SWAN: A Distributed Knowledge Infrastructure for Alzheimer Disease Research

+ Model
ARTICLE IN PRESS

Available online at www.sciencedirect.com
SCIENCE DIRECT®

Web Semantics: Science, Services and Agents on the World Wide Web xxx (2006) xxx–xxx

JOURNAL OF Web Semantics
www.elsevier.com/locate/websem

SWAN: A distributed knowledge infrastructure for Alzheimer disease research

Yong Gao [a], June Kinoshita [b], Elizabeth Wu [b], Eric Miller [c], Ryan Lee [c], Andy Seaborne [d], Steve Cayzer [d], Tim Clark [a,e,*]

[a] *MassGeneral Institute for Neurodegenerative Disease, Charlestown, MA 02129, USA*
[b] *Alzheimer Research Forum[1], MA, USA*
[c] *Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[d] *Hewlett-Packard Laboratories, Bristol BS34 8QZ, UK*
[e] *Initiative in Innovative Computing, Harvard University, Cambridge, MA 02138, USA*

Received 15 November 2005; accepted 1 May 2006

Abstract

SWAN – a Semantic Web Application in Neuromedicine – is a project to develop an effective, integrated scientific knowledge infrastructure for the Alzheimer disease (AD) research community, using the energy and self-organization of that community, enabled by Semantic Web technology. This infrastructure may later be deployed for research communities in other neuromedical disorders. SWAN incorporates the full biomedical research knowledge lifecycle in its ontological model, including support for personal data organization, hypothesis generation, experimentation, laboratory data organization, and digital pre-publication collaboration. Community, laboratory, and personal digital resources may all be organized and interconnected using SWAN's common semantic framework.

Keywords: Alzheimer disease; Biomedical research; Knowledge lifecycle; Ontology; Digital resource

1. Introduction

Neurodegenerative diseases are highly complex disorders. Researchers over the past 20 years have made significant progress in understanding Alzheimer disease and related neurological disorders. They have produced an abundance of data implicating diverse biological mechanisms in the etiology of such diseases. These include genes, environmental risk factors, changes in cell functions, DNA damage, accumulation of misfolded proteins, cell death, immune responses, changes related to aging, reduced regenerative capacity, and others. Yet there is still no clear agreement on the etiology of AD. Citation analysis from the Alzheimer Research Forum estimates that there are more than 40,000 citations in the PubMed database of relevance to neurodegenerative diseases, and 150–200 new studies are published each week.

The challenge of integrating so much data into testable hypotheses and unified concepts is clearly formidable. Researchers must strive to formulate testable hypotheses built on a corpus of research derived from multiple experimental modalities within many subfields of biomedicine and related areas, in all of which it is impossible to be expert simultaneously. The situations for Parkinson's, Huntington's, and ALS researchers are similar.

SWAN is an attempt to develop a practical, common, semantically-structured, web-compatible framework for scientific discourse using Semantic Web technology [1–3] applied to the problems of integrating multimodal scientific discourse, in the search for a cure for Alzheimer disease. The initial concept for SWAN was proposed in a talk at the W3C Semantic Web in Life Sciences workshop, October 2004 [4].

SWAN is intended to operate at the individual and community levels, enabling a system of interoperable personal and community knowledge bases. Individuals will use SWAN software as a personal tool to find, filter, and organize information. At the community level, the same software and the same ontological framework can be used to organize and curate the research of

* Corresponding author. Tel.: +1 617 947 7098; fax: +1 617 724 1480.
E-mail address: tim_clark@harvard.edu (T. Clark).
[1] www.alzforum.org.

1570-8268/$ – see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.websem.2006.05.006

a laboratory or an entire research community. Contextualized elements of the personal KB can be shared with the community at a low incremental cost. Community KB elements may also be shared with individuals and re-used in new contexts.

SWAN provides semantic interoperability of digital resources based on a common set of software and a common ontology of scientific discourse. This ontology is specified in an RDF Schema available on the web.[2] SWAN's content is intended to cover not just published literature, but all stages of the "truth discovery" process in biomedical research, from formulation of questions and hypotheses, to capture of experimental data, sharing data with colleagues, and ultimately the full discovery and publication process. This content is intended to be constructed and deployed by individual scientists working to organize their own data and knowledge, for their own benefit; in cooperation with community editors who collect, organize, and redistribute this knowledge.

