Applying Evolutionary Terminology Auditing to SNOMED CT

Journal:	*AMIA 2010 Annual Symposium*
Manuscript ID:	AMIA-1358-A2009.R1
Manuscript Type:	Paper
Date Submitted by the Author:	24-Jun-2010
Complete List of Authors:	Ceusters, Werner; NYS Center of Excellence in Bioinformatics & Life Sciences, Ontology Research Group
Primary Axis Classification:	I.A.2. Ontologies

Applying Evolutionary Terminology Auditing to SNOMED CT

Werner Ceusters, MD[1]
[1] **Center of Excellence in Bioinformatics and Life Sciences, SUNY at Buffalo, NY, USA**

Abstract

Evolutionary Terminology Auditing is a technique designed to measure quality improvements of terminologies over successive versions. It uses the most recent version of a terminology as a benchmark and assumes that changes in the underlying ontology correspond to changes in either that part of reality that is covered by the terminology, or the authors' understanding – if not the 'state of the art' in general – thereof. Applied to SNOMED CT over 18 versions, it reveals that at the level of the concepts minimal improvements are obtained and that the second assumption holds for far less changes than one would expect. It is recommended that future versions of SNOMED CT provide more explicit documentation for each introduced change.

Introduction

High quality ontologies are – or at least should be – the analogue of scientific textbooks in that they contain what is believed to be the case in a scientific domain, excluding what is not known or judged irrelevant. Appropriately developed ontologies have the advantage over textbooks that their content can be understood by software agents. Like textbooks change when the state of the art (SoA) in the domain covered by them changes, so should ontologies change accordingly. Changes in ontologies are required either (1) because of changes in reality (new diseases arise – AIDS, SARS, … - bacteria become resistant, new drugs are manufactured) or (2) because scientists come to discover what is already the case for some time but was unknown or judged irrelevant thus far (biomarkers, disease pathways) or (3) realize that earlier assumptions were wrong. The degree to which an ontology corresponds to the SoA as well as to the degree to which changes in successive versions correspond to changes in the SoA are therefore important markers to measure quality objectively. [1]

Most biomedical ontologies developed thus far are released either without any versioning information at all, or with information limited to what has changed in comparison with the previous version. It is thus left unspecified in new versions why alterations have been introduced, i.e. whether there are corresponding changes in reality or in the ontology authors'

understanding or representation of reality. This hampers the re-interpretation of data annotated by means of earlier versions. Furthermore, some ontologies provide documentation to the effect that a specific class was added in a certain version at a certain time, but fail to tell us since when in history there are believed to have been instances of that class. This poses problems to annotate patient data that have been collected prior to the new release because for some classes it might be such that the corresponding entities in reality did not yet exist (long time) before the inclusion of the class, while for others, that might be the case.

Two terminologies are, at first sight, notable exceptions to this lack of documentation: the Gene Ontology and SNOMED CT. In [2] we described how Evolutionary Terminology Auditing (ETA) [3] was applied to assess the quality of the former. Here we report on our experiences with SNOMED CT.

Background

ETA, as further explained, is based on determining how successive versions of a terminology do a better job in mimicking the structure of reality. It is a novel technique of which the theoretical foundations were outlined in [3] and the potential applicability to SNOMED CT assessed in [4].

SNOMED CT is a clinical reference terminology for annotating patient data designed to enable electronic clinical decision support, disease screening and enhanced patient safety. [5] It was first issued in 2002 following the merger of SNOMED-RT and Clinical Terms Version 3. SNOMED CT is structured around a taxonomy of what is called '*concepts*', where a concept is defined as '*a clinical meaning identified by a unique numeric identifier (ConceptID) that never changes*. [6, p14] Concepts are further associated with a variable number of elements such as their *relationships* to other concepts and the *terms* – linked to the concepts by means of *descriptions*. Whereas the descriptions provide the vocabulary to talk about the concepts (or what might be instances thereof when the vocabulary is used to annotate patient data), the concepts and relationships are supposed to be a representation of what exists, and is relevant for certain purposes in biomedicine.

