### Applications of OBI

A release candidate of OBI was published in November, 2009 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 four 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 most 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 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 controlled list of T cell assay types in the IEDB has now been mapped to OBI. The OBI terms utilized in the IEDB 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.

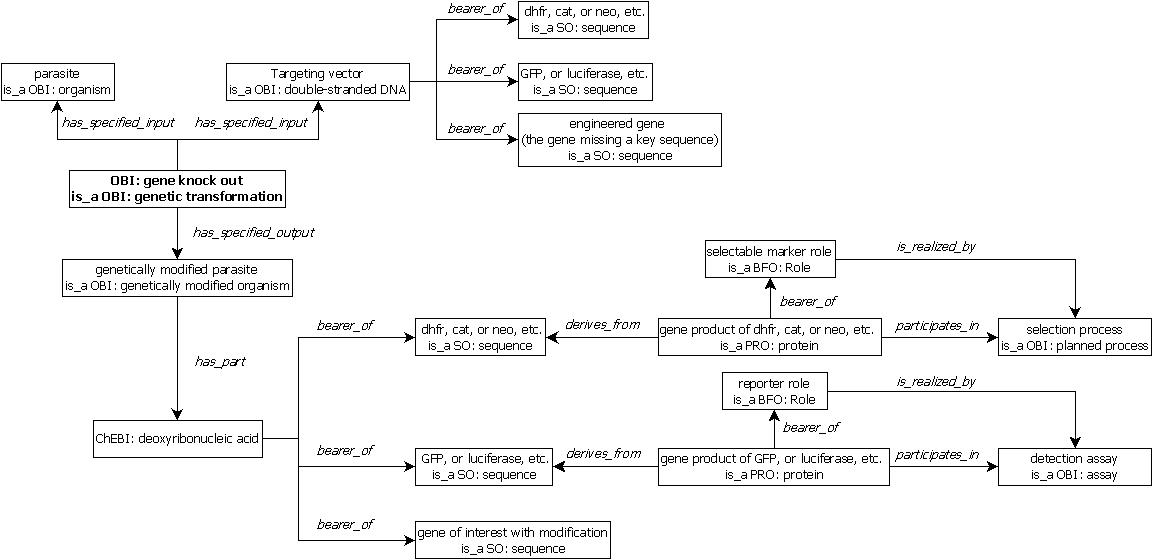
Using OBI as a source of assay terms in the IEDB has replaced plain lists of strings and given them expressive textual and logical definitions. This enables interoperability with other knowledge resources. For example, if RNA-seq data shows that T cells from asthmatic patients overexpress genes involved 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 genetically modified such as knocking out expression of the gene encoding the protein or tagging it with a fluorescent marker. Also of interest is the effect 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



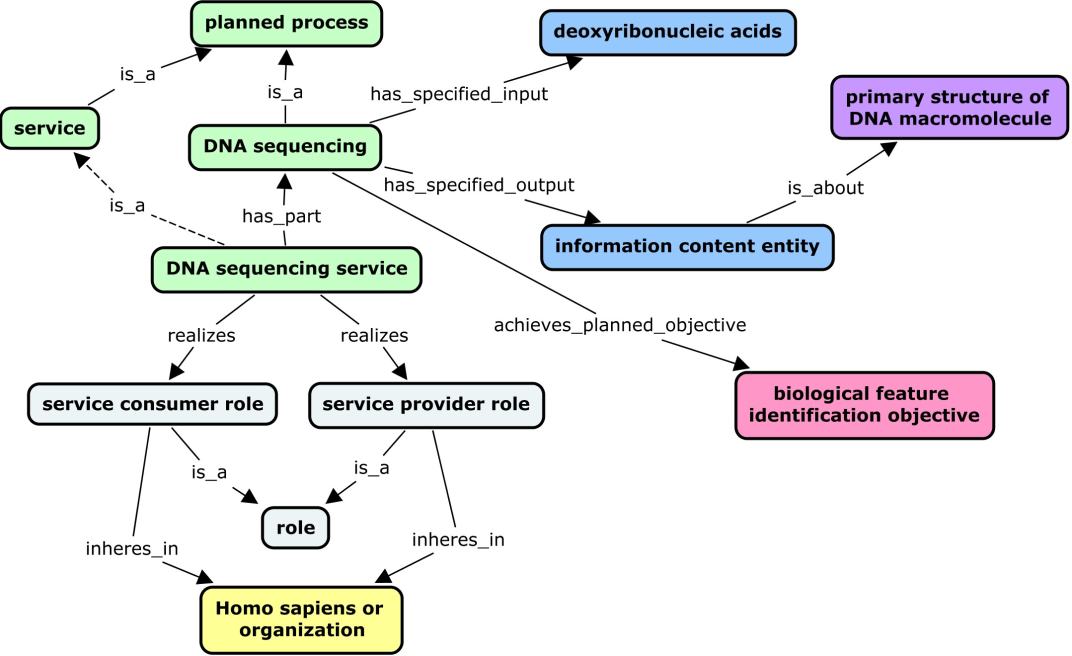
**eagle-i representation of research resources**

