

National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge

DANIEL L. RUBIN,¹ SUZANNA E. LEWIS,² CHRIS J. MUNGALL,^{3,10} SIMA MISRA,²
MONTE WESTERFIELD,⁴ MICHAEL ASHBURNER,⁵ IDA SIM,⁶
CHRISTOPHER G. CHUTE,⁷ HAROLD SOLBRIG,⁷ MARGARET-ANNE STOREY,⁸
BARRY SMITH,⁹ JOHN DAY-RICHTER,³ NATALYA F. NOY,¹
and MARK A. MUSEN¹

ABSTRACT

The National Center for Biomedical Ontology is a consortium that comprises leading informaticians, biologists, clinicians, and ontologists, funded by the National Institutes of Health (NIH) Roadmap, to develop innovative technology and methods that allow scientists to record, manage, and disseminate biomedical information and knowledge in machine-processable form. The goals of the Center are (1) to help unify the divergent and isolated efforts in ontology development by promoting high quality open-source, standards-based tools to create, manage, and use ontologies, (2) to create new software tools so that scientists can use ontologies to annotate and analyze biomedical data, (3) to provide a national resource for the ongoing evaluation, integration, and evolution of biomedical ontologies and associated tools and theories in the context of driving biomedical projects (DBPs), and (4) to disseminate the tools and resources of the Center and to identify, evaluate, and communicate best practices of ontology development to the biomedical community. Through the research activities within the Center, collaborations with the DBPs, and interactions with the biomedical community, our goal is to help scientists to work more effectively in the e-science paradigm, enhancing experiment design, experiment execution, data analysis, information synthesis, hypothesis generation and testing, and understand human disease.

This paper is part of the special issue of OMICS on data standards.

¹Stanford Medical Informatics, Stanford University, Stanford, California.

²Life Sciences Division, Lawrence Berkeley National Laboratories, Berkeley, California.

³Department of Molecular and Cell Biology, University of California, Berkeley, California.

⁴Department of Biology, University of Oregon, Portland, Oregon.

⁵Department of Genetics, University of Cambridge, United Kingdom.

⁶Department of Medicine, University of California, San Francisco, California.

⁷Department of Biomedical Informatics, Mayo Clinic, Rochester, Minnesota.

⁸Department of Computer Science, University of Victoria, Victoria, Canada.

⁹Department of Philosophy, University at Buffalo, Buffalo, New York.

¹⁰Howard Hughes Medical Institute.

INTRODUCTION

THE ADVENT of the Web and modern high-throughput techniques is having a major impact on the manner in which biomedical research is conducted. The explosive growth in biomedical data being generated by high-throughput experimental techniques has created tremendous opportunities for discovery by mining these data. Scientists increasingly rely on large online databases as a source of knowledge and data for exploration of new hypotheses. While this new e-science paradigm brings with it tremendous opportunities, it also poses significant challenges, because researchers are now confronted with huge data sets, bringing the urgent need for both people and computers to be able to make sense of massive quantities of heterogeneous data.

Ontologies—collections of formal, machine-processable and human-interpretable representations of the entities, and the relations among those entities, within a defined application domain—are helping researchers manage the information explosion by providing explicit descriptions of biomedical entities and an approach to annotating, analyzing the results of clinical and scientific research. Ontologies are useful because they provide regimentations of terminology that can support the reusability and integration of data and thereby support the development of useful systems for purposes such as decision support, data annotation, information retrieval, and natural-language processing.

The rising interest in the application of ontologies in biomedical research has created new challenges. The number of ontologies is proliferating, the domains covered by these ontologies are overlapping, and the modeling practices in building ontologies and ontology quality are subject to different sorts of variation. There are many existing ontologies used to describe biomedicine and that are used to annotate the data deriving from biological research and clinical practice. A methodology needs to be created to unify these ontologies, to align them, and to help biomedical researchers access them for use in annotating experimental data. At the same time, repositories that store annotations on experimental data and tools to enable researchers to analyze data in the context of these annotations are also required.

As interest in ontologies is rising in the biomedical community, there is an increasing need to provide researchers a portal to access and use existing ontologies. There is also a need to help ontology engineers recognize opportunities to reuse and extend existing ontologies and to teach ontology authors how to create new and robust ontologies. In our view, a collaboration among biomedical researchers, informaticians, and ontologists is needed in order to address these needs and to enable the community to create, maintain and use ontologies effectively.

We recently created the National Center for Biomedical Ontology, a National Center for Biomedical Computing of the NIH Roadmap. Our Center is a consortium of informaticians, biologists, clinicians, and ontologists that is developing technology designed to allow scientists to create, disseminate, and manage biomedical knowledge in machine-processable form using ontologies as the foundation for this task. Our goal is to assist researchers in performing their knowledge-intensive work, by providing tools and a resource enabling them to access, review, and integrate disparate knowledge and information resources in all aspects of biomedical investigation and clinical practice. A major focus of our work involves the use of biomedical ontologies to aid in the management and analysis of data derived from experiments, such as genetic sequence and phenotype data. We collaborate with projects that will benefit from the center's resources and that will stimulate the Center to create new tools and resources and propel our ongoing research and development activities.

