Journal of Biomedical Informatics 41 (2008) 739–751

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The SWAN biomedical discourse ontology

Paolo Ciccarese [a,d], Elizabeth Wu [b], Gwen Wong [b], Marco Ocana [a], June Kinoshita [b], Alan Ruttenberg [c], Tim Clark [a,d,e,*]

[a] Massachusetts General Hospital, Boston, MA 02129, USA
[b] Alzheimer Research Forum Foundation, Boston, MA 02109, USA
[c] Science Commons, Cambridge, MA 02139, USA
[d] Harvard Medical School, Boston, MA 02115, USA
[e] Harvard Initiative in Innovative Computing, 60 Oxford Street, 4th Floor, Cambridge, MA 02138, USA

ARTICLE INFO

Article history:
Received 5 September 2007
Available online 4 May 2008

Keywords:
Ontology
Semantic web
Scientific discourse
Biomedicine
SWAN
Alzheimer Disease
Knowledgebase

ABSTRACT

Developing cures for highly complex diseases, such as neurodegenerative disorders, requires extensive interdisciplinary collaboration and exchange of biomedical information in context.

Our ability to exchange such information across sub-specialties today is limited by the current scientific knowledge ecosystem's inability to properly contextualize and integrate data and discourse in machine-interpretable form. This inherently limits the productivity of research and the progress toward cures for devastating diseases such as Alzheimer's and Parkinson's.

SWAN (Semantic Web Applications in Neuromedicine) is an interdisciplinary project to develop a practical, common, semantically structured, framework for biomedical discourse initially applied, but not limited, to significant problems in Alzheimer Disease (AD) research.

The SWAN ontology has been developed in the context of building a series of applications for biomedical researchers, as well as in extensive discussions and collaborations with the larger bio-ontologies community.

In this paper, we present and discuss the SWAN ontology of biomedical discourse. We ground its development theoretically, present its design approach, explain its main classes and their application, and show its relationship to other ongoing activities in biomedicine and bio-ontologies.

1. Introduction

The SWAN project (Semantic Web Applications in Neuromedicine) aims to develop a practical, common, semantically structured framework for biomedical discourse initially applied, but not limited, to significant problems in Alzheimer Disease (AD) research. The SWAN project is the result of collaboration between the Alzheimer Research Forum (Alzforum) and informaticians at Harvard University, and 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 development stage of its first production-quality application [2–4]. 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 [5–7] 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 10-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 4000 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. SWAN attempts to model the scientific discourse about AD and its supporting evidence in a rich and extensible way that is compatible with the way

* Corresponding author. Address: Harvard Initiative in Innovative Computing, 60 Oxford Street, 4th Floor, Cambridge, MA 02138, USA. Fax: +1 617 724 1480.
E-mail addresses: tim_clark@harvard.edu, twclark@nmr.mgh.harvard.edu (T. Clark).

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.04.010

the domain of Alzheimer research functions as a technology-mediated knowledge 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 knowledgebase of major hypotheses, claims, evidence and discourse relationships in AD research. These applications are discussed more fully in Sections 3.2 and 5.

2. Scientific discourse and truth on the web

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 knowledge ecosystem is mediated by the technology of the web.

Philosophers of science have defined knowledge as "warranted true belief" [11]. The classical knowledge management definition of knowledge is a bit more limited: "information in context" [12]. This latter definition is insufficiently specific about evidence and the material basis of knowledge. Scientific knowledge strives to approximate objective truth, about a world that exists independently of our beliefs about it. Therefore scientific knowledge by its nature requires experimental validation—evidence—as a warrant for belief.

For scientific knowledge management systems, the context of information is its warrant for belief, while experiment in relation to theory and hypothesis supplies the criterion of truth. Discourse and social practice (of which it is a part) weave this whole together.

What we must know about scientific assertions is, (a) what warrant (context) is provided by the author through discourse; (b) whether the warrant is valid in the light of other work and its abstraction in theory (also explored through discourse); and (c) 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 organization schemas are therefore typically 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 knowledgebase is developed and maintained according to the SWAN ontology, which is presented here in overview. It is freely accessible on the web [13] and provides a formal basis in OWL [14] 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 prepublication 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. Development of the ontology

3.1. Purpose of the ontology and its implications for the design process

SWAN was designed to enable applications to fulfill a set of use cases in the knowledge ecosystem of biology while reusing artifacts from the ecosystem of web semantics. Its ontology represents the current semantic backbone of the SWAN project.

The purpose of the ontology is:

- To function as the schema of a distributed knowledgebase in Alzheimer Disease
- To link information in that knowledgebase with other information in biomedicine.

The SWAN knowledgebase, with associated software, processes and social networks is intended

- To enable both public and private scientific discourse to be usefully organized in a way which can support divergent viewpoints and become self-perpetuating.
- To highlight gaps, inconsistencies, and missing evidence in the discourse.
- To integrate scientific discourse as it evolves, with standard biological concepts such as genes, proteins, reagents, etc.
- To enable searching and discovery of new research findings across subdisciplines.
- To facilitate scientific discussion and collaboration mediated by the web.

Because it bridges two ecosystems, SWAN's design process is iterative and consultative.

We began our design with a tentative model of discourse and a projected set of use cases and applications, which we then piloted. The application design and information model was vetted and improved in focus groups of Alzheimer Disease researchers. Our ontology was then formalized and over time as we evolved the applications. We also consulted extensively with ontologists participating in the W3C Health Care and Life Sciences Task Force, the Open Biomedical Ontologies Foundry (OBO) and the National Center for Biomedical Ontology (NCBO), to develop and extend SWAN's interoperability with other ontologies in the biology domain.

Enhancements and changes to the ontology are discussed and approved by our software and information architecture team with our biological curators before being translated into a formal representation. This is a key factor in our development, as knowledge creation in SWAN requires scientists to participate in curation effort. Before updating the software tools we solicit approval from the scientists who use the SWAN tools, and ensure that they have a deep understanding of the conceptual framework and approve changes to the tools before we update them. In several key cases we also have consulted with leading scientists at the Massachusetts Alzheimer Disease Research Center to explore implications of proposed changes and to solicit ideas and responses. Further developments in this process are now being planned with the NCBO.

SWAN's ontology was developed in OWL-DL through both the Protégé Ontology Editor and Knowledge Acquisition system [15]