*The Knowledge Engineering Review*, Vol. 17:1, 77–80. © 2002, Cambridge University Press
DOI: 10.1017/S0269888902000334 Printed in the United Kingdom

# Ontologies in the life sciences

ROBIN McENTIRE

*GlaxoSmithKline; e-mail: Robin_A_McEntire@gsk.com*

**Abstract**

This paper describes the need for ontologies in the life sciences, some work that has been done already on the development of ontologies and languages for the exchange of life sciences objects, some important problems in the field where ontologies provide critical benefit and some of the organisations that are facilitating this work.

## 1 Introduction

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 *et al.*, 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 being used, either explicitly or implicitly, to develop ontologies with the aim of facilitating technical progress in the life sciences. Therefore the key to successfully leveraging these technologies is to foster the creation and sharing of public-domain ontologies and well-defined services for the exchange of these ontologies. The hope is that the creation of a set of agreed-upon, public-domain ontologies and services will provide support for the development of a set of tools for authoring, viewing and manipulating ontologies and, more importantly, the development of applications that use these ontologies and services.

Common, shareable ontologies provide the field with understandable, and approachable, semantics for the discussion and programmatic computation of life sciences objects. In order to easily exchange the ontological objects among applications there is a need for a single language, or a small set of interchangeable languages, that 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. The Bio-Ontologies Consortium (Bio-Ontology Consortium), an organisation whose goal is to facilitate the development of public-domain ontologies in the life sciences, made a recommendation for use of the DAML + OIL language as the interlingua of choice for exchanging ontologies in the life sciences (McEntire *et al.*, 2000).[10] In addition, well-defined services would allow easy programmatic access to life sciences objects and would also allow manipulation of those objects (to produce other life sciences objects). Stated in another way, a set of common, shared ontologies plus common exchange languages and well-defined services should foster an open environment in which best-of-breed tools and applications may compete.

https://doi.org/10.1017/S0269888902000334 Published online by Cambridge University Press

## 2 Current status of bio information

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.

One can 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), it makes more sense to provide 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 (meta-information). For example, a user agent (software or human) might ask what kind of ontologies a service supports or what kinds of query it will respond to. To effectively implement such an architecture it will be necessary to build ontologies that express the content-level objects, the life sciences entities, that will be moved through the system, and it will also be necessary to develop ontologies that express the service-level descriptions and the communication that takes place among services and/or agents. There has been much work already done on the development of languages to express these service-level descriptions, such as UDDI,[1] DAML-S,[2] ebXML,[3] WSFL[4] and others. What remains is to define the actual services that support application needs within the life sciences community.

Another area of the life sciences field in which ontologies can make a very practical and significant impact is the capture of annotations. Annotations are free-form text comments written by scientists about their work, and are a key source of valuable information in the life sciences arena. These textual annotations are "attached", or associated with, hard data, such as biological assays and other experimental results, and provide the scientist's interpretations of the hard data. There has been discussion in the field about providing a common semantics for these annotations to make them more consistent and, therefore, easier to read and review. The idea behind this common semantic annotations model is to provide a method for organising and managing high-value scientific (interpretive) information in a way that does not interfere with the pragmatic requirements of a high-throughput, automated data management facility as would be found in commercial organisations, such as large pharmaceutical companies or biotechs. Data models can be developed based on the throughput and storage needs of a facility, rather than on the "biologically correct" concepts within the field, so that the models are fairly generic and independent of annotative content. Separating the "interpretive" scientific content into a set of meaningful contexts and/or categories makes information much more visible than when it was "hidden" inside a body of text. The process of extracting information from text 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.

[1] http://www.uddi.org/.
[2] http://www.daml.org/services/daml-s/2001/05/.
[3] http://www.ebxml.org/.
[4] http://xml.coverpages.org/wsfl.html.

https://doi.org/10.1017/S0269888902000334 Published online by Cambridge University Press

## 3 Current ontology development

One of the early efforts to build consistent ontologies, or vocabularies, in the life sciences was by Dr Monica Riley. In 1993 she pioneered an effort to describe the *function* of the genome of Escherichia coli K-12 in a systematic and rigorous manner. The classification system that she developed was very influential and has served as the basis for the description of most sequenced bacterial genomes since then. However, our knowledge of *E. coli* has increased substantially, and it is now widely recognised that the original system is neither sufficient in scope nor sufficiently rich in structure for current needs. Riley's original work is currently being revised using the Gene Ontology resource that has been developed over the last three years by the Gene Ontology Consortium.[5]

