

SNOMED CT and formal ontologies

Gergely Héja, György Surján

National Institute for Strategic Health Research, Budapest, Hungary

Péter Varga

Eötvös Loránd University, Budapest, Hungary

Abstract

The analysis of the structure of SNOMED CT revealed several ontological and knowledge-engineering errors. The authors propose a methodology based on the formal top-level ontology DOLCE to correct these errors.

Key words: ontology, knowledge management, SNOMED CT, DOLCE

1. Introduction

The National Institute for Strategic Health Research is evaluating SNOMED CT [1] to be used in the Hungarian health care sector. It could be used for the following purposes:

- as a common reference base for coding systems (ICD10 [2], Hungarian adaptation of ICPM [3])
- providing common concepts for enabling the interoperability of health-care information systems in Hungary
- and, in principle enabling interoperability with the health-care systems of other EU member states. It is a very complicate task, requiring common terminology.

SNOMED CT is the candidate, because it is the most comprehensive clinical terminology system.

The first purpose does not require an exhaustive list of all possible medical concepts, rather “a set of building blocks and constraints – from which concepts can be composed” [4]. Consequently, we do not want to find the exactly matching concept to a given category of the coding system (which is practically not possible), but find those concepts which are the building blocks of this category. This approach is similar to that of SNOMED 3 [5]. Intelligent services (e.g. querying, coding support) require that the resulting conceptual system is suitable for automatic reasoning, i.e. the *is_a* hierarchy is clean, other roles (e.g. *part_of*, *acts_on*) are separated from the subsumption hierarchy, and that the complex concepts are formally defined. Since reasoning is a task of at least exponential time, the number of classes contained in the conceptual system should be kept as low as possible. It is obvious, that a system containing approximately 350 thousand concepts is not feasible for this task. Consequently, it has to possible to modularise SNOMED CT, by creating a core model, and extensions.

The support of interoperability instead requires an exhaustive concept list. The cause is that the current health-care systems typically use the concepts to fill in CVs (code value), and not to build expressions from them. Nevertheless modern health-care standards (such as HL7 or prEN 13606:2004) permit both approaches. There are two levels of interoperability: functional and semantic [6].

- functional interoperability: the ability of two or more systems to exchange information (so that it is human readable by the receiver)

- semantic interoperability: the ability for information shared by systems to be understood at the level of formally defined entities, so that the receiving system can process the information effectively and safely. It is essential for value-added EHR clinical applications (e.g. decision support) by reusing information.

Functional interoperability does not require a common terminology, only a common data model. Semantic interoperability is typically achieved by archetypes [7] and common terminology. In principle this terminology does not need to be an axiomatised system, however if any intelligent service (such as searching, classification, decision support) is needed, then the used conceptual system need to be suitable for automatic reasoning. However reasoning about a system containing several hundred thousand concepts is a problem not yet practically solved. It turns out that conceptual systems used for automated reasoning require different structure with less emphasis on coverage (the multitude of non-defined leaf categories) and greater emphasis on rich and well-organized high-level categories. Whilst SNOMED CT exhibits a proven excellence with respect to coverage, it is to be investigated whether it suffices the another criterion.

2. Material and Methods

2.1. DOLCE

The DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) a descriptive upper-level ontology is especially designed for ontology cleaning and interoperability [8]. It has taken several notions from OntoClean [9] which is a methodology created for ontology cleaning. As contrasted to other efforts aiming at the same tasks, such as SUO [10], it is based on sound theoretical grounds which made it well suited for our purposes.

As reflected by its acronym, DOLCE has a clear *cognitive bias*, in the sense that it aims at capturing the ontological categories underlying natural language and human common sense. Its authors do not commit to a strictly referentialist metaphysics related to the intrinsic nature of the world: rather, the introduced categories are thought of as cognitive artifacts ultimately depending on human perception, cultural imprints and social conventions. In other words it adopts a descriptive stance instead of the revisionistic attitude of scientific explanation. Consequently it is designed to suit not only scientific terminologies but also the “folk ontology”, i.e. the conceptual scheme of a given fragment of a given natural language.

It is an ontology of *particulars*, in the sense that its domain of discourse is restricted to them. Naturally there are universals in an ontology of particulars, in as much as they are used to organise them, but since they are not in the domain of discourse, they are not themselves subject to being organised.

DOLCE uses the so-called *multiplicative approach* which means that different entities can be *co-located* in the same space-time. The classical example is that of the vase and the amount of clay: the vase is constituted by an amount of clay.

