SNOMED CT's Problem List: Ontologists' and Logicians' Therapy Suggestions

Stefan Schulz[a,b], Boontawee Suntisrivaraporn[c], Franz Baader[c]

[a] *Freiburg University Hospital, Medical Informatics Department, Freiburg, Germany*
[b] *Pontificial Catholic University of Paraná (PUCPR), Master Program of Health Technology, Curitiba, Brazil*
[c] *Dresden University of Technology, Faculty of Computer Science, Dresden, Germany*

Abstract

After a critical review of the present architecture of SNOMED CT, addressing both logical and ontological issues, we present a roadmap towards an overall improvement of this terminology. In particular, we recommend the following actions: Upper level categories should be re-arranged according to a standard upper level ontology. Meta-class like concepts should be identified and removed from the taxonomy. SNOMED concepts denoting (non instantiable) individual entities (e.g. geographical regions) should be kept separate from those concepts that denote (instantiable) types. SNOMED binary relations should be reduced to a set of canonical ones, following existing recommendations. Taxonomies should be cleansed and split into disjoint partitions. The number of full definitions should be increased. Finally, we propose a new approach to modeling part-whole hierarchies, as well as the integration of qualifier relations into the description logic framework.

Keywords:

SNOMED CT, Ontologies, Description Logic

Introduction

The globalization of SNOMED CT offers a unique opportunity to bring together the following tendencies:

- The urgent need for a global standardized terminology for medicine and life sciences, suitable to cope with an immense flood of clinical and scientific information;
- An impressive legacy of systematized terminology;
- Ongoing efforts toward an ontological foundation of the basic kinds of entities in the biomedical domain;
- The increasing availability of logic-based reasoning artifacts suited for large terminological knowledge bases.

After a long-lasting embryonic and fetal period, the Standardized Nomenclature of Medicine, SNOMED, has seen the light of day as SNOMED CT (Clinical Terms). Since then, it has grown rapidly, in spite of some congenital and infancy problems which have raised the attention not only of domain experts, medical terminologists and software engineers but also of computer scientists and ontologists. Concerned about the conditions under which SNOMED CT is now reaching its adolescence, we recommend an in-depth health check.

Clinical advice will be required from specialists of Ontology and Logic. A careful follow-up of their counseling will be crucial for making SNOMED CT fit for the next decades.

According to the present status of its growth chart [1], SNOMED CT counts 300,000 concepts, about 770,000 English language descriptions, and 900,000 defining relationships. Each SNOMED concept is assigned to one of 19 top-level categories and each relation to one of 49 attribute types.

Methodology

The Ontologists' Investigation

Amongst the multiple definitions of what an ontology is [2], we here narrow the scope to *formal* ontologies. We consider a formal ontology a formal theory that accurately describes a domain in light of the kinds of entities it contains. Ontologists evaluate a terminology system by its agreement with ontological principles, such as the use of well-defined, unambiguous, and non-idiosyncratic types and relations, in accordance with existing standards.

The Logicians' Investigation

When dealing with ontologies or terminological systems, logicians are primarily interested in formal description languages in which the whole content of the terminology can be expressed. Such languages need to have sufficient expressivity for describing the domain entities. It must be suited for supporting truth-maintenance in the terminology design and maintenance process. It should be sufficiently intuitive for being used in a proper way by the curators of the terminology.

In the style of a clinical problem list, we will present first a thorough analysis of SNOMED CT's current status, under both ontological and logical perspectives. After that we will present a plan for adequate therapeutic measures which will be as little as invasive as possible, cost-effective, and with little impact on the patient's working capacity.

SNOMED CT's Problem List

Dystrophic Upper Level

Ontologically oriented terminology systems should be grounded upon an upper level that introduces fundamental distinctions such as between endurants and perdurants. Endurants are those entities that exist in their entirety at any point in time, perdurants, however, are never completely present at one single moment. Examples for endurants are objects and spaces, such as organs, cells, spaces, molecules, body fluids etc. Perdurants are traumatic events, courses of diseases, surgical interventions. Functions, dysfunctions, states, and processes, are frequently classified as non-material endurants. SNOMED CT's 19 top level categories preserve the legacy of the former SNOMED axes which do not easily agree with any formal upper level ontology.

