### Applications of OBI

In November 2009, the OBI ‘Philly’ release was published that supported several uses cases such as building data analysis workflows, querying databases queries, and representing investigations from detailed individual experiments to high level study designs. Since then, OBI has been adapted in practice, and is used in multiple projects including the five described next.

**Adding semantic expressivity to data stored in the IEDB**

The Immune Epitope Database (IEDB) catalogs experiments that characterize the location and function of immune epitopes in infectious agents, allergens, transplants and auto-antigens. Information is entered into the IEDB through author submissions and through manual curation of the scientific literature. Over 380,000 experiments have been entered into the IEDB to date and manual curation is on track to cover more than 95% of all relevant journal articles ever published by the end of 2011. Like many databases, information in the IEDB comes mostly in the form of values from controlled lists. Where available, existing ontologies were used as a source for terms on such lists, but for many types of information controlled vocabularies had to be developed by the IEDB team. Building and maintaining controlled vocabularies that deal with changing naming conventions over decades of scientific practice is a significant task. Worse, it is a thankless task, as the work has limited value outside of the IEDB itself.

To address this issue, the IEDB team has worked with multiple ontology developers to extend e.g. GO, ChEBI, PRO and PATO to replace IEDB internal controlled vocabularies. By far the largest contribution was made by OBI, which not only covers terms specific for experiments and investigations, but also provides the framework that explains how terms from other ontologies are related to each other in the context of an experiment. Figure X depicts how the list of T cell assay types used in the IEDB has been mapped to OBI. The T cell assay classes in OBI are constructed using logical definitions that tie them to GO terms representing the biological processes interrogated by the assays such as IFN-gamma production, and to more general experimental techniques represented in OBI such as ELISA or FACS assays.

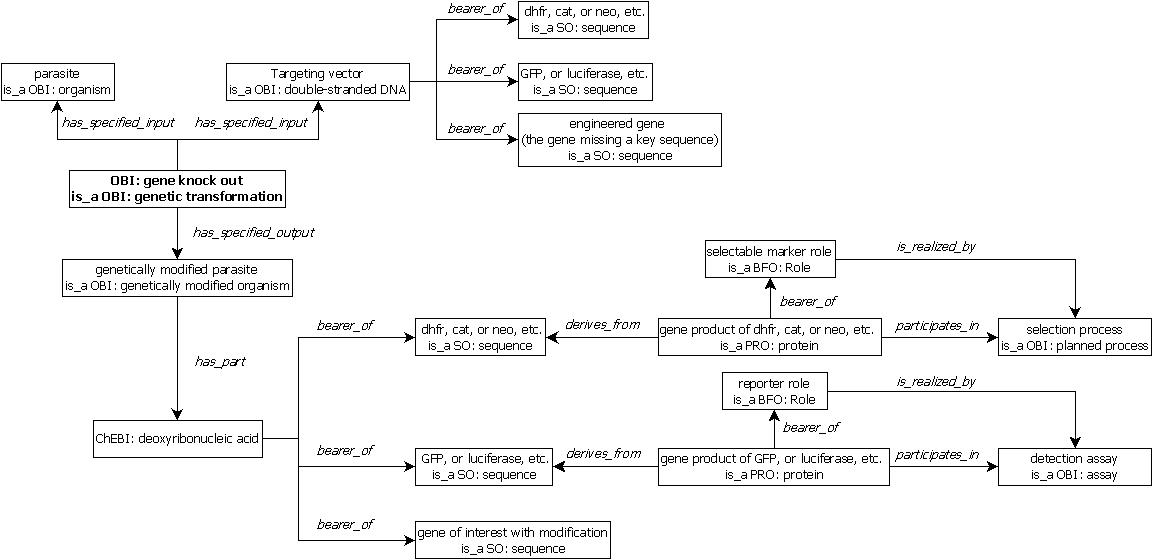
Using OBI as a source of assay terms in the IEDB has replaced plain lists of strings and given them expressive textual and logical definitions. This enables interoperability with other knowledge resources. For example, if RNA-seq data shows that T cells from asthmatic patients overexpress genes involved with particular GO biological processes, the IEDB data can now easily be interrogated for experiments that show T cell involvement in these biological processes following auto-antigen and allergen recognition. There are also multiple benefits in the maintenance of terms for the IEDB team. It is now easier to identify redundant entries as the use of GO terms provides synonyms to indicate that ‘X production’ is equivalent to ‘Y production’, and because logical reasoning flags equivalent entries upon their creation, Logical reasoning also organizes terms into a hierarchy without the need for human intervention. Such a hierarchy eases navigation of flat lists with hundreds of entries, and enables querying for more general terms such as ‘cytokine production assays’.

