Analyzing SNOMED CT's Historical Data: Pitfalls and Possibilities

Werner Ceusters, MD[1], Jonathan P. Bona, PhD[1]
[1]Department of Biomedical Informatics, University at Buffalo, Buffalo, NY

Abstract

SNOMED CT's Release Format 2 (RF2) has been announced as an improvement over its predecessor, for instance because of its more consistent and almost formal approach towards describing changes in components over different versions, as well as changes in the structure of SNOMED CT itself. We explore two sorts of changes that are only partially formalized in RF2: the relationships between associative relations and reasons for inactivations as expressed in Association Reference Sets and Attribute Value Reference Sets on the one hand, and the various patterns according to which semantic tags appearing in fully specified names change over subsequent versions with or without being related to inactivations. We propose a data conversion methodology that combines assertions about SNOMED CT components into history profiles and use elements of these profiles to build Formal Concept Analysis contexts to discover valid implications that can render implicit assumptions hidden in SNOMED CT's structure explicit.

Introduction

SNOMED CT [1], maintained by the International Health Terminology Standards Development Organization (IHTSDO) is a large healthcare terminology built around a concept-based ontology. Concepts are classified under several hierarchies, of which most of the top classes correspond roughly to the types of entities instances of which are encountered by clinicians during their work (body parts, diseases, substances, procedures, etc.) while other top classes correspond to types instantiated by descriptive elements of the SNOMED CT knowledge representation itself, for example classes denoted by terms such as '*inactive concept*', '*navigational concept'*, and '*core metadata concept'* [2]. In addition to *active components* – 'component' being the umbrella term used by IHTSDO for *concept* or *relationship* or *description* – SNOMED CT contains also *inactive components* which were active in one or more prior versions but at some point have been inactivated for one or other reason. Prior to July 2011 all releases were distributed in a format now known as '*Release Format 1 (RF1)*'. *Release Format 2* (RF2) was introduced in 2012 in order to (1) implement a more robust and consistent representation of versions in which changes are tracked in a uniform manner in the core files themselves; (2) introduce reference sets as a more easily extensible and maintainable replacement for the RF1 representations of subsets for specific uses such as not only mappings to other biomedical terminologies and classification systems, but also to improve RF1's history mechanism; and (3) create an added hierarchy to represent metadata about the structure of SNOMED CT itself [3, 4 p127].

The part of SNOMED CT that describes its own history has grown considerably over the years, as witnessed, for example, by the 105,313 inactive concepts – roughly 25% of the total concept count – which are annotated by 140,390 associations to other concepts or descriptions. For 99,489 of these inactive concepts reasons for their inactivations are provided. This comes on top of the 325,893 reasons for description-inactivations related to both active and inactive concepts. This raises the question whether the totality of assertions which are about changes in SNOMED CT rather than about external reality constitutes in and of itself a valuable resource to identify patterns that would allow the detection of mistakes in assertions about external reality that have thus far not been discovered. In other words: what can we learn about SNOMED CT's mistakes committed in the past to detect still existing mistakes and prevent new ones? This sort of quality improvement being the ultimate goal of our efforts, the work described in this paper is the first phase of this endeavor during which we explored the history information included in the RF2 distribution of the January 2016 version of SNOMED CT to identify pitfalls and possibilities that should be taken into account for the development of novel error detection methods.

Changes in SNOMED CT

The content of SNOMED CT evolves with each release. Once released, SNOMED CT components are persistent and their identifiers are not reused [4, p45]. When a component becomes inactive this is indicated by the value of the active field, a field which is present in all components. Components continue to be distributed even when they are no longer active. This allows a current release to be used to interpret data entered using an earlier release. Within RF2, all changes in components are represented in the corresponding files by adding a new row, with the same component ID, a new effective time and any necessary change in the component values. As an example, **Table 1** shows that the concept '301381004' with FSN '*Discomforting present pain (finding)*' was set to active in release 20020131 and to inactive in 20080131.

Table 1. Updates in the SNOMED CT concept file (RF2) for concept 301381004 with FSN '*Discomforting present pain (finding)*'.

conceptID	Effective Time	Active	ModuleID	Definitional Status
301381004	20020131	1	900000000000207008	900000000000074008
301381004	20080131	0	900000000000207008	900000000000074008

Legend: Active: (1) = active, (0) = inactive.

Table 2. Updates in the SNOMED CT relationships file (RF2) for the same concept 301381004

RelID	Effective Time	Active	Attribute	Target
126300024	20020131	1	Is a	Pain (finding)
126300024	20040131	0	Is a	Pain (finding)
126301023	20020131	1	Is a	Finding of present pain intensity (finding)
126301023	20080131	0	Is a	Finding of present pain intensity (finding)
657858027	20020131	1	Finding site	Structure of nervous system (body structure)
657858027	20060131	0	Finding site	Structure of nervous system (body structure)
2260209021	20030731	1	Interprets	Nervous system function (observable entity)
2260209021	20050131	0	Interprets	Nervous system function (observable entity)
2458913020	20040131	1	Is a	Discomfort (finding)
2458913020	20080131	0	Is a	Discomfort (finding)
2858465020	20060131	1	Finding site	Anatomical structure (body structure)
2858465020	20080131	0	Finding site	Anatomical structure (body structure)

Legend: RelID = Relationship identifier; Active: (1) = active, A(0) = inactive. Columns irrelevant for our purposes here are not shown. For readability, Attribute and Target identifiers have been replaced by their corresponding FSN – omitting '(attribute)' – in the most recent version studied (January 2016).

