MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-1000

Bridging the semantics gap between terminologies, ontologies, and information models

Stefan Schulz[a], Daniel Schober[a], Christel Daniel[b,c], Marie-Christine Jaulent[b]

[a]*Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany*
[b]*INSERM, UMR_S 872, eq.20, Descartes University, Paris, France*
[c]*ASIPSanté, Paris, France*

Abstract

SNOMED CT and other biomedical vocabularies provide semantic identifiers for all kinds of linguistic expressions, many of which cannot be considered terms in a strict sense. We analyzed such "non-terms" in SNOMED CT and concluded that many of them cannot be interpreted as directly referring to objects or processes, but rather to information entities. Discussing two approaches to represent information entities, viz. the OBO Information artifact ontology (IAO) and the HL7 v3 Reference Information Model (RIM), we propose an integrative solution for representing information entities in SNOMED CT, in a way that is still compatible with RIM and the IAO and uses moderately enhanced description logics.

Keywords:
SNOMED, Information models, Ontologies

Introduction

SNOMED CT, the emerging global health terminology standard is published by the International Health Terminology Standards Development Organisation (IHTSDO) as a "core general terminology for the electronic health record" [1]. It provides unified meanings for clinical terms from different languages by assigning them to concepts as language-independent identifiers of meaning. Terms are, according to ISO 1087, "*designations of defined concepts in a special language by linguistic expressions*" [2]. Although there are very different, partly contradicting approaches of which criteria should be used to classify a linguistic expression as a term, there is an increasing consensus of terms having both structural (noun phrases) and statistic properties (occurring with a certain frequency and specificity in written and oral communications) [3]. However, any cursory inspection of SNOMED reveals tens of thousands of entries for which it is at least debatable whether they should be regarded as terms along the above criteria, see Table 1:

Table 1-"Non-Terms" in SNOMED CT

#	SNOMED ID	"Term"
1	59000001	Surgical pathology consultation and report on referred slides prepared elsewhere
2	418577003	Take at regular intervals. Complete the prescribed course unless otherwise directed
3	39399006	Natural death with probable cause suspected
4	168383004	Helicobacter blood test negative
5	281581004	Poor condition at birth without known asphyxia
6	413241009	Suspicion of gastritis

Here, rather than to terms proper, SNOMED CT concepts correspond to more or less complex linguistic assertions, which include statements of facts, beliefs, and orders. This raises the hypothesis that these "concepts" fulfill tasks that differ from the provision of controlled terms.

Since SNOMED RT, CT's predecessor, description logics (DLs) [4], formal languages with a well-understood semantics, have been used to formally describe the meaning of SNOMED CT concepts in terms of the common properties of the particular things that instantiate them. We consider these formal descriptions as SNOMED CT's ontology component, considering ontologies as theories that attempt to give precise mathematical formulations of the properties and relations of real-world particulars [5].

Formal representations of electronic health record content have also motivated another line of effort, *viz.* the development of information models for messages and documents in the framework of HL7 Version 3 [6].

In this paper we want to explore the qualitative boundary between "terms" and "non-terms" in SNOMED CT. We postulate that only for the representation of concepts that are instantiated by objects in reality the current logic framework is appropriate, whereas for SNOMED CT concepts that are instantiated by information entities, this framework needs to be extended. We will investigate what kind of things SNOMED CT "non-terms" denote, in which parts of SNOMED CT they occur, and how they relate to clinical information models.

Materials and Methods

Description Logics

SNOMED CT uses a description logics dialect known as **`EL`**, we will shortly introduce. As a running example, we use the English term "Liver", which belongs to a concept uniquely identified by the number 181268008 and the human-readable name “Entire liver (body structure)”. SNOMED CT concepts are arranged in taxonomic (subsumption) hierarchies. This means that all instances of this concept (i.e. all individual livers) are also instances of its taxonomic parent identified by “272627002|Entire digestive organ (body structure)”. We express this as *Liver ⊑ Digestive Organ*. Beside the taxonomic arrangement the meaning of SNOMED CT concepts can be further described by the properties all their instances have in common. In the following example, we employ the ⊓ (“and”) operator and add a quantified role, using the existential quantifier ∃ (“exists”). For example, the expression *Inflammatory disease ⊓ ∃ has-location.Liver* extends to all instances that both instantiate *Inflammatory disease* and are further related through the relation *has-location* to some instance of *Liver*. This example actually gives us both the necessary and the sufficient conditions needed in order to fully define a class, e.g.:

