Life sciences domain analysis model

Robert R Freimuth,[1] Elaine T Freund,[2] Lisa Schick,[3] Mukesh K Sharma,[4] Grace A Stafford,[5] Baris E Suzek,[6] Joyce Hernandez,[7] Jason Hipp,[8] Jenny M Kelley,[9] Konrad Rokicki,[10] Sue Pan,[10] Andrew Buckler,[11] Todd H Stokes,[12] Anna Fernandez,[13] Ian Fore,[14] Kenneth H Buetow,[14] Juli D Klemm[14]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA
[2]3rd Millennium Inc, Waltham, Massachusetts, USA
[3]ScenPro, Inc., Richardson, Texas, USA
[4]Department of Pathology and Immunology, Washington University in St. Louis, St. Louis, Missouri, USA
[5]Computational Sciences, The Jackson Laboratory, Bar Harbor, Maine, USA
[6]Protein Information Resource, Department of Biochemistry and Molecular Biology, Georgetown University, Washington, DC, USA
[7]Global Clinical Data External Standards, Merck Sharp & Dohme Corp., Whitehouse Station, New Jersey, USA
[8]Department of Pathology, University of Michigan Health System, Ann Arbor, Michigan, USA
[9]Laboratory of Population Genetics, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, USA
[10]SAIC, Inc., McLean, Virginia, USA
[11]BBMSC, Wenham, Massachusetts, USA
[12]Department of Biomedical Engineering, Georgia Tech & Emory University, Atlanta, Georgia, USA
[13]Booz Allen Hamilton, Rockville, Maryland, USA
[14]NCI CBIIT, Rockville, Maryland, USA

Correspondence to
Dr Juli Klemm, NCI CBIIT, 2115 East Jefferson St., Rockville, MD 20852, USA;
klemmj@mail.nih.gov

Received 9 December 2011
Accepted 26 May 2012
Published Online First 28 June 2012

ABSTRACT

Objective Meaningful exchange of information is a fundamental challenge in collaborative biomedical research. To help address this, the authors developed the Life Sciences Domain Analysis Model (LS DAM), an information model that provides a framework for communication among domain experts and technical teams developing information systems to support biomedical research. The LS DAM is harmonized with the Biomedical Research Integrated Domain Group (BRIDG) model of protocol-driven clinical research. Together, these models can facilitate data exchange for translational research.

Materials and methods The content of the LS DAM was driven by analysis of life sciences and translational research scenarios and the concepts in the model are derived from existing information models, reference models and data exchange formats. The model is represented in the Unified Modeling Language and uses ISO 21090 data types.

Results The LS DAM v2.2.1 is comprised of 130 classes and covers several core areas including Experiment, Molecular Biology, Molecular Databases and Specimen. Nearly half of these classes originate from the BRIDG model, emphasizing the semantic harmonization between these models. Validation of the LS DAM against independently derived information models, research scenarios and reference databases supports its general applicability to represent life sciences research.

Discussion The LS DAM provides unambiguous definitions for concepts required to describe life sciences research. The processes established to achieve consensus among domain experts will be applied in future iterations and may be broadly applicable to other standardization efforts.

Conclusions The LS DAM provides common semantics for life sciences research. Through harmonization with BRIDG, it promotes interoperability in translational science.

BACKGROUND AND SIGNIFICANCE

To realize the promise of translational research, life sciences investigators, clinicians and informaticians must be able to meaningfully exchange information about remarkably diverse types of data. The potential rewards of this approach are matched by the difficulties in achieving them; the complexity of biological systems and the growing data deluge threaten to overwhelm researchers. It is even harder when the vast range of data and results are described in inconsistent and unclear terms. At best, this manifests itself as a distraction that decreases the efficiency of conducting the science; at worst, it is an obstacle to pursuing some of the most interesting and promising experimental paths. This lack of common semantics that spans domains of study impedes the flow of information and represents a significant challenge in translational research.

A domain analysis model (DAM) is an abstract, implementation-independent representation of the grammar, or semantics, of a domain. Both dynamic (business processes) and static (concepts) semantics of a domain may be described in a DAM;[1] this paper focuses on static semantics described in an information model. The model represents concepts and their relationships as classes, attributes and associations while using terms familiar to domain experts. The most important elements of a DAM include: (1) unambiguous definitions for the classes and their attributes, (2) the mutual exclusivity of concepts (eg, no two classes represent the same thing), (3) meaningful class names for the domain, (4) class associations with each other and (5) data types to unequivocally describe the intended use of an attribute.[2] A DAM provides a framework for communication among domain stakeholders and technical teams during development of information systems. As such, it can serve as a semantic foundation that may help lower barriers to information exchange.

The Life Sciences Domain Analysis Model (LS DAM[3]) is an information model representing the static semantics of hypothesis-driven and discovery-based research at the organismal, cellular and molecular levels (the dynamic semantics have been described elsewhere[4]). The creation of the LS DAM was initiated with the goal of harmonizing the semantics of a number of the cancer BioInformatics Grid (caBIG)® life sciences applications to support interoperability. The inputs to this effort have included the information models of these software applications as well as other foundational information models and data exchange formats within the life sciences domain, including FuGE-OM[5] and ISA-TAB.[6] The LS DAM builds from these efforts and extends them to support life sciences research.

The LS DAM modeling has also been informed by the Biomedical Research Integrated Domain Group (BRIDG) model,[7] which describes concepts important to protocol-driven clinical research and associated regulatory artifacts.[7 8] BRIDG is a collaborative effort of the National Cancer Institute (NCI), Health Level Seven (HL7), Clinical Data Interchange Standards Consortium and the Food and Drug Administration and is perhaps the most complete and widely adopted DAM in the biomedical research domain. The LS DAM shares many

classes and attributes with BRIDG, and together they support translational research through the shared semantics across the life sciences and clinical research domains.

