Prepublication version of chapter to appear in Jake Chen and Stefano Lonardi (eds.), *Biological Data Mining*, Boca Raton: Chapman Hall / Taylor and Francis, September 2009, 529-559.

Ontology-based Knowledge Representation of Experiment Metadata in Biological Data Mining

Richard H. Scheuermann[1,2], Megan Kong[1], Carl Dahlke[4], Jennifer Cai[1], Jamie Lee[1], Yu Qian[1], Burke Squires[1], Patrick Dunn[4], Jeff Wiser[4], Herb Hagler[1], Barry Smith[5], David Karp[3]

[1]Department of Pathology, [2]Division of Biomedical Informatics, [3]Division of Rheumatology, University of Texas Southwestern Medical Center, Dallas, TX; [4]Health Information Systems, Northrop Grumman, Inc., Rockville, MD; [5]Department of Philosophy, University of Buffalo, NY

Abstract: According to the PubMed resource from the U.S. National Library of Medicine, over 750,000 scientific articles have been published in the ~5000 biomedical journals worldwide in the year 2007 alone. The vast majority of these publications include results from hypothesis-driven experimentation in overlapping biomedical research domains. Unfortunately, the sheer volume of information being generated by the biomedical research enterprise has made it virtually impossible for investigators to stay aware of the latest findings in their domain of interest, let alone to be able to assimilate and mine data from related investigations for purposes of meta-analysis. While computers have the potential for assisting investigators in the extraction, management and analysis of these data, information contained in the traditional journal publication is still largely unstructured, free-text descriptions of study design, experimental application and results interpretation, making it difficult for computers to gain access to the content of what is being conveyed without significant manual intervention. In order to circumvent these roadblocks and make the most of the output from the biomedical research enterprise, a variety of related standards in knowledge representation are being developed, proposed and adopted in the biomedical community. In this chapter, we will explore the current status of efforts to develop minimum information standards for the representation of a biomedical experiment, ontologies composed of shared vocabularies assembled into subsumption hierarchical structures, and extensible relational data models that link the information components together in a machine-readable and human-useable framework for data mining purposes.

I. General Overview

Paradigm shift in biomedical investigation

The advance of science depends on the ability to build upon information gathered and ideas formulated through prior investigator-driven research and observation. Traditionally, the output of the international research enterprise has been reported in print format – scientific journal articles and books. The intended audiences for these scientific reports are other scientists who carefully read through the text in order to understand the rationale behind the arguments and experimental designs and thereby to gauge the merits of results obtained in addressing the proposed hypothesis. This approach has worked well thus far; it worked well during the advent of molecular biology, when many of the fundamental principles in biology were defined.

However, the last two decades have witnessed a paradigm shift in biomedical investigation, in which reductionistic approaches to investigation in which single functions of single molecules studied using tightly controlled experiments are being replaced by high throughput experimental technologies in which the functions of large numbers of biological entities are evaluated simultaneously. This shift in experimental paradigm was largely initiated when the U.S. Department of Energy, the U.S. National Institutes of Health and the European Molecular Biology Laboratory committed to the sequencing of the human genome. In addition to the information derived from the genome sequence itself, the human genome project spawned the development of new research technologies for high throughput investigation that rely on automation and miniaturization to rapidly process large numbers of samples and to simultaneously interrogate large numbers of analytes. For example, microarrays of probes for all known and predicted genes in the human genome are now commercially available to enable simultaneous measurement and comparison of the mRNA levels of all genes in biological samples of interest. The output of these high throughput methodologies is massive amounts of data about the biological systems being investigated, and this has lead to two challenges – how do we analyze and interpret these data, and how do we disseminate the resultant information in such a way as to make it available to (and thus discoverable by) the broader scientific community?

Data sharing standards

In order to maximize its return on investment, NIH established a policy for data sharing in 2003 (http://grants.nih.gov/grants/policy/data_sharing/), to the effect that for any project receiving more than $500,000 per year in NIH funding, the investigators must make their primary data freely available to the scientific community for re-use and meta-analysis. Although most journals now provide electronic versions of their print articles that also include supplemental files of supporting data, these data files are not always available through open access, nor is it easy to find relevant data sets through these sources. Thus, the U.S. National Center for Biomedical Informatics (NCBI), the European Bioinformatics Institute (EBI) and the Stanford microarray community have each established archives for gene expression microarray data – the Gene Expression Omnibus (www.ncbi.nih.gov/geo) [Barrett 2007], ArrayExpress (www.ebi.uk/array-

express) [Parkinson 2007], and the Stanford Microarray Database (http://genome-www5.stanford.edu/) [Demeter 2007], respectively. Several other institutes at NIH are supporting projects to develop more comprehensive data sharing infrastructures. The National Cancer Institute's caBIG project is working toward the development of vocabulary standards and software applications that will support data sharing using a distributed grid approach. The Division of Allergy, Immunology and Transplantation of the National Institute of Allergy and Infectious Disease (NIAID) is supporting the development of the Immunology Database and Analysis Portal (ImmPort) to serve as a sustainable archive for research data generated by its funded investigators (www.immport.org). The Division of Microbiology and Infectious Disease also of the NIAID has supported the development of eight Bioinformatics Resource Centers for Biodefense and Emerging/Re-emerging Infectious Disease to assemble research data related to selected human pathogens (www.brccentral.org) [Greene 2007, Squires 2008]. The goal of each of these projects is to make primary research data freely available to investigators in a format that will facilitate the incorporation of these data and the information derived there from into new research studies designed to extend previous findings.

Three new interrelated biological disciplines have emerged to address the challenges of data management and analysis – bioinformatics, computational biology and systems biology. While there is some debate about whether these are really distinct disciplines of biology, for the purposes of this chapter we will include in the domain of bioinformatics studies related to defining how laboratory data and biological knowledge relate to each other and how approaches to knowledge representation can aid in data mining for the discovery of new knowledge. We will also make the distinction between data retrieval and data mining, with the former being focused on identifying relevant data sets based on defined characteristics of the experiment (e.g. finding all experiments involving research participants with type 1 diabetes) and the latter being focused on identifying patterns in data sets (e.g. which single nucleotide polymorphisms correlate with the development of type 1 diabetes). Effective data retrieval requires the accurate and standardized representation of information about each experiment in an easily accessible format (e.g. in one or other standard relational database format). While data mining is also dependent on accurate and standardized representation of data, it is also further enhanced when the information incorporates previous knowledge in such a way as to enable identification of relevant patterns (for example through the use of Gene Ontology annotation to interpret gene expression patterns in microarray data).

Discussions of data mining tend to focus on the algorithmic portions of the technique. In this chapter we will focus the discussion on how data standards can help support more effective data mining by providing common data structures to support interoperability between data sources and to provide consistent points of integration between disparate data set types. When sharing data between individuals, the use of standards ensures an unambiguous understanding of the information conveyed. For computer programming, the use of standards is essential. In order to accomplish these objectives, data standards should be:

- Useful - provide an aid to storing, sharing and re-use of data;
- Consensual - agreed upon by a plurality of users;
- Flexible and evolvable - accommodate all forms of current and future data types;
- Comprehensible - understandable by users (and computers);
- Easy to implement - straightforward to use in software development;
- Widely adopted - they have to be used.