Hepatitis ≡ Inflammatory disease ⊓ ∃ has-location.Liver, with the equivalence operator ≡ telling that (i) each and every particular *Hepatitis* instance is also an instance of *Inflammatory disease* that is located in some instance of *Liver*, and *vice versa* (ii) that every instance of *Inflammatory disease* that is located at some *Liver* is an instance of *Hepatitis*.

SNOMED CT, in its current version is limited to the constructors summarized in Table 2.

Table 2-SNOMED CT’s logical constructors, corresponding to the description logics **EL**

DL Constructor		Meaning	Example
⊓	*E ⊓ F*	Intersection between *E* and *F*	*Acid ⊓ Organic Molecule*
∃	∃*r.G*	Existential restriction of the relation *r* by the filler *G*	∃*part-of.Liver*
⊑	*A ⊑ B*	*B* subsumes *A*	*Liver ⊑ Organ*
≡	*C ≡ D*	*C* and *D* are equivalent	*Organic Acid ≡ Acid ⊓ Organic Molecule*

It is not possible to express value constraints, e.g. that the relation *has-laterality* can only have the values *Right* and *Left*. It is equally impossible to express cardinalities, such as precisely defining a *Coronary bypass with three grafts*. And it is not possible to formulate negations, such as *Injury without infection*.

These restrictions can be tolerated as well as SNOMED restricts itself to the definition of the meaning of simple terms like "Hepatitis" or "Nephrotomy". It is, however, problematic, whenever more complex terms or whole statements as in Table 1 have to be compositionally represented.

Information models

Statements as illustrated in Table 1 typically belong to information models, such as underlying data acquisition templates, questionnaires and the like. Typical standards for clinical information models are open EHR archetypes [7] and HL7 version 3 information models [6]. The Reference Information Model (RIM) is the general structure that guarantees the coherence of the complex set of HL7 version 3 models, which rence of the complex set of HL7 version 3 models, which may be used in many contexts to describe particular administrative or clinical health care information. Table 3 contrasts what is typically represented by ontologies with what is typically represented by information models. For example the definition of the class *Act* in the HL7-supported code system is “*a record of something that is being done, has been done, can be done, or is intended or requested to be done*”.

Table 3-Ontologies vs. Information Models. In practice the distinction is less crisp. Especially the HL7 RIM contains many classes that can be assumed to represent non-informational entities.

Domain Ontologies	Information Models
Contain classes that have really existing domain entities (particulars) as members	Classes have information entities as members
Represent real-world particulars in terms of their inherent properties	Represent artifacts that are build to collect or annotate information
Can exist independently of information models as long as only the existence of particular things is recorded	Are required to record beliefs or states of knowledge about real things or types of things (as represented by ontologies)
Context independent	Context dependent

Examples are clinical observations, the assessment of health conditions, healthcare goals, treatment services, assisting, monitoring or attending, patient training and education services, editing and maintaining documents, and many others. *Acts* (besides *Entities* and *Roles*) are the pivots of the RIM; all domain information and processes are represented primarily in acts. Any profession or business, including healthcare, is primarily constituted of intentional actions, performed and recorded by responsible actors. An act-instance is a record of such an intentional action. The fundamental difference between such a RIM act instance and an instance of an ontology class (or also most SNOMED CT concepts) is to bring the aspect of recording and thus the person who edits EHR content into the picture. At least in theory, an instance of RIM:*Operation* refers to an information object which is “*about*” some type or concept, which not necessarily is instantiated. Representing discourse about operations that are being planned, postponed, or suspended is quite different from creating and instance of an ontology class *Operation*, as the latter one makes an existence claim which is often too strong.