Table 2 shows that during the life time of that concept, it underwent considerable changes in its reported relationships to other concepts after full DL classification. **Table 3** demonstrates how changes in the descriptions of concepts are similarly logged. Only one description record with the same descriptionID field is current at any point in time. The current record is the one with the most recent Effective Time before or equal to the point in time under consideration. If the active field is false ('0'), then the description is inactive at that point in time. If it is true ('1'), then the description is associated with the concept identified by the conceptId field (not shown in **Table 3**).

Table 3. Updates in the SNOMED CT descriptions file (RF2) for concept '274236006'

descriptionID	Effective Time	Active	Description Type	Term
410015012	20020131	1	Synonym	Asthenia [D]
410015012	20020731	0	Synonym	Asthenia [D]
666971011	20020131	1	FSN	Asthenia [D] (finding)
666971011	20030131	0	FSN	Asthenia [D] (finding)
1237162017	20020731	1	Synonym	Asthenia [D]
1472277017	20030131	1	FSN	[D]Asthenia (context-dependent category)
1472277017	20060731	0	FSN	[D]Asthenia (context-dependent category)
1489933012	20030131	1	Synonym	[D]Asthenia
2610401019	20060731	1	FSN	[D]Asthenia (situation)

Legend: Active: (1) = active, A(0) = inactive. Columns irrelevant for our purposes here are not shown. For readability, Description Type identifiers have been replaced by their corresponding term – omitting their semantic tag '(core metadata concept)'.

RF2 replaces the 'history mechanism' implemented in RF1 [5] by means of Historical Association Reference Sets (HARS) and Component Inactivation Reference Sets (CIRS). HARSs (**Table 4**) are used to indicate, for example, which deactivated concepts are in one way or another related to other active concepts, and CIRSs (**Table 5**) to indicate the reasons for inactivating a component – such as errors, duplication of another component, and ambiguity of meaning [4, p506]. Records that express such association are called *reference set members*. The primary purpose of these reference sets is to specify which (if any) of these associations should be followed in a fashion similar to following '*Is a (attribute)*' relations when determining whether to retrieve a record entry previously coded with a concept that has since then been inactivated. Whereas 'same as' and 'replaced by' associations can be followed unproblematically, the solution for ambiguous concepts related by 'possibly equivalent to' associations is less clear-cut [4, p654].

Table 4. Historical association reference set types in SNOMED CT (modified from [4, p509])

HARS name	**Use**
Possibly equivalent to (P)	From an ambiguous concept to one or more active concepts that represents one of the possible meanings of the inactive concept.
Moved to (T)	From a component to a namespace to which the component has been moved
Moved from (F)	From a namespace to the original component Identifier in its previous namespace.
Replaced by (R)	From an erroneous or obsolete inactive component to a single active replacement component.
Same as (S)	From a duplicate component to the active component that this component duplicates.
Was a (W)	From an inactive classification concept such as "not otherwise specified" to the active concept that was formerly its most proximal supertype.
Alternative (Z)	From an inactive classification concept derived from ICD-9 Chapter XVI 'Symptoms signs and ill-defined conditions' with the most similar active concept.
Refers to	From an inactive description which is inappropriate to the concept it is directly linked to but instead should refer to the concept referenced.

Table 5. Component inactivation set types for concepts (modified from [4, p506-507])

CIRS value	**Concept status and motivation**
Duplicate (D)	inactive because it has the same meaning as another Concept
Outdated (O)	inactive because it is an outdated concept that is no longer used.
Ambiguous (A)	inactive because it is inherently ambiguous either because of an incomplete FSN or because it has several associated terms that are not regarded as synonymous or partial synonymous.
Erroneous (E)	inactive because it contains an error
Limited (L)	active prior to Jan 2010, inactive since then because of unstable meaning within SNOMED CT
moved to (M)	inactive because moved to another namespace.
Pending move	active but in the process of being moved to another namespace

Methodology

SNOMED CT undergoes changes of various sorts with each release. Most changes are recorded explicitly in the sense that there is a formal mechanism through which changes of this type are documented in one or other component of the SNOMED CT distribution files. Some types of changes, implicit ones, lack such a formal mechanism but can be retrieved through the implementation of algorithms not documented in the SNOMED CT documentation. In this paper, we report on the evolution of activations and inactivations of component instances as examples of an explicit type of change, and on the evolution of semantic tags as an implicit type of change. Both types of changes required specific data reorganization strategies. The results of these conversions were then analyzed using Formal Concept Analysis.