The Center seeks to provide tools and methods to enhance the use of ontologies throughout biomedicine. To meet this objective, the Center is developing two major repositories of biomedical content: (1) Open Biomedical Ontologies (OBO), a comprehensive, online library of open-content ontologies and controlled terminologies; and (2) Open Biomedical Data (OBD), a database resource that will allow expert scientists to archive experimental data that is fully described (annotated) using the OBO ontologies and terminologies (Fig. 1). The Center is also developing software tools to enable researchers to access and use these resources effectively. The biomedical research community will access OBO and OBD via a system called BioPortal—a Web site and a suite of Web services that will enable both human users and computer-based agents to access the rich content that the Center and its collaborators will curate (Fig. 1). The Center is also developing a suite of tools (1) to permit ontology creators to align and merge ontologies and to evaluate

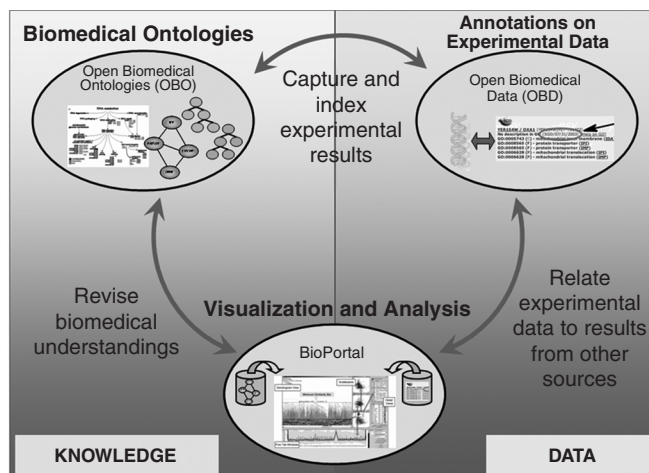


FIG. 1. Overview of ontology and annotation resources. (1) An integrated library of open biomedical ontologies (OBO). (2) A database of annotations on experimental data summarizing key attributes such as anatomy, phenotype, and genetic features (OBD). (3) A Web portal for accessing, visualizing, and analyzing experimental data informed using ontologies and annotations (BioPortal). These components inform one another: ontologies provide terms for data annotation; the annotations provide a foundation for data query and analysis; the results of these studies inform revisions to the ontologies.

their quality, (2) to enable researchers to view and analyze experimental data in the context of ontologies that were used to describe that data, and (3) to permit application developers to search, navigate, and visualize ontologies, and to access the Center's tools and resources through its Web services.

In the following sections, we discuss the current usage and unsolved challenges of ontologies in biomedical research, and we describe our approach to integrating and accessing diverse biomedical ontologies and creating comprehensive resources of annotations on experimental data. Our methods build on initial successful efforts of the GO consortium, though enhanced to address the lessons learned by their early efforts. We are leveraging our cumulated extensive experience in creating tools such as the Protégé and OBO-EDIT open-source ontology development platforms, terminology access and storage technology such as LexGrid (<http://informatics.mayo.edu/LexGrid/index.php?page=>), and ontology visualization components such as Jambalaya (Storey et al., 2002). We are also creating a public repository to store annotations on experimental data that use these source ontologies. Our tools and methods are driven by the needs of bench biologists and clinicians, and we will describe the DBPs that propel our technology. Finally, because exploiting the full potential of ontologies depends on engaging participation of the biomedical community in their development and evolution, we will describe our efforts in education and outreach to build and support communities of ontology content creators and application developers that will create future tools that build on the Center's technological foundation. Our ultimate goal is to provide researchers with tools and resources that will help them use ontologies to relate disparate information resources to leverage the knowledge contained in model organism databases to understand human disease.

ONTOLOGIES AND BIOMEDICAL DATA

Ontologies are currently at the Center of two major activities in biomedical research (Fig. 1). First, communities of researchers are creating and maintaining biomedical ontologies to represent the different types of entities and relations in different domains of biomedicine (ontology content curation). Second, biomedical experimentalists are using ontologies to annotate their data, enabling their data to be integrated with other researchers' data and permitting cross-species analyses through the experimental data annotations. Both ontology content curation and experimental data annotation present significant challenges that our Center is aiming to resolve.

Ontology content curation

Many individual groups and consortia have made important and substantial contributions to the development of biomedical ontologies; however, they have worked in an uncoordinated manner, and the field is fragmented. Numerous ontologies, vocabularies, and databases have been created, but they are not interoperable; they provide scientists with a confusing and conflicting array of terms to choose from when annotating their experimental data, and they require researchers to access and integrate disparate sources of information manually. These conditions create a major barrier to accessing and using expanding data repositories effectively. We will describe efforts our Center is undertaking to address these problems by unifying and relating biomedical ontologies.

Experimental data annotation

Ontologies are being created and used routinely in a rapidly growing number of online biomedical resources—for example, GO (Harris et al., 2004), SNOMED (Dolin et al., 2001), MGED Ontology (Whetzel et al., 2006), NCI Thesaurus (de Coronado et al., 2004), FuGO (Whetzel et al., 2006), and BioPAX (www.biopax.org/). Such resources have become valuable to large communities of scientists who can mine experimental data annotated using ontologies. However, the existing resources are largely confined to containing annotations from a single ontology. Researchers are often interested in performing cross-species analysis, using homology to discover new potential markers for disease or target for therapy; thus, a resource that combines data annotations from multiple organisms is desirable. In addition, there are no existing resources that relate newer ontologies being created, such as phenotype ontologies, to annotated experimental data. We will discuss our plans and initial work to create a resource that integrates annotated data across biomedical domains and links it with the sources ontologies to support cross-species analysis.

OPEN BIOMEDICAL ONTOLOGIES (OBO)

Despite the overwhelming importance of controlled terminologies and ontologies to biomedical research and to clinical practice, the world of terminologies and ontologies is chaotic at best. Terminological resources are created by a wide range of organizations, professional societies, and individual laboratories. Developers use different styles of defining terms, different knowledge-representation systems for encoding semantics, different conventions for declaring relationships, different serialization formats for writing concept definitions out to files, and different interfaces for enabling application programs to query concepts and relationships among concepts.

Because efforts in ontology development have thus far been uncoordinated and because individual ontologies are spread widely across the Internet in various degrees of formal coherence, a critical research focus of the Center is to unify and integrate ontologies and to raise the level of ontology design in ways that will provide biomedical scientists with the high quality shared ontology resources they need to annotate and access experimental data. Our goal is also to enhance the ability of biomedical scientists to analyze the resultant data by providing visualization and reasoning services (Fig. 1).