The Ontology – Epistemology Divide

We may be able, in theory, to draw a crisp line between what is the representation of real objects or processes on the one hand, and what represents information entities on the other hand. In current information models and ontologies this distinction is blurred, and users of both systems tend to be unaware of the very nature of things they represent. The resulting overlaps give rise to conflicting representations, which require sophisticated mitigation strategies (TermInfo). Such a mixed representation of the invariant (and possible definitional) properties of entities as they *are* (ontology) and how they are *seen / known / recorded* (epistemology) is prevalent in most biomedical terminology systems [8, 9].

Ontologies of Information Entities

Whether these epistemic aspects are considered relevant for ontology is a matter of definition. In the Information Artifact Ontology, under the OBO Foundry initiative [10], they are included in an ontology framework as information content entities, and their classes have representations of information as members. Information content entities are immaterial objects (more precisely: generically dependent continuants according to the Basic Formal Ontology, BFO [11]) that can be borne in material objects. So can the latter be a photographic print, and the former an (immaterial) photograph:

$$\mathit{PhotographicPrint} \equiv \mathit{MaterialEntity} \sqcap \exists\, \mathit{bearerOf}.\ (\exists\, \mathit{isConcretizationOf}.\ \mathit{Photograph})$$

Information content entities encompass documents, document parts such as sentences, texts, data, measurement results, serial numbers, datatypes, databases, and ontologies, and the processes in which they are created and consumed, totaling 131 classes. Information content entities are related by the relation *isConcretizationOf* to their material bearers, and by the relation *isAbout* to the things they denote.

There is a rough correspondence between IAO information content entities and the HL7 classes that derive from the class *Act*. In this context, *Act*, in contrast to its implicit meaning is to be understood as an information entity, i.e. *information about* a real act. This becomes obvious by the fact that HL7 acts can be modified by so-called mood or uncertainty codes.

The so-called *moodCode* in the information model distinguishes between acts that occurred and acts that are only planned (ordered, scheduled, rescheduled, etc.). Mood codes encompass intent, appointment, appointment request, promise, proposal, recommendation, resource slot, predicate, criterion, event criterion, expectation, goal, option, permission, permission request, risk.

The *uncertaintyCode* indicates whether the *Act* statement as a whole, with its subordinate components has been asserted to be uncertain in any way e.g., a patient might have had a cholecystectomy procedure in the past (but is not sure). When the uncertainty is associated with an *Observation.value* alone or other individual attributes of the class, such pointed indications of uncertainty should be specified by applying the *Uncertain Value – Probabilistic* (UVP)[1]or the *Parametric Probability Distribution* (PPD)[2] data type extensions to the specific attribute. Particularly if the uncertainty is uncertainty of a quantitative measurement value, this must still be represented by a PPD<PQ> in the value and NOT using the *uncertaintyCode*. Also, when differential diagnoses are enumerated or weighed for probability, the UVP<CD> must be used, not the *uncertaintyCode*. The use of the *uncertaintyCode* is appropriate only if the entirety of the *Act* and its dependent *Acts* is questioned. Finally, the attribute *negationInd* indicates that the *Act* statement is a negation of the *Act* as described by the descriptive attributes. For example, to test for "systolic blood pressure of 90-100 mm Hg," one would use only the descriptive attributes *Act.code* (for systolic blood pressure) and *Observation.value* (for 90-100 mm Hg). If one would also specify an *effectiveTime*, i.e., for "yesterday," the criterion would be more constrained. If the *negationInd* is true for the above criterion, then the meaning of the test is that a systolic blood pressure of 90-100 mm Hg yesterday does **not exist** (independent of whether any blood pressure was measured).

The IAO does not have so far a fine grained model of moods and probabilities such as the HL7 RIM, but its architecture does not preclude such an extension.

These examples show the crucial difference between a model of information and a model of reality. In the former, "*information related to an act*" can be subsumed by "*information related to a planned act*", whereas in a model of reality, i.e. an ontology in a narrower sense "*act*" and "*planned act*" are not related by taxonomic subsumption.

In the following we are studying several SNOMED CT concepts that clearly belong to the category of information entities. We critique their current representation and propose an alternative representation as information content entities.

