Using Multiple Reference Ontologies: Managing Composite Annotations

John H. Gennari[1], Maxwell L. Neal[1], Jose L.V. Mejino Jr.[2], and Daniel L. Cook[2,3]

Departments of [1]Medical Education & Biomedical Informatics, [2]Biological Structure, and Physiology & Biophysics[3], University of Washington, Seattle, WA

Abstract

There are a growing number of reference ontologies available across a variety of biomedical domains and current research focuses on their construction, organization and use. An important use case for these ontologies is annotation—where users create metadata that access concepts and terms in reference ontologies. We draw on our experience in physiological modeling to present a compelling use case that demonstrates the potential complexity of such annotations. In the domain of physiological biosimulation, we argue that most annotations require the use of multiple reference ontologies. We suggest that these "composite" annotations should be retained as a repository of knowledge about post-coordination that promotes sharing and interoperation across biosimulation models.

Connecting multiple reference ontologies

We define a reference ontology as a carefully-constructed ontology that aims to completely cover a specific realm or domain of knowledge[1,2]. By definition, such an ontology must be both broad and deep in its domain, and designed for reusability across multiple sorts of users and use cases. In biology, one goal of the OBO resource (http://obofoundry.org/) is to encourage the development of non-overlapping reference ontologies so that users can unambiguously access terms from such ontologies. In biology, an exemplar reference ontology is the Foundational Model of Anatomy (FMA)[2].

Ontologies are most effective when they are designed with specific use cases in mind. For many, the motivating use case has been *annotation*: users need to add unambiguous semantic metadata about their raw data, whether that data is from genomic research, clinical findings, or images. To date, the conception of these annotations has been relatively simple. For example, a gene expression level from some experimental result will be annotated in-line with a Gene Ontology (GO) id, or possibly a direct URI to the relevant GO term.

Annotations (even simple ones) provide a compelling justification for ontology development. Annotations allow external users, or even computer systems to explore and automatically align data and results across multiple sources. This use of annotations requires users to carry out two sorts of tasks: (1) annotating source data against ontologies, and (2) searching and integrating data from sources that use those ontologies for annotation. As others have pointed out, these tasks fit well into the intelligent information retrieval capabilities of the semantic web[3].

In this paper, we argue that this relatively simple use of annotation and ontologies can become very complex if annotations include multiple ontologies. Our domain of interest is in biosimulation, where researchers build models for understanding pathology or physiology. We show that when researchers annotate such models, they need to use multiple orthogonal ontologies. We present our preliminary architecture for these *composite annotations*, and describe prototype tools and ideas for the two user tasks described above: Annotating biosimulation models and then searching and integrating those models.

As we show, these annotations provide a solution to one case of the post- vs. pre-coordination problem: there are too many properties of too many biological entities to attempt to pre-coordinate all combinations. Instead, via composite annotations, users can post-coordinate concepts as needed, and store those combinations of terms across ontologies that are useful and relevant for their tasks. Without retaining this knowledge, ontology developers and end users are faced with a combinatorial problem—a cross product of terms across many large orthogonal ontologies.

The Biophysical Semantics of Biosimulation

For several years, our research group has been developing systems and ontologies for use with physiological biosimulation models. Recently, researchers have aimed at building a complete Physiome[4], a flexible integration of component models into large-scale or special-purpose biosimulations for application to clinical and investigatory problems. Toward this goal, a number of libraries of biosimulation models have been made available, notably BioModels (an SBML collection, http://www.ebi.ac.uk/biomodels/), the CellML repository (http://www.cellml.org), and the JSim library (http://physiome.org/jsim/).

The fundamental challenge for integrating and understanding biosimulation models is that although these models are based on classical physics and formally expressed in mathematics, the semantics of these models—the meaning of variables and equations—is usually only implicit in model computational code (e.g. naming conventions) or annotated using *ad hoc* in-line code comments. Although current best practices in biosimulation modeling include adherence to some annotation standards[5], these have not yet been widely adopted. We certainly applaud the use of OBO standards such as the FMA, ChEBI, GO, and the OBO Cell Type ontology. However, if annotations for biosimulation models are in-line, maintaining and searching over these annotations can be a challenge.