Four related standards will be discussed that, taken together, are necessary for unambiguous and consistent knowledge representation:

- proposals for the collection of *minimum data elements* necessary to describe an experiment (what information should be provided),
- common *ontologies* for the vocabulary of data values that will populate these data elements (how that information should be described),
- *data models* that describe the semantics of how the data elements and values relate to each other (how the information relates to each other), and
- standards that describe the common syntax (format) for *data exchange* (how the information should be transferred between information technology resources).

The chapter will end with an example of how these standards support a type of data mining that we term meta-mining, in which biological knowledge is integrated with primary experimental results for the development of novel hypotheses about the biological systems under evaluation.

II. Minimum Data Elements

The MAIME paradigm

Reports of experimental findings and their interpretations published in scientific journals routinely contain specific sections in which certain types of content are provided. The Methods section includes details about how specific assays and other procedures were performed and the materials to which those procedures were applied. The Results section contains information about the design of individual experiments and the data derived. The Introduction section sets the stage by summarizing the current state of the field and setting out the issues that remain unresolved and that will be addressed in the studies described. The Discussion section provides an interpretation of the experimental findings in the context of the body of knowledge outlined in the Introduction. The Abstract section summarizes the key points of the other sections. While this framework provides some general guidance as to what kind of information should be included in a scientific report, the details concerning what is to be included in each section are left to the authors to decide, thus resulting in considerable variability in the content, structure and level of detail of the information reported. Since there is general agreement that sufficient information should be provided to allow other investigators to reproduce the reported findings, the problem is not so much that important information is missing from scientific publications, but rather that the key information is provided in haphazard, unstructured,

and inconsistent ways, requiring readers to distill and organize the relevant content of interest to them.

While this approach to knowledge representation still has its place in the body of scientific investigation, the advent and widespread use of high throughput experiment methodologies has lead to the need to both capture the experimental results in archives and to describe the components of experiments in a standardized way in order to make the data more easily accessible. The importance of these kinds of minimum information check lists for describing the experiment metadata was recognized by the gene expression microarray community, and formalized in the Minimum Information About a Microarray Experiment (MIAME) recommendations [Brazma 2001]. The MIAME recommendations have since been adopted by many scientific journals and microarray archive databases as the de facto standard for reporting the experiment metadata associated with microarray results. Since then a variety of different communities have proposed similar minimum information check lists to capture the unique aspects of their favorite methodologies (e.g. MIFlowCyt for flow cytometry – http://flowcyt.sourceforge.net/) [Lee 2008].

MIBBI and MIFlowCyt

In order coordinate efforts in the development of these minimum information checklists, the Minimum Information for Biological and Biomedical Investigations (MIBBI) project (http://www.mibbi.org/) was established by a consortium of investigators from various research communities in order to standardize the content of these data standards and to encourage the re-use of common information elements across methodologies where appropriate [Taylor 2008]. Figure 1 shows the required minimum data elements identified in the MIFlowCyt standard. Many of the data elements included correspond to basic elements of experimental design (e.g. experiment purpose and dependent and independent variables) and assay procedures (e.g. biological sample source and treatment details); these kinds of data elements are included in most MIBBI standards as they serve as a common core for biological experiment descriptions. The MIBBI consortium is currently identifying these core data elements that should be consistently represented in all MIBBI-compliant minimum information standards. Other data elements that relate to the kinds of reagents that are used to measure analytes (e.g. fluorochrome reporters and antibody clone names), measurement instrument configuration details (e.g. flow cell and optical filters), and data analysis details (e.g. compensation and gating details) are more technology specific and may only be found as extensions to the common MIBBI core in a few related methodology standards.

Two important points about these kinds of minimum information standards are worth noting. First, there is a distinction between what information is necessary to reproduce an experiment, as detailed in the MIBBI standards, and the information that would be relevant to capture and support in a database archive. The latter correspond to a subset of MIBBI data elements that might specifically be represented in database tables and would be used to query the database, especially the dependent and independent variables of the experiment, the characteristics of the biological samples and their sources used in the

experiment, and the analytes being measured. While the other information is equally important to reproduce the experimental findings, they may not play important roles in the conduct of data meta-analysis. For example, while the instrument configuration details may be necessary to reproduce a particular data set, it is unlikely that one would search for data sets in the database archive based on the details of the optical path. Rather than capturing these details in specific database table columns, this information can be included in text documents that describe all of the protocol details.

The second important point about these minimum information standards relates to who should be responsible for providing the information. In the case of MIFlowCyt, some of the information is derived from the configuration of the instrument and analytical software used in the capture of the resulting data. This information would more appropriately be provided directly by the instrument and software packages themselves, rather than expecting the investigator to have to determine these details for every experiment they run. Thus, in the development of the MIFlowCyt standard, it was important to engage stakeholders from the instrument manufacturer and software developer communities so that they would agree to provide this information as a matter of course in the resulting output files.

Thus, by formalizing the details for the minimum data elements that should be included in the description of an experiment, a higher level of consistency in how to describe experiments can be obtained, both within and between different experimental methodologies. Consistent representation frameworks will facilitate the identification of related experiments in terms of health conditions being investigated, treatment approaches being varied, and responding variable being tested in order to support meta-analysis and re-use of related data sets.

Information Artifacts and MIRIAM

Although the biomedical minimum information standards were initially developed to support the description of wet lab experimentation, it became apparent that similar standards would be useful to support the work coming out of the bioinformatics research community in terms of the description of system models and data mining analysis. The BioPax [Luciano 2005], SBML [Hucka 2003] and CellML [Lloyd 2004] standards have been developed to provide the syntactic standards necessary to exchange biological pathway and systems descriptions. The Minimum Information Requested in the Annotation of biochemical Models (MIRIAM) [Le Novere 2005] was developed to ensure that sufficient information is included in the description of any biological model such that any results obtained from modeling could be reproduced by outside investigators. The Minimum Information About a Simulation Experiment (MIASE) extends MIRIAM to support model simulation. Recently, the European Bioinformatics Institute has established a set of resources based on the MIRIAM standards [Laibe 2007], including:

- *MIRIAM Database* - containing information about the MIRIAM data types and their associated attributes;
- *MIRIAM Web Services* - a SOAP-based application programming interface (API)

for querying the MIRIAM Database;

- *MIRIAM Library* - a library instruction set for the use of MIRIAM Web Services;
- *MIRIAM Web Application* - an interactive Web interface for browsing and querying MIRIAM Database, and for the submission and editing of MIRIAM data types.

