Quality Assurance of the Content of a Large DL-based Terminology using Mixed Lexical and Semantic Criteria: Experience with SNOMED CT

Alan Rector **Luigi Iannone** **Robert Stevens**
School of Computer Science, University of Manchester, Manchester UK
rector@cs.manchester.ac.uk

ABSTRACT

SNOMED-CT is a large medical terminology based on description logic and mandated for use in the US, UK and several other countries. The hierarchies are known to contain many errors, but have so far proved difficult to analyse or quality assure. We present a series of methods and lessons learnt from experience in quality assuring a "module" of SNOMED for specific applications that we expect to generalize both to SNOMED as a whole and to other large ontologies. They feature a) dependence on domain expertise b) starting from classes selected for relevance to specific applications, c) tracing all errors to their root and verifying repairs by reclassification d) extraction of manageable-sized "modules"; e) mixed semantic and lexical criteria, and f) extensive use of scripting. They aim to reduce the cognitive load on experts by a) looking initially upwards rather than downwards in the hierarchies, b) breaking up long lists of direct subclasses by introducing definitions for meaningful subcategories. Errors found range from simple mistakes to systematic errors in schemas.

Categories and Subject Descriptors

I.2.4 Knowledge Representation Formalisms and Methods – *representation languages*
D2.5 Testing and debugging – *debugging aids*
H.1.2 User/Machine Systems – *Human factors*

General Terms

Design, Human Factors, Standardization, Languages

Keywords

OWL, Quality Assurance, Ontologies, Description Logics, Modularization

INTRODUCTION

Quality assurance of large terminologies and ontologies built using description logics is a black art at best and claimed to be impossible at worst. However, SNOMED-CT [26] is now mandated as a terminology electronic health records in numerous countries including the USA, UK, Canada, Australia, and several countries in continental Europe. Quality assurance of SNOMED is therefore of great practical importance.

SNOMED[1] contains over 400,000 distinct concepts and more than a million lexical terms and synonyms. The core implementation is in a description logic equivalent to the OWL-EL profile[2] [2] without disjoint axioms.

For users and applications, the primary issue for quality in SNOMED is the inferred hierarchy. It is the inferred hierarchy that is published, affects the retrieval of information, and determines the scope of abstractions used in decision support. From most users' point of view, it is the inferred hierarchy that is the "meaning" of each concept. Other aspects of the representation are manifest to applications and users primarily via their effects on the inferred hierarchy.

Although widely known to contain many errors, [6] SNOMED's inferred hierarchies have been difficult to study because until recently: a) the source representation (known as the "stated form") was not released and the details of the classifier were proprietary, and b) most classifiers and tools could not handle terminologies of this size.

These difficulties have recently been overcome by four developments: a) publication by SNOMED of the stated form plus a Perl script to convert it to OWL; b) efficient methods for extracting "modules" for arbitrary "signatures" from ontologies which let us extract easily manageable sub-ontologies [11], c) a fast, open, robust OWL-EL classifier that integrates with Protégé (or other) environments using the OWL API[3] [16], d) the publication by the US National Library of Medicine of a "Core Problem List Subset" of the 8500 most commonly used SNOMED concepts from several prominent US hospitals[4] [9].

Using these new tools, we are seeking to use SNOMED in two practical applications:

- An industrial collaboration to develop clinical systems
- As a source for the "Foundation component" of the latest revision of the International Classification of Diseases (ICD-11).

To do so, it is necessary to quality assure the portions of SNOMED used to establish that they a) behave as expected

K-CAP'11, June 25-29, 2011, Banff, Alberta, Canada.
Copyright 2011 ACM .
Cite as: Rector, A., Iannone, L., & Stevens, R. (2011). Quality assurance of the content of a large DL-based terminology using mixed lexical and semantic criteria: Experience with SNOMED CT. Proceedings from K-CAP 2011, Banff, Canada. pp 57-64 (best paper prize)

[1] See website of governing body: http://www.ihtsdo.org

[2] http://www.w3.org/TR/owl2-profiles/#OWL_2_EL

[3] http://aehrc.com/hie/snorocket.html. NB.

[4] http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html

in applications, and b) are acceptable to the editors of the ICD.