Data reorganization of activations and inactivations

We combined the explicit assertions about (in)activations in components and history information present in CIRSs and HARSs into one format. We kept track of in which versions assertions were made – and possibly also changed – through the construction of a *history profile*. Such profile contains for each of the 29 versions from January 2002 to January 2016 a marker indicating whether the assertion was in that version absent (A) or present, in which case it was either active (Y) or inactivated (N). Since the very same concepts can not only appear as referenced component in one HARS member and as target component in another HARS member, but also appear in members of distinct HARSs, it was possible to compute clusters of concepts by randomly selecting a concept from a HARS member and recursively collecting all reference set members in which this concept appears with the goal of processing each associated concept in the same way until no more concepts can be found. **Table 6** contains assertions that were retrieved for one such cluster composed out of 5 related concepts. By 'assertion', we mean anything that inside SNOMED CT is explicitly or implicitly stated as applying to a concept. Examples of explicit assertions are the inclusion of the concept 324253001 in the 1st version ('Y' in first position of the history profile in row 8), the inactivation of that concept in the 3rd version ('N' in 3rd position, row 8) and the reactivation of it in the 5th version.

Table 6. History profile of SNOMED CT assertions related to a cluster of 5 concepts about Azithromycin dihydrate

Row	ConceptID	Attribute	Value	History profile (one character per version)
1	324253001	Duplicate		AAYYNNNNNNNNNNNNNNNNNNNNNNNNN
3		Duplicated by	375558000	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
3		Duplicated by	375559008	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
4		Duplicated by	375948007	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
5		Duplicated by	376025007	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
6		Semantic tag	product	AYYYYYYYYYYYYYYYYYYYYYYYYYYYY
7		Semantic tag	substance	YNNNNNNNNNNNNNNNNNNNNNNNNNNNN
8		Is active		YYNNYYYYYYYYYYYYYYYYYYYYYYYYY
9		Same-as	375559008	AAYNNNNNNNNNNNNNNNNNNNNNNNNNN
10		Same-as	375948007	AAAYNNNNNNNNNNNNNNNNNNNNNNNNN
11	375558000	Duplicate		AAYYYYYYYYYYYYYYYYYYYYYYYYYYY
12		Semantic tag	product	AYYYYYYYYYYYYYYYYYYYYYYYYYYYY
13		Is active		AYNNNNNNNNNNNNNNNNNNNNNNNNNNN
14		Same-as	324253001	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
15		Same-as	375948007	AAYYNNNNNNNNNNNNNNNNNNNNNNNNN
16	375559008	Duplicate		AAAYYYYYYYYYYYYYYYYYYYYYYYYYY
17		Duplicated by	324253001	AAYNNNNNNNNNNNNNNNNNNNNNNNNNN
18		Is active		AYYNNNNNNNNNNNNNNNNNNNNNNNNNN
19		Same-as	324253001	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
20		Same-as	375948007	AAAYNNNNNNNNNNNNNNNNNNNNNNNNN
21	375948007	Duplicate		AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
22		Duplicated by	324253001	AAAYNNNNNNNNNNNNNNNNNNNNNNNNN
23		Duplicated by	375558000	AAYYNNNNNNNNNNNNNNNNNNNNNNNNN
24		Duplicated by	375559008	AAAYNNNNNNNNNNNNNNNNNNNNNNNNN
25		Duplicated by	376025007	AAYYNNNNNNNNNNNNNNNNNNNNNNNNN
26		Semantic tag	Product	AYYYYYYYYYYYYYYYYYYYYYYYYYYYY
27		Is active		AYYYNNNNNNNNNNNNNNNNNNNNNNNNN
28		Same-as	324253001	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
29	376025007	Duplicate		AAYYYYYYYYYYYYYYYYYYYYYYYYYYY
30		Semantic tag	Product	AYYYYYYYYYYYYYYYYYYYYYYYYYYYY
31		Is active		AYNNNNNNNNNNNNNNNNNNNNNNNNNNN
32		Same-as	324253001	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
33		Same-as	375948007	AAYYNNNNNNNNNNNNNNNNNNNNNNNNN

Legend: *concepts in cluster with most recent FSN:* 324253001: Azithromycin 200mg/5mL oral suspension (product), 375558000: Azithromycin dihydrate 200mg/5mL suspension (product), 375559008: Azithromycin dihydrate 200mg/ 5 mL suspension (product), 375948007: Azithromycin dihydrate 200mg/5 mL suspension (product), 376025007: Azithromycin dihydrate 200mg/5 mL powder (product). *History profile*: 'A': absent, 'Y': active, 'N': inactive.

These assertions are explicit because there are corresponding effective time and activation status assertions in the concept table. Similarly, there is in the description table the assertion that the semantic tag of this concept was in the 1st version 'substance' (row 7), while 'product' in the 2nd version (row 6). Examples of implicit assertions are, for instance, the ones to the effect that in the 1st version there was no mention of a sameness-relation between this concept and 375559008 (row 9) nor 375948007 (row 10).

Data reorganization for the evolution of semantic tags

We computed for each concept the evolution of what in SNOMED CT is called *semantic tags*. Descriptions provide for each concept a *Fully Specified Name* (FSN) most of which '*end with a semantic tag in parentheses and which* *indicates the semantic category to which the concept belongs (e.g. clinical finding, disorder, procedure, organism, person, etc.)'* [4, p41]. It is further stated that *'the semantic tag helps to disambiguate different concepts which may be referred to by the same commonly used word or phrase*' [4, p41]. For example, it is the semantic tag '*morphologic abnormality*' in the FSN '*Hematoma (morphologic abnormality)*' that disambiguates the concept to which this FSN is assigned from a second concept with FSN '*Hematoma (disorder)*'. The former is intended to be used for what '*a pathologist sees at the tissue level*', while the latter '*represents the clinical diagnosis that a clinician makes when they decide that a person has a* "*hematoma*"' [4, p41]. As can already be seen in **Table 6** (rows 6 and 7) concepts can be assigned different semantic tags over time, changes which for some, as exemplified by **Table 7** are quite dramatic.