III. Ontologies

Ontologies versus controlled vocabularies

While the minimum data standards describe the types of data elements to be captured, the use of standard vocabularies as values to populate the information about these data elements is also important to support interoperability. In many cases, groups develop term lists (controlled vocabularies) that describe what kinds of words and word phrases should be used to describe the values for a given data element. In the ideal case each term is accompanied by a textual definition that describes what the term means in order to support consistency in term use. However, recently many bioinformaticians have begun to develop and adopt ontologies that can serve in place of vocabularies for use as these allowed term lists. As with a specific vocabulary, an ontology is a domain-specific dictionary of terms and definitions. But an ontology also captures the semantic relationships between the terms, thus allowing logical inferencing about the entities represented by the ontology and by the data annotated using the ontology's terms. The semantic relationships incorporated into the ontology represent universal relations between the classes represented by its terms based on knowledge about the entities described by the terms established previously. For example, if we use a disease ontology to annotate gene function that explicitly states that type 1 diabetes and Hashimoto's disease are both types of autoimmune diseases of endocrine glands, then we can infer that gene A, annotated as being associated with type 1 diabetes, and gene B, annotated as being associated with Hashimoto's disease, are related to each other even though this is not explicitly stated in the annotation through our previous knowledge of disease relationships captured in the structure of the disease ontology used for annotation. Thus, an ontology in the sense here intended is a representation of the types of entities existing in the corresponding domain of reality and of the relations between them. This representation has certain formal properties, enabling it to serve the needs of computers. It also employs for its representations the terms used and accepted by the relevant scientific community, enabling it to serve the needs of human beings. To support both humans and computers the terms used are explicitly defined using some standard, shared syntax. Through these definitions and through the relations asserted to obtain between its terms the ontology the ontology captures consensual knowledge accepted by the relevant communities of domain experts. Finally, an ontology is a representation of universals; it described what is general in reality, not what is particular. Thus, ontologies describe classes of entities whereas databases tend to describe instances of entities.

In recognition of the value that ontologies can add to knowledge representation, several groups have developed ontologies that cover specific domains of biology and medicine.

The Open Biomedical Ontology (OBO) library was established in 2001 as a repository of ontologies developed for use by the biomedical research community (http://sourceforge.net/projects/obo). As of August 2008 the OBO library includes 70 ontologies that cover a wide variety of different domains. In some cases, the ontology is composed of a highly focused set of terms to support the data annotation needs of a specific model organism community (e.g. the Plasmodium Life Cycle Ontology). In other cases, the ontology covers a broader set of terms that is intended to provide comprehensive coverage of an entire life science domain (e.g. the Cell Type Ontology). Since 2006, it has become possible to access the OBO library through their BioPortal http://www.bioontology.org/tools/portal/bioportal.html) of the National Center for Biomedical Ontology (NCBO) project, which also provides a number of associated software services and also access to a number of additional ontologies of biomedical relevance. . The European Bioinformatics Institute has also developed the Ontology Lookup Service (OLS) that provides a web service interface to query multiple OBO ontologies from a single location with a unified output format (http://www.ebi.ac.uk/ontology-lookup/). Both the BioPortal and the OLS permit users to browse individual ontologies and search for terms across ontologies according to term name and certain associated attributes.

OBO Foundry

While the development of ontologies was originally intended to advance consistency and interoperability in knowledge representation, the recent explosion of new ontologies has threatened to undermine these goals. For example, in some cases multiple ontologies that have been developed independently cover overlapping domains, thus leading to different terms or single terms with different definitions being used to describe the same entity by different communities. While this problem can be partly resolved by efforts to map between terms in different ontologies, in many cases the lack of a one-to-one mapping makes this problematic. Moreover, mappings themselves are difficult to construct, and even more difficult to maintain as the mapped ontologies change through time. A second problem is that in many cases the relationships between terms that are used to assemble a single ontology hierarchy are not described explicitly and are not used consistently. It is impossible to support inferencing based on the ontology if it is unclear how adjacent terms in the hierarchy relate to each other. Finally, some ontologies have been developed by small groups from what may be a highly idiosyncratic perspectives and thus may not represent the current consensual understanding of the domain in question.

In order to overcome these and other problems with the current collection of biomedical ontologies, several groups have proposed frameworks for disciplined ontology development (e.g. [Aranguren 2008]. The Open Biomedical Ontologies Foundry initiative (http://www.obofoundry.org/) was established in 2005 as a collaborative experiment designed to enhance the quality and interoperability of life science ontologies, with respect to both the biological content and its logical structure [Smith, 2007]. The initiative is based on the voluntary acceptance by its participant ontology development communities of an evolving set of design and development principles designed to maximize the degree to which their ontologies can support the broader needs of scientists.

These best-practice design principles can be roughly sub-divided into three broad categories – technical, scientific and societal (Table 1). Technical principles include requirements for the inclusion of a common set of meta-data (in a manner similar to the MIBBI-type specifications described above), and the use of a common shared syntax. Scientific principles include the requirement of orthogonality to the effect that the content of each Foundry ontology should be clearly specified and delineated so as not to overlap with other Foundry ontologies, and the requirement for a consistent semantic framework for defining the relations used in each ontology. (These scientific principles are described in more detail below.) Societal principles would include the requirement that the ontology be developed collaboratively, and that the resulting ontology artifact is open and freely available.

Table 1. OBO Foundry Principles as of April 2006

Technical	
	The ontology is expressed in a common shared syntax for ontologies (e.g. OBO format, OWL (Ontology Web Language) format).
	The ontology possesses a unique identifier space.
	The ontology provider has procedures for identifying distinct successive versions.
	The ontology includes textual definitions for all terms.
	The ontology is well documented.
Scientific	
	The ontology has a clearly specified and clearly delineated content.
	The ontology uses relations that are unambiguously defined following the pattern of definitions laid down in the OBO Relation Ontology.
Societal	
	The ontology is open and available to be used by all without any constraint.
	The ontology has been developed collaboratively with other OBO Foundry members.
	The ontology has a plurality of independent users.

As argued above, the great value in using ontologies to represent a particular data set lies in the background knowledge embedded in the relationships that link the terms together in the ontology. OBO Foundry ontologies are expected to utilize relations defined in the Relation Ontology (RO) [Smith 2005] to describe these relations (the first set of relations is depicted in Table 2).

Table 2. OBO Relation Ontology (RO)

Foundational	
	is_a
	part_of
Spatial	
	located_in

	contained_in
	adjacent_to
Temporal	
	transformation_of
	derives_from
	preceded_by
Participation	
	has_participant
	has_agent

The foundational relation that is used primarily in the assembly of OBO Foundry ontologies is the is_a relation that links parent and child terms in the ontology hierarchy. Parent and children terms in the is_a hierarchy can be thought of as having type-subtype relations similar to the genus-species relations in the species taxonomy. Several advantages arise out of building the ontology structure based on an is_a hierarchy of type-subtype relations, including:

- First, definitions of terms can be constructed using the genus differentia approach proposed by Aristotle, so that the definition of the term 'A' will take the form: 'An A is_a B that C's' in which A is a subtype (child) of (parent) type B with the special characteristic C that distinguishes instances of A from other instances of B. For example, 'type 1 diabetes is_a autoimmune disease of endocrine glands that involves the endocrine pancreas as the primary target.'
- Second, terms in a well-formed is_a hierarchy inherit characteristics from their parents through the property of transitivity. By defining type 1 diabetes as a subtype of autoimmune disease of endocrine glands, the term inherits characteristics that define all such autoimmune diseases, as well as characteristics of all diseases in general through the definitions of and other attributes associated with these parent terms. Indeed, adherence to the property of transitivity can be a good test for correct positioning of terms in the ontology hierarchy.

