The Use of Formal Ontology to Increase the Adaptability, Capacity and Efficiency of Natural Language Processing Software

Jonathan Simon[1], Matthew Fielding[2], Mariana Casella Dos Santos[2], MD

[1]Institute for Formal Ontology and Medical Information Science, Leipzig, Germany
[2]Language and Computing nv., Zonnegem, Belgium

Abstract: *The central hypothesis of the collaboration of L&C and IFOMIS is that the methodology of formal ontology will benefit application ontologies such as L&C's Linkbase. In this paper we discuss one of the general procedures to be implemented, with examples of several areas in which it has already brought greater clarification and perspicuity (clarification of ambiguity, allowing better future algorithm design, i.e. less human operator reliance, as well as a framework for a future translation hub) to the Linkbase ontology. The general procedure has been the implementation of a meta-ontological definition space, in which definitions of all concepts and relations of Linkbase are standardized in a framework of first-order logic. We then describe how this standardization effort has led to improvement of Linkbase's treatment of parthood relations, relation between processes and objects, treatment of absence, and of functions. Our description also points to ways in which application ontologies in general are forced to grapple with genuinely philosophical issues.*

General Procedures – Standardization

Linkbase is a medical domain ontology designed to integrate different medical terminologies and ontologies for use in Natural Language Processing applications. This task turns out to be staggeringly complex, since the different terminologies/ontologies to be integrated are often ambiguous and internally inconsistent, and mutually inconsistent to an even greater degree. Linkbase provides a central "hub" ontology, with fixed structured definitions into which external medical terminologies/ontologies may be embedded.[1]

BFO (Basic Formal Ontology) is a philosophically inspired top-level ontology.[2] For millennia, when we have encountered problems in understanding reality, we have turned to philosophers for solutions. Now, when we encounter problems in understanding how to *represent* reality, we must do the same. The cause of the aforementioned ambiguities and inconsistencies has been precisely the lack of a unified framework for understanding many of the basic formal relationships that structure reality (of object to process, of universal to particular, parthood, dependence, and so forth). BFO provides a coherent, unified understanding of these relationships. The implementation of BFO, therefore, as a top-level or "backbone" ontology for Linkbase, will not only provide a framework for the clarification of existing ambiguities and discrepancies in and between ontologies, but will also provide a template for future revision and augmentation of those ontologies. Thus, the implementation of a philosophically sound top-level ontology will provide the necessary link to successful integration, as well as be a useful guide for future algorithm development.[3]

The BFO ontology will provide Linkbase with standardized, formal (first-order) definitions of Linkbase elements (concepts and binary relations). This will disambiguate Linkbase itself, and isolate regularities which will facilitate axiomatic reasoning based on these formal definitions, and more generally the development of future algorithms. The standardization is an implementation of philosophical rigour in two dimensions. First, the first-order language used will be the language in which BFO is defined and axiomatized. Thus, the rigour of the BFO classification system is imported into Linkbase. This is a "metasystematic" importation of rigour, in that few changes are made to the elements themselves, but rather their place in a BFO-founded domain ontology is "tagged". The second dimension of rigour will be of the conceptual analysis variety. Linkbase itself may be viewed as an "object language" or a surface structure. It consists of a number of concepts and binary relations between them. Its axioms at this stage are therefore merely a list of instantiated binary relations. Yet 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" common to every such concept, relation, and axiom requires sound conceptual analysis. The BFO standardization provides for this. The analysis is to run as follows:

1) 1) For every Linkbase concept C, the definition is a mapping to a pair: <the universal named by C, the extension of the universal named by C>

2) For every Linkbase relation R(X,Y), the definition is a mapping to a Π_2 formula (where X and Y are variables ranging over Linkbase concepts):

For all x: x is the universal named by X or x is in the extension of that universal,

There is a y: y is the universal named by Y or y is an element in the extension of that universal, such that R*(x,y)

(where R*(x,y) is a relation in the formal language of BFO)
Certain relations will be tagged to specify that only the concept universal, or only its class of instances, serve as relata.
3) Axioms run according to the relation paradigm, with the variables replaced by specific Linkbase concepts.

The Π_2 structure was chosen since this turned out to be compatible with the intended reading of 99% of existing Linkbase elements.

The remainder of this article is devoted to describing a small selection of cases where this philosophical scrutiny has improved Linkbase.

