Revising the UMLS Semantic Network

Steffen Schulze-Kremer[a], Barry Smith[a,b] and Anand Kumar[a]

[a]*IFOMIS, University of Leipzig, Germany*
[b]*Department of Philosophy, University at Buffalo, NY, USA*

Abstract

The integration of standardized biomedical terminologies into a single, unified knowledge representation system has formed a key area of applied informatics research in recent years. The Unified Medical Language System (UMLS) is the most advanced and most prominent effort in this direction, bringing together within its Metathesaurus a large number of distinct source-terminologies. The UMLS Semantic Network, which is designed to support the integration of these source-terminologies, has proved to be a highly successful combination of formal coherence and broad scope. We argue here, however, that its organization manifests certain structural problems, and we describe revisions which we believe are needed if the network is to be maximally successful in realizing its goals of supporting terminology integration.

Keywords:

Terminology, Ontology, UMLS Semantic Network

Introduction

The January 2003AB version of the UMLS Metathesaurus, which is the total UMLS concept repository, includes some 900,551 concepts and 2.5 million concept names in its source vocabularies. [1] The associated Semantic Network (hereafter SN), consists of 134 Semantic Types together with 54 possible links between these types, and represents a high-level abstraction from the UMLS Metathesaurus. SN is a graph containing more than 6,000 edges organized into a double tree structure. An additional tree hierarchically classifies all available link types. We focus here on SN's role as a classificatory (*isa*) hierarchy, leaving for another place the scrutiny of SN's non-*isa* links. We aim to stay as close as possible to the intentions of SN's creators by focusing on potential problems which need to be addressed if the network is to serve its purpose of supporting integration of diverse biomedical information sources.

Ontology needs some most general term to designate everything which exists (all items, objects, beings, existents), and the term 'entity' has established itself in this role. Since SN's top-level division into *entities* and *events* corresponds in the ontological literature to that between *continuants* and *occurrents*, we here depart from SN usage and talk not of 'entities' and 'events' but rather of *continuant* and *occurrent entities*.

Continuants are entities which *endure*, or *continue to exist*: they preserve their identity from one moment to the next even while undergoing changes. Examples are physical objects (organs, cells, genes, molecules, …) – but also: shapes, qualities, dispositions, states, roles, spatial regions, body sites, and functions. Entities in the latter categories, too, endure self-identically through time. The function of a thermometer to measure temperature exists self-identically from one moment to the next – and it exists even at those times when it is not being exercised.

The *exercise* of a function, in contrast – like the *performance* of a role, the *execution* of a plan, the *application* of a therapy, the *realization* of a disposition – is an occurrent entity, an entity which *occurs* in a given interval of time. Occurrent entities (processes, events, activities, changes, histories) unfold themselves in time; they never exist in full in any single instant.

Ontologists distinguish also between *independent* and *dependent* entities, corresponding in first approximation to SN's distinction between *physical objects* on the one hand and *conceptual entities* and *events* on the other. To say that an entity is *dependent* is to assert that it requires a support from other entities in order to be sustained in existence. There is no mass or shape without some body, no cellular motion without some cell which moves. *Independent* entities require no such support, for they are themselves the substrates for qualities, dispositions, motions, functions and other dependent entities.

At those levels of granularity which are of concern in biomedicine, occurrents are always changes *of* or *in* some enduring entity or entities; thus they are always dependent entities. Of the four possible combinations yielded by our two divisions, therefore, only three are instantiated: dependent and independent continuant and occurrent. This system of three categories provides the top-level architecture *inter alia* for the DOLCE ontology developed within the framework of the Semantic Web Initiative as the first module of the Wonderweb Foundational Ontologies Library [2]. It underlies also a number of other ontological systems in current use, including LinKBase®, the large terminology-based medical ontology developed by the company L&C in Belgium [3].

