Representing SNOMED CT Concept Evolutions using Process Profiles.

Werner CEUSTERS[1] and Jonathan P. BONA
[a]*Division of Biomedical Ontology, Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo*

Abstract. SNOMED CT is a very large biomedical terminology supported by a concept-based ontology. In recent years it has been distributed under the new release format 'RF2'. RF2 provides a more consistent and coherent mechanism for keeping track of changes over versions, even to the extent that – in theory at least – any release will contain enough information to allow reconstruction of all previous versions. In this paper, using the January 2016 release of SNOMED CT, we explore various ways to transform change-assertions in RF2 into a more uniform representation with the goal of assessing how faithful these changes are with respect to biomedical reality. Key elements in our approach are (1) recent proposals for the Information Artifact Ontology that provide a realism-based perspective on what it means for a representation to be about something, and (2) the expectation that the theory of what we call 'process profiles' can be applied not merely to quantitative information artifacts but also to other sorts of symbolic representations of processes.

Keywords. SNOMED CT, ontological realism, changes in ontologies

1. Introduction

There are many differing views on what it means to do research conducted under the term 'ontology', on what ontologies as representational artifacts exactly are, on what the precise role of ontologies in information systems is, on what they should or should not be used for, and on what qualities or capabilities they should have [1]. Our view is that an ontology should be a faithful representation of the part of reality that it covers [2, 3]: looking through the c(/g)lasses of an ontology and how they are organized therein, should give us exactly the same view as if we were looking directly at the structure of the corresponding part of reality. This would hold both for T-box and A-box assertions, as well as throughout time. Imagine on one side of a room an aquarium with 12 fish, three of each of 4 types and some plants of various types, rocks, etc. and on the other side a holographic simulation of that aquarium and its relevant environment powered by a faithful ontology (including an A-box). If the simulation is synchronized from the start with the exact configuration of the aquarium at that time, though without access to the aquarium and its contents itself, then any change in the aquarium would happen in exactly the same way as in the simulation. If we would run the simulation faster, then we would see exactly what is going to happen in the

[1] Corresponding author. New York State Center of Excellence in Bioinformatics & Life Sciences. 701 Ellicott street, suite B2-160, Buffalo NY – 14203, USA. Email: ceusters@buffalo.edu

aquarium at a future time. To keep the simulation faithful, the maintenance contract for the aquarium and its contents, should go hand in hand with a maintenance contract for the ontology. If, for instance, fish of another type were to be added, then the ontology would need to be updated accordingly. These updates should be such that, by inspecting the ontology directly, we could find out exactly what happened in reality. This means also that the formalism, language, data structures, i.e. the entire representational machinery in and through which the ontology is expressed, must allow us to detect which changes in the ontology correspond to changes in reality, and which are purely ontology- or simulation internal. If the ontologist who maintained the ontology for this simulation is replaced by one who is color-blind and therefore changed the ontology so as to write out in words the names of typical fish colors on the avatars in the simulation, then we should not be forced to believe that the goldfish in our aquarium suddenly have the word 'orange' written all over their bodies.

It is this line of thinking that formed the basis of *Evolutionary Terminology Auditing*, a framework designed to measure quality improvements in ontologies over time using reality as benchmark by taking into account changes in reality itself, changes in our scientific understanding thereof, and pure editorial changes such as corrections of mistakes or changes in representation that are not inspired by changes in reality [4]. In [5, 6] this framework was applied to 18 versions of the *Systematic Nomenclature of Medicine / Clinical Terms* (SNOMED CT) [7] with the conclusion that changes to concepts over those versions do not necessarily correspond to improvements in quality, and that many changes are due to idiosyncrasies in the underlying ontology rather than to changes in the domain or in our scientific understanding. In [8], the method was found to have predictive power over future quality improvements in the Gene Ontology. It was also applied to the Basic Formal Ontology (BFO) [3] which led to a number of improvements to the framework itself [9].

In these past efforts we looked at consecutive versions of an ontology from the perspective of reality, the goal being to assess quality improvements of the ontology in terms of corresponding changes in reality. Here we look instead at mechanisms that an ontology can offer to let us see changes in reality in a reliable way by examining the changes in the ontology. We use as foundations the Basic Formal Ontology (BFO) [3] and recent proposals for the Information Artifact Ontology (IAO) [10] that provide a realism-based perspective for what it means for a representation to be *about* something. SNOMED CT is an ideal candidate for such analytical exploration as its distribution in the last few years includes a new release format known as '*RF2*' which is characterized by a more elaborate, and – as we will demonstrate unfortunately not yet totally – coherent and consistent representation of changes in its content to the extent that each newly released version includes all previous versions rolled up inside itself. Our exploration forms the basis for a long-term research objective to determine whether the totality of assertions about changes in SNOMED CT rather than about external reality constitutes in and of itself a valuable resource to identify patterns that allow detecting mistakes in assertions about external reality that have thus far not been discovered.

