Would SNOMED CT benefit from Realism-Based Ontology Evolution?

Werner M. Ceusters, MD[1], Kent A. Spackman, MD, PhD[2], Barry Smith, PhD[1,3]
[1] Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, NY, USA
[2] Oregon Health & Science University, Portland, OR, USA
[3] Department of Philosophy, University at Buffalo, NY, USA

Abstract

If SNOMED CT is to serve as a biomedical reference terminology, then steps must be taken to ensure comparability of information formulated using successive versions. New releases are therefore shipped with a history mechanism. We assessed the adequacy of this mechanism for its treatment of the distinction between changes occurring on the side of entities in reality and changes in our understanding thereof. We found that these two types are only partially distinguished and that a more detailed study is required to propose clear recommendations for enhancement along at least the following lines: (1) explicit representation of the provenance of a class; (2) separation of the time-period during which a component is stated valid in SNOMED CT from the period it is (or has been) valid in reality, and (3) redesign of the historical relationships table to give users better assistance for recovery in case of introduced mistakes.

Introduction

SNOMED CT is a clinical reference terminology for annotating patient data designed to enable electronic clinical decision support, disease screening and enhanced patient safety.[1] It was first issued in 2002 following the merger of SNOMED-RT and Clinical Terms Version 3 (CTV3, formerly known as the Read Codes). It is structured around 'concepts', where a concept is defined as '*a clinical idea to which a unique conceptId has been assigned*'.[2] We can think of the core components of the SNOMED terminology as forming a graph structure, whose constituent nodes are joined together by *is_a* relations representing the fact that all instances of a given child concept are also instances of its parent concept. Concepts themselves are represented by the nodes of the graph, which in SNOMED CT are also called 'classes'. Such nodes are labeled with the concept identifier associated with the concept that the class represents. They are further associated with a variable number of elements such as their *relationships* to other classes and the *terms* – linked to the classes by means of *descriptions* – that can be used to refer to them by means of natural language. Whereas some terms can be used to refer to several classes (homonymy), there is always one term, called the '*fully specified name*' (FSN), which is unique, and consists of a regular name suffixed (in parentheses) with a reference to what SNOMED CT calls the '*primary hierarchy*' of the class, the latter corresponding roughly to the top-level node of the including graph.[3]

The content of SNOMED CT evolves with each release. Types of changes involving the core components include the addition or deletion of classes, descriptions, and relationships. These changes are said to be '*driven by changes in understanding of health and disease processes; introduction of new drugs, investigations, therapies and procedures; new threats to health; as well as proposals and work provided by SNOMED partners and licensees*'.[4] A history mechanism keeps track of the changes over time on the basis of the following requirements: (1) graceful evolution rather than radical change, (2) the concept represented by the class does not change, (3) classes may become inactive but are never deleted, (4) concept identifiers are persistent over time and are never reused, (5) the link between a term and a class is persistent, so that if a term is no longer appropriate to a given class, then it is inactivated, and (6) recognition of redundancies.[5]

However, although the history mechanism does capture *what* changes have been introduced over time, it usually gives no reason for *why* such changes were made, nor does it help in assessing to what extent a specific release represents an improvement over its predecessors. If, for instance, a new disease class is added at a certain time, is that because (a) the disease denoted by the class did not exist earlier, or because (b) the disease has only recently been discovered? In case (a), the two versions would be equally faithful to the part of reality they were designed to represent; in case (b), the earlier version would be marked by the unjustified absence of the class that was added later.

As SNOMED CT becomes more widely used as a reference terminology on an international scale, the need for quality assurance becomes ever more urgent. We have proposed a method for quality

assurance of ontologies and terminologies that uses reality as benchmark by keeping track of whether changes in an ontology relate to (1) changes in the underlying reality, for example through the introduction of a new drug, (2) changes in our scientific understanding, for example of the effects of a given virus, (3) reassessment of what is relevant for inclusion in an ontology, or (4) encoding mistakes resulting from ontology curation.[6] Here, we report on a study performed to assess whether SNOMED CT and its users would benefit from the application of this method.