Parthood

There is a variety of disagreements between taxonomic systems that center on divergent uses of the relation of parthood.[4] In SNOMED, the concept "amputation of foot" subsumes the concept "amputation of toe", where it ought to be represented as an amputation of a part of the foot.[5] For most other classification systems the concept "amputation of foot" means an amputation of the whole foot (and thus does not subsume "amputation of toe"). Thus, identically named concepts have different denotations, and integration attempts that equated the two would be prone to error. Linkbase's understanding of the notions of parthood and proper part allows us to build an accurate representation in which both the SNOMED and the general concept of "amputation of foot" are recognized, but further, they are recognized as distinct, and their relation to each other is mapped.[6]
"Amputation of foot" is related to the concept "foot structure" (a part of the foot), and it subsumes two other concepts "complete amputation of foot" (related to the concept "foot") and "partial amputation of foot" (related to the concept "proper part of the foot").

Objects and Processes

In philosophical circles it is well understood that the universe of common sense contains two types of entities that relate differently to time.[7] 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. Such conflation is the source of many application problems.

Absences

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 bd(x) = bd(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 modelling 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.[8]

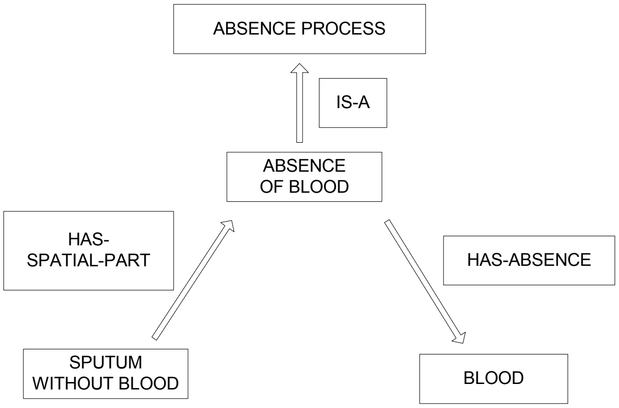

Figure 1: Previous representation of absence

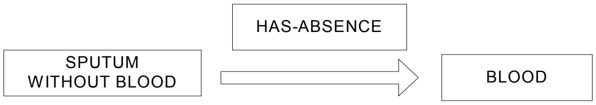

Figure 2: New representation of absence

Functions

Philosophers have been plagued since the days of Aristotle by the nature of functions. What do we mean when we say that "the function of the heart is to pump blood" or that "a function of the pancreas is to produce insulin"? There are two important features characterizing functions:

F1) Functions need not be realized. It is the function of Peter's heart to pump blood even when his heart has stopped beating.

From this it follows that we may not identify functions with processes of their realization.

F2) While we tend to hold as true assertions like "The function of the heart is to pump blood", there is in fact a hidden proviso: the function of the heart is to pump blood *properly*. When the heart beats improperly, it is (usually) malfunctioning.

The original Linkbase model, designed to conform with usage in natural language, contained inaccuracies relating to both of the above concerns. Information about function was represented primarily as a Linktype: Has-Function, which held between an entity concept and a process concept (for example, Heart_Has-Function_Beating-Process). A concept was a function concept whenever it was in the range of this link. In addition, for many concepts, such as 'Heart', a parent function concept, 'Function of Heart', was countenanced. This concept was a process concept and subsumed all processes that were functions of the heart. Thus the model violated rule F1) above.

Further, since conforming with natural usage was a key parameter, 'Heart_Has-Function_Beating-Process' and 'Pancreas_Has-Function_Production-of-Insulin' were taken as axioms. Yet this violates proviso F2), since 'tachycardia' -- a condition inducing malfunctional heartbeat, is subsumed under 'Beating Process', and thus under 'Function of Heart' and 'Insufficient Insulin Production', a malfunction -- is still subsumed under the concept 'insulin production process', which again implies that insufficient insulin production is a function of the pancreas. This is at best innacurrate, and at worst leads to contradiction, which arises if our inference engine contains a rule entailing that no dysfunction of X may also be a function of X.

Again, philosophical scrutiny demonstrated that a gross change in modelling strategy was required. Our new modelling meets the demands of the conditions above, thus avoiding the pitfalls of derivable contradiction, inaccuracy and lost information.

In the new analysis, we treat functions as particularized properties, whose behavior is much like that of states or attributes.[9] Importantly, function concepts such as 'Function of Heart' are no longer subsumed under process concepts, and thus are distinguished from the actual processes of their realization. A new link, 'Is-Realization-Of' links functions to processes realizing them. Since the extensions of function concepts are no longer processes, we are able to distinguish between 'proper' and 'improper' processes of a given type without the awkward terminological move of qualifying all function concepts with the word 'proper'. That is, we may keep the axiom 'Heart_Has-Function_Function-of-Heart' (we do not need to replace it with 'Heart_Has-Function_Proper-Function-of-Heart), since any inferences about subsumption rely on the additional information represented by the Is-Realization link, which filters, e.g., the proper from the deviant beating processes.

