AUSTRALIAN E-HEALTH RESEARCH CENTTR
www.csiro.au

A Note on Improving SNOMED CT Modelling Quality

Michael Lawley, Australian e-Health Research Centre, CSIRO

When modelling a concept there are often several ways to say the same thing. For example, Acute Appendicitis may be defined in terms of its proximal primitive parent, *Disease*, and the two grouped relationships: *Associated morphology = Acute inflammation* and *Finding site = Appendix structure*. Alternatively, it could be modelled as a child of *Appendicitis* with the refined relationship *Associated morphology = Acute inflammation*.

From a semantic point of view, these two approaches are equivalent. However, from a modeller's perspective there is a subtle distinction that is reflected in the amount and kind of work required when parts of the hierarchy get re-modelled. Specifically, in the first case, if *Appendicitis* gets re-modelled (directly or indirectly), then the equivalent changes need to also be made to *Acute Appendicitis* (and all other descendants of *Appendicitis*). On the other hand, the second approach requires no remodelling of *Acute Appendicitis*.

There is a tension here: modelling concepts in terms of their proximal primitive parents allows the Classifier to do the work of constructing the hierarchy, but it means that a significant amount of manual re-modelling may be required to propagate modelling changes (along with the associated effort of first identifying which concepts need to be remodelled). On the other hand, modelling in terms of other (closer) parents runs the risk of over-modelling concepts and stated subsumption relationships that are wrong, or need to be changed in response to restructuring/re-modelling elsewhere.

This raises the question of whether there are any heuristics to guide the choice of when to use the default strategy of modelling in terms of proximal primitive parents, and when to use a different modelling style.

An alternative perspective is to extend the modelling process beyond that of just encoding a stated form for a concept. In addition, the modeller would specify a set of invariants that should always be true for the modelled concept. In our example above, the modeller might state that *Acute Appendicitis* must always be a descendant of *Appendicitis*. Other invariants might state that the modelled concept must be disjoint with one or more other specified concepts.

This approach of allowing the modeller to record machine readable (and therefore checkable) invariants means that there is a much richer statement of how the modeller has conceived the meaning of the concept. This is very useful for later maintenance of the concept if and when remodelling is being undertaken by a different person. In addition, it provides a clear set of automatically checkable regression tests to be run against the inferred form of the terminology. Not only will these tests help preserve the quality of the terminology, but they can be used to identify the sets of concepts that may require consequential re-modelling. There is a strong correlation here with the benefit Unit Testing in software development provides for software refactoring and maintenance.

It is anticipated that the SNOMED CT Expression Constraint Language would be an appropriate foundation for the expression of such invariants. One could express the invariants as Boolean constraints (e.g., “`<<Appendicitis AND Acute Appendicitis IS NOT EMPTY`” and “`<<Clinical Finding AND <<Procedure IS EMPTY`”), or as queries that return the set of concepts violating the constraint (e.g., “`<<Acute Appendicitis MINUS <<Appendicitis`” and “`<<Clinical Finding AND <<Procedure`”).

Following from the discussion above are the following recommendations:

1. Enable modellers to record machine computable invariants;
2. Ensure that these invariants are checked by the authoring tooling; and
3. Develop a set of context-specific modelling strategies, e.g., for flexibility in the choice of stated parents.

Adopting these recommendations will result in a more stable and *automatically* maintainable terminology.