Concept Borderline Disorder

The Biomedical Informatics community has used the term "concept" to an extent that "Concept Systems" became a synonym of any ontology or knowledge engineering artifact. In the last years there has been a strive toward more terminological clarity which has challenged this tradition, arguing that the term "concept" is too ambiguous and obscures the representation of real-world entities [3,4] by ontologies.

We here use the term "concept" as a synonym for the nodes in SNOMED CT – as SNOMED CT does, but we normally refine it by talking about "classes", "meta-classes" and "individuals", which is – in our opinion – the most neutral parlance. Problematic under ontology scrutiny is that the wording of numerous SNOMED CT concepts suggests a meta-class interpretation, i.e. classes which classify classes, not individual entities of the world. Examples are "*SNOMED CT concept"*, "*Navigational concept", and "Additional values"*. This causes strange conclusions such as that *London* is an *Additional value*; my particular *Adverse reaction to premedication* is a *Navigational concept*; and my particular *Heartburn* is a *SNOMED CT concept*. Unfortunately – for lack of documentation – it remains open, whether these examples should be interpreted as normal classes with sloppy labels, or indeed as meta-classes.

Infestation by Individuals

There is no principle objection to why an ontology should not represent individuals, well distinguished from classes. There is, however, a broad consensus that the subsumption relation between two kinds, types or classes (*subtype-of* or *is-a*) is completely different from the instantiation relation between an individual and its type or class (*instance-of*). There are numerous SNOMED CT concepts that stand for individual entities (such as *Europe*, *Greater London*, *Binge eating scale*), i.e. instances of the class *Geographic Location*. However, an *instance-of* relation in missing in SNOMED CT

Relation Idiosyncrasy

Relations should be consistent and unambiguous in order to assist ontology developers and users in avoiding errors [5]. SNOMED CT does not formally define its relations, and by their names they can rarely be mapped to other ontologies. Some relations such as *Finding Site/Procedure Site* or *Specimen Substance* obviously specialize standard relations (e.g. *has-location, has-part*) which, on their part, are missing in SNOMED CT. Finally, there are a number of utterly fuzzy relations such as *Subject Relationship Context*. The problem here is that the more relations exist, the less one can expect agreement among users. Last but not least, there is the ontologically obscure *Relationship group* which was found to correspond to *has-part* between perdurants in most cases but which can also have other hidden meanings [6].

Taxonomic Dystrophy

Subclass hierarchies (taxonomies) should obey certain principles such as described in [7]. Accordingly, it must be criticized that numerous non-terminal SNOMED CT classes have one subclass only and that the number of classes with multiple parents is higher than necessary. Another kind of taxonomic dystrophy is the so-called "*is-a overloading"*, i.e., the use of taxonomic subsumption in order to express roles rather than generic properties, as already analyzed by [8], using the OntoClean methodology [9]. Sadly, this and other problems found nearly three years ago have not been fixed since then. An example for this is that *Bacterium* is subsumed by *Infectious Agent*: Not every individual bacterium is an infective agent, this is rather a role that can be played by bacteria under certain circumstances. Finally, as pointed out by [10], epistemological aspects should be kept apart from a clinical terminology. In SNOMED CT, there are numerous cases for "epistemology intrusion" such as *Newly diagnosed diabetes*. The point here is that the diabetes as such is no way of a different type by the fact that is has recently been diagnosed.

SEP Implants

So-called SEP triplets [11] are modeling artifacts which expand taxonomies by reified relations. For instance, $Femur_P$ is defined as the class of everything that is part of a *Femur*. $Femur_S$ is introduced as a common taxonomic parent of $Femur_P$ and *Femur*. Together, *Femur*, $Femur_P$ and $Femur_S$ form an SEP triplet. Such structures allow one to express part-whole relationships without explicitly using partitive relations. The main reason why SNOMED CT uses such structures is to enable the propagation of attributes along the aggregation hierarchy (part-whole relationships) in a parsimonious way. E.g., *Femur fracture* defined as a *Fracture* located at some $Femur_S$. Since $Femur_S$ subsumes *Neck of femur*, *Fracture of the neck of femur* is classified as a *Femur fracture*. SNOMED CT is replete of such reified classes, yet in an unsystematic and incomplete way. They are undefined and the terms assigned to them are often misleading. More precisely, we can consider taxonomic A_S-B_S links as kind of prostheses for missing *part-of* relations in the anatomy branch. However, they serve the needs of attribute propagation which is seen as an important asset in medical terminological reasoning.