**Designing smart, standardized submission forms for EuPathDB**

The Eukaryotic Pathogen Database (EuPathDB; http://eupathdb.org) project integrates genomic and functional genomics data from over 30 different protozoan parasite species. Protozoan parasites are a major cause of global human and veterinary infectious diseases, such as malaria, toxoplasmosis, cryptosporidiosis, Chagas disease, sleeping sickness and leishmaniasis. EuPathDB also aims to integrate data on specific isolates of parasites, their genotypes, and effects of genetic manipulation on the phenotype. However, currently available data on parasite isolates and genetic manipulation is highly heterogeneous and therefore hard to query and represent due to lack of community-accepted standards.

To better standardize data as it is being captured, the EuPathDB team and user communities decided to develop submission forms. OBI was chosen as the basis for these forms because it provides a framework for modeling the generation of the desired data through its use of planned processes in which external ontologies can easily be referenced. In OBI, description of the genotype of an isolate is done by referring to the process of specimen collection that resulted in the physical isolate, followed by the sequencing experiment performed on the isolate, which then resulted in information about its genome. We consulted with investigators performing these processes and established what was needed in the form. The result was a form that captures details about the process of creating an isolate specimen (where, when, and from what was it collected) and performing a sequencing assay (to obtain isolate sequence data). In this example, the terminology used for describing each instance of an isolate specimen is drawn mainly from other ontologies (e.g., GAZ, PATO). OBI is used for categories of terms needed (e.g., sequence data) and for relating the information collected on the form for loading into a database and subsequent data mining.

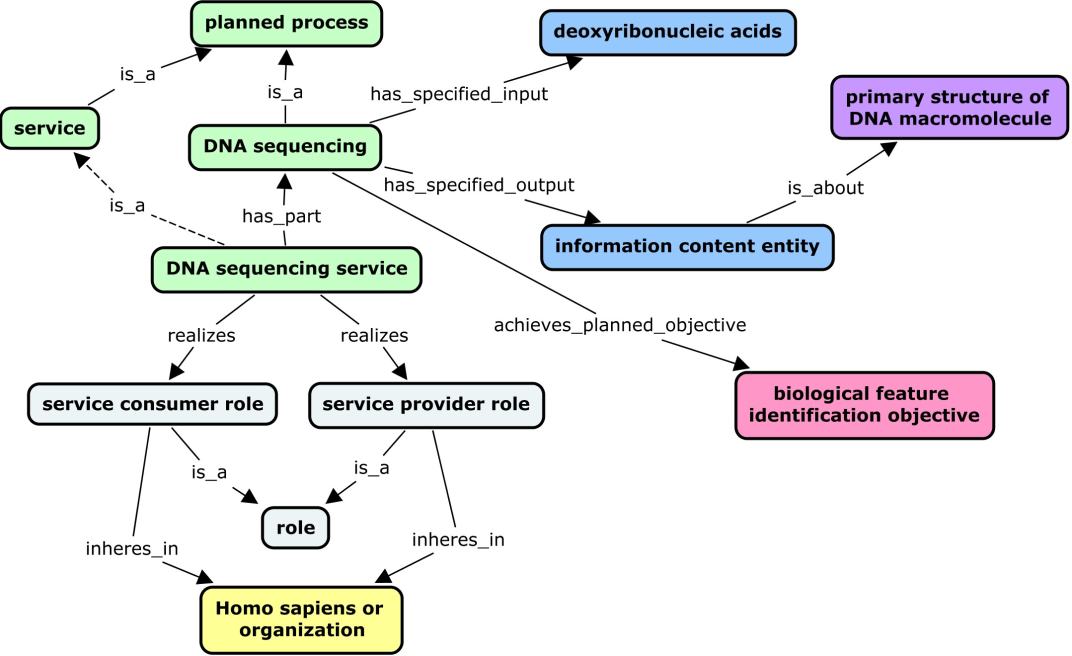
The same approach is being used for the more complex task of capturing phenotype information resulting from genetic modification of parasites. Insights into the function, location, and biological processes for parasite proteins of interest are found through genetic modification such as knocking out the expression of the gene encoding the protein or tagging it with a fluorescent marker. Also of interest is the effect of these modifications on the viability, infectivity, and proliferative ability of the parasite. The key planned processes used to model this case are genetic transformation and assay (both drawn from OBI). In addition to using OBI terms to relate collected information for EuPathDB databases, they will also be used to guide what is presented to the user on a web-based form to reduce the burden of filling out the form by only requesting relevant information. For example, in Figure Z, information provided on the



