion. The public beta release of SWAN's knowledge management tool will be hosted on the Alzheimer Research Forum website (http://www.alzforum.org) beginning in late 2007.

6.2 Directions in Ontology Alignment

SWAN as a software artifact, as an ontology, and as an element of the Semantic Web, is itself embedded in a community ecosystem devoted to representing and sharing knowledge in a more broad sense. We have begun to

investigate alignments with related work in other ontologies. Here we discuss some efforts relevant to SWAN from a preliminary survey. As we identify and define alignment use cases we will work to establish correspondence to these and other ontologies relevant to our domain.

The Open Biomedical Ontology (OBO) Foundry [25] is comprised of a subset of the ontologies [26] within a larger set of the OBO ontologies, including among them the Gene Ontology. Creators of OBO Foundry ontologies have committed to a (evolving) set of principles [27]. These principles are designed to encourage the development of a set of interoperable, logically well-formed, non-overlapping ontologies that accurately represent their domains. They use a shared set of relations, are collaboratively developed, and are based on a shared Basic Formal Ontology (BFO) [28].

Like SWAN, OBO's Ontology for Biomedical Investigations (OBI) [29] extends past the usual boundaries of the natural science ontologies. OBI does this because of the necessity of describing items which have been created for the purpose of doing experiments, such as instruments, and reagents; cognitive constructs, such as plans and protocols; as well as documentation associated with research, such as proposals, forms, reports, and data. In these last elements we find some direct overlap with some of the subject matter of SWAN.

The International Federation of Library Associations and Institutions convened a working group to develop requirements for bibliographic records. Their report [30] develops a theory of the central entities and relations that are relevant for representing bibliographic records and, as such, their concerns are relevant to an important function of SWAN –pointing to scientific literature.

Finally the Semantically Annotated Latex (SALT) project [31] concerns itself with the general problem of referring to and annotating semantic aspects of documents. One of the theories on which they base a portion of their ontology is Rhetorical Structure Theory (RST) [32]. RST theory provides analytic tools for analyzing how pieces of text in a single document relate to one another and what functions these relations accomplish in the process of communication between writer and reader. Although RST addresses relations among text in a single document, it may also be useful in providing a framework for intra-document relations of interest to SWAN.

7. CONCLUSION

The SWAN Ontology is a knowledge schema for personal and community organization and annotation of scientific discourse. Working bench scientists using the SWAN application will be able to organize key knowledge in their own specialties as a web of assertions whose relationships to one another and to their supporting evidence is well-characterized.

These assertions will be organized as metadata on the most commonly used digital resources representing unstructured scientific discussion, such as PDFs and web pages. They will be an important bridge between the scientific literature and concepts in several biomedical ontologies, and will be able to be published and shared in scientific web communities with relatively lightweight intervention by curators or editors.

SWAN is, by design, a mediating technology for working social networks of scientists. The authors believe it will enable a new level of knowledge organization to be created and shared by scientists themselves, as an integral part of their work activity.

8. ACKNOWLEDGMENTS

We are grateful to the Ellison Medical Foundation and to an anonymous foundation, for their generous support of the SWAN project.

Profound thanks to Brad Hyman (Harvard Medical School, Massachusetts General Hospital and Massachusetts Alzheimer Disease Research Center); Sean Martin, Ben Szekley and Lee Feigenbaum (formerly IBM Advanced Internet Technology Group); Dean Hartley (Rush-Presbyterian Hospital, Chicago); Carole Goble (University of Manchester); Kei Cheung (Yale Medical School); and to Barry Smith (University of Buffalo), for many valuable discussions.

We would also like to express our gratitude to Anne Young, Chief of Neurology at the Massachusetts General

Hospital (MGH), for her continuing support; and to Yong Gao (MGH) for his work in developing the initial proof-of-concept for SWAN.

9. REFERENCES

[1] Clark T, Kinoshita J. A pilot KB of biological pathways important in Alzheimer's Disease. W3C Workshop on Semantic Web for Life Sciences, Cambridge, MA, USA, October 2004.

[2] Gao Y, Kinoshita J, Wu E, Miller E, Lee R, Seaborne A, Cayzer S, Clark T. SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research. J Web Semantics 2006; 4(3).

[3] Wong GT, Gao Y, Wu E, Ciccarese P, Ocana M, Kinoshita J, Clark T. Developing SWAN, a shared knowledge base for Alzheimer's disease research. Abstracts, Society for Neuroscience 2006, Atlanta, GA.

[4] Kinoshita J, Strobel G. Alzheimer Research Forum: A Knowledge Base and e-Community for AD Research. In: Alzheimer: 100 Years and Beyond, Research and Perspectives in Alzheimer's Disease. Jucker M, Beyreuther K, Haass C, Nitsch R, Christen Y, editors. Berlin, Heidelberg, New York:Springer; 2006, p. 457-63.

[5] Lam YK, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, Crasto C, Morse T, Stephens S, Cheung KH. Semantic Web Meets e-Neuroscience: An RDF Use Case. Proceedings of International Workshop on Semantic e-Science, ASWC 2006. Beijing, China: Jilin University Press; 2006; p. 158-70.