2. SNOMED CT as a concept-based ontology

SNOMED CT – the name used to be an acronym for *Systematic Nomenclature of Medicine / Clinical Terms* but is now considered a mere brand name of a new product that grew out of this nomenclature – is developed by the International Health

Terminology Standards Development Organization (IHTSDO) and is claimed, probably rightly, to be the largest healthcare terminology currently available [7]. The *International Edition* released on January 31, 2016 is supported by an ontology consisting of 319,446 *active concepts* which are connected by in total 962,497 *active relationships* and described by 1,097,028 *active descriptions* which link 999,639 *terms* to these concepts. The relationships reported here are those generated by IHTSDO's EL++ description logic classifier on the basis of 655,312 *active* so-called *stated relationships* which have been directly edited by authors or editors prior to running the classifier on the logic definitions [11, p108].

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. Indeed, SNOMED CT is regularly updated [12], not only to correct mistakes, but also to reflect changes in biomedical science. Concepts are classified under several hierarchies. Most top classes correspond to the types of entities instances of which are encountered by clinicians during their work (body parts, organisms, 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'* [13]. Although the number of classes of this sort was originally – and is still – rather small, it is increasing as a result of the move from *Release Format 1* (RF1) to *Release Format 2* (RF2). The latter was introduced in 2012 to implement a more robust and consistent representation of versions including an added hierarchy to represent metadata about the structure of SNOMED CT itself [11 p127, 14].

At the heart of SNOMED CT is the notion of '*concept'* which in the SNOMED CT documentation is defined as '*a clinical idea to which a unique concept identifier has been assigned*'[11, p38]. What is represented by a specific concept cannot be determined on the basis of the identifier, but '*the meaning of a concept can be determined from relationships to other concepts and from associated descriptions that include human readable terms*' [11, p87]. Descriptions provide for each concept a *Fully Specified Name* (FSN): '*Each concept has at least one Fully Specified Name (FSN) intended to provide an unambiguous way to name a concept. The purpose of the FSN is to uniquely describe a concept and clarify its meaning*' [11, p40]. Furthermore: '*Each FSN term ends with a "semantic tag" in parentheses. The semantic tag indicates the semantic category to which the concept belongs (e.g. clinical finding, disorder, procedure, organism, person, etc.). The "semantic tag" helps to disambiguate different concepts which may be referred to by the same commonly used word or phrase*' [11, 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*"' [11, p41].

SNOMED CT's authors have noted – and have to a certain extent started to act upon, though not completely satisfactorily – the confusions around what '*concept'* might denote [15]. Despite their definition of '*concept*' as a clinical idea, the term is also stated to be a homonym for '*concept identifier*' as well as for '*the real-world referent(s) of the concept identifier, that is, the class of entities in reality that the*

concept identifier represents' [11, p127]. One consequence is that there are doubts about the sort of ontological commitments that are made by SNOMED CT authors and editors [16]. Another consequence is that SNOMED CT contains many ambiguities and competing interpretations of, for instance, pathological conditions and disorders [17].

Another consequence of this ambiguity, the one we address specifically in this paper, is that it also requires every occurrence of the word 'concept' in the SNOMED literature – and indeed, in the literature about concept-based ontologies in general – to be disambiguated in terms of whether it is used to denote something which is *outside* or *inside* the ontology. Tumors, procedures and other entities clinicians come in contact with while at work are *outside* SNOMED CT. Examples of something *inside* the SNOMED CT representation are the SNOMED CT concept identifier '*313029009*' and the corresponding FSN '*Brachytherapy – action (qualifier value)*', both of which are supposed to denote the method involved in what it takes for a procedure to be of a sort denoted both by '*384692006*' and by the term '*Brachytherapy procedure*'.

This ambiguity arises not only in the documentation but also in SNOMED CT itself! We can safely assume that the relationship (T1), between a procedure and a qualifier value, extracted from the SNOMED CT relationships file and rendered in human readable form by using FSNs is, as SNOMED CT puts it, about '*a class of entities in reality*', thus about something *outside* SNOMED CT. More concretely: the term '*Intracavitary brachytherapy (procedure)*' is *inside* SNOMED CT, but that what this term denotes and of which a specific brachytherapy procedure carried out on a specific patient is an instance (see section 3), is on the *outside*.

