

A Marketplace for Ontologies and Ontology-Based Tools and Applications in the Life Sciences

Robin McEntire
GlaxoSmithKline
Robin_A_McEntire@gsk.com

Eric Neumann
Beyond Genomics
eneumann@beyondgenomics.com

Peter Tarczy-Hornoch
University of Washington
pth@u.washington.edu

Prof. Carole Goble
University of Manchester
carole@cs.man.ac.uk

Paula Matuszek
GlaxoSmithKline
Robin_A_McEntire@gsk.com

Robert Stevens
University of Manchester
carole@cs.man.ac.uk

Terence Critchlow
Lawrence Livermore National Lab
critchlow1@llnl.gov

ABSTRACT

This paper describes a strategy for the development of ontologies in the life sciences, tools to support the creation and use of those ontologies, and a framework whereby these ontologies can support the development of commercial applications within the field. At the core of these efforts is the need for an organization that will provide a focus for ontology work that will engage researchers as well as drive forward the commercial aspects of this effort.

Keywords

ontologies, agents, life sciences.

1. Motivation

The number of data sources available and necessary for moving forward in the study of the life sciences is large and diverse and is growing at a remarkable rate. There is also tremendous growth in the field of computational methods and services for manipulating multiple, disparate, distributed, and heterogeneous sources of data. The proliferation of data formats and schemas for life sciences objects and the many methods for accessing or computationally manipulating those objects and services is a serious impediment to the efficient, effective, and scalable use of these data and services. However, the recent developments in web technology, semi-structured data query languages [Abiteboul 1999], data integration systems [Levy 2000], agent-based systems, communication protocols, object-oriented databases, and knowledge-representation technology are at a level of maturity at which they can be leveraged to help solve this problem. It is noteworthy that each of the technologies just mentioned is working, either explicitly or implicitly, on the development of ontologies to facilitate technical progress. Therefore, the key to successfully leveraging these technologies is to foster a marketplace for the exchange of ontologies, ontological tools, and applications that use these ontologies.

An ontology marketplace should meet the following set of objectives:

1. Define a set of services that provide well-characterized, uniform, and consistent access to data/information and services such as SwissProt, Prosite, GenBank, and other current and new data sources.
2. Define the ontologies that describe these lower-level data sources, instances of which will be provided as requested from the set of services defined above.
3. Adopt one language, or a small set of compatible (translatable) languages, for the exchange of these common ontologies.
4. Define ontologies that describe life sciences objects at higher levels of abstraction, and so span the ontologies representing individual data sources. These ontologies will provide representations to be used for integrating lower level ontologies and their data sources and will also serve as guides to workers in the life sciences area for rational definitions of life sciences objects.

5. Provide an environment for the development and exchange of applications in the life sciences that operate over common ontology descriptions.

Common, shareable ontologies provide the field with understandable semantics for the discussion and programmatic computation of life sciences objects. Exchange languages provide a common syntax, which enables the semantic representations to be moved from one site to another without undue, or inaccurate, translation of the definitions of the underlying objects. Well-defined services allow easy programmatic access to life sciences objects and to services that manipulate those objects (and produce other life sciences objects).

2. Terminology

Ontologies are defined in the literature in various ways with varying degrees of formality. One prevailing definition of an ontology is a specification of a conceptualization that is designed for reuse across multiple applications. By conceptualization, we mean a set of concepts, relations, objects, and constraints that define some domain of interest. More plainly we can say that ontologies are specifications of the concepts in a given field, and of the relationships among those concepts. By specification we mean the encoding of the specification to make it concrete, or computationally accessible.

One can argue at length about what is and is not an ontology [Gruber 1993; Guarino 1995]. Our view is that ontologies exist at several levels of complexity:

- A *controlled vocabulary* is an ontology that simply lists a set of terms.
- A *taxonomy* is a set of terms that are arranged into a generalization-specialization hierarchy. A taxonomy does not define attributes of these terms, nor does it define relationships between the terms.
- An *object-oriented database schema* defines a hierarchy of classes, and attributes and relationships of those classes.
- A *knowledge-representation system* based on first-order logic can express all of the preceding relationships, as well as negation, disjunction and more sophisticated constraints and behaviors of classes.

3. The Marketplace

The Ontology Marketplace will provide an environment in which both producers (ontology and application developers) and consumers (biotechs and pharmas) can interact. They can meet and trade with the goal of sharing the work, and the profit, of developing common, shareable ontologies and the tools and applications that use those ontologies. The marketplace will accomplish this duty by providing an environment which will facilitate the development of ontologies and the tools and applications that make use of them. Researchers, academics, and commercial developers will find in the Marketplace the data sources and tools they will need to pursue their own work, plus a venue for introducing their wares to the life sciences community.