The community members in SWAN, unlike those in a process such as Wikipedia,[3] are principally concerned with advancing their own research program. The incremental effort required to share knowledge from the team to the community will be relatively small, beyond that required in the standard publication process for scientific literature. We believe this will result in the creation of the highly facilitative knowledge-sharing networks argued for by the leadership of neuroscience research institutes at NIH [5].

2. The system-level use cases

The major SWAN system use cases are designed to be implemented as part of the existing scientific knowledge ecosystem—which includes scientists, scientific discourse, experiments, data, grant applications, publications, scientific databases, bibliographic databases, scientific ontologies, biomedical research collaborations, and scientific web communities.

SWAN's principal goal is to apply Semantic Web technology to this existing ecosystem in a way that can (a) enhance the productivity of the ecosystem as a whole (b) benefit each human constituency to ensure uptake and socialization (c) enable websites, individual scientists, and scientific laboratories to participate in virtual collaborations.

Primary System Use Case specify and implement a common semantic framework for scientific discourse across the knowledge ecosystem of science, compatible with the Web and with current approaches to managing scientific information. In this way, knowledge and discourse can be organized on a community website, a laboratory website, or a personal computer in mutually interoperable schemas.

Three *Supporting System Use Cases* further specify the primary use case:

- *Organize and annotate* digital scientific resources as integrated KBs across content types, using multiple ontologies.
- *Securely share* digital scientific resources including the ontologies and annotation generated in Use Case 1, from individuals to diverse communities and back again.
- *Provide integrated access* to digital scientific resources for a single scientist, a single community, or multiple communities, as a distributed knowledgebase, organized by the structures specified in Use Case 1.

3. Discussion

Biomedical researchers engage in certain typical patterns of activity in keeping up with the literature, developing hypotheses, planning research, applying for grants, analyzing data, and preparing for publication. These activities are common to the vast majority of researchers. They include

- Searching, reading, and thinking critically about the professional literature in their field.
- Formulating testable hypotheses consistent with the "story" or explanatory model.
- Finding possible connections amongst disparate data, creating a plausible explanatory "story" or model which can bridge gaps or open challenges in the existing body of knowledge.
- Designing experiments to test their hypotheses.
- Running the experiments.
- Collecting and analyzing experimental data.
- Interpreting data, e.g. by modifying the hypothesis, connecting it to other findings or hypotheses.
- Organizing personal collections of publications and related documents according to a relevant conceptual system to enable retrieval at a later date.
- Applying for grants to support their work (which typically involves presenting the model, hypotheses, and preliminary data).
- Communicating with other researchers, funding agencies, publishers, conference organizers, and local institutional management.
- Writing scientific articles for publication, preparing conference presentations, informal talks, and poster sessions.

Many of these activities are currently supported by public or private information systems, ranging from Google® to personal Excel® spreadsheets and personal bibliographic managers such as EndNote®. However, these tools all have their shortcomings from the knowledge ecosystem view, because they lack semantic constructs connecting the personal, community, and science-wide realms of discourse. Because digital resources in these spaces are largely organized using incompatible knowledge schemas, contextual information in the knowledge ecosystem is continually lost as it passes through human beings navigating point-and-click interfaces.

A public ontology is required for scientific communication—it establishes the terms of discourse. Biologists have been developing ontologies since at least the time of Aristotle. Private ontologies, inherently modifiable without discussion,

[2] Available at http://purl.org/swan/0.1. The trailing slash is significant. Also, depending upon how they deal with content types, some browsers may require a "view source" operation to see the RDF.

[3] Wikipedia: The Free Encyclopedia http://www.wikipedia.org.

