Journal of Biomedical Informatics 44 (2011) 48–58

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Toward an ontology-based framework for clinical research databases

Y. Megan Kong [a], Carl Dahlke [b], Qun Xiang [a], Yu Qian [a], David Karp [d], Richard H. Scheuermann [a,c,*]

[a] Department of Pathology, U.T. Southwestern Medical Center, Dallas, TX, United States
[b] Health Information Systems, Northrop Grumman, Inc., Rockville, MD, United States
[c] Division of Biomedical Informatics, U.T. Southwestern Medical Center, Dallas, TX, United States
[d] Division of Rheumatology, U.T. Southwestern Medical Center, Dallas, TX, United States

ARTICLE INFO

Article history:
Available online 10 May 2010

Keywords:
Ontology
Clinical trials
Biomaterial transformation
Assay
Data transformation
Conceptual model

ABSTRACT

Clinical research includes a wide range of study designs from focused observational studies to complex interventional studies with multiple study arms, treatment and assessment events, and specimen procurement procedures. Participant characteristics from case report forms need to be integrated with molecular characteristics from mechanistic experiments on procured specimens. In order to capture and manage this diverse array of data, we have developed the Ontology-Based eXtensible data model (OBX) to serve as a framework for clinical research data in the Immunology Database and Analysis Portal (ImmPort). By designing OBX around the logical structure of the Basic Formal Ontology (BFO) and the Ontology for Biomedical Investigations (OBI), we have found that a relatively simple conceptual model can represent the relatively complex domain of clinical research. In addition, the common framework provided by BFO makes it straightforward to develop data dictionaries based on reference and application ontologies from the OBO Foundry.

1. Introduction

The US National Institutes of Health are interested in maximizing the return on the public investment in biomedical research. This desire has led many institutes to develop policies that encourage sharing of data generated from research supported by this public funding. In this regard, the National Institute of Allergy and Infectious Disease (NIAID) has supported a number of bioinformatics initiatives designed to provide the infrastructure to capture and manage research data for re-use and re-analysis. The Bioinformatics Integration Support Contract (BISC) was awarded to develop a long-term sustainable archive of data generated by the ~1500 investigators supported by the Division of Allergy, Immunology and Transplantation (DAIT). DAIT investigators conduct a wide range of research program types, including basic research of immune system function, translational research to determine the underlying mechanisms of immune system disease and response to infection, and clinical trials to evaluate the safety, toxicity, efficacy and mechanisms of immune disease therapies and vaccination strategies. More recently, the National Center for Research Resources (NCRR) has developed the Clinical and Translational Science Award (CTSA) program [1], "to improve human health by transforming the research and training environment to enhance the efficiency and quality of clinical and translational research". At many CTSA institutions, comprehensive clinical research information systems are being developed for the electronic capture and use of clinical and translational research data at an enterprise level (e.g., [2,3]).

Through the BISC project, we have developed the Immunology Database and Analysis Portal (ImmPort; www.immport.org) as a web-based public resource to support not only the archiving of these valuable data sets, but also to support their integration with the biological knowledge contained in other public data repositories (e.g., GenBank, UniProt, the Immune Epitope Database) and their analysis using state-of-the-art data mining analytical tools. One of the biggest challenges in ImmPort design is how best to manage the data derived from the wide range of different experiment methodologies being used by DAIT-funded investigators, which includes everything from gene expression and SNP genotyping microarrays up through clinical trials, as well as methodologies that are somewhat unique to the immunology research domain (e.g., flow cytometry and ELISPOT). And so we have adopted a general strategy for database development in which our database structure is designed around the general features of any biomedical investigation, rather than based on experimental features that might be methodology specific.

* Corresponding author. Address: Division of Biomedical Informatics, Division of Translational Pathology, John H. Childers Professorship in Pathology, Department of Pathology, U.T. Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390-9072, United States. Fax: +1 214 648 4070.
E-mail address: richard.scheuermann@utsouthwestern.edu (R.H. Scheuermann).
URL: http://pathcuric1.swmed.edu/Research/scheuermann.html (R.H. Scheuermann).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.05.001

In addition to this practical design philosophy, ImmPort strives to ensure that the data and analytical infrastructure is maximally interoperable with other external databases and bioinformatics resources. Thus ImmPort has been an active participant and early adopter of many data standards development initiatives, including the development of minimum data standards like MIFlowCyt [4] and MIGen through the MIBBI consortium [5] and ontology standards like the Ontology for Biomedical Investigations (OBI) through the Open Biomedical Ontology (OBO) Foundry consortium [6].

In the development of a database support infrastructure for clinical research data in ImmPort, several related clinical data standards were evaluated for potential use. The Clinical Data Interchange Standards Consortium (CDISC, www.cdisc.org) is a global multi-disciplinary organization focused on the development of a set of clinical data standards to facilitate global clinical trial data interoperability and exchange [7,8]. CDISC leads several ongoing projects to develop standards for representation of the study designs and data elements of biomedical investigations, particularly clinical trials. These standards include the Study Data Tabulation Model (SDTM) for submitting clinical trial data to the FDA, the Biomedical Research Integrated Domain Group (BRIDG) model for clinical trials [9,10], and the Clinical Data Acquisition Standards Harmonization (CDASH) project for defining data elements that can be assembled into case report forms (CRFs). Because the CDISC standards have been developed to support information about ongoing clinical studies, they are not completely aligned with the needs of the ImmPort system, which is largely concerned with the results of studies and making study data available for data re-use and re-analysis. Therefore, ImmPort does not need to represent the administrative and regulatory aspects of a study that are important components of the CDISC standards. The SDTM standard, although it covers many of the data elements required by ImmPort, is a data transport standard based on SAS transport files and is therefore not a model for a data repository designed for data retrieval and integration with other reference data. CDASH provides many elements for modeling clinical encounters, but is incomplete for ImmPort's purposes because it does not model a study's schedule of events and does not support placement of clinical encounters on a study time line. The BRIDG model provides extensive support for monitoring the execution of the study protocol [9,10], which is also not needed for ImmPort. Finally, none of the CDISC models currently offer significant support for analysis of study data and the results derived from this analysis, and none offer significant support for the mechanistic studies being performed with specimens derived from study subject participants.

Through this work, we have considered how minimum data standards and ontology structures might be utilized to help inform the design of databases. It is important to be clear about the distinction between ontologies and conceptual models. Well-formulated ontologies are designed to describe universal classes of entities in reality and how these classes invariably relate to each other. The structure of ontologies should not be context dependent. In contrast, conceptual models which describe the entities and the relationships among the entities are focused on supporting instance-level data in which specific representatives of entity classes are described together with the characteristics that distinguish individuals from each other within the class. Thus, conceptual models need to be able to capture and integrate instance-level characteristics and context dependencies.