This marketplace might also be described as an instantiation of the semantic web [Berners-Lee 2000] specific to the life sciences domain. The notion of moving from a form that is human readable and human understandable to a form that is machine understandable, or a processable web, is much the same as moving toward a machine understandable collection of bioinformatic resources. Technologies such as CORBA and XML can describe the structure of resources and the infrastructure to support access to the resources, but the semantics of these resources must also be described, which requires ontologies. Therefore, the Semantic Web and this collection of bioinformatic resources are both driven by ontologies.

The Marketplace is composed of several elements, which are described in detail below.

3.1 Common Services

There are well over 500 public domain data sources of interest to life sciences researchers. Many of these "data sources" do more than just provide data; they also provide access to a wide range of services. Sequence

homology search engines are a good example. Given the differences in interfaces, syntax, and semantics between sites, there is no practical path for a given researcher or research team to use more than a few without a significant software development effort. Data warehouses, federated systems, and the like help, but only a little. The number of new sources coming online every year, and the number of changes to existing sources, is simply overwhelming.

We picture a genomics world in which scientists, search engines, bioinformatic applications, and soft-bots can browse and execute queries against a wide range of sites, with no significant per-site overhead. Rather than attempting to integrate these sources (thus allowing complex queries against few sites), we advocate providing just enough connective tissue to allow semi-intelligent agents or search engines to execute simplified queries against hundreds of sites. The connective tissue will take the form of a service-level description, or API. This API will provide a consistent look to underlying data sources in order to facilitate communication, and will carry request-specific information in an embedded, common form. The service-level description will support requests both to data in the underlying data source (content-level information) and to domain-specific information about the service itself (metainformation). For example, a user agent (software or human) might ask what kind of ontologies a service supports or what kinds of queries it will respond to. To accomplish this goal we will need to:

1. Enumerate relevant service types for bioinformatics
2. Prioritize services according to those whose availability would provide the most bang for the buck
3. Develop the ontologies that describe each data source
4. Produce service-oriented schemata that provide the "connective tissue" needed to access existing sites and services using a representation neutral format (e.g., ER / OO / UML diagrams [OMG 1997])
5. Use ontologies to support structured, cross-domain, semantically-consistent annotations (common semantic annotations model, CSAM)

The idea behind the common semantic annotations model is to provide a method for organizing and managing high-value scientific (interpretive) information in a way that does not interfere with the "business operations" requirements of an high-throughput, automated data management facility. Data models can be developed based on the throughput and storage needs of a facility, rather than on a "biologically correct" one, such that they are fairly generic and independent of annotative content. Separating the "interpretive" scientific content into the annotation branches therefore requires an ontology to manage the embedded concept types and relations within the annotations. This will permit the necessary shift from free-form text annotations to more machine-readable structures that are based on scientific concepts.

The initial goal is to take small but significant steps forward in the development of key services. We can not hope to end up with descriptions that will be complete enough to describe all the relevant services provided by all web sites of interest to the community - the semantic, syntactic and political challenges are too great. We do believe, however, that by starting with a few simple, well-constructed services, we will provide the base-level semantic and syntactic information necessary to support limited (but currently non-existent!) "1-click" browse and query access to many important services located across a wide range of sites.

3.2 Data Source Ontologies

As mentioned in the section above, life science data sources and services will present their underlying objects in the form of an ontology, and, most likely, in the form of a number of variants on an ontology to allow for differing uses for and views of the underlying objects.

We will promote the development of ontologies that represent specific data sources (traditionally, a *bottom-up* approach to ontology development) and also the development of ontologies that present a higher-level abstraction that represent the view of the world from the perspective of an application developer or the builder of a reasoning engine (traditionally, the *top-down* approach to building ontologies). In fact, the Ontology Marketplace is neutral with respect to bottom-up, top-down, or middle-out approaches to ontology

construction. The Marketplace will allow developers to target the area of greatest interest to them and dive in to the framework at a point most suitable to their needs.

The best way to facilitate the use of ontologies, once they are developed, is to register them in a publicly available source for subsequent download and use by the community. There are already existing web-based mechanisms for doing just this, the most prominent of which is UDDI [UDDI]. Taking advantage of already existing services is a clear advantage, and demonstrates the utility of associating our efforts with those of the web community.

3.3 Ontology Exchange Language

Over the last two decades, the knowledge representation and object-oriented database communities have developed languages that may be used for the expression of semantic database models. These languages share many elements in common, and are exemplified by the frame-knowledge representation systems used in the knowledge representation community. Frame systems have been used in many different bioinformatics projects (RiboWeb [Chen 1997], EcoCyc [Karp 1999], Tambis [Stevens 2000], etc) and provide the necessary representational constructs for modeling ontologies in the life sciences. Furthermore, frame systems have a significant history of use, and provide a stable representational paradigm [Fikes 1985].