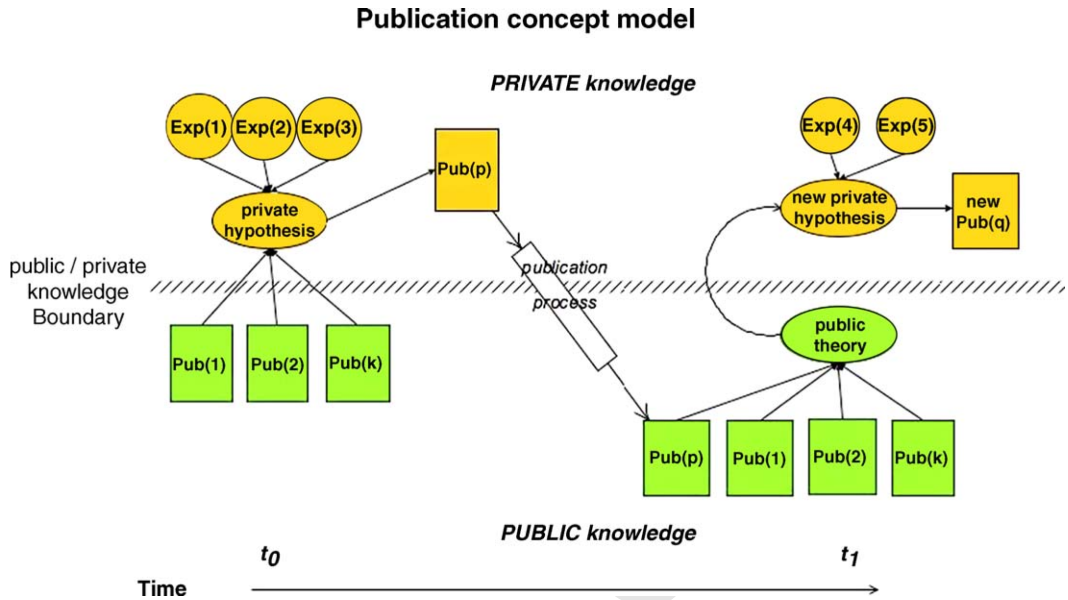

Fig. 1. Conceptual model of scientific publication.

are also required to support active research, in which new things and processes are constantly discovered, described, and named.

Clearly it is essential to incorporate shared public concepts and relationships into the organizational scheme, while also providing for personal differences or discoveries to be modeled and declared. What we are after here, from the viewpoint of the philosophy of science, is a formal way to represent potentially *incompatible* scientific models, which does not also force them to become *incommensurable*. To do this we require some public bridging ontology. In SWAN this is an ontology of reasoning and discourse.

Visser et al. discuss the problem of heterogeneous ontologies as barriers to system interoperability of varying severity [6] and discuss approaches to allowing heterogenous ontologies to communicate within a distributed system. This is essentially our problem, and we adopt an approach largely consistent with two of their proposed solutions (1) domain partitioning and (2) alternative domain views [7]. We will limit ontology mismatches to what Visser and Cui call *content heterogeneity* across a core set of structures.

Formally, SWAN adopts what Hausser calls the "+constructive" response in ontological model theory: in our ontological model, "the model-structure is part of the speaker-hearer" [8]. We recognize the act of cognition as seated *in individuals* practicing a scientific discipline in the material world... and make it part of our semantics. A significant part of this discipline is represented by scientific discourse. Hausser associates the [+constructive] interpretation particularly with the goal of analyzing language meaning, as opposed to the [−constructive] response, whose goal is "to characterize truth" and which he associates (exclusively) with science and mathematics. However, we do not make such a dichotomy. At least in biomedicine, discourse is not restricted to absolute propositions in which the author and context are either absent from the scene, or irrelevant to validation.

The [+constructive] model is in many ways implicit in bibliographic databases. GenBank [9] long ago[4] moved from a data model in which a consensus sequence was maintained, as "absolute truth", to a model accepting and publishing the varying experimental results of each researcher. This model therefore recognizes the speaker... but the hearer remains implicit. An explicit treatment of the hearer allows a collaboration network to be established.

Publication is a prominent part of the scientific discourse. Our notion is to join it with the supporting reasoning and evidentiary data in a knowledge schema. A conceptual model of knowledge acquisition and publication by an individual scientist is shown in Fig. 1. Documents (or *evidence*), and assertions upon documents, are fundamental objects in our system. Document assertions connect the discourse to its foundations, and concern the document characteristics, provenance, content, statements about the documents, categorization of the documents, and relationships to other documents and assertions.

We are not attempting to construct a formal computational language of biology. What we are attempting in our ontology is to increase the interoperability across various models specified in text, through establishing improved connections among documents and assertions about them.

Fig. 2 is a conceptual sketch of the relationship of scientific hypotheses, public and private ontologies, and documents. We believe that a successful knowledge infrastructure needs to support these relationships with special emphasis on public, private,

[4] Circa 1990, when GenBank was transferred from Los Alamos National Laboratories to the NCBI, and re-engineered.

