Reference Ontologies for Biomedical Ontology Integration and Natural Language Processing

Jonathan Simon[1], James Fielding[2], Mariana Dos Santos, MD[2], Barry Smith, PhD. [1,3]

[1]IFOMIS, Leipzig, Germany
[2]Language and Computing nv., Zonnegem, Belgium
[3]Department of Philosophy, University at Buffalo, Buffalo, U.S.A.

To be presented at EuroMISE 2004, Prague, April 12-15

The central hypothesis of the collaboration between Language and Computing (L&C) and the Institute for Formal Ontology and Medical Information Science (IFOMIS) is that the methodology and conceptual rigor of a philosophically inspired formal ontology greatly benefits application ontologies.[1] To this end LinKBase®, L&C's ontology, which is designed to integrate and reason across various external databases simultaneously, has been submitted to the conceptual demands of IFOMIS's Basic Formal Ontology (BFO).[2] With this project we aim to move beyond the level of controlled vocabularies to yield an ontology with the ability to support reasoning applications. Our general procedure has been the implementation of a meta-ontological definition space in which the definitions of all the concepts and relations in LinKBase® are standardized in a framework of first-order logic. In this paper we describe how this standardization has already led to an improvement in the LinKBase® structure that allows for a greater degree of internal coherence than ever before possible. We then show the use of this philosophical standardization for the purpose of mapping external databases to one another, using LinKBase® as translation hub, with a greater degree of success than possible hitherto. We demonstrate how this offers a genuine advance over other application ontologies that have not submitted themselves to the demands of philosophical scrutiny.

LinKBase® is one of the world's largest applications-oriented medical domain ontologies, and BFO is one of the world's first philosophically driven reference ontologies. The collaboration of the two thus initiates a new phase in the quest to solve the so-called "Tower of Babel" problem of ontology integration, and this paper reports its initial results.[3]

Introduction: L&C's LinKBase® and Basic Formal Ontology

For millennia, when we have encountered difficulties understanding reality, we have turned to philosophers for solutions. Why should we not do likewise

today? The return to a realist philosophy means a return to those foundations that reflect 2000 years of ontological research, but this in no way requires that we abandon our pragmatic perspective. In his *Physics*, Aristotle writes, "When the objects of an inquiry, in any department, have principles, conditions, or elements, it is through acquaintance with these that knowledge, that is to say *scientific knowledge*, is attained," and we would do well to keep such words in mind today when we seek to design an adequate ontological inventory of those basic elements that belong to the structure of reality.

LinKBase® is a biomedical domain ontology that has been designed to integrate terminologies and databases with applications designed for natural language processing and information retrieval. The ontology contains 543 different relation types (links), reflecting often subtle semantic differences. They are divided into different groups, including spatial, temporal and process-related link types. LinKBase® currently contains over 2,000,000 medical concepts with over 5,300,000 link type instantiations. Both concepts and links are language independent, but they are cross-referenced to about 3,000,000 terms in various languages. LinKBase® provides a central hub with fixed structured definitions into which external medical terminologies and databases, such as Swiss-Prot, SNOMED, and the Gene Ontology (GO), may be embedded. This task turns out to be complex endeavor, not least because the different terminologies or databases that are to be integrated are often internally and mutually inconsistent. Yet, as all these terminologies must essentially speak about the same reality, there is a common thread that runs through them and the LinKBase® methodology is based on the idea that it is possible to integrate them on the basis of a sound understanding of those basic categorical distinctions that are common to them all.

Basic Formal Ontology is a philosophically inspired top-level ontology which provides a coherent, unified understanding of these basic ontological distinctions and which is currently being implemented as a top-level open source backbone ontology for LinKBase®. BFO will provide a framework for mapping external ontologies, terminologies, and databases onto LinKBase® in a way that is designed to provide for successful integration, as well as to provide a useful guide for the future algorithm development that will allow for cross-ontology navigation.

Methods: Standardization

As ontologies and terminologies expand and are integrated together, it is natural that semantic consistency will become increasingly difficult to maintain. The cause of this difficulty is typically the ambiguities and inconsistencies that result from the lack of a standard unified framework for understanding those basic relations that structure our reality. The BFO formal ontology provides application ontologies with a set of standardized, first-order definitions for these ontological elements, definitions which can be exploited by reasoning applications, including applications designed for natural

language understanding. By disambiguating the ontological structures underlying informal definitions of insufficient precision, these formalizations can aid in the passage of domain knowledge between users and software agents, and thus improve coherence and adaptability in and between ontologies.