'*Intracavitary brachytherapy (procedure)*' (T1)
– '*Method (attribute)*'
– '*Brachytherapy – action (qualifier value)*'

'Actions by modality (qualifier value)' (T2)
– *'Is a (attribute)'*
– *'Action (qualifier value)*'

'Brachytherapy – action (qualifier value)' (T3)
– *'Is a (attribute)'*
– *'Actions by modality (qualifier value)*'

We can however no less safely assume that the triples (T2) and (T3) are to be interpreted as statements about how SNOMED CT classifies certain actions, perhaps in order to allow for easier browsing when SNOMED CT is used in some application as an interface terminology. These are thus statements about something *inside* SNOMED CT, rather than that 'actions by modality', on the outside, are a special kind of actions in and by itself of which brachytherapy actions are an example. These distinctions are important if we want to quantify reliably how much of external reality is represented in SNOMED CT and how SNOMED CT is qualitatively improving as a representation using reality as benchmark. For example, although (T2) and (T3) together use three concepts – (1) 'actions by modality', (2) 'action', and (3) 'brachytherapy' – only two of them, (2) and (3), correspond to an entity in reality.

3. SNOMED CT as an Information Content Entity

One way to address these issues is to perceive a version of SNOMED CT as an instance of an Information Content Entity (ICE), i.e. the sort of entity which is represented as the root of the IAO which is under development as a BFO-compatible ontology for information artifacts [10]. **Table 1** summarizes the definitions (D*n*) and elucidations (E*n*) as they crystalized out of several proposals in the past few years [10, 18-21]. They are themselves based in part on the terms ENTITY, GENERICALLY DEPENDENT CONTINUANT, MATERIAL ENTITY, QUALITY, FUNCTION and ROLE as well as the notions of specific and generic dependence as defined in BFO [3]. These definitions allow us to perceive a version of SNOMED CT as an ICE of which concretizations exist as INFORMATION ARTIFACTS in the form of, for example, a paper print out, or the portion of a hard drive which contains the RF2 distribution files each one of which can be rendered as a table on a computer screen by using appropriate software.

In this light, the PORTION OF REALITY (PoR) described by SNOMED CT, includes, from the ontological realist perspective as we perceive it [2]:

1. universals, such as, for instance, the universal denoted by the SNOMED CT concept identifier '126838000' which is further annotated by means of the *description* (with ID '126016') stating that the *term* 'neoplasm of colon' is an allowed term since the January 2002 version;
2. relations, for instance the formal subsumption relation which in SNOMED CT is represented by the concept identifier '116680003' and by the corresponding *term* 'Is a (attribute)';

Table 1. Core definitions and elucidations for representation and aboutness

INFORMATION CONTENT ENTITY (ICE) =def. an ENTITY which is (1) GENERICALLY DEPENDENT on (2) some MATERIAL ENTITY and which (3) stands in a relation of aboutness to some PORTION OF REALITY.	[18]	(D1)
INFORMATION QUALITY ENTITY (IQE) =def. a QUALITY that is the concretization of some INFORMATION CONTENT ENTITY.	[19]	(D2)
ARTIFACT =def. a MATERIAL ENTITY created or modified or selected by some agent to realize a certain FUNCTION or ROLE.	[10]	(D3)
INFORMATION ARTIFACT =def. an ARTIFACT whose FUNCTION is to bear an INFORMATION QUALITY ENTITY.	[10]	(D4)
REPRESENTATION =def. a QUALITY which ***is_about*** or is intended to ***be about*** a PORTION OF REALITY.	[20]	(D5)
MENTAL QUALITY =def. a QUALITY which ***specifically_depends_on*** an ANATOMICAL STRUCTURE in the cognitive system of an ORGANISM.	[20]	(D6)
COGNITIVE REPRESENTATION =def. a REPRESENTATION which is a MENTAL QUALITY.	[20]	(D7)
REPRESENTATIONAL UNIT (RU) = def. a smallest constituent sub-representation, including icons, names, simple word forms, or the sorts of alphanumeric identifiers we might find in patient records.	[21]	(D8)
x ***is_about*** *y* means**:** *x refers to or is cognitively directed towards y*. **Domain:** REPRESENTATIONS; **Range:** PORTIONS OF REALITY. **Axiom:** if *x* ***is_about*** *y* then *y* exists (veridicality).	[10]	(E1)
x ***concretizes*** *y* ***at*** *t* means: *x* is a QUALITY & *y* is a GENERICALLY DEPENDENT CONTINUANT & for some MATERIAL ENTITY z, *x* ***specifically_depends_on*** *z* at *t;* & *y* ***generically_depends_on*** *z* at *t*; & if *y* migrates from bearer *z* to another bearer *w* then a copy of *x* will be created in *w*.	[10]	(E2)
x ***is_a_representation_of*** *y* =def. *x* is a REPRESENTATION & *x* ***is_about*** *y* (where *y* is a portion of reality).	[10]	(D9)

3. instances, for example the one denoted by the concept ID '223502009' and corresponding *FSN* 'Europe (geographic location)'; and,
4. configurations, for instance the one directly referred to in the ICE concretized by the triple (T1), and the one indirectly represented by combining the triples (T3) and (T2) used as examples in section 2.

