# Replacing SEP-Triplets in SNOMED CT using Tractable Description Logic Operators

Boontawee Suntisrivaraporn[1]*, Franz Baader[1], Stefan Schulz[2], Kent Spackman[3]

[1]TU Dresden, Germany, {meng,baader}@tcs.inf.tu-dresden.de
[2]Freiburg University Hospital, Germany, stschulz@uni-freiburg.de
[3]Oregon Health & Science University, USA, spackman@ohsu.edu

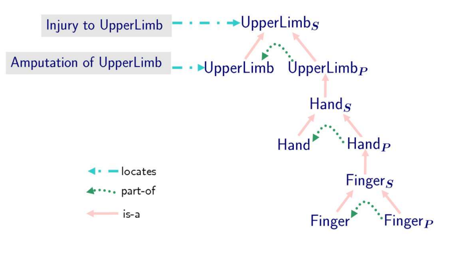
**Abstract.** Reification of parthood relations according to the SEP-triplet encoding pattern has been employed in the clinical terminology SNOMED CT to simulate transitivity of the part-of relation via transitivity of the is-a relation and to inherit properties along part-of links. In this paper we argue that using a more expressive representation language, which allows for a direct representation of the relevant properties of the part-of relation, makes modelling less error prone while having no adverse effect on the efficiency of reasoning.

## 1 Introduction

Description logics (DLs) [1] are a successful family of knowledge representation formalisms, which can be used to represent and reason about ontologies in a logically well-founded way. The Systematized Nomenclature of Medicine, Clinical Terms (SNOMED CT) [2] is a clinical terminology with a broad coverage of health care, which has been developed with the help of a rather inexpressive description logic dialect known as $\mathcal{EL}$ [3]. In $\mathcal{EL}$, one can build class descriptions using the operators *conjunction* ($C \sqcap D$) and *existential restriction* ($\exists r.C$). For example, the $\mathcal{EL}$ class description $\mathsf{Inflammation} \sqcap \exists \mathsf{has\text{-}location}.\mathsf{Appendix}$ describes a kind of inflammation characterized by its location being in some appendix. This description can be used as a definition (expressed by the DL symbol $\equiv$) for appendicitis: it constitutes both necessary and sufficient conditions for classifying a real world entity as being an instance of appendicitis. Classes defined this way are said to be *fully defined.* If only necessary conditions are given for a class, it is called *primitively defined* (expressed by the DL symbol $\sqsubseteq$). For instance, $\mathsf{LeftHand} \sqsubseteq \mathsf{BodyPart} \sqcap \mathsf{LeftLateral}$ is such a primitive definition.

DL systems provide their users with automated reasoning services, which can be used to infer implicit knowledge from the explicitly represented knowledge. In particular, they can *classify* an ontology, i.e., compute all the implied is-a relationships (i.e., subclass/superclass relationships, expressed by the *subsumption* symbol $\sqsubseteq$) between (names of) fully or primitively defined classes. The advantage of using the inexpressive DL $\mathcal{EL}$ for developing SNOMED CT is that classification is *tractable* (i.e., the is-a hierarchy can be computed in polynomial

* Funded by the German Research Foundation (DFG) under grant BA 1122/11-1.



$$\begin{array}{ll}
\mathsf{UpperLimb} & \sqsubseteq \mathsf{UpperLimb}_S \\
\mathsf{UpperLimb}_P & \sqsubseteq \mathsf{UpperLimb}_S \sqcap \\
 & \quad \exists \mathsf{part\text{-}of}.\mathsf{UpperLimb} \\
\mathsf{Hand}_S & \sqsubseteq \mathsf{UpperLimb}_P \\
\mathsf{Hand} & \sqsubseteq \mathsf{Hand}_S \\
\mathsf{Hand}_P & \sqsubseteq \mathsf{Hand}_S \sqcap \exists \mathsf{part\text{-}of}.\mathsf{Hand} \\
\mathsf{Finger}_S & \sqsubseteq \mathsf{Hand}_P \\
\mathsf{Finger} & \sqsubseteq \mathsf{Finger}_S \\
\mathsf{Finger}_P & \sqsubseteq \mathsf{Finger}_S \sqcap \exists \mathsf{part\text{-}of}.\mathsf{Finger}
\end{array}$$

**Fig. 1.** Complete SEP-triplets in SNOMED CT.

time). Efficiency and scalability of reasoning are very important for an ontology of the size of SNOMED, with about 370,000 classes. The disadvantage is that not all relevant properties can be explicitly expressed. In particular, $\mathcal{EL}$ does not allow to state that relations such as part-of are transitive, and consequently the reasoner does not take transitivity into account during classification. For example, even if the finger is defined to be part of the hand, and the hand to be part of the upper limb, an $\mathcal{EL}$ reasoner cannot deduce that the finger is part of the upper limb since it does not "know" that part-of is supposed to be transitive.

