ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Ontology Driven Data Collection for EuPathDB

Jie Zheng, Omar S. Harb, Christian J. Stoeckert Jr.

Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA, USA

Abstract. EuPathDB is a public resource of protozoan parasite genomic and functional genomic data. To address community needs, information on isolate specimens, and on genetic manipulation and phenotype, data will be collected directly from scientists. In order to facilitate data exploration, exchange, sharing and reuse, such data needs to be well-structured with standardized annotation. However, data collection in a uniform format remains challenging. In this report, we leverage existing ontologies to semantically represent the two cases of (1) isolate and (2) genetic manipulation and phenotype data with a focus on the needs/requirements of the EuPathDB community. Using ontology-based models, we designed submission forms and incorporated ontology terms for annotation with the goal of minimizing the burden on end users to submit standardized data.

Keywords: Ontology for Biomedical Investigation (OBI), ontology, EuPathDB

1 Introduction

Protozoan parasites are a major cause of global human and veterinary infectious diseases, such as malaria, toxoplasmosis, cryptosporidiosis, Chagas disease, sleeping sickness and leishmaniasis. Unfortunately current treatments are limited due to the rise of parasite drug resistance and immune evasion. The Eukaryotic Pathogen Database (EuPathDB; http://eupathdb.org) project integrates genomic and functional genomics data from over 30 different protozoan parasite species [1]. EuPathDB also integrates parasite isolate data, and genetic manipulation with resulting phenotype data but in a limited manner due to the heterogeneity of what is currently obtained. Desired information includes the geographic location of parasite isolate specimens collected, pathogen host information, and genetic manipulation and phenotype information associated with specific genes as these are important for parasite epidemiology, and vaccine anti-parasitic drug research. Currently, EuPathDB integrates isolate data from GenBank [2] which in turn, accepts sequence data and associated information directly from individual researchers. The submission to GenBank has clear requirements for sequence format and annotation. However, the completeness of information associated with sequence data is mainly dependent on the submitters. For example, when submitting isolate data to GenBank, there are no requirements for providing the parasite's host information, such as host organism, clinical information, and geographic location from where the isolate specimens were collected. This constitutes a big hurdle for EuPathDB and requires manual intervention to standardize, for example, the terms used for organism and geographic location. In addition, some important isolate information is lost during integration into EuPathDB due to its deposition as free text in the GenBank records. EuPathDB has some semi-structured genetic manipulation and phenotype data for *T. brucei*. Technically, genetic manipulation and phenotype data can also be collected through submission of User Comments available on EuPathDB gene pages. However, while such free text contributions are invaluable they are less useful for data exploration. For data integration, sharing and reuse, there is a need to collect isolate, genetic manipulation and phenotype data in a well-structured manner with standardized annotation.

Usage of the Gene Ontology (GO) [3] by model organism databases and many other resources has led to great success in data exchange, sharing, reuse and analysis. EuPathDB has also utilized GO and related ontologies (e.g., Sequence Ontology (SO) [4]) to

integrate genomic data annotation. The application of these ontologies enables EuPathDB to issue complex queries, such as retrieving functional genomic data across multiple species based on orthology [1].

The Ontology for Biomedical Investigations (OBI) is being developed for supporting consistent annotation of biological and clinical investigations [5]. It covers the terms to describe all aspects of an investigation including biological materials, protocols, generated data and type of analysis applied to the data. OBI is based on the Basic Formal Ontology (BFO) and follows Open Biomedical Ontologies (OBO) Foundry principles [6]. OBI is interoperable with other biomedical ontologies under the OBO Foundry umbrella since they are built on the basis of a common top-level ontology, BFO, and use a common set of relations. Each OBO Foundry ontology covers terms in a specific domain. For example, OBI focuses on experimental processes, GO is used for gene and gene product annotation and contains three main components, cellular component, molecular function, and biological process [3], and the Phenotype And Trait Ontology (PATO) defines terms for the description of phenotype and quality [7].

In this report, we describe how we apply OBI to model isolate and genetic manipulation with resulting phenotype data, generate effective submission forms, and provide terms for annotation from recommended OBO Foundry ontologies including OBI. Our experience demonstrates that starting with a semantic framework such as OBI is an effective approach of creating a well-structured data collection form.