The content of SNOMED CT evolves with each release. Types of changes involving the core components include the addition or deletion of concepts, descriptions, and relationships. A history mechanism keeps track of these changes over time thereby adhering to the well known requirements for terminology management proposed by Cimino. [7] This history mechanism captures *what* changes have been introduced over time, and partly *why* such changes were made (table 1). However, if a new concept is added at a certain time, no information is provided whether (a) the corresponding entity did not exist earlier, or because (b) it has only recently been discovered. In case (a), the two successive 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. Changes in concept status are a small fraction compared to the introduction of new concepts.

Methods

ETA distinguishes (1) what is *inside* a terminology, i.e. *representational units* (RU) from (2) what is *part of* the first-order reality toward which the terminology is directed, thereby assuming that entities in (1) are *about* entities in (2) [8]. The current version of ETA

(table 2) is based on 17 possible configurations of match or mismatch which are divided into two groups, labeled '*P*' and '*A*', denoting respectively the presence or absence of an RU. Each group is further subdivided on the basis of whether the presence or absence of an RU in a terminology is justified ('*P+*' and '*A+*') or unjustified ('*P-*' and '*A-*'). These configurations reflect the different kinds of mismatch between what the terminology authors *believe* to exist or to be relevant, and matters of *objective* existence and *objective* relevance-to-purpose. The encoding of a belief can be either correct (R+) or incorrect, either (a) because the encoding does not refer (¬R) or (b) because it refers to a domain entity other than the one which was intended (R-). The configurations P-9 and P-10 – new with respect to [3] – both involve an RU that denotes an intended and objectively existing domain entity that, however, is already denoted by another RU in the terminology (R++). Note that changes in the RUs because of changes in the representation formalism itself do not count as what we mean by 'changes' in this paper.

This typology can be used to assess the quality of a terminology as a whole. To do so, we would have to (1) inspect each RU in the terminology to determine what match/mismatch configuration it exhibits, and (2) examine its coverage domain to see what relevant

Table 1: **number of concept changes in SNOMED CT from Release Jan 2002 to July 2009.**

CT	Existing concept made …	N	Error Type
0	active: in current use	2,010	A-1
1	inactive: 'retired' without a specified reason	1,993	P-1
2	inactive: withdrawn because duplication	9,711	P-9
3	inactive because no longer recognized as a valid clinical concept (outdated)	1,348	P-1
4	inactive because inherently ambiguous.	5,829	P-4
5	inactive because found to contain a mistake	1,204	P-1
6	active with limited clinical value (classification concept or an administrative definition)	4,461	A-1
10	inactive because moved elsewhere	14,406	P-6
11	pending move		P-6
	TOTAL	40,962	

Legend: CT: concept status as defined in SNOMED CT; N: cumulative number of changes in all versions studied; Error Type: corresponding error in previous version according to the typology described in Table 2

Table 2: Typology of expressions included in and excluded from an ontology in light of relevance and relation to external reality

	Reality		Understanding		Encoding		E
	OE	ORV	BE	BRV	Int.	Ref.	
P+1	Y	Y	Y	Y	Y	R+	0
A+1	N	-	N	-	-	-	0
A+2	Y	N	Y	N	-	-	0
P-1	N	-	Y	Y	Y	¬R	3
P-2	N	-	Y	Y	N	¬R	4
P-3	N	-	Y	Y	N	R-	5
P-4	Y	Y	Y	Y	N	¬R	1
P-5	Y	Y	Y	Y	N	R-	2
P-6	Y	N	Y	Y	Y	R+	1
P-7	Y	N	Y	Y	N	¬R	2
P-8	Y	N	Y	Y	N	R-	3
P-9	Y	Y	Y	Y	Y	R++	1 2
P-10	Y	N	Y	Y	Y	R++	2
A-1	Y	Y	Y	N	-	-	1
A-2	Y	Y	N	-	-	-	1
A-3	N	-	Y	N	-	-	1
A-4	Y	N	N	-	-	-	1

Legend: OE: objective existence; ORV: objective relevance; BE: belief in existence; BRV: belief in relevance; Int.: intended encoding; Ref.: manner in which the expression refers; E: number of errors when measured against the benchmark of reality. P/A: presence/absence of term. R: encoding (see text for details)

