An Overview of the SWAN 1.0 Discourse Ontology

Paolo Ciccarese
Harvard Medical School
& Mass General Hospital
Charlestown, MA 02129
+1 857-272-6627
paolo.ciccarese@gmail.com

Elizabeth Wu
Alzheimer Research Forum
82 Devonshire St
Boston, MA 02109
+1 617-928-3447
ewu1@comcast.net

Tim Clark
Initiative in Innovative Computing
Harvard University
Cambridge, MA 02138
+1 617-947-7098
tim_clark@harvard.edu

ABSTRACT
In this paper, we provide an overview of the SWAN 1.0 ontology for scientific discourse.

Categories and Subject Descriptors
H.2.8 [**Information Systems**]: Database management: Database applications: scientific databases
H.3.4 [**Information Systems**]: Information storage and retrieval: Systems and software: World Wide Web (WWW)
J.3 [**Computer applications**]: Life and medical sciences

General Terms
Standardization, Languages

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. The SWAN project is the result of a collaboration between the Alzheimer Research Forum (Alzforum) and informaticians at Harvard University, Massachusetts General Hospital and IBM. The initial concept has been 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 development stage of its first production-quality application [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].

The SWAN project has built on Alzforum's successful ten-year history as a scientific web community and strong social network [9,10] (currently with over 4,000 registered members) to construct a semantically-structured network of hypotheses, claims, dialogue, publications and digital repositories. 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.

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.

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 such as [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 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 definition of knowledge is, "information in context" [16] – a constructivist answer. For scientific knowledge management systems, the context is the warrant, while discourse and experiment supply the criterion of truth. What we must know about scientific assertions is, what warrant (context) is provided by the author, and how can we validate (replicate) this context for ourselves through experiment, in a continuous evolutionary process.

Current practices in providing warrant are poorly adapted to the reality evolved over the past decade – that most scientific discourse now takes place mediated by digital artifacts accessed on the Web. This is because information content is not transferred 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 1.0 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. SWAN CLASSES

3.1 The Root Class 'SWANThing'

Every conceptual entity in SWAN is a sub class of SWANThing. SWANThing defines the provenance of data. Besides the creation date, it records the curators of the entered knowledge and the persons who entered it. The curator performs the process of structuring the knowledge coming from a resource, for instance from a hypothesis published in the Alzforum website or in a journal article, suitable for encoding in the SWAN conceptual framework. The curator can be the same person who authored the hypothesis, or other users working independently.

3.2 SWAN 'DigitalResource'

In SWAN, digital resources represent (typically unstructured) resources outside the SWAN environment. These can be journal articles, published comments, news, a web page about a gene, as well as 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 that are fundamental for representation of scientific discourse mediated by the Alzforum website. The managed digital resources are:

- Journal articles/news/comments/images
- Newspaper articles/news/images
- Web pages/articles/news/comments/images

All these classes are representative of the original sources. Every element contains a set of attributes and relationships useful to uniquely define the resource and to give a sufficient set of information to the users who deal with it.

SWAN represents (for public access) only 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 many attributes useful for improving search and data mining capabilities. Aside from copyright issues, the general idea is to duplicate the least possible information required to guarantee the necessary functionality, and to enable proper data integration when needed. Thus, for a web page the content will not be duplicated in SWAN – we will endure the risk of losing the resource if the web page is not maintained over time.

WWW 2007, May 8--12, 2007, Banff, Canada.

In the current version of the ontology we are still not considering resources as manuscripts-in-process. That functionality will be integrated later on. As we envision the next iteration of this ontology, a manuscript could represent, for instance, an idea that is under development for a journal article. It would represent an outlier in front of the other resources, as it is not the result of a publishing process, but an embryonic form of a publication. It will typically be an entity belonging to the private space of the user and may also contain an abstract and a full text, in case it is necessary to make it public, as it cannot be publicly found in digital format. Other digital resources that will probably be integrated in the future versions of the SWAN ontology include files of data, images, and database entries from the user's personal workspace.

3.3 'People' in SWAN

When referencing the listed authors of an external source like a journal article, most of the time, we have only a text giving us the whole name and often the person's title all in one string (as is the case with PubMed). Therefore, in SWAN we create an instance of a Person class. This class is characterized by a textual label only. Through this label alone, it is not possible to uniquely identify the person. When disambiguation of persons can be performed, it is possible to have, through the subclass KnownPerson, a better-defined Person entity, enabling capabilities such as links to homepages or to existing vCards.

