MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI230932

Setting the Scene to Link SNOMED CT to Realism-Based Ontologies

Anuwat PENGPUT[a,1] and Werner CEUSTERS[a]
[a]*Department of Biomedical Informatics, University at Buffalo, USA*
ORCiD ID: Anuwat Pengput https://orcid.org/0000-0002-0273-1531, Werner Ceusters https://orcid.org/0000-0002-2676-8689

Abstract. In a proof of concept study, we assessed the feasibility of designing a first-order logic (FOL) framework capable of translating SNOMED CT's terminological view on patient data as referencing *concepts*, into the realism-based view of the Basic Formal Ontology and the Ontology for General Medical Science according to which patient data represent *instances of types*. Because within the subject domain of this study, SNOMED CT's terminological coverage was excellent, and its EL++ axioms can be automatically translated into FOL as well as the antecedent part of bridging axioms between SNOMED CT and realism-based ontologies, we conclude that this is an area of R&D that deserves further attention and that may lead to new ways of federating terminologies with ontologies.

Keywords. SNOMED CT, Basic Formal Ontology, CLIF

1. Introduction

SNOMED CT is a large clinical terminology organized on the basis of *meanings*. The meaning of each *term* is either provided through an individual *concept*, or by at least one *axiom* expressed in the description logic (DL) language EL++ [1]. These axioms are used by a classifier to organize the terms in a directed acyclic graph with the goal to check the internal coherence and consistency of the term collection with respect to the underlying concept model. The reliability of this approach is in general determined by the expressivity of the DL used, the experience of authors in crafting axioms without violating the syntax and semantics of the DL, and by the capability of the classifier to deal with the expressivity [2]. However, neither logic or logical axioms used to express meanings can guarantee that the concept model which serves as foundation for a terminology – or even ontology – is faithful to the portion of reality one intends to describe [3]. Faithfulness to reality requires the concepts to have at least clear counterparts in reality, i.e. to have *ontological commitment*. Terms such as 'black bile' and 'miasma' do have a meaning, and their meaning can be expressed using logic, but they do not have a counterpart in reality. The lack of ontological commitment or at least the fuzziness thereof in SNOMED CT has regularly been pointed out [4,5].

The Basic Formal Ontology (BFO) – now an ISO standard [6] and fully axiomatized in First Order Logic (FOL) – and ontologies derived therefrom such as the Ontology for

[1] Corresponding Author: Anuwat Pengput, Department of Biomedical Informatics, University at Buffalo, 701 Ellicott St, Buffalo NY, 14203, USA, email: anuwatpe@buffalo.edu.

General Medical Science (OGMS) [7] differ from SNOMED CT in having a very precisely defined ontological commitment. Terms in BFO and OGMS are not organized on the basis of meanings, but on the basis of *types* identified following a domain-independent ontological theory [8]. The resulting typology is quite different from SNOMED CT's concept model which is primarily clinical practice oriented, thereby including pragmatic considerations (e.g. clinicians must be able to easily find their way in it to annotate patient data) and epistemological ones (e.g. when something is annotated as a disease, it must have been found, thus diseases are *clinical findings*) while BFO only includes ontological considerations (e.g. diseases must exist prior to finding them).

An interesting question is whether an integration between SNOMED CT and BFO-based ontologies such as OGMS might bring benefits. Both Hogan and Schulz argue it is at least desirable, but where the former claims it is impossible because of mutual incompatibility of the different underlying categorization principles [9], the latter sees room for a partial integration as exemplified by his proposal to reinterpret SNOMED CT's clinical finding concepts as denoting something that for BFO would be an occurrent [10]. El-Sappagh proposed a mapping from SNOMED CT's upper level concepts to BFO and OGMS [11], but this effort is seriously flawed, in part due to misunderstandings about the underlying models on either side, and in part due to the limitations of the description logic used for the mapping. We, too, believe that a lot can be gained from combining SNOMED CT's terminological richness with the realism-based logic offered by BFO and OGMS, be it not by means of a direct integration. To justify our belief, we started a proof of concept study to assess the feasibility of, and effort required to design a logical framework that acts as a mediator and is able to coherently reference BFO-based ontologies on the one hand, and SNOMED CT on the other hand.

2. Methods

Data elements from 551 subjects from the Kalasin Province in Thailand that participated in the *Cholangiocarcinoma Screening and Care Program* (CASCAP) established by the Khon Kaen University [12], were merged with available clinical data obtained from the Kalasin Provincial Public Health Office (UB-IRB approval STUDY00006059 of June 9, 2022). Source data consisted of (1) diagnoses expressed in ICD10, (2) values selected from a controlled vocabulary designed for reporting verbal screening and echography findings, and (3) free text diagnoses and patient occupations not covered in the controlled vocabularies but provided as comments to data entry fields labeled 'other …'. Following the Referent Tracking guidelines [13], each data element was considered to be a reference to an entity – or configuration of entities – in the real world, each entity being of a *most specific type* (MST) referenceable in a realism-based ontology. The next steps were then: (1) to construct for each MST a *corresponding meaning expression* (CME) using only SNOMED CT codes, either by selecting a perfectly matching code, or through post-coordination in line with SNOMED CT's machine-readable concept model (MRCM), thereby keeping track of any problem encountered; (2) to extract for each SNOMED CT code used all active relationships and, recursively, also of all their subsumers; (3) to create in FOL, using the same dialect of the Common Logic Interchange Format (CLIF) as used for the BFO, for each MST axioms that bridge the realism-based view with SNOMED CT's concept-based view; and finally (4) to classify the problems encountered in a number of meaningful categories for each one of which possibly a general remediation strategy might be designed.