Case Study

We center our forthcoming discussion on four SNOMED term cases (C1-C4) which, in our view, represent epistemic states rather than ontological concepts:

C1: ***Absent nose*** (111317000) is stated to imply: *Congenital malformation* ⊓ ∃ *FindingSite. Nasal Structure*

C2: ***Heart operation planned*** (183983001)[3]. This concept is in SNOMED CT's *Situation with explicit context* branch and is fully defined as

∃ *rg.*(
∃ *Associated procedure.Operation on heart* ⊓
∃ *Procedure context.Planned* ⊓
∃ *Temporal context. Current or Specified* ⊓
∃ *Subject relationship context. Subject of record*)

C3: ***Operation on heart, rescheduled***. (64915003|: 272125009|=58334001), This is a postcoordinated concept, refining operation on heart by using the qualifier *Priority* with the value *Rescheduled*, in DL notation: *Operation on heart* ⊓ ∃ *Priority. Rescheduled.*

C4: ***Suspected gallstones*** (390926006). This concept is also in SNOMED CT's *Situation with explicit context* branch and is fully defined as

∃ *rg.*(∃ *Associated finding.Gallstone* ⊓
∃ *Finding context.Suspected* ⊓
∃ *Temporal context. Current or Specified* ⊓
∃ *Subject relationship context. Subject of record*)

Case critique

All four concepts have in common that in their definition they are related to other concepts that are definitely not, or not nec-

[1] A generic data type extension used to specify a probability expressing the information producer's belief that the given value holds.

[2] A generic data type extension specifying uncertainty of quantitative data using a distribution function and its parameters (mean, standard deviation)

[3] *rg* means „role group", cf [12].

essarily, instantiated. SNOMED CT's description logics notation, however, by using existentially quantified roles ($\exists$), asserts the existence of at least one instance of the concepts in question. So does the expression

$\exists$ *FindingSite. Nasal structure* formally assert that some instance of *Nasal structure* exists, whereas the intended meaning is exactly the contrary. Similarly, the expression *Operation on heart* $\sqcap \exists$ *Priority. Rescheduled* states that there is a heart operation, whilst the intended meaning refers to some heart operation in the future, which still includes the case that there will not be any operation at all (e.g. due to worsening conditions of the patient). The same argument holds for the planned heart operation. Regardless the syntactic difference (the rescheduled operation is a operation, whilst the planned operation isn't), the expression

$\exists$*rg.* ($\exists$ *Associated procedure.Operation on heart*)

is a necessary condition for *Heart operation planned*, i.e. the plan implies its execution, which is certainly not always the case. In exactly the same way, the definition of *Suspected gallstones* leads to the conclusion that there exist real gallstones even in case a doctor registers a suspicion only.

What is wrong with these concept definitions? There is no doubt that there must be a way to refer to "something" which does not exist now, which existed in the past, or which may exist in the future. But statements about non-existence are not terms, although they syntactically include terms. Ideally, they should be represented in an information model, which is distinct from the ontology, or is expressed in an "information Entity" branch in the same ontology. However, there are strong reasons why application builders want to have "real" concepts as well as whole assertion in one and the same representational artifact such as SNOMED CT. So has it been a precondition for the use of this standard with in the UK National Health Service, that the former CTV3 terminology was fused with SNOMED RT. One characteristics of CTV3 (the successor of the former Read Codes) was its abundance of epistemic laden concepts such as in our examples.

Case remodeling

We here propose alternative representations based on the information artifact ontology, using information content entities such as Plan and Suspicion. All the four concepts *Absent nose*, *Heart operation planned*, *Operation on heart, rescheduled*, and *Suspected gallstones* represent information content entities. In order to make this clear (and because the language is often misleading), we slightly rename the concepts to *Patient without nose, Plan of heart operation, Rescheduled plan of heart operation, Suspicion of gallstones.*

To express this adequately, we need to enhance our description language by the constructors given in Table 4.