Partition Agenesis

Partitioning means that sibling classes are declared mutually disjoint, as it can be assumed with the 19 SNOMED CT top-level categories. On lower levels, sibling overlap is a general

phenomenon that gives rise to complex taxonomic graphs, making the maintenance of the ontology difficult.

Description Asthenia

As advocated by [12, 13] for biomedical ontologies, taxonomies should be founded upon the Aristotelian principle of *genus* (the common properties of members in the subsuming class) and *differentiae* (the properties that distinguish each instance of the subsumed class from the genus). According to [7], half of the classes are primitive ones, i.e. they have no criterion which distinguishes them from their super-classes. Besides cases in which Aristotelian definitions are difficult or impossible (e.g. in anatomy), other ones are obviously defined as primitive in order to obviate undesired inferences, such as classifying *amputation of the foot* as a kind of *amputation of the lower limb*. In other cases, there is no obvious reason for primitive definitions: For instance, *severe asthma* could be fully defined as *asthma* (genus), characterized by the value *severe* of the attribute *severity* (differentia).

The Qualifier Syndrome

SNOMED CT qualifiers, such as *laterality*, *severity*, *onset* and *course* are relations used for constraining post-coordination for a further refinement of a class [14]. For example, *asthma* allows 12 different values for the qualifier *course* and six for the qualifier *severity*. Only a small subset of all SNOMED CT relations are used as qualifiers, and it seems that these relations are never used for different purposes. On the other hand, those relations which are used in definitions never feature as qualifiers. So we have the strange situation in which the qualifier *severity* is allowed for the class *asthma*, but is not used for defining its subclass *severe asthma*. For the latter one, *severity* is allowed with its whole range of values, so that the formation of a post-coordinated class *severe asthma* with *severity* = *mild* is possible. There are innumerous examples that show that the value ranges of the qualifiers are not well adapted to the characteristics of the class they belong to.

SNOMED CT's Treatment Plan

The assessment of SNOMED CT's health status by both specialties has revealed the following trade-off: Ontologists strive for a comprehensive account of reality, they may use the whole inventory of logics for describing it with the precision and expressiveness they deem adequate. In contrast, logicians point at the computational properties of full logics, which are prohibitive for any large scale implementation, let alone the issues of modeling and maintenance expenses. A good compromise is given by description logics (DLs), a family of decidable fragments of the first-order logic, which have a clean and intuitive syntax (without the need for free variables), cf. [15]. The description logics implementations available now as well as in future only cover part of this expressiveness. This results in an only partial ability of the system to perform the inferences which should ideally be expected. It is therefore not desirable (and not even feasible from a practical point of view) to use the whole logical machinery for describing entities in a large terminology system. As a result of these underspecifications, one has to be aware of unintended models, but this is unavoidable. However, description logics are relatively well studied, and computationally cheap dialects can be tailored to the actual needs of the ontology under scrutiny. We here sketch the specification of a logic which seems to be well suited to support most modeling and reasoning requirements of SNOMED CT.