The Gene Ontology (GO) is the most successful ontology effort to date within the life sciences domain. The Gene Ontology resource is being developed as a comprehensive basis for the annotation of three orthogonal attributes of gene products: their molecular functions, their roles in biological processes and their cellular or subcellular localisations. The Gene Ontology now contains over 10,000 terms with full definitions for 1800 of these terms. The GO is being developed and used by groups working on genes for mouse, Saccharomyces, Drosophila, Arabidopsis, Schizosaccharomyces pombe, Dictyostelium, Plasmodium (and other parasitic protozoa) and human. In addition, the SWISS-PROT/InterPro team at the European Bioinformatics Institute (EBI) is a member of the GO Consortium, and GO is being used within the commercial realm by several genomics companies for gene annotation, and by several pharmaceutical companies for in-house work. The Gene Ontology is currently expressed in a proprietary language; however, an effort is underway which will express all of GO in the DAML + OIL language, a description logic developed by a joint US-European committee associated with the DARPA DAML project.

In addition to GO, the life sciences community is engaged in the development of a number of other ontologies: MGED, the Microarray Gene Expression Database Working Group;[6] the Tambis Ontology (TaO), for bioinformatics and molecular biology;[7] RiboWeb,[8] an ontology describing ribosomal components, associated data and computations for processing those data. A full list of current ontologies in the life sciences can be found at the Bio-Ontologies Consortium website.[10]

In addition to these consortia and working groups, there are commercial organisations that have created products for the life sciences based on ontologies. Several of the products offered use ontologies to augment text-mining engines for information extraction or document identification. Others have created large ontologies from various knowledge sources, including scientific journal articles, and provide sophisticated query and reasoning over these knowledge bases.

Finally, there are organisations whose purpose it is to create standards or promote the development of common, shareable ontologies and service descriptions. The most prominent of these organisations are the Obejct Management Group's Life Sciences Research (LSR) Domain Task Force[9] and the Bio-Ontologies Consortium.[10] The purpose of the OMG LSR is to standardise models and interfaces for software tools, services, frameworks and components in life sciences research. The objective of the Bio-Ontologies Consortium is the identification and promotion of a practical set of technologies, tools and best practices that will aid in the exchange of concepts, representations and knowledge management in life sciences.

## 4 Conclusion

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

[5] http://www.geneontology.org/.
[6] http://www.cbil.upenn.edu/Ontology/MGED_ontology.html.
[7] http://img.cs.uman.ac.uk/tambis.
[8] http://smi-web.stanford.edu/projects/helix/riboweb.html.
[9] http://www.omg.org/homepages/lsr.
[10] http://www.bioontology.org.

https://doi.org/10.1017/S0269888902000334 Published online by Cambridge University Press

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 organise and bring coherence to this massive body of information will rely on ontologies to provide standard terminologies (such as the current GO effort and MedDRA), semantics (e.g. DAML + OIL[11] or UML (Object Management Group, 2001)) and intelligent cross-relational annotations (CSAM). These ontologies will form the basis of an environment that will clearly define both research and commercial opportunities in a data-rich life sciences community.

## References

Abiteboul, S, Suciu, D and Buneman, P, 1999, *Data on the Web: From Relations to Semistructured Data and XML* Morgan Kaufmann Series in Data Management Systems.

Berners-Lee, T, 1999, *Weaving the Web: the Original Design and Ultimate Destiny of the World Wide Web* Harper San Francisco.

Bio-Ontologies Consortium, `http://www.bioontology.org/`.

Chen, RO, Felciano, R, Altman, RB, "RIBOWEB: linking structural computations to a knowledge base of published experimental data" *Proceedings of the International Conference on Intelligent Systems for Molecular Biology 1997* 84–87.

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*., 2000, "Gene Ontology: tool for the unification of Biology" *Nature Genetics* **25** 25–29.

Karp, PD and Chaudhri, V, *XOL Specification*, `http://www.ai.sri.com/pkarp/xol/`.

Karp, P, Riley, M, Paley, S, Pellegrini-Toole, A and Krummenacker, M, 1999, "EcoCyc: electronic encyclopedia of E. coli genes and metabolism" Nuc. Acids Res. 27(1) 55–58.

Levy, AY, 2000, "Logic-based techniques in data integration" in J Minker (ed.) Logic Based Artificial Intelligence Kluwer Publishers.

McEntire, R, et al., 2000, "An evaluation of ontology exchange languages for bioinformatics" Proceedings of International Conference on Intelligent Systems for Molecular Biology 2000.

Object Management Group, 2001, Unified Modeling Language (UML), Version 1.4, `http://www.omg.org/technology/documents/formal/uml.htm`.

Stevens, R et al., 2000, "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources" Bioinformatics 16(2) 184–186.

[11] `http://www.daml.org/2001/03/daml+oil-index`.

https://doi.org/10.1017/S0269888902000334 Published online by Cambridge University Press