Method

We formulated a series of consensus ontological principles and used these as the basis for a thorough audit of the SN. We present here samples of the results of applying the following principles to SN's classificatory hierarchy:

P1. entities in different highest-level categories (independent continuant, dependent continuant, occurrent) should not be combined within a single class;

P2. *objects* should not be combined within a single class with the *roles* they play or with the *functions* they exercise;

P3. *entities in reality* should not be combined within a single class with our *knowledge about* or with our *concepts of* such entities;

P4. what is *concrete* (what exists in space and time and enters into causal relations) should not be combined within a single class with what is *abstract* (for example with abstract spatial regions, measures, and the like);

P5. classifications should respect the factor of *time*; for example classes should be assigned in a way that is consistent with the fact that continuant entities endure through time.

We show that failure to respect these principles implies suboptimal reasoning capabilities: valid inferences will be blocked and invalid inferences will be admitted.

Results

We focus on the graph generated by SN's class-subclass subsumption links. Working up from the lowest nodes of this graph, we consider the question of compliance with the five basic principles mentioned above, and draw conclusions as to possible revisions of the network. At the same time we add commentary pointing to other potential problems in SN's organization, sometimes drawing on examples of assignments of SN-types within the UMLS. We mark only problematic cases, using the designation P*n* to mark potential problems connected with non-compliance with the corresponding principle.

Plant – (Alga): *Plant*, like *algae*, is an independent continuant. By allowing *plant roots isa plant*, *plant leaves isa plant*, etc., UMLS runs together *isa* with *part-of* relations. *Pollen isa plant* may reflect a failure to do justice to the factor of time (it holds only in a specific stage of the reproduction cycle). P5

Vertebrate – (Amphibian; Bird; Fish; Reptile; Mammal) / Animal – (Invertebrate; Vertebrate) / Organism – (Plant; Fungus; Virus; Rickettsia or Chlamydia; Bacterium; Animal; Archaeon): classified under independent continuants.

Fully Formed Anatomical Structure – (Body Part, Organ, or Organ Component; Tissue; Cell; Cell Component; Gene or Genome): These again are independent continuants. *Fully Formed Anatomical Structure* is defined by SN as: "An anatomical structure in a fully formed organism; in mammals, for example, a structure in the body after the birth of the organism." One potential problem with this definition pertains to time. How would it allow one to code terms in literature about e.g. pressures on the heart during birth? From the definition, it is clear that 'structure' is intended to denote a concrete physical object (an independent continuant). Note, however that the term is used elsewhere in the SN (for example under 'Chemical', below) to denote something abstract, namely the way an object or thing is organized, its *Bauplan* or the arrangement of its parts. A minor point: the term carries the erroneous suggestion that it is the entity itself (rather than the organism within which it is housed) that is fully formed.

There are some odd features of the use of *Body Part, Organ, or Organ Component* in the Metathesaurus; thus while 'hand' is a *Body Part, Organ, or Organ Component*, 'fingers' is classified under *Body Location or Region*. Such problems sometimes arise due to the fact that the UMLS merges source vocabularies whose classification hierarchies have different rationales. This cannot be a reason here, however, for while source vocabularies like MeSH and different flavors of SNOMED classify both *hand* and *fingers* as *Body Region* or *Body Region Structure*, Digital Anatomist classifies both as *Body Part Subdivision*. In any case both 'hand' and 'finger(s)' should be classified in the same way.

Gene or Genome is defined as: "A specific sequence … of nucleotides along a molecule of DNA or RNA. A genome is, however, much more than a sequence of genes, and a eukaryotic genome contains only gene fragments. P4

Anatomical Abnormality – (Congenital Abnormality; Acquired Abnormality): *Anatomical abnormality* is defined as: "An abnormal structure, or one that is abnormal in size or location." Thus understood, the term embraces both dependent continuants (such as abnormalities in shape or position of the uterus) and independent continuants (such as an acquired fistula). This categorial ambiguity should be eliminated. P1

Anatomical Structure – (Embryonic Structure; Fully Formed Anatomical Structure; Anatomical Abnormality): *Anatomical Structure* is defined as: "A normal or pathological part of the anatomy or structural organization of an organism." Note that in the phrase 'structural organization', the term 'organization' is not used in conformity with SN's own definition (see below) as meaning 'social organization'. Rather it is used to mean an entity's *Bauplan*. The latter, however, would be, not a concrete three-dimensionally extended independent thing but rather some dependent abstract feature which gives shape and functionality to an entity of this sort. Then, however, it should *not* subsume *liver* or *leukocyte*.