It includes a set of achievements from philosophical conceptual analysis in the past century. Although the members of the set are not exempt from any debate (an unlikely situation in philosophy), they are usually agreed upon and definitely they possess a high explanative potential. Example of such achievements is the distinction of colour as something perceived (the inner mental state of perceiving a colour) from colour as such (abstract colour quality as distinct from light waves), the so-called qualia-theory. As the example shows, general acceptance does not necessary mean triviality.

DOLCE is based on the distinction between *endurants* (also called *continuants*) and *perdurants* (also called *occurents*). Endurants are characterized as entities that are ‘in time’, they are ‘wholly’ present (all their proper parts are present) at any time of their existence. On the other hand, *perdurants* are entities that ‘happen in time’, they extend in time by accumulating different ‘temporal parts’, so that, at any time *t* at which they exist, only their temporal parts at *t* are present.

DOLCE is augmented by ontology modules which together form the so called DOLCE-Plus. One of these modules, the Descriptions and Situations (DnS) is of particular importance for us, since it contains the DOLCE’s abstraction of socially constructed entities (like finding etc.) There exists a transcription of DOLCE-Plus to description logics formalism called DOLCE-Light-Plus

2.2. Analysis of SNOMED CT in the framework of DOLCE

We made our analysis on the January 2006 release of SNOMED CT using the Clue Browser 5.5. We reviewed several concepts, their parents and children in the subsumption hierarchy. SNOMED CT concept labels are in *italic*, while DOLCE category names in **bold**. The following problems have been found:

- mixing the subsumption relation with other roles (e.g. part of). For example *haemoglobin* subsumes *haemin*, in fact *haemin* is a constituent of *haemoglobin*. Likewise *exacerbation of asthma attack* is subsumed by *asthma*, rather it should be a temporal part of it.
- misplacing concepts in the hierarchy. For example *smoker* (intuitively it should be a kind of **agent**) is subsumed by *tobacco smoking behaviour – finding* (a subclass of **role**, which is a **concept**). However *smoker* and *smoker (finding)* are synonyms. It seems that the concepts contains several disjoint entities. We prefer the approach taken by the DOLCE ontology: a distinction should be made between an agent (e.g. a particular human who is smoking at this moment) and a behaviour as a part of a conceptual structure (“tobacco smoking behaviour” which can be played by many agents and which ultimately can be a subclass of finding). This is why the aforementioned extension module to DOLCE introduces the **role** class, instances of which are **played-by** instances of **agent** class. Similarly *severe asthma* is not a kind of asthma, but a kind of *asthma finding*. Likewise the parent of *polycarbonate* (which not one type of chemicals, but a group) is *polymer*, whereas it should be *synthetic polymer*.
- medical errors. *Disease*, *observation* and *finding* are subsumed by *clinical finding*, which is erroneous even from medical point of view. Disease, finding, observation and complaint are disjoint medical concepts, consequently none of them can subsume another one. Likewise *therapeutic response* is a kind of *function (observable entity)*, which is subsumed by *observable entity*. A response on a therapy should not be considered as a “built in” function of living beings. Moreover, function is not always observable entity, in that cases it should be classified as a **description**.
- categories taken form classification systems. For example, *pneumonia in other diseases classified elsewhere* is clearly taken from a classification system, since no physician would write down such an expression, consequently they should be omitted (however such concepts are marked as “ConceptStatus Limited”). Similarly *asthma causes daytime symptoms most days (finding)* is probably also taken from classification system (and it is marked as

“ConceptStatus Limited”). Moreover, if every major classification system (such as LOINC [11] and the variants of ICD and ICPM) would be added to SNOMED CT the resulting system would be too large to manage. Consider UMLS [12] as an example, which contains approximately one million concepts. Such a system is practically unmanageable.

- problems with naming. For example, *additional pathologic finding in tumor specimen (observable entity)* has a synonym *additional pathologic finding*. It is clear that the two terms have different meanings. *Inflammation (a pathological process)* has a synonym *inflammation (qualifier value)*. A process should not be considered as a qualifier.
- the intended meaning of the categories is not clear. Like in the case of *smoker*, the precise intended meaning of the concept can only be guessed from its terms and from its place in the hierarchy. However, the terms seem to denote disjoint entities, and the hierarchy is so erroneous, that the translation of certain SNOMED CT categories to another language is very debatable.
- lot of unnecessary concepts. Is it truly necessary that a medical terminology contains terms such as *mars bar* or *Kit Kat*?
- mixing ideas with real world entities. It seems that the mystery of UFOs has not been solved, *unidentified flying object device (physical object)* is classified as an *air AND/OR spacecraft*, which is subsumed by *transport vehicle*. It seems problematic that SNOMED interferes with decisions about the actual existence of entities. We prefer the above-mentioned descriptive stance of DOLCE, which does not make any such assumption and leaves it to the user to decide whether UFOs are socially constructed, mental or physical objects.
- roles (such as part of) are represented also as concepts. This approach prohibits the direct conversion to any formalism based on first-order logic, and, *a fortiori*, to any description logic formalism, such as OWL DL.
- underspecification. The other cause which prohibits the automatic translation of the obtained distribution to a DL language is that the roles are not quantified. It has to be decided whether to use existential or universal quantification.
- polihierarchy. For example, *alcoholic beverage* (through its parent *ingestible alcohol*) is subsumed by *central depressant*, *ethyl alcohol* and *psychoactive substance of abuse – non-pharmaceutical*. From a philosophical point of view, none of these subsumptions is true. Alcoholic drinks contain ethyl alcohol, which plays a role of depressant and substance of abuse. It is likely that most polihierarchies in SNOMED CT follow this pattern, consequently they should be eliminated.
- We find that the categories describing disease courses are also lacking solid ground, consider *acute on chronic* which is both subsumed by *acute* and *chronic*.