The Center is building a new version of OBO, a large ontology library that is being seeded using the entries in the existing OBO site on SourceForge (<http://obo.sourceforge.net/>). The SourceForge OBO has brought together many of the common biomedical ontologies and made them available for download in their native formats. These formats include OBO-EDIT, Protégé frames, and OWL. However, they can currently be accessed only in the tool that was used to create them (e.g., OBO-EDIT, Protégé frames, and Protégé-OWL). No platform exists to read all these formats and give users access to the contents of the different ontologies. No functionality exists to search for individual terms, or to align and map related ontologies.

To build the new OBO resource that is required to solve this problem, the Center is performing research and development in the following key areas, which will be described in the following sections:

- Methods for indexing and categorizing ontologies
- Methods for relating different ontologies and terminologies to one another and for creating mappings among them
- Methods for visualizing ontology libraries and their collective contents
- Semantic Web technology enabling access to the virtual library of OBO ontologies

Indexing ontologies

The Center's new OBO resource will contain a very broad range of biomedical ontologies and terminologies. It is impractical to develop a single ontology that will integrate all ontologies contributed to OBO, or to expect all or even a large fraction of the ontologies to use a similar set of properties and distinctions. While many ontologies will use a variety of syntaxes—such as OBO format, Protégé, RDF, and OWL—we need a common representation of the terminological aspect of ontologies (represent the entities in ontologies as terms) to support functionality such as search and retrieval of the entities in ontologies meeting user needs. Terminological aspects encompass the representation of terms (synonyms, terms in different languages and lexical variations), their relationships to types of biological entities and to the relations between them.

Our approach to indexing the diverse ontology content in OBO is to first map biomedical entities in the ontologies to vocabulary terms, and second, to define a shared model about how the terminologies are represented. To that end, we are adopting the LexGrid model, which draws from the communities engaged in building Description-Logic ontologies, widely used clinical terminologies, and the heritage of terminology services over the past decade. While not a blind superset of every model encountered, it reasonably abstracts functional similarity across terminology models to achieve a pragmatically parsimonious but broadly generalizable model of how terminologies and ontologies can be formatted, aligned and used in annotations.

Terminology services enable uniform access to terminologies and ontologies. Terminology services such as the Object Management Group (OMG) Terminology Query Service (Solbrig et al., 1998) and HL7's Common Terminology Services (CTS; <http://informatics.mayo.edu/LexGrid/index.php?page=ctsspecdetail>) specification were developed as Application Programming Interface (API) specifications that are intended to describe the basic functionality needed by software implementations to query and access terminological content. We will be able to exploit the LexGrid terminology services to access a broad range of terminologies and ontologies stored in OBO. We will also be able to provide essential ontology access functionality such as term indexing and search. The LexGrid platform is particularly powerful in terms of search, supporting several varieties of search, including an edit-distance-based approximate search and a search based on metaphone patterns and synonymy.

Ontology alignment and difference

Some OBO ontologies relate to overlapping or even identical domains. For example, there are several different ontologies describing anatomy, even within single species. An important functional requirement for making the OBO library of ontologies useful to the biomedical community is to align related ontologies in such a way as to enable users to map-related terms in a coherent fashion.

The PROMPT suite of tools are extensions to the Protégé ontology platform that provides ontology alignment and merging services. Protégé (Noy et al., 2003) is an open-source ontology development platform having a plug-in architecture, enabling it to access a variety of ontologies in different formats and providing a platform to build advanced ontology analysis tools such as PROMPT. PROMPT can identify identical or similar entities in ontologies that should be mapped between ontologies. The tool uses the structure of the topology of entities in an ontology and relations among them, as well the names of the entities themselves (Fig. 2). The PROMPT tools also provide the initial heuristics for comparing any two ontologies.

We are using PROMPT to analyze the OBO ontologies, storing the relationships among the ontologies and terminologies. We are developing GUI interfaces to PROMPT to provide the interface to display mappings. In order to use PROMPT, the OBO ontologies are imported into Protégé. There are many plug-ins

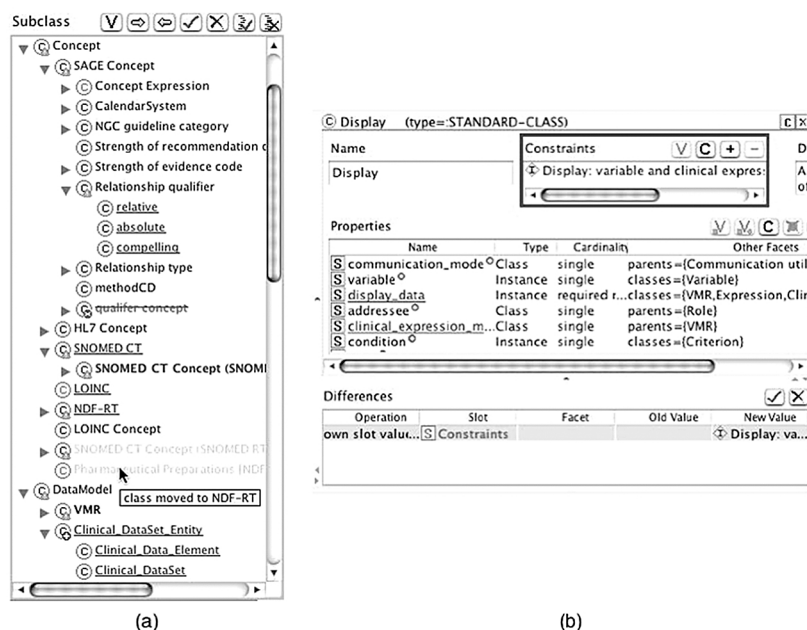


FIG. 2. Displaying ontology differences in PROMPT. (a) Class hierarchy changes. Different styles represent different types of changes: Added classes are underlined; deleted classes are crossed out; moved classes are grayed out in their old positions and appear in bold in their new ones; tooltips provide additional information. (b) Individual class changes.