3. Results

Table 1 shows for the various components how many MSTs were found and in what way they were translated into SNOMED CT by using either a single code or post-coordination, and in the latter case, with or without violation of the MRCM rules. Violations were introduced when codes sensibly to be used were available, yet not within the domain-range restrictions of any sensible attribute. The table also shows how many of the MSTs could be precisely described using available SNOMED CT codes. Thanks to SNOMED CT's compositional nature, and the substantial occurrence of 'history of' and 'absence of' in the source data, only 366 codes were required for the translation. The transitive closure set for these codes over all relationship types amounted to 11.149 relationship records covering 1680 additional distinct SNOMED CT codes including 36 attributes. Table 2 provides three example axioms to demonstrate how in one logical framework SNOMED CT's concept-based view is used together with BFO's references to types and particulars without any need for direct mapping between concepts and types.

Table 1. Translation of Most Specific Types (MST) from data into SNOMED CT codes or expressions.

Data Component	MST	Single code	Post-coordination		Precise
			No MRCM violation	MRCM violation	
Verbal screening					
controlled vocabulary	76	21 (28%)	48 (63%)	7 (9%)	73 (96%)
free text diagnoses	207	195 (94%)	12 (6%)		207 (100%)
free text occupation	95	95 (100%)			66 (69%)
Echography vocabulary	122	44 (36%)	67 (55%)	11 (9%)	93 (76%)
EHR ICD10 diagnoses	27	27 (100%)			27 (100%)
Total	527	382 (73%)	127 (24%)	18 (3%)	466 (88%)

Table 2. Example axioms in CLIF for bridging SNOMED CT to OGMS and BFO.

```
(forall (x y) (if (individual-of x y) (and (particular x) (concept y))))                     (axiom 1)
(forall (x)  (iff (individual-of x sctid-312104005-cholangiocarcinoma-of-biliary-tract)      (axiom 2)
                  (and (individual-of x sctid-64572001-disease)
                       (exists (y z) (and (sctid-363698007-finding-site x y)
                                          (individual-of y sctid-34707002-biliary-tract-structure)
                                          (sctid-116676008-associated-morphology x z)
                                          (individual-of z sctid-70179006-cholangiocarcinoma))))))
(forall (x y z)                                                                               (axiom3)
   (if (and (individual-of x sctid-64572001-disease) (sctid-363698007-finding-site x y)
            (individual-of y sctid-34707002-biliary-tract-structure)
            (sctid-116676008-associated-morphology x z)
            (individual-of z sctid-70179006-cholangiocarcinoma))
       (and (= x z) (exists (rx ry t)
            (and (occupies-spatial-region x rx t) (occupies-spatial-region y ry t) (rcc-overlap rx ry t)
                 (instance-of x ogms-disorder t) (instance-of y ogms-bodily-component t))))))
```

4. Discussion

Table 1 shows that SNOMED CT offers very good coverage for diagnoses and occupations in the researched domain, with minimal post-coordination and near maximal preciseness. 'Preciseness' reflects the degree to which an MST can be expressed in SNOMED CT without information loss nor use of external concepts. Demographic screening requires more post-coordination (72%), while coverage and preciseness is less for reporting echography findings. These findings suggest that there is at least in this domain little need to acquire content from other terminologies. Doing so in our earlier work using the ROBOT tool for incorporating terms from the NCI Thesaurus led to too many inconsistencies [14]. Since our goal is to have a good ontological account of the domain, the few missing elements can better be defined directly following the principles of Ontological Realism [8], or, of course, added to SNOMED CT in a later version.

