ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Towards an Ontology for Conceptual Modeling

James P. McCusker, Joanne Luciano, Deborah L. McGuinness

Tetherless World Constellation, Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

Abstract. Conceptual modeling can be viewed as a way of expressing human understanding of a body of knowledge. This view can be viewed as distinct from standard notions of data modeling and ontology, which seek to directly describe data and reality. We define our notion of conceptual interoperability, give use cases and requirements for it, and introduce the Conceptual Model Ontology (CMO), which satisfies the discussed use cases and requirements. We show how, using a common vocabulary, conceptual models can be used to tie together data at the level of conceptual interoperability. Finally, we introduce an implementation of CMO in the semantic web Biomedical Informatics Grid (swBIG), a linked data proxy for cancer Biomedical Informatics Grid (caBIG) models, semantic metadata, and data.

1 Introduction

The relationship between entities, the idea of the entities, and their information representation has come to the forefront of ontology and information modeling because conceptual models [1] have become critical in encoding human understanding of information [2]. To support this, layered representations of information models, such as the conceptual, logical, and physical model layers, [3] [4] have become common practice in many modeling disciplines. We seek a meta-modeling ontology that can easily express human understanding of entities and their data in terms of independent, reusable vocabularies that can be annotated onto conventional ontologies in a way that does not computationally disturb the annotated ontology (does not produce any undesired inferences) and does not require modification of the ontology or the data it represents. Our goal is to provide a way to use these sorts of annotations to satisfy certain use cases for conceptual interoperability [5].

2 Background

Biomedical ontologies have, for more than ten years, worked towards interoperability of data through use of verified categories, as has been the case in the Gene Ontology [6] and other OBO Foundry ontologies [7], reference models, as has been the case in Health Level 7 (HL7) Reference Information Model [8] and openEHR [9], and common vocabularies, such as SNOMED-CT, LOINC, and NCI Thesaurus. However, integration across these ontologies has identified a number of challenges surrounding the strategies that were used to produce the ontology.

Ontologies are often created using one of two primary influences: linguistic or realist. Linguistic influences on ontologies stem from how people talk about and understand entities, whereas realist influences on ontologies stem from a focus on scoping by including only things that have scientific evidence of existence in the real world [10]. We refer to HL7-RIM as being linguistically influenced because it is primarily concerned with communication of human-generated records between entities. Ontologies with realist influence attempt to model reality as it is, and only model things for which there is scientific evidence [11]. The Basic Formal Ontology (BFO) [12] is possibly the most rigorous example of an ontology with realist influences. A realist strategy can provide a framework for relating other strategies, including conceptual models. Smith *et al.* [2] have developed a three-layer system for things in the world, our ideas of them, and representations of those ideas. Specifically, they define the following levels of entities that are involved in ontologies:

Level 1: the objects, processes, qualities, states, etc. in reality (for example on the side of the patient);

Level 2: cognitive representations of this reality (for example, on the part of researchers and others);

Level 3: concretizations of these cognitive representations in representational artifacts (for example, textual or graphical).

In a conceptual model, we consider Level 1 entities to be realist classes and properties. Level 2 entities can be classes and properties or concepts. Level 3 entities are terms, which are expressed as lexicographical labels for classes or concepts. We define a conceptual model as a set of Level 2 entities where each "represents" a Level 1 or Level 2 class or property [1]. Classes and properties that are also concepts can therefore represent themselves. We define logical models to be collections of classes and properties from either Level 1 or 2. Level 2 entities that are both classes or properties and concepts therefore exist both in the conceptual model (those assertions that treat them as concepts) and in the logical model (those assertions that treat them as classes or properties).

We seek to use conceptual models to achieve conceptual interoperability of data. Conceptual interoperability is the use of models of human understanding, or conceptual models, to provide interoperability commensurate with the level of alignment between conceptual models [13, 5, 14]. Two goals that we seek for conceptual interoperability are:

- Make similar but distinct data resources available for search, conversion, and inter-mapping in a way that mirrors human understanding of the data being searched.
- Make data resources that use cross-cutting models, such as HL7 v. 3 RIM [1] and provenance models (such as PML [15]), interoperable with domain-specific models without explicit mappings between them.