2 Method

The following steps were applied to generate the submission form for collecting data from individual investigators.

2.1 Semantically Represent Data For Collection Using OBI

This step involved identifying the categories of data and information required to sufficiently characterize isolate specimens or genetically modified parasite phenotypes, followed by OBI-based modeling of the defined categories to capture their interconnectedness and relationships.

2.2 Generate the Submission Form Based on the Ontology Model

Using the ontology model as a guide, categories from step 1 (section 2.1) were organized logically with related categories grouped together in the two forms. For the isolate submission form, category order was as follows: isolate information (species, type, etc.), location of collected sample (geographic location, country, province, city), isolation source (host organism sample, or environmental sample), and nucleotide sequence information (sequence name, type, sequence). The genetic manipulation and phenotype submission form is organized as two main sections: genetic manipulation including genetic modification method, markers and/or reporters used in the modification; and assays used for investigation of the impact on the organism or the cellular location, molecular function, or biological process associated with the gene product.

We then determined which OBO library ontologies to use for various types of data and identified the terms in an ontology needed for standardized annotations. Some types of data, such as organism species and country, have a long list of ontology terms. In this case, we provide commonly used terms by surveying existing datasets.

The choice of format for the submission forms was based on feedback from EuPathDB end users with experience with these types of datasets.

3 Result

We first describe the data to be collected and then describe the generation of the submission forms for acquiring isolate data and genetic manipulation and phenotype information.

3.1 Isolate Submission Form

Parasites are isolated from samples and typed by their contained nucleotide sequences. This data and associated meta-data are important for epidemiological research related to parasite spread and resistance. The meta-data include information specific for the isolate specimen (species in the isolate, geographic location,

collection source), and host organism specific information (species, age, and clinical information). We describe data of interest and their relations using OBI and other ontologies.

The graphical representation in Fig. 1 allows effective and rapid identification of connections and relationships between the various items in the submission forms. For example, viewing the middle left section of the graphic illustrates how an isolate specimen is a subclass of *specimen*, while specimen ID is a *symbol* used to uniquely identify the isolate specimen studied. The specimen encompasses the target species of interest (isolate) which is a subclass of *organism* (bottom left of Fig. 1) and other microorganisms and is the output of *specimen creation process* (middle left of Fig. 1). Moving up the graphic in Fig. 1, the input of the *specimen creation process* is either the environmental source or host material (both subclass of *material entity*) that is *located in* a specific *geographic location*. In turn, the host material (upper middle of Fig. 1) is *part of* a host *organism* that *has qualities* such as *sex*, *age*, and *symptoms* (upper right of Fig. 1). The sequencing type assay (center of Fig. 1) is a subclass of *assay* and is linked to the isolate specimen as a *specified input* and its *specified output* is nucleotide *sequence data* for DNA from the isolate. The *textual entity*, locus/product name and sequence description (lower middle of Fig. 1), specify the *sequence*. The GenBank ID (bottom right of Fig. 1) is a subclass of *symbol* that is used to identify the *sequence* submitted to GenBank.

Submission of sequence data to GenBank and obtaining a GenBank ID(s) are essential steps for scientists sharing their data directly or via publications. Hence, it is critical that our submission form provides a convenient method to make data available in both GenBank and EuPathDB. In fact, enabling easy submission to GenBank through use of the form is a major incentive for community use. Therefore, essential to developing our submission form is developing a parser that generates GenBank "ready" files yet retains the added structure and standardized terms. It is anticipated that although EuPathDB will initially collect isolate data with the form, the data will first be delivered to GenBank to ensure it is properly archived and tagged with GenBank IDs. Subsequently, isolate data will be retrieved from GenBank for integration into EuPathDB and querying through its web sites.

A spreadsheet (Excel) format was used for the isolate specimen submission form because such data are generally already collected in this format according to our community advisors familiar with this data type. The spreadsheet format also facilitates submission of multiple nucleotide sequences associated with the isolates which can be cumbersome on typical web-based forms.