Note that the representation formalism used by SNOMED CT is not able to let us distinguish universals from instances [13]. Configurations are formally represented through records in, for instance, the relationships file. This includes configurations formed by ICEs themselves such as those denoted by records in Historical Association Reference Sets (section 4.2). Relations are implicitly represented as such by being subsumed by the concept with FSN '*Attribute (attribute)*' and explicitly through their specific position in records of, for example, the relationships file in RF2.

To avoid the confusions arising from the word 'concept' as used in the SNOMED CT documentation, we will use in this paper the term '*SNOMED CT concept*' – or '*concept*' for short – *exclusively in the ICE sense*, i.e. to denote a representational element *inside* the SNOMED CT representation. If this representational element succeeds in being about something (see D9), we will denote that something by terms such as 'the corresponding PoR' or 'the corresponding universal'. This holds also for the other SNOMED CT components such as descriptions and relations. These terms will exclusively be used to denote representational elements *inside* SNOMED CT.

4. Changes in SNOMED CT

4.1. Additions and deactivations

The content of SNOMED CT evolves with each release. The types of changes made include the addition and inactivation of concepts, descriptions, and relationships as well as updates in definitions, and to a certain extent also the provision of motivations for these changes. Once released, SNOMED CT components are persistent and their identifiers are not reused [11, 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. Whereas in RF1 the history mechanism was only used to annotate changes in concepts and descriptions, RF2 annotates changes in a consistent fashion for all components, though only for changes that occurred since the January 2002 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 2** shows that the concept '301381004' with FSN '*Discomforting present pain (finding)*' was set to active in release 20020131 and to inactive in 20080131. **Table 3** shows that during the life time of that concept, it underwent considerable changes in its reported relationships to other concepts after full DL classification. It must however be noted that the SNOMED CT documentation remains silent on whether these reported changes are syntactical changes, effectual changes or a combination thereof. From **Table 3** alone it can indeed not be assessed whether the relationship '*Isa – Pain (finding)*' is truly inactivated, or whether it is still active in the historical transitive closures, something that can be computed on the basis

of the history information available in RF2. Between 2009 and 2011 there were typically more effectual changes (74%) than ineffectual ones (26%); within the removals there was a high number of ineffectual changes (37%) whereas in the additions there were on average more effectual changes (84%) than ineffectual ones (16%) [22]. Note, however, that '*effectual change*' in [22] is to be understood as a pure change *inside* SNOMED CT from one version to another, and not as an assertion that an effectual change corresponds to a change in reality or SNOMED CT's authors' knowledge thereof.

Table 4 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 4**).

Table 4 points out another weakness in the concept-orientation adhered to by SNOMED CT, and its consequent reliance on '*meanings*' and all problems that arise therefrom [23]. The SNOMED CT documentation states that '*only limited changes may be made to the* "*term*" *field, as defined by editorial rules*' [11, p145]. This is consistent with the view that '*the meaning of a concept can be determined […] from associated descriptions that include human readable terms*' [11, p87]. This editorial rule is also used as an 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*' [11, p393]. Yet, a change from '*finding'* to '*context-dependent category*' (later renamed '*situation'*) is precisely a move from one top-level category to another. Despite this change, the concept was not deactivated! This can only be explained – unless it was a mistake introduced in 2003 and detected prior to the release of the July 2009 version in which this concept became deactivated – if we assume that the SNOMED CT editors at that time clearly realized that whatever they change *inside* SNOMED CT does not have an impact on how matters are on the *outside*.

Thus what stays fixed – modulo the appearance of truly new entities such as new drugs, mutated viruses, and, perhaps new disorder types caused by newly developed techniques or chemicals – are the entities on the outside, the portions of reality denoted by some SNOMED CT component on the inside. This holds, of course, also for the massive number of changes that occur at the level of the SNOMED CT relationships: although they clearly change '*the meaning*' of the concept in many cases, they are still, from a realist perspective, intended to denote the very same PoRs.

Table 2. 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 3. 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, 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 4. 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, 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)'.