Resources such as the Gene Expression Omnibus (GEO) [16], ArrayExpress [17], and caArray [18] all contain separate logical models, but rely on related conceptual models, MAGE for ArrayExpress and caArray [19] and MIAME for GEO [20]. By encoding this model over each resource with a common vocabulary, it could then become possible to search across all resources using a single query, or easily convert data from one resource to another. Similarly, conceptual interoperability could enable the ability to search for patient history across domain-specific databases using queries that only talk about patient history, as we show in our Translational Research Provenance Vision [21] for biomedical experiments.

2.1 Relevant Ontologies and Frameworks

We leverage properties and classes from the BFO [10] and Information Artifact Ontology (IAO), [22] which are implementations of the scientific realist perspective on developing ontologies. We also leverage SKOS [23] as a basis for simple common vocabularies and associating conceptual models with those vocabularies. This work was based on practical issues surrounding mapping semantics from the cancer Biomedical Informatics Grid (caBIG) [24] into the semantic web. We have in the past worked on converting caBIG's layered semantics into OWL [25, 26] with success; however, the representation is limited to caBIG applications. Additionally, the mapping could not produce a one-to-one mapping between UML attributes and OWL properties, resulting in complex, unintuitive models.

3 Conceptual Interoperability Use Cases and Requirements

We divide the possible use cases of conceptual interoperability into three groups: search (or query), conversion, and direct mapping. Each of these use cases can be tailored to specific applications and additional requirements based on the level of interoperability needed. These use cases are necessarily abstract, and represent decompositions of uses cases such as testable hypothesis generation into component tasks.

Search: A user would like to perform queries with no knowledge of the underlying model. For example, "List the Education Level of all Persons in a dataset." or "Find me all Tissue Specimens from Persons with an Adverse

[1] Health Level 7 Version 3 Reference Information Model [8]

Event while taking Drug *x*." That "Drug *x*" is actually a class of drugs should not have to be a concern to the user.

Conversion: A user would like to convert instance [2] data from one logical model to another with a certain level of fidelity. This can be between domain models, or between a domain model and a cross-cutting model, such as a provenance model. For example, when events of *Clinical Service* occur with a given *Date*, dynamically create a record of *Vital Status* of *Alive* on that *Date*. These data are critical for tools like Kaplan-Meier survival curves [27], but availability of encounter data can be scattered across multiple organizations and systems that use different internal models.

Mapping: A user would like to create an automated mapping between two logical models. For example, take existing caBIG data models and align them with the BRIDG (Biomedical Research Integrated Domain Group) model [28]. This would occur when it is desirable for the Annotation and Image Markup [29] class *Person* to be automatically mapped as subclass of *bridg:Person*[3] because of their mutual relationship with *ncit:Person*[4].

We have identified a number of requirements for tools that would support these use cases:

Common Vocabulary: Conceptual models must use a common vocabulary that is distinct from any particular conceptual model. This is to allow portability of vocabularies between models, and prevent the reliance on one particular representation that might favor one logical model over another.

Distinction from Logical Models: A conceptual model and its vocabulary must not be represented in the same metamodel as a logical model. Doing so in metamodels that support reasoning may allow for direct inferences between conceptual and logical layers. This can have unintended consequences, for example, in cases where the logical and conceptual models are both expressed in OWL. If the logical model has classes that are subsumed by conceptual model classes, then it no longer becomes clear whether the instance is referring to an instance of a thing, or an instance of an idea of a thing.

Natural, Idiomatic Expression: A conceptual modeling framework must support natural, idiomatic expression of the actual data in its natural form. This means that there must never be any need to modify a logical model or its data in order to allow annotation of a conceptual model onto it.

Types, Properties, and Relations: A conceptual modeling framework must provide a way to express relationships between types, properties, and relations.

Additional Relationships: Most concepts have inter-relationships that can assist in improving conceptual interoperability. Any framework must provide a way of expressing these additional relationships.