Role, quality, function and type

During ontology development, there is often difficulty in trying to define which sort of characteristic should be used as the primary differentia between sibling subtypes. For example, let's consider the type: old, faded blue Dodge Caravan™ minivan airport shuttle that we might want to represent in a vehicle ontology. Should it be considered to be a subtype of old vehicles, a subtype of faded blue vehicles, a subtype of Dodge vehicles, a subtype of Dodge Caravan™ vehicles, a subtype of minivan vehicles, a subtype of airport shuttle vehicles, or a subtype of all of these parent terms?

In order to address this issue, we need to discuss the distinctions between roles, qualities, functions and types. A role is a special attribute that an entity can be made to play by societal choice. The 'airport shuttle' attribute is an example of a role that the driver has assigned to the vehicle. It is not an inherent property of the vehicle that distinguishes it from other vehicles. Indeed, any vehicle could be used as an airport shuttle. The role is also frequently a transient attribute. For example, at the end of the driver's shift the

'airport shuttle' might transform into a 'soccer practice shuttle', and then back into an 'airport shuttle' the next day. The transient, subjective nature of roles makes them a poor choice for primary differentia in ontology hierarchies.

In the example, 'old' and 'faded blue' describe *qualities* of the vehicle. Qualities are not acquired by choice in the natural world. We cannot choose our age or the color of our skin. And yet qualities are frequently transient in nature. At one point the vehicle was a new, bright blue Dodge Caravan™. Thus, basing annotations on terms distinguished based on quality characteristics would mean that the annotation would not be invariant and would have to be undated continually to deal with these changes over time. In addition, entities can be described on the basis of a whole range of different quality characteristics – height, width, length, volume, shape mass, color, age, smell, etc. Without selecting a single defining characteristic, the ontology would explode into a hierarchy of multiple inheritance, in which specific terms would have parents like: old things and blue things, large things and oily-smelling things, and so on.

We conclude that type-subtype is_a relations should be based on properties of the entity that are invariant. A Dodge Caravan™ will always be a minivan regardless of whether it is used as an airport shuttle, whether it is old or new, whether it is painted red or blue, and so forth. It will never be a sports car.

For a more biological example, consider the use of the EcoR1 enzyme in the construction of a recombinant plasmid in a genetic engineering experiment. We would consider EcoR1 to be a **type** of protein with a molecular **function** restriction endonuclease activity (GO:0015666). In its normal context, EcoR1 plays a **role** in the DNA restriction-modification system (GO:0009307) that protects an organism from invading foreign DNA by nucleolytic cleavage of unmethylated foreign DNA. However, EcoR1 can also play another **role** in an experimental context, its use to open up a double-stranded circular plasmid to accept the insertion of a foreign DNA fragment in a cloning experiment. At the end of the cloning experiment we may want to change the **quality** of the enzyme from active to inactive through a denaturation process in order to prevent it from realizing its function any further. Thus, while the type of protein and its designed function haven't changed, its role can change based on the process it is involved in and its quality can change dependent on its physical structure state in this case. By precisely distinguishing between roles, functions, qualities and types we can support the accurate representation of entities in the their normal states and in artificial experiment contexts, and accurate reasoning about theses entities in these different contexts.

Orthogonality

The second major scientific OBO Foundry principle relates to the concept of orthogonality. The Foundry is striving toward complete coverage of the entire biological and biomedical domain using one and only one term for a given entity. However, it would be virtually impossible to build a single ontology that covers the entire scope of this domain in a reasonable amount of time with a manageable group of developers who have the requisite expertise in all disciplines of biology. For these reason, the OBO

Foundry has adopted a modular, iterative approach in which smaller subdomains are defined that then become the focus of activity for groups of ontology developers with appropriate expertise. For example, the Chemical Entities of Biological Interest (ChEBI) ontology is being developed by biochemists and organic chemist with knowledge of chemical structure and function in order to cover the small molecules (e.g. drugs, metabolites, peptides, etc.) of interest to biologists. While this modular approach addresses the challenges of biology domain coverage, it brings the problem of potential overlap between subdomains being developed by different groups. Thus, ontologies that are part of the OBO Foundry must submit to the principle of orthogonality in which a given biological entity is covered by one, and only one, ontology. In cases, where potential overlap exists, negotiation and consensus building is used to assign terms to a given ontology.

In order to bring some level of consistency in the definition of a subdomain module, the biology domain can be divided into partitions based on two axes (Figure 2). The first axis relates to size/granularity, e.g. from molecules to organelles to cells to tissues to organisms to populations to ecosystems. The second axis reflects the general types of entities found in reality as represented in the Basic Formal Ontology (BFO). At the highest level, these entities can be broken down into continuants and occurrent. Continuants are further subdivided into dependent and independent continuants. Continuants exist throughout time. An independent continuant exists on its own without any dependence on another entity; these are the physical objects like tables, cups, proteins, organs, etc. A dependent continuant exists throughout time, but requires adherence in an independent continuant to exist; these are the qualities, roles and functions like the color blue, which only exists I the context of a physical entity. Occurrents are processes, like driving and replication, which exist in a defined time period with a start point and end point. Thus, we can subdivide the biology domain based on grid in which one axis corresponds to size/granularity and the other to time and entity dependencies (Figure 2).

Through the OBO Foundry initiative, the goal is to achieve, in the long term, complete coverage of the entire biological domain with single terms for each entity of interest, in which terms are initially linked together using foundational relations (is_a and part_of) into a single hierarchy in a given ontology, which is developed by groups of subdomain experts and reflects the consensual knowledge of the general relations of types of entites in that domain. These terms can then be used to annotate database records in an unambiguous way that supports inference based on the consensual knowledge incorporated into the ontology structure and thus supports database interoperability.

IV. Data Models

The Immunology Database and Analysis Portal (ImmPort) is being developed to serve as a long-term sustainable archive for data being generated by investigators funded by the Division of Allergy, Immunology and Transplantation (DAIT) of the U.S. National Institute of Allergy and Infectious Diseases (NIAID). DAIT funds a wide range of basic

laboratory and clinical research studies and so the ImmPort system must be able to handle everything from genotyping and gene expression microarray data to clinical trials of new vaccines and therapeutic strategies. As such, ImmPort must be able to manage data associated with a variety of different experiment methodologies and must be able to effectively integrate these data through common metadata features and/or results characteristics. For example, investigators may want to aggregate data from any experiment in which type 1 diabetes is being investigated, or experiments in which some characteristic of particular gene (e.g. TLR4) was found to be significantly associated with the independent variable of the experiment. Thus, the many of the requirements for knowledge representation associated with ImmPort are distinct from those associated with other database archives focused on data from single experiment approaches, like dbGAP for human genetic association data or GEO and ArrayExpress for microarray data. The challenge to support such a wide range of research data lead us to implement a database structure that would reflect the common features of a biomedical investigation.