to Protégé for importing different ontology formats, and we will use these plug-ins to access the diverse OBO ontology content. For example, a plug-in has recently been developed to access ontologies stored in OBO-EDIT (Gennari, 2005). An example of an OBO-EDIT ontology imported into Protégé is shown in Figure 3. Other workers have imported the Gene Ontology into Protégé (Yeh et al., 2003), and work is ongoing to enhance the Gene Ontology with using logical definitions expressed in OWL syntax (Mungall, 2004; Wroe et al., 2003). The MGED ontology is represented in the Web Ontology Language (OWL), and it can be read into Protégé using the OWL plug-in (Knublauch et al., 2004).

We have built an extension to Protégé to read OBO-EDIT ontologies and to translate them into an ontology in Protégé (comprising classes, slots, instances, and slot values; Fig. 3). Just as the LexGrid technology will provide terminological indexing and search services for ontologies in OBO, the Protégé platform will provide alignment and difference-detection services to relate OBO ontologies. In addition, we anticipate exploiting other existing Protégé extensions to use with OBO ontologies such as ontology visualization and Web service interfaces to OBO.

Ontology visualization

In order to make the OBO ontologies accessible to a broad audience of users in the community through a common interface, we are building a Web portal (BioPortal). In addition to providing access to OBO content, BioPortal will provide a single entry point to access all of the Center's resources and tools. For OBO ontologies, BioPortal will provide a graphical user interface and multiple different ways for users to browse and search the ontologies.

In addition to providing basic tree-based ontology navigation, BioPortal will provide more advanced paradigms of ontology visualization using Protégé extensions. Jambalaya is a Protégé plug-in that provides an extensible, flexible, and scalable visualization environment for exploring, navigating, and understanding ontologies. Figure 4 shows how concepts and relationships in Jambalaya are represented using a graph metaphor. Classes and instances are represented as nodes in a graph; different types may be distinguished

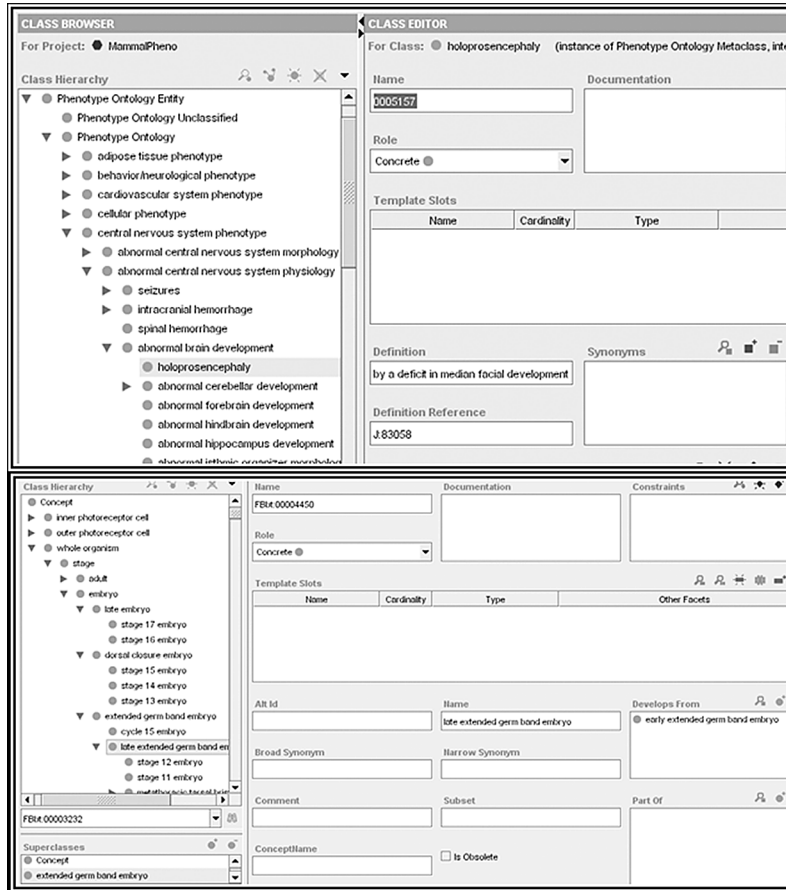


FIG. 3. OBO ontologies in two different formats accessed using Protégé. **(Top)** The mammalian phenotype ontology (originally in OBO-EDIT format and imported using the OBO-EDIT extension to Protégé). **(Bottom)** The Drosophila anatomy ontology (originally in OBO-EDIT format and imported with the OBO-EDIT extension to Protégé). The hierarchy of classes appears on the left in the Protégé GUI, and the values of slots for a selected class appear on the right. The contents of both ontologies are now accessible in the same common format.

using different color hues. Directed edges (arcs) are used to show relationships between classes and instances, such as *is-a* relationships between classes in the concept hierarchy, instance-of relationships relating instances to classes, and slot relationships between classes and instances.

It is also possible to use other user defined slot types for nesting nodes. For example, in an anatomy ontology, the user defined *part-of* relationship may be a more appropriate relationship for nesting nodes than the *is-a* relationship. The nested views are an important aspect of this tool, as nesting is one aspect that improves the scalability of the views. Users can drill in and view the Protégé forms at multiple levels of abstraction and edit directly within the zoomable view used by Jambalaya. The user can also use hyperlinks within the forms to navigate around the graphical view. The navigation mechanism, combined with the nested graphical view, helps users form a mental map of an unfamiliar ontology. Jambalaya was specifically designed to support navigation and comprehension of large ontologies. It has support for showing multiple views that can be customized, to a certain extent, according to a user's needs and user tasks.