These requirements come from previous experience with modeling layered semantics using OWL [25] where relating models with reference terminologies expressed in the same language (OWL 1) proved problematic.

4 The Conceptual Model Ontology

The Conceptual Model Ontology (CMO)[5] is a metamodel for representing conceptual models and their inter-relationships to logical models and vocabularies. Core to the CMO are these three classes:

cmo:Type. An abstract or general idea inferred or derived from specific in stances, representing a set of those instances.

cmo:Quality. The conceptual representation of anything that is a property (a thing that is inherent in an entity, like eye color) or an attribute (a thing that has been assigned, or attributed, to an entity, like name or identification number). *cmo:Quality* is the union of those two sets, so issues relating to determining if a quality is an attribute or property are not relevant here.

[2] Instances here and in the rest of the paper informally refer to OWL Individuals, in particular, Type 1 individuals in reality.

[3] BRIDG: Biomedical Research Integrated Domain Group. http://bridgmodel.org

[4] ncit: NCI Thesaurus. http://ncit.nci.nih.gov

[5] http://purl.org/twc/ontologies/cmo.owl

cmo:Relation. A concept representing the relationship between two independent entities.

Each of these classes are subclasses of *skos:Concept*, which is in turn is asserted in CMO to be a subclass of *iao:information content entity* [22]. These concepts are considered Level 2 entities from Smith *et al*. [2]. Concepts are tied to logical model entities through the *cmo:represents* property, a subproperty of *iao:is about*. Entities in logical models can either be concepts or Universals (Type 1 entities). *cmo:Universal* is a subclass of *bfo:independent_continuent* and *cmo:Fiat Entity* is a subclass of *bfo:generically_dependent_continuent*. Both *cmo:Universal* and *cmo:FiatEntity* have requisite classes, qualities, and relations, and are intended to be types that are punned on to OWL classes and properties in the OWL 2 metamodeling pattern [30]. The class hierarchy is displayed in Figure 1. Classes that are not universals are usually considered to be themselves concepts, and are metamodeled as *skos:Concepts*. These classes, since they are themselves concepts, in a very real sense represent themselves. However, it is impossible for a universal to represent itself, since universals are not considered to be concepts.

Subproperties of *skos:broadMatch* are provided to provide relationships between CMO concepts and common vocabularies. We provide *cmo:hasPrimaryConcept* and *cmo:has Qualifier* to allow for more nuanced composition, for example allowing "Tissue Specimen" to have a primary concept of "Specimen" and a qualifier of "Tissue".

We use SKOS as a basis for CMO because of its following properties. SKOS concepts unambiguously align to the definitions of concepts that we are using (as Level 2 entities), while OWL is ambiguous in its definitions of "class", it could either be considered a set or a concept. This is important, because we seek to draw a distinction between concepts as they exist in conceptual models, and the sets of things that they represent. Alternatively, remaining in OWL DL means that to use OWL classes as a common vocabulary would mean either creating instances of that class or punning that class to an instance. Punning the class means that the instance no longer has any semantics associated with it, and would need to either be given the type of the OWL class to regain semantics, or be given secondary semantics using an alternative structure. Here we do exactly that by giving the instance semantics using SKOS. Giving the class as a type of the instance in the conceptual model is also problematic, because it conflates being a thing of a type and being the idea of a type. The idea of a cat is not a cat, and when creating a conceptual metamodel that integrates with instance data, it is important to maintain that distinction.

cmo:Type relates to *cmo:Quality* through the use of *cmo:hasQuality* and its inverse, *cmo:qualityOf*. Qualities can have *cmo:values CanBe* assertions which provide the set of possible values for that quality. *cmo:Relation* has source (*cmo:hasSourceRole*) and target (*cmo:hasTargetRole*) types which help describe how those entities are related. Taken together, these qualities and relations form the structure of a conceptual model. The relations of CMO are outlined in Figure 2. By tying into existing common vocabularies, CMO-based concept models can be easily aligned along those vocabularies, as we will show below.