Efforts to harmonize the LS DAM and BRIDG models highlighted the challenges of defining a common language across the life science and clinical research domains. A good example is the term 'protocol', which can have multiple meanings within and across these domains. In clinical research, the concept of protocol refers to the clinical study. Conversely, protocol in life sciences refers to a set of instructions for a specific experimental technique. Through the use of DAMs, subject matter experts can recognize familiar concepts and find precise definitions that remove ambiguity about meaning.

The availability of a shared, well-defined description of domain semantics is a valuable resource for domain stakeholders and technical teams to support the development of interoperable information systems. Derivations of the model can be used for implementation and, in addition, the model can be used generally to support a shared understanding of the domain among cross-functional teams. As such, it can be used to help lower the barriers to information exchange in biomedical research. Preharmonizing software applications' semantics to the LS DAM in a coordinated fashion would promote interoperability and give researchers the ability to integrate heterogeneous knowledge in a coherent fashion in order to pursue hypotheses that cross domain boundaries and were not previously possible. The result is increased productivity in research.

OBJECTIVES

The overall objective of the LS DAM project is to construct a technology-agnostic information model to be broadly used as a standard for semantic interoperability within the life sciences domain. The information model should represent life sciences research supporting (1) in vivo experiments, (2) ex vivo, in vitro or in situ experiments and (3) in silico research. The model should contain classes and attributes important for describing hypothesis-driven and discovery-based research at the organismal, cellular and molecular levels. In addition, by harmonizing with BRIDG, the two models can be utilized together to support the broad range of information necessary to describe translational research.

MATERIALS AND METHODS

The NCI caBIG®[9] is a collaborative information network with a shareable, interoperable infrastructure providing tools for cancer research. To support interoperability in the life sciences domain and in translational research, the caBIG® program sponsored a working group to construct the LS DAM. The LS DAM team, which assumed ownership of the project, consisted of a technical analyst who facilitated discussions and guided the modeling activity and a core group of 10 life science subject matter experts with varied experience in biological research, information modeling and familiarity with pre-existing caBIG® project models.

The LS DAM work has and continues to involve stakeholders from NCI's caBIG® project, HL7's Clinical Genomics Work Group (HL7 CGWG)[10] and the BRIDG project. More information about the project stakeholders, timelines, as well as the mechanisms for community input and feedback is posted on the LS DAM wiki.[11]

Standards used in the LS DAM

The group used the following standards in the construction of the LS DAM:

- The International Organization for Standardization 21090[12] standard provided a set of data type definitions for representing concepts commonly found in healthcare and medical research.
- The Unified Modeling Language an industry standard,[13] was used to create representations of life sciences concepts that can be readily understood by domain experts. Modeling was done in Enterprise Architect (EA).[14]
- Class and attribute names and definitions were based on standard terminologies, primarily the NCI Thesaurus (NCIt).[15]

LS DAM sources of information

Many of the classes and attributes included in the LS DAM are drawn from an inventory of established caBIG® project information models. By comparing models from multiple sources, shared classes and attributes representing areas of interest were identified, thus providing a foundation of classes on which to build. This bottom up approach to modeling promotes the reuse of elements already in use by researchers. Specific resources from which classes and attributes were drawn include caTissue (biobanking data management),[16] caArray (microarray data management),[17] caBIO (molecular annotations),[18] caMOD (cancer models database),[19] caLIMS (laboratory information management system),[20] caNanoLab (nanotechnology data sharing portal),[21] Annotation and Image Markup[22] and the National Biomedical Imaging Archive.[23]

A primary requirement for the LS DAM is harmonization with the BRIDG model to promote interoperability in support of translational research. To that end, the LS DAM team has adopted the semantics already defined and validated in the BRIDG model, whenever possible, reflected in the reuse of a number of fundamental classes including Person, Organization and Equipment.

The content and structure of the LS DAM also incorporates well-known life sciences models and exchange formats from independent modeling efforts including ISA-TAB[6] and FuGE.[5] In particular, these sources were used to identify and address critical gaps during the modeling process. For example, the LS DAM classes *ExperimentalParameter* and *ExperimentalFactor* are modeled following the ISA-TAB fields Factor Value and Parameter Value.[24]

The LS DAM team has also employed a top down approach to modeling, analyzing use cases, scenarios and other resources developed by working groups across caBIG® to foster collaboration and define interoperability requirements. For example, analysis of the caBIG® Enterprise Use Case EUC01_Translational Research[25] revealed the need to include *BiologicSpecimen, Protocol, Subject, Biomarker, Finding, Image* and *ImageAnnotation* classes. Similarly, analysis of the Integrative Cancer Research Scenario 8 (overlay of protein array data on the regulatory pathways with links to patient and cell culture)[26] supported the inclusion of *Pathway, Biologic Specimen, Protein, Data, Finding, Protocol, Experiment* and *Image* classes. Subject matter experts provided additional scenarios, which were helpful in identifying classes and attributes in specific areas of interest. Examples include *InvitroCharacterization* and *InvivoCharacterization* from nanotechnology informatics and *CellLine* and *CellCulture* from pathology imaging. Additionally, the LS DAM team has drawn on a business architecture model for life sciences,[4] which describes high-level business processes of the life sciences domain.

RESULTS

The LS DAM v2.2.1 is an information model that describes concepts central to life sciences research. Within the model, there are a number of logical groupings of related classes which

are referred to here as 'core areas'. Core areas currently supported by the LS DAM include Experiment (conducting an experiment), Molecular Biology (molecular biology entities), Molecular Databases (information management) and the Specimen (collection, processing and tracking of a specimen). The primary classes of each core area are described in tables 1–4. Figure 1[27] highlights the core areas of the LS DAM, illustrating the complete current model beyond the classes described here. A full view of the model is available in EA[28] and in HyperText Markup Language.[29]

Experiment core

The Experiment core is described in table 1 and figure 2.[27] This portion of the LS DAM describes the concepts that pertain to conducting experiments in both hypothesis-driven and discovery-based research in the life sciences domain. The core classes support scientists performing in vivo, in vitro and in silico research by representing the components that support planned and actual activities conducted as part of an experiment, the inputs (eg, protocols), and outputs (eg, data), thus supporting reproducibility. Members of the HL7 CGWG participated in the development of this core area, which is in the process of a formal review and feedback by the HL7 community.