[6] Cheung KH, Lam YK, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, Crasto C, Morse T, Stephens S. AlzPharm: A Light-Weight RDF Warehouse for Integrating Neurodegenerative Data. ISWC 2006, Athens, Georgia.

[7] Lam YK, Marenco L, Clark T, Gao Y, Kinoshita J, Shepherd G, Miller P, Wu E, Wong G, Liu N, Crasto C, Morse T, Stephens S, Cheung KH. Semantic Web Meets e-Neuroscience. BMC BioInformatics 2007, (In press).

[8] Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH. Advancing translational research with the Semantic Web. BMC Bioinformatics, 2007, 8(Suppl 3)**:**S2.

[9] Kinoshita J, Clark T. Alzforum: Towards an e-Science for Alzheimer Disease. In: Crasto C, editor. Neuroinformatics. Humana Press (in press).

[10] Clark T, Kinoshita J. Alzforum and SWAN: The Present and Future of Scientific Web Communities. Briefings in Bioinfomatics (in press).

[11] Davenport T, Prusak L. Information Ecology: Mastering the Information and Knowledge Environment. Oxford University Press, 1997.

[12] Brown JS, Duguid P. The Social Life of Information. Cambridge: Harvard Business Review, 2002.

[13] Nardi BA. Activity Theory and Human-Computer Interaction, IN: Nardi, B. editor, Context and Consciousness: Activity Theory and Human-Computer Interaction. Cambridge: MIT Press, 1996.

[14] Ficenec D, Osborne M, Pradines J, Richards D, Felciano R, Cho R, Chen R, Liefeld T, Owen JJ, Ruttenberg A, Reich C, Horvath J, Clark T. Computational Knowledge Integration in Biopharmaceutical Research. Briefings in Bioinformatics; 2003 4(3):260-78.

[15] Klein PD. Concept of Knowledge IN: Craig E, editor, The Routledge Shorter Encyclopedia of Philosophy. Abingdon, Oxfordshire, UK: Routledge; 2005, p. 525.

[16] Davenport T, Prusak L. Working Knowledge. Boston:Harvard Business School Press; 1984.

[17] Ciccarese P, Wu E, Kinoshita J, Wong G, Ocana M, Clark T. SWAN 1.0 Ontology. 2007 [http://purl.org/swan/1.0/]

[18] OWL Web Ontology Language. Smith M, Welty C, McGuiness D, editors. W3C; 2004. [http://www.w3.org/TR/owl-guide/]

[19] The Friend of a Friend (FOAF) project. [http://www.foaf-project.org/]

[20] Hausser R. The Four Basic Ontologies of Semantic Interpretation. Tenth European-Japanese Conference on Information Modeling and Knowledge Bases, Saariselkä, Finland, IOS Press, Amsterdam, The Netherlands; 2000.

[21] Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford E, Lush MJ. The HUGO Gene Nomenclature Database, 2006 Updates. Nucleic Acids Res. 2006 January 1; 34(Database issue): D319–D321. Published online 2005 December 28. doi: 10.1093/nar/gkj147.

[22] Feigenbaum L, Martin S, Roy MN, Szekely B, Yung WC. Briefings in Bioinformatics 2007 8(3):195-200, doi:10.1093/bib/bbm017

[23] Marchesi V. An alternative interpretation of the amyloid Aβ hypothesisn with regard to the pathogenesis of Alzheimer's disease. Proc Natl Acad Sci U S A. 2005 Jun 28; 102(26):9093-8.

[24] Lesné S, Koh MT, Kotilinek L, Kayed R, Glabe CG, Yang A, Gallagher M, Ashe KH. A specific amyloid-beta protein assembly in the brain impairs memory. Nature 2006 Mar 16; 440(7082):352-7.

[25] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Consortium TO, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Shah N, Whetzel PL, and Lewis S, The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. In Review, 2007.

[26] The OBO Foundry Ontologies. 2007 [cited; Available from: http://obofoundry.org.]

[27] OBO Foundry Principles. 2006 [cited; Available from: http://www.obofoundry.org/crit.shtml]

[28] Grenon P, Smith B, Goldberg L, Biodynamic ontology: applying BFO in the biomedical domain. Studies in health technology and informatics, 2004. 102: 20-38.

[29] OBI Sourceforge Site. 2007 [cited; Available from: http://sourceforge.net/projects/obi/.]

[30] IFLA Study Group on the Functional Requirements for Bibliographic Records. and International Federation of Library Associations and Institutions. Section on Cataloguing. Standing Committee., Functional requirements for bibliographic records : final report. UBCIM publications. 1998, München: K.G. Saur. viii, 136 p.

[31] Groza T, Handschuh S, Möller K, and Decker S. SALT: Semantically Annotated LaTeX for Scientific Publications. in 4th European Semantic Web Conference, 2007. Innsbruck, Austria.

[32] Mann WC and Thompson SA, Rhetorical structure theory: a theory of text organization. ISI Reprint Series; ISI/RS-87-190. 1987, Marina del Rey, Ca: Information Sciences Institute. 82 p.