5 Implementation

The Conceptual Model Ontology is currently used as a backbone for "semantic web for the Biomedical Informatics Grid" (swBIG). This tool is currently available as a prototype RESTful service [31] that converts requests for resources from linked data URIs to caGrid service calls to requisite grid endpoints. This service uses a representation of NCI Thesaurus [32] converted to a SKOS representation using OWLtoSKOS. This representation addresses some, but not all of the concerns of Shulz *et al*. [33], and provides the ability to reason over concepts as instances in property value sets as well as in conceptual models. The retrieval operations are very simple and are documented on the swBIG web site. The source UML models are very closely mapped to preserve generalizations, attributes, and associations. Class and value typing on attributes and associations (using domain and range) and cardinality are preserved. When permissible values are listed for an attribute as part of the Common Data Element (CDE) [35] [36], an OWL ObjectProperty is created with a range of an enumeration class of the permitted

concepts (not strings). The concepts for classes, attributes, and properties as represented in CDEs are modeled using CMO. Instance data is generated using the model to determine and query associations and convert values to concepts when a permissible value mapping is used.

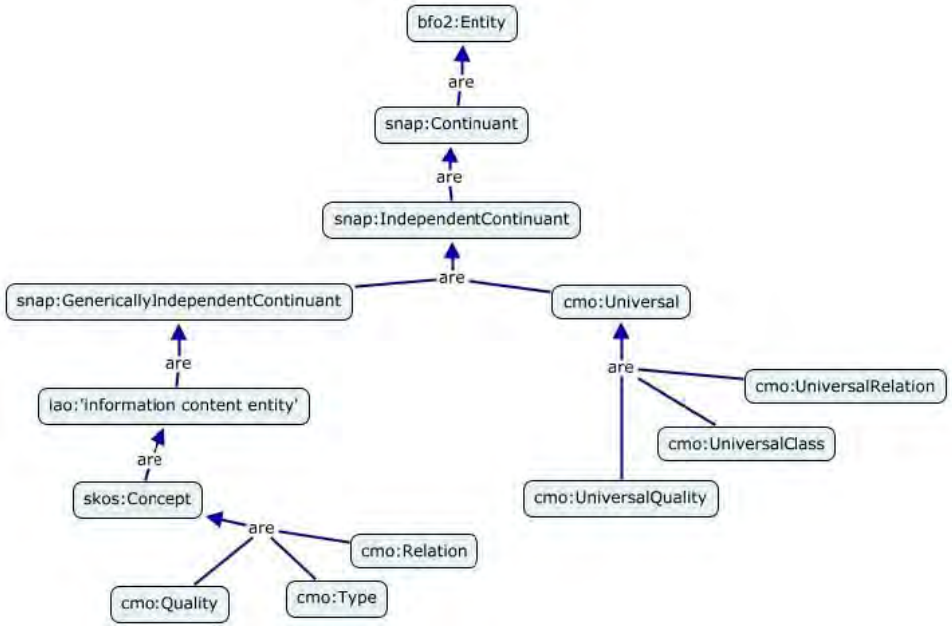

Figure 1. CMO Classes and their relationships with BFO, IAO, and SKOS. All diagrams are generated using CMap COE (http://coe.ihmc.us), and follow its labeling conventions.