Molecular Biology core

The Molecular Biology core (table 2) represents the molecular components of cells, such as nucleic acids and proteins. Genomic and proteomic annotations and experimental findings can be linked to the entity classes in the Molecular Biology core. This core and other relevant sections of the model will be extended through our ongoing collaboration with the HL7 CGWG.

Molecular Databases core

The Molecular Databases core (table 3) supports the association of identifying reference information held in various databases to the molecular entities (eg, proteins, genes, messenger RNAs, single nucleotide polymorphisms), interactions and pathways under investigation. By providing an information link to one or more reference databases, such as Ensembl,[30] GenBank,[31] or Kyoto Encyclopedia of Genes and Genomes,[32] important annotations can be associated to provide unambiguous identification of these entities.

Specimen core

The Specimen core of the LS DAM (table 4) contains classes and attributes describing the collection, processing and storage of a specimen. While the LS DAM supports handling of any type of specimen, this core currently emphasizes biological specimens. Classes in the Specimen core represent specimen collection and processing as well as the physical location of specimens.

LS DAM relationship to BRIDG

The integration of basic and clinical research findings required in translational research studies is often hampered by the distinct vocabularies used in these disciplines.

Many information models have been developed to support specific life sciences platforms, but few provide sufficient reach into clinical research sciences to support bench to bedside research. To address this challenge, a fundamental requirement for LS DAM modeling has been to maintain harmonization with the BRIDG model of clinical research on the touch points between the domains.

As such, of the 130 classes in LS DAM v2.2.1, 56 originate from BRIDG 3.0.2 classes. Shared concepts between the two models include people, organizations, places, materials and activities; indeed, several BRIDG classes and attributes have been adopted verbatim into the LS DAM, including *Organization* and *Person*. In addition, there are cases where the models share a superclass, but due to the unique semantics in each domain, each model has distinct subclasses. By way of example, the Unified Modeling Language class diagram in figure 3 illustrates harmonization of the LS DAM with BRIDG on the class of *Subject*. The shared superclass *Subject* provides the relationship between a person who is participating in a clinical trial (a BRIDG *StudySubject)* and that same person who has provided consent on a specimen collection protocol (an LS DAM *SpecimenCollectionProtocolSubject*).

Model validation

The LS DAM team evaluated the model to validate its support for requirements in the life sciences research domain, prioritized by stakeholders participating in the caBIG® program. This includes alignment with existing models support for representative research scenarios, and representation of information

Table 1 Experiment core classes with their definitions are listed

LS DAM class	Definition
BiologicalEntity	Any individual living (or previously living) being
BiologicalEntityGroup	A collection of biological entities
Data	A collection or single item of factual information, derived from measurement or research or other data, from which conclusions may be drawn
Equipment	An object intended for use whether alone or in combination for diagnostic, prevention, monitoring, therapeutic, scientific and/or experimental purposes
Experiment	A coordinated set of actions and observations designed to generate data, with the ultimate goal of discovery or hypothesis testing
ExperimentalFactor	An independent variable manipulated by the experimentalist with the intention of affecting biological systems in a way that can be measured by an assay
ExperimentalItem	Entities used in the execution of an experiment
ExperimentalParameter	Any factor that defines a system and determines (or limits) its performance; note: a parameter may have a default value and should not be represented by another class in the model (eg, a gene list)
ExperimentalStudy	A detailed examination or analysis designed to discover facts about a system under investigation; systems may include intact organisms, biological specimens, natural or synthetic materials, diseases, and pathways
Finding	An interpretation of results of an experiment
Organism	A grouping or classification of living things
PointOfContact	A person or organization (eg, helpdesk) serving as the coordinator or focal point of an activity or program
Protocol	A composite activity that serves as a rule which guides how activities should be performed
Software	A set of coded instructions which a computer follows in processing data, performing an operation or solving a logical problem upon execution of the program

LS DAM, Life Sciences Domain Analysis Model.

Table 2 Molecular Biology core classes with their definitions are listed

LS DAM class	Definition
AminoAcidPhysicalLocation	A physical location within an amino acid sequence
AminoAcidSequence	A representation of the linear arrangement of the molecules known as amino acid residues
Chromosome	A structure composed of a very long molecule of DNA and associated proteins (eg, histones) that carries hereditary information
DNASequence	A representation of the linear arrangement of deoxyribonucleotides comprising a DNA polymer
Exon	A portion of a gene sequence that is transcribed into the final mRNA product
Gene	A functional unit of heredity which occupies a specific position (locus) on a particular chromosome, is capable of reproducing itself exactly at each cell division, and directs the formation of a protein or other product
GeneticVariation	Deviation(s) in the nucleotide sequence of the genetic material of an individual from that typical of the group to which the individual belongs or deviation(s) in the nucleotide sequence of the genetic material of offspring from that of its parents
Genome	An assembly of the DNA sequence for the entire genome for a given organism
Intron	A portion of a gene sequence that is transcribed but excised from the mature mRNA during processing
MessengerRNA	A representation of (eg, as in public resources such as NCBI RefSeq or EBI Ensembl) a member of the class of RNA molecules that contains protein-coding information in its nucleotide sequence that can be translated into the amino acid sequence of a protein
NucleicAcidPhysicalLocation	A physical location within a nucleic acid sequence
Protein	A representation of (eg, as in public resources such as UniProt or NCBI RefSeq) an organic macromolecule composed of one or more chains (linear polymers) of α-L-amino acids linked by peptide bonds and ranging in size from a few thousand to over 1 million Daltons
RNASequence	A representation of the linear arrangement of ribonucleotides comprising an RNA polymer

LS DAM, Life Sciences Domain Analysis Model.

within well-established public information resources. The evaluation included comparisons of the following to the LS DAM: Common Biorepository Model (CBM),[33] whole slide imaging research scenarios developed by experts in pathology imaging and several public molecular biology databases. These three validation efforts are described below.