RUs are missing. Because the magnitude of a mistake in an undesirable configuration is maximally 5, we would give each best case configuration encountered a score of 5, while each deviation there from would receive the difference between 5 and the corresponding penalty for the corresponding sort of deviant case. The total score would be the ratio of the sum of the scores obtained for each present RU, over the sum of 5 times the number of RUs present and 4 times the number of RUs missing. The latter is because all missing RUs have an error magnitude of 1, and 5-1=4. The general formula is:

$$\frac{\sum_{i=1}^{n} (5 - e_i)}{5n + 4m}$$

in which e_i stands for the magnitude of the error (if any) for a given corresponding RU, n for the number of RUs present in the terminology and m for the number of RUs unjustifiably absent.

To perform our analysis, we used the versions from SNOMED CT released between January 2002 and July 2009. Applying this methodology to a terminology the size of SNOMED CT would be an impossible task. But here, for demonstration purposes, we assume naively that with each release, its authors assume in good faith that all its constituent expressions are of the correct type: active concepts should be of type 'P+1' while inactive ones either 'A+1' or 'A+2'. The further assumption that the authors advance with each release the terminology as complete, i.e. as containing RUs designating *all* PoRs deemed relevant to SNOMED CT's purpose, does very likely not hold but adopting it allows us to use a new version as a benchmark for all previous ones, while still remaining faithful to the realist agenda.

To avoid individual inspection of each term and concept, we applied a number of principles to project a change made in each version onto an error – if any at all – in all previous versions. First, if a newly introduced RU was never inactivated, there had to be an unjustified absence in each version prior to the addition, and a justified presence starting with the version in which the addition was introduced. Second, if an RU was found to have been made inactive and this action was never undone, there was a justified absence both prior to the introduction of the corresponding RU and after it was inactivated (including the version in which the RU was made inactive), and an unjustified presence in each version that contained the RU. If a RU, made inactive previously, was found to be re-introduced, then there must have been an unjustified absence prior to the addition, a justified presence after the addition until

the RU was inactivated, again an unjustified absence after the latter change, and finally a justified presence from the point of re-introduction onwards.

For those cases in which SNOMED CT provides a reason for the change, a mapping was established as outlined in Table 1 for changes in concept status. A similar mapping was performed for changes in the status of descriptions. Inactivation of descriptions because of inactivation of the corresponding concept was considered to reflect a justifiable absence and was thus not counted as an error. Adding or removing relationships were taken into account as well, but not changes in the refinability status or their inclusion or withdrawal from a role group.

Results

The changes that SNOMED CT underwent in its core components during the period studied are enormous: 8,361,989, of which 583,292 at the level of concepts and 1,528,653 concerning descriptions (including all introductions in the first release).

Several comparative tables and statistics were generated, only few of which are presented in this communication. We also focus here primarily on changes in the concept table.

Table 3 demonstrates how the metrics just described can be used to obtain two distinct, yet closely related views. Read horizontally, the table shows for four versions how the quality of a specific version deteriorates in light of the state of the art represented in a more recent version. Read vertically, it shows how much of the state of the art in a more recent version was already accounted for in a previous version. As can be inferred from the formula of our metric and the principles for quantifying the error involved in mismatches, each version considers itself to be perfect as witnessed by the series of "100%" along the diagonal of the matrix.

Figures 2 and 3 depict these two views graphically over all versions analyzed. The trend lines marked with triangles, squares and circles correspond to changes in the concepts, descriptions and relationships respectively. The trend lines without markers depict the overall changes.

Discussion

Under the assumptions entertained, the figures seem to indicate that with respect to concepts, only small quality improvements are introduced with each new version, i.e. roughly 2% with an overall quality improvement of about 16% since 2002. This need not be a negative finding for two reasons: (1) the

Table 3: ETA-based comparison of 4 versions of SNOMED CT.