It is the *computationally tractable* Description Logic $\mathbf{EL}^{++}$ [16], which is the underlying DL **EL** (conjunction and existential quantification) of SNOMED CT enhanced by general class inclusions (GCIs), complex role inclusions (CRIs), the empty class $\bot$, nominals (individuals as classes), restricted concrete domains, and ABox (extensional knowledge about individuals). To give an example, the **EL** class expression $\mathit{Inflammation} \sqcap \exists \mathit{has\text{-}location}.\mathit{Appendix}$ denotes the set of all individuals belonging to the class *Inflammation* and relating via the relation *has-location* to some individual of the class *Appendix*. The example constitutes both necessary and sufficient conditions for the class *Appendicitis* and thus can be used as its definition. Not only can GCIs be used to add supplement constraints that are beyond capacity of definitions, together with $\bot$, class disjointness can also be expressed ($\mathit{BodyPart} \sqcap \mathit{ClinicalFinding} \sqsubseteq \bot$). Briefly speaking, CRIs make feasible several ontologically important constraints on roles, including role hierarchy ($\mathit{proper\text{-}part\text{-}of} \sqsubseteq \mathit{part\text{-}of}$), role reflexivity ($\varepsilon \sqsubseteq \mathit{part\text{-}of}$), role transitivity ($\mathit{part\text{-}of} \circ \mathit{part\text{-}of} \sqsubseteq \mathit{part\text{-}of}$), and right identity on role ($\mathit{has\text{-}location} \circ \mathit{part\text{-}of} \sqsubseteq \mathit{has\text{-}location}$). Concrete domains help connect classes in the medical domain to values in concrete domains such as numbers and strings. A simple example is $\mathit{Minor} \equiv \mathit{Person} \sqcap <_{18\mathrm{year}}(\mathit{age})$, i.e., a minor is defined to be a person with less than 18 years of age. Considering the scale of SNOMED CT, it is inevitable to pay attention to the complexity of the logic employed. It has been shown both theoretically [16] and empirically [17] that the **EL** DL family is computationally cheap and adequate in terms of ontological expressive power.

Rectification of the Upper Level

There are several proposals of upper level ontologies, such as BFO, DOLCE, GFO, and SUMO, which roughly coincide in their upper level categories. The ontologists' recommendation is to refer as much as possible to a commonly accepted upper ontology. From the logicians' point of view, the choice is of minor importance. A major preoccupation is, however, that the upper level be expressible in terms of the logics supported. Note that with regard to upper-level organization there are still several controversial points, first of all the ontological account for disease (delimited from *courses of disease*, but also from *sign* and *symptom*). This is currently subject to ontological inquiry, and SNOMED CT could be a good testbed for this.

Isolation of Meta-classes

The mentioned borderline problems derive from the fact that most SNOMED CT concepts coincide with classes, describing real-world entities, whereas a minor part corresponds to meta-classes, defining and describing the proper SNOMED CT terminology. These two aspects must be kept apart. However, the proposed logic does not properly support a meta-class level for reasoning. This is not really a problem because we

see the necessity of meta-classes more as a kind of housekeeping feature for which annotation functions such as in Protegé and OWL can be used and in which additional *rdf* attributes can be introduced.

Increasing Tolerance of Individuals

Geographic locations, scales, etc., which are individual entities, should be related as individuals to the corresponding classes and could be related via a certain relation to other individuals. As mentioned, **EL^{++}** supports nominals and ABoxes. Both are closely related and are helpful when information about individuals is to be included. For instance, the Siamese cat could be specified to have been originated in Thailand, with Thailand a nominal, not a class ($SiameseCat \sqsubseteq \exists country\text{-}of\text{-}origin.\{Thailand\}$). An ABox comprises assertions about individuals by means of *instance-of* (concept assertions) or *related-by* (role assertions) relations, e.g., $London \in GeographicLocation$ and $(London, England) \in has\text{-}location$, respectively. To keep it simpler without the need of introducing individuals, we could limit ourselves to the reference to geographical entities in terms of location classes. In this case, reifications of the type $A_L \equiv \exists has\text{-}location.A$ with $A \sqsubseteq GeographicLocation$ may be discussed for the sake of parsimony. Then, for instance, the fact that London is located in England could be indirectly expressed by $London_L \sqsubseteq England_L$.

Reconstruction of Relations

SNOMED CT relations should be reduced to a minimum of canonical ones, starting with the OBO relations [5]. Relationship groups should be substituted by the corresponding relation, most likely *has-part*. One should also give a clear account of the algebraic features of each relation in terms of reflexivity, symmetry, and transitivity. Furthermore, relations should be further defined in terms of domain ($\exists r.\top \sqsubseteq D$) and range ($\top \sqsubseteq \forall r.R$)[1] restrictions.