1. The CBM: The CBM is a representation of summarized biorepository information developed collaboratively by vendors and academic institutions to meet the goals of data sharing across these stakeholders. To evaluate the degree of shared semantics between these models, a mapping exercise was performed comparing classes and attributes of the models. This analysis confirmed that the semantics represented in the CBM is fully consistent with those described in the LS DAM. Therefore, information systems based on the CBM would have the potential to share data with information systems based on the LS DAM.
2. Whole slide pathology imaging scenarios: Subject matter experts in pathology imaging developed representative scenarios describing whole slide imaging with multi-site assessments. Mapping these scenarios to the LS DAM led to the identification of several gaps within the model. A number of these gaps were addressed by: (1) adding classes to support information about cultured cells; (2) modifying classes to distinguish an animal and an animal model; and (3) including support for identifying the individual person executing an activity. The remainder of the gaps will be addressed in future extensions of the model.
3. Molecular biology databases: To evaluate the comprehensiveness of the Molecular Biology core, the LS DAM team surveyed several public databases providing genetic variation, genomic and proteomic information. More than 200 entities, data types, roles and outcomes from the study of biological systems were identified in these public databases. Most of these entities were mapped directly to the existing Experiment core or Molecular Biology core classes in the LS DAM. A handful of gaps were identified and several of these have been addressed by adding attributes to classes in the Molecular Biology core; other gaps will be resolved in the future.

Additional validation of the LS DAM will occur over time as the model is used to facilitate information exchange in life sciences and translational research. For example, the successful application of BRIDG to addressing the interoperability gaps across the caBIG Clinical Trials Suite (CCTS) is a motivator for the LS DAM project. The CCTS is an enterprise clinical trials system comprised of a collection of interoperable modules that cover a broad range of key functions in cancer clinical trials management including patient registration, patient scheduling, adverse events reporting, lab analysis and clinical data management. The BRIDG model provides the foundation for achieving interoperability among these applications.

Several of the use cases that guide CCTS development could be extended to include translational research, thereby incorporating concepts that are supported by the LS DAM. For example, the CCTS 'Register Subject' scenario describes the workflow and information exchange between the patient registration system and other applications at the time that a subject is registered on a study. This scenario could easily be extended to include specimen collection, which would require information to be sent to a tissue banking system that is based on the LS DAM. The

Table 3 Molecular Databases core classes with definitions are listed

LS DAM class	Definition
GeneIdentifier	An identifier for a gene within some data source
MessengerRNAIdentifier	The unique identifier assigned to a transcript; it is used to uniquely identify a gene transcript in a given system, for example, ENSEMBL and RefSeq
PathwayIdentifier	The identifying information assigned to a pathway by a public data source such as KEGG
ProteinIdentifier	A unique identifier assigned to a protein in a system, for example, the ENSEMBL identifier, the RefSeq identifier and the UniProt Knowledgebase identifier
SingleNucleotidePolymorphismIdentifier	The unique identifier assigned to a reference SNP which is used to uniquely identify a SNP in a given system, for example, dbSNP

LS DAM, Life Sciences Domain Analysis Model.

Table 4 Specimen core classes with definitions are listed

LS DAM class	Definition
BiologicalSpecimen	Any material sample taken from a biological entity, including a sample obtained from a living organism or taken from the biological object after halting of all its life functions; a biospecimen sample can contain one or more components including but not limited to cellular molecules, cells, tissues, organs, body fluids, embryos and body excretory products
Container	An object that can be used to hold things (eg, box, tube, slide, rack)
Material	A physical substance
PerformedMaterialProcessStep	The completed act of processing a material (ie, a specimen or a nanomaterial) which includes but is not limited to freezing, thawing, spinning, embedding, dividing and aliquot; a specimen may be the result of a specimen processing step (ie, aliquot or division of a specimen)
PerformedObservationResult	The finding obtained by observing, monitoring, measuring or otherwise qualitatively or quantitatively recording one or more aspects of physiological or psychological processes
PerformedSpecimenCollection	The completed action of gathering samples that may be used for subsequent analysis (eg, blood draw)
PerformedSpecimenPlacement	The completed action of moving a specimen from one storage location to another
Place	A bounded physical location which may contain structures
SpecimenCollectionProtocol	A set of procedures that govern the collection of biospecimens
StorageEquipment	Equipment that is used for storage purposes

LS DAM, Life Sciences Domain Analysis Model.

feedback from development efforts that utilize the LS DAM will be used to further improve and extend the model.

DISCUSSION

To effectively exchange information within and across disciplines, the barriers to semantic interoperability must be overcome. These barriers include ambiguous meanings in the representations of concepts (eg, names, definitions, data types and the associations among these classes) and in the description of business processes. The LS DAM has been developed to provide a common representation of the static semantics that is important for the conduct of hypothesis-driven and discovery-based research in the life sciences domain.

Utility of LS DAM

Development of the LS DAM was initiated to address the need to harmonize the semantics across information systems developed through the caBIG® program. However, the value of the LS DAM extends beyond caBIG® program-sponsored tool development and can be used as the basis for implementing information systems and data exchange formats across the life sciences domain, thereby facilitating interoperability by providing a shared understanding of common concepts. Harmonizing the semantics of software applications to the LS DAM provides a foundation for the integration of knowledge provided by these applications. However, while the LS DAM provides common semantic meaning for concepts in life sciences research, coordination among development teams is required to ensure that constraints and extensions of the LS DAM are implemented consistently across applications.

