17 September 2012

Jamie Ferguson

Peter Hendler

**Semantic Node Labeling**

**Summary:**

To solve the problems introduced when conflicting logic types exist in clinical information models, explicit labeling of model components for intensional logic can improve machine level interoperability and enable semantic web tools to be used safely.

**Introduction:**

Our first white paper on logic in clinical modeling[[1]](#footnote-1) described serious operational problems that occur when clinical information models confuse intensional and extensional logic without a boundary between the different types of logic in the model. Most clinical models based on the HL7, openEHR, and ISO 13606 families of standards use Object Oriented (OO) extensional logic based on the “Closed World Assumption” (CWA)[[2]](#footnote-2). These information models usually have query characteristics that are incompatible with the intensional logic used in SNOMED CT which is based on the “Open World Assumption” (OWA)[[3]](#footnote-3). Interoperability of clinical information between systems and between organizations requires explicit knowledge about where each logic type exists, i.e., to be explicit about when each logic type may be used. This paper proposes a method of identification of logic in clinical information models that may be used in multiple standards and modeling paradigms, and which would make it safe for SNOMED CT users to adopt the models by unambiguously labeling the model components in which the specific logic types reside.

**About clinical information in SNOMED CT**

SNOMED CT is based upon the description logic EL+ which is an intensional logic[[4]](#footnote-4) similar to the “Web Ontology Language” OWL. OWL is among the most widely used Semantic Web technologies[[5]](#footnote-5). To query data in these types of logic one does not use SQL, nor any other object query languages. Instead the intensional logic requires that queries are executed using software classifiers which are also known as tableaux reasoners, reason engines, rules engines, or simply “reasoners.” Many reasoners are available that are used in clinical decision support, clinical research, and clinical analysis with SNOMED CT. Using this type of logic reduces cost, adds value, and dramatically reduces the technical and operational effort to produce clinical analytical results by allowing for the inference of new information that mathematically and logically follows from the stated axioms of SNOMED CT and the clinical information model. Reasoners that are used with SNOMED CT cannot be used in any part of a clinical information model that is based on the extensional OO type of logic.

**About Clinical Information Models:**

Abstract information models usually must account for the “what, where, who, why, when, and how” of the information they intend to represent. When SNOMED CT is used in clinical documentation and other medical information, generally it represents the “what.” As noted above, clinical information in SNOMED CT cannot be processed effectively or efficiently if it is all mixed up with other data that are based on a fundamentally different system of logic; but how can one separate the wheat from the chaff? One way to maintain this separation would be to create a clinical statement that would represent the “where, who, why, and when” in an OO extensional model component, and strictly prohibit the “what” from being represented in this OO part whenever SNOMED CT is used. The Observation class of a clinical statement has an attribute named Code. If it is agreed to allow only SNOMED CT or a SNOMED CT extension in this “Code” attribute, and further specify that it must come from only certain hierarchies within SNOMED CT (for example Observables and Clinical Findings), then one could safely use a reasoner to perform analyses using powerful subsumption queries on the model. For example one easily could find all patients that have an *“autoimmune disease with finding site lung and morphology fibrosis”*. A model designed this way is “correct” and one could safely obtain the benefits of using reasoners in clinical work.

Some models incorrectly and inappropriately mix up the two different kinds of logic, which presents problems that may have multiple solutions. The main problem can be most easily recognized by the use of “what” words like “Systolic Blood Pressure” in the extensional OO part of an information model that also may use SNOMED CT. This is an incorrect mixing of logic types because, in this case, Systolic Blood Pressure is a term defined in SNOMED CT that also is defined anew in an extensional model. Although it may be an unintentional result from the perspective of the model author, this model is reinventing SNOMED CT while simultaneously creating confusion and seriously complicating its use. Instead, the official definition of the “what” of the model should be taken from a standard vocabulary code bound to the object, e.g. SNOMED CT.

One possible alternative solution might be to use entirely intensional logic in all parts of a model; but this does not work primarily because available ontologies are not intended to be used for e.g. the “who” and “when” parts of models, and because with current technology they do not perform well for millions of patient records. On the other hand, models in which the use of SNOMED CT is prohibited do not have this problem. Generally, it could pose real dangers to patients and clinicians if the SNOMED CT term and the newly defined term were confused. It would be best never to use such “what” words in a clinical model that also may use SNOMED CT, but when they remain, then we propose they always should be an interface term or human readable label for the concept. Much of the model may be completed in the extensional OO style, but the “what” part usually should use an ontology like SNOMED CT and this must be known explicitly.