4.2. Replacements

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 5**) are used to indicate, for example, which deactivated concepts are in one way or another related to other active concepts, and CIRSs (**Table 6**) to indicate the reasons for inactivating a component – such as errors, duplication of another component and ambiguity of meaning [11, 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 [11, p654].

Table 5. Historical association reference set types in SNOMED CT (modified from [11, p509])

HARS name	Use
POSSIBLY EQUIVALENT TO	From an ambiguous concept to one or more active concepts that represents one of the possible meanings of the inactive concept.
MOVED TO	From a component to a namespace to which the component has been moved
REPLACED BY	From an erroneous or obsolete inactive component to a single active replacement component.
SAME AS	From a duplicate component to the active component that this component duplicates.
WAS A	From an inactive classification concept such as "not otherwise specified" to the active concept that was formerly its most proximal supertype.
ALTERNATIVE	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 6. Component inactivation set types for concepts (modified from [11, p506-507])

CIRS value	Concept status
Duplicate	inactive because it has the same meaning as another Concept
Outdated	inactive because it is an outdated concept that is no longer used.
Ambiguous	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	inactive because it contains an error
Limited	active prior to Jan 2010, inactive since then because of unstable meaning within SNOMED CT.
moved to	inactive because moved to another namespace.
Pending move	active but in the process of being moved to another namespace

Interestingly, the very same concepts can not only appear as source concept in one HARS member and as target concept in another HARS member, but also appear in members of distinct HARSs. This allows the computation of association networks 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 concept in the same way until no more concepts can be found.

5. Discussion: towards process profiles for changes in SNOMED CT components

As instances of ICE, thus continuants, components have a *history* – an occurrent process – in which they participate for the entire time of their existence. This is comparable to the history of an organism, i.e. the process in which an organism participates for the entire temporal period during which it exists. For organisms, there is a process of shorter duration with can be qualified as *life*, the process in which the organism participates for the entire time it is alive and which is an occurrent-part [3] of the organism's history. In a similar sense, a component can be perceived as being alive or dead, when declared to be active or inactive respectively. Furthermore, depending on the type of component, it can be alive or dead in different ways. While a concept is active, it can be 'fully' alive or, when it is marked for a pending move, 'dying' (**Table 6**). Prior to 2010, it could also be alive in a 'limited' way. In [24], process profiles were identified as something that is not numerically but qualitatively 'the same' in distinct

processes such as the ‘same’ temperature change of two rocks in our aquarium when the water temperature changes. These processes each have as part an instance of a quality process profile of exactly the same (determinate) type, i.e. ‘*that part of a process which serves as the target of selective abstraction focused on a sequence of instances of determinate temperature qualities*’ [24]. It is speculated in [24] that the theory of process profiles can be applied not merely to quantitative information artifacts but also to other sorts of symbolic representations of processes. It is this that we try to achieve with respect to changes that occur in SNOMED CT components, including memberships in HARSs and CIRSs. Although SNOMED CT’s RF2 format is more coherent than its predecessor at the syntactic level, it requires more restructuring of the data to arrive at a uniform view of what changed in relation to a specific concept, and from there to infer what might have happened on the side of the corresponding PoR (in case there is one).

Table 7 uses 5 concepts (C1 … C5) as examples of how to construct process profile representations (PPRs) in a (nearly) uniform way for the various sorts of changes the concepts – from this perspective – underwent. Each PPR consists of 29 characters, 1 for each version, each one representing the status of some quality-like feature that can be ascribed to the concept. The column ‘Attr.’ represents those features at a level on a par with ‘temperature’, ‘color’, etc. For the rows with neutral background, the combination of what appears in the ‘Attr.’ and ‘Value’ columns represents those features at the most determinate level that we were able to measure, comparable to ‘37.2 centigrade temperature’. Here ‘FSN+T-367’, f.i, means that the term ‘General symptom NOS (finding)’, the 367th term (randomly numbered) out of 999,639 terms was ‘measured’ as the determinate value for ‘FSN’ (since we used a FSN-thermometer, not a Synonym-thermometer). The table shows that this quality-like feature was found to inhere in the concepts C1 through C4, whereby, as can be determined from the respective PPRs, the histories of C1, C2 and C3 all share some occurrent-part which instantiates the same most-determinate process profile universal, and they do this at the same time (starting from the 14th version). ‘A’ in this case stands for ‘active’, while ‘_’ means that there is at the respective time no instance of the quality-like feature inhering in the concept. C4, in contrast, exhibits a different PPR for this feature, one that is the result of a start in the 9th version. For the rows in grey background, the value in ‘Value’ does not correspond to a measurement at a specific point in time, but with a most-determinate PPR type itself. ‘DSP-05’, for instance, is one out of 34 most-determinate PPR types for the quality-like feature ‘Dstatus’ (definitional status). It is C1, C2, and C3 that exhibit an instance of this type.