Several different data model frameworks have been developed over the years – the hierarchical model (used in IBM's IMS database management system), the network model (used in IDS and IDMS), the relational data model (used in IBM DB2, Oracle DBMS, Sybase, Microsoft Access), the object-oriented model (used in Objectstore and Versant) and hybrid object-relational models [Ramakrishnan 2003]. Depending on the application, each of these frameworks has its advantages and disadvantages, but the relational data model has become widely adopted for databases such as ImmPort in that it can handle complex interrelated data in an efficient manner.

The development of a database involves six major steps:

- Requirements Analysis – in which an understanding of what the data include and how it will be used are defined;
- Conceptual Database Design – in which the entities and their relationships are defined;
- Logical Database Design – in which the conceptual design is converted into a logical model based on the data model framework chosen;
- Schema Refinement – in which the logical model is analyzed to identify potential problems and refined accordingly;
- Physical Database Design – in which the logical model is converted into a physical database schema that incorporate design criteria to optimize system performance;
- Application and Security Design – integration of the database with other software applications that need to access the data.

Because the uses of biomedical data varies by data type and the specific requirements of the user communities involved, it is virtually impossible to development a single physical database design that will efficiently meet all user requirements. And yet, in order to support interoperability between database resources common approaches for data representation is essential. For these reasons several groups have worked on the development of data models that capture the kinds of entities and relationships in

biomedical data that are universal and application independent at the conceptual level. These conceptual models can then be refined in their conversion into physical database schemas optimized to support specific use case applications while still incorporating the common entity types and relationships that allow effective data sharing between data users and data providers [Bornberg-Bauer 2002].

BRIDG

Several groups have attempted to develop a data model based on this principle. In the clinical domain, the Biomedical Research Integrated Domain Group (BRIDG) project is a collaborative initiative between the National Cancer Institute (NCI), the Clinical Data Interchange Standards Consortium (CDISC), the Regulated Clinical Research Information Management Technical Committee (RCRIM TC) of Health Level 7 (HL7), and the Food and Drug Administration (FDA) to develop a semantic model of protocol-driven clinical research [Fridsma 2008]. It was developed to provide an overarching model that could be used to harmonize between various standards in the clinical research and healthcare domains. It includes representation of "noun things" like organizations, participants, investigators, drugs, and devices, 'measurement things' like physical exam assessments, and 'interpretation things' like adverse event determination.

FuGE

In the basic research domain, the FuGE (Functional Genomics Experiment Model) is a generic data model to facilitate convergence of data standards for describing high-throughput biological experiments [Jones 2007]. Development of FuGE was initially motivated by analysis of a well-established microarray data model: MAGE-OM. The initial goal of FuGE is to deliver a more general model than MAGE-OM by removing concepts that were specific to microarray technology. After receiving a wide range of use cases from different communities on describing experimental designs across multi-omics and conventional technologies, FuGE developers further generalized the model and added more placeholders in a way to broaden the application scope of the model. Current FuGE v1 model aims at not only a generic database schema but also a data exchange standard, if the model can be widely adopted. In order to capture not only common but also domain-specific experimental information, FuGE is designed to be extensible. Its core model consists of a set of generic object classes to represent the common information in different laboratory workflows and experimental pipelines used in high-throughput biological investigations, while domain-specific extensions of the FuGE core classes are needed to capture specific information requirements in the domain. Due to its extensible characteristic in providing formal and generic data representations, FuGE has been adopted by the Microarray and Gene Expression Data (MGED) Society, the Human Proteome Organization - Proteomics Standards Initiative (PSI), Genomics Standards Consortium (GSC), Metabolomics Standards Initiative (MSI) etc.

Object classes in FuGE v1 are organized under two namespaces: common and bio. There are six packages under namespace common: Audit, Description, Measurement, Ontology, Protocol, and Reference. Bio namespace consists of four packages: ConceptualMolecule, Data, Investigation, and Material. Each package has a set of predefined classes that can

be either used to describe experimental information directly or re-used in domain-specific extensions through inheritance, depending on the nature of the information and the class to be used. Not only entities like instruments, samples, and data files but also relationship among the entities can be described. For example, sample and data can be linked together in FuGE through protocol and protocolApplication, the latter of which provides a flexible binding between input and output as well as related parameters and the use of instrument. To facilitate data sharing, ontology terms and external resources can be referenced in FuGE to annotate or directly describe the objects. Another benefit of using FuGE is if an extension of FuGE follows MDA (Model-Driven Architecture) standard, a FuGE-compliant database schema (XSD) as well as software code can be automatically generated. A typical extension example is FuGEFlow (http://wiki.ficcs.org/ficcs/FuGEFlow), a recent extension of FuGE to capture MIFlowCyt-compliant information for flow cytometry experiments. The generated data schema, called Flow-ML, is planned to be used to help building flow cytometry databases and exchanging flow cytometry experimental information among different labs and institutions.

Ontology-Based eXtensible Data Model (OBX)

In order to leverage the efforts of the MIBBI and OBO Foundry community, we have recently developed a data model – the Ontology-Based eXtensible Data Model (OBX) that reflects many of the design principles incorporated in these data standards. Of particular importance to ImmPort database development, a consortium of different research communities has been working on the development of an ontology of terms needed to represent experiment metadata - the Ontology for Biomedical Investigation (OBI; http://purl.obofoundry.org/obo/obi/). The OBI ontology is focused on those aspects of biology that are directly related to scientific investigations – experiment design, protocol specification, biomaterial isolation/purification, assays for the measurement of specific analytes in specimens, instruments and reagents used in these measurement assays, etc. The OBI ontology hierarchical structure has been built on a BFO framework and points to, but does not include, terms from other OBO Foundry ontologies. Several important concepts have been incorporated into the OBI structure, including the typing of assay, biomaterial transformation and data transformation processes based on the kinds of entities that serve as inputs and outputs to those processes, the typing of investigation specifications based on objectives, and the importance of precisely defining the roles played by the various different process components (e.g. specimen, reagent, principle investigator). While the OBI ontology is focused on representing general features of entity classes used in investigations, OBX is focused on capturing instance level information about specific investigations.

A UML diagram of the high level entities represented in OBX is depicted in Figure 3. The major axis in the model includes objects – events – qualities, which reflects the major branches of the BFO, namely independent continuants – occurrents – dependent continuants (Figure 3A). Objects include entities like biological specimens, laboratory equipment, assay reagents, chemicals, etc. Procedures are types of Events, and are defined based on their inputs and outputs. For example, biomaterial transformation

procedures have biomaterial objects as inputs and outputs, whereas assay procedures have biomaterial object inputs and quality outputs (i.e. the assay results). In addition to describing the specific input entity to a given procedure, the role that the entity plays in the procedure is also captured. This allows for the distinction between the use of an antibody as an assay reagent versus the role of an antibody as an analyte to be measured in a given specimen. Every event occurs in a spatial-temporal context and so the Event table is linked to the Context table to capture these event attributes. In the case of OBX, all events also occur during the conduct of a research study, which is specified through a study design; the study and the study specification are kept distinct to accommodate deviations from the original design.