A further extension of the logics including concrete domains (in this case numeric values) will be necessary if probabilistic values are to be represented such as UVP and PPD in HL7 RIM. This is already possible, e.g. using data properties in Protégé, but it is not yet covered by off-the-shelves terminological reasoners such as Fact++ and Pellet.

Table 4-Additional description logics constructors

DL Constructor		Meaning	Example
$\neg$	$\neg A$	Negation of *A*	*Base* $\sqsubseteq \neg$ *Acid*
$\forall$	$\forall r.G$	Value restriction of the relation *r* by the filler *G*	*Hand* $\sqsubseteq$ $\forall$*has-Laterality.* (*Left* $\sqcup$ *Right*)
$\sqcup$	$A \sqcup B$	Union of A with B	

Case remodeling

Coming back to the running examples, we propose the following representations:

C1: ***Person without nose***:

Human $\sqcap \neg$ *hasPart. Nasal Structure*

C2: ***Plan of heart operation*** (183983001):

Plan $\sqcap \forall$ *isAbout. Operation on heart*
with *Plan* being an information content entity. The universal quantifier $\forall$ means that this plan can only be realized by a heart operation. In contradistinction to the existential quantifier $\exists$ the formula does not assert that there must be an operation for each and every plan.

C3: ***Rescheduled plan of heart operation***:

Plan $\sqcap \forall$ *isAbout. Operation on heart* $\sqcap$
$\exists$ *hasQuality.Rescheduled*

Alternatively:

Plan $\sqcap$ ($\forall$ *isAbout. Operation on heart*) $\sqcap$
$\exists$ *participantOf.Rescheduling*
with *Rescheduling* being an event.

C4: ***Suspicion of gallstones***.

Suspicion $\sqcap \forall$ *isAbout. Gallstones*
with *Suspicion* being an information content entity.

Variations

There may be a need to distinguish simple instantiations (e.g. asserting that there is an instance of *Gallstones*) from a record of a finding (i.e. that some physician has diagnosed gallstones).

Note that all version of SNOMED CT until now, have placed *Gallstones* (a material entity), together with processes like *Myocardial infarction*, *Headache* and *Hypercholesterolemia* into an epistemology-infested *Findings* hierarchy.

The subtle difference between instantiations and findings is that there are undiagnosed diseases just as there are false diagnoses (which continue being diagnoses even being false). These special cases should be accounted for in a medical record, and the terminology should provide the means for this. We propose a solution using again the example C4.

We may want to distinguish between:

- C4a: A diagnosis "Gallstones" whatsoever
- C4b: A confirmed diagnosis "Gallstones"
- C4c: A suspected diagnosis "Gallstones"
- C4d: A false diagnosis "Gallstones"
- C4e: Gallstones that have not been diagnosed

In all these cases diagnoses are information content entities. According to the diagnosing person they can be subdivided in terms of medical diagnosis, nursing diagnoses, etc.

C4a: $\mathit{Diagnosis} \sqcap \forall\ \mathit{isAbout.Gallstones}$

C4b: $\mathit{Diagnosis} \sqcap \forall\ \mathit{isAbout.Gallstones} \sqcap \exists\ \mathit{isAbout.\ Gallstones}$

C4c: $\mathit{Diagnosis} \sqcap \forall\ \mathit{isAbout.\ Gallstones} \sqcap \exists\ \mathit{hasQuality.Suspected}$

C4d: $\mathit{Diagnosis} \sqcap \forall\ \mathit{IsAbout.} \perp$

C4e: $\mathit{Gallstones} \sqcap \neg\exists\ \mathit{inv(IsAbout).Diagnosis}$

The examples show the possibilities but also the limitations of using the proposed description logics. If we wanted to represent quantitative statements, e.g. in C4c that there is a probability of 0.1 that the diagnosis is true, then we would need to include numeric values as data properties. As C4d shows, there is no possibility to distinguish between different kinds of false diagnoses. From a HL7 point of view, the establishment of a diagnosis is an observation, a sub-class of the class *Act* defined as "An act that is intended to result in new information about a subject." Being a sub-class of the class *Act*, the class Observation inherits of the attributes of the class Act including *moodCode*, *uncertaintyCode* and *negationInd*. In addition UVP or PPD data type extension may be used to express respectively a probability expressing the information producer's belief that the given qualitative observation value holds or the uncertainty of quantitative data using a distribution function and its parameters.