The resultant standardization reflects an implementation of philosophical rigor along two dimensions. First, it establishes internal consistency on the basis of precise analyses of the concepts involved. Ontologies such as LinKBase® (as well as SNOMED and GO) are viewed as object languages with a certain "surface structure." They consist of systems of concepts joined together in binary relations such as is-a and part-of. For the most part however, these relations and concepts are given only in natural language and their grammatical form leads to various ambiguities. Thus, the project of defining a unique "deep structure" to which every such concept, relation, and axiom can be mapped requires sound conceptual analysis. The standardization effort gives us a methodology with which to identify and repair internal inconsistencies and ambiguities in LinKBase®.

The second dimension of rigor requires the use of the standard first-order logical language in which also the concepts of BFO are defined and axiomatized. In this way the rigor of the BFO classification system is imported into an ontology from the outside. This importation is meta-ontological, in the sense that changes are not made directly within the external ontology itself; rather, their place in the BFO re-articulated domain ontology, in this case LinKBase®, is marked via an external mapping algorithm in a way that provides the degree of consistency required to navigate between different third-party ontologies such as GO and SNOMED.

The analysis runs as follows:

1. For every concept C, the definition consists in a mapping to a pair: < the class named by C, the extension of the class named by C >

2. For every relation R(X,Y), the definition consists in a mapping to a logical formula of the following form: For all x such that x is in the extension of the class named by 'X', there is a y such that y is an element in the extension of the class named by 'Y', and R*(x,y). (where R* is a relation in the formal language of BFO, for example part-of)

Axioms, which are essentially instantiated relations, are defined by a mapping similar to the definition of relation presented above, differing only in that the variables are replaced by specific concepts within the ontology.

In the remainder of this essay we accomplish two goals. We first witness ways in which the philosophical insights afforded by this standardization has allowed us to disambiguate the LinKBase® ontology itself. Second, we discuss the way that BFO standardization has assisted in the ontology

integration effort, by adding structural information according to the BFO standardization methodology and the development of the MaDBoKS software.

Results: Part One

Objects and Processes in LinKBase®

In philosophical circles it is well understood that the universe of common sense contains two types of entities that relate differently to time. There are on the one hand objects: tables, chairs, countries, and people. These entities are said to *endure* through time, which means that they do not have temporal parts, but rather are wholly present at every moment in which they exist. On the other hand are processes like brain surgeries, heart attacks, lives. These are said to *perdure* through time, which means that they do have temporal parts, such as the first half of the surgery, the last phase of the heart attack, one's childhood. This distinction is not adequately made in existing applications ontologies and taxonomies. In particular when the ultimate tribunal for those ontologies are natural language practices, it becomes very important to identify the ambiguity in terms like 'injury', 'dilation', and 'dislocation'. For these seeming concepts are each in fact two distinct concepts. We speak both of an injury as a perdurant ("when did that injury occur?") and as an endurant ("That injury looks terrible"). Likewise with kinds of injuries, like dislocations: "The dislocation of his shoulder occurred yesterday" vs. "The doctor reduced the dislocation." Indeed, in the medical domain it is commonplace for a sort of process and the state resulting from that process to share a name:

"Dilation" may stand for the process of dilation, i.e. of becoming broader: "Once in place, a small balloon tip is inflated for a few seconds to *dilate* the artery." Or, it may stand for the dilated, broadened structure: "*Dilation* of the posterior mitral ring was corrected."

Here the philosophical distinction between endurants and perdurants allows us to maintain the separation of concepts which would otherwise be, and often are, conflated. By implementing this distinction into the LinKBase® top level, we have been able to recognize these instances of homonymy when they appear. We thereby avoid modeling errors that emerge in contexts where relationships between links and concepts are inferred.

Absences in LinKBase®

It is a tenet of contemporary philosophy that absences are not entities, but the lack of entities. Yet LinKBase® must represent natural medical language concepts like "absence of bacteriuria (bacteria in the urine)", and "sputum without blood". Further, while less common, medical texts may feature reference to absences without a specified location of absence, because the location is determined by context

The straighforward approach, and the approach that LinKBase® formerly used, violated the philosophical tenet mentioned above, and construed these absences as special kinds of entities, namely processes of absence. With this approach, it was necessary to specify more about the processes in question. What kind of process is an absence? What is its duration? Who are its participants? How do we know when two descriptions of absences actually refer to the same entity?