2.3. Correcting the problems

To correct these problems we advise the following methodology:

- If SNOMED CT is to be used as a computational resource, it should be totally reorganised. First of all, it should be modularised, with clear separation of the high-level ontology (core model describing medicine in general) and the domain-specific extensions (e.g. cardiology). The core model SNOMED CT should be reduced to a combinatory conceptual system like SNOMED 3, and

that system should be aligned to a formal top-level ontology (e.g. DOLCE). During the building of the core model, each category should be carefully analysed carefully to be placed in the correct place in the hierarchy.

- The core model should be applicable for formal consistency checking. That means it should be a highly axiomatised ontology. According to the principles of “good ontology building” the asserted hierarchy should be a unihierarchy. Multiple inheritance should be computed by the automatic reasoner based on the formal definitions of the categories.
- The full SNOMED CT terminology should be mapped to this core model, the computed hierarchical relations should be marked differently as the asserted relations.
- The concepts should have a natural language description to facilitate translation (an also ontological analysis).
- During the maintenance of the terminology, each new concept added to the conceptual system should be analysed, and correctly placed into the core model (if it is a fundamentally new concept), or into the extended terminology (if it is a composite entity).

3. Discussion

The ontological analysis of SNOMED CT showed so many errors that we do not consider it as a candidate for the formal core ontology of the Institute. The main reason for that is the need for formal definition of several classification systems (ICD10, Hungarian adaptation of ICPM). The formal definition of the categories of these systems enable formal consistency checking (aiding maintenance), (semi)automatic interconnection of the classification systems (e.g. creating rules between diseases and procedures) and supporting statistical analysis (by determining the appropriate categories to a query (e.g. “injuries of hand”)).

There are two possible solutions:

- Correct the errors by performing the ontology cleaning methodology described in Section 2.3.
- Devise an own core ontology, and for the sake of interoperability connect its concepts to SNOMED CT (if possible). Work in this direction is already under way, the Institute is working on a core anatomy model derived from the Foundational Model of Anatomy [13]. The other modules of the core ontology could be derived from SNOMED CT (or SNOMED 3).

However, the decision of this question requires further analysis.

4. Conclusion

The analysis of the structure of SNOMED CT revealed several ontological and knowledge-engineering errors. The authors proposed a methodology based on the formal top-level ontology DOLCE to correct these errors. Since the SNOMED CT proved to be an ontology of low quality

5. References

1. Information materials about SNOMED CT can be found at <http://www..snomed.org>
2. International Statistical Classification of Diseases and Health Related Problems, WHO, Geneva, 1992
3. International Classification of Procedures in Medicine, WHO, Geneva, 1978

4. Rogers JE, Rector AL. Extended Core model for representation of the Common Reference Model for procedures, GALEN-IN-USE project 1997
5. Coté RA, Rothwell DJ et al. (eds.). SNOMED International College of American Pathologists Northfield Illinois USA 1993
6. CEN/ISSS eHealth Standardization Focus Group "Current and future standardization issues in the e-Health domain: Achieving interoperability", draft V8.2, 2004, pp. 35-36
7. Beale T. Archetypes: Constraint-based Domain Models for Future-proof Information Systems. 2000. Available at <http://www.deepthought.com.au/it/archetypes.html>.
8. Information material about DOLCE can be found at: <http://www.loa-cnr.it/DOLCE.html>
9. Guarino N and Welty C. Evaluating Ontological Decisions with OntoClean. Communications of the ACM. 2002; 45(2): 61-65.
10. Information materials about SUO can be found at <http://suo.ieee.org/>
- 11 Information materials about LOINC can be found at <http://www.regenstrief.org/loinc/>
12. Information materials about UMLS can be found at <http://www.nlm.nih.gov/research/umls/>
13. 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