Ontology terms that can be used in the data annotation are provided as drop-down lists in the submission form. These include, OBO ontologies, NCBI Taxon, OBI, Environment Ontology (EnVO), PATO, Ontology for General Medical Science (OGMS) and Gazetteer (GAZ). Having standardized annotation such as these in GenBank will greatly reduce or eliminate the need for manual intervention to integrate isolate data from GenBank into EuPathDB.

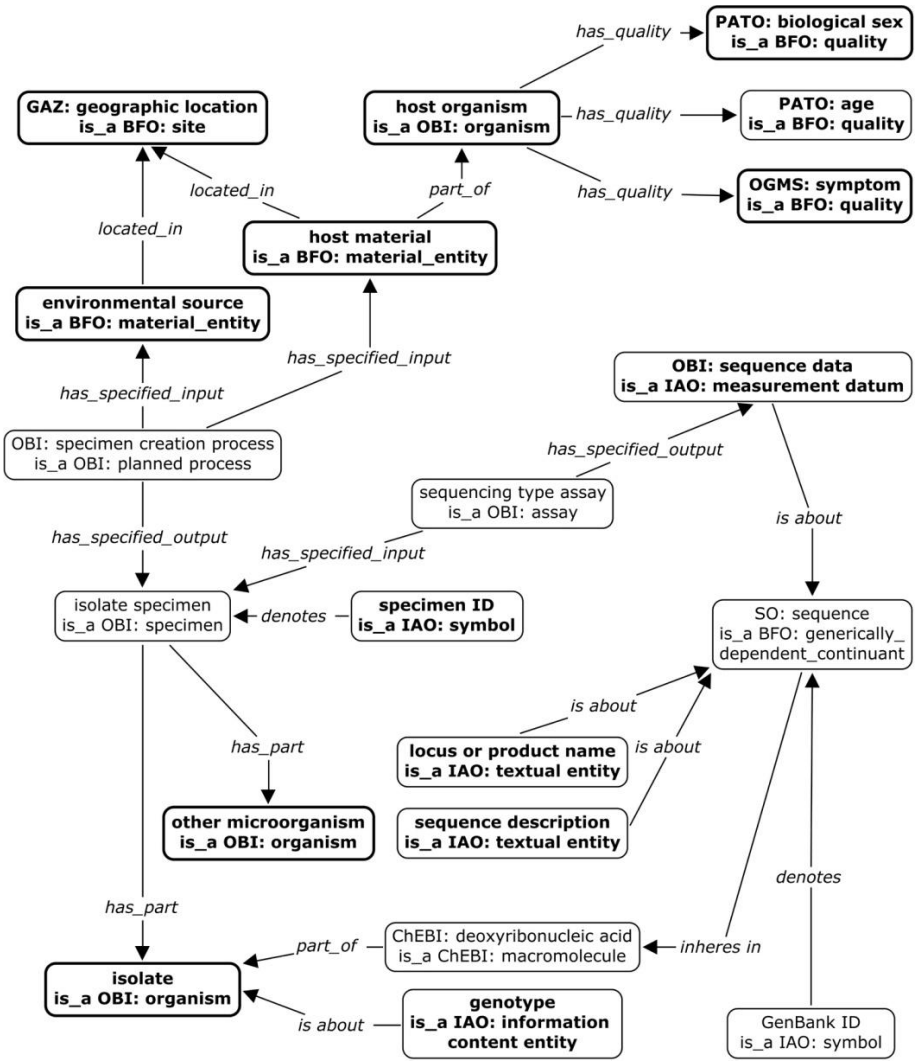

Figure 1. Ontology-based representation of sequence related isolate data. The ontology terms are indicated by using ontology name abbreviation as prefix. Italicized text represents relations. The data collected in the submission form are in the bold font. The fields require ontology terms are in thick border box. IAO stands for Information Artifact Ontology. ChEBI stands for Chemical Entities of Biological Interest ontology.