The LS DAM has recently reached a breadth of scope needed to support the development of information systems, and several development projects have begun using the model to standardize the semantic representation of concepts relevant to their focus areas. The caLIMS v.2.0,[20] an open source laboratory information management system, contains localized LS DAM concepts relevant to a laboratory environment, such as equipment, data, experiment and biospecimen. In addition, the model for the Molecular Annotation Service,[34] a resource for accessing molecular annotations from curated data sources, contains classes and attributes derived from the Molecular Biology core (eg, genes, proteins and genetic variations). Both project teams have

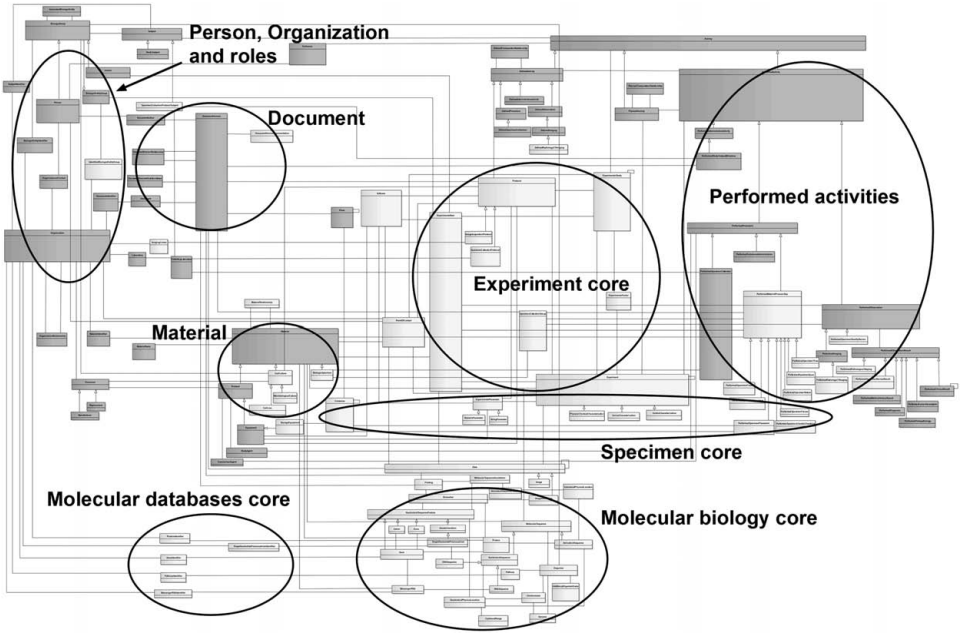

Figure 1 The Unified Modeling Language class diagram of the Life Sciences Domain Analysis Model (LS DAM) highlighting core areas. The full LS DAM class diagram is shown to illustrate the breadth and depth of the current model. Circles designate regions of the diagram containing classes related to core areas including those described here: Experiment core, Specimen core, Molecular Biology core and Molecular Databases core. Dark gray classes are harmonized to the Biomedical Research Integrated Domain Group. Light gray classes are unique to the LS DAM. Adapted from the Experiment Model Implementation Guide[27] with NCI CBIIT permission.

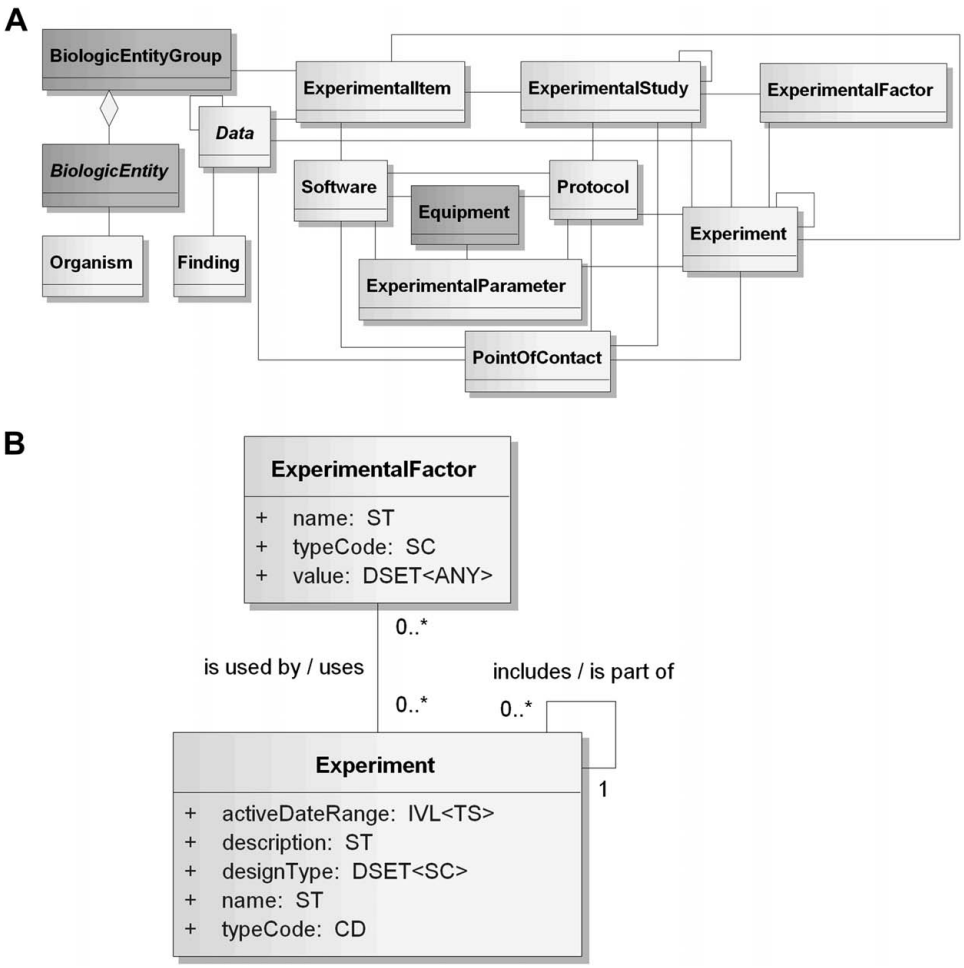

Figure 2 (A) Diagram of the Experiment core classes. Classes included in the Experiment core are shown with their associations. Dark gray classes are harmonized to the Biomedical Research Integrated Domain Group. Light gray classes are unique to the Life Sciences Domain Analysis Model. Adapted from the Experiment Model Implementation Guide[27] with NCI CBIIT permission. A HyperText Markup Language view of the Experiment core is available. (B) Example Experiment classes. *Experiment* and *ExperimentalFactor* classes are shown with attributes, associations, cardinalities and data types to illustrate details in the Life Sciences Domain Analysis Model which are not shown in the diagram of the Experiment core.

provided feedback that has been used as input to direct the evolution of the model.