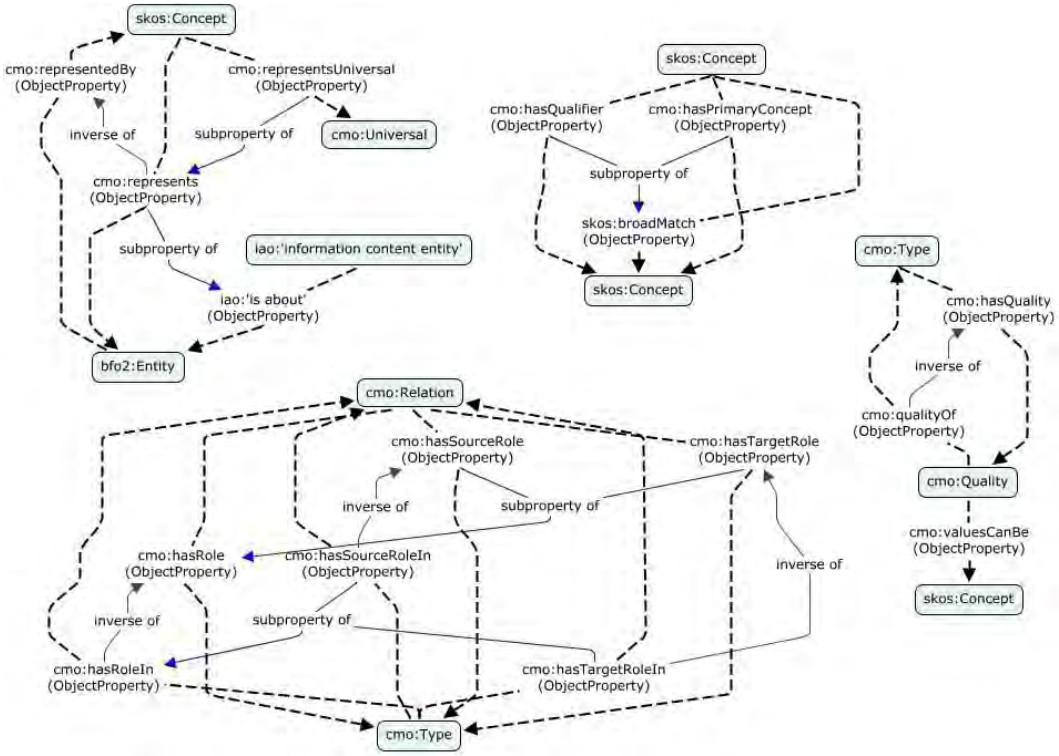

Figure 2. CMO properties and how they integrate with SKOS and IAO.

```
PREFIX cmo: <http://purl.org/twc/ontologies/cmo.owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#> PREFIX ncit:
<http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>

select count(distinct ?person) as ?count ?value
where {
  ?person a [cmo:representedBy [skos:broader ncit:Person]].
  ?person ?prop [rdfs:label ?value].
  ?prop cmo:representedBy [skos:broader ncit:Education_Level].
}
```

Figure3. Query 1: "Return the distribution of education level of all persons in the HINTS 2005 grid service." This query is performed by only using terms from a common vocabulary, NCI Thesaurus.

6 Evaluation

The Conceptual Model Ontology addresses all of the conceptual interoperability use cases and requirements. For example, the Query use case is satisfied with queries such as the one in Figure 3, which queries for the number of survey participants with a given level of education. The results of this query are displayed in Figure 4.

Conversion of data can be handled using rules such as the one in Figure 5. This example illustrates how CMO can be modified depending on the requirements of the task. The built-in semantics of CMO are kept minimal so that rules based on it can be tailored to the needs of the task. Some applications may require very strict conceptual alignment, while other applications may require a looser coupling in order to meet requirements. Models can also be mapped directly onto each other as shown in Figure 6.

CMO also satisfies conceptual interoperability requirements. Common vocabularies are distinct from the conceptual and logical models. Existing ontologies in OWL can be annotated without modification or change to existing semantics. While CMO is used to express semantics from caBIG, it is not limited to caBIG models. CMO provides a simple way to express relationships between types, properties, and relations. Finally, because it uses SKOS-based common vocabularies, CMO allows additional relationships to be asserted between those concepts. For example, it is possible to assert that birds can fly at the conceptual level with direct assertions that have no automatic inference. Performing this using concepts means that this can be compared against instances without triggering consistency exceptions, such as the case with flightless or injured birds.

7 Future Work

We are currently investigating the use of CMO models to provide automated mappings of caBIG data elements into the BRIDG clinical model. This effort has seen some initial success, and work continues. Additionally, we will explore the use of CMO to represent domain-specific models in relation to a common model of provenance as envisioned in McCusker and McGuinness [21] including conceptual representations of biomedical experiments. We also are exploring the use of a common vocabulary to provide a unified view of existing provenance models and domain models in terms of provenance. We hope to do this with the Translational Medicine Ontology [37]. We plan to provide satisfaction of additional use cases as well.