Furthermore, we employ an 'X_Displaces_Y' link between process types: this holds whenever X processes prevent Y processes from occurring. If X processes realize a certain function F, and Y processes do not, then Y processes will be dysfunctional with respect to F. An example, for the pancreas: 'Insufficient Production of Insulin' displaces 'Function of Producing [an adequate amount of] Insulin', but it does not displace 'Function of Glucose Metabolization'. Dysfunctions are now relativized in a manner that preserves information lost in the original analysis.

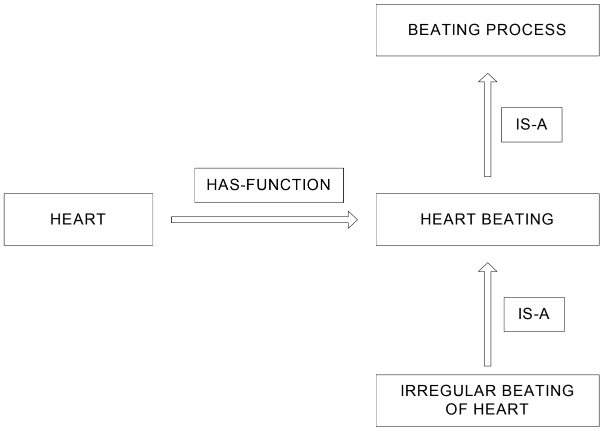

Figure 3: Previous representation of function

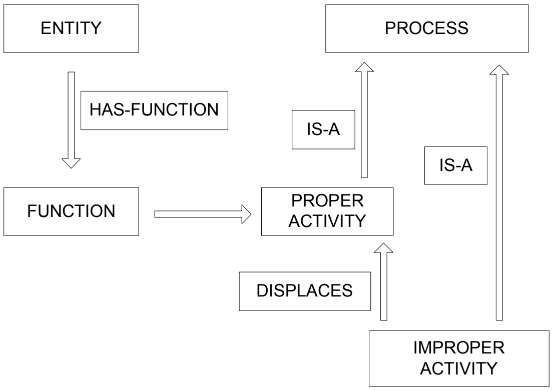

Figure 4: New representation of function

Conclusions

The philosophical restructuring of Linkbase is in its infancy, yet we already have several examples demonstrating the use of philosophical knowledge and technique in improving the database: we have seen examples in which changes were made leading to enhancement of internal consistency and efficiency, as well as steps towards a general database translation hub.

If early success is 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. For what the results cited here demonstrate is not simply that there have been lucky and isolated circumstances under which this integration happened to produce results. Rather, there is a pattern here: Ad hocness characterizes so many features of so many medical ontologies, and this is the main cause of both the tower of babel problem and many local algorithm development constraints. Yet, this ad hocness does not arise by accident. Rather, it has developed because applied ontologists are forced to make uninformed decisions about complex classification issues; indeed the same issues that philosophers have been pondering for millennia.[10] Yet in applications the importance of philosophically sound solutions is often obscured by the temptation of immediate solutions to localized problems. In this way the forest is lost for the trees; the larger integration problems are rendered unsolvable and ad hoc solutions foster further ad hoc problems.

It is thus a tangled web we weave, when we seek to create applied ontologies without a basis in formal ontology. Yet, as our negligence here caused the problem, so shall our renewed vigilance solve it.

References

1. Flett A,Dos Santos M, Ceusters W. Some ontology engineering processes and their supporting technologies. Siguença, Spain, October 2002. EKAW2002.
2. Smith B. Basic formal ontology, http://ontology.buffalo.edu/bfo
3. Montayne F, Flanagan J. Formal ontology: the foundation for natural language processing. January 2003. http://www.landcglobal.com
4. Hahn U, Schultz S, Romacker M. An ontological engineering methodology for part-whole reasoning in medicine.1998. citeseer.nj.nec.com/hahn98ontological.html
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. Simons P. How to exist without temporal parts. The Monist 2000, 83.3, 419-436.
8. Van Geyt L, Martens P, Terzic B, Flanagan J. Get more out of your unstructured medical documents. December 2002. http://www.landcglobal.com
9. Millikan RG. Language, thought and other biological categories. Cambridge: MIT Press, 1984.
10. Smith B. Ontology. In: Floridi L. (ed.), Blackwell guide to philosophy, information and computers. Oxford: Blackwell, 2003.

Acknowledgments

We are grateful for helpful comments from Barry Smith and Dirk Siebert. This work was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation.