Formal Ontology for Biomedical Knowledge Systems Integration

James M. Fielding[a], Jonathan Simon[b], Barry Smith, PhD.[bc]

[a]*Language and Computing nv., Zonnegem, Belgium and Philadelphia, PA*
[b]*Institute for Formal Ontology and Medical Information Science, Leipzig, Germany*
[c]*Department of Philosophy, University at Buffalo, NY*

Abstract:

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 will greatly benefit software application ontologies. 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). With this, 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 mapping external databases with a greater degree of coherence than hither. We then show how this offers a genuine advance over other application ontologies that have not submitted themselves to the demands of philosophical scrutiny.

Keywords:

Biomedical Systems Integration, Formal Ontology, Ontology Development, Biomedical Data-Mining, Natural Language Processing, Concept Formation

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 retreival. 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.[1] 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® methodlogy 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[2] 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[3].

Ontological Distinctions

We begin by reviewing a small number of the fundamental ontological distinctions that form the basis of our methodology. These distinctions will serve as examples in the case studies cited below.

Universals vs. Particulars

As realist philosophers in the Aristotelian tradition we distinguish between universals (also called classes, kinds, species, or types) and particulars (individuals, instances, or tokens). An example of a universal would be the species "Malaria" that a doctor studies in medical school, or the general function "to boost insulin production." An example of a particular would be *this* malaria present in *this* blood sample, or the function of *this* gene to boost insulin production in *these* beta cells in your pancreas.

Endurants vs. Occurrents

Among both universals and particulars, we can further distinguish between what are called endurants and occurrents. These two sorts of entities relate differently to time. Endurants are those entities which *endure* through time and are wholly present at each moment of their existence. Examples of endurants are people, cells, and chromosomes. All of these kinds of entities, and all of their parts, maintain their identity from one moment to the next, even while undergoing familiar sorts of changes.

Occurrents, on the other hand, are those sorts of entities that are never fully present at any one given moment in time, but instead *unfold* themselves in successive phases, or temporal parts. Examples of occurrents are processes, activities, events, such as a morning run, a court session, or cellularization.

In a parallel fashion, where your arm is a part of you, and your hand is a part of your arm, your youth is a part of the process which is your life, and your first day at school is a part of your youth. But it is important to note here that parthood never crosses these boundaries – parts of endurants are always endurants and parts of occurrents always themselves occur.

Dependent vs. Independent

Some entities have the ability to exist without the ontological support of other entities. These are entities such as people, cells, or molecules. These sorts of entities we call independent. On the other hand, there are entities that require the existence of entities of the first sort for their own existence: a morning run needs a runner, a viral infection is dependent on the virus and on the organism infected.

All occurrent entities require an independent entity upon which to inhere; in other words, there is no process without a substance, but within the category of endurant entities there are both dependent and independent entities: thus the function of an organ depends on the existence of the organ itself.

General Procedures

In the remainder of this article, we first describe our general program of standardization, and what we have achieved so far. Following this, we discuss a small selection of cases where the BFO structure has illuminated inconsistencies in third-party ontologies so that they may be coherently modelled in the LinKBase® ontology and brought to greater clarity and perspicuity. Drawing examples first from SNOMED, GO, and LinKBase® itself, we describe how this structure has already aided in our external mapping of the SNOMED and GO ontologies, and how this has introduced a greater level of consistency and expressiveness to these ontologies.

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, defintiions 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 SNOMED and GO are viewed as an object language 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 philosophically driven formal ontology approach provides for this.[4]

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.

The analysis runs as follows:

1. For every concept C, the definition consists in a mapping to a pair: < the universal named by C, the extension of the universal 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 the universal named by X or x is in the extension of that universal, there is a y such that y is the universal named by Y or y is an element in the extension of that universal, and R*(x,y). (where R* is a relation in the formal language of BFO, for example part-of)

3. 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.

Isolating Problems of Internal Consistency

The intent is not to remodel SNOMED or GO. Rather it is to *integrate* these varying terminologies on the basis of the fact that such integration requires a certain degree of consistency. By adding structural information in a way that removes inconsistencies, we have been able to map these databases to the LinKBase® ontology. This will be discussed further below.

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. 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.[7] 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,"[8] 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®. 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.[9]

Mapping SNOMED

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 modelling 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 distinciton 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.[9] 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

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 threee simultaneously into the expansion of BFO motivtaed by the formal-ontologically extended top level.[10]

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.

Conclusions:

Notes toward an industrial philosophy

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 nomenclaturs to digital systems, to make a series of uninformed decisions about complex logical 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, Mariana Dos Santos 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 Humbolt Foundation.

References:

[1] Flett A, Dos Santos M, Ceuster W. Some Ontology Engineering Processes and their Supporting Technologies. Siguença, Spain, October 2002. EKAW 2002.

[2] Smith, B. Basic Formal Ontology, http://ontology.buffalo.edu/bfo

[3] Monteyne F, Fanagan J. Formal Ontology: The Foundation for Natural Language Processing. January 2003. http://www.landcglobal.com

[4] see further: http://ontology.buffalo.edu/

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

[6] Smith, B, Rosse, C. The Role of Foundational Relations in the Alignment of Biomedical Ontologies. http://ontology.buffalo.edu/medo/isa.doc

[7] Smith B, Williams J, Schulze-Kremer S. The Ontology of the Gene Ontology. Proceedings of the AMIA Symposium 2003, forthcoming.

[8] Gene Ontology general documentation. http://www.geneontology.org/doc/GO.doc.html

[9] 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

[10] Smith B. Mereotopology: a theory of parts and boundaries. Data & Knowledge Engineering 1996; 20: 287-303.

Correspondence:

James Fielding may be contacted at matthew@landc.be

Language and Computing
Hazenakkerstraat 20A
B-9520, Zonnegem
Belgium
Tel. +32 (0)53/68.95.55