The LS DAM can be considered a partner model to BRIDG and taken together these harmonized models support the semantics of translational research. By way of example, one can use these information models in combination to support preclinical trials involving model organisms. Classes and attributes related to in vitro and in vivo characterization are found in LS DAM, while classes and attributes related to study protocols and regulatory artifacts are found in BRIDG. Classes related to animals and specimens span both models. While practical considerations motivate the development and maintenance of LS DAM and BRIDG models by separate teams, maintaining their harmonization is a priority and the project teams are in regular dialog to ensure continued alignment and reuse of classes and attributes.

The LS DAM provides a necessary semantic baseline to support interoperability. However, it is important to emphasize that working interoperability requires the parties exchanging information to agree on several aspects, including how the model is constrained (eg, using the same set of attributes), the version of data types and the code systems used for coded values. To this end, organization-wide adoption can be supported by implementation guides to ensure consistent use of model components, data types and vocabularies. Consistent adoption on a larger scale can be obtained through the use of a standard specification produced by organizations such as Clinical Data Interchange Standards Consortium and HL7. The LS DAM can be used to inform the development of data exchange models by standards development organizations.

Robust documentation of the LS DAM is critical for development teams to make appropriate use of the model. To this end, each version of the model and its associated artifacts is published on a wiki landing page.[3] The current release contains: the full LS DAM model published from the EA modeling tool in EA,[28] HyperText Markup Language[29] and rich text format[35] formats; Model Documentation containing the model specifications; Release Notes that include the logic and rationale behind the modeling decisions made and information on future directions of the effort; and an Experiment Model Implementation Guide that describes the goals and approach to the development of the Experiment core, as well as explanations of the concepts included in the core, and a sample instantiation of the Experiment core to support the described sample scenario.

Challenges

A consistent challenge in developing the LS DAM has been the community-specific use of terms within and across domains. For example, the terms 'study', 'method' and protocol can have different meanings to a bench scientist or clinical researcher, and can even have a somewhat different meaning to researchers within each discipline. As a result, the LS DAM may use a different term to name a concept from what an individual investigator might choose. The model provides an unambiguous definition for each concept to assure concordance of these synonyms. Classes that originate from BRIDG use the BRIDG definition, while the definitions for other classes are generally chosen from the NCIt. In instances where the LS DAM team determined that existing definitions in the NCIt or BRIDG needed modifications to fit the LS DAM's domain, proposed

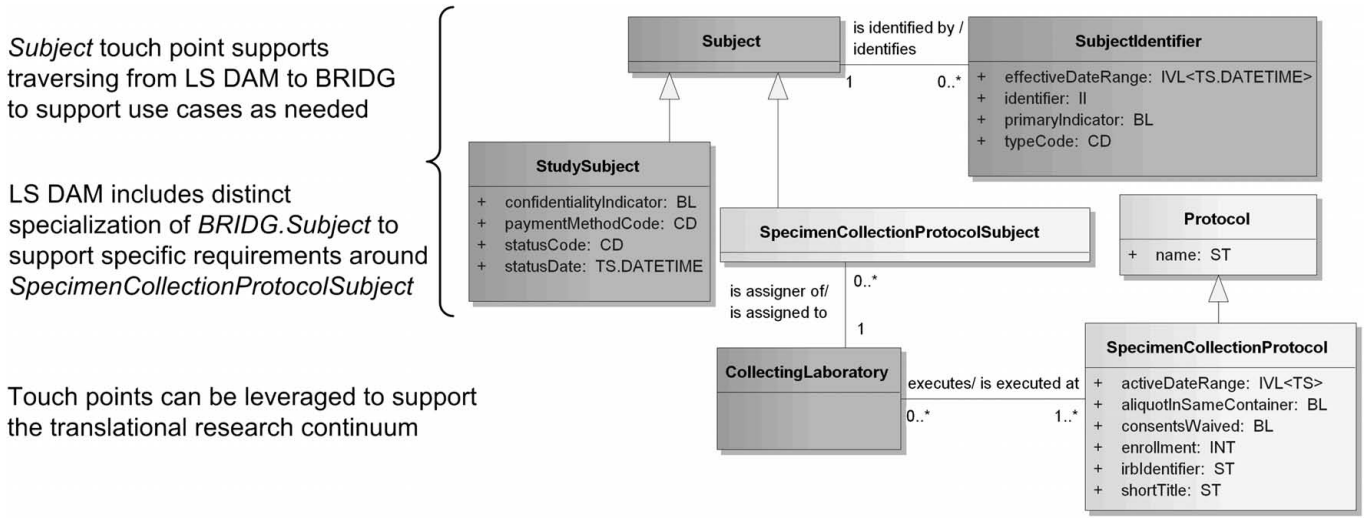

Figure 3 The Biomedical Research Integrated Domain Group (BRIDG) and the Life Sciences Domain Analysis Model (LS DAM) share *Subject*. The BRIDG *Subject* class is an example of a touch point between BRIDG and the LS DAM which facilitates traversing and simultaneous utilization of both models. Dark gray classes are harmonized to BRIDG. Light gray classes are unique to the LS DAM. Adapted from the Experiment Implementation Guide[27] with NCI CBIIT permission.

changes have been submitted to the BRIDG Semantic Coordination Committee[7] or the NCIt team.[36]

While a significant number of LS DAM classes are harmonized with BRIDG, some concepts cannot be harmonized due to differing domain requirements. In these cases, the LS DAM introduces new concepts with unique names to eliminate ambiguity and address the needs of life science researchers. For example, 'specimen' is defined as a role in BRIDG 3.0.3[7] (eg, an entity may play the role of a specimen in an *Experiment)*, but the concept of a specimen as a physical entity is integral to the biobanking community. As such, the LS DAM represents a specimen as an entity to be processed, cataloged and stored, in addition to playing a role of *ExperimentalItem* in an *Experiment*.