Version	RU	T0201	T0407	T0701	T0907
T0201	Concepts	100.00%	91.05%	88.17%	84.17%
	Descriptions	100.00%	85.10%	79.44%	70.10%
	Relationships	100.00%	45.87%	41.84%	35.96%
	TOTAL	100.00%	62.92%	58.83%	52.09%
T0407	Concepts		100.00%	96.51%	91.93%
	Descriptions		100.00%	92.48%	81.53%
	Relationships		100.00%	67.09%	54.44%
	TOTAL		100.00%	78.22%	66.73%
T0701	Concepts			100.00%	95.22%
	Descriptions			100.00%	88.11%
	Relationships			100.00%	70.83%
	TOTAL			100.00%	79.39%
T0907	Concepts				100.00%
	Descriptions				100.00%
	Relationships				100.00%
	TOTAL				100.00%
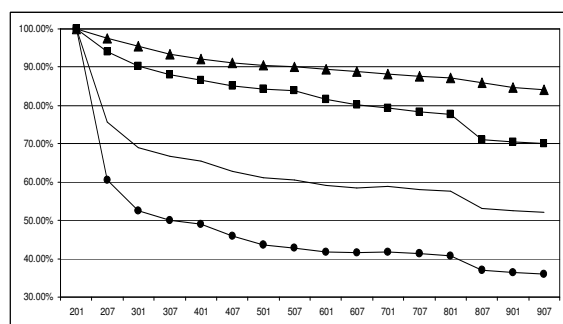
Figure 1. Evolutionary view on the Jan 2002 release of SNOMED CT since its inception ('201') until the July 2009 release ('907').
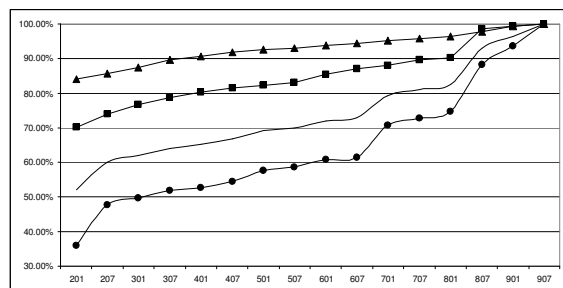
Figure 2. Evolutionary view on the relative increase in quality of SNOMED CT since its inception ('201') until the July 2009 release ('907').

proposed metric becomes less sensitive when the size of the terminology increases, and (2) it might very well be that SNOMED CT 'got it right' from the very beginning, since, after all, its real foundations were created almost 50 years ago. However, as suggested earlier, also a lack of resources to make necessary changes can be responsible. Changes in the descriptions exhibit larger improvements: 30% over the past 8 years. The biggest gains seem to be obtained in the relationships. However, several reflections need to be made.

For concepts, our analysis principles used thus far treat all new introductions as being unjustifiably missing in earlier versions. This is adequate for most types of concepts, except for pharmaceutical products – new products come on the market constantly – and certain information artifacts such as newly constructed rating scales or named guidelines and protocols: when such entities come into existence after the release of a SNOMED version, then absence of corresponding RUs in that and earlier versions is, of course, justifiable. Mistaking a justifiably absent concept for an unjustifiably present one for reasons of non-existence (P-1) makes a difference in error rate of 0 versus -3. The move of SNOMED CT to migrate brand-named products to extensions eliminates this problem, although the presence of brand-named products in versions before migration occurred needs to be judged as an unjustifiable presence for relevancy reasons (P-6, error: -1).

A second concern is the mapping between SNOMED CT's documented reasons for status changes and our reality-based interpretation. The main problem here is that the SNOMED documentation does not contain enough information on what precisely motivated its authors to introduce changes of a certain type, this on top of the fact that the status labels are rather ambiguous. Only status 'duplicate' can directly be translated into our P-9 configuration. For status changes 1, 3, 4 and 5 (Table 1) matters are less clear.