Embryonic Structure is: "An anatomical structure that exists only before the organism is fully formed." This definition is problematic for reasons outlined already under *Fully Formed Anatomical Structure* and *Anatomical Abnormality*. P1, P4, P5

Medical Device – (Drug Delivery Device) / Manufactured Object – (Medical Device; Research Device; Clinical Drug): When we refer to medical devices we can refer either to a physical object or to the *role* this object plays in some context. Under the former heading we are referring to an independent continuant, under the latter to a dependent continuant. We propose introducing into SN a new higher-level category of *role*. This would allow also a more adequate treatment of terms such as *doctor*, *nurse*, *patient*, etc., as also of *manufactured object* and *chemical* (see below and [4]). P1, P2

Lipid – (Steroid; Eicosanoid) / Organic Chemical – (Nucleic Acid, Nucleoside, or Nucleotide; Organophosphorus Compound; Amino Acid, Peptide, or Protein; Carbohydrate; Lipid) / Chemical Viewed Structurally – (Organic Chemical; Element, Ion or Isotope; Inorganic Chemical): All of the above are independent continuants. SN's dichotomy *Chemical Viewed Structurally* and *Chemical Viewed Functionally* reflects, we believe, not a genuine classificatory subdivision, but rather a distinction between types of classification. *Structure* yields one classification, *functions* a second, and to put these two together leads to problems of the same sort which we would face if we were to divide the class of people into: *tall people, people who play tennis,* etc. Chemicals viewed functionally should be treated not as special types of chemicals, but rather in terms of special types of roles or functions. P2

Pharmacologic Substance – (Antibiotic): *Pharmacologic Substance* is of course properly to be classified under *Substance*. Here again, however, there is a dimension of *function* that has to be addressed. A given pharmacologic substance will be unable to perform its function for example when its expiry date has elapsed. We propose distinguishing *Pharmacologic Substance* as an independent continuant, *Pharmacologic Function* as a dependent continuant, and *Pharmacologic Action* as an occurrent. P2, P5

Biologically Active Substance – (Neuroreactive Substance or Biogenic Amine; Hormone; Enzyme; Vitamin; Immunologic Factor; Receptor) / Chemical Viewed Functionally – (Pharmacologic Substance; Biomedical or Dental Material; Biologically Active Substance; Indicator, Reagent, or Diagnostic Aid; Hazardous or Poisonous Substance) / Chemical – (Chemical Viewed Structurally, Chemical Viewed Functionally): *Chemical Viewed Functionally* is defined as: "A chemical viewed from the perspective of its functional characteristics or pharmacological activities." For reasons given above, we suggest that the information captured under this node should be represented in the SN via the addition of new upper-level categories of *role* or *function*. A feature like "hazardous" is more properly classified as a role than as a type of substance, since the same substance may be hazardous in some contexts but not in others, and at low concentrations it may even be beneficial. P2, P5

Substance – (Body Substance; Chemical; Food): *Substance* is defined as: "A material with definite or fairly definite chemical composition." Here again *food* should more properly be classified in terms of a substance's special role in a certain context. P2

Organism Attribute – (Clinical Attribute): The types in question refer to dependent continuants. See discussion under *conceptual entity* below.

Finding – (Laboratory or Test Result; Sign or Symptom): A *Finding* is: "That which is discovered by direct observation or measurement". A *Sign or Symptom* is: "An observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation." These terms reflect a running together of ontology and epistemology. [5] Classifying *Symptom* as a subclass of *Finding* amounts to the assertion that "observable manifestations of a disease" exist only when they are "discovered by direct observation or measurement". This, however, means that your symptoms do not exist until they are cognized as such. *Symptom* should be reclassified under the heading *Organism Attribute.* P1, P3

Molecular Sequence – (Nucleotide Sequence; Amino Acid Sequence; Carbohydrate Sequence): These are independent continuants.

