Accepted May 8, 2010 for Bio Ontologies 2010

AO: An Open Annotation Ontology for Science on the Web

Paolo Ciccarese[*]*, Marco Ocana*[**]*, Sudeshna Das*[*]*, and Tim Clark*[*‡]

[*]Harvard Medical School and Massachusetts General Hospital, Boston MA; [**]Balboa Systems, Newton MA

ABSTRACT

We present the Annotation Ontology (AO), an open ontology in OWL for annotating scientific documents on the web. AO supports both human and algorithmic content annotation. It enables "stand-off" (separate) metadata anchored to specific positions in document text by any one of several methods. In AO, the document may be annotated but is not required to be under update control of the annotator. AO contains a provenance model to support versioning, and a set model for specifying groups and containers of annotation.

1 INTRODUCTION

Much current work in biomedical ontologies now focuses on detailed formal classification of objects, functions and processes, using description logics [1-3]. This approach creates a set of fixed categories for searching and navigating ontology-annotated content on the web whether in standard journal publications or in web "collaboratories" [4].

However, we currently lack a robust common set of methods for linking text in new scientific publications to ontological elements, with full annotation provenance. Given such a facility, formal ontologies can serve as schemas for extremely rich stores of metadata on web documents, linking new scientific content across scientific specializations and collaboratories. One fundamental requirement for such methods, if they are to become widely used, would be a formal specification of its metadata. Seminal lines of research in distributed link services [5] and in conceptual open hypermedia [6] have explored this area, without yet to our knowledge publishing an annotation metadata specification meeting requirements for the semantic web.

Not only subject area ontologies, but also a straightforward annotation ontology and a framework for generating annotations with algorithmic assistance, are required, if we are to capture emergent knowledge in new publications, linking the "frozen" consensus thinking embodied in ontologies, across domain boundaries, to the latest discoveries about the natural world most of interest to working scientists. All three elements are needed to successfully expand the collaboratories model across related, linked domains.

We have developed an annotation ontology specifically designed to support content linking in collaboratories. Content in collaboratories has the great advantage of providing a strong focus to the collected, evolving discourse. Specialists accessing material in such a focused web community – such as PD Online [7] (http://pdonlineresearch.org), or Alzforum [8], (http://www.alzforum.org) – will not need to wade through extraneous material on cardiology, drug addiction, hematology, and so forth. Essentially what these communities do is dramatically improve the signal-to-noise ratio for specialists, making the information explosion in science nearly tractable within a given specialty.

Annotation – either marking up contributions with comments, or more importantly, with relevant concepts and entities from biomedical ontologies – provides a technological boost to "strategic reading" for members of such communities [9,10] and selectively breaches established specialist focus boundaries and semantic barriers where required [11].

Existing ontologies and vocabularies which can serve as a basis for such annotation are particularly abundant in the biomedical field and are often expressed in OWL/RDF [12] or in SKOS [13], with OWL/RDF now apparently the most favored option. Subjects for ontological structuring include biological processes, molecular functions, anatomical and cellular structures, tissue and cell types, chemical compounds, and biological entities such as genes and proteins.

We take proteins as a typical example. There are a number of database resources that catalog and identify proteins. UniProt is certainly the most popular but, at the moment, is not available in OWL format (i.e. as a description logic). The PRO Ontology [14] is a project which represents a growing proportion of the content of UniProt and other protein databases as declarations in OWL, and is interoperable with other OBO Foundry ontologies - such as the Sequence Ontology [15] and the Gene Ontology [16] - that provide representations of protein qualities. This interoperability facilitates cross-species comparisons, pathway analysis, disease modeling, and the generation of new hypotheses through data integration and machine reasoning.

Our annotation ontology was motivated by these requirements and use cases. It has also been influenced by an analysis of strengths and weaknesses of earlier work by Swick et al. in the Annotea Project [17], discussed below.

Annotea was developed as a Web-based shared annotation system based on a general-purpose open RDF infrastructure,

[‡] To whom correspondence should be addressed: tim_clark@harvard.edu