Our mapping is the best estimate that we could make on the basis of an analysis of a sample of 1000 randomly selected concepts (n=264) and descriptions (n=736) that underwent a status change of some sort, the goal of the analysis being to find some underlying principles. It turned out that all concepts with the status 'outdated' in our sample involved organisms, the change probably being introduced because of re-classification in the biology domain. We found them replaced by other concepts that nevertheless carry the original preferred name of the outdated concept as a synonym. The majority of concepts stated to be inactivated for reasons of 'ambiguity' do in our opinion not look ambiguous at all, as further witnessed by the fact that some of them have been replaced by a concept with an identical name, in addition to a more specific one. An example is 'breech extraction (procedure)' that became replaced by 'breech extraction (procedure)' and 'total breech extraction (procedure)'. If this line of thinking is to be

taken seriously, then each concept which has 'children' is ambiguous. We assume that the main reason for this state of affairs is the correction of inadequate original assignments of synonyms such as 'partial X' and 'total X' for just 'X'. We did not find any principle underlying the assignment of 'inactive, reason not specified' and 'erroneous'. For the latter case, we spotted a few typographic mistakes, an issue which has little to do with whether or not there are corresponding entities in reality. For type 1 inactivation, we spotted, for example, occurrences where an earlier inactivation for reason of duplication was changed into an inactivation for unspecified reason (e.g. 'biological test (procedure)').

For sure, the assumptions described in the methods section are not valid from one version to another and the statistics obtained need to be assessed in that light. Lack of resources might for instance prevent changes to be introduced although the authors know it has to be done at some point. Having a better insight in the concrete reasons for change, would give a more accurate application of our proposed metric. This is certainly the case for the relationships, although here further work can be done: the disappearance of a relationship in a newer version might not be a real disappearance since the relationship might still be inferred from the graph structure underlying SNOMED CT. Figuring this out, however, requires a lot of computer effort and time, a project that is still ongoing.

Conclusion

ETA answers two questions: (1) how much is a new version of a terminology better than any previous version and (2) to what degree do terminology changes reflect evolutions in the underlying domain or the terminology authors' understanding thereof. The answer to the first question, in the context of SNOMED CT, seems to be: not much, at least not for the concepts. This is in contrast to our findings on the application of ETA to the Gene Ontology for which the same assumptions were used [2]. The answer to the second question is less straightforward. Close inspection of the documented motivations for status changes and new additions which 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;…*' [9, p38] reveals that the majority of them have little to do with changes in the domain or altered understanding thereof, but rather with the idiosyncrasies of SNOMED CT's representational framework: the distinction between 'concepts' and 'terms' is far less absolute than one would expect.

Our recommendation is that the SNOMED CT authors provide for future versions greater insight into the underlying reasons for changes they introduce and that they do this in a way that supports computation. Above all, we hope that our findings lead to further introspection on the appropriateness of the concept-based approach [1] for a resource as famous as SNOMED CT, or that, at least, more attention is given to the lack of ontological commitment. [10]

Acknowledgements

References

1. Smith B. From concepts to clinical reality: An Essay on the Benchmarking of Biomedical Terminologies. Journal of Biomedical Informatics. 2006;39(3):288-98.
2. Ceusters W. Applying Evolutionary Terminology Auditing to the Gene Ontology. Journal of Biomedical Informatics. 2009;42(3):518-29.
3. Ceusters W, Smith B. A Realism-Based Approach to the Evolution of Biomedical Ontologies. *Proceedings of the 2006 AMIA Annual Symposium*. Washington DC: American Medical Informatics Association 2006:121-5.
4. Ceusters W, Spackman KA, Smith B. Would SNOMED CT benefit from Realism-Based Ontology Evolution? *Proceedings of the 2007 AMIA Annual Symposium*. Chicago IL: American Medical Informatics Association; 2007:105-9.
5. Donnelly K. SNOMED CT: the advanced terminology and coding system for eHealth. In: Bos L *et al.* eds. *Studies in Health Technology and Informatics Vol 121*. Amsterdam: IOS Press 2006:279 - 90.
6. IHTSDO. SNOMED Clinical Terms® User Guide - January 2010 International Release (US English) 2010.
7. Cimino JJ. Desiderata for controlled medical vocabularies in the 21st century. Methods of Information in Medicine. 1998;37(4-5):394-403.
8. Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a reference terminology for ontology research and development in the biomedical domain. *KR-MED 2006, Biomedical Ontology in Action*. Baltimore MD, USA 2006.
9. IHTSDO. SNOMED CT® Technical Reference Guide - Jan. 2010 International Release. 2010.
10. Schulz S, Cornet R. SNOMED CT's Ontological Commitment. In: Smith B, ed. *International Conference on Biomedical Ontology*. Buffalo NY, 2009:55-8.