3.2 Genetic Manipulation and Phenotype Submission Form

Observed phenotype(s) that can be linked to specific genetic modifications are valuable for the development of novel anti-parasitic drugs. In addition, knowledge of when in the parasites life cycle an observed phenotype occurs and has an effect is important. To this end, we represent the data of interest using ontologies (mainly OBI and GO), as illustrated in Fig. 2.

Phenotype data may refer to the impact of gene modification on four possible observed features:

- Quality of the organism
- Cellular location of gene product
- Molecular function of gene product
- Biological process of gene product

The genetically modified parasite is a subclass of *genetically modified organism* and is generated by *genetic transformation* process (top section of Fig. 2). The left and lower sections of Fig. 2 illustrate that *assays* are performed to examine the genetically modified parasite about *organismal quality* (eg. viability), *cellular component* the gene product is *located in*, effects on its *molecular function*, or *biological process* it *participates in* at specific *lifecycle stage*.

The OBI terms will be used in the genetic modification method and assay fields. The GO terms will be listed in cellular component, molecular function and biological process fields and Ontology for Parasite Lifecycle (OPL) will be used for annotation of lifecycle stages.

Driven by our communities' needs genetic manipulation and phenotype submission will occur using a web form since EuPathDB plans to collect these data directly from specific locations at the web site (Gene pages containing information that can be linked to genetic manipulation and phenotype). One big advantage of a web form is that it can change dynamically based on the users input. For example, not all kinds of genetic modification methods use selectable markers and/or reporters. Once a user selects a specific modification method like gene knock out or gene knock in, the form dynamically displays questions about marker(s) and reporter(s). The ontology terms used for annotation are also changed depending on the input data that can short the term list. For example, different parasites may have different lifecycle stages. A subset of lifecycle stage terms will be listed based on which kind of parasite the gene is derived from. All these can minimize the efforts of the end users during the submission process.

We have collected comments on the submission forms from some of the end-users. The forms have been adjusted based on the feedback. The isolate submission has been approved by the end users and will be distributed to more parasite researchers.