**A new proposed solution, Semantic Node Labelling (SNL)**

The case above describes an example of the explicit separation of logic types in a model by an informal rule. The extensional logic – the OO part – is the base of the model, and the intensional logic would be limited to only one or more nodes in the model, in this case the “code” of the Observation class. If it were also known to the modelers and the users of this model that only SNOMED CT would be allowed in this part of the clinical information model, and only specific SNOMED CT hierarchies, then the users of the model would know that reasoners and powerful subsumption searches may be used, but that they must be limited to SNOMED CT-compatible logic and only in that one designated node of the model.

We can generalize this principle by explicit labeling of the model components, or Semantic Node Labeling (SNL). Explicit labeling would greatly simplify the analysis of models and enable machine interoperability of the SNOMED CT information without special knowledge of the intent of the author of the model. Users of SNOMED CT and others who would like to take advantage of the power of semantic web technologies like SPARQL and OWL, or the EL+ intensional logic of SNOMED CT, could do so without danger.

In SNL we propose that clinical information models incorporate conditional and optional metadata tags that may be bound to any node in any clinical model:

* **Vocabulary or Coding System**   
  This existing **conditional** element usually is **required** in the condition that a standard vocabulary or known coding system is used. This element is already found in most models either at the node, statement, cluster, entry, or element levels. These data may be formatted differently in various families of standards, yet the identification of standard vocabularies and coding systems is widely understood. When an ontology that uses intensional logic is indicated as the sole vocabulary then this vocabulary element draws a bright line around a model component where the open world assumption exists. More detailed information about the ontology frequently is useful and is detailed below.
* **Intensional Logic**   
  This proposed **optional** element specifies intensional logic that is present by indicating intensional logic tools that are allowed for the model component. The range of values can include terms for logic or tools such as SPARQL, OWL, EL, EL+, and SNOMED CT. If the **intensional logic** tag is present then it is known which reasoners can safely be used to classify this particular node. Clinical models separately may identify SNOMED CT as the coding system at an elemental level but this logic tag at the component, cluster, entry, or node level may include one or more data elements.
* **Ontology**   
  This proposed **optional** tag will specify that only this specific ontology is used. In many cases we would expect it to be SNOMED CT. This would also allow for any extension terms as long as they follow the same rules and have the same roles as the parent ontology.
* **Hierarchies**   
  This proposed **optional** tag would indicate which hierarchies are allowed at this particular node.
* **Post-coordination-allowed (Boolean)**This proposed **optional** tag if “false” indicates that values must be single codes, or if “true” then post coordinated expressions are allowed.

**Conclusion:**

Users of clinical information models who take advantage of intensional logic with tableaux reasoners and other ontology classifiers have a recognized problem: there sometimes is no way to know when, if, or how such logic can be safely used except by having special knowledge of the intent of the author of a model. This is because there is no mutually agreed way to separate the two kinds of logic that occur in clinical models, and any mutually agreed way to do this in one particular model would not carry over to other models.

A small set of metadata tags can conditionally and optionally be attached to any model component or node in any extensional OO based clinical information model to designate the node as capable of being reasoned over with intensional logic tools (such as SNOMED CT subsumption for example). These metadata tags could be attached to any node in a model, but we anticipate these labels will mainly, at least in the beginning, be attached to the “what” nodes in clinical statement models. We call these tags Semantic Node Labeling or SNL.

The proposed tags for Semantic Node Labelling are:

* vocabulary (any standard vocabulary)
* intensional logic (RDF, OWL-DL, EL, EL+, SNOMED CT, etc.)
* ontology (SNOMED CT, etc.)
* hierarchies (Clinical Findings, Observables etc.)
* post-coordination-allowed (TRUE / FALSE)

Any model component or node without any SNL tags, or with an extensional vocabulary tag, is understood to be incapable of intensional reasoning. Any model that has a node or nodes tagged to indicate an appropriate ontology (for example the code attribute of an Observation class in a clinical statement) could be evaluated with the more powerful semantic web intensional logic such as subsumption queries. For many models this will mean one can use SNOMED CT to its full extent and advantage, but it allows for other ontologies to be used, as well as leaves the door open to allow for intensional logic to be used in the future in other parts of the models. By adding these tags a user of clinical information models would be able to take advantage of the intensional logic features contained in any model.

1. <http://www.ringholm.com/docs/05000_Clinical_Models_and_SNOMED.htm> [↑](#footnote-ref-1)
2. <http://en.wikipedia.org/wiki/Closed_world_assumption> [↑](#footnote-ref-2)
3. <http://en.wikipedia.org/wiki/Open_world_assumption> [↑](#footnote-ref-3)
4. <http://www.cs.ox.ac.uk/ian.horrocks/Publications/download/2007/BaHS07a.pdf> [↑](#footnote-ref-4)
5. <http://www.w3.org/2001/sw/wiki/Main_Page> [↑](#footnote-ref-5)