In addition, all of the above ontologies are for biological structure and physical entities. For physiological modeling, it is important to also represent the principles by which such entities participate in processes. Recently, we have developed the Ontology of Physics for Biology (OPB)[6], an ontology of the physical properties and physical laws by which biological processes occur. As such, it is orthogonal to strictly structural representations (e.g., FMA, ChEBI) in that it represents the physical properties that reside in structural entities. Thus, in biosimulation models, the elements of interest necessarily include both reference to structural entities of biology (E.g. blood, muscle, or smaller entities such as glucose or oxygen.) as well as ***properties*** of those entities (e.g., flow, mass, or chemical concentration). In the next section, we provide specific examples of these composite annotations.

Example composite annotations

As a simple example, consider a common concept used in many cardiovascular biosimulation models: Aortic blood pressure. This concept may be mapped to differently named variables (Aop, AP, PAorta, etc) in different models. To integrate models that share this concept, these variables would have to be annotated with both the anatomical entity (blood-in-aorta) as well as the physical property that is modeled: fluid pressure. This is a simple example, because it involves just two reference ontologies, the FMA and the OBP, and because fluid pressure is a property of the FMA entity blood-in-aorta.

As a slightly more complex example, consider the concentration of oxygen in the blood of the aorta. This entity (which might be used by many different biosimulation models) needs three ontologies: ChEBI, for oxygen, the OPB, for chemical concentration, and the FMA, for blood in the aorta. If we omit any of these three ontologies, our representation is inaccurate or even erroneous. If we are not explicit about chemical concentration then we might be discussing (for example) the flow of oxygen in the aorta. If we omit the aorta, we might be discussing concentration of oxygen in the vena cava. Finally, we obviously need ChEBI for oxygen as there are many chemicals of interest in the aortic blood (e.g., calcium ion concentration).

Finally, annotations become most complex in models that are multi-scale. Consider a model that includes glucose concentration in beta cells. It may matter a great deal whether that concentration is cytoplasmic, extracellular, arterial, or venous. Potentially, such a concept might need five reference ontologies: cell component (e.g., GO cell component), cell type (e.g. the OBO CellType), as well as the FMA, the OPB, and ChEBI.

Effectively, composite annotations are recording "cross-products of interest" over the participating reference ontologies. Thus, one could imagine a set of tuples for pathway level biosimulation that were {OPB x ChEBI x FMA} or perhaps {OPB x ChEBI x GOCellComponent}. However, the vast majority of such tuples would be nonsensical or not of interest for a particular model or group of biosimulation researchers (e.g., momentum of oxygen in the skull bone). In addition, our composite annotations need internal structure—formal terms that describe the relationship between, for example, blood and the aorta ("contained-in"). The research questions we raise deal with how to create, store, and retrieve for reuse, these sort of composite annotations.

Managing annotations: SemSim for biosimulation

For a single biosimulation model, we have developed an approach to composite annotation we call a "SemSim model" (for Semantic Simulation)[7,8]. SemSim models are OWL-based ontologies that capture the computational and semantic aspects of a biosimulation model, and they include a set of annotations for that particular biosimulation model. At most, there is one annotation per variable and equation in the source code. For variables, these are *composite annotations*, where each annotation has the structure we diagram in Figure 1.

Biosimulation model variables, such as "PAorta", are annotated by first mapping them to physical properties, such as pressure, flow, concentration, etc. These properties are defined in the OPB, and referenced in the composite annotation. It is these properties that take on numeric values during any given simulation run. As Figure 1 shows, these properties are then connected to the physical entities (via "has property" links) which then point to entities in struc-

