# An Ontology of Biomedical Investigations

### Authors

The OBI consortium (in alphabetical order): Ryan Brinkman1, Bill Bug2, Helen Causton3, Kevin Clancy, Cristian Cocos, Mélanie Courtot1, Dirk Derom4, Eric Deutsch5, Liju Fan6, Dawn Field7, Jennifer Fostel8, Gilberto Fragoso3, Frank Gibson9, Tanya Gray, Jason Greenbaum10, Pierre Grenon, Jeff Grethe7, Yongqun He11, Mervi Heiskanen, Tina Hernandez-Boussard7, Philip Lord, Allyson Lister12, James Malone, Elisabetta Manduchi, Luisa Montecchi, Norman Morrison3, Chris Mungall13, Helen Parkinson3, Bjoern Peters10, Matthew Pocock, Philippe Rocca-Serra3, Daniel Rubin, Alan Ruttenberg14, Susanna-Assunta Sansone3, Richard Scheuermann15, Daniel Schober16, Barry Smith17, Larisa N. Soldatova18, Holger Stenzhorn, Christian J. Stoeckert Jr.19, Chris Taylor3, Randi Vita10, John Westbrook, Joe White, Patricia L. Whetzel7, Stefan Wiemann, Jie Zheng19

### Affiliations

1BC Cancer Agency, Vancouver, Canada; 2National Center for Microscopy Imaging Research, UCSD, CA, USA; 3The European Bioinformatics Institute, Cambridge, UK; 4Victoria University of Wellington, New Zealand; 5ISB, Seattle, WA, USA; 6Ontology Workshop, LLC, Columbia, MD, USA; 7Stanford University Medical School, Stanford, CA, USA; 8SRA Global Health Sector, Durham, NC, USA; 9Abcam plc, Cambridge, UK; 10La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA; 11University of Michigan, Ann Arbor, USA; 12Newcastle University, Newcastle upon Tyne, UK; 13Lawrence Berkely National Lab, Berkely, CA, USA;14Science Commons, Cambridge, MA, USA; 15Southwestern Medical Center, USA; 16IMBI, University Medical Center, Freiburg, Germany; 17University at Buffalo, Buffalo, NY, USA; 18Aberystwyth University, Wales, UK; 19University of Pennsylvania School of Medicine, Philadelphia, PA, USA

## Abstract

The goal of OBI is to enable a formal representation of biomedical investigations that captures the experimental evidence on which their findings are based. The scope of OBI includes: materials made in and produced for investigations, research objectives, experimental protocols, roles of people in investigations and processing and publication of data gathered in investigations. Use of OBI will allow comparison of experimental data from the wide array of scientific disciplines represented by domain experts in the OBI consortium. OBI follows the principles laid out by the OBO foundry, and integrates tightly with other foundry candidate ontologies, such as GO (www.geneontology.org) and ChEBI (www.ebi.ac.uk/chebi/) whose terms are used to describe biological reality. The use of OBI by the scientific community to represent or annotate their investigations within electronic data resources will facilitate interdisciplinary data synthesis, enable access to their data on the semantic web and improve third-party understanding of information related to life-science and clinical investigations.

## Introduction

Information derived in biomedical investigations is increasingly being captured in structured electronic format and made available through public database resources as a complement to its reporting in traditional print journal publications. For many years, different formats to describe experiments have been developed in parallel, as research communities tended to focus on one type of methodology at a time [1, 2]. However, even though the ‘many communities – many standards’ development methodology meant that each community could create something focused on their needs, these standards overlapped - the common features of different investigations were being described in different ways. This situation lead to difficulty for representing cross- disciplinary experiments, and for those who would do integrative meta-analyses that spanned the work of multiples scientific communities, such as translational medicine efforts. In order to work with different database systems, key features of an investigation need to be described in a common, unambiguous manner. Minimally, this requires shared standards for the representation of the source and preparation of samples used in an investigation, assay methods, and data analyses applied in the course of coming to published conclusions.

To address this need, here we report the development of the Ontology for Biomedical Investigations (OBI; [http://purl.obolibrary.org/obo/obi.owl](http://purl.obofoundryobolibrary.org/obo/obi.owl)), designed to enable consistent representation of investigations across the full range of experiment-based biomedical research. For those unfamiliar with the use of ontologies OBI functions first as a well-designed terminology. As such, OBI provides a standardized vocabulary for entities in biomedical investigations and a textual definition for each entity and is therefore suitable for term or keyword based indexing of data associated with investigations. OBI goes beyond being simply a terminology by also including logical definition that formally describe how different entities relate to each other, and which support more powerful computational use, such as enhanced querying capability. The terms and relations in OBI were defined in an open, community-driven process, which follows the principles promulgated by the Open Biomedical Ontology (OBO) Foundry[3]. With participation from 19 different experimental communities, OBI strives for a cross-disciplinary consensus covering investigations involving basic, translational and clinical research.

The use of OBI by the scientific community to represent or annotate their investigations within electronic data resources such as public data repositories and laboratory information management systems (LIMs) will alleviate need for lower accuracy text mining and enable semantic web searches and third-party understanding of information related to life-science and clinical investigations. While many of the members of the OBI Consortium are involved in the development of public data repositories where OBI is already being incorporated, we welcome and encourage broader community participation in this effort.

## Results

### Development process and framework

OBI was developed to be complementary to, and integrated with, a framework of existing ontologies in the biomedical domain. For example, if an investigation includes taking a human tissue sample from a specified anatomical entity, the definition of the anatomical entity is referenced from the Foundational Model of Anatomy [4]. New terms are only created in OBI if they denote entities specifically related to investigations, such as the process of *taking a sample from an organism.*

Re-using existing ontologies and integrating them into a single framework poses several challenges. Contacts from OBI to other ontology development activities were necessarily established early, and OBI became one of the driving projects in the establishment of the OBO Foundry [3].The OBO Foundry oversees and maintains a set of key principles designed to facilitate integration of biomedical ontologies. Any ontology developed within the OBO Foundry is made freely available in a standardized format, is developed with broad representation and input by the relevant communities, and is orthogonal to the domains covered by other OBO Foundry ontologies. This latter principle ensures that disparate communities do not develop multiple terms referring to the same entity. OBI will tie together different OBO Foundry projects and apply them to the annotation of biomedical investigations.