This paper reports the methods used and the lessons learned in this process. Our overall approach has several distinguishing features, in that it:

- *Starts from what domain experts identify as key concepts.*
- *Aims to reduce the cognitive load on domain experts.*
- *Breaks the ontology into modules that classify quickly*
- *Uses the classifier interactively.*
- *Combines semantic and lexical methods.*
- *Makes extensive use of scripting*

By contrast with our focus on content and hierarchies, previous studies of SNOMED have focused on issues of its "ontological" structure, *e.g.* [5, 21, 22, 24], have used pattern based techniques to try to detect regularities and departures from them, *e.g.* [27], have tested inter-rater reliability, *e.g.* [1, 7, 25], or compared coverage, but not precision, in recall against various controls, *e.g.* [8], but have not focused on the clinical content or inferred hierarchies themselves.

Using our approach, we identified a number of error types with regard to content and inferred hierarchies:

- *Simple mistakes* – which may lead to widespread incorrect inferences.
- *Misunderstandings of the semantics of concepts and attributes*[5] *as implemented in the description logic* – leading to unintended inferences.
- *Fundamental errors in the modelling schemas*.
- *Over-literal definitions* – leading to over-generalised concepts that do not correspond to common usage.
- *Incomplete modelling* – leading to missing inferences.
- *Attempts to fix erroneous inferences without tracing them to their roots* – leading to "kluges" that result in "helter-skelter modelling."
- *Lack of normalisation of complex segments* – leading to tangled and inconsistent hierarchies.

MATERIALS

The study was conducted using the SNOMED IHTSDO[6] release of 31 Jan 2010, and later rechecked against the release of 31 July 2010. Transformation to OWL used SNOMED's supplied PERL script. As a signature, we used the UMLS Core Problem Subset as of July 2010.

For visualisation and editing, we used Protégé 4.0[7] with the SNOROCKET[8] EL^{++} classifier[16]. All results were checked using at least one of the classifiers packaged with Protégé – Pellet and/or FaCT++.

For modularization, we used the tools built into the OWL API 3[9] via a publicly available standalone tool.[10] Diff and patch tools used OWLPatch[11]

For mixed lexical and semantic searches and scripting, we used OPPL-2 in [14, 15] in Protégé 4.1. OPPL is a scripting language to query and process sets of axioms in OWL ontologies. Details of the of OPPL and its syntax as used in this paper are available from its website.[12]

In order to check that the errors found were not the result of any of the transformations or use of OWL, all errors were verified to exist in the full SNOMED distribution using at least one of two browsers: SNOB[13] and the *de facto* standard CliniClue Explorer[14].

All changes made are to experimental internal versions, but will be submitted to the SNOMED organisation. All have been made within the limits of SNOMED's formalism and the SNOROCKET classifier used by SNOMED. In our suggested repairs, SNOMED's naming conventions have been adhered to where possible. Supplementary material is available http://www.owl.cs.manchester.ac.uk/snomed.

METHODS AND EXAMPLES

Extracting modules

The Core Problem List Subset, as published, is a simple list of roughly 8500 SNOMED classes without hierarchies or description logic structure. This set of classes was, therefore, used as a "signature," to extract a "module" from the full SNOMED stated form.

A "signature" is a set of entities. A "module" is a subset of the entities and axioms in an ontology sufficient to make all inferences about the entities in the "signature" as would have been made in the full ontology [11] – *i.e.* in our case, the same inferred hierarchies as in the full SNOMED.

The resulting module consisted of roughly 35,000 classes – less than ten per cent of the SNOMED total. The module classifies in under 30 seconds using SNOROCKET and under four minutes using Pellet or FaCT++.

Selecting starting points

Even 8500 main classes in a total OWL model of 35000 are too many to understand easily. The classified hierarchy presents a tangled forest. Where to start?

Rather than browsing the terminology as a whole, our experts found it reduced their cognitive load to look at specific key concepts relevant to their application. Accordingly, a set of initial starting points was selected from the two projects. All were of major clinical importance – hypertension (high blood pressure), diabetes, pneumonia, myocardial infarction (heart attack), head injury, etc. Is-

[5] SNOMED's term for what OWL calls "properties"

[6] Licensing for SNOMED varies by country, but it is available for research and academic use world-wide. For access, contact national authority or the SNOMED managing body, the International Health Terminology Standards Organisation (IHTSDO), http://www.ihtsdo.org.