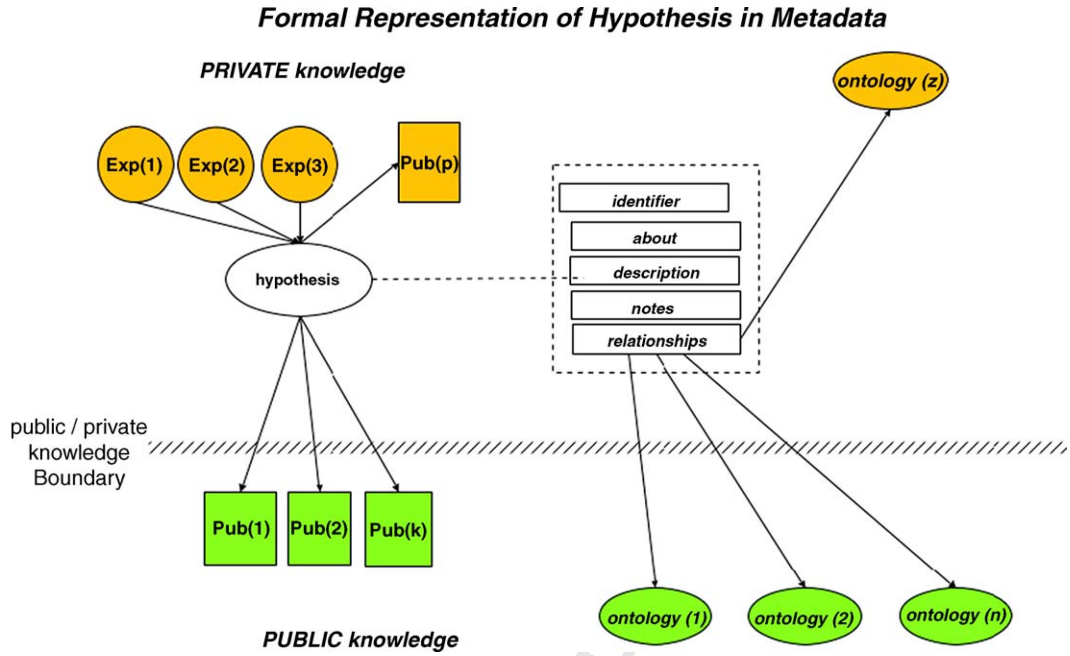

Fig. 2. Representation of hypotheses as metadata.

and shared knowledge definitions; and to support evolution and transition between these states.

4. Socialization

Successful socialization of our system is the key to success, because it is powered by individual scientists. In our view, socialization has the following three basic requirements:

1. Scientists can use the system to organize their own personal data, gaining efficiency, and insight into their own processes and project history.
2. A convergent public view of data is supported through publication of private views.
3. A researcher may combine what he/she knows, with what the public view of data in our system provides, to discover *something surprising and new*.

We have attempted to support all three elements in the design of SWAN.

Currently, papers are generally published as one-dimensional units, meaning there are little to no links or associated information besides the references cited. Yet there is a whole host of information that is not transmitted with a paper. Some journals provide useful links to additional support/supplemental materials, which cannot be included in the paper due to the word limit imposed by the editor. These limits help the journal publish more papers per issue (i.e. more cost effective), but severely limit the scientist trying to duplicate experiments by the lack of information.

Some investigators provide their own website to post additional information. Other beneficial information may include images, tables, data base links (e.g. AlzGene), websites links (e.g. Alzheimer's Research Forum), collaborator information, previously published and non-published data (this may be a problem due to copyright issues), and detailed methods, including specifics on reagents (which can be a non-trivial issue).

This additional information would give the paper multiple dimensions by embedding this associated information within the paper (when opened electronically) and/or providing links to other information that is too large to embed. This concept is an expansion of the orange to green transition seen in the right-hand portion of Fig. 1. Clearly, all the information under "Private knowledge" space is not transmitted in the publication process for many reasons, including the motivation and the ability to collect this information in a standardized way. If a researcher is collecting this additional information in a software program during the building of a "Private hypothesis" (Fig. 1 top-half), knowing that it will be used for their publication (bottom half), then it will provide strong motivation for its use. Additionally, if the data structure becomes a standard way to relay information to other researchers, investigators will support its use (e.g. Word or Excel documents).