Table 7. Examples of changes in semantic tag assignment over time

conceptID	Most recent FSN	Changes in semantic tags
66076007	Chewable tablet (qualifier value)	(substance)|(product)|(qualifier value)
66402002	Peritoneal dialysis education (procedure)	(procedure)|(regime/therapy)|(procedure)
68433009	Childhood (finding)	(function)|(observable entity)|(finding)
69736008	Vocational assessment (procedure)	(procedure)(regime/therapy)|(regime/therapy)|(procedure)
70409003	Mouthwash (qualifier value)	(substance)|(product)|(qualifier value)
70444001	Recessive gene (substance)	(function)|(observable entity)|(substance)
70790008	Absence of nausea and vomiting (situation)	(finding)|(context-dependent category)|(situation)
73669007	Kung fu (qualifier value)	(qualifier value)|(observable entity)|(qualifier value)
73905001	Sees flickering lights (finding)	(qualifier value)|(observable entity)|(finding)

Legend: semantic tags are written between brackets. '|' indicates a transition from one (or more tags) to another. History profiles are omitted in this table.

Data analysis

Exploratory statistical analyses were performed to find associations, or unexpected lack thereof, between the various sorts of assertions derived from the conversions. Specifically to mention here is Formal Concept Analysis (FCA), a mathematical theory for understanding the structure of data given as a set of objects described in terms of attributes they possess, which is done by representing the data as a concept lattice [6]. Every FCA concept – we will use explicitly the term 'FCA concept' to distinguish it from SNOMED CT concepts – has its *extent* (the set of objects that fall under the FCA concept) and its *intent* (the set of attributes that together are necessary and sufficient for an object to be an instance of the FCA concept. In [7], for instance, attributes were defined on the basis of the normal forms of pre-coordinated SNOMED CT expressions. For our analyses here, we created FCA attributes and corresponding objects on the basis of two contexts: (1) the co-occurrence of SNOMED CT HARS and CIRS attributes as described in **Table 4** and **Table 5** throughout the history of SNOMED CT concepts, and (2) the evolution of semantic tags over time. While FCA concept lattice diagrams have visualizing power when applied to domains with a small number of concepts and attributes governed by a simple organizational structure, they are rather useless in case of more complex situations as the one explored here. More useful here is the computation of *attribute implications* where an implication asserts a certain relationship between two attribute sets which are respectively called *premise* and *conclusion*: an implication is valid in the data set if every object that has all attributes from the premise of the implication also has all attributes from its conclusion. For example, if 'being mammal' and 'being vertebrate' would be attributes used to correctly describe animals, then '*being mammal* → *being vertebrate*' would be computable as being a valid implication. The set of all valid implications can be reduced to a smaller set – the Duquenne–Guigues base – from which all other implications follow semantically [8], and a set of approximate implications known as Luxenburger base [6]. The latter are valid for a specified percentage of FCA concepts; for example, for a particular zoo it could be found that 85% of the vertebrates on display are mammals. For both contexts, we assessed whether implications correspond to SNOMED CT's editorial policies in relation to inactivations and SNOMED CT's concept model.

Results and discussion

Concept ambiguity and *possibly-equivalent-to* associations

SNOMED CT's technical implementation guide [4] contains indications that certain changes in components go hand in hand with changes in HARSs and CIRSs. For instance, from the description of what it means for an inactive concept to be *possibly equivalent to* another concept (**Table 4**) one can assume that such concepts are asserted as being ambiguous in a CIRSs. We found that not to be the case for 4 concepts. On the other hand, although it is allowed for an inactive concept to be *possibly equivalent to* only one other concept – we found 7815 of such cases – it is for many of these cases hard, if not impossible, to find out, especially algorithmically through some automated procedure, why the change has been made in this specific way. It is clear that 'Pyogenic arthritis of lower leg (disorder)' is ambiguous in the sense that it does not specify in which joint specifically the arthritis is located, but then the question is why it has only been associated with 'Knee pyogenic arthritis (disorder)'. Another example is 'Distal interphalangeal joint structure of third finger (body structure)' which has been asserted as being *possibly equivalent to* another concept with exactly the same FSN. Inactivation because of ambiguity is stated to be '*because it is inherently ambiguous either because of an incomplete FSN or because it has several associated terms that are not regarded as synonymous or partial synonymous*' (**Table 5**). Since there was nothing wrong with the FSN it must have been because of the synonyms. Indeed, inspection reveals that for the inactivated concept there is the synonym 'Distal interphalangeal joint of third digit of forelimb' which is not to be found in the target concept, thus removing the ambiguity of whether the concept denotes a human body part in addition to an animal body part. This reasoning, unfortunately, does not hold for '391651001: Gluten-free/wheat-free baguette (product)' which was rendered ambiguous in the 5th release and made *possibly equivalent to* 407775004, a then newly introduced concept with exactly the same set of terms (**Table 8**). Relevant questions are (1) what motivated the SNOMED CT editor to introduce the new concept, and (2) why to use the inactivation because of ambiguity rather than because of duplication? The history mechanism is not able to provide arguments.