In the study reported here, we have attempted to investigate whether it would be possible to integrate these two components of knowledge representation in such a way as to leverage the class-level structural characteristics provided by a set of well-formulated reference ontologies as an underlying common framework that could then be extended in a consistent fashion to incorporate the instance-specific details. We have specifically applied this strategy to the representation of clinical research data, including the study design components found in clinical protocols, clinical assessment results captured in case report forms and laboratory results obtained from the evaluation of derived human specimens. The end result is the Ontology-Based eXtensible (OBX) data model.

2. Methods and results

In this section, the general framework of OBX model, the components of the OBX model (biomaterial transformation, assay, data transformation, composite process and study design.) and the physical database implementation of the OBX model are presented.

2.1. General framework

Two reference ontologies were chosen as the foundation for OBX design – the Basic Formal Ontology (BFO) and the Ontology for Biomedical Investigation (OBI). The BFO (http://www.ifomis.org/bfo) was originally conceived of by Smith and Grenon as an upper level ontology that could serve as a framework to support the development of domain-specific ontologies for scientific research [11]. The BFO structure is based on the central dichotomy between objects (continuants) and processes (occurrents), reflecting their distinct relationships with time. Continuants endure through time and retain some notion of their identity even while undergoing various kinds of changes. Occurrents unfold in time and can be defined to include temporal start and end points. Continuants can be further sub-divided into those physical objects that exist independent from other entities – independent continuants (e.g., organs, tissues, cells, molecules, etc.), and entities that depend on physical objects for their existence – dependent continuants (e.g., the color red, the investigator role, the ribonuclease molecular function). The OBI (http://purl.obofoundry.org/obo/obi/) builds upon BFO, extending the core structure by describing those entities that are specific to the biomedical research domain. For example, occurrent is extended to include subtypes of various process like biomaterial transformation, assay and data transformation; independent continuant is extended to include biomaterial and instrument; dependent continuant is extended to include investigator role, analyte role and evaluant role. Both BFO and OBI have been built using a strict *is_a* hierarchy of type/subtype relations and are compliance with the principles for ontology development best practices promulgated by the OBO Foundry (http://www.obofoundry.org/crit.shtml).

In order to determine if the structure of these ontologies could be used to build a database that could support the management of a wide range of data derived from clinical and translational research studies, we extracted the core structure of the OBI extension of the BFO and developed a conceptualization of the core components as a starting point for data modeling (Fig. 1A). The central component of the core conceptual model is the *Event* table, which includes descriptions of the actual events that happened during the study. These actual events may or may not be planned. A planned event is a realization of *Procedure Specification*; this separation allows for situations in which the actual event deviates from what was planned, including protocol deviations and adverse events of critical importance to the clinical research domain. Each event may also include one or more objects that play defined input and output roles. Each event also occurs in a specified time context. And finally, each event occurs in the context of a specific study that describes the actual realization of a study design.

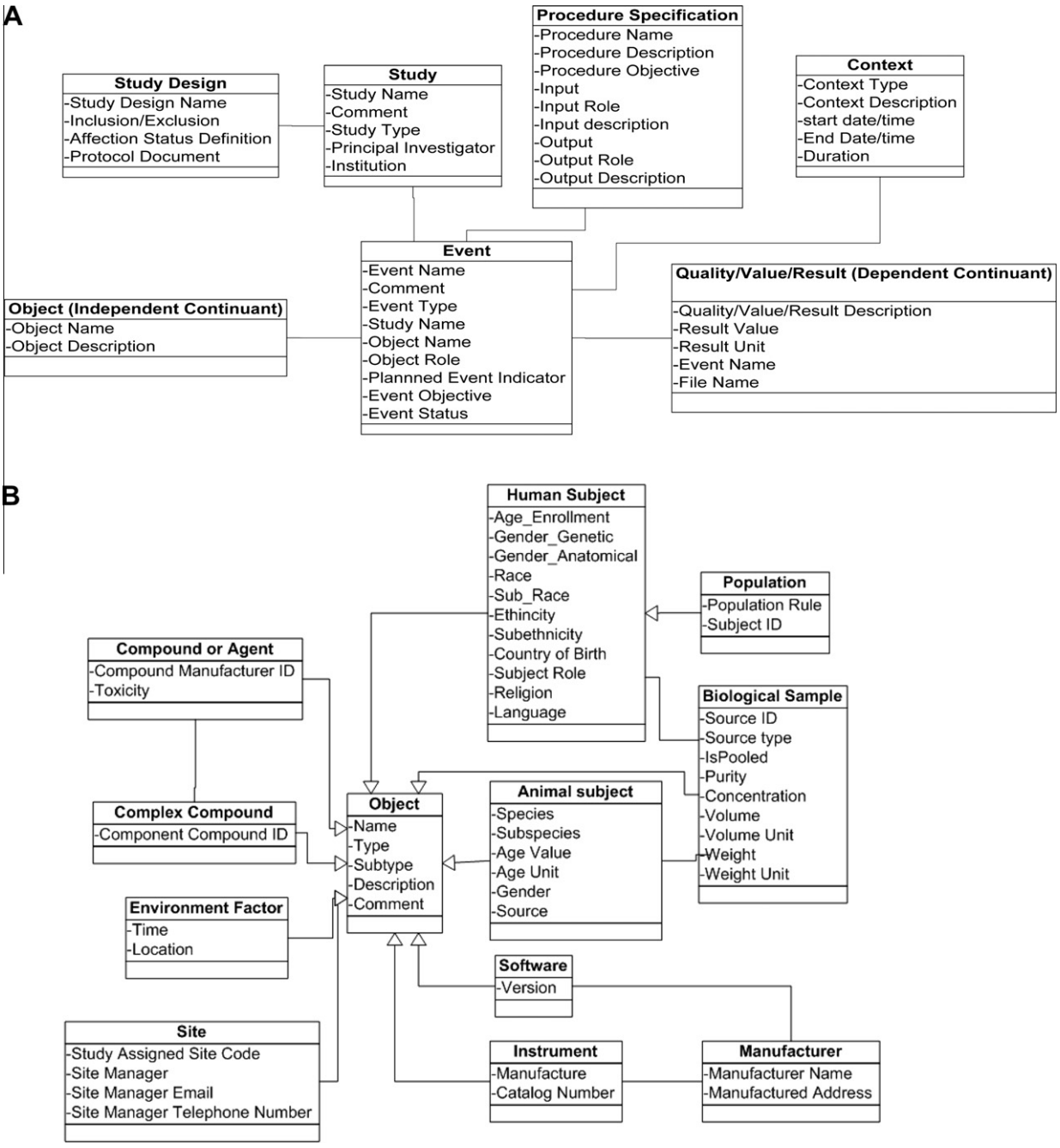

Fig. 1. OBX conceptual model representation. (1A) The schematic representation of the OBX core conceptual model showing high level concepts of Study Design, Study, Independent Continuant (Object), Event, Dependent Continuant (Quality, Value, or Result) and Context (for each event). (1B) UML model for the Independent Continuant Object. Human subject, Population (grouping of human subject), Animal Subject, Biological Sample (from human or animal subject), Compound, Site, Instrument, and Software are all specific objects modeled in OBX. A complete UML representation of the resulting OBX conceptual model with cardinality restrictions can be found at http://pathcuric1.swmed.edu/Research/scheuermann/OBX.html.