Objectives

The purpose of our research was to assess whether the various reasons for change just sketched are indeed applicable in the context of SNOMED CT, and if so, to lay down guidelines for a more detailed study with the goal of developing recommendations for improving SNOMED CT's history mechanism in such a way that it can accommodate these reasons for change and thus support the quality assurance of the terminology in the future.

Material and methods

We used the January 2007 US version of SNOMED CT and focused our attention on changes reflected in the 'Concept Table', the 'Descriptions Table', the 'Component History Table'. A 'Historical Relationships Table' was created on the basis of the component history tables that were shipped with each new release. We performed a basic exploratory statistical analysis of the various types of changes currently recorded in SNOMED CT to uncover trends and outliers with respect to variables such as number of changes per class, types of changes kept track of, and so forth. We used this analysis to assess the size of the problem, if any, in general, and to identify patterns indicative of ontological errors. We then studied some of these cases in detail and used them to identify the nature of possible problems.

Results

Global findings

The history mechanism tracks a number of different types of status through which SNOMED CT classes may evolve. Table 1 shows the number of classes in release 2007-01-31, grouped by the types of status currently tracked. It indicates that the number of changes is very large. They result in a pool of 'useful' (i.e. active and non-limited) classes comprising 75% of the whole terminology. We found accordingly a 69% pool of useful descriptions. Where there are occurrences in the concept status table of all types of concept status listed in the SNOMED CT Technical Reference Guide[4], this is not the case for the possible values for description status. Values that are not used in the description status table include 'duplicate', 'outdated', 'erroneous', 'inappropriate', and 'moved elsewhere'. However, as shown in Table 2, there are a few description change records in the Component History Table that do include such phrases. It shows also that over the last 10 revisions 38% of the changes concerning classes were additions (48,075, CT="0") as compared to 62% (77,351, CT="1" or "2") which were class modifications, of which over 17% were cases of class retirement (21,691). The percentage of additions versus modifications varies widely from one release to another. As indicated by Table 3, the same class or description can undergo several modifications over time. Some modifications introduced at a certain time, may become invalidated at later time.

In cases where release changes do reflect a change that is external to SNOMED CT, we are left uninformed about whether the change was in reality, or in our understanding thereof. Thus we find that the class *Ehrlichia risticii* was removed from release 2007-01-31 because it was deemed to be 'outdated', which means: '*withdrawn from current use because it is no longer recognized as a valid clinical concept*'.[4] This leaves open whether it is only *now* that the concept is deemed to be no longer valid, perhaps because the corresponding species died out, or whether the concept in question was *never* valid,

ST	Concept Status	N	%
0	active in current use	281,693	75.37%
6	active with limited clinical value (classification concept or an administrative definition)	27,200	7.28%
1	inactive: 'retired' without a specified reason	6,832	1.83%
10	inactive because moved elsewhere	1,091	0.29%
2	inactive: withdrawn because duplication	40,018	10.71%
3	inactive because no longer recognized as a valid clinical concept (outdated)	1,199	0.32%
4	inactive because inherently ambiguous.	14,694	3.93%
5	inactive because found to contain a mistake	1,004	0.27%
	TOTAL	373,731	100%

Table 1: Distribution of SNOMED CT concepts' status in release 2007-01-31.