In order to overcome such limitations in DLs without transitive relations, the SEP-triplet encoding was proposed in [4]. In the next section, we will briefly sketch this approach, and also show that, in addition to transitivity reasoning, it can encode inheritance of properties along part-of links. For example, injury to finger can thus be classified as subclass of injury to hand. We will then point out some disadvantages of the SEP-triplet encoding, and propose to replace SEP-triplets by the direct representation of transitive relations in the DL $\mathcal{EL}^+$ [5]. In addition to transitive relations, $\mathcal{EL}^+$ can also express so-called right-identity rules [6], which can be used to explicitly represent inheritance of properties along part-of and other relations. In spite of its higher expressive power compared to $\mathcal{EL}$, reasoning in $\mathcal{EL}^+$ is still tractable [3, 5]. In fact, we will see that not only does the replacement make the classification reasoning faster, but it also helps simplify the ontology structure and thus ease the modelling and maintenance.

## 2 SEP-Triplets in SNOMED CT

SEP-triplets are extensively employed in the anatomical part of SNOMED CT. Figure 1 illustrates the encoding technique with an example. The left-hand side of the figure provides a graphical representation, whereas the right-hand side shows the formal representation in $\mathcal{EL}$. For every proper SNOMED class, called *entity* class (E-class) in the following, there are two auxiliary classes, the *structure* class (S-class) and the *part* class (P-class). In the example, we have three entity classes: Finger, Hand and UpperLimb, and thus three triplets. Intuitively, the E-class is supposed to be instantiated by entire anatomical objects (such as my hand), and the P-class by the proper parts of the referred objects (such as any part of my hand). The S-class, finally, is instantiated by both entire objects and their parts. This intuition explains the is-a links from the E-class and the P-class to the S-class, as well as the part-of link from the P-class to the E-class. The main idea underlying the SEP-triplet approach is to represent a part-whole relationship between two entity classes not by a part-of link between the E-classes, but rather by an is-a link between the S-class of the "part" and the P-class of the "whole". It should be noted, however, that the formal representation of the intuition underlying the three classes of the SEP-triplet approach is in fact limited to these links, and thus only consequences that follow from the presence of these links can be drawn. This is, however, sufficient to simulate transitivity of part-of through the inherently transitive relation is-a: $\mathsf{Finger} \sqsubseteq \mathsf{Finger}_S \sqsubseteq \mathsf{Hand}_P \sqsubseteq \mathsf{Hand}_S \sqsubseteq \mathsf{UpperLimb}_P \sqsubseteq \exists \mathsf{part\text{-}of}.\mathsf{UpperLimb}$ allows us to conclude that every finger is part of some upper limb.

Since characteristics are inherited along the is-a hierarchy, the SEP-triplet encoding also allows us to simulate inheritance of characteristics along the part-of hierarchy. In our example, by connecting an injury via a location link to the *S-class*, we can ensure that 'injury to finger' is classified as 'injury to hand' and 'injury to upper limb'. To suppress such inheritance along the part-of hierarchy (viz., 'amputation of finger' should not be classified as 'amputation of hand' or 'amputation of upper limb'), one needs to connect via location to the *E-class*.

There are, however, several problems with the SEP-triplet encoding. First, from a formal ontological point of view, it partially conflates the is-a hierarchy with the part-of hierarchy, which is dangerous since the two relationships are completely different by nature [7]. In SNOMED, it has indeed turned out that is-a links can be ambiguous, i.e., it is not always clear whether they are introduced as part of the SEP-triplet approach, or are supposed to represent a genuine generalization relationship. Second, the SEP-triplet approach is error prone since it works correctly only if it is employed with a very strict modelling discipline. In SNOMED, triplets are often modelled in an incomplete way, in particular, the P-class and the part-of link to it from the E-class are missing in most cases. In addition, the auxiliary S-class is often used as if it were a proper entity class; for instance, incorrect links to this class rather than the E-class may result in unintended consequences like the classification of 'amputation of finger' as a subclass of 'amputation of upper limb'. Third, the approach introduces for every proper class in the ontology two auxiliary classes, which results in a drastic increase in the ontology size.