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 ot recognizes 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.

A number of efforts now exist to address this problem, namely the eagle-i Consortium, the Neuroscience Information Framework (NIF), VIVO, and the Biositemaps project. The eagle-i Consortium (www.eagle-i.org/home) is developing a semantic framework and suite of ontology-driven tools to help researchers find unpublished and rarely shared biomedical research resources, including reagents, instruments, services, model and non-model organisms, protocols, biospecimens, human studies and research opportunities via a semantic search portal [1]. NIF is an ontology-based, federated semantic search system that indexes and queries over 75 public databases that are relevant to neuroscience, including large (Gemma; [2]) and small (CCDB; [3]) databases [4]. The VIVO [5] project addresses the issue from a different perspective, linking resources to the researchers who use and produce them. VIVO has developed an ontology driven semantic search application to allow discovery of experts and collaborators based on their interests, activities and outputs of their research. The Biositemaps project [6] has focused on indexing resources developed through the National Center for Biomedical Computing (NCBC) program and Clinical and Translational Science Awards (CTSA) Informatics institutions. Resource annotators employ the Biositemaps information schema, which includes use of the Biomedical Resource Ontology (BRO), to annotate the resources locally and then it is made available in an online search tool.

Recognizing the need for a common, shared and semantically linked representation of research resources, eagle-i-, NIF, VIVO, and Biositemaps have collaborated together 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 re-used by these four projects. Thus far, alignment efforts have focused on instruments, services, organismal genotypes and antibodies. Below we illustrate our process for representing services in OBI, and discuss approaches to facilitate the use of OBI within software applications**.**

**Services.** 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 (Fig x3).



**Figure x3. DNA sequencing service as an example of a service design pattern**.

In designing a hierarchy of service types, two principles for classification were discussed: (a) A process-based approach that uses the nature of the process performed by the service (e.g. analysis versus production versus storage, etc); (b) An input/output-based approach that relies on the nature of the input and output of the service (e.g. material versus data). There was consensus that classifying services first according to their input/output was more intuitive and useful, and that the type of process performed would be a secondary axis for classification. The service hierarchy is shown in Figure x4. Modeling services in such a way will support query via the component processes (for example the DNA sequencing technique, Figure x3), the material entities involved in the service (such as DNA), and the objective of the service (sequence identification).



**Figure x4.** Asserted service hierarchy. The primary axis of classification is the input and the output of the planned process.

**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 [7] 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.

- Common considerations: alternative labels / shortcut relations

- Potential for reducing # of use-cases, and list all in table

[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] http://www.chibi.ubc.ca/Gemma/

[3] Martone ME, Gupta A, Wong M, Qian X, Sosinsky G, Ludaescher B, Ellisman MH (2002). A cell centered database for electron tomographic data. J Struct Biol. 138:145-155.

[4] 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.

[5] 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.

[6] 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.

[7] 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>

**ISA tools and annotation of functional genomics experiments**

Use case 3: ISA-OBI synergy to provide a common annotation framework for reporting biological studies

Funders and journals require that researchers share their data, and encourage the enrichment and standardization of experimental context or metadata [ref-1] to make datasets reliable and reusable, particularly if used as the underpinnings of future investigations. This situation demands better annotation at source (by data generators or community-based curation efforts), using open source software with automated content validation [ref-2]. The ISA community is an example of a growing number of international public and internal resources that implement and contribute to the development of the ISA tools [ref-3] - the ISA-Tab format and ISA software components - in diverse life science scenarios, including toxicogenomics, stem cell genomics, environmental genomics and metagenomics [ref-4].

