A Strategy for Improving and Integrating Biomedical Ontologies

Cornelius Rosse, MD, DSc,[a,b] Anand Kumar, MD,PhD,[d] Jose LV Mejino Jr, MD,[a]
Daniel L Cook, MD, PhD,[a,c] Landon T Detwiler,[a] Barry Smith, PhD,[d,e]
Structural Informatics Group, Departments of Biological Structure[a],
Medical Education and Biomedical Informatics[b], and Physiology and Biophysics[c],
University of Washington and
IFOMIS, University of Saarland, Saarbrücken, Germany[d],
Department of Philosophy, SUNY at Buffalo[e]

The integration of biomedical terminologies is indispensable to the process of information integration. When terminologies are linked merely through the alignment of their leaf terms, however, differences in context and ontological structure are ignored. Making use of the SNAP and SPAN ontologies, we show how three reference domain ontologies can be integrated at a higher level, through what we shall call the OBR framework (for: Ontology of Biomedical Reality). OBR is designed to facilitate inference across the boundaries of domain ontologies in anatomy, physiology and pathology.

Keywords: ontology integration, top-level ontology, domain ontology, terminology, biomedicine

Introduction and Background

Ontology, a branch of philosophy, is the theory of what exists in all areas of reality. The term has of late acquired currency also in the biomedical domain, where a number of terminologies, developed primarily for encoding or annotating biological or clinical data, are now commonly referred to as ontologies. Unfortunately, most such terminologies are not supported by theories, and they fall short to varying degrees of conforming to sound ontological practice.[1,2] Even those computational representations of biological entities that were designated as ontologies from their inception exhibit similar shortcomings.[3]

Ontologies in biological and medical domains may be sorted into three categories:

1. Formal, top-level ontologies, such as DOLCE[4] and Basic Formal Ontology (BFO)[5], which provide domain-independent theories largely through a framework of axioms and definitions, involving categories such as *continuant*, *process* and *boundary* and relations such as *is_a* (for subtype) and *part_of.* They are marked by a high degree of representational adequacy and are designed primarily to be used as controls on the remaining two types of ontologies.

2. Domain reference ontologies, such as the Foundational Model of Anatomy (FMA),[6] which declare a theory about a particular domain of reality and make use of the machinery of formal ontology for sorting and interrelating the entities that exist in this domain. Reference ontologies are general-purpose resources designed to generalize to other domains (e.g., from the human to other species) and also to support a range of different types of research and clinical applications.

3. Terminology-based application ontologies, which are systems of terms (or ‘controlled vocabularies’) purpose-built and designed to meet particular needs, such as annotating biological databases (e.g., the Gene Ontology and other OBO ontologies[7]) or the medical record (e.g., ICD-10, SNOMED).

Biomedical ontologies are being developed in ever growing numbers; but unfortunately there is still too little attention paid by the various separate groups involved to results already obtained by other groups working in neighboring or even overlapping fields. Thus although several anatomy terminologies exist already for both human and murine species, new anatomy terminologies were recently developed for each of these species in the NCI Thesaurus[8] without any apparent reference to this existing work.

The need for ontology alignment and integration is however widely accepted. The attempts to realize this goal have been limited thus far to the creation of mere mappings (often at the purely syntactical level) horizontally between one terminology-based ontology and another. Since many existing ontologies are of poor quality, however, the provision of such mappings represents no scientific advance. To achieve the necessary improvements, we advocate vertical integration between ontologies in the three categories, of a sort designed also to achieve improvements in quality of the aligned terminology-based application ontologies. The type 1 formal, top-level ontologies should provide the validated framework for type 2 (reference) ontologies that represent the domains of reality studied by the basic biomedical sciences. The latter should then in turn provide the scientifically tested framework for a variety of type 3 (terminology-based) ontologies developed for specific application purposes.