The OBX core conceptual model was then used as a framework for representing information about particular entities that need to be described in the clinical research database component of ImmPort and how they relate to each other in subtype tables. (N.B. For the sake of clarity, we use the term "universal entity" to describe the classes of things in reality, "particular" to describe the specific member of that class, "data entity" to describe the information artifact in the conceptual model about the universal entity, and "data instance" to describe the information artifact value about the particular member.) Again, we relied on OBI/BFO ontology design principles to capture the specific distinctions between the different entities.

The following class types and subtypes have been modeled in this way:

- Object – population, population arm, human subject, animal subject, biological sample, compound, complex compound, software, instrument, site (Fig. 1B).
- Biomaterial transformation – substance merging, device intervention, surgery intervention, biosampling process, environment exposure process (Fig. 2).
- Assay – subject assessment, lab test, questionnaire, medical history taking, ECG (Fig. 3).

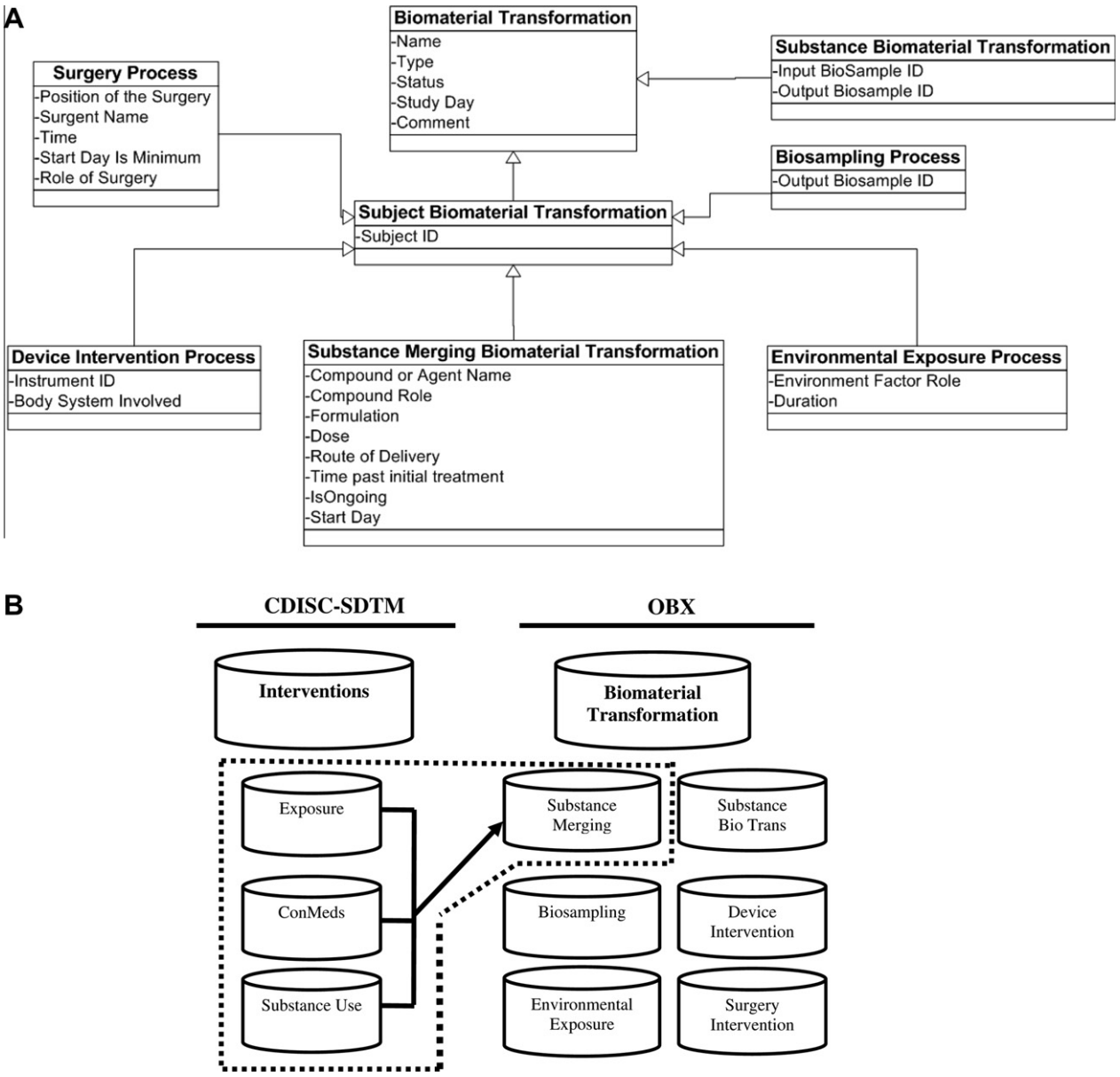

Fig. 2. Biomaterial transformation in OBX Model. UML model of the OBX biomaterial transformation class (2A) and a comparison of the Intervention domain of the CDISC-SDTM with the biomaterial transformation component of the OBX model (2B). The arrow illustrates that the three Intervention classes – Concomitant Medications, Exposure and Substance Use – in CDISC_STDM map to one Substance Merge class in OBX.

- Data transformation – diagnosis process, research data analysis, outcome measure determination process, baseline characteristic designation process, protocol deviation determination (Fig. 4).

In each case, the subtype tables contain attributes that are specific to the given subtype. In some cases we have made practical decisions to directly include data entities as attributes within tables in order to optimize database performance even though they could be indirectly linked through table joining procedures. For example, the *Human Subject* class contains attributes about phenotypic qualities of this particular entity, such as age and gender; it also contains the information about country of birth, which is another independent continuant. While these data entities could be represented in separate dependent continuant and independent continuant tables, we chose to include them as attributes in the *Human Subject* table because they are so commonly linked with subject participants in clinical and translational research. The drawback to this approach is that the specific relations between the subject class of the table and the associated attributes cannot be easily represented in traditional object-oriented database tables, and must be inferred during downstream analytical processing.