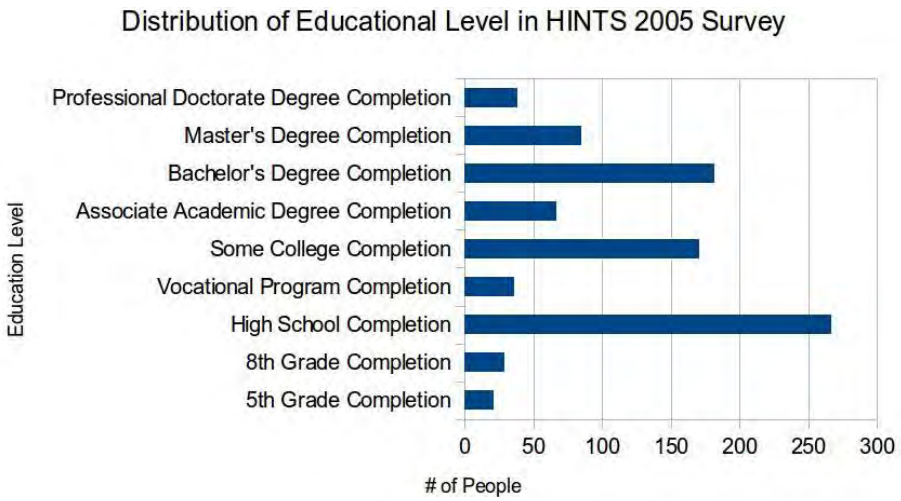

Figure 4. Distribution of educational level in the HINTS 2005 survey. These data were gathered using the query in Figure 3.

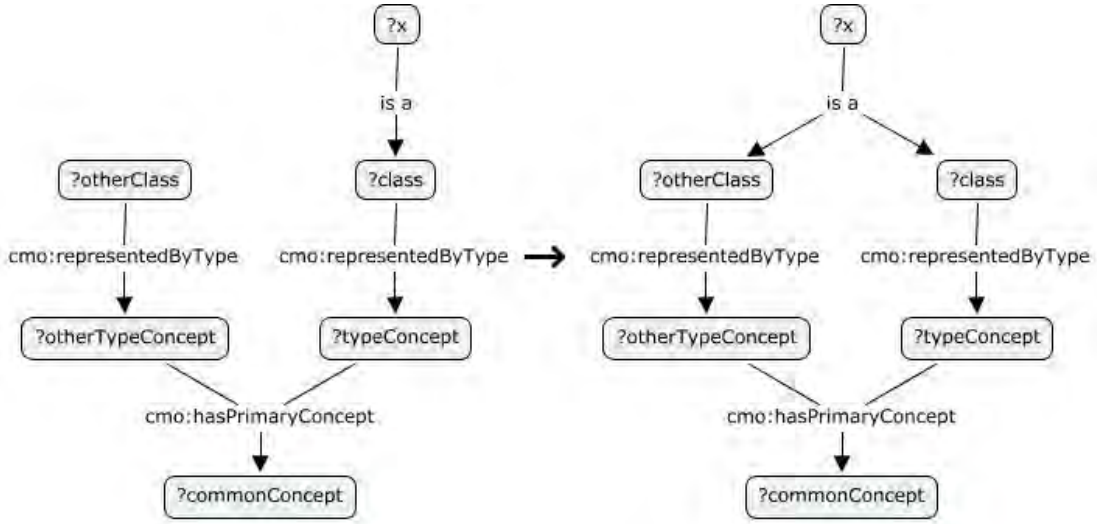

Figure 5. Mapping data from one logical model to another. By identifying that a "parent" class hangs directly off of a broader term of a "child" class, an instance of the "child" class can be given the type of the "parent".