To provide a foundation upon which to anchor OBI development, BFO [5] was used as a top level classification of entities, and relations were taken from the OBO relations ontology (RO, [6]). These upper-level ontologies ease interoperability of OBI terms with external ontologies based on the same foundation, and they were chosen specifically as they were most aligned with other OBO Foundry candidate ontologies.

Terms were collected by each of the 19 community within the OBI consortium. In addition, term submissions were also accepted through a public tracker. Before terms were incorporated as entities into OBI, they were checked to ensure that they are in scope, not already present, and sufficiently well defined.

A detailed description of the import processes, the metadata and ID policies, the distributed development process, conflict resolution, the release process, tools used for versioning and the use of automated reasoners is given in the Methods section.

### OBI classes and relations

The OBI 1.0 release has XXX classes and YYY relations. The high level class organization of OBI is depicted in Figure 1. The upper level consists of the BFO classes *material entity*, *process*, *role* and *function*, and the class *information content entity.* . Throughout this text, *italics* are used to indicate a reference to a class or a relation. This section gives an overview of several higher-level OBI classes and relations that outline the scope of OBI and illustrate the modeling approach. Formal textual and logical definitions of each class are provided in the Supplemental Material.



Figure 1 – High level structure of OBI classes

#### Material entity

The class *material entity* encompasses all entities made up of matter. It was created in BFO at the request of OBI developers as the union of object, object part and object aggregate. Use of these BFO classes is appropriate if all entities are of similar scale and a simple three level of granularity representation is adequate. However entities relevant to the domain of biomedical investigations, span sizes from molecule to ecosystem. Depending on the level of granularity chosen these might be considered unitary objects or aggregates of smaller parts. By using *material entity* as our root class we avoid committing to the currently inadequate granularity schema in BFO.

Several subclasses of *material entity* are imported from external ontologies: Classes under *organism* are imported from the NCBI taxonomy [7], *anatomical entity* and its subclasses are imported from the Foundational Model of Anatomy [4] or the Common Anatomy Reference Ontology [8], and *molecular entities* hierarchy are imported from ChEBI [ref]. Based on these imports, OBI-specific sub-classes were constructed. For example, *chemical entities in solution* such as a *PBS buffer* can be defined by referencing the molecules from which they are made of (defined in ChEBI), and specifying their *molecular concentration* asa quality. Table X gives the complete class definition of PBS buffer.

Table Example meta data for class Assay. \* indicates mandatory meta data.

|  |  |  |
| --- | --- | --- |
| **Annotation property** | **Example** | **Useage** |
| Preferred Term\* | Assay | The concise, meaningful, and human-friendly name for a class or property preferred by the ontology developers that reflects community usage, or disambiguates the term if necessary |
| Definition\* | A planned process with the objective to capture information about an evaluant. | The official OBI Aristotelian definition, explaining the meaning of a class or property |
| Definition editor\* | PlanAndPlannedProcess Branch | The name of the editor |
| Definition Source\* | OBI Branch derived | An unambiguous and traceable reference to the source of the definition |
| Curation Status\* | Ready for release | The curation status of a class or property, one of uncurated, meta data incomplete, meta data complete, pending final vetting, ready for release |
| Example | Measure the wavelength of light emitted by excited Neon atoms. | A phrase describing how a term should be used |
| Alternative term | Scientific observation, measuring | A synonym for the class or property |

Material entities that come into existence as the result of intentional acts, for example in as the output of some protocol, or which are manufactured, are *processed material.* For example a *specimen* such as a *blood sample* collected in a clinical protocol is a *processed material* as is a *device* which is manufactured with the intent to perform a *function* such asa *flow cytometer*. Some *material entities* such as a *protein* can be produced as the result of applying a protocol, or even manufactured, for example in a recombinant expression system or by synthesis, but are also naturally expressed in an organism. In order to avoid having to assert multiple superclasses, *processed material* is logically defined with necessary and sufficient conditions, namely as exactly those material entities that *specified output of* *processing material* . When OBI is run through an automated reasoner, any class for which it can be determined that every member is a *specified output of processing material* will be inferred as a subclass of *processed material*.

#### Process

OBI’s definition of *processed material* implies the existence of a *process* by which something is intentionally made, *i.e.,* initiated by an agent (typically a person) in order to achieve a certain goal. OBI captures such goals by defining a subclass of *information content entity – objective specification –* subclasses of which specify these goals[[1]](#footnote-1)*.* An example of such a specification is the sentence “Transform the final reaction into competent cells.” part of a site directed mutagenesis protocol[[2]](#footnote-2). Processes that are carried out to meet objective specifications are *planned processes*, and constitute the majority of processes represented in OBI.

Near the top of OBI’s process hierarchy there are several classes that represent realizations of broad objectives, including *material separation,* instances of which are processes such as those where a cell sorter is used to separate out CD8+ cells from a blood sample, and *material combination,* instances of which are processes such as when a chemical is added to a cell culture.

It is within OBI’s scope to represent successfully completed *investigations*, a type of *planned process*. Successful *investigations* may produce negative results, but processes that result in falsified data or during which there were gross errors are not considered successful investigations and are therefore considered to be out of scope for OBI. Therefore, every *planned process*, since it is successful, has achieved one or more goals. OBI defines the relation *achieves planned objective* as holding between instances of *planned processes* and the corresponding *objective specifications*.