In addition to implementing the Jambalaya visualization methodology, we will be creating other ontology visualization tools, driven by the needs of our biological projects, such as image-based ontology navigation and ontology alignment visualization. These tools will be accessible through the BioPortal.

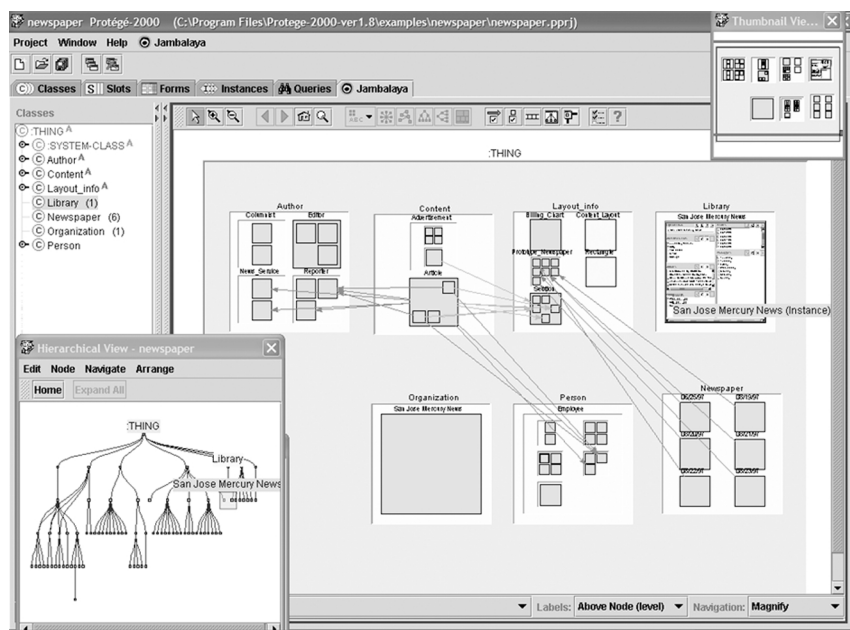


FIG. 4. The Jambalaya plug-in for Protégé displays an ontology using multiple views, including a main view that displays the classes and instances as a nested zoomable graph, supported by Protégé's class tree widget (**upper left**), a flat tree view (**bottom left**), and a "thumbnail" overview (**upper right**).

Semantic Web technology

In order for the rich ontology content in OBO to be maximally useful, a diversity of Web applications in cyberspace will need to be able to access and refer to ontology content. Furthermore, many biomedical resources, such as model organism databases, are actively annotating their data and curating the biomedical literature using terms from OBO ontologies. Such distributed applications and resources require a technology to refer to OBO ontologies, and the Semantic Web provides a solid foundation to build such functionality.

We are developing Web services to expose the core BioPortal functionality to agents in cyberspace. Our methodology will provide both human users and intelligent software agents with comprehensive access to the virtual library of ontologies in OBO. We hope to encourage use of this platform to create novel and intelligent applications for using these ontologies and for analyzing annotated data in OBD. By providing access to the center's ontologies and data through Web services, we hope to encourage development of a broad range of semantic Web applications that leverage the center's resources as well as other relevant biomedical knowledge available on the Internet.

OPEN BIOMEDICAL DATA (OBD)

The OBO ontologies in isolation have limited interest; their particular value is as a structured knowledge resource to be used for annotating biomedical data, and to create explicit representations of biomedical knowledge that humans and computers can understand and process. However, the current tools and approaches for using ontologies for annotating data are cumbersome and costly, and our goal is to create tools to streamline this process and exploit the value of these annotations analytically to formulate new hypotheses from large-scale results in different species.

The center is developing methodologies and tools for annotating experimental data with terms from OBO ontologies. The center is also creating a resource to consolidate these annotations, Open Biomedical Data (OBD), a *database resource* for storing, visualizing, and analyzing the ontology-based annotations that are

linked to primary experimental data. We are also creating tools to populate OBD with the annotations on data, and developing methods for browsing, visualizing, and analyzing those annotations for hypothesis generation.

Our specific tasks in creating the OBD resource include creating the following tools and resources:

- Tools to capture, and to describe rigorously, experimental data using ontologies
- Methods to reconcile changes in annotations and to update them as the source ontologies change
- Resources and design methods to help biomedical investigators to store, view, and compare annotations of biomedical research data
- Algorithms and tools to permit scientists to link and compare annotated data across organisms, leveraging cross-species homology to learn about genes and diseases in humans based on data in model organisms

Data annotation tools

A significant problem with current annotation methods is the complete dissociation of the process of data annotation from the process of building ontologies and terminologies in the first place. Investigators, examining a piece of new information, will very often find that a completely new annotation term is needed. Under present conditions, they must enter into a negotiation with ontology-content personnel in order to revise or extend the ontology. These negotiations can be lengthy and involve multiple groups and review committees. At the same time, it is the biomedical scientist who is annotating the data who will ultimately be called upon to provide the information needed to extend the ontology. This lengthy negotiation process interrupts the natural course of annotation that would allow investigators to use the terms that they need in a timely manner.

To address this problem, we are incorporating into our annotation tools the functionality that will make this fundamental dialog more efficient. We are extending the OBO-EDIT ontology editing tool to enable the rapid creation of data annotations. This new tool will enable users to suggest new ontology terms as they perform data annotation, which will streamline the annotation process. Furthermore, the annotation tool will be integrated into the OBO platform, providing efficient access to the relevant ontologies and new versions of those ontologies.