CT	ST	020731	030131	030731	040131	040731	050131	050731	060131	060731	070131	Total
0	0	7456	7765	8067	4266	4578	2588	1699	2411	2112	3029	43971
0	1	0	238	0	0	0	0	0	0	0	0	238
0	2	9	3382	28	37	103	6	3	0	0	0	3568
0	3	0	0	2	0	0	0	0	0	0	0	2
0	4	0	12	1	59	0	1	0	0	0	0	73
0	5	0	4	1	0	0	0	0	0	0	0	5
0	6	4	23	13	112	8	42	16	0	0	0	218
1	0	27	282	68	226	821	222	15	48	18	29	1756
1	1	1140	39	19	24	22	14	1	8	7	54	1328
1	10	16	885	0	0	6	50	28	87	20	0	1092
1	2	1327	1684	821	989	1262	462	233	392	322	757	8249
1	3	4	8	319	393	14	58	13	29	298	65	1201
1	4	1116	730	696	533	222	320	170	218	369	477	4851
1	5	21	290	373	66	30	79	46	32	58	56	1051
1	6	3	53	368	3866	31	5	14	4	2	24	4370
2	0	11766	7919	2175	1942	3069	903	656	8706	5785	1126	44047
2	10	0	2	0	0	0	0	0	0	0	0	2
2	2	4	4	1	0	0	12	6	4	0	0	31
2	6	1090	135	36	202	205	29	10	5960	1691	15	9373
Activated		20346	16177	10727	10614	8712	3789	2410	17129	9608	4223	103735
Inactivated		3637	7278	2261	2101	1659	1002	500	770	1074	1409	21691
Added		7469	11424	8112	4474	4689	2637	1718	2411	2112	3029	48075
Changed		16514	12031	4876	8241	5682	2154	1192	15488	8570	2603	77351
Total		23983	23455	12988	12715	10371	4791	2910	17899	10682	5632	125426

Table 2: Changes related to SNOMED concepts from version 2002-07-31 to version 2007-01-03 as listed in the Component History Table. CT= Change Type (0=added, 1=status change, 2=minor change); ST= status type

because a corresponding species never existed at all but that it is only now that science has come to this insight. The additional information that we find for this case in the Historical Relationships Table does not add more clarity. We learn that the outdated class was 'replaced by' the newly added class '*Neorickettsia risticii'*, but, as for all classes added, no reason is given. Is this a new species that evolved from the former one? Is it a species that has long existed already, but has just been discovered? It is only on the basis of external information not distributed as part of the new release that we learn about a recommendation issued in 2001 to reclassify the genus of the organism because the previous classification of its genus was flawed.

One would legitimately be surprised to find that in release 2003-01-31, 238 classes were *added* (not *changed*) with the status 'retired', i.e. inactivated without any specified reason, 3382 *added* as 'duplicate', 12 as 'ambiguous', and even 4 as 'erroneous' (Table 2). We found in total 5402 of such classes over the entire SNOMED CT history. An example is the class with FSN '*Green peppercorn RAST test (procedure)*' added as 'duplicate' in 2004-07-31 and declared 'same as' '*Piper nigrum (unripe seed) specific IgE antibody measurement (procedure)*' which was added in 2003-01-31 and has among its associated terms '*Green peppercorn RAST test*'. The reason for these strange additions turns out to be that, during the first few releases of SNOMED CT, the UK was continuing to make updates to the 4-byte and 5-byte versions of the Read codes. New additions to these necessarily involved 4-byte Read code identifiers, and different 5-byte Read code identifiers. These were added to the SNOMED CT tables to maintain 100% inclusion of all the Read codes ever issued. As a result, some new rows in the Concepts table were (intentionally) duplicates, or they were ambiguous, from the start.

Case study: saquinavir

An extreme case is class 324847008 with the current FSN '*Saquinavir (free base) 200mg capsule (product)*' which underwent 8 modifications. Although being a true outlier, it is an interesting case to demonstrate the various types of changes introduced. This class, referred to in what follows as C1, was introduced in the first version of SNOMED CT – it was not present in either SNOMED-RT or CTV3 –, and was initially associated with the FSN '*Saquinavir (free base) 200mg capsule (substance)*' and the preferred term'*Saquinavir (free base) 200mg capsule*'. This original FSN was 'retired without any specified reason' in release 2002-07-31 and replaced