While processes can have many participants, objective and plan specifications typically specify which participants are particularly important, calling them specified inputs and specified outputs. OBI defines two additional relations, *has specified input* and *has specified output*, relating process instances to such participants. Other than that these participants are mentioned in plan specifications, we can infer that specified outputs must necessarily be present at the end of the planned process in which they participate. The focus on the desired result of a process allows the use of OBI to model exchangeable paths to a similar outcome.

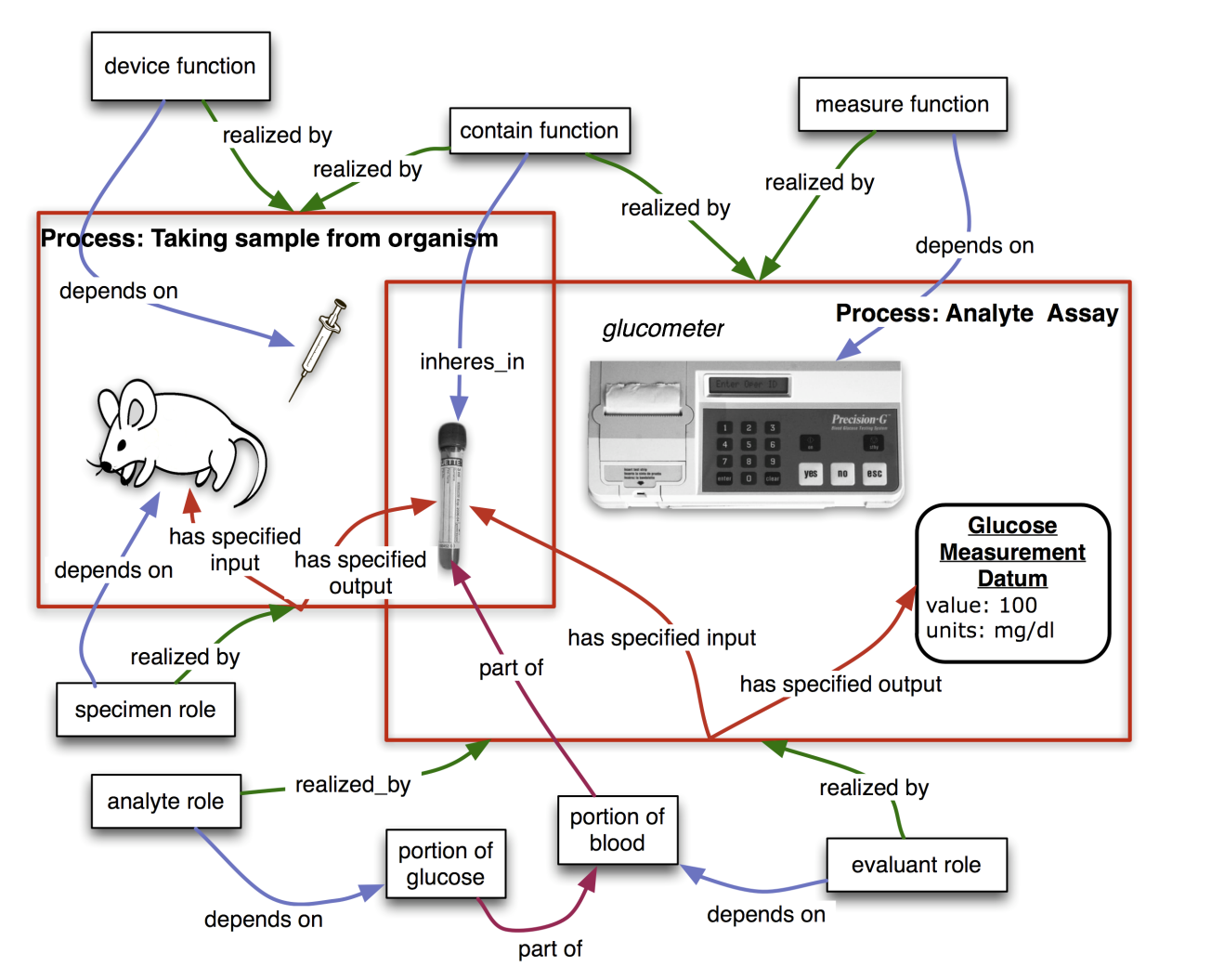
Figure 2 gives an example how several planned processes of drawing blood and measuring the glucose concentration in it are modeled in OBI.

*Information processing* isa subclassof *planned process* distinguished by having information as the input and output of the process. For example, the OBI class *hierarchical clustering* is defined as a *process* which takes as input a collection of objects and builds a hierarchy of clusters. *Hierarchical clustering*, in practice, can be used to achieve two different objectives; to cluster data (OBI class: *class discovery data transformation*) and to partition data (OBI class: *partitioning data transformation*). A hierarchy of planned processes based on their achieved objective permits the user to see immediately how *information processing* processes can be applied to achieve an objective specification.

#### Information artifacts

Any ontology of biomedical investigations needs to cover entities related to information, such as the design of the investigation, collected data, and reports of investigation results. Such entities were not originally in the scope of BFO, and they pose challenges to a realism-based ontology because of X. To facilitate the inclusion of information related entities in OBI while at the same time ensuring that this work will be generally usable outside of the context of biomedical research, the Information Artifact Ontology (IAO, [ref]) was created to serve as a high-level layer between BFO and OBI. IAO is maintained by OBI developers together with other scientists interested in modelling information. It encompasses terms such as *plan specification* and *objective specifications*, which capture the information content of a plan / objective as can be written down on paper. IAO also includes *report* such as the content of a journal article or a patent application, *report elements* such as *report* *table*, *report* *figure*, *scatterplot*, and *software*, which is a series of encoded instructions that can be directly executed by a CPU, or transformed into a form that can be.

Terms for information artefacts that relate specifically to investigations are maintained in OBI. They include *protocol* and *study design*, which are types of plans specifications for either a single procedure or an entire investigation. An important part of many *study designs* is the identification of *dependent* and *independent variables*. Also central to an investigation is the data generated. This includes *measured data item*, which is obtained directly in an assay, and *derived data item,* which is processed through further data transformations. The information generated as the outcome of an investigation is captured by *study result* and *study interpretation*, which follows the distinction typically made in journal articles between the Result and Conclusion section.