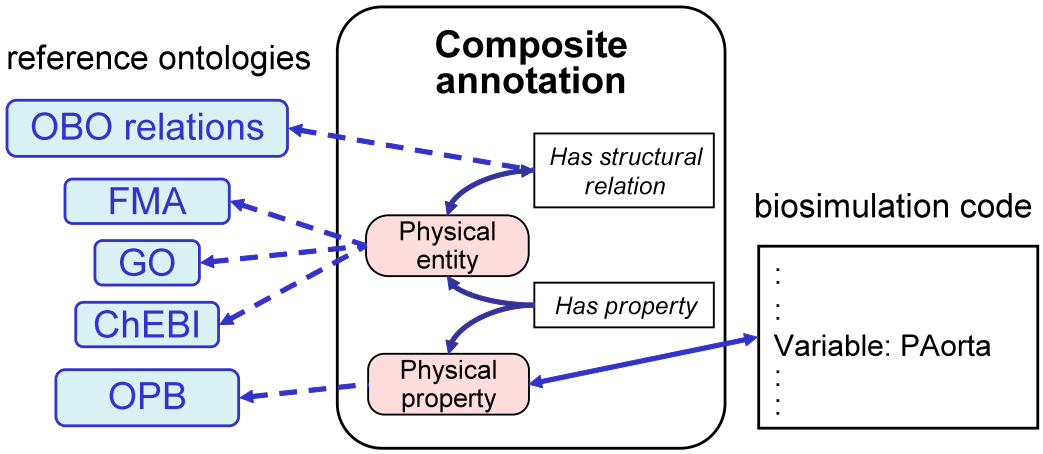

Fig 1. The structure of our composite annotations, which connect variables in simulation code to a set of reference ontologies. A SemSim model is a collection of these annotations, for a set of variables as used in a specific piece of biosimulation code.

tural reference ontologies. If there is more than one entity and more than one structural ontology (e.g., oxygen in blood), then these are connected via "has structural relation" links, and then connected to the OBO relations ontology.

In contrast to current annotation practices, our SemSim models are external entities to both the reference ontologies and the biosimulation code. One advantage of this design is that the source code can remain un-modified, an important feature when annotating legacy code. A second advantage is that we can easily collect the set of all annotations as a repository for search and reuse.

Creating composite annotations

For biosimulation researchers, the structure shown in Figure 1 should be largely invisible. Thus, we have developed prototype tools that hide this representational complexity and help users author and create composite annotations from biosimulation source code. For creating annotations, our prototype tool, SemGen, parses the source code to find instances of variables, and then prompts the user for search terms to use with particular ontologies. The system then queries these ontologies to find exact matches and IDs for the terms, and finally generates the composite annotation as part of a SemSim model.

As we develop a larger repository of annotations for biosimulation models, our SemGen system can better assist users. For, example, if a model uses a variable that captures "cytoplasmic glucose concentration in pancreatic beta cells", then this annotation could require five searches across five participating reference ontologies. However, if some other user has already created a similar or related annotation, then the SemGen system can return a list of these as soon as the user enters any one of these terms. E.g., as soon as "glucose" is entered, the system could return a list of all prior glucose annotations, and one of these may be a close or perfect match for the user.

Because there are relatively few biosimulation models available, the number of useful composite annotations for models is small, at least compared to the cross product of the cardinality of the reference ontologies. Thus, annotators help us carry out post-coordination of terminologies: the composite annotations are created only on an as-needed basis, and then stored in a repository for reuse.

Using annotations to search and merge models

As we alluded to earlier, there are two sorts of user tasks for annotations. In addition to creating composite annotations (e.g., with SemGen), users need to search annotations and their models, and then perhaps merge or adapt models created by others. Reusing and adapting others' models is common in biosimulation engineering, but currently, this work is manual, costly, error-prone, and typically requires extensive communication and collaboration between bioengineers.[8]

In prior publications, we have presented early results that show how our SemSim approach would make model merging semi-automatic.[7,8] Although promising, this preliminary work avoids some of the broader indexing and retrieval challenges for a repository of composite annotations. In particular, for semantic web use cases, composite annotations need (a) a unique name or URI, and (b) indices for appropriately efficient retrieval. We can assume that each reference ontology term (such as "FMA: blood in

aorta") already has a URI. Thus, although unwieldy, one could use URIs for composite annotations that simply consist of concatenations of the URIs of each reference ontology term.