Cleansing of Taxonomies

All classes should have at least one sibling, otherwise they should be merged with their super-class. Multiple inheritance should be reduced to a minimum. Wherever an *is-a* link is inferable from defined classes it should be omitted for the sake of brevity and clarity. For instance, *Acute type B viral hepatitis* is fully defined as *Type B viral hepatitis* which is *acute*. By the definitions of *Type B viral hepatitis*, *Hepatitis*, and *Acute hepatitis* a DL classifier can compute the subsumption between *Acute type B viral hepatitis* and *Acute hepatitis*, so that this *is-a* link does not have to be included.

SEP Explanation and Substitution

The extra nodes should be fully defined. Although irrelevant under ontological scrutiny, they may be preserved for reasons of backward compatibility. A full definition of S and P nodes, however, requires distinguishing between *proper-part-of* which is transitive and irreflexive, and the broader relation *part-of* which is transitive and reflexive.

So we can fully define $A_P \equiv \exists proper\text{-}part\text{-}of.A$, together with $A_S \equiv \exists part\text{-}of.A$. However, our language does not allow to enforce irreflexivity of a relation, so that the following GCI might be added where required: $\exists proper\text{-}part\text{-}of.A \sqcap A \sqsubseteq \bot$. With the right identity rule $has\text{-}location \circ part\text{-}of \sqsubseteq has\text{-}location$ we then get the right inference in the femur example. False inferences, such as the classification of an *amputation of the foot* as an *amputation of the lower limb*, can then be prevented by introducing a subrelation of *has-location*, *viz*. *has-exact-location* for which the right identity rule dos not apply.

Taxonomy Partitioning Operation

Although we do not advocate pure single hierarchies as an ontological engineering dogma, we nevertheless recommend the use of one classifying principle for all subclasses of every class as far as possible. This will yield clear partitions at each level, which helps maintain and better understand the terminology. Partitioning a taxonomy generally requires negation statements of the form $A \sqsubseteq \neg B$. Such a restricted negation statement is in fact equivalent to a disjointness axiom of the form $A \sqcap B \sqsubseteq \bot$ which is available in our DL dialect.

Revitalization of Full Definitions

We suggest the revision of primitive SNOMED CT classes, especially the elimination of misspecifications that are, obviously, the reasons that SNOMED CT classes which could be fully defined, are still kept as primitive ones. Where possible, full definitions should be introduced. The introduction of full definitions generally brings to light hidden misspecifications as soon as the ontology is classified. The use of a terminological classifier is therefore of utmost heuristic importance in the process of building and maintaining SNOMED CT. This requires, however, that SNOMED CT moves to the DL format as the primary format in which all editing is performed.

Qualifier Transplant

The realm of qualifiers which is kept somewhat apart from the rest of SNOMED CT sheds light on an intricate problem which complicates the move from a database of frame-like view towards description logics. Whereas the former systems assume a closed world, description logics assume an open world. As a practical consequence, this means that once there are relation types and classes, any relation may be asserted between any individuals unless this is explicitly precluded. SNOMED CT's approach of providing qualifiers with well-defined value restrictions for controlling the building of post-coordinated classes in description logics extends the capabilities of the logics we use. Strictly spoken, the job of description logics is to define classes by logic descriptions but not to provide constraints for the definition of new classes or the handling of individuals. There are in principle two different ways we can deal with this problem. Firstly, we can handle the constraints as provided by the qualifiers outside description logics. This would mean that we abandon the goal of proposing that the whole content of SNOMED CT be expressed in description logics. The second way is to resort to a more

[1] A controlled use of the universal quantifier $\forall$ in these cases has no negative impact on the computational properties.

expressive DL dialect, at the price of performance of the implementations. As a possible way out of this dilemma we suggest the following. On the one hand, we maintain the **EL** specification for SNOMED CT class definitions, but on the other hand we add an additional layer using DL value constraints. This second layer would then be invisible for the DL reasoner, but it can be used as a resource for those applications in which this information is needed, e.g. to constrain data entry by adaptive pick lists etc., which had been the main rationale for the SNOMED CT qualifiers. Similar to what for meta-classes, this kind of housekeeping information can be realized with the help of annotation functions.