Table 8. Inactivation of '391651001: Gluten-free/wheat-free baguette (product)'

ConceptID	Attribute	Value	History profile (one character per version)
391651001	AMB		AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
	Is active		AAYYNNNNNNNNNNNNNNNNNNNNNNNNN
	Poss-equivalent-to	407775004	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
407775004	Semantic tag	product	AAAAYYYYYYYYYYYYYYYYYYYYYYYYY
	Is active		AAAAYYYYYYYYYYYYYYYYYYYYYYYYY

Duplication and *same-as* associations

Table 4 and **Table 5** let us reasonably assume that when a duplicate concept is discovered, it is inactivated with the reason for inactivation being stated as 'duplicate' and that at the same time a *same-as* association would be created. That is indeed the case for 44,113 inactivations. Yet, we found 1449 cases where a concept is stated to be duplicate without a corresponding same-as association, and 3453 cases in which a same-as association was created without a duplicate assertion.

We also found cases in which concepts are stated to be duplicates, yet denote clearly distinct entities. The most notorious case is the one in which '34759008: Urethral catheter, device (physical object)' is stated to be duplicated by 73 other concepts, each one of which denoting nevertheless a more precisely specified type of catheter, for example '349499005: Bard 10mL balloon 22Ch 1658 2-way all-silicone male length urethral Foley catheter' and '349501002: Bard 10mL balloon 24Ch 1265LV 2-way Teflon coated male urethral Foley catheter'. Several of these catheters are nevertheless, by means of other concepts, listed as descendants of '34759008'.

Mining implication and association rules for concept inactivations

The observation that – with one exception: whenever a 'moved to' association is asserted in a HARS, there is a corresponding 'moved to' inactivation asserted in a CIRS – there are no simple consistent relationships between reasons for inactivation and associations motivated us to mine for possible relationships using Formal Concept Analysis. In our case here, we used as FCA concepts every combination of CIRS and HARS membership encountered for all (active and inactive) SNOMED CT concepts as depicted in **Table 6**, without, however, including history profile information. We used the uppercase characters written between brackets in the first columns of **Table 4** and **Table 5** – with the exception of 'pending move' and 'refers-to' which were not included in this analysis – to name our FCA concepts. For example, the FCA concept 'DLSW' was attached to every SNOMED CT concept for which throughout its history it was at least once stated to be duplicate (D) and limited (L), as well as associated with other concepts by means of a same-as (S) and was-a (W) association. 85 such FCA concepts were found, in total covering all 424,759 active and inactive SNOMED CT concepts.

Table 9 shows the positive Duquenne-Guigues and Luxenburger base of implications between HARS and CIRS assertions. For example, the implication '< 3 > ADW ==> S' states that if a SNOMED CT concept has ever been annotated as being ambiguous, duplicate and enjoying a was-a association to some other SNOMED CT concept, then it is also the case that this concept has been annotated as having a same-as association. The '<3>' indicates that this implication corresponds to 3 of the 85 FCA concepts encountered which when calculated back to SNOMED CT concepts covers only 4 cases. As another example, the implication '< 5 > P E ==> A R' states that whenever a SNOMED CT concept has been asserted to be possibly equivalent to some other concept as well as being erroneous, it is also the case that that SNOMED CT concept has been annotated as being ambiguous and having been replaced by some other concept. This is the case for 5 FCA concepts which cover in total 15 SNOMED concepts.

The Luxenburger base implication '< 24 > P =[92%]=> < 22 > A' states that for 92% of the FCA concepts which correspond to SNOMED CT concepts that have been annotated with a possibly equivalent to association it is the case that the corresponding SNOMED CT concepts have also been annotated as being ambiguous. Similarly, '< 20 > E =[90%]=> < 18 > R' states that for 90% of the FCA concepts with an E-attribute, there is also an R attribute. When these two implications are assessed towards the SNOMED CT concepts to which the FCA concepts apply, then we find 18,413 SNOMED CT concepts under the 22 FCA concepts for which the P ==>A implication holds, and 1,397 SNOMED CT concepts under the 18 FCA concepts for which E ==> R holds. These counts are many magnitudes higher than the 15 SNOMED CT concepts for which P E ==> A R holds. This raises the question whether this huge discrepancy is indicative for something being wrong with all or some of the assertions made in relation to these 15 concepts. And it leads to the more general question whether the differences in counts observed between SNOMED CT concepts covered by Duquenne–Guigues implications in contrast to Luxenburger implications form the basis for a novel method of quality control that to the best of our knowledge has thus far not been applied to SNOMED CT.