Data annotation resources

FlyBase (Drysdale et al., 2005) and the ZFIN database (Sprague et al., 2003) are model-organism databases that store experimental and computed data related to their respective organisms. These two projects have recently begun annotating experimental data using ontologies in OBO that provide detailed information about anatomy or phenotype. However, no public resources have yet been developed to allow users to view or query these annotations. We therefore have built a prototype of a resource to allow users to browse these annotations, called Open Biological Data (OBD).

A screen shot of a prototype of OBD is shown in Figure 5. The annotations contained within OBD are derived from ontology annotations that were originally created by the content curators of the contributing model organism databases, FlyBase and ZFIN. The curators of these databases have created these annotations by reading the published literature and by annotating genes and alleles with the appropriate ontology terms to describe those genes and alleles.

The user can browse the annotations in OBD and search OBD for genes and alleles that contain particular patterns of annotations. For example, one can view all annotations on a particular allele to identify biomedical knowledge pertinent to that allele (Fig. 5, left). Figure 5 demonstrates that the selected allele is associated with lethal viability of the embryo, abnormal embryonic head, and morphogenesis of embryonic epithelium, among other phenotypes. Alternatively, the user can query OBD for genes/alleles containing particular annotations, such as those associated with abnormal embryonic head (Fig. 5). Queries such as the latter could be very helpful in identifying the state of genetic knowledge pertaining to particular phenotypes.

Our approach to ontologies and experimental data is notable in terms of combining efforts to unify ontologies (OBO) with a resource that compiles ontology annotations on experimental data (OBD). Related efforts have focused on either building particular domain ontologies (Harris et al., 2004; Ashburner et al.,




DAG | DB Statistics

Search for text: in phenotype data Go

Total Phenotype Characters: 165

Showing: 1..50 | 51..100 | 101..150 | 151..165

| allele | entity | attribute | value |
|-------------|-------------------------|---------------------|---------------|
| FBai0145168 | arista | Shape | Abnormal |
| FBai0145168 | actin filament | Shape | Irregular |
| FBai0145168 | actin filament | **NO TERM** | **NO TERM** |
| FBai0145168 | microtubule | **NO TERM** | **NO TERM** |
| FBai0033484 | **NO TERM** | Viability | Lethal |
| FBai0033484 | eclosion | Process | Arrested |
| FBai0033484 | development | FloodingSensitivity | Slow |
| FBai0033484 | wing morphogenesis | Process | Arrested |
| FBai0033484 | adult behavior | Behavioral_activity | Uncoordinated |
| FBai0145168 | **NO TERM** | Shape | Branched |
| FBai0033484 | longitudinal connective | Process | Arrested |
| FBai0033484 | sensory epithelial cell | Process | Arrested |



DAG | DB Statistics

Search for text: in phenotype data

Allele: Ca- α 1D^{AR66}

ID FBai0063160
Gene Ca- α 1D
Organism Drosophila melanogaster (fruit fly)

| entity | attribute | value |
|---------------------------------------|-------------|----------|
| embryo | Viability | Lethal |
| embryo | Viability | Lethal |
| embryo | Viability | Lethal |
| embryonic head | Qualitative | Abnormal |
| embryonic head | Qualitative | Abnormal |
| embryonic head | Qualitative | Abnormal |
| morphogenesis of embryonic epithelium | Process | Abnormal |
| morphogenesis of embryonic epithelium | Process | Abnormal |
| morphogenesis of embryonic epithelium | Process | Abnormal |
| embryonic dorsal epidermis | Qualitative | Abnormal |
| embryonic dorsal epidermis | Qualitative | Abnormal |
| embryonic dorsal epidermis | Qualitative | Abnormal |
| flizkorper | Process | Normal |
| flizkorper | Process | Normal |
| flizkorper | Process | Normal |

FIG. 5. Open Biomedical Data (OBD). This resource collects annotations on experimental data using OBO ontologies. **(Left)** Ontology annotations on alleles. The annotations consist of entities, attributes, and/or values (EAV). **(Right)** Detailed view showing all annotations on a particular allele in the EAV format.

2000; Sheth et al., 2004) or on creating model organism databases containing annotations (Westerfield et al., 1999). By unifying the breadth of knowledge resources, we hope to enable scientists to use cross-species knowledge to gain new insights into the basis for human disease.

DRIVING BIOLOGICAL PROJECTS

While our center is undertaking many biocomputational research efforts internally, it is vital for the success of our endeavor that we collaborate with biomedical researchers to stimulate the development of new tools and ensure that our efforts are relevant to their scientific needs. Consequently, a major objective of the Center is to reach out to the scientific community to provide tools and methodologies that will promote biomedical research, and to work with these scientists to drive future development of the center's resources and technologies. The common theme among these projects is the need to use ontologies to describe the particular biomedical knowledge in their domain, to create explicit descriptions (annotations) of their data, and to analyze the data using those annotations and the ontologies from which they are drawn to glean biomedical insights or new hypotheses.

The Center currently supports three driving biological projects that serve as testbeds for the Center technology and provide feedback on all activities within the Center. We shall seek additional collaborations in the future, covering a broader scope of biomedicine and domains of experimental investigation.

Linking mutations in Drosophila to human disease

In this project, researchers are annotating mutations of *Drosophila*, giving priority to those genes whose human homologs are associated with disease and with zebrafish orthologs, applying a Phenotype and Trait Ontology (PATO) that currently is undergoing development in the biomedical community. The investigators are using PATO to curate the human homologs of the annotated *Drosophila* genes, using records from the Online Mendelian Inheritance in Man (OMIM) database as the main informative source for this curation.

Relating zebrafish phenotypes to human disease genes

Researchers affiliated with the ZFIN database are developing PATO to provide annotations for phenotypes resulting from zebrafish gene mutations. Together with the *Drosophila* group, they are using PATO to identify human orthologs of the annotated zebrafish genes they curate, using records from the Online Mendelian Inheritance in Man (OMIM) database.