For our logical framework to be applicable, we assume that when an experienced biomedical ontologist strictly adhering to Ontological Realism and a gastroenterologist adhering to SNOMED CT are discussing a concrete medical case, they are referencing the very same entities on the side of the patient, for instance the carcinoma in the patient's biliary tract. That is so, even when they disagree about how that carcinoma, that patient and that gallbladder wall are to be appropriately classified, be it in SNOMED CT or in some realism-based ontology. Schulz's proposal is also based on this assumption [10], but whereas his approach is to have SNOMED CT adepts accept that certain entities are particulars that instantiate occurrents as per BFO's view, thus requiring f.i. 'carcinoma of gallbladder' to be interpreted as '*having* a carcinoma of the gallbladder', our approach is to let SNOMED CT's view and BFO's view happily co-exist; not in one *ontological* framework, but in one *logical* model-theoretic framework capable of exploiting what SNOMED CT offers *terminologically* and realism-based ontologies *ontologically*. This is set up as follows. Because individuals in BFO's domain of discourse enjoy either one of the unary relations *universal* or *particular*, we include in the domain of discourse of our framework – not in BFO's domain of course – individuals unary related as *concept*. What sorts of things universals and particulars are, is described in manuscripts about the BFO, but not in its axioms. That they are quite distinct is however explicitly axiomatized, as well as in what sorts of n-ary relations they can figure and in what argument positions thereof. A similar treatment is given to *concept*: whereas in BFO the time-indexed ternary *instance-of* relation is used to assert of what universal some particular is an instance during what temporal-region, we add the binary relation *individual-of* – one may quibble about the name – to assert under what concept some particular is classified (Table 2, axiom 1), while we leave it to the SNOMED CT authors to explain what they mean by 'concept'. It is then easy to automatically translate SNOMED CT's EL++ axioms into the sort of FOL-axioms of our framework as exemplified by axiom 2 in Table 2. Finally, axiom 3 in Table 2 exemplifies how the logical framework translates the application of a concept from SNOMED CT to a phenomenon on the side of a patient, in this case the presence of a cholangiocarcinoma, into a representation that (1) enumerates explicitly the entities that must exist for the use of the SNOMED CT concept to be faithful to reality, and (2) how these entities relate to each other. It is crafting axioms of this type, more precisely the consequent part, which turns out to be much more time-consuming. The reason is that there is no one-to-one mapping possible between concepts and types, nor between relations in either view. In the cholangiocarcinoma case of axiom 3, SNOMED CT requires three concepts to be referenced figuring two distinct attributes, whereas the realism-based view posits the existence of five entities amongst which also two, but totally different sorts of relations hold. Also, SNOMED CT's attribute *finding site* corresponds here with the time-indexed version of *overlap* in the region connection calculus (RCC) defined as any of the RCC-relations except *disjoint* and *touches*. But for a foreign object in the stomach, surely a different RCC-relation would be more appropriate, while also BFO's time-indexed *located-in* relation would work. Future work will include applying the approach in line with other BFO-based ontologies and express SNOMED CT content within their domains in exactly the same way.

5. Conclusions

Though our study covers only a small domain, results are indicative for the feasibility of our approach. We thus believe that clinicians can continue to use SNOMED CT as they do now, and envision that the application of this sort of framework as an extension to SNOMED CT 'behind the scene' may lead to more powerful secondary analytics. Some might pity that the entire framework is beyond the capabilities of OWL-DL reasoners, but in light of the many inappropriate uses thereof [2], we argue this to be an advantage.

References

[1] Rodrigues JM, Schulz S, Mizen B, Trombert B, Rector A. Scrutinizing SNOMED CT's Ability to Reconcile Clinical Language Ambiguities with an Ontology Representation. Stud Health Technol Inform. 2018;247:910-4.

[2] Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Inform. 2019 Mar;100S:100002, doi: 10.1016/j.yjbinx.2019.100002.

[3] Smith B. Against Fantology. In: Reicher ME, Marek JC, editors. Experience and Analysis. Vienna: HPT&ÖBV; 2005. p. 153-70.

[4] 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.

[5] Schulz S, Cornet R, Spackman K. Consolidating SNOMED CT's ontological commitment. Appl Ontol. 2011;6(1):1-11, doi: 10.3233/AO-2011-0084.

[6] International Standards Organisation. ISO/IEC 21838-2:2021 Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO). 2021.

[7] Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit Transl Bioinform. 2009;2009:116-20.

[8] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010;5(3-4):139-88. doi: 10.3233/AO-2010-0079.

[9] Hogan W. Aligning the top-level of SNOMED-CT with Basic Formal Ontology. Nature Precedings. 2008. doi: 10.1038/npre.2008.2373.1.

[10] Schulz S. SNOMED CT x BFO: can the gap between legacy terminology and foundational ontology be bridged? ICBO/FOIS; September 16, 2021; Bolzano, Italy2021.

[11] El-Sappagh S, Franda F, Ali F, Kwak KS. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak. 2018;18(1):76. doi: 10.1186/s12911-018-0651-5.

[12] Khuntikeo N, Chamadol N, Yongvanit P, Loilome W, Namwat N, Sithithaworn P, et al. Cohort profile: cholangiocarcinoma screening and care program (CASCAP). BMC Cancer. 2015;15:459. doi: 10.1186/s12885-015-1475-7.

[13] Ceusters W. The Place of Referent Tracking in Biomedical Informatics. In: Elkin PL, editor. Terminology, Ontology and their Implementations: Teaching Guide and Notes. Cham: Springer International Publishing; 2022. p. 39-46.

[14] Pengput A, Diehl AD. Ontology Representation for Cholangiocarcinoma. International Conference on Biomedical Ontology; September 25-28; Ann Arbor, 2022.