Table 9. Partial Duquenne-Guigues and Luxenburger base of implications between HARS and CIRS assertions

Positive implications from the Duquenne–Guigues base			Luxenburger base >80%
< 19 > M ==> T;	< 3 > A D W ==> S;	< 1 > M T E ==> R;	< 24 > P =[92%]=> < 22 > A;
< 19 > T ==> M;	< 1 > R O W ==> L;	< 5 > A E ==> P R;	< 20 > E =[90%]=> < 18 > R;
< 9 > P S ==> A;	< 2 > Z ==> M T;	< 5 > P E ==> A R;	< 10 > P R =[90%]=> < 9 > A;
< 4 > A R S ==> P;	< 1 > M T A L ==> P;	< 7 > S E ==> R;	< 10 > A R =[90%]=> < 9 > P;
< 1 > M T A D ==> P;	< 8 > P L ==> A;	< 2 > A P R S E ==> D;	< 9 > A L =[89%]=> < 8 > P;
< 3 > A R D ==> P S;	< 2 > R S L ==> D;	< 8 > D E ==> R;	< 9 > P D =[89%]=> < 8 > A;
< 3 > P R D ==> A S;	< 9 > D L ==> S;	< 2 > O E ==> R;	< 7 > W E =[86%]=> < 6 > R;
< 2 > P O ==> A;	< 1 > R O L ==> W;	< 1 > R S O E ==> D;	< 7 > R S E =[86%]=> < 6 > D;
< 1 > A R O ==> P;	< 1 > O W L ==> R;	< 5 > L E ==> R;	< 6 > A W =[83%]=> < 5 > L;
< 3 > D O ==> S;	< 1 > S F ==> D;	< 2 > A R W ==> P E;	< 6 > A W =[83%]=> < 5 > P;
< 2 > M T W ==> S L;	< 1 > D F ==> S;	< 2 > R S W ==> D E;	< 11 > A S =[82%]=> < 9 > P;
< 5 > P W ==> A;	< 2 > A R L ==> P E;	< 3 > R D W ==> E;	< 5 > A W L =[80%]=> < 4 > P;

Semantic tag evolutions

We found in total 285 patterns of the sort exemplified in **Table 7** according to which SNOMED CT concepts underwent changes in the semantic tags assigned to them. A change from no semantic tag at all to a semantic tag (43 patterns) counted – under one perspective – also as a change. There were no patterns with more than 3 changes over time under either perspective. Changes in semantic tags can happen for a number of reasons. One is a change in SNOMED CT's concept model, for instance when distinctions are made that didn't exist in earlier versions, or different interpretations were introduced (e.g. the product / substance distinction). Such changes have a global impact on large parts of the ontology. Another reason is that concepts were in one or other way erroneous and had to be corrected. The SNOMED CT documentation states for instance that '*only limited changes may be made to the* "*term*" *field, as defined by editorial rules*' [4, p145]. This is consistent with the view that '*the meaning of a concept can be determined […] from associated descriptions that include human readable terms*' [4, p87]. This editorial rule is also used as argument for not retiring the concept to which it is attached in cases where the FSN undergoes minor changes. Indeed, '*Minor changes in the FSN are those changes that do not alter its meaning. A change to the semantic type shown in parentheses at the end of the FSN may sometimes be considered a minor change if it occurs within a single top-level hierarchy (e.g. a change from a finding tag to a disorder tag, or a change from a procedure tag to a regime/therapy tag), but a move to a completely different top-level hierarchy is regarded as a significant change to* *the Concept's meaning and is prohibited*' [4, p393]. It was therefore hypothesized that changes in semantic tags within the history of a concept would strongly correlate with the concept being inactive in the most recent version. A first analysis summarized in the right half of **Table 10** under the perspective of 'no tag → semantic tag' constituting a change shows that this is indeed the case for concepts in which only one such change occurred: where the expected ratio of active versus inactive concepts is 75.2%/24.8%, the observed ratio is 28.5%/71.5%. The Cramer's V statistic over the entire table being 0.55, suggests a strong correlation. That there are more than expected active concepts with more than one change might be explained by the corrections of mistakes made during an earlier change. However, as further inspection revealed that the majority of one-time changes were changes from no semantic tag to a semantic tag, we recalculated the matrix under this 2nd perspective (left half of **Table 10**) only to find out that under this perspective no conclusive association is present (Cramer V=0.09). It was also at that time that our attention was drawn to the fact that a large amount of semantic tags were assigned to the FSN of concepts that were already inactive since many earlier versions!