The basic sciences of anatomy, physiology, pathology, microbiology, genome science and molecular and developmental biology have served for many generations as the foundations of clinical medicine. The results of these basic sciences are applied in all application domains of health care and biomedical research. Developers of application ontologies intended to support activities in biomedical research, education and health care should now similarly be in a position to reuse these results and to this end reference ontologies for each of the corresponding basic science domains are urgently needed. At the same time an approach to ontology development based on vertical correlations between the three types of ontologies, by providing for the inheritance of sound ontology construction principles, should promote horizontal interoperability between different ontologies of types 2 and 3 in ways which will support new types of automatic inference across the different domains of biomedicine.

In this paper we describe an example for this new paradigm of ontology development. We report on the process of integrating a domain reference ontology, exemplified by the FMA, into the framework of BFO, a formal, top-level ontology, and illustrate the benefits that can be realized through such integration. The following two sections introduce BFO and the FMA and related evolving ontologies. We then propose as a hypothesis the first iteration of an Ontology of Biomedical Reality (OBR).

Basic Formal Ontology

BFO is a formal, top-level ontology based on tested principles for ontology construction, which subdivides reality into two orthogonal categories[5]. The SPAN ontology takes account of *occurrents*, processual entities (events, actions, procedures, happenings) which unfold over a span of time from their beginning to their ending. The complementary SNAP ontology ranges over *continuants*, the participants in such processes, which are entities that endure during the period of their existence. We can think of the SNAP ontology as a snapshot of the continuant entities existing at some given instant of time. The SPAN ontology, in contrast, surveys the processes unfolding over a given interval of time. Anatomy is a science that studies biological continuants; physiology studies biological occurrents. Pathology, on the other hand, is concerned with structural alterations of biological continuants and with perturbations of biological occurrents which together are ultimately manifested as diseases. Since processes cannot exist without their participants, occurrents are entities that *depend* on corresponding continuants.

Entities can thus be sorted also into *independent* and *dependent* categories. Cells and organs are independent continuants; a cell surface or a state of an organism (e.g. of being alive) is a dependent continuant. In addition to the orthogonal SNAP and SPAN categories, BFO draws distinctions also between *instances* (individuals, tokens, particulars) and *universals* (categories, types, kinds, classes), and furnishes formal definitions for its high-level categories, and for the relations which link them.[9-11]

FMA and Associated Ontologies

The Foundational Model of Anatomy was initially intended as a terminology to enhance the anatomical content of UMLS, but in the course of its development was gradually transformed into a reference ontology for anatomy.[12] Independent evaluations indicate that in its current state the FMA satisfies a comprehensive suite of requirements deemed to be fundamental for a sound ontological representation of a life science domain.[3,13]

The FMA comprehends the structure of the idealized human body – it deals with canonical anatomy – and ranges over those classes of anatomical entities (anatomical universals) which exist in reality through their instances. The root of the FMA's anatomy taxonomy (AT) is `anatomical entity` and its dominant class is `anatomical structure`. `Anatomical structure` is defined as a material entity which has its own inherent 3D shape and which has been generated by the coordinated expression of the organism's own structural genes. This class includes material objects that range in size and complexity from biological macromolecules to whole organisms. However, the definition excludes tumors and other pathological lesions as well as those foreign organisms and their parts, as well as non-biological entities, which are introduced into the organism. Portions of body substance (e.g., a portion of cytosol, intercellular matrix, lymph) and immaterial entities (spaces, surfaces, lines and points) are represented in the FMA in terms of their relation to anatomical structures. More than two million instantiations of over 150 relations interrelate AT's more than 72,000 classes in several relational networks.[12,14] Consistent with its foundational nature, the FMA is providing a template for two evolving biomedical domain ontologies: the Physiology Reference Ontology (PRO)[15] and the Pathology Reference Ontology (PathRO).

Ontology of Biomedical Reality

To demonstrate how vertical integration with a sound top-level ontology can support horizontal integration of type 2 ontologies we here present OBR, a federation of interdependent ontologies which range over the domains of biomedical reality traditionally studied by the basic biomedical sciences (Figure 1). It is rooted in the principles embodied in BFO and the FMA and also amalgamates the evolving PRO and PathRO and is, like all of these, purpose-neutral.