Analyzing evidence in HIV clinical trials

In this project, researchers are developing a principled way to structure, summarize, and visualize the primary evidence from certain HIV/AIDS clinical trials by creating an ontology describing clinical trial data and producing annotations using those ontologies for clinical-trial data on mother-to-child-transmission and structured treatment interruption. The goal will be to use the Center's resources to help determine whether, how, when, and to whom to administer preventive strategies, in hopes of decreasing the burden of HIV/AIDS worldwide.

COORDINATION AND DISSEMINATION OF BEST PRACTICES IN ONTOLOGY DEVELOPMENT AND USE

An important activity of the center beyond tool and resource development involves formulating, testing, and communicating best practices of ontology development to biomedical scientists in ways designed to advance the creation of high-quality shared ontologies and coordination in the development and use of such ontologies through the OBO library and through training in the Center's technologies. By promoting adherence to rigorous logically-coherent ontology design practices, we can ensure that both ontologies and

the data annotated in their terms can be integrated in ways that can support automatic reasoning and error-checking in addition to satisfying existing needs in information retrieval and terminology regimentation.

To this end, our Center hosts highly interactive workshops and tutorials, where Center staff participate in educational activities designed to lead to enhanced ontology content and more usable and useful ontology resources for the biomedical community. This year we are hosting workshops in ontology development for biomedical imaging (spanning molecular level to clinical imaging), phenotypes, and diseases.

CHALLENGES AND SUCCESSES

There are many challenges in embarking on our ambitious mission, both on a technical and cultural level. From the technical perspective, it could be extremely difficult to unify the breadth of existing ontologies, given that they exhibit such wide variations in content, expressivity, and quality. The first challenge is that of reconciling the varying formats for storing ontologies. At present, OBO is a simple posting of a collection of ontologies and mark-up language descriptions. While this is a step in the right direction, it is still an incomplete solution.

Biomedical researchers often discover ontology websites only through word of mouth, and, once there, they are unable to evaluate the content and relate those ontologies to other existing ontologies. Our Center is already addressing this problem by integrating the OBO ontologies within the framework of a single representation. The ontologies formulated within this framework will be required, by degrees, to satisfy more stringent criteria assuring quality and consistency (Smith et al., 2005). We are finding the LexGrid and Protégé technologies very useful, and we are optimistic that these technologies will enable the Center to make the diverse OBO ontology content more accessible and useful to the biomedical community.

Another technical challenge is to make the contents of OBO ontologies and tools to view and analyze them accessible to researchers through a single simple interface. We are creating BioPortal for this purpose. The chief design goal of BioPortal is to make ontologies available online and to simplify ontology access and use. We have already created solutions to problems such as ontology difference, alignment, versioning, and search. BioPortal will access this suite of the Center's ontology tools through components already successfully implemented in Protégé.

Visualization and navigation of large ontologies poses another important technical challenge. Tree-based visualization methods work well for small ontologies, but they are impractical for large ontologies. Through components such as Jambalaya, we anticipate being able to offer several alternative paradigms to help ontologists navigate large ontologies and view complex relationships. In this regard, it is beneficial to have close collaborations with our DBPs, as they provide an active community of users who are committed to working with our Center to create tools and methods to enable their work and to test the quality of their results.

Managing the vast amount of experimental data (OBD) that is annotated with terms from the OBO ontologies and making it useful to biomedical researchers is also technically challenging. Most biomedical researchers who will be interested in accessing the center's resources to help them tackle their biomedical problems will be very interested in using the power of ontologies to analyze experimental data and relate their data to results in other organisms, even if they have no particular interest in the ontologies themselves. In fact, it would be beneficial to hide the ontologies from such researchers altogether.

In addition to the technical difficulties, there are cultural challenges. Our center brings together three different scholarly domains: (1) biomedical researchers and clinicians, (2) computer scientists and bioinformatics specialists, and (3) philosophers and logicians specializing in ontology research. While their shared interest in ontologies produces many synergies, their different backgrounds, interests, and scientific perspectives can cause potential conflicts.

For example, biomedical researchers are using ontologies to enable and streamline their work, and they have a need for rapid development and deployment of ontologies. At the same time, early versions of their ontologies often have problems such as inconsistent definitions of relationships, inconsistency with other existing ontologies, and poorly defined classes, which is problematic for ontologists who are encouraging the community to coalesce on a small set of orthogonal reference ontologies that have been carefully constructed and robustly evaluated. Bioinformaticians are extending existing ontologies or creating their own

customized application ontologies to serve the needs of biomedical researchers, while computer scientists align existing ontologies in an attempt to unify the overlapping efforts and converge toward a small number of robust and powerful reference ontologies.

In addition to the intersections among these communities interested in creating ontologies, there are interactions with the community of ontology users—people who wish to partake of the benefits bestowed by ontologies but who would prefer to be shielded from the ontological details that consume the computer scientists and philosophers. Thus, we anticipate two distinct groups of users of the Center's resources: those who create ontologies and who will be interested in ontology quality and in the Center's visualization, alignment, reasoning, and authoring tools, and those who wish to access the Center's annotated data to explore scientific hypotheses without needing to become an ontology expert.

We have planned to cope with these different perspectives by developing customized views within BioPortal to provide access to the tools and visualizations appropriate to the type of user. The bench scientist will see summaries of OBD phenotype results and will be able to query using searches transparently guided by ontologies without needing to know they are being used (Fig. 5). Likewise, ontology authors and computer scientists will access and visualize the ontologies themselves as well as the ontology management tools to help them develop and improve the ontologies (Figs. 2 and 4). We also anticipate a community of application developers who will want to access the Center's ontologies and data annotations to create new algorithms and applications to support scientific discovery. To support these users, we are also creating Web services to put the center's resources on the Semantic Web.