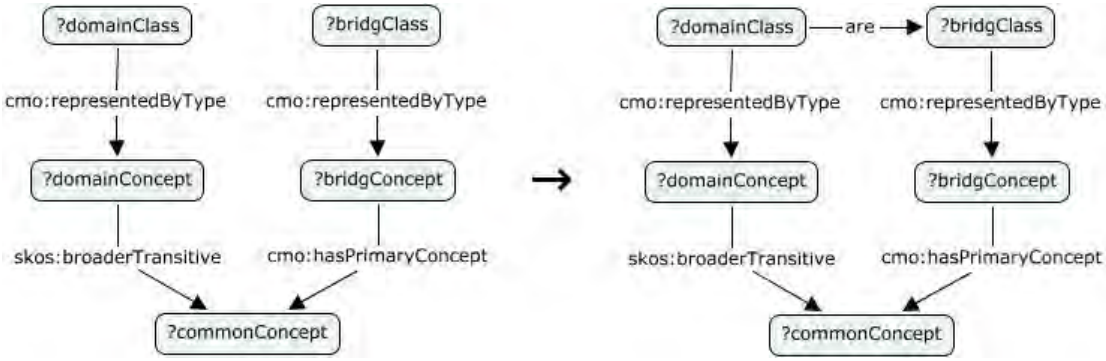

Figure 6. Mapping logical models directly on to each other can be accomplished by discovering relative relationships of the classes within the common vocabulary. The left hand side of the figure shows the precondition for mapping one class to another. The right side is the final state, where the added "are" arc represents the assertion that *?domainClass* is now a subclass of *?bridgClass*.

Finally, CMO does not yet provide a way to map between different levels of granularity. One model may represent a relationship as a direct link, while another may provide an intervening class which provides more information. It would be useful for CMO to include a property for how these levels relate.

8 Conclusion

Conceptual models can play a significant role in automated semantic interoperability, because they can allow the integration of data from across logical models without the need for direct integration of logical models. The Conceptual Model Ontology can support important uses cases in conceptual interoperability and is being used to represent existing semantics from a large software development program (caBIG). CMO is currently available for use with instance data using the swBIG linked data proxy. Finally, CMO is not limited to caBIG models, but can be applied to any logical model expressed in OWL.

References

1. Ceusters, W., Smith, B.: Foundations for a realist ontology of mental disease. Journal of Biomedical Semantics **1**(1) (2010) 10
2. Smith, B., Kusnierczyk, W., Schober, D., Ceusters, W.: Towards a reference terminology for ontology research and development in the biomedical domain. In: Proceedings of KR-MED. Volume 2006., Citeseer (2006) 57–65
3. Melnik, S., Decker, S.: A layered approach to information modeling and interoperability on the web. In: Proc. of the ECDL'00 Workshop on the Semantic Web. (2000)
4. Luciano, J., Stevens, R.: e-Science and biological pathway semantics. BMC bioinformatics **8** (Suppl 3) (2007) S3
5. Tolk, A.: What comes after the semantic web-pads implications for the dynamic web. In: Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation, IEEE Computer Society (2006) 55
6. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., et al.: Gene ontology: tool for the unification of biology. Nature genetics **25**(1) (2000) 25
7. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11) (2007) 1251–1255
8. Schadow, G., Russler, D., Mead, C., McDonald, C.: Integrating medical information and knowledge in the HL7 RIM. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (2000) 764
9. Kalra, D., Beale, T., Heard, S.: The openehr foundation. Studies in Health Technology and Informatics **115** (2005) 153–173
10. Smith, B.: Beyond concepts: ontology as reality representation. In: Formal Ontology In Information Systems: Proceedings of the Third International Conference (FOIS-2004), IOS Press (2004) 73–84
11. Smith, B., Ceusters, W.: Ontological realism as a methodology for coordinated evolution of scientific ontologies. Applied Ontology **5** (2010) 139–188
12. Grenon, P., Smith, B.: SNAP and SPAN: Towards dynamic spatial ontology. Spatial Cognition & Computation **4**(1) (2004) 69–104
13. Tolk, A., Muguira, J.: The levels of conceptual interoperability model. In: Proceedings of the 2003 Fall Simulation Interoperability Workshop, Citeseer (2003) 007
14. Dobrev, P., Kalaydjiev, O., Angelova, G.: From Conceptual Structures to Semantic Interoperability of Content. Conceptual Structures: Knowledge Architectures for Smart Applications (2007) 192–205
15. McGuinness, D., Ding, L., Pinheiro da Silva, P., Chang, C.: Pml 2: A modular explanation interlingua. In: Proceedings of AAAI. Volume 7. (2007)
16. Edgar, R., Domrachev, M., Lash, A.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research **30**(1) (2002) 207
17. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G., et al.: ArrayExpressa public repository for microarray gene expression data at the EBI. Nucleic Acids Research **31**(1) (2003) 68
18. Bian, X., Klemm, J., Basu, A., Hadfield, J., Srinivasa, R., Parnell, T., Miller, S., Mason, W., Kokotov, D., Duncan, M., et al.: Data submission and curation for caArray, a standard based microarray data repository system. (2009)
19. Spellman, P., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bern-hart, D., Sherlock, G., Ball, C., Lepage, M., et al.: Design and implementation of microarray gene expression markup language (MAGE-ML).