The root of OBR is the universal biological entity, which is primitive (thus not defined). A distinction is then drawn between the classes biological continuant and biological occurrent, the definitions of which are inherited from BFO, whose SNAP-SPAN framework incorporates a methodology for reasoning simultaneously with both continuants and occurrents.[5] Next we distinguish in both the continuant and occurrent categories classes of entities that range over single organisms and their parts and associated processes (instances of organismal entity) from those that range over aggregates of organisms and over processes in which multiple organisms participate (instances of extra-organismal entity). The FMA, PRO and PathRO comprehend organismal entities only.

Independent organismal continuants

We first subdivide the class organismal continuant into independent and dependent subcategories. Extrapolating from the FMA's principles, independent organismal continuants have mass and are material, whereas dependent continuants, which are immaterial, do not have mass. Relying on the FMA's definition of anatomical structure, we distinguish anatomical (normal) from pathological (abnormal) material entities; the latter resulting from processes other than those governed by the organism's structural genes. We also extend the FMA's representation of material entities by distinguishing structures from portions of body substances on the basis of the possession by the former of their own inherent 3D shape. The representation of blood, bile, etc., in terms of portions caters to the need to give an account of those processes in PRO and PathRO which involve as participants, for example, different portions of blood, inspired air, pus, exudate or cytoplasm.

Within the class anatomical structure we make a distinction between canonical anatomical structures, which exist in the idealized organism, and variant anatomical structures, which result from an altered expression pattern of normal structural genes without health related consequences for the organism (e.g., presence of a middle lobe of left lung,