**Research Resource Discovery**

An important challenge in biomedical research is the ability to find relevant scientific resources. The majority of resources generated in the course of biomedical research are never publically documented in the literature, catalogs, or lab web sites. Those that are advertised through such channels have proven challenging to discover in a reproducible and efficient way, as they are dispersed across a complex information landscape, described using inconsistent standards or language, and lack information connecting them to related entities such as organisms, genes, genotypes, site of action, biological processes, or diseases. For example, a search for an antibody that can be used to recognize the GFAP protein in rat brain in an immunostaining assay is seemingly simple, yet difficult to accommodate. Giving researchers semantically linked information about available resources can reduce time-consuming and expensive duplication of resource development and promote reuse of existing resources and collaboration. The eagle-i [1], NIF [2], VIVO [3], and Biositemaps [4] projects all have as their mission to improve search and discovery of research resources. Since developing multiple incompatible standards would negate the vision of seamless resource navigation, these organizations have collaborated in aligning their ontologies by reusing common terms and definitions, with the goal of migrating to usage of the same ontology classes and URIs.

Due to its scope, quality, and community development practices, OBI was identified as the ideal ontology in which to develop a common representation of research resources to be reused by these four projects. Thus far, alignment efforts have focused on instruments, services, organismal genotypes and antibodies. Fig. 3x illustrates the representation of a DNA sequencing service within OBI as an example of a unified representation for reuse across the different resource discovery systems.



**Figure 3x. DNA sequencing service as an example of a service design pattern**. Services are typically offered by core laboratories, which perform some planned process (for example, an assay or a material production) for a customer. Like planned processes, services are also linked to entities such as protocols, instruments, specimens, reagents, and objectives. However, key features that distinguish services from the planned processes they perform are the additional processes entailed in a service (order placement, billing, etc.). In addition, services necessarily have service providers and service consumers as participants. To capture these distinctions and allow query using these search facets, a design pattern was used in which a service *has\_part* some planned process and *realizes* some ‘service consumer role’ and ‘service provider role’ which *inheres\_in* a person or organization.

**Shortcut relations.** From an application point of view, eagle-i and NIF sometimes required a simplified representation from that available within OBI. For instance, a single property to relate a service to a core laboratory providing that service was required by the eagle-i user interface, rather than using OBI’s composed relation built from two properties to connect an organization to the service it provides (‘*organization’* *‘bearer\_of’’* some ‘*service provider role’* and ‘*realized\_by’* some ‘*service’*). We logically collapsed this complex statement with a single property linking a service to its provider (‘*service provider’* ‘*provides\_service’* ‘*service’*) where ‘*service provider’* is defined as follows: (‘*organization’* or *‘Homo sapiens’*) and (‘*bearer\_of’* some *‘service provider role’*). This need to simplify complex relation chains is a common issue in using ontologies for data collection applications, and ‘shortcut relation’ approaches like the ones suggested in [5] should be exploited. The reuse of OBI within NIF and eagle-i has provided a valuable use case for developing best practices around reuse of ontologies within software applications. Use of OBI across different applications will allow publication of resources as Linked Open Data under a common representation. Investigators will have improved navigability of potential research resources, identification of colleagues for collaboration, and exploration of novel research connections.

[1] Torniai, C., Bashor, T., Bourges-Waldegg, D., Corday, K., Frost, H.R., Johnson, T., Segerdell, E., Shaffer, C.J., Stone, L., Wilson, M.L., Haendel, M.A. (2010) eagle-i: an ontology-driven framework for biomedical resource annotation and discovery. Bio-Ontologies 2010: Semantic Applications in Life Sciences, ISMB 2010 (Boston, MA).

[2] Gupta A., Bug W., Marenco L., Qian X., Condit C., Rangarajan A., Müller H.M., Miller P.L., Sanders B., Grethe J.S., Astakhov V., Shepherd G., Sternberg P.W., Martone M.E. (2008) Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). Neuroinformatics. Sep;6(3):205-17.

[3] Krafft, D.B., Cappadona, N.A., Caruso, B., Corson‐Rikert, J., Devare, M., Lowe, B.J., VIVO Collaboration (2010) VIVO: Enabling National Networking of Scientists. Proceedings of the WebSci10: Extending the Frontiers of Society On‐Line, April 26‐27th, 2010, Raleigh, NC: US.