## 3 Replacing SEP-Triplets by Using the DL $\mathcal{EL}^+$

The DL $\mathcal{EL}^+$ extends $\mathcal{EL}$ with relation inclusions of the form $r_1 \circ \ldots \circ r_n \sqsubseteq s$, which express that the composition of the relations $r_1, \ldots, r_n$ must be interpreted as a subset of the relation $s$. These inclusions generalize several expressive means useful in bio-medical ontologies: *(i)* transitivity of $r$ as $r \circ r \sqsubseteq r$, *(ii)* reflexivity of $r$ as $\epsilon \sqsubseteq r$ (where $\epsilon$ stands for the empty composition), *(iii)* relation hierarchies as $r \sqsubseteq s$, and *(iv)* right-identity rules as $r \circ s \sqsubseteq r$. It has been shown in [3] that

$$\mathsf{Finger} \sqsubseteq \mathsf{BodyPart} \sqcap \exists \mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand} \tag{1}$$

$$\mathsf{Hand} \sqsubseteq \mathsf{BodyPart} \sqcap \exists \mathsf{proper\text{-}part\text{-}of}.\mathsf{UpperLimb} \tag{2}$$

$$\mathsf{UpperLimb} \sqsubseteq \mathsf{BodyPart} \tag{3}$$

$$\mathsf{AmputationOfFinger} \equiv \mathsf{Amputation} \sqcap \exists \mathsf{has\text{-}exact\text{-}location}.\mathsf{Finger} \tag{4}$$

$$\mathsf{AmputationOfHand} \equiv \mathsf{Amputation} \sqcap \exists \mathsf{has\text{-}exact\text{-}location}.\mathsf{Hand} \tag{5}$$

$$\mathsf{AmputationOfUpperLimb} \equiv \mathsf{Amputation} \sqcap \exists \mathsf{has\text{-}exact\text{-}location}.\mathsf{UpperLimb} \tag{6}$$

$$\mathsf{InjuryToFinger} \equiv \mathsf{Injury} \sqcap \exists \mathsf{has\text{-}location}.\mathsf{Finger} \tag{7}$$

$$\mathsf{InjuryToHand} \equiv \mathsf{Injury} \sqcap \exists \mathsf{has\text{-}location}.\mathsf{Hand} \tag{8}$$

$$\mathsf{InjuryToUpperLimb} \equiv \mathsf{Injury} \sqcap \exists \mathsf{has\text{-}location}.\mathsf{UpperLimb} \tag{9}$$

$$\mathsf{proper\text{-}part\text{-}of} \circ \mathsf{proper\text{-}part\text{-}of} \sqsubseteq \mathsf{proper\text{-}part\text{-}of} \tag{10}$$

$$\mathsf{proper\text{-}part\text{-}of} \sqsubseteq \mathsf{part\text{-}of} \tag{11}$$

$$\mathsf{part\text{-}of} \circ \mathsf{part\text{-}of} \sqsubseteq \mathsf{part\text{-}of} \tag{12}$$

$$\epsilon \sqsubseteq \mathsf{part\text{-}of} \tag{13}$$

$$\mathsf{has\text{-}exact\text{-}location} \sqsubseteq \mathsf{has\text{-}location} \tag{14}$$

$$\mathsf{has\text{-}location} \circ \mathsf{proper\text{-}part\text{-}of} \sqsubseteq \mathsf{has\text{-}location} \tag{15}$$

**Fig. 2.** A re-engineered extract of SNOMED CT without SEP-triplets.

the presence of such axioms does not increase the complexity of reasoning—classification in $\mathcal{EL}^+$ is still tractable.

When replacing the SEP-triplet encoding by the direct representation of transitivity of the part-of relation, we must be careful not to disrupt the rest of the ontology. Especially since the proper classes representing entire anatomical objects as well as the auxiliary S- and P-classes are used by definitions in other parts of the ontology, we must still be able to describe them if needed. Most importantly, we must be able to deduce the same consequences from the direct representation that could be drawn from the SEP-triplet encoding.