To be able to import all the possible types of authors coming from PubMed, the class CollectiveName has been defined. This class works like the Person class. It is made up solely of a label until such time as the label can be recognized and better detailed as an Organization. Again, when importing bibliographic references from PubMed, instances of CollectiveName will be strings representing organizations. In SWAN, besides representing potential authors of digital resources, an organization (e.g., PubMed, Alzforum, etc.) may represent the authoritative source for knowledge elements such as the digital resource or its metadata.

3.4 'DiscourseElements': the Core of SWAN

Discourse elements classes represent the core of the SWAN Ontology. Through such classes it is possible to use self-annotated discourse as a bridging ontology connecting the many specialized research sub-domains contributing to AD research and to research in general. The advantage of this approach is that the bridging ontology will automatically track the knowledge as it emerges, and is not required to make "value judgments" about the proper bridging level concepts. The bridging level, in fact, becomes concrete as speech acts, which are documented only as to what is said, and its logical and / or evidentiary relationship to other statements. The bridging level is, in Hausser's terminology, a "+constructive" ontology [19], that is, an ontology about what is said, rather than about agreed-upon objective facts.

The ResearchStatements in the ontology characterize digital resources which themselves contain statements in the informal ontology of English or other languages. Each ResearchStatement may also be linked dynamically 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, fluid, multi-hued, 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. The connecting point in SWAN's ontology for externally defined ontological categories is the "Concept", which functions as a kind of "adapter" in allowing these links to be made.

The SWAN discourse elements are:

- research statements: a claim or an hypothesis
- research questions: topics under investigation
- comments: personal annotation or collective discussion

DiscourseElements have a simple set of attributes (besides the creation date coming from the super-class SWAN Thing, a title and a description) but a very important variety of relationships. In order to give an idea of the properties involving the discourse elements we take into consideration an example of a research statement creation and in particular of a hypothesis.

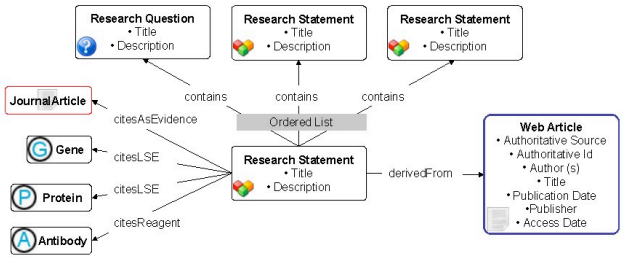

Figure 1 – Some example relationships of a DiscourseElement

In **Figure 1** we depict a possible instantiation showing some relationships between the primary research statement and other SWAN entities.

In this particular case, the research statement (an hypothesis, actually) 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 from scratch by an author in the SWAN environment.

The "content" of ResearchStatements is then composed of an ordered list of other DiscourseElements This "has part" relationship is defined by "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 life science entities/reagents through "citesLifeScienceEntity" or "citesReagent".

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

Therefore, it is possible to have three cases:

1. A new research statement from scratch. This is shown in **Figure 2**. The research statement can be detailed in a title and description and it is possible to relate it to other discourse elements 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 provenance will be defined by the curator - which could correspond with the original author, or with a knowledge base editor.

2. Full reuse of an existing research statement. In this case it is possible to include in the primary research statement an already existing discourse element as it is. The provenance of such discourse elements is maintained. But the connection between the primary research statement and the already existing discourse element could have a different creator.

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.

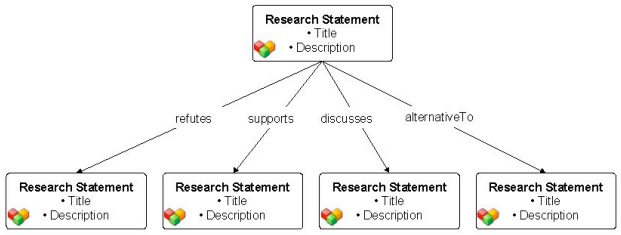

Figure 2 – Examples of logical relationships among discourse elements

Other interesting use cases come from the idea of commenting that in SWAN will have the form of the Comment entity. A Comment upon a ResearchStatement can mirror a comment that has been published on a journal or on a web site, or can be created from scratch. 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 in the SWAN workbench directly. The comment is always "inResponseTo" some other discourse element (because we are modeling dialogue) and such relationship can be characterized more fully through a supports/discuss/refutes relationship.