Genome biology 3(9) (2002)

20. Edgar, R., Barrett, T.: Ncbi geo standards and services for microarray data. Nature biotechnology **24**(12) (2006) 1471
21. McCusker, J.P., McGuinness, D.L.: Explorations into the Provenance of High Throughput Biomedical Experiments. Provenance and Annotation of Data and Processes (2010) 120–128
22. Ruttenberg, A., Smith, B., Ceusters, W.: The Information Artifact Ontology (2008)
23. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference. (2008)
24. Von Eschenbach, A., Buetow, K.: Cancer Informatics Vision: caBIG. Cancer informatics **2** (2006) 22
25. McCusker, J.P., Phillips, J., Beltrán, A., Finkelstein, A., Krauthammer, M.: Semantic web data warehousing for caGrid. BMC bioinformatics **10**(Suppl 10) (2009) S2
26. Gonzalez-Beltran, A.: Ontology-based Queries over Cancer Data. (2010)
27. Efron, B.: Logistic regression, survival analysis, and the Kaplan-Meier curve. Journal of the American Statistical Association **83**(402) (1988) 414–425
28. Fridsma, D., Evans, J., Hastak, S., Mead, C.: The BRIDG project: a technical report. Journal of the American Medical Informatics Association **15**(2) (2008) 130–137
29. Rubin, D., Mongkolwat, P., Kleper, V., Supekar, K., Channin, D.: Annotation and image markup: Accessing and interoperating with the semantic content in medical imaging. Intelligent Systems, IEEE **24**(1) (2009) 57–65
30. Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4) (2008) 309–322
31. Richardson, L., Ruby, S.: RESTful web services. O'Reilly Media, Inc. (2007)
32. Sioutos, N., Coronado, S., Haber, M., Hartel, F., Shaiu, W., Wright, L.: NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics **40**(1) (2007) 30–43
33. Schulz, S., Schober, D., Tudose, I., Stenzhorn, H.: The Pitfalls of Thesaurus Ontologization–the Case of the NCI Thesaurus. In: AMIA Annual Symposium Proceedings. Volume 2010., American Medical Informatics Association (2010) 727
34. Hesse, B., Moser, R., Rutten, L., Kreps, G.: The health information national trends survey: research from the baseline. Journal of Health Communication **11** (2006) 7–16
35. Warzel, D., Andonyadis, C., McCurry, B., Chilukuri, R., Ishmukhamedov, S., Covitz, P.: Common data element (CDE) management and deployment in clinical trials, American Medical Informatics Association (2003)
36. Kunz, I., Lin, M., Frey, L.: Metadata mapping and reuse in caBIG. BMC bioinformatics **10**(Suppl 2) (2009) S4
37. Dumontier, M., Andersson, B., Batchelor, C., Denney, C., Domarew, C.: The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. In: Proceedings of the 13th ISMB2010 SIG Meeting" Bio-Ontologies. (2010) 120–123