2.2. Biomaterial transformation

The process of biomaterial transformation, which is defined as events with one or more biomaterials as inputs and outputs, is differentiated into merging, biosampling and transformation subtypes in OBX (Fig. 2A). One example of an important merging type of event is substance intervention (*Substance Merging Biomaterial Transformation*), whose description includes details about the type of compound included, and the formulation, dose and route of delivery used. In the case of *Biosampling Process* (e.g., blood draw or

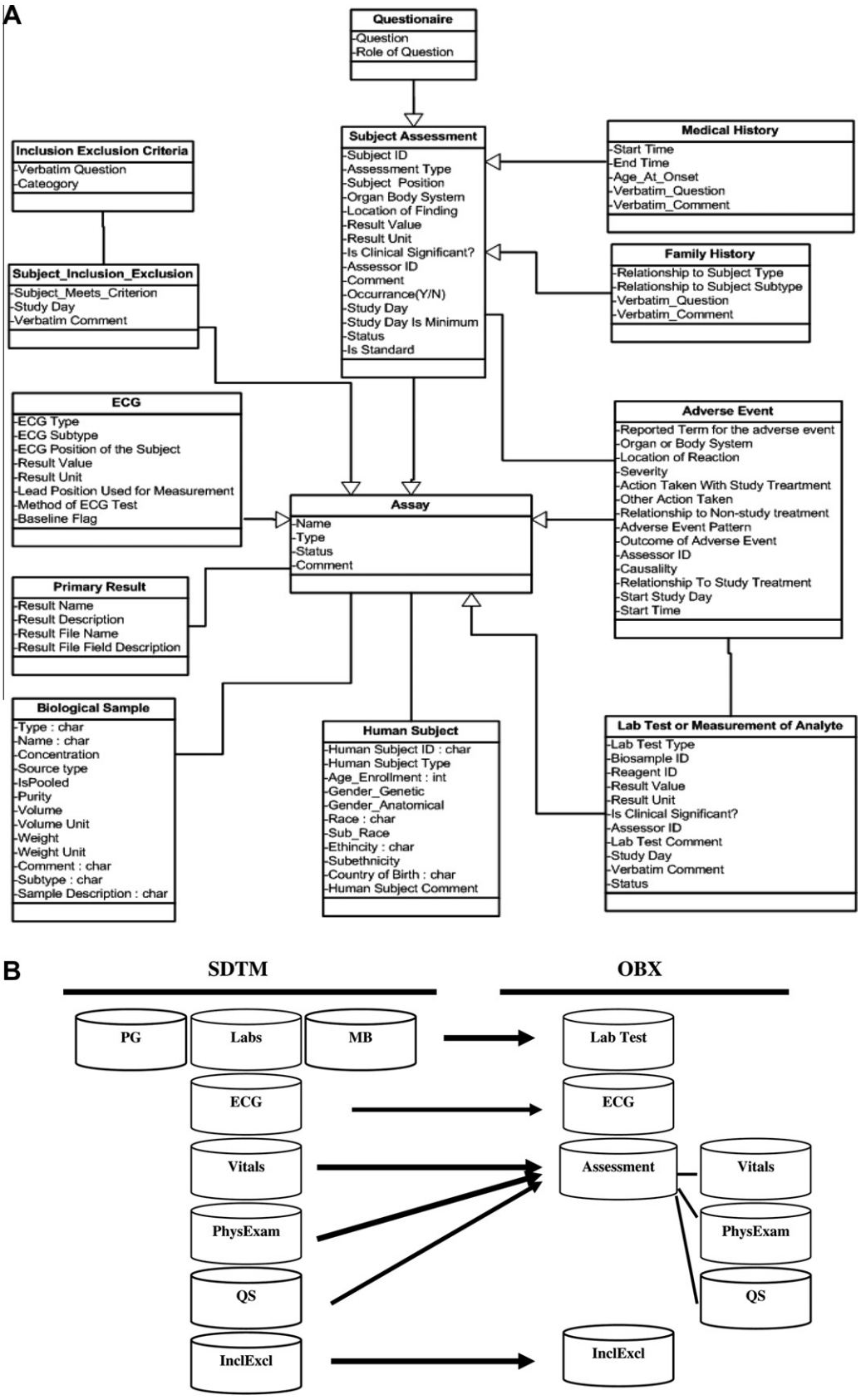

Fig. 3. Assay in OBX Model. UML model of the OBX Assay class (2A) and a comparison of the Findings domain of the CDISC-SDTM with the Assay component of the OBX model (2B). PG, pharmacokinetics; MB, microbiology; QS, questionnaire; and InclExcl, inclusion/exclusion.