Figure 2 shows the part of the re-engineered ontology that corresponds to our example. First, note that we now distinguish between the part-of relation (which is reflexive and transitive) and the proper-part-of relation (which is transitive and a sub-relation of part-of).[1] The direct representation of transitivity allows us to draw the same consequences as in the SEP-triplet approach (e.g., that the finger is part of the upper limb), but dispenses with the auxiliary classes. Whenever any of the P- and S-classes are needed (e.g., since they occur in other parts of the ontology) they can be pre-coordinated as fully defined classes, as illustrated here for the class hand: $\mathsf{Hand}_P \equiv \exists \mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand}$ and $\mathsf{Hand}_S \equiv \exists \mathsf{part\text{-}of}.\mathsf{Hand}$. Note that we need no explicit is-a relationships among the three nodes in a triplet. Because part-of is reflexive, it is inferred that $\mathsf{Hand} \sqsubseteq \exists \mathsf{part\text{-}of}.\mathsf{Hand} \sqsubseteq \mathsf{Hand}_S$. Analogously, $\mathsf{Hand}_P \sqsubseteq \exists \mathsf{proper\text{-}part\text{-}of}.\mathsf{Hand} \sqsubseteq \exists \mathsf{part\text{-}of}.\mathsf{Hand} \sqsubseteq \mathsf{Hand}_S$, since part-of is a super-relation of proper-part-of.

In order to allow for inheritance of characteristics along the proper-part-of hierarchy, we must explicitly state this inheritance property by a right-identity rule (see (15) in Fig. 2). To avoid unintended inheritance of characteristics (e.g.,

[1] A more precise modelling, which expresses that part-of has to be interpreted as reflexive closure of proper-part-of is not possible since it would cause intractability.

in the case of amputation), we use two distinct relations: has-location, which is inherited from a part to its whole, and has-exact-location, a sub-relation of has-location, which is not inherited that way. Intuitively, has-exact-location associates an event with a location in which it happens as a whole, for instance, 'amputation of upper limb' happens exactly to the upper limb as a whole and not just any part of it. In contrast, has-location relates an event to any containing spatial location it occurs in, i.e., either part or whole of the specified location. For instance, 'injury to upper limb' happens to the upper limb as a whole or any of its parts.

The proposed re-engineering has been put into practice by experimenting with the anatomy fragment of SNOMED CT. Although the SEP model has been adopted in SNOMED CT, it is incomplete in the sense that many SEP-triplets consist of only one or two nodes, and the correct is-a and part-of links are not always present. For this reason, it required a considerable effort to locate and complete all triplets, in order to enable a correct replacement. However, the obtained results are quite promising: by our re-engineering, the number of anatomical classes dropped from 54,380 to 18,125, and the time needed by our CEL reasoner (version 0.94) [5] from 900.15 seconds to 18.99 seconds. An empirical analysis of our proposed re-engineering of the entire SNOMED CT ontology still needs to be done, however. In particular, this will show how the introduction of right-identity rules to enable inheritance of characteristics along the aggregation hierarchy and the introduction of two different relations for location influence classification time.

## References

1. F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook. Theory, Implementation, and Applications.* Cambridge, U.K.: Cambridge University Press, 2003.
2. SNOMED *Clinical Terms.* Northfield, IL: College of American Pathologists, 2006.
3. F. Baader, S. Brandt, and C. Lutz. Pushing the $\mathcal{EL}$ envelope. In *Proc. of the Nineteenth Int. Joint Conf. on Artificial Intelligence (IJCAI-05)*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.
4. S. Schulz, M. Romacker, and U. Hahn. Part-whole reasoning in medical ontologies revisited: Introducing SEP triplets into classification-based description logics. In C. G. Chute, editor, *Proc. of the 1998 AMIA Annual Fall Symposium* pages 830–834. Hanley & Belfus, 1998.
5. F. Baader, C. Lutz, and B. Suntisrivaraporn. CEL—a polynomial-time reasoner for life science ontologies. In U. Furbach and N. Shankar, editors, *Proc. of the 3rd Int. Joint Conf. on Automated Reasoning (IJCAR'06)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 287–291. Springer-Verlag, 2006.
6. K. A. Spackman. Managing clinical terminology hierarchies using algorithmic calculation of subsumption: Experience with SNOMED-RT. 2000. OHSU Technical Report.
7. J. Patrick. Aggregation and generalisation in SNOMED CT. In *Proc. of the 1st Semantic Mining Conf. on SNOMED CT*, Copenhagen, Denmark, 2006.