The foundation of the LS DAM was built from classes and attributes from existing models. Since the LS DAM and project models were under concurrent development, the LS DAM team worked closely with stakeholders to inform the evolution of the LS DAM. This approach led to maximal reuse of established classes and attributes despite concurrent development of the implementation models that were used as a reference, thereby fostering interoperability between existing tools and those to be developed.

Next steps

The LS DAM has reached a level of maturity to support implementation efforts but will continue to evolve with the requirements of researchers and developers. Several gaps and additional requirements have been identified and will be addressed through continued development and maintenance of the model. Receiving and addressing feedback from development teams that have used the LS DAM, like the caLIMS and Molecular Annotation Service teams, are critical to maintaining the quality and utility of the model. Additionally, feedback received from the HL7 CGWG collaborators is a priority. They asked for increased level of detail in genetic variation to support clinical genomics reporting in translational medicine. The LS DAM will also evolve to address the needs of the community in response to rapid technological advances and changing interests in life science research.

The HL7 CGWG is using the LS DAM Experiment core to inform development of the Omics DAM. As part of our continued collaboration, the HL7 CGWG will bring gaps forward to the LS DAM team to be added to support their requirements. These gaps will be assessed for harmonization into the LS DAM.

In addition to harmonization with BRIDG and collaboration with the HL7 CGWG, the LS DAM team is also seeking to expand collaborative efforts with other standard development groups. Several suggestions for future activities have been gathered, prioritized and are available on the LS DAM wiki.[37] Additional community input to this list is welcomed.[38] Harmonization of classes and attributes from existing standards will reduce repetitious efforts while expanding the LS DAM's support of life sciences and translational research.

Finally, it is important to note that the open nature of the LS DAM and associated artifacts allows adopting institutions to locally extend the model to meet the specific needs of their environment. While there is no requirement to do so, harmonization of such extensions back into the primary LS DAM, or simply posting such changes to the LS DAM wiki, will broadly benefit the entire community.

CONCLUSIONS

The LS DAM is a reference model providing a shared view of the static semantics of the domain of hypothesis-driven and discovery-based research at the organism, cell and molecular levels in order to facilitate human understanding and computable interoperability. The LS DAM provides a rich resource of shared semantics helping the life sciences community integrate, mine and reuse data. With its touch points to BRIDG, the LS DAM facilitates information sharing and data integration for translational science. As BRIDG strives to be the 'semantic bridge' between clinical research and care,[10] the LS DAM begins to bridge clinical research and life sciences research, enabling the 'bench to bedside and back' paradigm.

Ultimately, the usefulness of the LS DAM will be measured in terms of stakeholders and adoption of the model. The LS DAM fills a gap in standard efforts and is supported by a substantial and accessible set of documentation. As the model continues to evolve by addressing feedback through a collaborative process

involving representatives from several different perspectives, it will increasingly serve the needs of the community.

Acknowledgments The authors thank Patti Kwong and Sharron Lewis for support; the HL7 CGWG and BRIDG SCC groups; Smita Hastak, Rakesh Nagarajan, Fred Prior, Charlie Mead, John Koisch and Steven Freund for helpful feedback.

Contributors RRF, ETF, BES contributed domain expertise, provided analysis, contributed to model development, provided feedback on draft models, and contributed to writing and editing the manuscript. LS provided analysis, model development expertise, and contributed to writing and editing the manuscript. MKS, GAS, J Hipp, JMK contributed domain expertise, provided analysis, contributed to model development, provided feedback on draft models, contributed use cases, and contributed to writing and editing the manuscript. J Hernandez provided analysis, contributed to model development, provided feedback on draft models, contributed use cases, and contributed to writing and editing the manuscript. KR provided analysis, contributed to model development, provided feedback on draft models, and reviewed and provided feedback on the manuscript. SP provided analysis, contributed to model development, provided feedback on draft models, contributed use cases, and reviewed and provided feedback on the manuscript. AB contributed domain expertise, provided feedback on draft models, contributed use cases, contributed to editing the manuscript, and reviewed and provided feedback on the manuscript. THS contributed domain expertise, contributed to model development, contributed use cases, and reviewed and provided feedback on the manuscript. AF, IF contributed domain expertise, provided feedback on draft models, and reviewed and provided feedback on the manuscript. KHB reviewed and provided feedback on the manuscript. JDK contributed domain expertise, provided feedback on draft models, and contributed to writing and editing the manuscript. RRF, ETF, LS, MKS, GAS and BES authors have contributed equally.

Funding This work was supported by contracts from Booz Allen Hamilton and SAIC-Frederick on behalf of the National Cancer Institute as part of the caBIG® Integrative Cancer Research Workspace.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All of the data and information artifacts produced in this work are publically available.