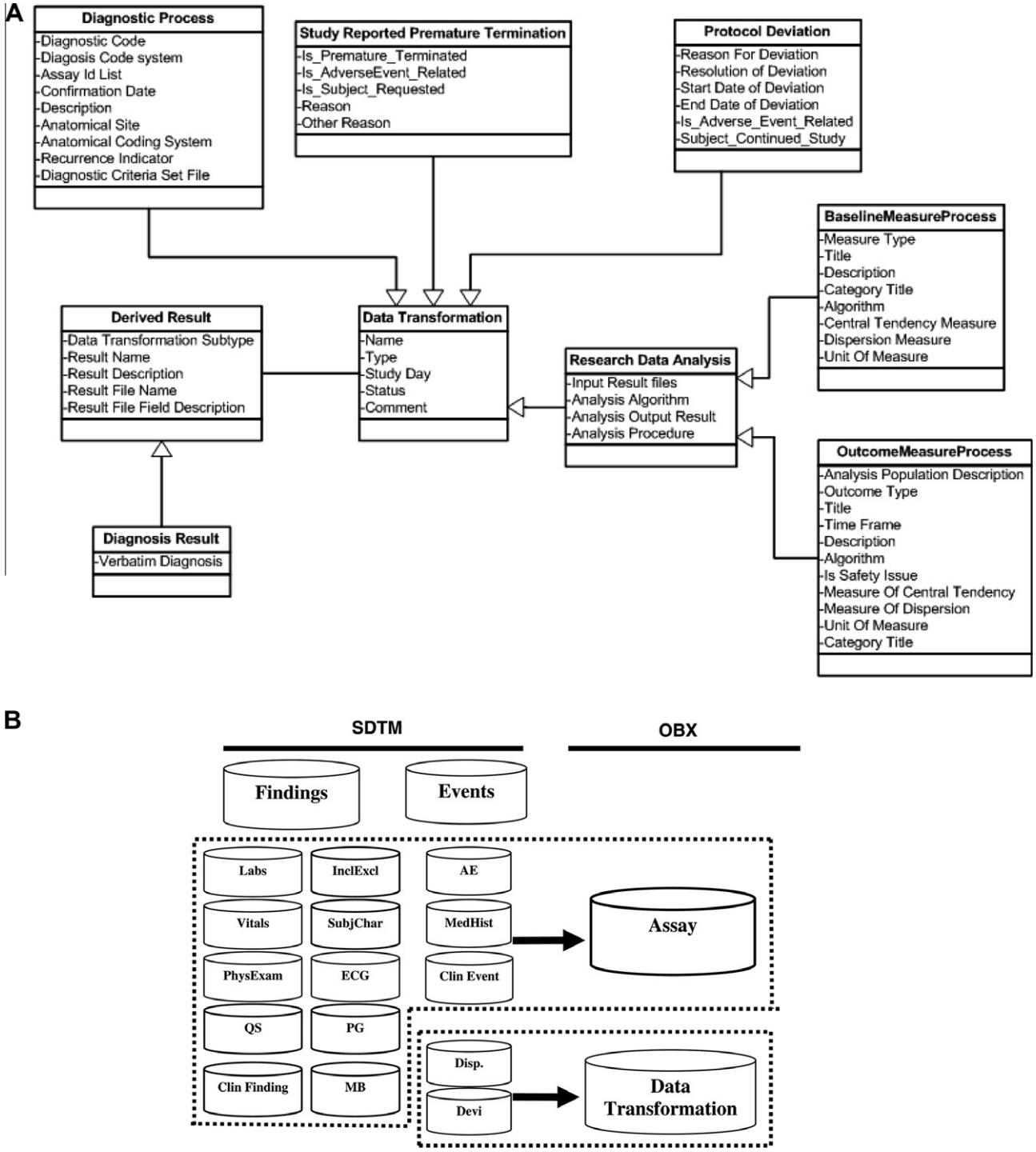

Fig. 4. Data transformation in OBX Model. UML model of the OBX data transformation class (4A) and a comparison of the Findings and the Events domains of the CDISC-SDTM with the assay and the data transformation component of the OBX model (2B). PG, pharmacokinetics; MB, microbiology; QS, questionnaire; InclExcl, inclusion/exclusion; AE, adverse event; Disp, disposition; and Devi, deviation.

saliva collection), the input subject and the output biosample are specified in the conceptual model. The *Substance Biomaterial Transformation* (e.g., incubation of a specimen) describes the transformation from one biomaterial to another biomaterial. In this way, a wide variety of different events can be defined by describing the event type, the input and output continuants and the roles that they play in the process.

We compared the OBX approach to the data representation approaches in CDISC. The Study Data Tabulation Model (SDTM) is one of the data standards developed by CDISC, which has been adopted by the US Food and Drug Administration (FDA) to be the standard format for clinical trial data to be submitted to the FDA. In SDTM, observations collected during the study are divided into three classes: Interventions, Events, and Findings. Interventions class captures investigational treatments and is further divided into three domains (Fig. 2B): Concomitant Medications (ConMeds), Exposure, and Substance Use (Subst Use).

The OBX model places CDISC-SDTM Interventions class under *Biomaterial Transformation* given that both the input and output in the Interventions class are biomaterials. The inputs for Concomitant

Medications, Exposure or Substance Use are subjects of the study and the substances. The substances taken by the subject are called concomitant medications, investigational drugs or self-administered substances for Concomitant Medications, Exposure or Substance Use, respectively. The outputs of these interventional processes are also subjects of the study. However, instead of using three different domains to represent essentially the same process as in CDISC-SDTM, OBX recognizes that the difference between Concomitant Medications, Exposure, and Substance Use is the role that the substance plays in this *Substance Merging Biomaterial Transformation* (see Fig. 2A). By adding a *Compound Role* attribute, the *Substance Merging Biomaterial Transformation* class encompasses the information that is captured in all three SDTM interventional domains.

2.3. Assay

The assay process is defined in OBI as events with one or more biomaterials as inputs and data as outputs (http://obi-ontology.org/page/Main_Page). In OBX, the *Assay* class is differentiated into *Subject Assessment, Lab Test, ECG, Adverse Event,* and *Subject Inclusion/Exclusion* classes (Fig. 3).

Subject Assessment and *Lab Test* differ in that lab tests involves a specimen (biosample) as input and frequently utilize reagents for measurement purposes (substance or compound chemicals) whereas assessments involve a subject participant as input and do not typically utilize reagents. Physical Exam, Medical History, Family History and Questionnaire are examples of different types of assessments. By using the "assessment type" or the "lab test type" attribute, assessment or lab tests that are specific for certain studies (e.g., atopic dermatitis skin assessment) can be captured by OBX. Hematology, Urine Test and Pregnancy Test are examples of types of *Lab Test*. *ECG* is captured as a special type of assessment since it involves attributes such as lead position and baseline flag that are quite different from other types of assessment.

The OBX conceptual model puts the CDISC-SDTM Findings domain under *Assay* given that the input of the Findings domain is biomaterial and the output is data. The SDTM pharmacokinetics, lab test and microbiology domains correspond to the *Lab Test* in OBX whereas the SDTM Vital Signs, Physical Exam and Questionnaire map to *Assessment* in OBX. The SDTM inclusion/exclusion class maps to *Subject Inclusion/Exclusion* class in OBX. In addition, part of the SDTM Event domain, including adverse events, medical history and clinical event, maps to the *Assay* domain in OBX (see Fig. 3B). The rest of the Event domain in CDISC-SDTM including study deviation and disposition map to the *Data Transformation* domain in OBX (see Fig. 4B).

2.4. Data transformation

The process of *data transformation* is an event in which both the input and the output are data (Fig. 4). The *Research Data Analysis* (e.g., microarray data analysis) process is a good example of data transformation. Starting from *Primary Result* (e.g., .cel file which contains the fluorescent intensity of all the spots on the slide) collected in the study, the study investigator may get a set of *Derived Result* (e.g., a list of differentially expressed genes) after certain steps of data processing, or draw conclusions from the study. *Baseline Measure Process*, defined as a data transformation process the output of which is a table of demographic and baseline data for the entire trial population and for each arm or comparison group (www.clinicaltrials.gov, [12]), captures basic statistical measures of the study population such as average age and gender proportions using the demographic data as input. *Outcome Measure Process*, defined as a data transformation process whose output is a table of values for each of the outcome measures by arm (www.clinicaltrials.gov, [12]), specifically captures study result for each study arm using data collected about individual enrolled in each arm as input. The *Diagnosis Process* uses available data collected as input for the decision making process. The *Protocol Deviation* (part of the Event domain in CDISC-SDTM) and the *Study Reported Premature Termination* classes start by looking at available recorded data and give the output as protocol deviation or early termination, respectively.

2.5. Composite process

Even though OBX differentiates clinical processes as biomaterial transformation, assay and data transformation, certain clinical events are combinations of these three basic process types (Fig. 5). For example, a clinical encounter (e.g., visit) may be composed of a blood draw (a biomaterial transformation), a lab test (assay) and a diagnosis (data transformation), or a data analysis pipeline may involve several steps of data transformation. We call these "composite process" or "complex event". OBX currently includes two approaches for grouping events. The first is the *Actual Visit*, which groups events occurring for a single subject in a single clinical encounter. An actual visit may reference a planned visit and will reference actual events. If the actual visit is associated with a planned visit then the actual events will often correspond to the planned events of the planned visit. The second approach is the *Panel of Events*. A panel of events specifies a set of associated events and allows multiple instances of that event set to be associated with actual events. An example of an event panel is rush immunotherapy with subject assessments in which each event set consists of a serious of immunotherapy exposures followed by two subject assessments checking for adverse events. The entire panel consists of a *Sequence of Events* set. There would be as many event sets in a subject's rush immunotherapy event panel as there were immunotherapy exposures.