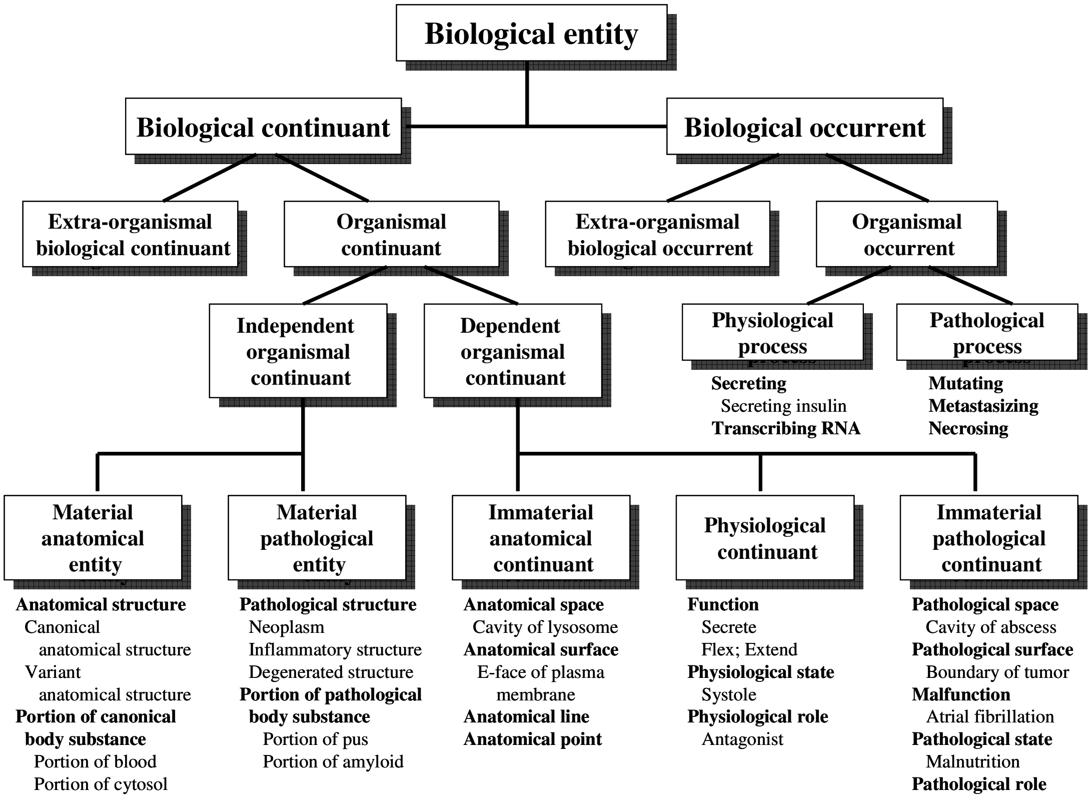

Figure 1. Ontology of Biomedical Reality

accessory spleen, or dextrocardia). In contrast, most instances of `pathological structure` come into being through the transformation of normal anatomical structures by processes other than the expression of the organism's normal complement of genes (e.g., abscesses, tumors, granulomas, etc.).

We retain the traditional paradigm of sorting pathological structures into categories on the basis of etiology and the pathogenic processes that generate them (e.g., congenitally abnormal structure, neoplasm, inflammatory structure, and so on).

Since tumors or portions of pus do not exist in the domain of entities represented in canonical anatomy, `pathological structure` and `portion of pathological body substance` are not subordinated in OBR to the class `material anatomical entity`, but rather to its sibling class `material pathological entity`.

Dependent organismal continuants

In analogy with their normal anatomical counterparts, we recognize also pathological spaces and boundaries such as surfaces and lines. These immaterial anatomical and pathological entities are dependent continuants, since their existence depends on corresponding independent continuant entities. Thus without the stomach, the surface and cavity of the stomach cannot exist, and the same dependence prevails for the cavity of an abscess on the abscess or for the surface of a tumor on the tumor.

Examples of different classes of dependent continuants are the functions, states and roles in the domains of physiology and pathology. Functions are certain sorts of potentials of independent anatomical continuants for engagement and participation in one or more processes through which the potential becomes realized. Whereas processes unfold in time, the function (e.g., the potential of a cell to synthesize a particular protein) is a continuant, since it, too, endures through time and it exists even during those times when it is not being realized. The function to secrete, though dependant on the cell, is independent of the actual processes (of function*ing*) that realize this function.

Whether or not a function becomes realized depends on the physiological or pathological state of the associated independent anatomical continuant. By physiological and pathological state we mean a certain enduring constellation of values of an independent continuant's aggregate physical properties. These physical properties are represented in the Ontology of Physical Attributes (OPA), which provides the values for the physical properties of organismal continuants. The states of these continuants can be specified in terms of specific ranges of attribute values and a change in these values will become manifest as an observable and measurable process.

The independent continuants that participate in a physiological or pathological process may play different roles in the process (e.g. as agent, co-factor, catalyst, etc.). Such a process may transform one state into another; for example a physiological into another physiological, or into a pathological state. One pathological state is transformed into another as a pathological structure or disease develops.

Organismal Occurrents

Paralleling the classification of independent organismal continuants, we distinguish among organismal occurrents between physiological and pathological processes (Fig.1). The transformations of one physiological state into another are instances of `physiological process`, whereas processes that transform a physiological into a pathological state, or one pathological state into another, are instances of `pathological process`. The relative balance of the two kinds of processes as they evolve over time results either in the maintenance of health or in the pathogenesis of material pathological entities and thus in the establishment and progression of diseases. Transformation of a pathological state into a physiological one, manifest as healing or recovery from a disease, comes about through physiological processes that successfully compete with and ultimately replace pathological processes. Function is restored.

Processes are extended not only in time but also in space by virtue of the nature of their participants; their location is the location of their participants. The synthesis of insulin, for example, takes place within a beta cell because the protein synthetic apparatus with the potential (function) to engage in this synthesizing process is included in the beta cell as part.