**Figure 2. Measuring the glucose concentration in blood.** The largeboxes represent processes and contain their participants. The *taking sample from organism* process takes place first. In this process, a syringe is used as a device to draw blood from the mouse which bears the specimen role. At the end of this process, a tube contains the blood specimen. In a second step, that blood will be used as evaluant in an analyte assay in which the concentration of glucose in the blood will be measured. A glucometer device is used to make this measurement. The analyte role inheres in the glucose molecules scattered throughout the blood specimen. The objective of this planned process is to analyze the analyte (glucose) concentration. This modeling of a specific use case is used as the basis for generation of pattern templates [ref QuickTerms] allowing us to automate following additions of similar analyte measuring assays.

#### Roles and Functions

A role is defined in OBI through two properties: firstly, by the entity which bears the role, and secondly by the process through which the role is realized. A role is not essential to the entity bearing the role, but rather is defined by the circumstances under which the role is realized. For example, humans can bear specific roles relating to investigations. These roles include the *principal investigator role*, realized by leading an investigation, and the *author role*, realized by writing a *report*. Additionally, OBI contains roles defined by the study design of an investigation, such as *study subject role* which can inhere in humans, rodents, plants, sections of tissue, or any object of the study design. Similarly, the *reference role* can be borne by different entities that are used as a reference for other entities being studied. OBI also contains roles defined in specific experimental procedures, such as the *cloning insert role* or *administered material role*.

Functions differ from roles in that they are intrinsic to an entity based on its structural organization. Thus the function of a heart to pump blood is true for all living hearts. Functions within the scope of OBI inhere in devices created for research purposes. For example, a *test* *tube* and a *cage* have the *contain function*, which maintain certain material entities in space, while a preparative centrifuge has a *separation function* is to separate material entities.

#### Organization

*Organizations* such as companies, regulatory agencies, and research institutes need to be modeled in OBI. Placing *organization* into the BFO hierarchy proved controversial, as good arguments were made for treating it either as a material entity or an immaterial kind of social construct (related to other legal entities, which are not currently well described in BFO). To end this controversy, we defined an organization by all the things that are true about it: an organization is a continuant entity which can play roles, has members, and has a set of organization rules. Members of organizations are either organizations themselves or individual people. Members can play specific organization member roles that are determined in the organization rules. The organization rules also determine how decisions are made on behalf of the organization by the organization members.

#### Relations

Most of the relations used in OBI are taken from RO including relations which have been proposed for inclusion into RO but are not yet released (RO\_proposed). When new relations are defined between classes in OBI, these are based on RO relations wherever possible. For example, OBI defines the relation *has specified output* as a sub-relation ofthe RO relation *has participant,*  which relates a process, a continuant, and a time at which the continuant participates in some way in the process [Smith, 2005]. *Has specified output* more specifically requires that the process is a *planned process*, that time t is at the end of the process, and that the presence of the participant is required in the *plan specification* which is realized by the *planned process*.

Another example for an OBI relation is the *is proxy for* relation between two continuant instances c1 and c2, where, within a protocol application, measurement of c1 is used to determine the measurement of c2. A position on a gel *is proxy for* mass and charge of molecule, and a measurement of glucose oxidase activity *is proxy for* the glucose present in a sample.

The OBI-created relations are in effect proposals for inclusion in RO. If over time a relation in OBI is accepted in RO, then OBI will deprecate its own relation and use the one provided by the import of the relation from RO.

### Use Cases

#### In order to evaluate OBI, we describe the use of the ontology against three use cases. The first demonstrates use of the ontology to enable workflow construction via restrictions placed on ontology classes, the second illustrates the rich querying mechanisms that the ontology enables and the third shows the coverage OBI provides for annotating experimental designs.

#### Constructing a Genomic Data Analysis Workflow

Our first use case focuses on the annotation of a GenePattern software module for the purposes of creating a data analysis workflow. GenePattern is a powerful scientific platform which provides access to over 90 genomic analytic and visualization tools (Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP GenePattern 2.0 Nature Genetics 38 no. 5 (2006): pp500-501). This provides interoperability between many popularly used tools within *in silico* analysis, including GeneCluster, Matlab, Bioconductor and many others. Of particular interest to users of this resource is the ability to define a protocol using generic terms that can give appropriate options for each step. For the use case considered here, the user wishes to define a workflow in which the steps 'normalization' then 'class discovery' occur and then finally, visualization of these results. Figure XX illustrates the workflow steps for this use case.

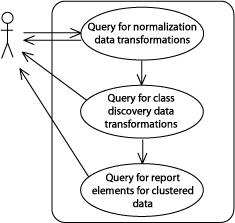


Figure XX. The GenePattern use case for defining a workflow using OBI classes.

In OBI, *data transformation* classes are minimally defined by the following elements: the input data; the output data produced from this process; and the *objective specification* which the data transformation is used to achieve. Note that within OBI, data transformations can have more than one objective, reflecting the fact that a single process can be used to achieve different goals. For this query, data transformations with the objective of normalization and, subsequently, classes with the objective class discovery are required. Finally, data visualization classes that are able to render clustered data (i.e. the output of a class discovery data transformation) are required to satisfy the use case. This is made possible because the output and input data types are modeled explicitly in OBI and relevant restrictions placed on each data transformation class. To complete this workflow, the query of data visualization classes where input data type is clustered data is used. Importantly, the user is able to define generic protocols referencing OBI concepts, instead of specific identifiers such as LSIDs, for the workflow steps. This provides the advantage of a rich set of results (e.g. all class discovery data transformations) and ensures only valid classes that satisfy the query are returned (e.g. only those data visualizations that are able to render clustered data), filtering out the invalid and thereby ensuring that answers are accurate.