In this context, ISA can rely on OBI ontology in several ways. In its most basic form, ISA tools uses OBI as a controlled vocabulary to describe assays and instruments, as well as biological material and data manipulation protocols. In practice, ISAconfigurator tool ties up relevant OBI entities in ISA table definitions, via calls to NCBO bioportal [ref-5]. Figure Xa details how assay tables can be declared while figure Xb shows how protocols, used in a canonical ISA workflows, may be typed with OBI process classes for each assay worflow.

When used in ISAcreator, the ISA format editor, those configurations guide users in their annotation process. ISA approach has been validated by managing experimental metadata from various domains, from environmental gene surveys (OBI:to be submitted) used in microbial diversity studies to multi-omics toxicogenomics studies [ref-6,7,8] and single nucleotide resolution nucleic acid structure mapping assays (OBI\_0000870) (Rocca-Serra et al, submitted). We have also demonstrated how the tools could be used recording flow cytometry data and are currently working with the MIBBI [ref-9] project to extend to other domains and convince of the usefulness of OBI as core resources for data integration.

In a more advanced use, ISAvalidator, reading ISAconfigurations, relies on OBI to carry out semantic validation and error detection by inspecting protocols referenced between ISA node elements. 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\_xxx) should be present in the data matrix file if a false discovery rate correction method (OBI\_0200163) is declared).

Finally, a more complex use of OBI in ISA infrastructure can be found in the ISAconverter RDF conversion component. This module exposes experimental metadata to the world of Linked Data. OBI then provides backbone classes for RDF representation of ISA studies while OBI object properties can be use to make explicit the relations between ISA syntactic elements without necessarily having to resort to the OWL expressivity and possibly complexity.

1.

PMID:19815759

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.

2.

PMID: 18769432

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, Rhee SY. Big data: The future of biocuration. Nature. 2008 Sep 4;455(7209):47-50.

3.

PMID:20679334

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.

4. ISA community and case studies: http://www.isa-tools.org/case\_studies.html

5.

PMID: 19483092

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.

6.

PMID: 20888850

Ellinger-Ziegelbauer H, Adler M, Amberg A, Brandenburg A, Callanan JJ, Connor S, Fountoulakis M, Gmuender H, Gruhler A, Hewitt P, Hodson M, Matheis KA, McCarthy D, Raschke M, Riefke B, Schmitt CS, Sieber M, Sposny A, Suter L, Sweatman B, Mally A. The enhanced value of combining conventional and "omics" analyses in early assessment of drug-induced hepatobiliary injury. Toxicol Appl Pharmacol. 2011 Apr 15;252(2):97-111. Epub 2010 Oct 1.

7.

PMID: 21081137

Matheis KA, Com E, Gautier JC, Guerreiro N, Brandenburg A, Gmuender H, Sposny A, Hewitt P, Amberg A, Boernsen O, Riefke B, Hoffmann D, Mally A, Kalkuhl A, Suter L, Dieterle F, Staedtler F. Cross-study and cross-omics comparisons of three nephrotoxic compounds reveal mechanistic insights and new candidate biomarkers. Toxicol Appl Pharmacol. 2011 Apr 15;252(2):112-22. Epub 2010 Nov 21.

8.

PMID: 21315101

Boitier E, Amberg A, Barbié V, Blichenberg A, Brandenburg A, Gmuender H, Gruhler A, McCarthy D, Meyer K, Riefke B, Raschke M, Schoonen W, Sieber M, Suter L, Thomas CE, Sajot N.A comparative integrated transcript analysis and functional characterization of differential mechanisms for induction of liver hypertrophy in the rat. Toxicol Appl Pharmacol. 2011 Apr 15;252(2):85-96. Epub 2011 Feb 26.

9.

PMID:18688244

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Michael Clark A, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK Jr, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novère N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ Jr, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.Nat Biotechnol. 2008 Aug;26(8):889-96.