Thus, the high-level core set of entities includes processes/events, process specifications, objects and qualities. Each of these high-level tables is then connected to a series of lower level, related tables. The high-level tables include attributes in common to all related entities, whereas the low-level tables include attributes that are specific to the related entity types. For example, the Objects table is connected to a series of sub-tables for Human Subjects, Organ/Tissue, Instrument, Compounds/Agents, etc. (Figure 3B), which include attributes specific for the given entity type (e.g. Compound Manufacturer for the Compound/Agent entities). Assays include both clinical assessments/ physical exams as well as laboratory tests. Therapeutic intervention is a type of material transformation in which the input is a human subject and a compound formulation and the output is a treated subject. Diagnosis is a type of data transformation in which the variety of data inputs from laboratory test results and clinical assessments are processed (i.e. transformed) into a diagnosis by the clinician/diagnostician. A detailed description of the OBX UML model can be found at http://pathcuric1.swmed.edu/Research/scheuermann/OBX.html.

An example of how the OBX framework can be used to represent a laboratory experiment is shown in Figure 4. The experiment comes from a published study in which the immune response to influenza virus infection is assessed by measuring the levels of interferon gamma in the lungs of infected mice. The first protocol application is the generation of a lung homogenate from infected mice, which can be thought of three ordered biomaterial transformations – the infection of the mouse, the removal of the lung specimen and the generation of the lung homogenate. In each case the output from the previous process serves as the input for the subsequent process. The process inputs are described as playing specific roles, e.g. host and infectious agent. The first process is a merging (mixing) type of biomaterial transformation, whereas the second two are partitioning (enriching) type of biomaterial transformations. The second protocol application is the measurement of interferon gamma levels in the lung homogenate, which is composed of three ordered sub-processes – the ELISA assay used to derive output data (OD590) as a surrogate of the analyte (interferon gamma) concentration in the specimen, the standard curve interpolation data transformation in which the OD590 value is transformed into interferon gamma mass amount, and final a simple mathematical data transformation to convert the mass amount into an analyte concentration for the original input specimen. Again the input components are described

to play specific roles in the processes, including evaluant, analyte, analyte detector, reagent reporter, comparator, etc. The final protocol application includes a series of data transformation sub-processes to determine if the concentrations of interferon gamma in the lung are significantly different between uninfected and infected mice.

Several features of this approach to the representation of this experiment are worth noting:

- First, this approach emphasizes the role that the experiment processes play in linking entities together. For example, if a different type of assay were used to measure the interferon gamma analytes or a different standard curve were used for the interpolation step, different concentrations of interferon gamma would be obtained. Thus, in order understand the result it is critical to know how the processes that were used to derive it.
- Second, the approach shows how specimen qualities are determined – through combinations of assays and data transformations.
- Third, the relationships between biomaterial objects are captured and can be used to transfer quality information up the biomaterial chain. Thus, the concentration of interferon gamma is both a quality of the lung specimen as well as the mouse source.
- Fourth, the model is focused on capturing the key entities necessary to identify relevant data sets based on the structure meta-data and on common approaches for re-analysis, namely to search for patterns in the experiment results (assay output qualities) based on difference in the input assay variables.
- Fifth, one of the big advantages of taking this approach for database representation is that it naturally interoperates with the ontology terms from the OBO Foundry, which can be used to describe the types of specimens (FMA), chemical therapeutics (ChEBI), analytes (PRO), assay types (OBI), protocol application roles (OBI), etc.
- Finally, in some cases a database resource may not be interested in capturing all of the details for each sub-process in a protocol application in a structured way. For example, the database resource may only what to parse into the table structure selected process inputs (the virus, the mouse strain, the lung specimen) and selected outputs (interferon gamma concentration and the t-test p-value result). In this case, the other entities necessary to describe the derivation of the outputs from the inputs described in the database record must be described in a text document, similar to the experiment description provided in the methods section of the paper.

V. Ontology-based data mining

The ontology-based approach to knowledge representations offers many significant opportunities for new approaches to data mining that go beyond the simple search for patterns in the primary data by integrating information incorporated in the structure of the ontology representation. We term these approaches ‘meta-mining’ because they represent the mining and analysis of integrated knowledge sets derived from multiple,

often disparate, sources. Meta-mining approaches can be used for a wide range of different data mining activities, including indexing and retrieval of data and information, mapping among ontologies, data integration, data exchange, semantic interoperability, data selection and aggregation, decision support, natural language processing applications, and knowledge discovery [Rubin 2008, Bodenreider 2008]. For example, Cook et al. have describe the use of ontological representations to infer causal chains and feedback loops within the network of entities and reactions in biological pathway representation [Cook 2007]. O'Conner et al. have described the use of ontological representation of clinical trials information to support temporal reasoning in electronic clinical trials management systems [O'Conner 2008]. Here we focus on two specific examples. In the first example of meta-mining, the role of ontologies in the description of the experiment meta-data for the identification and use of related data sets will be discussed. In the second example of meta-mining, the use of Gene Ontology-based biological process gene annotation for the interpretation of gene expression microarray results [Lee 2006] and protein interaction network structure [Luo 2007] will be discussed.

Metadata mining

The goal of establishing experiment data archives like ArrayExpress and ImmPort is to allow the re-use of data derived from previous experimentation in the interpretation of new experiment results. Indeed, scientists currently do this in an informal, subjective way in the discussion sections of their journal articles where they interpret their experiment result in the context of the current state of scientific knowledge in the relevant biological discipline. One of the goals of meta-mining is to approach this integrative interpretation in an objective, computational manner. For example, let's imagine that you have recently completed a gene expression microarray experiment in which you have determined gene expression levels in a series of samples from pancreatic specimens of rats immunized with insulin to induce type 1 diabetes through an autoimmune mechanism in the presence and absence of treatment with the immunosuppressive drug cyclosporin. While gene expression microarray have revolutionized the way we do gene expression analysis by allowing the simultaneous assessment of all genes in the organisms genome, it is still hampered by the presence of natural biological and experimental variability, resulting in relatively high false positive and false negative rates. One approach for addressing these inaccuracies is to compare your data with related data sets under the assumption that any discoveries made with independent, related data sets are likely to be real and relevant. So how does one determine which data sets are 'related' in a comprehensive, objective way. This is where ontology-based representation of experiment meta-data can play a valuable role.

A simple approach would be to look for data sets derived from identical experiment designs, but the chances that there are sufficient numbers of these data sets in the public domain tends to be relatively small. And so, we would like to extend what we would consider to be 'related' data sets to include those that are 'similar' in experiment design. In this case we have used microarrays as the assay methodology for quantifying mRNA transcript levels. However, other types of assay methodologies used for assessing transcript levels, like RT-PCR, SAGE and even northern blotting, would provide similar

data that could be useful for meta-mining. Using an ontology like OBI to describe assay types would allow for this kind of definition of similar assay. The pancreas specimen used for microarray assessment is a type of endocrine organ. We might be interested in incorporating data derived from experiment using other types of endocrine organs, e.g. thyroid gland, adrenal gland, ovaries, etc. Using organ terms derived from the Foundational Model of Anatomy for the annotation of the specimen derived from the biomaterial transformation step would allow this kind of inference about "similarity' to be made. The rat is used as an experiment animal model because it has a similar anatomy and physiology as humans, as do other mammalian species. Related species could be identified using a species taxonomy, like the NCBI Taxonomy, as the basis for organism annotation. Type 1 diabetes is an autoimmune disease of endocrine glands, as are Graves' disease, Hashimoto's thyroiditis, Addison's disease; experiment that investigate these types of diseases are also likely to be helpful in the interpretation of your results. Indeed, we might also want to include any autoimmune disease (e.g. lupus, multiple sclerosis, etc.) for this purpose. Use of an ontology like the Disease Ontology would facilitate the identification of experiments based on these kinds of relationships. Finally, the use of an ontology like ChEBI for the identification of other immunosuppressive compounds could further extend the meta-mining analysis.