2.6. Study design

The set of classes comprising a study design allows representation of many elements of a study's protocol in OBX (Fig. 6). The *Study Design* class provides attributes describing the study including the study title, study type (observational or interventional), a summary of the study, principal investigator, condition of focus, planned study outcomes, etc. Two subclasses of the *Study Design* class, *Observational Study Design* and *Interventional Study Design*, augment the *Study Design* class with descriptive attributes that are specific to observational or interventional studies. Many of the attributes in the study design classes correspond to the study attributes described in the document Protocol Data Element Definitions (DRAFT) available at ClinicalTrials.gov [13]. For study *Arm* and *Period*, we have incorporated the concepts from the CDISC-SDTM and the BRIDG model. An arm is a grouping of study subjects that either share the same characteristics (e.g., case or control for observational study) or get the same treatments (e.g., interventional or placebo in interventional study). A period marks a section of the study time line having a specific purpose in the course of the study, e.g., screening, treatment, or follow-up. A study time line has at least one period. If a study has only one period as is the case with many observational studies, then the name of the period is "entire study". The *Visit* class is used to construct a study's schedule of events. It links together *Arm*, *Period* and *Planned Event* classes and specifies the time window within which a set of planned events occurs. The *Event Plan* class can be further categorized into the *Biomaterial Transformation Plan* (e.g., blood draw plan), *Assay Plan* (e.g., lab test plan or assessment plan) and the *Data Transformation Plan* (e.g., statistical analysis plan). In Clinical studies, an ac-

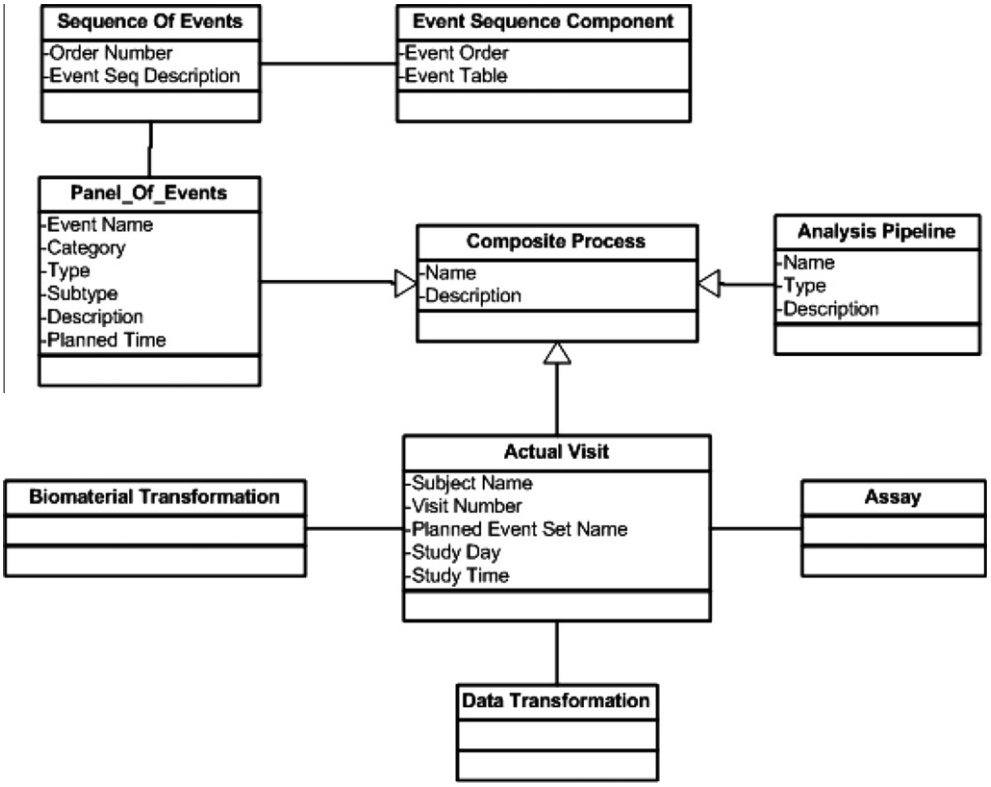

Fig. 5. Composite Process in OBX Model. OBX captures the composite process (sequence of event) through a panel of event (Panel of Events) and the order of the events (Sequence of Events).

tual biomaterial transformation, assay or data transformation recorded during the course of a study may or may not correspond to an *Event Plan*.

2.7. Database schema

The physical implementation of the conceptual model is the database schema. The database implementation of OBX allows the widely varied contents of the diverse clinical data to be loaded and retrieved for display. The database schema was ontologically derived, but the process of filling the database was not ontology driven per se; that is, the loading process was not looking for conceptual matches between the contents of a study and the contents of an ontology. Rather, it was required to allow easy mapping of syntactic structure of the study data sets into the syntactic structure of the database schema.

A major characteristic of the growing collection of clinical studies is that the number of defined data entities in the study data sets tends to grow without bounds as studies are added to the collection. Every study is likely to describe its own set of assessments, for example. Thus a challenge facing the database was to find a mechanism for managing the proliferation of data entities in the study data sets without proliferating the tables in the database schema by using a one to one match to the different data entities in the study data sets.

The data entity management is achieved by partitioning the OBX conceptual model into common and study specific parts:

- Data entities that represent common components of all studies such as visit plan, subject demographics, or adverse events are modeled as data entities with a well-defined finite set of attributes (row model). This modeling takes advantage of the emergence of reporting standards (e.g., CDISC) for common events such as adverse events or ECGs.
- Data entities such as subject assessments and lab test panels that are study specific, or whose numbers of attributes vary from study to study (such as Vital Signs where the same assessment can be called Physical Exam in one study and Vital Sign in another study) are modeled using the data Entity-Attribute-Value (EAV) approach.

An EAV model may be thought of as a 90° rotation of the row model used by classic relational database tables. In a row model, each row corresponds to data entities about a particular of a universal entity class, each column corresponds to the representation of an attribute of the entity and the cells formed by the intersection of columns and rows contain entity attribute values corresponding to data instances about those particular entities. In an EAV model, one row represents a cell from the row model in which the value is tagged with entity class and attribute class labels. In any data instance of a row model, the number of attributes and the identity of the attributes of an entity are fixed. The EAV model allows an indefinite number of data instances of attributes of any type to be associated with a data entity and also allows many different data entities to be represented in the same form. Note that an EAV model is not equivalent to an RDF triples data store. Entity, attribute, and value records tend to contain information about a particular quality of a particular entity, which is not equivalent to subject, predicate, and object components of an RDF triple. The EAV model is more constrained in its expressivity. As mentioned above, a primary goal for developing the database schema for the ImmPort clinical data repository has been to control the multiplication of modeled data entities needed to represent the different universal classes of entities; therefore, we used both the row model and the EAV model to balance the flexibility needed and the performance desired.