Conclusion

Numerous SNOMED CT concepts are representations that are more adequately described by complex linguistic statements than by domain terms in a stricter sense. These complex statements address epistemic notions, i.e. information about the user and the context, which clearly extends the realm of ontology. Those SNOMED CT concepts that correspond to "real" terms can generally be defined using the very inexpressive logic **EL**, currently used for SNOMED CT.

In our finding that there are numerous SNOMED "non-term" concepts that cannot be adequately represented giving the current restrictions of SNOMED CT's logic, we are close to the analysis done by Rector & Brandt [13]. Just as we do, they defend the (controlled) use of a more expressive description logic, analyzing a similar scope of concepts as we do. However, the model they propose is different. By understanding findings, procedures, and observables as *situations* they manage to solve the negation problem. Yet their approach reaches short when it comes to uncertainty, such as speculative diagnoses, or plans that have not yet been executed at the time of recording.

Our approach comes closer to what is possible to encode using the HL7 RIM, where medical record entries can be modified in terms of "mood codes" like *Event, Goal, Risk, Expectation, Intent*, or uncertainty codes such as *Possibly done* or *Probably done*. It is also consistent with the Information Artifact Ontology, which, however, lacks detail for representing diagnostic statements. Thus, using one single representation formalism, our proposal brings different worlds together: real-world, heterogeneous terminologies, HL7 information models, as well as philosophically founded ontologies.

Acknowledgements:

This work was funded by the EU 7th FP project DebugIT (ICT-2007.5.2-217139).

References

[1] IHTSDO (Intern. Health Terminology Standards Development Organisation). Systematized Nomenclature of Medicine - Clinical Terms. http://www.ihtsdo.org/snomed-ct. Last accessed: March 2^{nd}, 2010.

[2] ISO 1087. Terminology Work. International Standards Organization.

[3] Wermter J. Defining and Collocations and Terms. In: Wermter J. Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods, chapter 2. 2009, http://www.dart-europe.eu/full.php?id=159166. Last accessed: March 2^{nd}, 2010.

[4] Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The Description Logic Handbook. Theory, Implementation, and Applications (2nd Edition). Cambridge: Cambridge University Press, 2007.

[5] Hofweber T. Logic and Ontology, Stanford Encyclopaedia of Philosophy, 2004. Available from: http://plato.stanford.edu/entries/logic-ontology. Last accessed: March 2^{nd}, 2010.

[6] HL7 version 3, Jan 2009 ballot package, 2009, http://www.hl7.org/v3ballot2009jan/html/welcome/environment/index.htm. Last accessed: March 2^{nd}, 2010.

[7] Garde S, Knaup P, Hovenga E, Heard S. Towards semantic interoperability for electronic health records. Methods Inf Med. 2007; 46(3): 332–343.

[8] Ingenerf J, Linder R. Assessing applicability of ontological principles to different types of biomedical vocabularies. Methods Inf Med. 2009; 48(5): 459–467.

[9] Bodenreider O, Smith B, Burgun A (2004). The Ontology-Epistemology Divide: A Case Study in Medical Terminology. Int. Conf. on Formal Ontology and Information Systems (FOIS 2004). Amsterdam: IOS-Press, 185–195.

[10]IAO Information Artifact Ontology. http://code.google.com/p/information-artifact-ontology/ Last accessed: March 2^{nd}, 2010.

[11]Basic Formal Ontology. http://www.ifomis.org/bfo. Last accessed: March 2^{nd}, 2010.

[12]Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp. 2002: 712–716.

[13]Rector AL, Brandt, S. Why Do It the Hard Way? The Case for an Expressive Description Logic for SNOMED. JAMIA 2008; 15: 744–751.

Address for correspondence

Stefan Schulz, IMBI, University Medical Center Freiburg, Stefan-Meier-Str. 26, D-79104 Freiburg, Germany. Email: stschulz@uni-freiburg.de