Thus, the use of ontology-based knowledge representation to define the qualities (disease state, immunosuppressive) of the input entities (rat, cyclosporin) playing specific roles (therapeutic) in the material transformation that gives rise to the specimen output (pancreas), which serves as the input to the assay (gene expression microarrays) that results in the measurement of mRNA transcript levels output allows one to extend the analysis for associations between the dependent and independent variable in a range of related experiments through this ontology-driven meta-mining approach.

Knowledge integration

The second example of meta-mining that takes advantage of the knowledge incorporated into the semantic structure of the ontology comes from the well-established approach of using the Gene Ontology (GO) [Ashburner 2000, Harris 2004, Diehl 2007] annotation of gene products to analyze experiment data sets. The classic example of this comes from the use of the Gene Ontology in the interpretation of gene groups derived from gene expression microarray data clustering. A common goal of microarray data analysis is to group genes together based on similarity in their expression pattern across different experiment conditions, under the assumption that genes whose expression correlates with the experiment condition being investigated are likely to be involved in a relevant underlying biological process. For examples, genes whose expression pattern correlates with the cancer phenotype of the specimens might be expected to be involved in cell proliferation control. The GO Consortium has performed two activities of relevance for this use case. The GO developed by the consortium includes three comprehensive term hierarchies for biological processes, molecular functions and cellular components. The consortium has then curated the scientific literature and annotated gene products with GO terms from this knowledgebase.

Several groups have developed approaches for utilizing GO annotation as a means for identifying relevant biological processes associated with gene expression clusters derived from microarray data by assessing whether specific GO annotations are over-represented in the gene cluster (e.g. [Lee 2006] and http://geneontology.org/GO.tools.shtml). The CLASSIFI algorithm not only assesses the co-clustering of the primary GO annotations for genes in a cluster, but also captures the parent terms from the GO hierarchy for this assessment [Lee 2006].

The analysis of the data sets from B cell stimulated with a panel of ligands illustrates how the semantic structure of the GO hierarchy allowed the discovery of important biological processes that would not have been readily apparent from the use of a flat vocabulary for the gene function annotation [Lee 2006]. B lymphocytes were isolated from mouse spleen and treated with a variety of different ligands to simulate natural environmental stimuli. RNA isolated from these specimens was evaluated by gene expression microarray to measure the expression level of a large cross section of genes in the mouse genome. The ~2500 genes that were differentially expressed were grouped together into 19 gene clusters based on their expression pattern in response to three important B cell stimuli – anti-CD40, lipopolysaccharide (LPS) and anti-Ig. The CLASSIFI algorithm was then used to determine if any categories of genes were overrepresented in any of the gene clusters based on an analysis of their GO annotations. For example, Gene Cluster #18 includes genes that were upregulated in response to anti-Ig but not anti-CD40 or LPS, and contains 191 of the 2490 genes in the entire data set, 7 of the 10 genes annotated with the GO term 'monovalent inorganic cation transport', and 24 of the 122 genes annotated with the GO term 'transporter activity'. In the latter case, the probability that this degree of co-clustering would have occurred by chance is $\sim 9 \times 10^{-6}$. The result of this data mining exercise was the hypothesis that stimulation of B lymphocytes through their antigen receptor using anti-Ig results in the activation of a set of specific transporter processes involving receptor endocytosis and intracellular vesicle trafficking to facilitate antigen processing and presentation. Subsequent experimental studies confirmed this hypothesis.

The important point is that this data mining results would not have been possible without the semantic structure of the GO, which was used to infer relationships between the genes in the gene clusters based on prior knowledge of the interrelationships of biological processes. Figure 5 shows the hierarchical structure of a small portion of the GO biological process branch focused on transporter processes. The genes found in Cluster #18 that gave rise to the cluster classification of 'transporter activity' are listed next to the specific GO term with which they are annotated. At most only three genes were annotated with a given GO term. However, many genes are annotated with terms that are closely related within this small region of the GO biological process hierarchy. By incorporating the GO hierarchy in the analysis, CLASSIFI allowed the discovery of these relationships, which would not have been possible with a flat vocabulary.

VI. Concluding Remarks

In order to take maximum advantage of the primary data and interpretive knowledge derived from the research enterprise is has become increasingly important to agree on standard approaches to represent this information in a consistent and useful format. Many international consortia have been working toward the establishment of standards related to what kind of information should be captured, how the information should be described and how the information can be captured in database resources. The combined effect is that experimental data from biomedical investigation is becoming increasingly accessible to re-use and re-analysis and thus is playing increasingly important roles in the discovery of new knowledge of the workings of biological systems through improved approaches to data mining.

VII. Acknowledgements

This work was supported by the National Institute of Allergy and Infectious Diseases - N01AI40041 and N01AI40076.

Bibliography

Aranguren ME, Antezana E, Kuiper M, Stevens R. “Ontology Design Patterns for bio-ontologies: a case study on the Cell Cycle Ontology” BMC Bioinformatics 2008, 9(Suppl 5):S1.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium” Nat Genet. 2000 May;25(1):25-9.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. “NCBI GEO: mining tens of millions of expression profiles--database and tools update” Nucleic Acids Res. 2007 Jan;35(Database issue):D760-5.

Bodenreider O. “Biomedical ontologies in action: role in knowledge management, data integration and decision support” Yearb Med Inform. 2008; 67–79.

Bornberg-Bauer E, Paton NW. “Conceptual data modeling for bioinformatics” Briefings in Bioinformatics 2002; 3:166-180.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S,

Stewart J, Taylor R, Vilo J, Vingron M. “Minimum information about a microarray experiment (MIAME)-toward standards for microarray data” Nat Genet. 2001 Dec;29(4):365-71.

Conenello GM, Zamarin D, Perrone LA, Tumpey T, Palese P. “A single mutation in the PB1-F2 of H5N1 (HK/97) and 1918 influenza A viruses contributes to increased virulence” PLoS Pathog. 2007 Oct 5;3(10):1414-21.

Cook DL, Wiley JC, Gennari JH. “Chalkboard: ontology-based pathway modeling and qualitative inference of disease mechanisms” Pac Symp Biocomput. 2007:16-27.

Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA. “The Stanford Microarray Database: implementation of new analysis tools and open source release of software” Nucleic Acids Res. 2007 Jan;35(Database issue):D766-70.