REFERENCES

1. **HL7 International.** *Domain Analysis Model*. 2008. http://wiki.hl7.org/index.php?title=Domain_Analysis_Model (accessed 21 Sep 2011).
2. **Ambler SW.** *Agile Modeling. UML 2 Class Diagram Guidelines*. http://www.agilemodeling.com/style/classDiagram.htm (accessed 22 Sep 2011).
3. **National Cancer Institute.** *Life Sciences Domain Analysis Model (LS DAM)*. 2011. https://wiki.nci.nih.gov/x/cxRIAQ (accessed 21 Sep 2011).
4. **Boyd LB,** Hunicke-Smith SP, Stafford GA, *et al*. The caBIG® life science business architecture model. *Bioinformatics* 2011;**27**:1429–35.
5. **FuGE | Home Page.** *Functional Genomics Experiment (FuGE) Home Page*. http://fuge.sourceforge.net (accessed 28 Apr 2011).
6. *ISA Infrastructure*. http://isatab.sourceforge.net (accessed 21 Sep 2011).
7. **BRIDG.** *Biomedical Research Integrated Domain Group*. 2011. http://www.bridgmodel.org/ (accessed 21 Sep 2011).
8. **Fridsma DB,** Evans J, Hastak S, *et al*. The BRIDG project: a technical report. *J Am Med Inform Assoc* 2008;**15**:130–7.
9. **National Cancer Institute.** *caBIG® Cancer Biomedical Informatics Grid®*. https://cabig.nci.nih.gov/ (accessed 21 Sep 2011).
10. **HL7 International.** *Clinical Genomics*. 2009. http://www.hl7.org/Special/committees/clingenomics/overview.cfm (accessed 21 Sep 2011).
11. **National Cancer Institute.** *IRWG LS DAM Feedback*. 2012. https://wiki.nci.nih.gov/x/ClHbAQ (accessed 8 Feb 2012).
12. **International Organization for Standardization.** *International Standards for Business, Government and Society*. http://www.iso.org/iso/home.html (accessed 21 Sep 2011).
13. **Unified Modeling Language (UML®).** *OMG®, We Set the Standard®*. http://www.omg.org/spec/UML/ (accessed 21 Sep 2011).
14. **Sparx Systems.** *Enterprise Architect*. 2011. http://www.sparxsystems.com/products/ea/index.html (accessed 21 Sep 2011).
15. **National Cancer Institute.** *Enterprise Vocabulary Services. NCI Thesaurus*. 2011. http://ncit.nci.nih.gov/ncitbrowser/ (accessed 21 Sep 2011).
16. **National Cancer Institute.** *caTissue Suite*. 2011. https://cabig.nci.nih.gov/tools/catissuesuite (accessed 21 Sep 2011).
17. **National Cancer Institute.** *caArray - Array Data Management System*. 2011. https://cabig.nci.nih.gov/tools/caArray (accessed 21 Sep 2011).
18. **National Cancer Institute.** *caBIO Wiki Home Page*. 2011. https://wiki.nci.nih.gov/x/4Q9y (accessed 21 Sep 2011).
19. **National Cancer Institute.** *Cancer Models Database (caMOD)*. 2011. https://cabig.nci.nih.gov/tools/caMOD (accessed 21 Sep 2011).
20. **National Cancer Institute.** *caLIMS v2 Wiki Home Page*. 2011. https://wiki.nci.nih.gov/x/2oMYAQ (accessed 21 Sep 2011).
21. **National Cancer Institute.** *caNanoLab*. 2010. https://cananolab.nci.nih.gov/caNanoLab/ (accessed 21 Sep 2011).
22. **National Cancer Institute.** *Annotation and Image Markup (AIM)*. 2010. https://cabig.nci.nih.gov/tools/AIM (accessed 21 Sep 2011).
23. **National Cancer Institute.** *National Biomedical Imaging Archive*. 2010. https://cabig.nci.nih.gov/tools/NCIA (accessed 21 Sep 2011).
24. **ISA Infrastructure.** *ISA-TAB Format Specification*. 2008. http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf (accessed 22 Sep 2011).
25. **National Cancer Institute.** *caBIG® Enterprise Use Cases*. 2009. http://gforge.nci.nih.gov/docman/index.php?group_id=576&selected_doc_group_id=4588&language_id=1 (accessed 22 Sep 2011).
26. **National Cancer Institute.** *ICRi Scenarios*. 2009. https://wiki.nci.nih.gov/download/attachments/39912351/ICR_Interoperability_Scenarios_20090212.doc (accessed 22 Sep 2011).
27. **National Cancer Institute.** *Experiment Model Implementation Guide Life Sciences Domain Analysis Model v2.2*. 2011. https://wiki.nci.nih.gov/download/attachments/23401587/Experiment+Model+Implementation+Guide.docx (accessed 28 Apr 2011).
28. **National Cancer Institute.** *LS DAM 2.2.1 Model Download*. 2011. https://wiki.nci.nih.gov/download/attachments/23401587/LSDAM_2_2_1.EAP (accessed 22 Sep 2011).
29. **National Cancer Institute.** *LS DAM*. 2011. http://ncientarch.info/sandbox/LSDAM2_2_1/ (accessed 22 Sep 2011).
30. **Flicek P,** Amode MR, Barrell D, *et al*. Ensembl 2011. *Nucleic Acids Res* 2011;**39**:D800–6.
31. **Benson DA,** Karsch-Mizrachi I, Lipman DJ, *et al*. GenBank. *Nucleic Acids Res* 2008;**36**:D25–30.
32. **Kanehisa M,** Goto S, Furumichi M, *et al*. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010;**38**:D355–60.
33. **National Cancer Institute.** *Common Biorepository Model (CBM)*. 2010. https://cabig.nci.nih.gov/workspaces/TBPT/CBM/ (accessed 19 Sep 2011).
34. **National Cancer Institute.** *Molecular Annotation Service Case Study*. 2011. https://wiki.nci.nih.gov/x/vZv9AQ (accessed 19 Sep 2011).
35. **National Cancer Institute.** *LS DAM 2.2.1 Model Documentation*. 2010. https://wiki.nci.nih.gov/download/attachments/23401587/LSDAM2_2_1+ModelDocumentation.rtf (accessed 22 Sep 2011).
36. **National Cancer Institute.** *NCI Term Form*. 2011. https://wiki.nci.nih.gov/x/ZwVyAQ (accessed 8 Feb 2012).
37. **National Cancer Institute.** *Suggested LS DAM Future Activities: Compilation of Items and Background*. https://wiki.nci.nih.gov/download/attachments/23401587/LS+DAM+2_2Future+Activities+Compilation.docx?version=1&modificationDate=1304986410000 (accessed 22 Sep 2011).
38. **National Cancer Institute.** *IRWG LS DAM Feedback*. 2011. https://wiki.nci.nih.gov/x/ClHbAQ (accessed 18 Sep 2011).