Processes are perdurants, entities located in spacetime. They thus have boundaries, volumes, and locations ("the surgery took place in the operating room"). An adequate inference engine will know various things about bounded objects: it will know, for example, that if the boundary of object x is different from the boundary of object y, then x cannot be the same object as y. Now it is clear that in a natural language data extraction application, information about the boundary of an absence would be specified via a description like "an absence *in the liver*."

Philosophical scrutiny (one of whose functions is to test adaptability by demanding responses to creative counterexamples) tells us that the treatment of absences as processes is unstable, in that a reasoning engine attempting to handle and infer information about absences so construed runs the risk of deriving contradictions. This possibility arises when we wish to recognize the identities of differently described absences. 'The book was absent from my apartment' and 'The book was absent from my bedroom' seem to refer to the same absence. However, as soon as we instruct our inference engine to consider the two absences here described as identical, we will encounter inconsistency. For the system will record both that the absence has as boundary: my apartment, and that it has as boundary: my room. But this is a contradiction, since x=y implies boundary_of(x) = boundary_of(y). Thus if a treatment of absence concepts and relations in LinKBase® is to be perfectly general, and is not to rely on every absence concept coming with its own preset location, then we cannot construe absences as processes. So how do we treat them?

Another tenet of philosophy is: distinguish the particular from the universal. When we say "There is an absence of bacteria in the patient's urine" we clearly are not saying *of* the bacteria in the urine, that *it* is not there. Rather, we are saying *of the universal*: bacteria in the urine, that *it* has no *instances* in the patient's urine. Following the intuition here, the current modeling eliminates concepts of absence themselves. Rather, relations of absence (like the absence of bacteria in the urine) are construed as links between the relevant bacteria concept, and the urine concept, but here it is the universal of the former that is involved: "if x is the bacteria universal, y is an instance of urine, then x has no instance located in y." This technique allows us to make inferences very naturally that would be artificial and error prone with the absences-as-entities model. We no longer need to answer the question of whether the absence of the book from my apartment is the same absence as that of the book from my room. We may naturally infer that there is an

absence of the book from my room, given that there is an absence of the book from my apartment. This follows from our general knowledge of location and parthood.

Along with improving our reasoning power, this solution improves our representation structure, rendering applications involving absences more elegant and simple. The old representation of absences as processes blocked us from directly linking two entities where one entity is "absent in" the other entity, but rather a third concept had to be created, the process of "absence of entity" which related both entities. Thus, to represent the concept "sputum without blood" the concept "absence of blood" had to be created to be related to "blood" (the absent entity) and to "sputum without blood" (the location of the absence).

By representing absence as a relation between the "absent entity" and the "entity from which the related entity is absent" we avoid creating a third unneeded concept, and reduce the distance between the related concepts to one relation instead of two. (E.g. the concept "sputum without blood" can be represented with a direct link to the concept "blood", which will be interpreted in formal language as: "The blood universal has no instance located in (the patient's) sputum"). The distance between concepts, and between links, on parent child trees, is relevant to many LinKBase® applications [4]

Results: Part Two

SNOMED and the "Parthood" Relation

Identically named concepts and relations often have very different denotations. The degree of internal consistency required to apply the BFO standardization accurately to an ontology requires that these terms be disambiguated. One common variety of disagreement within a taxonomic system centers on divergent uses of the relation "parthood." In SNOMED, for example, the concept "amputation of toe" is a special case of the concept "amputation of foot."[5] But while the toe certainly is a part of the foot, the amputation of the toe certainly is not an amputation of the foot. The former ought to be represented either as a *part of* an amputation of the foot, or alternatively, as an amputation of *part of* the foot. Depending on the context, these are two very different sorts of things.[6]

SNOMED here runs together endurants and occurrents. It runs together that element of parthood associated with the foot, an entity that endures in time, with that parthood associated with an amputation, an event that occurs in time. It is for reasons such as these that these two dimensions of parthood must be kept apart.

Objects and Processes within GO

GO is divided into three disjoint hierarchies: the *cellular component*, *biological processes*, and *molecular function* ontologies.[7] The first, equivalent to that of anatomy in the medical domain, is an ontology of

endurants. It allows users to access the physical structure with which a gene or gene product is associated. A biological process, on the other hand, is defined in GO as "a phenomenon marked by changes that lead to a particular result, mediated by one or more gene products." This ontology is therefore a hierarchy of occurrents.