Diehl AD, Lee JA, Scheuermann RH, Blake JA. “Ontology development for biological systems: immunology” Bioinformatics. 2007 Apr 1;23(7):913-5.

Fridsma DB, Evans J, Hastak S, Mead CN. “The BRIDG project: a technical report” J Am Med Inform Assoc. 2008 Mar-Apr;15(2):130-7.

Greene JM, Collins F, Lefkowitz EJ, Roos D, Scheuermann RH, Sobral B, Stevens R, White O, Di Francesco V. “National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics” Infect Immun. 2007 Jul;75(7):3212-9.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R; Gene Ontology Consortium. “The Gene Ontology (GO) database and informatics resource” Nucleic Acids Res. 2004 Jan 1;32(Database issue):D258-61.

Hucka M, Bolouri H, Finney A, Sauro H, Doyle J, H K, Arkin A, Bornstein B, Bray D, Cuellar A, Dronov S, Ginkel M, Gor V, Goryanin I, Hedley W, Hodgman T, Hunter P, Juty N, Kasberger J, Kremling A, Kummer U, Le Novère N, Loew L, Lucio D, Mendes P, Mjolsness E, Nakayama Y, Nelson M, Nielsen P, Sakurada T, Schaff J, Shapiro B, Shimizu T, Spence H, Stelling J, Takahashi K, Tomita M, Wagner J,

Wang J. “The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. Bioinformatics” 2003; 19:524-531.

Jones AR, Miller M, Aebersold R, Apweiler R, Ball CA, Brazma A, Degreef J, Hardy N, Hermjakob H, Hubbard SJ, Hussey P, Igra M, Jenkins H, Julian RK Jr, Laursen K, Oliver SG, Paton NW, Sansone SA, Sarkans U, Stoeckert CJ Jr, Taylor CF, Whetzel PL, White JA, Spellman P, Pizarro A. “The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics” Nat Biotechnol. 2007 Oct;25(10):1127-33.

Laibe C, Le Novere N. “MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology” BMC Systems Biology. 2007; 1:58

Lee JA, Sinkovits RS, Mock D, Rab EL, Cai J, Yang P, Saunders B, Hsueh RC, Choi S, Subramaniam S, Scheuermann RH; Alliance for Cellular Signaling. “Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR)
Stimulation” BMC Bioinformatics. 2006 May 2;7:237.

Lee JA, Spidlen J, Atwater S, Boyce K, Cai J, Crosbie N, Dalphin M, Furlong J, Gasparetto M, Goldberg M, Hyun B, Jansen K, Kollmann T, Kong M, Lary D, Leif R, McWeeney S, Moloshok TD, Moore W, Nolan G, Nolan J, Nikolich-Zugich J, Parrish D, Price J, Purcell B, Qian Y, Selvaraj B, Simmerson M, Smith C, Tchuvatkina O, Wilkinson P, Wertheimer A, Wilson C, Scheuermann RH, Brinkman RR. “MIFlowCyt: The Minimum Information about a Flow Cytometry Experiment” Cytometry: Part A. 2008 in press.

Le Novère N, Finney A, Hucka M, Bhalla US, Campagne F, Collado-Vides J, Crampin EJ, Halstead M, Klipp E, Mendes P, Nielsen P, Sauro H, Shapiro BE, Snoep JL, Spence HD, Wanner BL. “Minimum Information Requested in the Annotation of biochemical Models (MIRIAM)” Nature Biotechnology. 2005; 23(12):1509-1515.

Lloyd C, Halstead M, Nielsen P. “CellML: its future, present and past” Progress in Biophysics & Molecular Biology. 2004; 85:433-450.

Luciano J “PAX of mind for pathway researchers” Drug Discovery Today. 2005; 10(13):937-42.

Luo F, Yang Y, Chen CF, Chang R, Zhou J, Scheuermann RH. “Modular organization of protein interaction networks” Bioinformatics. 2007 Jan 15;23(2):207-14.

O'Connor MJ, Shankar RD, Parrish DB, Das AK. “Knowledge-data integration for temporal reasoning in a clinical trial system” Int J Med Inform. 2008 Sep 12.

Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A. "ArrayExpress--a public database of microarray experiments and gene expression profiles" Nucleic Acids Res. 2007 Jan;35(Database issue):D747-50.

Ramakrishnan R, Gehrke J. "Database Management Systems, 3rd Edition" McGraw-Hill Co., New York 2003.

Rubin DL, Shah NH, Noy NF. "Biomedical ontologies: a functional perspective" Brief Bioinform. 2008 Jan;9(1):75-90.

Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. "Relations in biomedical ontologies" Genome Biol. 2005;6(5):R46.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ; OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration" Nat Biotechnol. 2007 Nov;25(11):1251-5.

Squires B, Macken C, Garcia-Sastre A, Godbole S, Noronha J, Hunt V, Chang R, Larsen CN, Klem E, Biersack K, Scheuermann RH. "BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and virulence" Nucleic Acids Res. 2008 Jan;36(Database issue):D497-503.

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, Novère NL, 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.

Figure Legends

Figure 1 - Minimum Information about a Flow Cytometry Experiment (MIFlowCyt). Version 07.09.13 of the MIFlowCyt standard. *Those data elements that are common to most, if not all, MIBBI minimum information standards. †Those data elements that are relatively unique to the flow cytometry methodology.

Figure 2 – OBO Foundry candidate ontologies. Domains of biological reality (highlighted in light blue) are defined based on the intersection between entity types (columns; yellow) as defined by their relationship with time as proposed in the Basic Formal Ontology (BFO) and granularity (rows; pink). Candidate OBO Foundry ontologies for each domain are listed - Foundational Model of Anatomy (FMA), Common Anatomy Reference Ontology (CARO), Cell Ontology (CL), Gene Ontology (GO), Chemical Entities of Biological Interest (ChEBI), Sequence Ontology (SO), RNA Ontology (RnaO), Protein Ontology (PrO), Common Physiology Reference Ontology (CPRO), and Phenotypic Qualities Ontology (PaTO).

Figure 3 - the Ontology-Based eXtensible Data Model (OBX). See text for a more detailed description. (A) The high-level general entity tables of the OBX model including the individual table attributes. (B) Low-level entity tables of specific Object types with subtype-specific attributes.

Figure 4 – OBX-based representation of a virus infection experiment. An OBX-based representation of an experiment from a recent publication [Conenello 2007] in which mice are infected with influenza virus and the amount of interferon gamma in the lung is assessed as a measure of the host immune response to viral infection. Individual sub-processes are defined based on the specific inputs and outputs of the sub-process together with the roles that each component plays. The three major types of sub-processes – biomaterial transformation, assay, and data transformation - are described. The ordered set of sub-processes forms a specific protocol application defined by its objective (e.g. cytokine quantification).

Figure 5 – Gene Ontology Hierarchy in the Mining of Gene Expression Microarray Data. A piece of the GO biological process hierarchy that includes the cellular transport terms is displayed with GO terms listed in the yellow boxes. Genes that have been found in the Gene Cluster #18 [Lee 2006] and are annotated with the specific GO term are listed *italic* to the right of the GO term box. See text for details.