by '*Saquinavir (free base) 200mg capsule (product)*'. The same type of modification was applied to all other substances. This explains the high number of changes in release 2002-07-31, in which there were a number of minor concept changes without concept retirement (11,766, see Table 2). In 2003-01-31, C1 was declared inactivated because of duplication with another class and as a consequence, its associated terms were annotated as being descriptions for a retired class. The duplicating class was '*Saquinavir mesylate 200mg capsule (product)*' (C2) which had been earlier added in release 2002-07-31. At that time, it was also noticed that a third class (C3), with the FSN '*Saquinavir (as mesylate) 200mg capsule (product)*', had been already included in the first version of SNOMED CT (again without a prior appearance in either SNOMED-RT or CTV3). This class, too, was rendered inactive as a duplicate of C2.

With release 2004-01-31, the SNOMED CT authors changed their minds. They re-activated C3 while deactivating C2, thereby still declaring that both are representations of the same concept by means of the SAME AS relation in the Historical Relationships Table. At the same time they reactivated C1, thereby deeming it to be no longer a duplicate of C2. From then on, C1 started to lead a life of its own. It became deactivated once again in 2004-07-31, being considered a duplicate of C3. It was reactivated in 2005-01-31, and deactivated (for the third time) in 2005-07-31, thereby again being declared a duplicate of C3, whose FSN in 2004-07-31 was changed to '*Saquinavir mesylate 200mg capsule (product)*', surprisingly (or not?) the very same FSN which was assigned to C2, which had been retired in 2004-01-31.

In 2007-01-31, a new class, with conceptId 422836001, (C4) was added to SNOMED CT, and was given the FSN '*Saquinavir mesylate 200mg capsule (product)*'! C2, still deactivated because of duplication, was then 'replaced' by C4 with the motivation that it (C2) contains an error, while C3was also deactivated and 'replaced' by C4 for the same reason. At the same time, the class (C5) with FSN '*Saquinavir 200mg capsule (product)*' which was incorporated in the first version of SNOMED CT as an active class – although with a status of having 'limited' clinical value – through the integration of CTV3, was inactivated for being 'ambiguous', and accordingly further annotated in the Historical Relationship Table as being 'may be a' C4 and 'may be a' C1.

Mods	Classes		Descriptions	
N	N	%	N	%
1	230,738	61.74%	869,736	83.40%
2	120,913	32.35%	158,488	15.20%
3	19,972	5.34%	12,871	1.23%
4	2,030	0.54%	1,728	0.17%
5	74	0.02%	43	0.00%
6	3	0.00%	3	0.00%
7	0	0.00%	2	0.00%
8	1	0.00%	0	0.00%
Total	373,731	100%	1,042,871	100%

Table 3: Distribution of the number of classes and descriptions according to the number of modifications they underwent over time.

Discussion

There are at least three use cases that justify the existence of SNOMED CT's history mechanism. One is the support that it can give to users who want to update data annotated with clinical codes from a previous version to conform to codes from the latest release. The Historical Relationships Table allows this to be done automatically for the relationships 'same as' and 'replaced by', while it can generate triggers for classes that have been found to be ambiguous. A second use case is internal quality control: awareness of past mistakes may prevent SNOMED CT authors from making similar mistakes in the future. The third use case relates to SNOMED CT's ambition of being a '*reference terminology for clinical data*', defined as '*a set of concepts and relationships that provides a common reference point for comparison and aggregation of data about the entire health care process, recorded by multiple different individuals, systems, or institutions*'.[7] To serve as such a common reference point, a terminology should faithfully capture the state of the art in the domain which it is intended to serve. Several studies have shown that (static) releases of SNOMED CT perform well in terms of *coverage* of the biomedical domain,[8,9] but there has also been criticism of the way in which (like other terminology systems) it runs together (1) what is the case in the domain, (2) what clinicians believe and (3) what clinicians communicate[10,11]. We have argued that these distinctions should be made more explicit. It is now clear that such criticisms can also be applied to SNOMED CT's history mechanism. Though the analysis we performed is not as yet complete, we accumulated sufficient evidence to show that changes in successive versions of SNOMED CT were often driven neither by changes believed to have occurred in the corresponding part of reality nor by changes in our scientific understanding of that part of reality to which the given SNOMED CT classes are supposed to refer. Activating and deactivating the C1 class