#### Use-Case #2: The Immune Epitope Database

As a second use case, we chose the Immune Epitope Database (IEDB, Peters, Nat Rev Immunol. 2007 Jun;7(6):485-90), which catalogs experiments manually curated from the scientific literature that characterize immune epitopes. Over 250,000 experiments from over 6,000 literature references have thus far been entered into the database. Each experiment is described by entries in over 300 database table columns. The IEDB website implements a comprehensive query by example interface to retrieve this data. However, many queries cannot be formulated directly as query by examples, and can require extensive manual post-processing.

It is therefore desirable to allow the user to formulate more expressive queries against the IEDB. As a translation of IEDB data into OBI format exists, this use case evaluates if meaningful queries can be formulated directly in OBI. BP to ADD: To achieve this application specific extension ontology was build extending OBI, but within the OBI ID space. The specific queries considered in this use case are:

1) what assays have been used to measure IFN-gamma production?

“assay and has\_specified\_output some ‘measurement datum’ and ‘is about’ some ‘IFN-gamma production’ “

2) what investigations have been used to study both human and non-human T cell responses?

“investigation and has\_part some (realizes some ‘host role’ and ‘role of’ some ‘homo sapiens’) and has part some (realizes some ‘host role’ and ‘role of’ some ‘mus musculus)”

The IFN-gamma production term is taken from GO, the organism terms (homo sapiens, mus musculus) reference the NCBI taxonomy, while the remaining terms are taken from OBI. Executing this query retrieves XXX (need to finalize results as we finish export. We are struggling with Protégé’s ability to handle large files).

Importantly, such a query mechanism can also be used to seamlessly query across multiple resources. For example, it would be possible to query for certain processes in GO, and link that to a query of the IEDB for assays that give information about these processes.

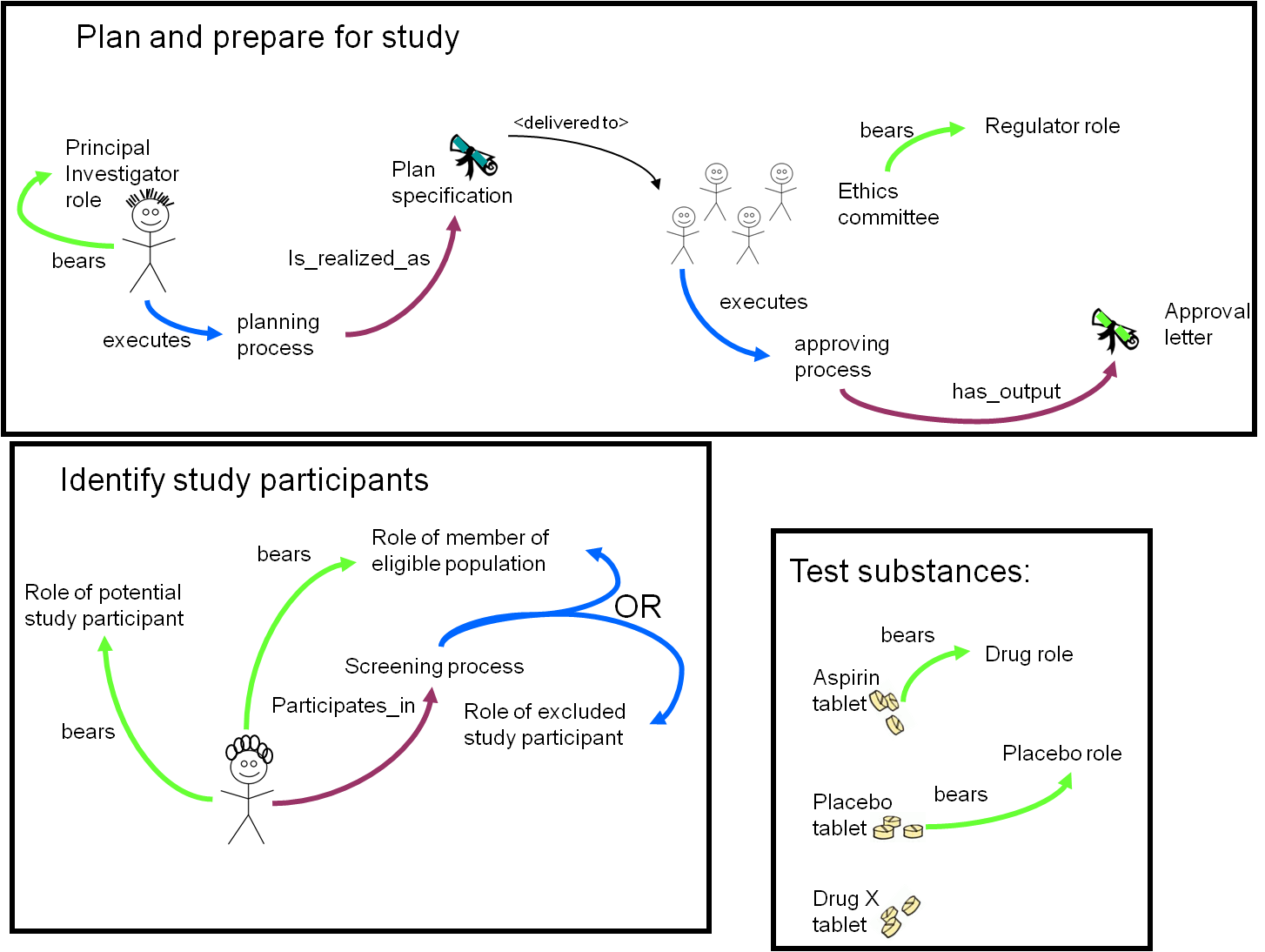
***Alternative / secondary IEDB use case:*** To achieve validity and consistency of the manual curation between the different members of the curation team, a set of curation rules was established (Vita, Cytometry A. 2008 Nov;73(11):1066-70) which is enforced through peer review. It is desirable to enforce curation rules computationally in order to reduce the time spent on peer review, and to decrease the likelihood of curation errors being overlooked. In order to unequivocally communicate such rules between curators and system engineers, the rules are expressed as OBI statements which are meaningful to domain experts, and exact enough to be translated into code by programmers. The specific use case we consider is to formulate a curation rule on how to curate the assay antigen in an infectious challenge assay in OBI. Challenge assays measure the protective efficacy of an immune response against an infectious agent by exposing the host to the infectious agent, and measuring’.