Publishing is one of the major factors motivating researchers, because it is closely tied to securing funding and promotion. Publications are a snapshot of an individual's thoughts and experiments, and of the evolution of scientific thought as a whole. As indicated in the bottom of Fig. 1, time is the *X*-axis. The process depicted here represents a unit of time (although variable) which repeats itself over a scientist's life manyfold. Often what is lost in this process is how these units became connected and any information that never made it to publication. This could be due to lack of time, funding, technical problems, incorrect hypothesis or lack of acceptance by the scientific community for a certain line of reasoning. Much of this information is kept

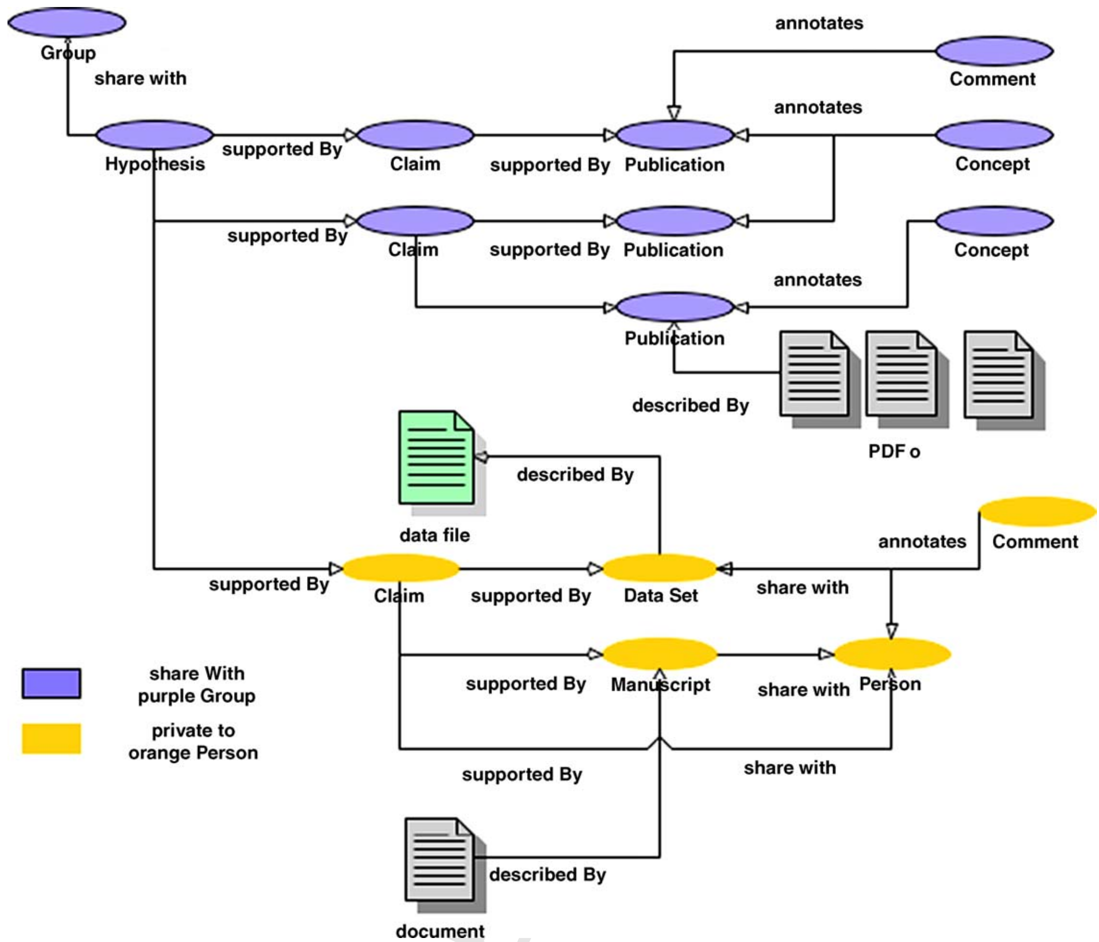

Fig. 3. SWAN semantic relationships.

as "Private knowledge" cloistered in notebooks or the archives of the brain.

Providing a platform to document ideas that succeeded (i.e. published), failed or were never evaluated has a very significant scientific value allowing current or future generations to extend, avoid, or develop these ideas. Such a model could either have a historical perspective built on years of accumulated knowledge or may be a de novo idea based on a new observation.