A pure EAV triple store containing thousand of assertion types of data entities would be just as unfathomable as a database schema

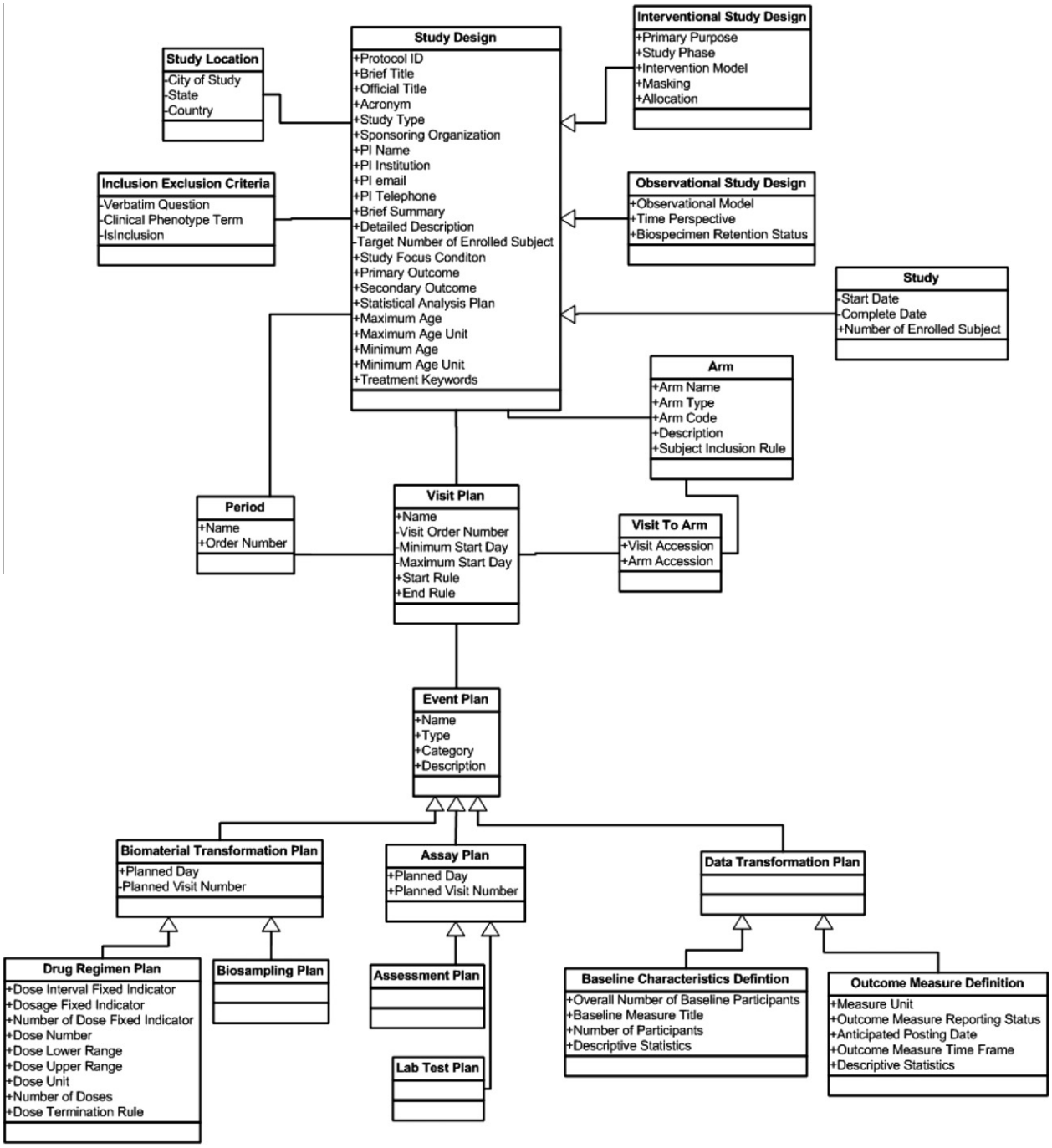

Fig. 6. Study Design in OBX Model. OBX captures the study type, inclusion exclusion criteria, study descriptions, location of the study, study Arm, study period, visit and the planned events for each visit.

that contained hundreds of row models of data entities. The ImmPort approach to making an EAV model usable involves the following modifications to a pure triple store:

- The EAV model is partitioned so that assays (lab tests and subject assessments), biomaterial transformations (such as drug interventions, biological material sampling, and environmental exposure), and data transformations (including data derived from lab tests and assessments) each have their own set of tables. In any area of the EAV model you know the type of data entities with which you are dealing.
- The scope of an EAV row is expanded beyond a single attribute value pair by including attributes that can be optionally associated with the main attribute of the row. So an EAV row for an assessment component will contain not only the component name (the attribute identifier) and result value, but also associated units of measure, results categories and so forth. The set of additional associated attributes provided for a single EAV row depends on the partition of the EAV model in which the row exists.
- The rows of the EAV model that constitute a data entity's attributes are identified by grouping tables. For example, the *Assessment Panel* table groups the individual attributes and values that make up an assessment, and the *Lab Test Panel* groups all

the lab tests that are normally composed into panels. These grouping tables provide a single row identifying a data entity in the EAV model and also are used to associate a data entity with the study time line so that time line references do not have to be pushed down to the individual attribute level.
- The model provides support for describing complex events such as a rush immunotherapy regimen, and for creating identified groupings of the lower level intervention and assessment events which comprise the complex event.

In practice the team working with loading the ImmPort repository has found that mapping many different types of data entities into the multi-level EAV database schema is a straightforward process. So the unbounded extensibility of the EAV model is working in practice to control the proliferation of entities in the ImmPort repository while retaining easy access to entity identities provided by the single row entity identifiers in the grouping tables.

3. Discussion

The Ontology-Based eXtensible (OBX) conceptual model was developed to support the implementation of the clinical research database component of the ImmPort system. ImmPort has used available standards such as CDISC-SDTM and BRIDG to guide development of the clinical study aspects of the OBX conceptual model. Certain specific parts of these standards, such as the SDTM standard for modeling a study time line using study days, were adopted as standards for the ImmPort repository. In addition, these standards have provided critical guidance in defining the scope of data elements to incorporate into the OBX model of biomedical investigations and in defining the attributes of these data elements. For example, the SDTM Adverse Event domain provided most of the descriptive attributes of adverse events used for the OBX adverse event domain.

The OBX model has now been implemented as the database for capturing clinical research data in ImmPort. We have successfully mapped components of a variety of clinical studies from the Atopic Dermatitis Vaccinia Network, Immune Modeling Centers and the Immune Tolerance Network into this model representation. The complexity of these study ranges from observational studies, to phase II multi-arm interventional studies. Based on this exercise, we have refined the conceptual model to ensure that we can not only describe information about the basic entity classes in a clinical study, e.g., human subjects, biosamples, assays, assessments and assessment results, but also the more complex components of a clinical study, e.g., protocol deviations, adverse events, study arm specifications and composite events like the clinical visit. The fact that a variety of studies from different sites have been successfully loaded into a database based on OBX supports the idea that the approach achieved the level of extensibility and interoperability sought for the ImmPort system. We have also successfully implemented clinical data user interfaces to allow users to view, query and download data from ImmPort (www.immport.org).