Spatial Concept – (Body Location or Region; Body Space or Junction; Geographic Area; Molecular Sequence): *Spatial Concept* is defined in SN as: "A location, region, or space, generally having definite boundaries", and it thus falls under the category of independent continuants. However, the use of the term 'concept' is then inappropriate. For concept are dependent continuants; concepts exist in people's minds; they are abstract entities which depend on concept-users. There have been concepts only during the period in which there have been concept-using organisms.

Independent continuants like geographic areas and molecular sequences, in contrast, do not exist in people's minds. We surmise that the mentioned subclasses are grouped together under *Spatial Concept* because they are held to share the following characteristics: a) they are extended in space and b) their boundaries are determined not by any underlying physical discontinuities but rather by demarcations introduced by human subjects. The referent of 'Bethesda, MD' satisfies these conditions, and so also does *knee-joint* – but then so also does *hand*, which is not classified by UMLS as a conceptual entity. Molecular sequences, too, can be regarded as the products of demarcation or segmentation. We believe that it is appropriate to include in a top-level ontology the distinction between entities which have bona fide boundaries and fiat entities (entities which exist because of demarcations which we draw). The human body and the liver have bona fide boundaries, while the boundaries of your hand or finger are fiat in nature. [6] But this does not mean that the entities in question are special sorts of concepts. P1, P3, P4, P5

Functional Concept – (Body System): *Functional concept* is defined in SN as: "A concept which is of interest because it pertains to the carrying out of a process or activity." This definition runs together functions, which are dependent continuants (for example the function of your heart, which is to pump blood), and occurrents, which are the realizations or executions of such functions.

Body system is defined as: "A complex of anatomical structures that performs a common function." *Functional concept* is subsumed by *concept*. A body system is however not a concept. (Concepts do not perform functions or have physical parts.) Rather, a body system is a certain physical part of an organism, whose (fiat) demarcation is functionally determined. Body systems – and the associated functions – existed for many millennia before there were concept-using organisms. P1, P3, P4, P5

Idea or Concept – (Temporal Concept; Qualitative Concept; Quantitative Concept; Spatial Concept, Functional Concept): *Idea or Concept* is defined as: "An abstract concept, such as a social, religious or philosophical concept." Note that this definition is circular, since the term 'concept' appears also in the definiens.

Given our arguments above, the entities currently coded under *Spatial concept* and *functional concept* should not be included among the subclasses of *idea or concept*. *Temporal concept*, too, seems to us to be misclassified. As continuants are located in spatial regions, so occurrents are located in temporal intervals. This analogy is lost if both spatial and temporal terms are treated under the UMLS-ST equivalent of the category of continuants. P1, P3, P4

Occupation or Discipline – (Biomedical Occupation or Discipline) / Organization – (Health Care Related Organization; Professional Society; Self-help or Relief Organization) Group – (Professional or Occupational Group; Population Group; Family Group; Age Group; Patient or Disabled Group): *Group* is defined as: "A conceptual entity referring to the classification of individuals according to certain shared characteristics." Groups on the one hand endure in time (and survive changes in their members), but on the other hand they are dependent upon demarcations or classifications. Thus they belong to the category of dependent continuants; but again: they are not concepts. *Professional or Occupational Group* is defined as: "An individual or individuals classified according to their vocation." This makes it clear that *Group* subsumes also individual members, which yields consequences like: 'a member of a physicians organization *isa* physicians organization', 'an individual is an age-group'. A better approach, again, would be to classify types such as *doctor*, *nurse*, *patient*, etc., not in terms of groups but rather in terms of *roles* or *functions* performed by corresponding individuals. P1, P2, P3, P4

Intellectual Product – (Regulation or Law; Classification): Note that the use of the term 'conceptual entity' here does not cause the problems it causes elsewhere; intellectual products are indeed such as to depend for their existence on the existence of concept-using organisms.