Conclusions and Discussion

We propose the Ontology of Biomedical Reality as a high-level framework for integrating in robust fashion those more narrowly tailored ontologies developed to meet the needs which arise in specific domains of life science and health care. Ontology integration through OBR complements current reform efforts, for example in the framework of the OBO (Open Biomedical Ontologies) consortium to bring about a greater degree of formal rigor in the definition of biomedical relations.[10] We hope that it will greatly improve on the prevailing practice of mapping between the leaves of structured vocabularies. The benefits of integration through OBR are entailed by the sound ontological structure of BFO and of the FMA, which the OBR framework imposes on existing and emerging ontologies in disparate domains, and this in turn lays the ground for the drawing of inferences across the boundaries of domain ontologies.

OBR integrates the high-level classes of three domain reference ontologies, which take account of normal and perturbed biological processes along with the material and immaterial entities that participate in these processes, encompassing the entire granular spectrum from biological macromolecules to the whole organism. The integration calls for subordinating selected classes of reference domain ontologies to corresponding classes of OBR. For example, the FMA classes anatomical structure and portion of canonical body substance are subordinated to the OBR class independent organismal continuant, which also subsumes the PathRO classes pathological structure and portion of pathological body substance. The OBR class dependent organismal continuant subsumes the class immaterial anatomical entity from the FMA, as well as function and physiological state from PRO and malfunction and pathological state from PathRO. Only PRO and PathRO contribute to the OBR class biological occurrent.

The need for OBR arose in the course of our work on the Virtual Soldier Project (VSP), which calls for reasoning about traumatic injuries and prediction of their outcomes. Such reasoning must bridge the divide between FMA, PRO and PathRO, which together with OPA, constitute the VSP knowledge base (VSKB). The latter provides input for the amendment of mathematical models of physiological functions in order to refine their ability to simulate the behavior of functional systems and predict the outcomes of traumatic injuries. The VSP serves as the motivator and evaluation domain for OBR. Preliminary results attest to the potential and power of an alignment of ontologies along the lines here proposed, suggesting that the advantages afforded by OBR will translate into multiple areas of biomedical research and health care, including the electronic health record.

Acknowledgements

This work was supported in part by a grant from the DARPA Virtual Soldier Project, Contract # W81XWH-04-2-0012, the Wolfgang Paul Program of the Alexander Von Humboldt Foundation and the Project "Forms of Life" of the Volkswagen Foundation.

References

1. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based Error Detection in SNOMED-CT®. Medinfo 2004;2004:482-6.

2. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. *Methods of Information in Medicine* 2005; in press.

3. Smith B, Köhler J, Kumar A. On the application of formal principles to life science data: A case study in the Gene Ontology. *DILS 2004: Data Integration in the Life Sciences*. 2004;124-139.

4. DOLCE:**http://www.loucnr.it/DOLCE.html**

5. Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. In DM Pisanelli (ed.*), Ontologies in Medicine*, Amsterdam:IOS Press, 2004,20-38.

6. Foundational Model of Anatomy: **http://fma.bistr.washington.edu**

7. Open Biomedical Ontologies: **http://obo.sourceforge.net/**

8. NCI Thesaurus: **http://www.nci.nih.gov/cancerinfo/terminologyresources**

9. Smith B, Rosse C: The role of foundational relations in the alignment of biomedical ontologies. Medinfo 2004;2004:444-448.

10. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A., Rosse C. Relations in biomedical ontologies. Genome Biology 2005, 6:R46

11. Donnelly M, Bittner T, Rosse C. A formal theory for spatial representation and reasoning in biomedical ontologies. *IFOMIS Reprints* 2005: 1.

12. Rosse C, Mejino JLV Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform, 2003 Dec;36(6):478-500.

13. Zhang S, Bodenreider O. Law and order: Assessing and enforcing compliance with ontological modeling principles. *Computers in Biology and Medicine* 2005: in press.

14. Mejino JLV, Rosse C. Symbolic modeling of structural relationships in the Foundational Model of Anatomy, in *U. Hahn (ed.), Proceedings of KR-MED 2004* (First International Workshop on Formal Biomedical Knowledge Representation), 48-62.

15. Cook DL, Mejino JLV, Rosse C. Evolution of a Foundational Model of Physiology: Symbolic representation for functional bioinformatics. Medinfo 2004;2004:336-340.