3.1. Characteristics of the OBX model

During the refinement process, several advantages of the OBX approach have been noted. The relatively simple structure of OBX has made it relatively easy to add new class tables to the database schema without disrupting the pre-existing structure. The logical framework used provides a consistent mechanism for linking component data entities together. It is relatively easy to re-use data entity tables as needed in generating primary key-foreign key relationships. The fact the OBX is based on the logical framework of BFO/OBI allows for its obvious integration with ontology term use as values for specific data elements in the database record instances in the future.

OBX also allows for the integration of clinical data (CRF data) and mechanistic experiment data. An experiment is a type of *Assay* (input is biomaterial and output is data). Specifically, most laboratory experiment would correspond to a *Lab Test* because of the use of reagents in the study of specimens. In this way, OBX can put the experiment inside the clinical study on the study time line, which then greatly facilitates the clinical data analysis.

3.2. OBX model and BRIDG model

The purpose of OBX is to archive completed clinical studies and to facilitate data integration and data re-use. Therefore the patient management information found in BRIDG (http://bridgmodel.org/) designed to monitor ongoing clinical trials, such as HealthCareProvider, HealthCareFacility, LegalSponsor, Resource Provider, OversightAuthority and RegulatoryAuthority, is considered to be outside the scope of OBX. In the process of OBX development, we made the clear distinction between describing a process and the result of the process. We believe capturing the process is important for recapitulating the process and for understanding the results. BRIDG, however, tends to focus on the research result but not the process that gives rise to the result. For example, in BRIDG, the class PerformedClinicalResult is derived from of PerformedObservationResult class. However, the actual clinical process (maybe a lab test), which gives rise to the PerformedClinicalResult or the PerformedObservationResult is not captured. Similarly, the BRIDG PerformedDiagnosis class only captures the final diagnosis, but the process through which the investigator makes that specific diagnosis is not recorded, i.e., the input parameters for this data transformation process are not captured.

3.3. Implementation of the OBX model

We have recently completed a physical database schema based on this model, which is made freely available at www.immport.org. We are now in the process of preparing a complete documentation of the database schema that includes a description of the design principles detailed here, definitions for each of the data element classes, suggested sources of value sets for use in populating specific database records based on standard vocabularies or ontologies, and an implementation guide.

The OBX database schema is now being used to support the capture, managements, query of a wide range of different clinical research studies in the ImmPort system for the National Institute of Allergy and Infectious Disease. Databases built upon the OBX framework could be implemented more extensively at academic health centers focused on clinical and translational research, including those institutions that are part of the NIH-funded Clinical and Translational Science Award (CTSA) program. Indeed, the North and Central Texas Clinical and Translational Sciences Institute is in the process of evaluating the OBX model to store data derived from its clinical and translational research programs. Having said this, many of the CTSA and other institutions conducting these kinds of research programs are in the process of developing their own database infrastructures based on other common conceptual models (e.g., BRIDG) or their own home-brewed approaches. Ideally, one would want to compare the different models to determine whether any provide significant advantages over the others, perhaps by comparing the results of usage metrics for data represented in each [14]. However, it would take a great deal of effort to perform such a comprehensive comparative analysis. In addition, the fact that many extant systems have already been built upon other models and that regulatory agencies (e.g., the FDA) have already proposed the adoption of specific models add to the

complexity of the outcome of any such comparison. Thus, the most practical stance is that OBX can be considered as an alternative model for capturing managing and using clinical and translational research data. Groups considering the implementation of database systems will need to evaluate the alternative strategies based on the theoretical underpinnings of the different approaches, their specific needs, and their existing infrastructure. But one of the advantages of adoption of OBX by other organization interested in managing clinical research data is that it would support data sharing, system interoperability and semantic query of data content from the Immunology Database and Analysis Portal.

3.4. Future directions

In order to extend this initial modeling work into a useful artifact in support of clinical research data management and interoperability data sharing, we are in the process of combining the UML model with a data dictionary and an implementation guide, such that sufficient information is provided for its use in database development. We are also integrating a vocabulary service component, through the NCBO BioPortal (http://www.bioontology.org/) or the EBI Ontology Lookup Service (www.ebi.ac.uk/ontology-lookup/), based on relevant OBO Foundry ontologies to provide the preferred value sets for the data elements described in the OBX model and to support the power of reasoning provided by the use of ontologies as the source for the structured vocabularies.

Acknowledgments

We would like to thank the OBI consortium for helpful discussion about biomedical investigations that forms the basis for the described work. Supported by NIH N01AI40076 and U54RR023468.

References

[1] Califf RM, Berglund L. Principal investigators of national institutes of health clinical and translational science awards. Linking scientific discovery and better health for the nation: the first three years of the NIH's clinical and translational science awards. Acad Med 2010;85(3):457–62 [PubMed PMID: 20182118].

[2] Kush R, Alschuler L, Ruggeri R, Cassells S, Gupta N, Bain L, et al. Implementing single source: the starbrite proof-of-concept study. J Am Med Inform Assoc 2007;14(5):662–73 [Epub 2007 June 28. PubMed PMID: 17600107; PubMed Central PMCID: PMC1975790].

[3] Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at Mayo clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc 2010;17(2):131–5 [PubMed PMID: 20190054].

[4] Lee JA, Spidlen J, Boyce K, Cai J, Crosbie N, Dalphin M, et al. International society for advancement of cytometry data standards task force. MIFlowCyt: the minimum information about a flow cytometry experiment. Cytometry A 2008;73(10):926–30. PubMed PMID: 18752282; PubMed Central PMCID: PMC2773297.

[5] Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 2008;26(8):889–96 [PubMed PMID: 18688244; PubMed Central PMCID: PMC2771753].

[6] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251–5 [PubMed PMID: 17989687; PubMed Central PMCID: PMC2814061].

[7] Hammond WE, Jaffe C, Kush RD. Healthcare standards development. The value of nurturing collaboration. J AHIMA 2009;80(7):44–50 [quiz 51-2. PubMed PMID:19663144].

[8] Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. Methods Inf Med 2009;48(5):408–13 [Epub 2009 Jul 20. PubMed PMID: 19621114].

[9] Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inform Assoc 2008;15(2):130–7 [Epub 2007 Dec 20. PubMed PMID: 18096907; PubMed Central PMCID: PMC2274793].

[10] BRIDG release 2.2 static elements report; 2009. Available from: http://gforge.nci.nih.gov/frs/?group_id=342.

[11] Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 2004;102:20–38 [PubMed PMID: 15853262].

[12] ClinicalTrials.gov. Basic results data element definitions (DRAFT); 2009. Available from: http://prsinfo.clinicaltrials.gov/results_definitions.html.

[13] ClinicalTrials.gov. Protocol data elements definitions (DRAFT); 2009. Available from: http://prsinfo.clinicaltrials.gov/definitions.html.

[14] Nahm M, Zhang J. Operationalization of the UFuRT methodology for usability analysis in the clinical research data management domain. J Biomed Inform 2009;42(2):327–33 [Epub 2008 November 6. PubMed PMID: 19026765; PubMed Central PMCID: PMC2737809].