Conceptual Entity – (Organism Attribute; Finding; Idea or Concept; Occupation or Discipline; Organization; Group; Group Attribute; Intellectual Product; Language): *Conceptual entity* is defined as: "A broad type for grouping abstract entities or concepts." We have seen above that many of the subclasses of this node are neither abstract entities nor concepts and that they belong to no single higher-level ontological category. *Organism Attribute*, *Occupation or Discipline*, *Organization*, *Group* and *Group Attribute* are entities belonging not to the abstract realm of concepts but rather to the real world of space, time and causality, and they should be reclassified accordingly. P1, P4

Behavior – (Social Behavior; Individual Behavior): SN draws no clear boundary between *Social* and *Individual Behavior*. 'Racism' is classified under both headings, 'smoking' and 'singing' as *Individual Behavior,* 'walking' as *Social Behavior*. Here again we postulate that SN's problematic treatment of individual human beings will be improved via the addition of a category of role. P2

Health Care Activity – (Laboratory Procedure; Diagnostic Procedure; Therapeutic or Preventive Procedure): An *activity* is an occurrent, and the same holds for the *carrying out* of a laboratory procedure. *Procedures* themselves, however, are dependent continuants. A procedure is an intellectual product that endures identically through time and *is-realised-in* its successive applications, perhaps undergoing changes as methods and equipment are refined. P1, P5

Research Activity – (Molecular Biology Research Technique): Here again one needs to distinguish between a *technique* (which is a dependent continuant) and the *application of a technique*, which is an activity (an occurrent). The former stands to the latter not in an *isa* relation but rather in an *is-realised-in* relation. P1, P5

Occupational Activity – (Health Care Activity; Research Activity; Governmental or Regulatory Activity; Educational Activity) / Activity – (Behavior; Daily or Recreational Activity; Occupational Activity; Machine Activity): All of the above are occurrents.

Human-Caused Phenomenon or Process – (Environmental Effect of Humans): Phenomena such as aviation and cultural evolution (both examples given by UMLS) are dependent continuants. Processes are occurrents. When 'water pollution' and 'deforestation' are classified by UMLS as *Environmental Effect of Humans*, this leaves open whether we are dealing with the *phenomenon* (continuant) or the *act* (occurrent) of water pollution or deforestation. To resolve such problems, *Phenomenon* and *Process* and their respective subclasses should be classified separately. P1, P5

Organism Function – (Mental Process): *Organism Function* is a dependent continuant. *Mental Process* is defined as: "A physiologic function involving the mind or cognitive processing." This definition, too, mixes function and process. P1

Molecular Function – (Genetic Function) / Physiologic Function – (Organism Function; Organ or Tissue Function; Cell Function; Molecular Function): Physiologic Function is defined as: "A normal process, activity, or state of the body." This mixes process (occurrent) and state (continuant). P1

Disease or Syndrome – (Mental or Behavioral Dysfunction; Neoplastic Process): We believe that *neoplastic process* is misclassified as a *disease or syndrome*. As SN elsewhere recognizes, a 'disease' is to be distinguished from the processes which constitute a 'disease history'. But only a continuant can have a history. An occurrent does not have a history because it *is* a history. To this degree, therefore, SN recognizes diseases as continuants: they are conditions which endure through time while undergoing changes (for example by becoming chronic, fibrous, malignant). On the other hand, however, diseases are listed by SN as a subclass of *event.* 'Complex of symptoms', too, in being included under *Disease or Syndrome,* is classified by SN as a subclass of *event*; symptoms themselves however are classified under *finding*, and thus under *conceptual entity*.

When 'disease history' is classified under *Health Care Activity* this is a case of running together the history or course of a disease on the side of the patient (ontology) with the act of eliciting that history (epistemology). Similar problems pertain to the

UMLS classifications of ‘natural history of disease’ under *Finding* and of ‘cancer patient’ under *Mental or Behavioral Dysfunction* (the latter a reflection of the odd umbrella term ‘cancer patients and suicide and depression’). P1, P3, P4, P5

Pathologic Function – (Disease or Syndrome; Cell or Molecular Dysfunction; Experimental Model of Disease): Occurrent processes, activities and responses are here combined together with continuant states and conditions. Pathologic functions of organisms are classified in the same way as pathologic functions of body systems. Yet the former are classed by SN as physical objects, the latter as conceptual entities. P1

Biologic Function – (Physiologic Function; Pathologic Function) / Natural Phenomenon or Process – (Biologic Function): Here again the definition of *Biologic Function* mixes *process* (occurrent) and *state* (continuant). P1