[7] http://protégé.stanford.edu

[8] http://aehrc.com/hie/snorocket.html

[9] http://sourceforge.net/projects/owlapi/

[10] http://owl.cs.manchester.ac.uk/snomed

[11] http://owl.cs.manchester.ac.uk/patch/)

[12] http://oppl2.sourceforge.net/

[13] These methods guarantee that the classification will be the same in the module as in the complete SNOMED

[14] http://www.cliniclue.com

sues were identified by the clinical experts from their effects in the applications or by comparison with external sources; alternative solutions suggested by the OWL experts, some of whom had some clinical knowledge; and the results vetted by the clinical experts.

Looking up the hierarchy

For each class, we started by looking up the hierarchy rather than down. Hierarchies usually fan in going up and out going down. The maximum number of direct superclasses of our key starting point classes was four, whereas the number of direct subclasses was often over ten and sometimes up to twenty. The maximum number of significant ancestors of our key starting points was twelve; the maximum number of descendents, several hundred. Consequently, the entire upward hierarchy can generally be viewed graphically as shown in Figure 1, whereas the downwards hierarchy rarely can be.

To the experts, oddities in ancestors tend to jump out. For example, hypertension is not considered to be a disorder of soft tissues, nor is it found under this heading in any standard reference source. Similarly, diabetes is not considered a disease of the abdomen, although it is found in the upward hierarchy for diabetes.

Omissions are more difficult to spot, but experts have well-established expectations. That myocardial infarction was not classified as a form of ischemic heart disease – *i.e.* disease of the blood supply to the heart – was spotted independently by several experts and had major effects on applications.

Tracing anomalies to their root: Analysis by repair

Having found an error, the question is how to repair it. Since SNOMED is formulated in a simple description logic – OWL-EL less disjointness – identifying the reason for the inference is rarely difficult. If it is, the Protégé explanation facilities [13] usually find the problem quickly.

Although in some cases which axiom to change is obvious, in many cases a choice must be made. The error found can be at the end of one or more paths up the hierarchy. For example, as Figure 1 shows, "*Hypertensive disorder, systemic arterial (disorder*)"[15] is inferred to be under both *Finding* and *Disorder of Soft Tissue*, following two different paths, each step of which seems superficially plausible. However, none of our experts agreed that *Hypertension* should be under anything related to *Soft tissue* – either finding or disorder.

Tracing the cause of both inferences to their root, it was found that they both followed from the axiom that the *site*[16] of *Hypertensive disorder* was *some Artery*[17] and that *Arteries* were classed as *Soft tissues*. Together, these axioms led to the unwanted inferences.

Such cases require discussion with the experts. In this case, they decided, after looking at the diseases classified under hypertension, that *Hypertension* was a systemic disorder that should be *sited* in the cardiovascular system as a whole rather than individual arteries. Therefore, the axiom was changed so that *Hypertension* was *sited* simply in the *Cardiovascular system.* This gave rise only to the inference that *Hypertension* is a *Disorder of the cardiovascular system,* and eliminated both paths related to *Soft tissue*.

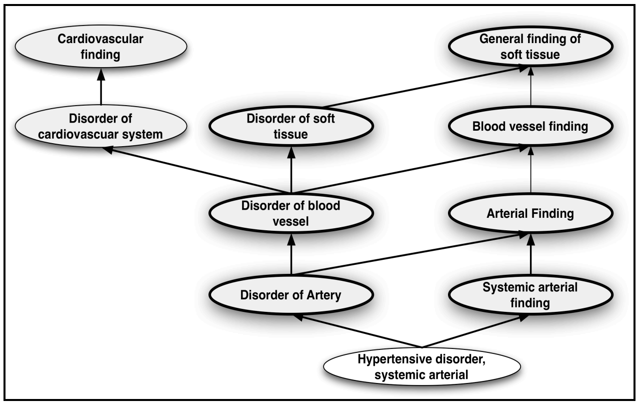

Cardiovascular finding
General finding of soft tissue
Disorder of cardiovascuar system
Disorder of soft tissue
Blood vessel finding
Disorder of blood vessel
Arterial Finding
Disorder of Artery
Systemic arterial finding
Hypertensive disorder, systemic arterial

Figure 1: Upwards hierarchy for Hypertensive disorder before repair (from OWLViz view in Protégé[18])

Looking down the hierarchy

Although looking up the hierarchy is cognitively efficient, it is not always sufficient. Even after the repairs above, there are thirteen direct subclasses of *Hypertensive disorder*. In most browsers, these are merely ordered alphabetically. Are they all correct? Are any missing? Where to begin? We used three methods.