There are however some confusions over the role of the molecular function hierarchy. While GO defines molecular function as "the action characteristic of a gene product," it is clear that functions do not occur, but rather endure; the function of a gene or gene product exists identically for as long as its bearer exists and is present at all times, even if that function is never realized. Even mutant genes retain their function. Thus for example, "signal transducer activity" remains the function of the EPO_HUMAN protein even though the latter is incapable of performing the signal transduction process.

Molecular functions and biological processes are obviously closely related. The function "signal transducer activity" certainly *involves* performing "signal transduction" in some sense; yet in GO this relationship is undefined. The authors of GO have attempted to clarify this relationship, stating, "a biological process is accomplished via one or more ordered assemblies of molecular functions," in order to suggest that the relation is one of agency. Here, functions *initiate* biological processes, but this would suggest that they share in a relation of parthood, which GO on the other hand explicitly rules out. For GO's authors insist, correctly in our view, that parthood only holds between entities of the same hierarchy. So long as the associated relations continue to conflate the distinct categories of function and process within the ontology, however, GO's architecture will continue to constrain the sorts of reasoning systems which it can support.

Mapping Ontological Elements: Applying External Consistency

The Mapping Databases onto Knowledge Systems tool (or MaDBoKS) is an extension of the LinkFactory® ontology management system that administers and generates mappings from external databases onto LinKBase®.[8] This mapping mediates the data contained in the external database in a manner that expands the hub ontology, leaving the structure of the foreign ontology untouched. The MaDBoKS system is designed in such a way that all implicit and explicit relationships between data from the different databases are mapped to the ontology. Administration of the mapping mediates the data contained in the different databases in such a way that it is associated with ontological information and the ontology is thereby virtually expanded with data and relations. The mapping tool can map column data as well as cell record data in such a way as to carry relationships over into the ontology. The MaDBoKS system meets the requirement that the ontology does not change upon coupling or decoupling of the databases. In this manner the ontology management system, LinkFactory®, is able to navigate across problematic

definitions and relations within an external database using the BFO standardization as translation mechanism.

Mapping SNOMED to LinKBase®

LinKBase® understands not only the notion of "part", but also "proper part", "part-of", "part-for" and "has-part". These refinements allow us to build an accurate representation in which various distinctions in the conception of "amputation of foot" discussed earlier are recognized as distinct and their relation to each other can be mapped. The distinctions rest on the formal notion of parthood, along with an understanding of the interplay of classes and their instances crucial to the modeling of this formal notion and its relatives. Class X is part-of Class Y whenever every instance of Class X is a *part for* some instance of Class Y. Class Y *has part* Class X whenever every element of Class Y has some element of Class X as part. Class X is *part of* Class Y whenever Class X is part for Class Y, and Class Y has part Class X. The further distinction between parts and proper parts lies on the instance level: individual x is a proper part of individual y whenever x is a part of y, but x is not identical to y. Where the toe is both a part *and* a proper part of the foot, the foot is a part, and not a proper part, of itself. In LinKBase®, these distinct parthood relations are captured, with part-of as the root relation. Further, LinKBase® contains a concept named "structure," designed to be relativised to embed information about parthood in the concept space, as well as in the relation space. If X is a class, then there is a concept "X structures" which is such that it subsumes all and only those classes that stand in the part of relation to X. For example, both the toe and the foot itself are subsumed by the concept "foot structures."

This configuration is then mapped to the SNOMED ontology, where "amputation of foot" is related to the concept "foot structure" (any part of the foot including the foot itself), which subsumes two further concepts "complete amputation of foot" and "partial amputation of foot" (related to the concept "proper part of foot"). In this way we maintain a hierarchical structure that subsumes both the toe and the foot without reducing either one to the other, thus allowing each to be related to different, and possibly incommensurable concepts without the problematic inconsistencies derived through inherited criteria.

Mapping GO to LinKBase®

During the conceptual analysis phase, we carefully investigated the top-layer concepts of the three GO sub-domains that act as our gateway between the LinKBase® concepts and GO terms. We identified the more general concepts of GO in LinKBase® and created new concepts in those cases where suitable equivalents were not already recognized. In this way we are able to relate GO's molecular function hierarchy to the two other GO hierarchies by

integrating all three simultaneously into the expansion of BFO motivated by the formal-ontologically extended top level.