In addition, the explosive growth of the world wide web and the languages and tools associated with it are a tremendous force in the current technology infrastructure and cannot be ignored when selecting a language for the exchange of ontologies. The DAML program and its community, which includes membership from the W3C and the Semantic Web Consortium, has taken on the challenge of developing a language for the representation of ontologies for the web. The language developed and recommended by this group is DAML+OIL, which provides the rich semantics of description logics in an XML-based syntax. It is worth noting that the DAML+OIL language had its roots in the XOL language [Karp XOL], which was developed within the bioinformatics community. DAML+OIL is still evolving as it is applied to new problem sets; however, we believe that DAML+OIL represents the current best language for the exchange of ontologies in the life sciences and would note that the Bio-Ontologies Consortium has recommended it as the language to be used for the exchange of ontologies in the life sciences.

3.4 Ontology-based Application Development

The Ontology Marketplace will also support the development of applications in the life sciences. Application development will be significantly eased if descriptions of underlying data sources are provided in a higher-level, object-oriented form. However, the major advantage that the Ontology Marketplace can offer is access to various data sources in a simple fashion. One of the major resource-bottlenecks facing the application development community is the time needed to provide programmatic access to the several, or possibly many, data sources needed to perform adequately to solve problems.

Application developers will be interested in higher level ontologies that provide a more abstract description of life sciences objects, which are not wedded closely to the objects, and ontologies, represented in the databases themselves. There is an inherent mismatch between these high-level descriptions of life science objects and the underlying databases, which were often built for a specific task. The advantage that the Ontology Marketplace offers is that application developers as well as data source ontology developers can build their ontological descriptions in parallel.

Also, there is significant research being conducted in the area of reasoning over bio-information. This area holds much promise, but is currently held back by the dearth of available descriptions for life sciences information. The Ontology Marketplace will provide a test bed for these researchers to explore new algorithms and methods for reasoning over large amounts of life sciences information. In addition, languages such as DAML+OIL are ready-built for reasoning engines, since they naturally allow for knowledge-base construction.

4. Commercial Opportunities

It is the belief of the authors that over time, the well-defined services and the well-understood, easily exchangeable ontologies provided in the Ontology Marketplace will create opportunities for vendors to compete to provide new applications and services as well as new data source services. This will, in turn, drive forward work in technologies (ontology authoring tools, data mediation/integration agents, knowledge representation, and advanced reasoning tools, to name a few) relevant to the life sciences. As we have seen in other technology sectors, the greatest driver for the adoption and use of a technology is a strong marketplace. Commercial organizations within the life sciences community are eager for new applications that operate over large quantities of scientific data, and that also have the ability to integrate with that data appropriate information of other kinds, such as scientific journal articles and patents. For example, there is currently work being done with text mining tools to extract pertinent information from, or provide pointers to relevant data in, large bodies of scientific and competitive intelligence text documents. This process could be significantly aided by the terms and categories that comprise an ontology. In other words, by focusing on the greatest business impact, the Ontology Marketplace will yield the greatest impact on our industry as a whole.

5. Summary

Bioinformatics, cheminformatics, and other research areas in the life sciences are knowledge based disciplines. As such they are highly dependent on large volumes of data of many different kinds. With the capture of data by means of high-throughput technology and the ever-increasing numbers of patents, scientific journal articles, and other text documents, the work of finding relevant information for a given area of research has become a daunting task. Efforts to organize and bring coherence to this massive body of information will rely on ontologies to provide standard terminologies (such as the current GO [Gene Ontology Consortium 2000] effort and MedDRA), semantics (e.g., UMLS) and intelligent cross-relational annotations (CSAM). These ontologies will form the basis of a marketplace to clearly define commercial opportunities in a data-rich life sciences community.

6. References

- Abiteboul, S., Suciu, D., Buneman, P., Data on the Web: From Relations to Semistructured Data and Xml, Morgan Kaufmann Series in Data Management Systems, 1999
- Berners-Lee, Tim, Weaving the Web; the Original Design and Ultimate Destiny of the World Wide Web,
- Chen, R.O., Felciano, R., Altman, R.B., RIBOWEB: linking structural computations to a knowledge base of published experimental data. *Proc Int Conf Intell Syst Mol Biol.* 1997;5:84-7.
- Fikes, R., Kehler, T. 1985. The Role of Frame-Based Representation in Reasoning *Communications of the Association for Computing Machinery*, 28(9):904-920.
- Gene Ontology Consortium, et al, Gene Ontology: tool for the unification of Biology, *Nature Genet.* 25: 25-29, 2000
- Karp, P.D., Chaudhri, V, *XOL Specification*, <http://www.ai.sri.com/pkarp/xol/>.
- Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A., Krummenacker, M. 1999. EcoCyc: Electronic Encyclopedia of *E. coli* Genes and Metabolism. *Nuc. Acids Res.*, Vol 27 No 1, pp 55-58.
- Levy, A.Y., Logic-Based Techniques in Data Integration, Logic Based Artificial Intelligence, Edited by Jack Minker. Kluwer Publishers, 2000
- Object Management Group, 1997. *UML Semantics, Version 1.1*, <ftp://ftp.omg.org/pub/docs/ad/97-08-04.pdf>.
- Stevens, Robert, et al, 2000, TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources *Bioinformatics* pp. 184-186. volume 16, number 2
- UDDI, <http://www.uddi.org/>