A Comment has a form similar in certain respects to a ResearchStatement. It can be composed by an ordered list of discourse elements, it can refute, support or discuss other discourse elements and it can be alternative to some other discourse element. It can cite life science entities or reagents as well as digital resources. As with all the other discourse elements, Comments can present an ordered list of authors. Some instantiable relationships of Comment are shown in **Figure 3**.

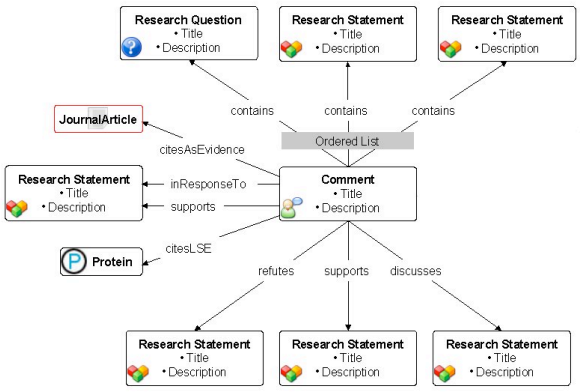

Figure 3 – Examples of possible relationships between Comment and other SWAN classes.

The last discourse element in Version 1.0 is the ResearchQuestion. A ResearchQuestion can be contained in another discourse element or it can be "motivatedBy" another discourse element. ResearchQuestions are open topics of investigation where dialogue is initiated and experiments performed.

3.5 Concepts, LSEs and Reagents

An important contribution to scientific discourse in Alzheimer Disease (AD) research and in many other biomedical contexts is integration of scientific statements with external resources such as concepts coming from the Gene Ontology (GO), genes coming from Entrez Gene, as well as antibodies detailed in the Alzforum antibody database. In order to keep under control the evolution of the SWAN environment, we decided to create an instance in SWAN for all those external resources, a unique individual for each unique entity.

According to the same approach that guided us in the definition of digital resources, we store the minimum set of information allowing search and entity recognition by the user. Such instance will then point to all the external resources such as web pages in Entrez Gene or HUGO Nomenclature Committee as well as RDF that is going to be provided, in a near future, by Alzforum for the antibody database.

The current version of the ontology provides the capability of referencing the following entities:

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

3.6 Tags

As already mentioned, SWAN includes support for personal data organization. One of the mechanisms provided is the well-known concept of tag. Besides a textual label, the Tag class presents another attribute that can be used to possibly define the tag type. Thus, if a user is defining a custom free text tag, the type will be "custom" but if the tag is coming from a terminology or taxonomy it will have the type "MeSH" for instance. The latter case is particularly useful when we import records from PubMed. Together with the bibliographic record, PubMed provides some keywords encoded through the MeSH terminology. The way it is possible to encode such terms related to a digital resource is another application of the Comment class. In fact, given that the terms connected to a journal article in PubMed are authored by the PubMed organization, we decided to treat them as comments as they are an expression of a re-elaboration of the original work.

3.7 Qualifiers

Qualifiers are predefined tags that can be applied only in specific context. The most important example is given by the resource statements that can be qualified as Claim or Hypothesis originally by the curator. But the mechanism of qualifiers as well as the one of tags allows any user to tag or qualify the entities as he/she prefer fostering personal knowledge organization.

3.8 Versioning and Evolution

The SWAN ontology has been designed to support the knowledge life cycle, including evolution of research statements as knowledge evolves. When a curator is defining a new entity, for instance a research statement, automatically this will become the first version of the entity. If the same user is changing this research statement, we will have a new version of the entity. If another user wants to start from that entity to define his own version it will become a new entity "evolvedFrom" the original research statement but with another curator/author.

4. Applying the Ontology in Practice

In **Figure 4** below we illustrate a small section of an example applying the SWAN ontology to real scientific discourse in which there is substantial uncertainty and conflict over correctness of competing models of AD pathology. Biologists and science curators on our team worked numerous such examples in detail, and in parallel with development of both the ontology itself and the software which will allow scientists to apply the ontology in their daily work. As an element of implementing the SWAN project, we are in the process of annotating several dozen large-scale current hypotheses in AD research which will be provided as an initial content store to our user community via the Alzforum. Quality assurance in development of the annotation will be provided by scientific staff of the Massachusetts Alzheimer Disease Research Center, Massachusetts General Hospital (http://www.massgeneral.org/neurology/MADRC).