We believe that users may want to search the annotation repository in a variety of ways. Thus, it seems likely that these annotations will need to be indexed with all of their component terms. To continue with the glucose example, users may want to begin with glucose, or pancreatic beta cells, or "cytoplasmic glucose" and therefore, all of these should be indexed, so that the system can retrieve the full term regardless of how the user searches.

Another complication is that SemSim annotations, as currently implemented, include a pointer directly to the source code variable name. Thus, even if two models mean exactly the same thing by "aortic pressure", the annotations would still be different because they would refer to different source code variable names (possibly in different biosimulation languages). This design leads to challenges for automatically finding duplicate concepts and for merging models with such shared concepts.

Managing orthogonal ontologies: OBO relations

The management of multiple ontologies for annotating biosimulation models is just a specific example of managing multiple orthogonal ontologies. This issue is faced by the OBO set of ontologies, and partially addressed by the OBO Relation Ontology. This ontology provides the formal relations needed to describe ***how*** the structural entities in a composite annotation relate to each other. For example, for cytoplasmic glucose concentration in beta cells, we can say precisely that we are referring to the cytoplasm (GO CellComponent) that is "part of" (OBO relation) the pancreatic B cell (OBO CellType).

Thus, the OBO relations ontology provides the ability to appropriately link entities across OBO ontologies that pertain to structural entities. However, this ontology does not include relationships appropriate for connecting non-structural ontologies such as the OPB. How should the notion of "pressure" be related to the concept of "blood"? In our SemSim approach, we currently use the generic "has property" relation for such links.

Pragmatically, our initial work has focused on managing and building composite annotations. We certainly use the OBO relation ontology where appropriate, but as a first goal, building a corpus of useful composite annotations will be a significant contribution, and can ease the task of biosimulation model integration.

Discussion and conclusions

In this paper, we describe composite annotations to represent entities of interest to biosimulation modelers. In addition, we propose that these annotations can be used as a way of storing knowledge about post-coordination, so that useful terms such as "concentration of oxygen in blood of aorta" can be easily retrieved or created on-the-fly. Elsewhere, we demonstrated the value of such annotations for merging biosimulation models, and here, we raise issues and propose possible solutions for building a semantic web repository of such composite annotations.

In support of the Physiome vision, the biosimulation research community is working to integrate models to build larger and more complex models (with the expectation that such models are more predictive and useful). We argue that reference ontologies and tool support could provide significant assistance with this work. However, a key first step to integrating models is a solid understanding of the semantics of model variables and equations. We propose that a repository of composite annotations could both make annotation of additional models easier, as well as allow researchers and systems to find variables that share common semantics across biosimulation models.

Acknowledgements

This work was partially funded by NIH grants #R01 HL087706-01 and #T15 LM007442-06. We also thank Michal Galdzicki for contributions to these research ideas.

References

1. Brinkley, J.F., et al., *A framework for using reference ontologies as a foundation for the semantic web.* AMIA Annu Symp Proc, 2006: p. 96-100.
2. Rosse, C. and J.L.V. Mejino, *A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy.* Journal of Biomedical Informatics, 2003. **36**: p. 478-500.
3. Sahoo, S.S., et al., *An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence.* Journal of Biomedical Informatics, 2008. **41**(5): p. 752-765.
4. Hunter, P.J. and T.K. Borg, *Integration from proteins to organs: the Physiome Project.* Nat Rev Mol Cell Biol, 2003. **4**(3): p. 237-43.
5. Le Novere, N., et al., *Minimum information requested in the annotation of biochemical models (MIRIAM).* Nat Biotechnol, 2005. **23**(12): p. 1509-15.
6. Cook, D.L., et al., *Bridging Biological Ontologies and Biosimulation: The Ontology of Physics for Biology.* AMIA Annu Symp Proc, 2008: p. 136-140.
7. Gennari, J.H., et al., *Integration of multi-scale biosimulation models via light-weight semantics.* Pac Symp Biocomput, 2008. **13**: p. 414 – 425.
8. Neal, M.L., et al., *Advances in semantic representation for multiscale biosimulation: A case study in merging models.* Pac Symp Biocomput, 2009: p. 304-315.