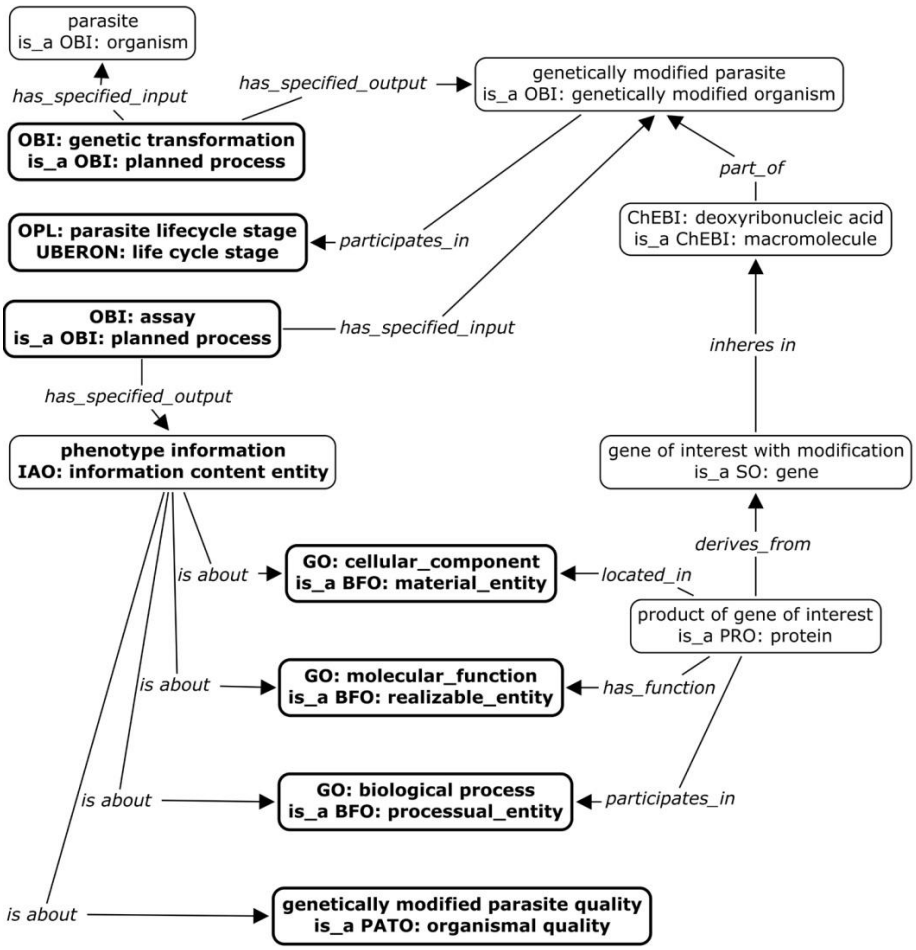

Figure 2. Ontology-based representation of phenotype data. The ontology terms are indicated by using ontology name abbreviation as prefix. Italicized text represents relations. The data collected in the submission form are in the bold font. The fields require ontology terms are in thick border box. IAO stands for Information Artifact Ontology. ChEBI stands for Chemical Entities of Biological Interest ontology.

4 Discussion

Consistently-applied annotations will facilitate data integration, sharing, and (re)analysis. In recent years, ontologies have been widely used in supporting the consistent annotation of the biological data. However, it is still a big challenge to collect standardized data directly from scientists. Complicated submission forms and the use of ontology terms for annotation often inhibit scientists from submitting their data. However, collection of low quality data is less useful for sharing, exchanging and analysis. Our goal is to minimize the efforts of the submitters and capture important data with standardized annotation at same time. In this report, we apply a semantic framework to submission form designs. It is an efficient approach to organize the data in a logical fashion and to identify appropriate bio-ontologies to be used in annotation.

The isolate form will allow scientists to submit multiple sequences to GenBank without going through the GenBank submission process which can be a challenge to scientists especially when submitting many isolate records. Using the genetic manipulation and phenotype collection form to capture crucial genetic modification and phenotype data using ontology terms will facilitate the ability of EuPathDB to share and exchange data with other genetically modified parasite phenotype databases, such as RMgmDB (http://pberghei.eu/). Furthermore, high quality isolate and genetic modified parasite related phenotype data, will enable users of EuPathDB to perform integrated searches of the types: "Compare sequence data from *Plasmodium* isolates that are restricted to East Africa to those from West Africa", "List genes that when knocked out result in a defect in parasite growth during the intraerythrocytic cycle", "List genes fused to green fluorescent protein (GFP) that when expressed are located in the cell membrane".

Currently the submission forms are in the prototype stage. We will distribute the isolate submission form to EuPathDB user communities and incorporate the genetic manipulation with associated phenotype form into EuPathDB websites. Based on user feedback, the forms and underlying ontology-model will be improved to achieve the goal of collecting critical information from individual scientists about their experiments that is structured yet is not burdensome, and provides incentives (such as facilitated submission to the GenBank archive).

Acknowledgments

We thank Dr. G Robinson, Dr. R Chalmers, Dr. CJ Janse, Dr. G. Widmer, Dr. L. Xiao, and Dr. SM Khan for their valuable comments on the submission forms.

This research is supported by NIH grant 5R01GM93132-1 and by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health Award NO1-AI900038C Contract No. HHSN272200900038C.

References

1. Aurrecoechea C., Brestelli J., Brunk B.P., Fischer S., Gajria B. et al.: EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res 38(Database issue): D415-419 (2010)
2. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W.: GenBank. Nucleic Acids Res 37(Database issue): D26-31 (2009)
3. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H. et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25-29 (2000)
4. Eilbeck K., Lewis S., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M.: The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology 6:R44 (2005)
5. Brinkman R.R., Courtot M., Derom D., Fostel J.M., He Y. et al.: Modeling biomedical experimental processes with OBI. J Biomed Semantics 1 Suppl 1: S7 (2010)
6. Smith B., Ashburner M., Rosse C., Bard J., Bug W. et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25(11): 1251-1255 (2007)
7. Mungall C.J., Gkoutos G.V., Smith C.L., Haendel M.A., Lewis S.E. et al. Integrating phenotype ontologies across multiple species. Genome Biol 11(1): R2 (2010)