[4] Tenenbaum JD, Whetzel PL, Anderson K, Borromeo CD, Dinov ID, Gabriel D, Kirschner B, Mirel B, Morris T, Noy N, Nyulas C, Rubenson D, Saxman PR, Singh H, Whelan N, Wright Z, Athey BD, Becich MJ, Ginsburg GS, Musen MA, Smith KA, Tarantal AF, Rubin DL, Lyster P. (2010, Oct 16). The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. Journal of Biomedical Informatics. Epub ahead of print, doi: 10.1016/j.jbi.2010.10.003.

[5] Mungall CJ, Ruttenberg A, Osumi-Sutherland D. Taking shortcuts with OWL using safe macros. Nature Precedings. 2010. <http://precedings.nature.com/documents/5292/version/1/files/npre20105292-1.pdf>

**Harmonize the annotation across different functional genomics resources**

Terminologies, minimum information checklists, and exchange formats are increasingly used in the structuring and annotation of data sets, but in many cases the lack of coordination between their developers, who normally cater only to their own community, brings significant challenges. This diversity of standards hinders discovery, because only those able to navigate the various available terminologies and formats can associate and integrate information scattered across incompatible databases. The importance of unlocking the value of shared datasets to accelerate discovery requires new models for the way we collaborate [1, 2, 3].

The Investigation, Study, Assay (ISA) framework is an exemplar approach that illustrates the concept of *synergizing* as a way to build productive infrastructural relationships among data producers, consumers and service communities. The core elements of the collaborative ISA framework are an open source ISA software suite and an extensible hierarchical data structure [4], whose model has informed the structure of OBI and provided terms for populate its classes. ISA and OBI have been implemented by a growing number of international public resources [listed at 5] that cater to data as diverse as stem cell, toxicogenomics, environmental gene surveys, microbial diversity studies, and a variety of metabolomics and metagenomics-based studies to maintain cross-domain compatibility in the way the experimental context is described.

In its most basic form, OBI is used as a source of controlled terminology to describe assays and instruments, as well as biological material and data manipulation protocols. One of the ISA software component, ISAconfigurator, ties up relevant OBI entities in ISA table definitions, via calls to NCBO BioPortal [6]. Figure Xa details how assay tables can be declared while figure Xb shows how protocols, used in canonical ISA workflows, may be typed with OBI process classes. Uploaded to the ISAcreator, the editor component, these configurations guide the users in the annotation process, enabling curation at the sources. This novel approach offers experimentalists the opportunity to start taking data management into their own hands, becoming active consumers of community terminologies. In a more advanced use, the ISAvalidator component reads the configurations files and relies on OBI to carry out semantic validation and error detection, *e.g.*, by inspecting protocols referenced between elements of the ISA structure. Developments are ongoing to further check the semantics of data matrices by inspecting reported data transformations and data matrix quantitation types (e.g. a *q-value* (OBI\_ 0001442) should be present in the data matrix file if a *false discovery rate correction method* (OBI\_0200163) is reported).

The combination of ISA and OBI illustrates the power of a ‘horizontal’ synergistic approach to engender networks that go beyond individual life science domain- or assay technology-focused communities. As the ISA community works to expose the experimental metadata to the world of Linked Data, OBI is set to provide the backbone classes for RDF representation and its object properties are used to make explicit the relations between ISA syntactic elements.

[1] Data's shameful neglect, Nature 461, 145 (2009)

[2] Dealing with data, Science, 331, 692 (2011).

[3] Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka AM, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J. Megascience. 'Omics data sharing. Science. 2009 Oct 9;326(5950):234-6.

[4] Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K, Field D, Harris S, Hide W, Hofmann O, Neumann S, Sterk P, Tong W, Sansone SA.ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics. 2010 Sep 15;26(18):2354-6. Epub 2010 Aug 2.

[5] ISA communities and data resources: <http://www.isa-tools.org/case_studies.html>

[6] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W170-3. Epub 2009 May 29.

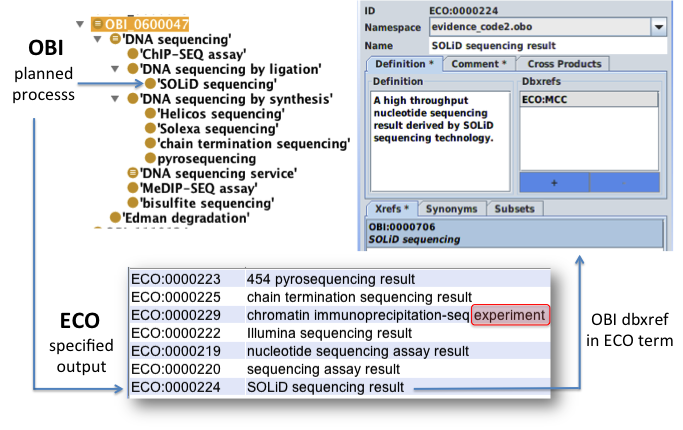
**Expanding types and clarifying the meaning of “evidence” in the Evidence Code Ontology**