#### Use-Case #3 JF

In this example, a clinician is studying the effects of drug X on platelet aggregation, measured by measured by prothrombin time, the time it takes blood to clot. The example study also contains two control groups, a positive control group who receive aspirin and a negative control group who receive a placebo. This study is modeled on a study reported in an article describing aspirin effects on gingivitis: http://www.rdhmag.com/display\_article/217392/56/none/none/Colum/Gingivitis-and-aspirin, with embellishments to illustrate features of OBI not mentioned in the article. In panel 1, the principal investigator plans the study, secures approval from the institutional ethics committee, and selects patients using inclusion and exclusion criteria defined in the study. The principal investigator has assembled the materials for the study, the three tablets which have been manufactured to have similar appearance so that the study personnel, the agents who carry out the study, are blinded to each treatment. In panel 2, a baseline prothrombin time is measured for each participating patient. This produces original measured data which are given to a statistician who carries out a data transformation to produce a statistical description of the PT times. The PI uses this information to assign patients to cohorts with similar ranges in PT. In panel 3, a registered nurse (RN) gives each subject a tablet every day for seven days. At the end of this time the PT time is measured and used for analysis for significant changes associated with the three regimens. The results are then reported in a journal article.

By representing this use case in OBI we are able to associate the details of the study design and execution with the journal article reporting the conclusion. OBI captures the details of tablet appearance, tablet composition, role of the tablets (test substance, positive control, negative control), and the roles of the study subjects. The qualifications of the study personnel (registered nurse is a qualification regulated by the government; the PI has secured approval from an institutional ethics committee) are also captured so that the data can be re-used by others without questioning the appropriateness of the study execution. The time trigger allows OBI to refer to sequential events in the study (which of the PT time measurements were taken before or after the treatment, the relative timing of any adverse events noted by the RN). The data associated with each patient and the details of the data transformation can also be captured in OBI.







Use-Case #4: PRS

**Use Case: ensuring standard-compliant annotation at community level: MIBBI requirements and ISA tools**

Biology, just like Physics, is now a data intensive field of science. High throughput imaging techniques, parallelized approaches in molecular biology mean that thousands of biological and molecular entities can be routinely assayed. However, instrument output is only so useful unless decorated with contextual experimental information*.* For data managers, the crux lies in achieving consistent and sufficient recording of those tokens of information. Communities have established checklists (Taylor et al, 2008) for an array of techniques and domains devised at ensuring minimal amount of reporting. Achieving consistency however requires bringing together a syntactic framework and a semantic landscape for the description of experimental work. A suite of tools (isatab.sf.net), *ISAconfigurator* and *ISAcreator* provide such capability: *ISAconfigurator* enables implementation of MIBBI requirements leveraging ISA-Tab syntax (Sansone et al, 2008). Data entry fields (ISA-Tab elements) can be configured to consume ontological artifacts, available from brokering services such as OLS and Bioportal (Cote et al, 2008; Noy et al, 2009). This includes restricting to particular portion of any given artifact, for example, the instrument class and children from OBI. OBI is an essential resource as it recapitulates the different life science techniques, their field of application, their inputs and outputs. In the context of reporting instances of experiments, sample collection and assays, OBI provide the semantic anchoring required to achieve consistency in all aspects of the technical components. Once community vetted, MIBBI based ISA configurations can be distributed and can initiate a community based archiving effort using *ISAcreator* to generate ISA-TAB formatted archives. Those may be persisted to the BioInvestigation Index database, which represents an attempt to move away from technology centric data silos while providing the capability of triaging results based on the technology used to produce them and the kind of measurement carried out. Essentially it enables assembly of candidate datasets for meta analysis through MIBBI compliance and reliance on OBI potentially allows to search repositories more effectively for instance to identify all experiments focusing on protein profiling irrespective of the technique used to perform the profiling. Such tools enabling reposition of experimental data while complying with community-validated efforts are praised by funding agencies, publishers and reviewers alike

References:

[Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project.](http://www.ncbi.nlm.nih.gov/pubmed/18688244?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum)

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.

Nat Biotechnol. 2008 Aug;26(8):889-96.

PMID: 18688244

[The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?".](http://www.ncbi.nlm.nih.gov/pubmed/18447634?ordinalpos=2&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum)

Sansone SA, Rocca-Serra P, Brandizi M, Brazma A, Field D, Fostel J, Garrow AG, Gilbert J, Goodsaid F, Hardy N, Jones P, Lister A, Miller M, Morrison N, Rayner T, Sklyar N, Taylor C, Tong W, Warner G, Wiemann S; Members of the RSBI Working Group.

OMICS. 2008 Jun;12(2):143-9.

PMID: 18447634

http://isatab.sf.net

[The Ontology Lookup Service: more data and better tools for controlled vocabulary queries.](http://www.ncbi.nlm.nih.gov/pubmed/18467421?ordinalpos=3&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum)

Côté RG, Jones P, Martens L, Apweiler R, Hermjakob H.

Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W372-6.

PMID: 18467421

[BioPortal: ontologies and integrated data resources at the click of a mouse.](http://www.ncbi.nlm.nih.gov/pubmed/19483092?ordinalpos=2&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum)

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA.

Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W170-3.