The example in **Figure 4** shows a small section of the metadata developed around two current hypotheses of AD etiology, originating from Vin Marchesi at the Yale Medical School (Marchesi, Intracellular a-Beta Dimers Hypothesis) [20] and the group of Karen Hsiao-Ashe at the University of Minnesota (Lesné et al, a-Beta*56 Hypothesis) [21]. Both scientists attempt to develop models which explain literally thousands of research observations tying plaque deposits of 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 the SWAN ontology 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 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 exterts 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 logical relationship 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 a framework for making serendipitous discoveries of research previously unknown to them.

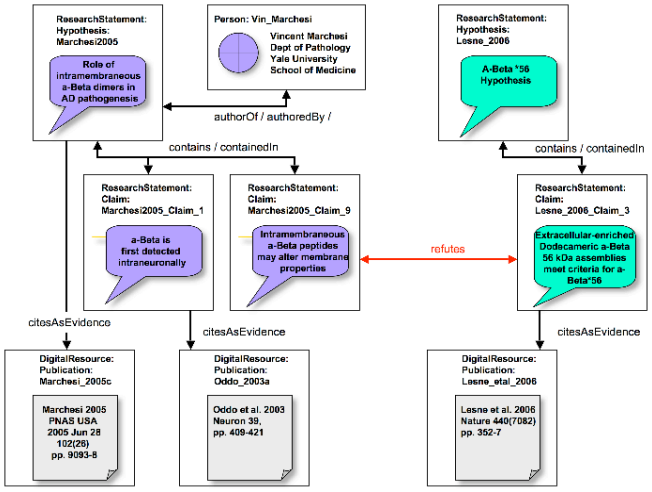

Figure 4 – How the ontology is instantiated for conflicting hypotheses, claims and evidence – example of refutation.

3.9 SWAN's Relationship to Other Efforts in the Biomedical and Bio-ontology Communities

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 (quality control of hypothesis content); the W3C Health Care and Life Sciences Task Force (development of AD and Parkinson's Disease research-based use cases and an interoperability demonstration); and the Sense Lab Group at Yale School of Medicine's Department of Medical Informatics. 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 mid 2007.

5. 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 each other 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.

6. ACKNOWLEDGMENTS

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

Thanks as well to Sean Martin, Ben Szekley and Lee Feigenbaum (IBM Advanced Internet Technology Group); Brad Hyman (Harvard Medical School and Massachusetts General Hospital) and Carole Goble (University of Manchester) for many valuable discussions.

The SWAN project team members are June Kinoshita, Elizabeth Wu, Gwen Wong, Marco Ocana, Paolo Ciccarese, Ben Szekley, and Tim Clark.

7. REFERENCES

[1] Clark T and 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, et al. SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research. J Web Semantics, 2006, 4(3).

[3] Wong GT, Gao Y, Wu E, et al. Developing SWAN, a shared knowledge base for Alzheimer's disease research. Abstracts, Society for Neuroscience 2006, Atlanta, GA.

[4] Kinoshita J and Strobel S. 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. Ed. Jucker M, Beyreuther K, Haass C, Nitsch R, Christen Y. Springer, Berlin, Heidelberg, New York, 2006, pp. 457-463.

[5] Lam YK, Marenco L, Clark T, et al. 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, pp. 158-170.

[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 (2007) 'Semantic Web Meets e-Neuroscience', BMC BioInformatics (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 (2007) Advancing translational research with the Semantic Web. BMC Bioinformatics (in press).

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

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

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

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

[13] Nardi BA. Activity Theory and Human-Computer Interaction, in Nardi, B. (ed.) 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, and Clark T (2003) 'Computational Knowledge Integration in Biopharmaceutical Research'. Briefings in Bioinformatics, VoL 4(3), pp 260-278.

[15] Klein PD (2005) 'Concept of Knowledge' in Craig E (ed) The Routledge Shorter Encyclopedia of Philosophy. Abingdon, Oxfordshire, UK: Routledge, 2005, p. 525.

[16] Davenport T and Prusak L (1998) Working Knowledge. Harvard Business School Press: Boston, MA.

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

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

[19] 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.

[20] Marchesi V (2005) 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.

[21] Lesné S et al. (2006) A specific amyloid-beta protein assembly in the brain impairs memory. Nature 2006 Mar 16; 440(7082):352-7.