An immediate example of this program's value could be seen in a student–teacher relationship, in transmitting the teacher's view of a particular subject to a naive student. If the student wants to understand this view it would useful if he or she was able to see a model of this hypothesis containing all the information gathered together to support this idea. This project has the potential to build a program that would allow the collection of thoughts, data, and experiences over a lifetime, creating a scientific life history. Most of this data will be collected in the "Private knowledge" space, but is built on the *Publication Model* described above.

A significant question is, when will one allow their private world to become public? At a minimum, scientists would be inclined to release this "Private knowledge" at the end of their scientific careers. Nonetheless, without the effort to collect this highly valuable knowledge it is doomed to be lost forever. Additionally, some of the payoff of the collection of this "Private knowledge" would not always be immediate, but would be the beginning of a knowledge base that would grow, benefiting future generations. These two models are not mutually exclusive, but in fact are intertwined because the "Publication Model" is an element repeated over time giving a "Scientific Life History." The value of collecting this information cannot be underestimated and to our knowledge has not been done in a systematic manner that would be searchable.

5. The SWAN pilot

The SWAN pilot project has three major components, which are intended to work together as an integrated whole.

- SWAN ontology.
- Semantic Bank & faceted browser.
- SWAN Information Management Tool (SwIM).

The SWAN ontology permits knowledge content from multiple stages of the scientific discovery life-cycle to be represented in the W3C Resource Description Framework (RDF), in a way that can support electronic pre-publication group sharing and collaboration, as well as personal and community knowledge base construction. The current version of this early schema 11 (Clark, Gao et al. [10]) can be persistently referenced on the web for re-use by other applications. Fig. 3 gives an example of how the schema instantiates a Hypothesis with supporting Claims and evidence, combining public (community) and private information.

Several information categories created and managed in SWAN are defined as subclasses of Assertion. They include Publication, Hypothesis, Claim, Concept, Manuscript, DataSet, and Annotation. An Assertion may be made upon any other Assertion, or upon any object specifiable by URL. For example, a scientist can make a Comment upon, or classify, the Hypothesis of another scientist. What makes this something more than an intellectual exercise is that linking to objects "outside" SWAN by URL allows one to use SWAN as metadata to organize – for example – all one's PDFs of publications, or the Excel files in which one's laboratory data is stored, or all the websites of tools relevant to Neuroscience.

Each Assertion has a set of information including the speaker–hearer pairing (owner and persons it may be shared with); abstract; citations to other Assertions or miscellaneous URIs. Depending upon the subclass it may include some or all the "citation" information normally associated with a journal article. It may also reference a content image, such as a PDF; and an entry in a public bibliographic database. Citations to other Assertions may be evidentiary, inclusive, or referencing. Evidentiary Citations are used in asserting that some Assertion is evidence supporting a Hypothesis, Claim, or other Assertion. Inclusive Citations are used to specify the Assertions which belong to a Collection. Referencing Citations are used wherever a reference to something is made for a purpose other than those previously described.

Annotation may be structured or unstructured. Structured annotation means attaching a Concept (tag or term) to an Assertion. Unstructured annotation means attaching free text. Assertions may be imported from Alzforum, Pubmed, EndNote bibliographies previously exported in XML, RDF N3 serialization, and from other SWAN-RDF stores, using SwIM. Assertions may also be exported in RDF or in EndNote-compatible XML. SWAN Assertions may be organized by placing them in a Collection.

SWAN uses a speaker–hearer core ontological model. Therefore, Persons and Groups need to be defined as sources and targets of discourse for each Assertion. Groups are named collections of Persons. Persons are a subclass of Group containing only a single Person.

Concepts are nodes in controlled vocabularies, which may also be hierarchical (taxonomies). Concepts natively supported include special Alzforum categories, MeSH terms, and Gene Ontology 12 13 (Harris et al. [12]) categories. Genes and Pro-