If we return to the EPO_HUMAN protein example from earlier, we see now that LinKBase® is able to appropriate this example and model the relations with a greater degree of clarity, essentially mirroring the BFO defined structure. The connection between a GO protein and its activity in LinKBase® is captured by a "has-function" relation, and the connection between an activity and its corresponding processes is captured by the LinKBase® "realization" relation. The former reflects the relation between a substance and its function, and the latter, that between a function and its actualization. Clearly, this latter relation is skew to the whole/part relation, which is properly left exclusive to each hierarchy.

In this manner not only is GO consistently mapped to LinKBase®, but the expressiveness of GO itself has been expanded without any major alterations required in its core structure.

Discussion

Our LinKBase® ontology is a representation of the medical domain. By mapping more specialized information sources like GO and protein databases, we were able very quickly to expand the reach of our ontology and hence achieve a database warehousing system within which all mapped databases stand automatically in the right sort of relation to each other in such a way that a global view of the dispersed information is made possible. The MaDBokS system can be used to graft databases onto the ontology and thereby make the latter useable for a variety of applications. The flexibility of the MaDBoKS system and the speed with which databases can be integrated allows the prototyping of different integration protocols in relation to different sets of databases and hence enables a fine-tuning of the integration process for specific applications such as data-mining and information extraction.

The BFO-driven restructuring of LinKBase® is still in its infancy, yet we already have examples demonstrating increased adaptability through the application of philosophical knowledge and techniques. We have demonstrated examples in which changes were made leading to an enhanced internal consistency, allowing the level of access necessary for a general database translation hub.

If early successes (like the integration of GO into a MaDBoKS extention of LinKBase®) are any indicator, we have great reason to expect that the thoroughgoing integration of BFO and LinKBase®, of which the above results are merely preliminary groundwork, will greatly enhance the capacity of LinKBase® to effect direct integration between foreign ontologies such as SNOMED and GO. For the results cited here are not isolated instances but rather illustrations of a general pattern. There are reasons for the ad hoc features of many biomedical ontologies, the main cause of the so-called

"Tower of Babel" problem of interoperability. These features have developed because ontologists and terminologists were forced, in moving from printed dictionaries and nomenclatures to digital systems, to make a series of uninformed decisions about complex ontological issues, indeed about the very same issues that philosophers have been pondering for millennia. To date, the importance of philosophical scrutiny in software application ontologies has often been obscured by the temptation to seek immediate solutions to localized problems. In this way the forest is lost for the trees, and larger integration problems are rendered unsolvable. Ad hoc solutions foster further ad hoc problems.

It is thus a tangled web we weave when we seek to create application ontologies without a basis in philosophically sound formal theories. The philosophically sound formalism of LinKBase® enables it to support the integration (and thereby, the untangling) of data from different external data sources in a transparent way, capturing the exact intended semantics of the database terms, and filtering out erroneous synonyms.

Acknowledgements

We are grateful for the helpful comments of Dirk Siebert, Werner Ceusters, and Jean-Luc Verschelde. This work has been supported by the Language and Computing Research Division and the Wolfgang Paul Program of the Alexander von Humboldt Foundation.

Further References

1] Flett A, Dos Santos M, Ceusters W.: Some Ontology Engineering Procedures and their Supporting Technologies. EKAW2002 (2003)
[2] Smith B.: Basic Formal Ontology. http://ontology.buffalo.edu/bfo (2002)
[3] Montayne F, Flanagan J.: Formal Ontology: The Foundation for Natural Language Processing http://www.landcglobal.com (2003)

[4] Van Geyt L, Martens P, Terzic B, Flanagan J. : Get More Out of Your Unstructured Medical Documents. http://www.landcglobal.com (2002)

[5] SNOMED (Systematized Nomenclature for Medicine) http://www.snomed.org

[6] Smith, B. : Mereotopology : A Theory of Parts and Boundaries. Data & Knowledge Engineering (1996) 20 :287-303

[7] GO (Gene Ontology General Documentation) http://www.geneontology.org/doc/GO.doc.html

[8] Verschelde J.L., Dos Santos M, Deray T, Smith B, Ceusters W. : Ontology-assisted Database Integration to Support Natural Language Processing and Biomedical Data-mining. Journal of Integrative Bioinformatics, forthcoming.