6. Related work

Computer scientists and logicians have developed a number of theoretical approaches to deal with logical changes in description logic based ontologies. For instance, in [25] a model-theoretic semantics for ontology versioning based on first-order-logic is proposed that can be applied to ontologies expressed in RDF and OWL. [26] reports on the development of a Multi-version Ontology REasoner (MORE) based on using temporal logics to perform reasoning across multiple versions of ontologies. MORE was tested on small ontologies in two different domains. In [27], a change detection approach for OWL based on a logical change definition language and temporal logic is proposed. [28] presents a tool for tracking and visualizing differences

between two versions of an ontology. [29] describes an interactive tool for visualizing and exploring ontology changes that offers both overview and concept-based analyses.

In [30] the rate of changes in SNOMED CT was characterized and quantified from 2002 to 2005, finding that most changes were occurring among relationships, and in particular subsumption relationships, and concluding that implementers must '*carefully examine mechanisms for handling this degree of change*'. By examining changes in SNOMED CT over three years as recorded in the Component History and Concept Model with a focus on the subset of concepts in the NLM CORE Problem List, four types of changes (present in over 40% of the target concepts over the studied timespan) were identified that are likely to impact health recordkeeping [31]. In [32], an approach is presented to identify idiosyncrasies such as relation reversals (a particularly dramatic type of structural change) in the evolution of SNOMED CT, finding 48 such reversals since 2009. [33] demonstrates how changes between two SNOMED CT versions affected a majority of concepts used in a legacy mapped interface terminology, including unexpected effects of structural changes in SNOMED CT, and argues for a consideration of impact on such implementations as part of terminology development. Motivated by [33], [12] presents indicators that can be computed to assess whether an upgrade from one version to the next would be worth the effort.

Table 7. Uniform representation of changes in SNOMED CT components using process quality profiles.

S	FSN	Attr.	Value	Value label	Process Profile Representation (PPR)
C1	GS NOS	Dstatus	DSP-05		DDDDDDDDDDDDDDDDDDDDDDDDDDDDD
C1	GS NOS	Reason	CIP-15		DDDDDDDDDDDDDDDDLLLLLLLLLLLL
C1	GS NOS	FSN	T-367	GS NOS (finding)	_____AAAAAAAAAAAAAAA
C1	GS NOS	Same-as	C4	GS NOS (finding)	_AAAAAAAAAAAAAAAAAAAAAAAAAAAA
C1	GS NOS	Was-a	C5	GS (finding)	_____AAAAAAAAAAAA
C2	GS NOS	Dstatus	DSP-05		DDDDDDDDDDDDDDDDDDDDDDDDDDDDD
C2	GS NOS	Reason	CIP-15		DDDDDDDDDDDDDDDDLLLLLLLLLLLL
C2	GS NOS	FSN	T-367	GS NOS (finding)	_____AAAAAAAAAAAAAAA
C2	GS NOS	Same-as	C4	GS NOS (finding)	_AAAAAAAAAAAAAAAAAAAAAAAAAAAA
C2	GS NOS	Was-a	C5	GS (finding)	_____AAAAAAAAAAAA
C3	GS NOS	Dstatus	DSP-05		DDDDDDDDDDDDDDDDDDDDDDDDDDDDD
C3	GS NOS	Reason	CIP-15		DDDDDDDDDDDDDDDDLLLLLLLLLLLL
C3	GS NOS	FSN	T-367	GS NOS (finding)	_____AAAAAAAAAAAAAAA
C3	GS NOS	Same-as	C4	GS NOS (finding)	_AAAAAAAAAAAAAAAAAAAAAAAAAAAA
C3	GS NOS	Was-a	C5	GS (finding)	_____AAAAAAAAAAAA
C4	GS NOS	Dstatus	DSP-20		PPPPPPPPPPPPPPPDDDDDDDDDDDDD
C4	GS NOS	Reason	CIP-18		LLLLLLLLLLLLLLLLLLLLLLLLLLLL
C4	GS NOS	FSN	T-367	GS NOS (finding)	________AAAAAAAAAAAAAAAAAAAA
C4	GS NOS	FSN	T-258	GS NOS (cont-dep. category)	AAAAAAAADDDDDDDDDDDDDDDDDDDD
C4	GS NOS	Is a	C5	GS (finding)	AAAAAAAAAAAAAAAADDDDDDDDDDDD
C4	GS NOS	Same-as	C1	GS NOS (finding)	_____AAAAAAAAAAAA
C4	GS NOS	Same-as	C2	GS NOS (finding)	_____AAAAAAAAAAAA
C4	GS NOS	Same-as	C3	GS NOS (finding)	_____AAAAAAAAAAAA
C4	GS NOS	Was-a	C5	GS (finding)	_____AAAAAAAAAAAA
C5	GS	Dstatus	DSP-03		PPPPPPPPPPPPPPPPPPPPPPPPPPPP
C5	GS	FSN	T-368	GS (finding)	________AAAAAAAAAAAAAAAAAAAA
C5	GS	FSN	T-277	GSs (cont-dep. category)	AAAAAAAADDDDDDDDDDDDDDDDDDDD