Table 10. Associations between number of semantic tag changes and inactivations

Frequency										
Overall %	**No tag → semantic tag = no change**					**No tag → semantic tag = change**				
Row %	Observed			Expected		Observed			Expected	
Column %	Active	Inactive	Total	Active	Inactive	Active	Inactive	Total	Active	Inactive
Changes	1	0		1	0	1	0		1	0
0	292,823	90,798	383,621	288,508	95,113	292,666	42,038	334,704	251,719	82,985
	68.9%	21.4%	90.3%	67.9%	22.4%	68.9%	9.9%	78.8%	59.3%	19.5%
	76.3%	**23.7%**		**75.2%**	**24.8%**	**87.4%**	**12.6%**		**75.2%**	**24.8%**
	91.7%	86.2%		90.3%	90.3%	91.6%	39.9%		78.8%	78.8%
1	25,027	14,403	39,430	29,654	9,776	25,182	63,156	88,338	66,436	21,902
	5.9%	3.4%	9.3%	7.0%	2.3%	5.9%	14.9%	20.8%	15.6%	5.2%
	63.5%	**36.5%**		**75.2%**	**24.8%**	**28.5%**	**71.5%**		**75.2%**	**24.8%**
	7.8%	13.7%		9.3%	9.3%	7.9%	60.0%		20.8%	20.8%
2	1,543	107	1,650	1,241	409	1,545	114	1,659	1,248	411
	0.4%	0.0%	0.4%	0.3%	0.1%	0.4%	0.0%	0.4%	0.3%	0.1%
	93.5%	**6.5%**		**75.2%**	**24.8%**	**93.1%**	**6.9%**		**75.2%**	**24.8%**
	0.5%	0.1%		0.4%	0.4%	0.5%	0.1%		0.4%	0.4%
3	53	5	58	44	14	53	5	58	44	14
	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	91.4%	**8.6%**		**75.2%**	**24.8%**	**91.4%**	**8.6%**		**75.2%**	**24.8%**
	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%		0.0%	0.0%
Totals	319,446	105,313	424,759	319,446	105,313	319,446	105,313	424,759	319,446	105,313
Statistics			Chi^2:	3476.874				Chi^2:	130479.8	
			Cramer's V:	**0.090474**				Cramer's V:	**0.554243**	

A second observation was that certain change patterns occur frequently within a smaller subset of semantic tags. One such subset is the one formed by the semantic tags disorder, finding, situation, morphologic abnormality, event and navigational concept. We constructed again a formal concept analysis context on the basis of 13 attributes: 2 in relation to each of the semantic tags in the subset just sketched, each such attribute reflecting whether the tag is one which occurs in a non-terminal position or a terminal position in a change pattern, and one reflecting whether the concept is active in the last version. We computed the number of SNOMED CT concepts that are described by means of this FCA context (**Table 11**, not showing, however, the breakdown in active/inactive concepts). This table shows, for instance, that – within this context as specified – what finally became tagged as events, where primarily tagged as findings in some earlier version, as well as to a large extent disorders, with the exception of 25 concepts that started with a semantic tag outside the subset of 6.

We also computed the corresponding Duquenne–Guigues base with the number of SNOMED CT concepts that are covered by these implications (**Table 12**). Although this cluster applies to 20,867 concepts, there are not many cases that are covered by the positive implications in the Duquenne–Guigues base. The implication with the largest number of FCA concepts (i.e. 4) s ==> A covers only 40 SNOMED CT concepts. It tells us that – within this context – all SNOMED CT concepts that had a situation in non-terminal position are active. From **Table 11** we can read that 14 of these concepts were finally tagged as findings, and 26 as navigational concepts. The low coverage of positive implications suggests that hard rigor in semantic tag change patterns is hard to come by.

Table 11. Distribution of semantic tag change patterns within a subset of 6 semantic tags

	Terminal tag:	disorder	situation	Morphologic abnormality	finding	event	Navigational Concept	
	Non-terminal tag	D	S	M	F	E	N	
d	Disorder	278	141	0	299	1303	185	2206
s	Situation	0	0	0	14	0	26	40
m	morphologic abnormality	9	0	1	0	0	0	10
f	Finding	1650	472	0	1	8124	215	10462
e	Event	271	0	0	0	0	2	273
n	navigational concept	0	0	0	0	0	0	0
o	Other start tag	71	4464	99	2949	25	268	7876
	Total concepts:	2279	5077	100	3263	9452	696	20867

Table 12. Positive implications from the Duquenne–Guigues base of an FCA context built out the semantic tags specified in **Table 11**

< 4 > s ==> A	40	< 1 > e N ==> A d	2	< 1 > A m ==> M	1
< 1 > o E ==> A	27	< 1 > A s f ==> N	6	< 1 > m M ==> A	1
< 4 > d f ==> A	20	< 1 > A d s ==> F	5	< 1 > f F ==> A	1

Limitation: what qualifies as semantic tags?

The SNOMED CT documentation available from the IHTSDO webserver provides insufficient information on what the precise set of semantic tags the SNOMED CT editors are working with might be. The information that a semantic tag is that what appears at the end of a FSN between brackets [4, p41] is not reliable. Historically, FSNs didn't have a semantic tag at all, as this was apparently introduced later as witnessed by the many changes in descriptions to that end. Parsing anything that terminates a FSN between brackets leads to many false positives in older concepts. For many of those, manual inspection is required for disambiguation. But even then it is not always obvious especially in light of the occurrence of FSNs that apparently enjoy 2 semantic tags: **Figure 1** depicts all terms which we assume to be (or have been at some point) semantic tags and what they are collocated with. For example, we found 393 FSN assertions in which the semantic tag 'body structure' collocates with other semantic tags, i.e. 'morphologic abnormality' (11 occurrences), 'surface region' (82 occurrences) and 'combined site' (300 cases).