PMID: 19483092

## Discussion

OBI provides a way of sharing a common meaning, or semantics, when describing experiments produced by different research groups, institutes, or even experiment types. OBI is a large-scale, community-driven ontology developed to merge experimental terminology across disparate biological disciplines as well as maximize the re-use of existing ontologies outside of the scope of OBI. This is the first OBO Foundry ontology to address the challenge of producing an orthologonal ontology where many overlapping ontologies already existed. As such, a number of challenges were encountered: limitations in the OWL language were discovered and addressed, gaps in the chosen ULOs were identified and filled , and a method for the ongoing collaborative creation, maintenance and release of OBI in a decentralized setting was developed.

During the development of OBI, it was discovered that there were some inherent limitations with the importing of entire ontologies in OWL. Firstly, current editing tools are not effective for working with very large ontologies such as the NCBI Taxonomy [3]. Secondly, some ontologies used by OBI may not be aligned with OBI’s design, for instance if BFO is not used as a ULO or OWL-DL is not used as a language. Importing large or incompatible ontologies in their entirety could lead to inconsistencies or unintended inferences. Instead, the MIREOT import mechanism was developed [ref], which allows for the partial import of external ontology terms into OBI. IAO was developed to include information artifacts within the BFO structure when it became apparent that BFO did not contain the necessary concepts. Collaboration across multiple countries and time zones is problematic for small groups, and becomes quickly unmanageable when 19 communities are involved. In order to efficiently discuss and develop OBI, a combination of teleconferences, versioning and division of labor was used. Additionally, integrating classes from other ontologies raised the need for cross-talk between the different teams building OBO Foundry resources and not just those building OBI. This led to the organization of OBI / OBO Foundry development meetings (held at the EBI with support from BBSRC and possibly ELIXIR project) to precisely identify common issues faced by practitioners and developers.

Prior to the development of OBI, researchers could not use a single common language to describe life-science. Instead, terms were distributed across multiple ontologies and controlled vocabularies from different biological communities. Existing alternatives to OBI include EXPO (ref) and the MGED Ontology (Whetzel et al. Bioinformatics 2006 PMID= 16428806) however these are community specific and more limited scope. Members of both projects are actively working on mapping their existing terms into OBI. By providing standard methods of describing (the semantics of OBI) and structuring (FuGE, http://fuge.sourceforge.net [ref]) data, as well as suggesting a minimal content checklist (MIBBI, http://www.mibbi.org [ref]), sharing experimental data and metadata irrespective of data type becomes easier and more productive. Instead of many standards from many communities, OBI, FuGE and MIBBI provide a new ‘many communities – one standard’ model that simplifies experimental description. Providing researchers with a standard syntax, semantics, and content for experimental description eases computational analyses, encourages collaborations, and aids reproducibility of experiments.

OBI can be used to tag existing data, and provides computationally-amenable definitions for more complex programmatic usages of OBI. In the same way that GO has opened up a whole new level of searching and analysis via the annotation of genes and gene products, OBI provides a method of tagging experimental descriptions with terms that are both understandable to the human reading about the experiment and the computer program designed to analyse 1000s of experiments. Humans, however, have biases about the meaning and usage of commonly-used terms that are not shared across communities. OBI as a result employs terms and definitions designed to avoid these confusions with the cost of sometimes sounding as though the terms were written primarily for computers. To address human accessibility without losing precision, OBI will employ annotation properties such as ‘alternative\_term’ for community preferences.

Add: Future work, conversion of evidence codes in GO into OBI.

OBI was created to be extensible, and to make use of existing ontologies wherever possible. OBI supports future developments of new technologies for use in the biomedical research domain.

## Acknowledgments

This work is partially supported by grant funding from the National Institute of Biomedical Imaging and Bioengineering, National Human Genome Institute, NIH (R01EB005034, NIH P41 HG003619), EC EMERALD project (LSHG-CT-2006-037686), EC FELICS, MUGEN, BBSRC (BB/E025080/1, BB/C008200/1), NERC-NEBC, EU IP CarcinoGenomics (PL 037712), EU NoE NuGO (NoE 503630), CARMEN project EPSRC (EP/E002331/1), the Michael Smith Foundation for Health Research

## Methodology

### Choice of language and metadata conventions

OBI development has spanned the last 5 years and is achieved by means of weekly teleconferences and twice yearly face-to-face meetings. To date 19 communities and XXX researchers have been actively involved in the development process. The requirement for remote development required formalization of the concept meta data (per concept author, progress information etc) conventions as a lightweight meta data schema tailored to development needs. The complete specification of the metadata scheme is available [<http://obi-ontology.org/page/OBI_Minimal_metadata>].

OBI is developed using the Ontology Web Language 1.0 (OWL) [9] as OWL provides richer semantic support than OBO format - the other commonly used alternative in the biomedical domain. The OWL-DL subset was selected to gain maximum expressivity while allowing machine readabilility and use of reasoning for maintenance and development. The metadata scheme is implemented as OWL annotation type properties.

Concurrent development of OBI is necessary given the number of developers. OBI development is therefore organized into branches, with each sub-team working independently. Each branch is organized by subject area, such as data transformations, instruments or roles. Using this strategy tasks such as redundancy or inconsistency checking are manageable, as members in each branch can focus on a subset of terms.

We also required version control, offline development, use of different tools and editors, and script-based augmentation of the ontology content. A review of the existing collaborative ontology development tools failed to identify a single application that met OBI’s requirements. Instead, we split the ontology into multiple OWL files, one per branch, and rely on subversion to manage version control, distributed and offline development, as well as logging history for change management. The current development version of OBI is available from subversion at [ref].

### Integration with existing ontologies

OBI was developed to be complementary to, and integrated with, a framework of existing ontologies in the biomedical domain. The description of experiments and investigations provided by OBI ties into the description of ‘natural’ biomedical information, as produced by the Gene Ontology for gene function *[10]* and others for example in for cell types and anatomy. Contacts to other ontology development activities were therefore established early, and OBI was one of the driving projects in the establishment of the OBO Foundry [3]*.* The desire to re-use existing diverse ontologies and integrating them into a single framework posed several challenges.