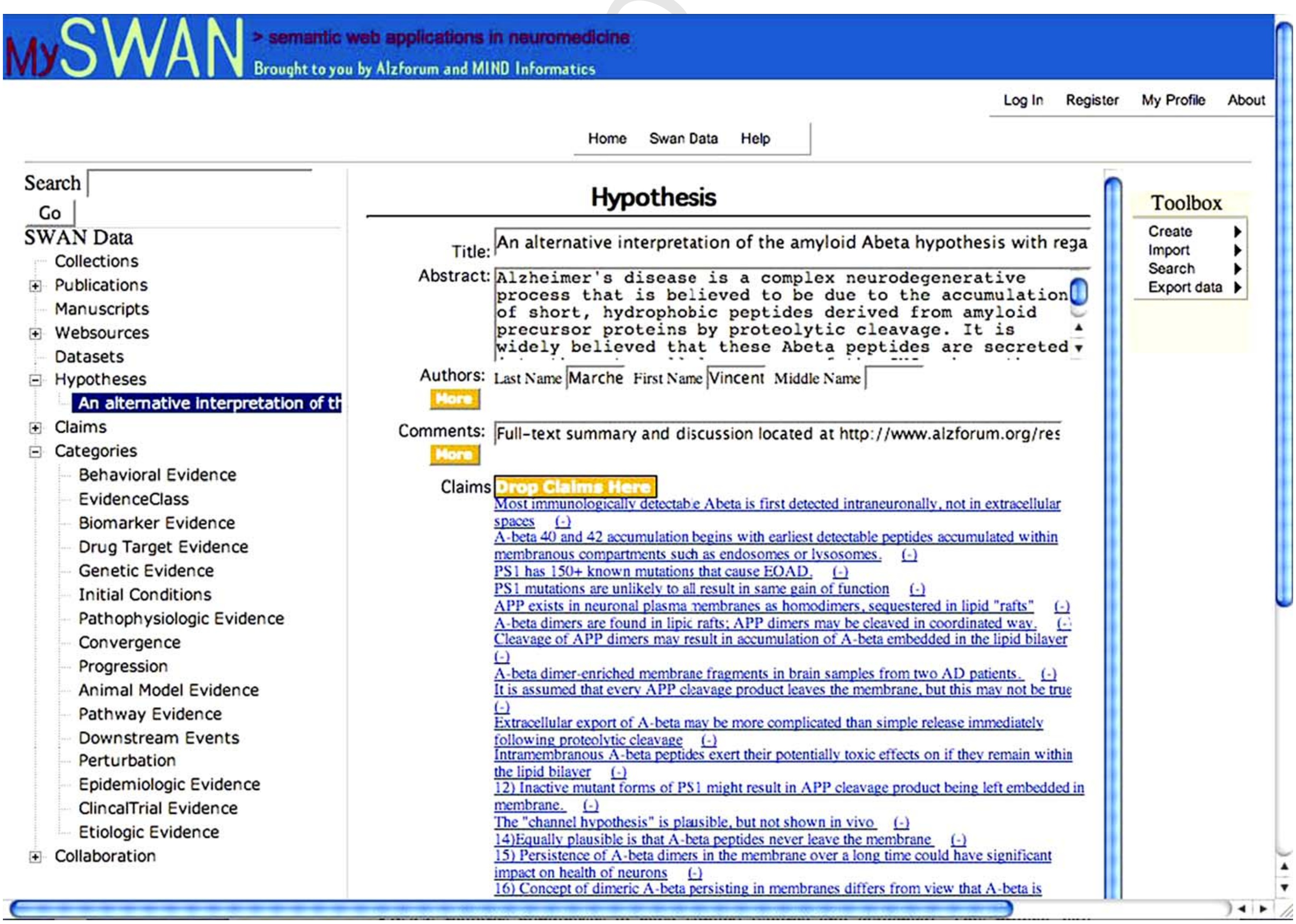

Fig. 4. mySWAN browser snapshot.

teins are considered Concepts in SWAN, as are Organism names. Personal concepts may be added by the user.

A SWAN Collection is a set of Assertions. Typically a Collection might include publications, annotations, statements of Hypotheses and supporting evidence, and so forth.

Alzforum website may be extracted, transformed to SWAN-RDF, and stored in a Semantic Bank repository. This is an RDF knowledge base, which can be queried and displays its contents in the browser. The current SWAN Semantic Bank is a prototype of one way SWAN's information can be published on the web in a directly accessible and queryable form. This is an extension of previous work at MIT on the Simile project [13].

A pilot version of the SWAN Information Management Tool (Fig. 4) has been developed to allow hypotheses, concepts, publications and other information to be annotated, linked to fundamental documents, and organized by annotators and/or individual scientists. These objects are stored in SWAN-RDF form in a personal or community semantic repository. This tool is a simplified version of what will eventually be used by scientists to manage their personal data, or by a laboratory or community website to manage shared data.