Legend: '**S**' source concept. Concept identifiers were abbreviated for space reasons: C1=139169008, C2=139174000, C3=161914002, C4=161919007, C5=267022002. FSNs of concepts are abbreviated to 'GS' for 'General symptom'. Dstatus=concept definition status. DSP=description status profile, CIP=concept inactivation profile. T=term. Individual characters in PPR are abbreviations of SNOMED CT properties: 'A'=active, 'L'=limited value, 'P'=primitive, 'D'=defined, '_'=no value present.

7. Conclusion

Many efforts have been made to measure the amount and type of changes occurring between SNOMED CT versions. To our best knowledge, a method based on the representation of process profiles has thus far not been attempted. The results we obtained in our exploration are promising although more work on our side towards further harmonization is required. In any case, when in 2011 we asked ourselves the question whether with RF2 SNOMED CT's future is bright [14], we were not able to answer it. Now we believe we can: when complemented with an approach as proposed here, it is! We strongly recommend any ontology to be distributed using such an improved RF2 format – or semantic equivalent along the lines described here – since without such mechanisms data annotated in terms of previous versions lose value dramatically.

Acknowledgments

This work was supported in part by Clinical and Translational Science Award NIH 1 UL1 TR001412-01 from the National Institutes of Health, by grant R21LM009824 from the National Library of Medicine (NLM), and by grant 1R01DE021917-01A1 from the National Institute of Dental and Craniofacial Research (NIDCR). 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] F. Fonseca, "The Double Role of Ontologies in Information Science Research," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 6, pp. 786-793, 2007.

[2] B. Smith, and W. Ceusters, "Ontological realism: A methodology for coordinated evolution of scientific ontologies," *Applied Ontology*, vol. 5, no. 3-4, pp. 139-188, 2010.

[3] R. Arp, B. Smith, and A. D. Spear, "Building ontologies with basic formal ontology," The MIT Press,, 2015, p. 1 online resource.

[4] W. Ceusters, and B. Smith, "A Realism-Based Approach to the Evolution of Biomedical Ontologies," *Biomedical and Health Informatics: Proceedings of the 2006 AMIA Annual Symposium*, pp. 121-125, Washington DC: American Medical Informatics Association, 2006.

[5] W. Ceusters, "Applying Evolutionary Terminology Auditing to SNOMED CT," *AMIA Annu Symp Proc*, vol. 2010, pp. 96-100, 2010.

[6] W. Ceusters, K. A. Spackman, and B. Smith, "Would SNOMED CT benefit from Realism-Based Ontology Evolution?." In Teich JM, Suermondt J, Hripcsak C. (eds.), American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy, Chicago IL, 2007;:105-109..

[7] Donnelly K, "SNOMED CT: The Advanced Terminology and Coding System for eHealth," *Studies in Health Technology and Informatics - Medical and Care Compunetics 3. Vol 121*, Bos L, Roa L, Yogesan K *et al.*, eds., pp. 279 - 290, Amsterdam: IOS Press, 2006.

[8] W. Ceusters, "Applying Evolutionary Terminology Auditing to the Gene Ontology," *Journal of Biomedical Informatics; Special Issue of the Journal of Biomedical Informatics on Auditing of Terminologies*, vol. 42, no. 3, pp. 518-529, 2009.

[9] S. Seppälä, B. Smith, and W. Ceusters, "Applying the Realism-Based Ontology-Versioning Method for Tracking Changes in the Basic Formal Ontology.," *Formal Ontology in Information Systems*, Frontiers in Artificial Intelligence and Applications P. Garbacz and O. Kutz, eds., pp. 227-240, 2014.

[10] B. Smith, and W. Ceusters, "Aboutness: Towards Foundations for the Information Artifact Ontology," in International Conference on Biomedical Ontology, Lisbon, Portugal, 2015, pp. 47-51.