A top level ontology can provide guidance in how individual terms or whole ontologies interrelate and is useful when integrating external ontologies. BFO [ref] was chosen as the top level ontology as it is stable, publicly available in OWL syntax, aligned to existing OBO Foundry ontologies, and because there is a responsive community of developers and users of it [cite bfo-discuss list]. The OBO Relations ontology[ref] was chosen provide relations for OBI, and these were extended where required.

Four key ontologies are directly imported using the owl:import mechanism, including BFO and RO mentioned above, plus the ro\_bfo bridge mapping to set ranges and domains for properties using BFO classes, and the Information Artifact Ontology (IAO, [ref]).

The *owl:imports* mechanism was not suitable for all external ontologies for several reasons: Firstly, current editing tools are not effective for working with very large ontologies such as the NCBI Taxonomy [3], therefore direct OWL import is impractical. Secondly, some ontologies used by OBI are activly develoed and may not be aligned with OBI’s design (e.g., not using BFO as an upper ontology, or not using OWL DL). Importing such ontologies as a whole could lead to inconsistencies or unintended inferences. Instead, the MIREOT import mechanism was developed [ref], which is based on the practice of the Gene Ontology (GO) [5] in that the intended meanings of classes remain stable. Even when the source ontology is repaired or reorganized, the effects of such modifications do not change the intended meaning of terms.

### Term requests and triaging for concurrent development

The majority of terms in OBI were collected from the various communities that intend to use it. Term lists were solicited and collated by a representative of each community (OBI consortium page] and terms were also submitted through a public term tracker [(http://sourceforge.net/tracker/?group\_id=177891&atid=886178). Each term was then assessed to determine if it was in the scope of OBI. For example, terms like ‘peptide’ and ‘antibody’ are needed to describe immunological experiments describe an experiment, but were available in ChEBI [11] and GO respectively, and were therefore imported (see above) rather than added to OBI directly.

Submitted terms were distributed to an appropriate OBI branch for curation and inclusion in the OWL file and interaction between branches was achieved through conference calls.

### Release and Quality Control

Users require a traceable, static version of OBI. We have therefore established a monthly release process where multiple branch development OWL files are merged into a single file, which makes it easier to use and view the ontology in currently available tools. A second version of the ontology is released in which classes are organized in the inferred hierarchy. Each release is versioned, and is available from a specific URI using the release date as a tag. The most up-to-date file prior to monthly release is always available from [http://purl.obolibary.org/obo/obi.owl](http://purl.obolibaryfoundry.org/obo/obi.owl). Releases of OBI are also uploaded to the NCBO BioPortal [12] and the OBO Foundry repository [3].

Checks are made prior to release to assure compliance of the released version with OBI policies. Specifically, classes are identified that do not comply with the minimal metadata policy, have invalid OWL syntax or lead to inconsistency when reasoning using Pellet [13] or Fact++[ http://owl.man.ac.uk/factplusplus/]. Identifier format and deprecation policies are also enforced. OBI identifiers are prefixed with “New classes are automatically assigned a permanent identifier in the format OBI\_0000001” on their first release. Similarly, the release process verifies that all IDs that were present in the previous release are still in use. This conforms with the GO deprecation policy that OBI has adopted. Deleted classes are moved under the *ObsoleteClass* in the OBI hierarchy and stored in a separate file.

### References

1. Jones, A.R., et al., *The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics.* Nat Biotechnol, 2007. **25**(10): p. 1127-33.

2. Spellman, P.T., et al., *Design and implementation of microarray gene expression markup language (MAGE-ML).* Genome Biol, 2002. **3**(9): p. RESEARCH0046.

3. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.

4. Rosse, C.a. and J.L. Mejino, Jr., *A reference ontology for biomedical informatics: the Foundational Model of Anatomy.* J Biomed Inform, 2003. **36**(6): p. 478-500.

5. Grenon, P. and B. Smith, *SNAP and SPAN: Towards Dynamic Spatial Ontology.* Spatial Cognition & Computation: An Interdisciplinary Journal, 2004. **4**(1): p. 69-104.

6. Smith, B., et al., *Relations in biomedical ontologies.* Genome Biol, 2005. **6**(5): p. R46.

7. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2006. **34**(Database issue): p. D173-80.

8. Haendel, M.A., Neuhaus, F., Osumi-Sutherland, D., Mabee,P.M. and M. Mejino Jr. J.L.V., C.J. and Smith, B., *CARO - The Common Anatomy Reference Ontology*, in *Anatomy Ontologies for Bioinformatics: Principles and Practice*, D.D.a.R.B. Albert Burger, Editor. 2008, Springer: New York.

9. Horrocks, I., Patel-Schneider,P.F and van Harmelen, F., *From SHIQ and RDF to OWL: The making of a web ontology language.* Journal of Web Semantics, 2003. **1**(1): p. 7-26.

10. Blake, J.A. and M.A. Harris, *The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis.* Curr Protoc Bioinformatics, 2008. **Chapter 7**: p. Unit 7 2.

11. Degtyarenko, K., et al., *ChEBI: a database and ontology for chemical entities of biological interest.* Nucleic Acids Res, 2008. **36**(Database issue): p. D344-50.

12. Musen, M., et al., *BioPortal: Ontologies and Data Resources with the Click of a Mouse.* AMIA Annu Symp Proc, 2008: p. 1223-4.

13. Evren Sirin, B.P., Bernardo C Grau, Aditya Kalyanpur, Yarden Katz, *Pellet: A practical OWL-DL reasoner* Web Semantics: Science, Services and Agents on the World Wide Web, 2007. **5**(2): p. 51-53.

1. We represent *objective* *specifications* rather than objectives because we understand the former better than the latter. This is a case of a repeated pattern in OBI – when faced with difficult representation problems we sometimes define related terms that might be better understood, and which are still adequate to fulfill anticipated uses cases. [↑](#footnote-ref-1)
2. http://www.stanford.edu/~loening/protocols/Site\_Directed\_Mutagenesis.pdf [↑](#footnote-ref-2)