Conclusion

SNOMED CT has undoubtedly come a very long way since its original conception as a mere nomenclature for pathology [9, 10]. The IHTSDO has been working very hard on developing editorial and technical principles for updating SNOMED CT and on training its terminologists in applying the principles faithfully. Furthermore, the distribution format RF2 presents itself as a formidable resource to obtain a deeper insight in how SNOMED CT evolved. The exploratory analyses we have performed as part of the work described here and which are of sorts that to our best knowledge have thus far not been described in the literature, made us aware of certain possibilities, but nevertheless revealed many pitfalls in attempting to derive from SNOMED CT's history mechanism what the before mentioned principles exactly might be, or whether they are indeed applied consistently. Whether it is the methodology proposed here itself, or a lack of, for instance, discriminatory power in the reasons for inactivation – one could even wonder why no reasons are given for the addition of new concepts –, is something that needs further to be researched. Nevertheless, it is at this stage of our work possible to formulate the following concrete recommendations towards the IHTSDO: (1) formalize the relationships between semantic tags and SNOMED CT concept hierarchies, (2) implement in the authoring environment mechanisms to prevent and detect incoherent and missing CIRS and HARS records, and (3) provide reasons for not only inactivations, but also activations, which reflect whether changes are purely internal in SNOMED CT (e.g. because of changes in the concept model) or external (changes in the covered domains).

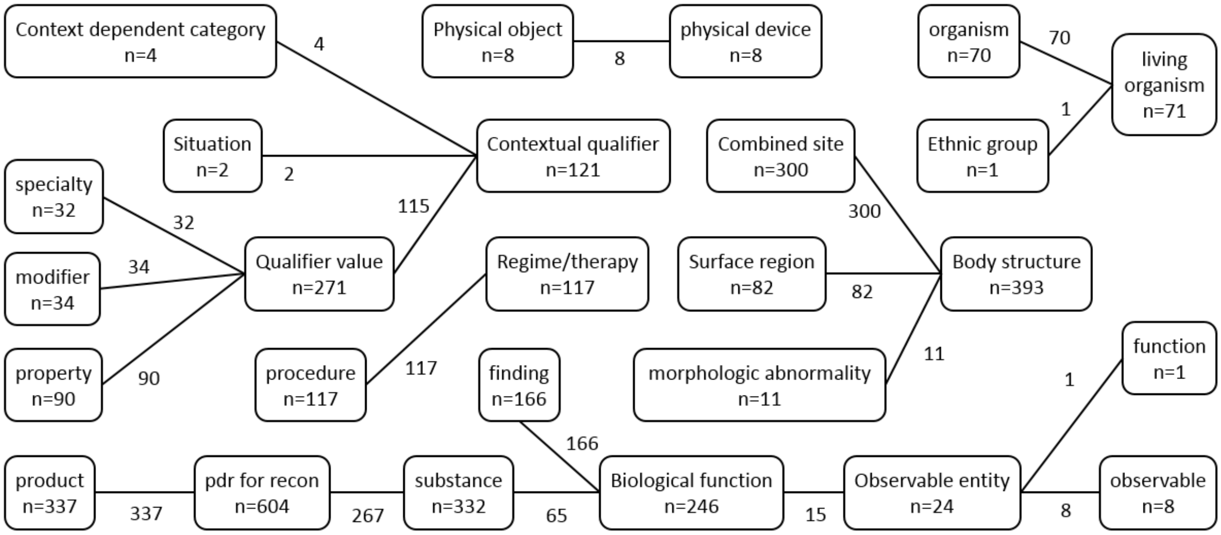

Figure 1. Co-occurrence of candidate semantic tags in FSNs. Numbers inside nodes tally the FSNs in which the candidate semantic tag in the node collocates with another tag. Numbers along the edges tally the FSNs in which the tags in the connected nodes collocate with each other.

Acknowledgments

This work was supported in part by Clinical and Translational Science Award NIH 1 UL1 TR001412-01 from the National Institutes of Health, and by grant R21LM009824 from the National Library of Medicine (NLM). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCR, the NLM or the National Institutes of Health.

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, editors. Studies in Health Technology and Informatics - Medical and Care Compunetics 3 Vol 121. Amsterdam: IOS Press; 2006. p. 279 - 90.
2. Schulz S, Suntisrivaraporn B, Baader F. SNOMED CT's problem list: ontologists' and logicians' therapy suggestions. Stud Health Technol Inform. 2007;129(Pt 1):802-6.
3. Ceusters W. SNOMED CT's RF2: Is the future bright? Stud Health Technol Inform. 2011;169:829-33.
4. IHTSDO. International Health Terminology Standards Development Organization - SNOMED CT® Technical Implementation Guide - January 2015 International Release (US English). 2015. p. 757.
5. Ceusters W. Applying Evolutionary Terminology Auditing to SNOMED CT. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:96-100.
6. Ganter B, Stumme G, Wille R. Formal concept analysis foundations and applications. Berlin: Springer,; 2005. Available from: SpringerLink http://dx.doi.org/10.1007/978-3-540-31881-1.
7. Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):89-102.
8. Kuznetsov SO, Obiedkov S. Some decision and counting problems of the Duquenne–Guigues basis of implications. Discrete Applied Mathematics. 2008;156(11):1994–2003.
9. Major P, Kostrewski BJ, Anderson J. Analysis of the semantic structures of medical reference languages: part 2. Analysis of the semantic power of MeSH, ICD and SNOMED. Medical informatics = Medecine et informatique. 1978 Dec;3(4):269-81.
10. Sommers SC. Systematized nomenclature of pathology. Pathol Microbiol (Basel). 1967;30(5):826-7.