('*Saquinavir (free base) 200mg capsule (product)*') had nothing to do with the appearance or disappearance of that product from the market. Both saquinavir mesylate (Invirase) and saquinavir (Fortovase) were already approved by the FDA (on December 6, 1995, and November 7, 1997, respectively) as antiretroviral protease inhibitors that act by blocking a protein that HIV needs to replicate itself.[12] This activation and deactivation had nothing to do, either, with any change in our scientific understanding of these protease inhibitors. Users of SNOMED are left to attempt to infer from insufficient information what the motivation might have been not only for the changes mentioned but for a wide variety of other sorts of changes, including all additions of classes to SNOMED CT.

Conclusion

Clearly, the history mechanism implemented in SNOMED CT provides insight into how the system has evolved over time, giving information primarily about the sorts of actions its authors undertook in the course of time. It is an interesting resource into which a great deal of effort has been invested, but to the best of our knowledge it has not thus far been acknowledged in the literature or been the subject of research. However, we conclude that this mechanism in its current form does not do justice to needs of SNOMED CT as a reference terminology of international scope. Our study conducted thus far is sufficient to show that there is indeed a problem with the sorts of interpretations that can be given to stated changes, and that the nature of the problem is, at least in part, a running together of what is the case and what is believed to be the case, which we have shown to be accompanied with the concept orientation in terminology development. At this stage of the study, it is too early to make detailed recommendations on how the current mechanism might be improved without impinging negatively on the work that has been done. But we believe that benefit can be gained at least by adding mechanisms (1) to represent the provenance of a class more explicitly; (2) to separate the time-period during which a component is believed to have been valid in SNOMED CT from the period it is believed to be (or has been) valid in reality since the latest release; (3) to redesign the historical relationships table in such a way that it serves to provide assistance for recovery for example to users who have employed codes later found to contain errors.

References

1. Donnelly K. SNOMED CT: The Advanced Terminology and Coding System for eHealth. In: Bos L, Roa L, Yogesan K, O'Connell B, Marsh A, Blobel B, eds. Studies in Health Technology and Informatics - Medical and Care Compunetics 3. Vol 121: IOS Press; 2006:279 - 290.
2. College of American Pathologists. SNOMED Clinical Terms® Guide - Abstract Logical Models and Representational Forms - V5. 2006.
3. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. Artificial Intelligence in Medicine 2007;39(3):183-195.
4. College of American Pathologists. SNOMED CT® Technical Reference Guide – January 2007 Release. 2007.
5. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inform. Medicine. 1998;37:394-403.
6. Ceusters W, Smith B. A realism-based approach to the evolution of biomedical ontologies. Proc. AMIA Symp. 2006; 2006:121-125.
7. Spackman KA, Campbell KE, Côté RA. SNOMED RT: A reference terminology for health care. In: Masys DR, ed. The Emergence of Internetable Health Care: Systems that Really Work. Proc. AMIA Symp. 1997:640-644.
8. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. Proc. AMIA Symp. 2003; 2003:699-703.
9. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc. 2006;81(6):741-748.
10. Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: A case study in medical terminology. Formal Ontology and Information Systems; 2004:185-195.
11. Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED CT®. In: Fieschi M, Coiera E, Li Y-CJ, eds. MEDINFO. 2004, p. 482-486.
12. Anti-HIV agents. Saquinavir--switching from Fortovase to Invirase. TreatmentUpdate. April-May 2004;16(3):8-9.