Phenomenon or Process – (Injury or Poisoning; Human-caused Phenomenon or Process; Natural Phenomenon or Process): *Injury or Poisoning* is defined as: “A traumatic wound, injury, or poisoning caused by an external agent or force.” A wound (continuant) should be distinguished from a wounding (occurrent), a state of being poisoned (continuant) from an act of poisoning (occurrent). P1

Event – (Activity; Phenomenon or Process): Phenomena and processes should not be classified together. P1, P5

Discussion

The inclusion of the opposition *Chemical Viewed Structurally* and *Chemical Viewed Functionally* suggests that SN might be better interpreted as classifying not entities but rather the concepts we have of such entities. The concepts we use when referring to chemicals can after all be divided quite naturally under these two headings. Then, however, the root nodes of SN should be not: *Entity* and *Event*, but rather: *Entity Concept* and *Event Concept*, and the latter should themselves be re-assigned to the position of daughters of a new root *Concept*. A restructuring along these lines would however in other ways conflict radically with SN’s current architecture. Above all, it would contradict the fact that *Idea or Concept* is already itself a subnode of *Conceptual Entity*. It would also contradict explicit statements to the effect that SN is ‘an upper-level ontology … in which all concepts are given a consistent and semantically coherent representation’. [7]

Conclusion

A number of proposals have been advanced to increase SN’s effectiveness as a terminology integration platform that can support enhanced reasoning and information retrieval. Thus [8] argues that UMLS lacks the requisite granularity, semantic types and relationships for comprehensively and consistently representing anatomical concepts in machine readable form. [9] and [10] propose enhancing the efficiency of UMLS-based reasoning systems via a clustering of SN nodes to yield more coarse-grained partitions of the network.

Our proposal is that SN’s power to support terminology-based reasoning can be enhanced through a reclassification along the lines sketched in the above. As an example of how such a reclassification would support inferences currently blocked, consider the way in which SN currently views tissues and cells as *physical* parts of organs, but views these organs themselves as mere *conceptual parts* of body systems, which are in turn *conceptual parts* of fully-formed anatomical structures, which are in turn *physical* parts of organisms. When we reclassify *Body System* as a *Physical Entity*, there is no longer a need for the distinction between *conceptual* and *physical part-of* relations. Reasoning systems can thus exploit the full power of mereology, including the rules governing transitivity of *part-of*.

The proposed reclassification would lead also to a system of semantic types for which definitions can be formulated which are at one and the same time both more intuitive and also more rigorously formalizable than existing definitions, thereby making it easier to train and monitor those with the task of assigning semantic types to new and existing source-terminologies.

Acknowledgements

Work on this paper was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation. Our thanks go also to Olivier Bodenreider.

References

[1] http://umlsinfo.nlm.nih.gov.

[2] http://www.loa-cnr.it/DOLCE.html.

[3] http://www.landcglobal.com.

[4] Burgun A, Bodenreider O, Le Duff F, Mounssouni F, Loréal O. Representation of roles in biomedical ontologies. Proc AMIA Annu Symp 2002;:86-90.

[5] Kumar A, Smith B. The Unified Medical Language System and the Gene Ontology. Proc KI 2003. In press.

[6] Smith B. Fiat objects. Topoi, 2001; 20: 131-148.

[7] McCray A. An upper level ontology for the biomedical domain. Comp Functional Genomics 2003; 4: 80-84.

[8] Rosse C, Ben Said M, Eno KR, Brinkley JF. Enhancements of anatomical information in UMLS knowledge sources. Proc Annu Symp Comput Appl Med Care. 1995;:873-7.

[9] Halper MH, Chen Z, Geller J, Perl Y. A metaschema of the UMLS based on a partition of its semantic network. Proc AMIA Symp. 2001;:234-8.

[10] Chen Z, Perl Y, Halper M, Geller J, Gu H. Partitioning the UMLS semantic network. IEEE Trans Inf Technol Biomed. 2002; 6(2): 102-8.

Address for correspondence

Steffen Schulze-Kremer, Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Härtelstr. 16, 04107 Leipzig, Germany. URL: http://ifomis.de.