[11] IHTSDO, "International Health Terminology Standards Development Organization - SNOMED CT® Technical Implementation Guide - January 2015 International Release (US English)," 2015, p. 757.
[12] W. Ceusters, "SNOMED CT revisions and coded data repositories: when to upgrade?," *AMIA Annu Symp Proc*, vol. 2011, pp. 197-206, 2011.
[13] S. Schulz, B. Suntisrivaraporn, and F. Baader, "SNOMED CT's problem list: ontologists' and logicians' therapy suggestions," *Stud Health Technol Inform*, vol. 129, no. Pt 1, pp. 802-6, 2007.
[14] W. Ceusters, "SNOMED CT's RF2: Is the future bright?," *Stud Health Technol Inform*, vol. 169, pp. 829-33, 2011.
[15] B. Smith, "Beyond concepts: ontology as reality representation," *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, pp. 73-84, Amsterdam: IOS Press, 2004.
[16] S. Schulz, and R. Cornet, "SNOMED CT's Ontological Commitment," *ICBO: International Conference on Biomedical Ontology*, B. Smith, ed., pp. 55-58, Buffalo NY: National Center for Ontological Research, 2009.
[17] S. Schulz, A. Rector, J. M. Rodrigues *et al.*, "Competing interpretations of disorder codes in SNOMED CT and ICD," *AMIA Annu Symp Proc*, vol. 2012, pp. 819-27, 2012.
[18] W. Ceusters, "An information artifact ontology perspective on data collections and associated representational artifacts," *Stud Health Technol Inform*, vol. 180, pp. 68-72, 2012.
[19] B. Smith, T. Malyuta, R. Rudnicki *et al.*, "IAO-Intel: An Ontology of Information Artifacts in the Intelligence Domain," *CEUR Workshop Proceedings*. pp. 33-40.
[20] W. Ceusters, and B. Smith, "Foundations for a realist ontology of mental disease," *Journal of Biomedical Semantics*, vol. 1, no. 10, pp. 1-23, 9 December 2010, 2010.
[21] B. Smith, W. Kusnierczyk, D. Schober *et al.*, "Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain," *KR-MED 2006, Biomedical Ontology in Action.*, Baltimore MD, USA 2006.
[22] R. S. Gonçalves, B. Parsia, and U. Sattler, "Facilitating the analysis of ontology differences," in Joint workshop on knowledge evolution and ontology dynamics (EvoDyn) 2011, pp. 20-35.
[23] B. Smith, "Against Fantology," *Experience and Analysis*, M. E. Reicher and J. C. Marek, eds., pp. 153-170, Wien, 2005.
[24] B. Smith, "Classifying Processes: An Essay in Applied Ontology," *Ratio (Oxf)*, vol. 25, no. 4, pp. 463-488, Dec 1, 2012.
[25] J. Heflin, and Z. Pan, "A model theoretic semantics for ontology versioning," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2004, pp. 62-76.
[26] Z. Huang, and H. Stuckenschmidt, "Reasoning with multi-version ontologies: A temporal logic approach," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2005, pp. 398-412.
[27] P. Plessers, O. De Troyer, and S. Casteleyn, "Understanding ontology evolution: A change detection approach," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 1, pp. 39-49, 3//, 2007.
[28] N. F. Noy, S. Kunnatur, M. Klein *et al.*, "Tracking Changes during Ontology Evolution," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2004, pp. 259-273.
[29] M. Hartung, T. Kirsten, A. Gross *et al.*, "OnEX: Exploring changes in life science ontologies," *BMC Bioinformatics*, vol. 10, pp. 250, 2009.
[30] K. A. Spackman, "Rates of Change in a Large Clinical Terminology: Three Years Experience with SNOMED Clinical Terms," *AMIA Annual Symposium Proceedings*, vol. 2005, pp. 714-718, 2005.
[31] D. Lee, R. Cornet, and F. Lau, "Implications of SNOMED CT versioning," *International Journal of Medical Informatics*, vol. 80, no. 6, pp. 442-453, 6//, 2011.
[32] S. Tao, L. Cui, W. Zhu *et al.*, "Mining Relation Reversals in the Evolution of SNOMED CT Using MapReduce," *AMIA Summits on Translational Science Proceedings*, vol. 2015, pp. 46-50, 03/23, 2015.
[33] G. Wade, and S. T. Rosenbloom, "The impact of SNOMED CT revisions on a mapped interface terminology: Terminology development and implementation issues," *Journal of Biomedical Informatics*, vol. 42, no. 3, pp. 490-493, 6//, 2009.