Conclusion

We have subjected the current version of SNOMED CT to an in-depth diagnostic examination under the aspects of ontology and logic. SNOMED CT's clinical picture exhibits mostly chronic problems most of which can be treated in a conservative but yet determined fashion. Some of the problems require a more invasive intervention. We recommend the elaboration of a treatment plan, the definition of priorities, and the allocation of resources. Altogether, the cost of this treatment will be considerable, and it requires specialists from the field of ontology and description logics; nonetheless, it is a good investment for assuring the SNOMED CT's long lasting fitness and its increasing ability to stand the upcoming challenges of medical documentation and standardization.

Acknowledgments

This work was supported by the EU Network of Excellence Semantic Interoperability and Data Mining in Biomedicine (NoE 507505). Additionally, the first author was supported by a research fellowship (550830/05-7) from the Brazilian Research Council (CNPq/Brazil).

References

[1] SNOMED CT® Technical Reference Guide – Jan. 2006, College of American Pathologists, Northfield, IL.

[2] Kuśnierczyk, W. Nontological Engineering. FOIS 2006 - Proc of the 4th Intl Conf on Formal Ontology in Information Systems; 2006. p. 39-50.

[3] Klein GO, Smith B. Concept Systems and Ontologies. [updated 2006 Oct 6; cited 2006 Dec 4]. Available from http://ontology.buffalo.edu/concepts/ConceptsandOntologies.pdf

[4] Smith, B. Beyond concepts, or: Ontology as reality representation FOIS 2004 - Proc of the 3rd Intl Conf on Formal Ontology in Information Systems; 2004. p. 73-84.

[5] Smith B, Ceusters W, Klagges B, Kцhler J, Kumar A, Lomax, J et al. Relations in biomedical ontologies, Genome Biology. 2005;6(5).

[6] Schulz S, Hanser S, Hahn U, Rogers, J. The semantics of procedures and diseases in SNOMED CT, Methods Inf Med. 2006; 45(4): 354-358.

[7] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A case study in SNOMED CT. KR-MED 2004 - Proc of the 1st Int Workshop on Formal Biomedical Knowledge Representation; 2004. p. 12-20.

[8] Spackman KA, Reynoso G.Examining SNOMED from the perspective of formal ontological principles. KR-MED 2004 - Proc of the 1st Int Workshop on Formal Biomedical Knowledge Representation; 2004. p. 72-80.

[9] Guarino N, Welty CA. An overview of OntoClean. In: Staab S, Studer R. editors. Handbook on Ontologies, Berlin: Springer; 2004. p. 151-171.

[10] Ingenerf J, Linder R. Ontological Principles Applied to Biomedical Vocabularies. FCTC 2006 - Workshop on Foundations of Clinical Terminologies and Classifications; 2006 Apr 8: Timişoara, Romania.

[11] Schulz S, Hahn U. Part-Whole Representation and Reasoning in Biomedical Ontologies, Artificial Intelligence in Medicine. 2005: 34(3), p. 179-200.

[12] Michael J, Mejino JL, Rosse C. The role of definitions in biomedical concept representation. Proc AMIA Symp; 2001. p. 463-467.

[13] Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. Proc MEDINFO; 2004. p. 444-448.

[14] Dolin RH, Spackman KA, Markwell D. Selective retrieval of pre- and post-coordinated SNOMED concepts. Proc. AMIA Symp; 2002. p. 210--214.

[15] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The Description Logic Handbook. Theory, Implementation, and Applications, Cambridge, U.K.: Cambridge University Press; 2003.

[16] Baader F, Brandt D, Lutz C. Pushing the EL Envelope. IJCAI-05 - Proc of the 19th Intl Joint Conf on Artificial Intelligence; 2005. p. 364-369.

[17] Baader F, Lutz C, Suntisrivaraporn B. CEL–A Polynomial-time Reasoner for Life Science Ontologies. IJCAR06 - Proc of the 3rd Intl Joint Conf on Automated Reasoning; 2006. p. 287-291.

Address for correspondence

Stefan Schulz, Medical Informatics, Freiburg University, Stefan-Meier-Str. 26, 79104 Freiburg, Germany, stschulz@uni-freiburg.de