SwIM allows knowledge elements (Assertions) from the individual repository to be constructed; linked to existing digital resources such as Excel files and PDFs; organized; and shared to the community space, with specific collaborators—or kept private.

SwIM attempts to provide a pragmatic knowledge modeling capability to scientists, based on observations and discussion of how they actually do their work and what would be useful. For example, other more elaborate and elegant approaches have been developed to modeling scientific claim [14]. Our approach limits the model's complexity to what we feel can be of immediate benefit to a working scientist in preparing a grant application or writing a paper.

SwIM permits linking any Assertion to an arbitrary URL as the underlying object the assertion is made upon. This means for example that a concept map can be constructed of useful Websites (WebPage is a class in the SWAN vocabulary) and published as RDF metadata—which can itself be stored in a Semantic Bank and viewed through a browser as a resource ontology.

6. Conclusion

The primary goals of SWAN are to provide an improved structure for public discourse between laboratories, to enable "surprise" connections between groups working (possibly unknowingly) on related matters, to synthesize scientific results across the AD community, and to enable a better "organizational memory" within individual laboratories.

We are not building an informatics model of biology. Such an effort would lag perpetually behind the science. It could be of little use to specialists because cutting-edge research – at least in biomedicine – tends to produce controversy before it produces a single accepted model.

What we are after is to build an extensible model of digital resources in the process biologists themselves follow, through which they endeavor to construct accurate models of biological phenomena. We will then use this model to create tools biologists can use to accelerate the process of discovering new knowledge, by removing barriers to effective discourse and increasing the interconnectedness of new discoveries.

Uncited reference

[11].

Acknowledgements

We thank Dean Hartley, Brad Hyman, Zane Hollingsworth, Marian DiFiglia, Lars Hernquist, Dora Kovac, David Grossman, Sean Martin, Ralph Swick, Sherrilynne Fuller, Jim Hendler, and Carol Goble, for thoughtful discussion, constructive criticism, and moral support. We are also grateful for the support of the Ellison Medical Foundation, which funded portions of this research.

References

[1] T. Berners-Lee, A Roadmap to the Semantic Web, 1998.
[2] T. Berners-Lee, J. Hendler, et al., The Semantic Web, Scientific American, May 2001.
[3] J. Hendler, Science and the semantic web, Science 299 (2003) 520–521.
[4] T. Clark, J. Kinoshita, A pilot KB of Biological Pathways Important in Alzheimer's Disease, W3C Workshop on Semantic Web for Life Sciences, Cambridge, MA, USA, October 2004.
[5] T.R. Insel, N.D. Volkow, et al., Neuroscience networks: data-sharing in an information age, PLoS Biol. 1 (1) (2003) E17.
[6] P.R.S. Visser, D.M. Jones, et al., An Analysis of Ontology Mismatches; Heterogeneity versus Interoperability, AAAI 1997 Spring Symposium on Ontological Engineering.
[7] P.R.S. Visser, Z. Cui, Heterogeneous Ontology Structures for Distributed Architectures, ECAI-98 Workshop on Applications of Ontologies and Problem-Solving Methods.
[8] R. Hausser, The four basic ontologies of semantic interpretation, in: Tenth European-Japanese Conference on Information Modelling and Knowledge Bases, Saariselka, Finland, IOS Press, Amsterdam, The Netherlands, 2000.
[9] D.L. Wheeler, T. Barrett, et al., Database Resources of the National Center for Biotechnology Information, Nucleic Acids Res. 33, Database Issue, 2005, pp. D39–D45.
[10] T. Clark, Y. Gao et al. SWAN Vocabulary Version 0.1. Boston, MA, MIND Center for Interdisciplinary Informatics (URL: http://purl.org/swan/0.1/, Accessed April 30, 2006).
[11] M. Ashburner, C.A. Ball, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[12] M. Harris, J. Clark, et al., The Gene Ontology (GO) Database and Informatics Resource, Nucleic Acids Research 32, Database Issue, 2004, pp. D258–D261.
[13] E. Miller, R. Swick, An overview of W3C semantic web activity, Bull. Am. Soc. Inf. Sci. Technol. 29 (4) (2005) 8–11.
[14] S.J. Buckingham Shum, V. Uren, et al., Modelling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues, Knowledge Media Institute, 2004, pp. 1–42.