The Evidence Code Ontology (ECO; http://www.obofoundry.org/cgi-bin/detail.cgi?id=evidence\_code) is a rich ontology for making statements about types of evidence, where the root class *evidence* is defined as “a type of information that is used to support an assertion.” Originally ECO grew out of a list of evidence types that were used to qualify Gene Ontology (GO) annotations at participating databases, such as The *Arabidopsis* Information Resource (TAIR; http://www.arabidopsis.org), with respect to different experimental types.1 The original intent of ECO was to provide a set of evidence codes that could be used in conjunction with GO terms, and which were not to be considered outside of the context of GO. In its first release on SourceForge ECO contained some 125 terms, which contained the subset of well-known GO terms (<http://www.geneontology.org/GO.evidence.shtml>).

Eventually, however, the need has arisen for more detailed evidence types for use not only by model organism databases that perform GO annotation and projects related to GO such as Phylogenetic Annotation and INference Tool (PAINT; http://wiki.geneontology.org/index.php/PAINT), but also by external groups such as the Universal Protein Resource (UniProt; http://www.uniprot.org), Ontology of Microbial Phenotypes (<http://www.microbialphenotypes.org>), Ascidian Network for *In Situ* Expression and Embryological Data (ANISEED; <http://aniseed-ibdm.univ-mrs.fr>), and various independent laboratories engaged in independent research programs. The maintainers of ECO have been simultaneously engaged in creating terms to meet the needs of users while attempting to broaden the scope of ECO to allow for general compatibility with groups outside of GO. To that end, we performed revisions of ECO which have resulted in the creation of Aristotelian definitions, the removal of overt references to GO, and the addition of a second root class called *assertion method*, defined as “a means by which a statement is made about an entity.” All terms in ECO now are a type of evidence, an assertion method (manual or automatic), or an internal cross product of *assertion method* and *evidence*, i.e. how a given type of evidence is applied to make a given statement. At present, there are 257 terms in ECO.

Despite progress in clarifying what ECO represents, a critical area that needs more work is to distill what is meant by “evidence” itself. Indeed, at least four meanings of “evidence” are present in the ontology at present (with example term names in italics): class of evidence (e.g. *phylogenetic evidence, experimental evidence,* or *expression pattern evidence*), type of analysis performed (e.g. *combinatorial analysis, protein binding experiment,* or *tissue ablation*), type of data output from an experiment (e.g. *flow cytometry data, chain termination sequencing result,* or *sequence alignment evidence*), or the statement of fact that the data support (e.g. *motif similarity* or *structural similarity*). Clearly there is a need to distinguish among the currently used meanings of “evidence” in order to achieve better internal consistency, which will result in an ontology that is easier to develop and maintain, as well as understand by users. To this end, we are using the carefully defined classes of OBI to refine our term meanings and will eventually create ECO terms based on OBI terms to expand the types of evidence the ECO covers. Likewise, OBI will benefit from incorporating the types of experiments and processes used to generate evidence types found in ECO, that are not yet found in OBI. In this way both ontologies will benefit.

ECO and OBI are orthogonal ontologies because OBI is built around planned processes, whereas ECO classes are information artifacts. We intend for many ECO *evidence* types to be defined as specified outputs of OBI *planned process* and its subtypes, including *assay* and *data transformation*.For example, ECO:0000224 *SOLiD sequencing result* would be a specified output of the OBI class *SOLiD sequencing* (Figure). The ECO term would carry the database cross reference (dbxref) to its OBI counterpart *SOLiD sequencing*.



**Figure**. Many ECO terms will be created or validated against OBI, by making ECO terms specified outputs of OBI planned processes. ECO:0000224 SOLiD sequencing result is an output from the *planned process SOLiD seuquencing* in OBI. The connection will be made in the xref field of the ECO term, shown here in OBO-Edit. Viewed in light of OBI, terms like ECO:0000229 *chromatin immunoprecipiation-seq experiment,* highlighted in red, immediately appear as inconsistent with other terms in ECO, and would be targets for correction to achieve consistency.

Although we don’t expect that all types of evidence in ECO will have counterpart elements in OBI (for example *curator inference* will not), we still believe that it is feasible to make certain domains within ECO and OBI completely synergistic.

1. [Nucleic Acids Res. 2004 January; 32(Database issue): D258–D261]