By bringing together these different communities and perspectives to create our Center, we are optimistic that we will be able to focus our activities and align their differing perspectives and objectives. In creating our Center, we brought together principal participants in the GO consortium with leaders in knowledge representation and ontology theory. We have already begun to reap the benefits of close interactions among this mix of expertise within our Center. For example, an initial task we have begun undertaking is to bring our different perspectives to the task of defining metrics of ontology quality, with the goal of creating a minimal set of standards to help communities that develop ontologies to produce higher quality ontologies by ensuring correct choices at early stages of development. The Center's educational workshops are run by members of all three of these communities, who do not yet always agree on specific modeling aspects of ontology; however, by engaging in frequent and lively discourse, we hope that difficult issues will be identified, enabling us to build consensus over time.

CONCLUSION

In conclusion, we believe that our Center will benefit the scientific community in developing technologies and resources focused on integrating biomedical ontologies and data. Making the diversity of ontologies and annotations on data using those ontologies accessible to scientists and computer applications is a core goal of our Center. These activities will stimulate the development of biomedical ontologies. Our work will also improve the quality of ontologies by allowing content creators to align and compare OBO ontologies with other ontologies as well as permitting them to interact with communities of ontology developers through workshops hosted by the Center. Biomedical data annotation databases such as OBD as well as tools to link annotations across species are also important resources that our Center will deliver, and they could bring enormous benefits to biomedical research by enabling the e-science paradigm, as well as prioritizing and streamlining laboratory experimentation. Finally, through education and outreach activities, we hope to broaden the base of expertise throughout the scientific community and raise the level of discourse and technology development in the field of biomedical ontology.

ACKNOWLEDGMENTS

We would like to acknowledge the enormous contributions of all the talented members of our Center, (<http://bioontology.org/overview-team.html>). We wish to acknowledge and thank Amelia Ireland of the Eu-

ropean Bioinformatics Institute, Hinxton, Cambridge, UK, for her contributions in creating and maintaining OBO. We also wish to thank the entire OBO community of users for their contributions to building and improving this growing resource. This work was supported by the National Center for Biomedical Ontology, under roadmap-initiative grant U54 HG004028 from the National Institutes of Health.

REFERENCES

- ASHBURNER, M., BALL, C.A., BLAKE, J.A., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29.
- COMMON TERMINOLOGY SERVICES. (2006). Available at: (<http://informatics.mayo.edu/LexGrid/index.php?page=ctsspecdetail>).
- DE CORONADO, S., HABER, M.W., SIOUTOS, N., et al. (2004). NCI thesaurus: using science-based terminology to integrate cancer research results. *Medinfo* **11**, 33–37.
- DOLIN, R.H., SPACKMAN, K., ABILLA, A., et al. (2001). The SNOMED RT procedure model. *Proc AMIA Symp* 139–143.
- DRYSDALE, R.A., CROSBY, M.A., GELBART, W., et al. (2005). FlyBase: genes and gene models. *Nucleic Acids Res* **33**, D390–D395.
- FLYBASE DATABASE. (2003). The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res* **31**, 172–175.
- GENNARI, J. (2005). *The Protege DAG-EDIT Plug-In*, University of Washington.
- HARRIS, M.A., CLARK, J., IRELAND, A., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, D258–D261.
- KNUBLAUCH, H., FERGERSON, R.W., NOY, N.F., et al. (2004). The Protege OWL plugin: an open development environment for semantic web applications. *Semantic Web ISWC 2004 Proc* **3298**, 229–243.
- LEXGRID. (2006). Available at: (<http://informatics.mayo.edu/LexGrid/index.php?page=>)).
- MUNGALL, C.J. (2004). OBOL: integrating language and meaning in bio-ontologies. *Comp Funct Genomics* **5**, 509–520.
- NOY, N.F., CRUBEZY, M., FERGERSON, R.W., et al. (2003). Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu Symp Proc* 953.
- SHETH, A., YORK, W., THOMAS, C., et al. (2004). Semantic web technology in support of bioinformatics for glycan expression. In: *W3C Workshop on Semantic Web for Life Sciences: 2004* (Cambridge, MA).
- SMITH, B., CEUSTERS, W., KLAGGES, B., et al. (2005). Relations in biomedical ontologies. *Genome Biol* **6**, R46.
- SOLBRIG, H., and BRINSON, T. (1998). Lexicon query service: RFP response. In: *3M Health Information Systems* (ProtocolSystems, Murray, UT), pp. 1–175.
- SPRAGUE, J., CLEMENTS, D., CONLIN, T., et al. (2003). The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res* **31**, 241–243.
- STOREY, M.A., NOY, N.F., MUSEN, M.A., et al. (2002). Jambalaya: an interactive environment for exploring ontologies. In: *Intelligent User Interfaces: 2002* (ACM Press, San Francisco).
- WESTERFIELD, M., DOERRY, E., KIRKPATRICK, A.E., et al. (1999). Zebrafish informatics and the ZFIN database. *Methods Cell Biol* **60**, 339–355.
- WHETZEL, P.L., BRINKMAN, RR., CAUSTON, H.C., et al. (2006). Development of FuGO: an ontology for functional genomics investigations. *OMICS (this issue)*.
- WHETZEL, P.L., PARKINSON, H., CAUSTON, H.C., et al. (2006). The MGED ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics* **22**, 866–873.
- WROE, C.J., STEVENS, R., GOBLE, C.A., et al. (2003). A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput* 624–635.
- YEH, I., KARP, P.D., NOY, N.F., et al. (2003). Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics* **19**, 241–248.

Address reprint requests to:

Dr. Daniel L. Rubin
Stanford Medical Informatics
MSOB X-215
Stanford, CA 94305

E-mail: rubin@smi.stanford.edu