Combined lexical and semantic search

If omissions are sufficiently glaring, experts may spot them despite the presentation. Following up such intuitions may then lead to significant repairs.

For example, looking down the list of hypertensive diseases, experts immediately noticed that hypertensive renal disease was present but, contrary to their expectations and requirements for the application, neither hypertensive heart nor eye disease ("hypertensive retinopathy") were present. This raised the question of how many other hypertensive disorders might be missing.

To find out, we used a method related to that proposed by Campbell [4]. We first used a simple "Find" to determine that numerous candidates containing the string "hypertensive" existed. We then did a complete search using an OPPL script. The script in Figure 2 selects all the terms containing "hypertensive" but not classified under "Hypertensive disorder, systemic arterial" and places them under an arbitrary class "Candidate". When reclassified the top candidate classes are as shown in Figure 3.

[15] 38341003|Hypertensive disorder, systemic arterial (disorder)|

[16] 363698007 | Finding site (attribute)|

[17] 281159003 | Systemic arterial structure (body structure)|

[18] Larger image at: http://www.cs.man.ac.uk/~rector/papers/KCAP-2011-QA/Fig-1-Enlarged.pdf

Out of the list of "Candidates" shown in figure 3, the four highlighted classes were identified by experts as kinds of systemic arterial hypertensive disorder; the remainder were a mixture of other conditions whose names happen to contain the word "hypertensive".

```
?C:CLASS=MATCH(".*[Hh]ypertensive.*")  //Find all classes with names
SELECT ?C SubClassOf Thing               // of form …hypertensive…
 // check if they are not already classified under hypertensive disorder-
WHERE FAIL ?C SubClassOf 'Hypertensive disorder, systemic arterial (dis-
order)'
BEGIN ADD ?C SubClassOf Candidate END;  // if no, add to list of
                                         //candidates
```

Figure 2: OPPL script to identify suspect classes on mixed lexical and semantic criteria[19]

Viewing the classified hierarchy rather than the flat list is important. For example, the single class "Hypertensive heart disease" in Figure 3 has twenty-five descendants. Showing the classified view reduces the cognitive load on experts. Furthermore, only the top classes need to be edited, the remainder will "inherit" the correction.

```
▿ Candidate
      ➢   'Antihypertensive adverse reaction (disorder)'
      ➢   'Antihypertensive allergy (disorder)'
      ➢   'Antihypertensive overdose (disorder)'
          'Blind hypertensive eye disease (disorder)'
          'Hypertensive heart disease (disorder)'
          'Hypertensive retinopathy (disorder)'
      ➢   'Poisoning by antihypertensive agent (disorder)'
          'Portal hypertensive gastropathy (disorder)'
          'Pulmonary hypertensive arterial disease (disorder)'
          'Pulmonary hypertensive venous disease (disorder)'
          Ulcer of skin caused by ischemia due to hypertensive dis-
          ease (disorder)'
```

Figure 3: Results of lexical search in Figure 2 after reclassification. Highlighted classes are semantically kinds of hypertensive disorder.

Repairing and Regularizing schemas

The experts noted that all of the highlighted classes in Figure 3 were complications rather than kinds of hypertension and, furthermore, that the original list of subclasses contained two better treated as complications than kinds.

Of the six complications of hypertension identified – four lexically and two semantically in the original subclass hierarchy – there were four patterns. One was linked to hypertension by the property *due to*[20], another by *associated with*[21], a third in no way at all, and the others as subclasses.

Looking further revealed that, although there was no class for *Hypertensive complications* in SNOMED, there was a class for *Diabetic complications,* which was defined as those diseases that were *associated with some Diabetes*. For uniformity, it was therefore decided to use the property *associated with* rather than *due to*.

To complete the repair, therefore, a new class for *Hypertensive complication* was defined as *Disorder associated with Hypertensive disorder* and the descriptions of all classes identified as hypertensive complications edited to fit this pattern.

However, there was a further problem. These actions changed meaning of the class *Hypertensive disorder* so that it now corresponded to hypertension *per se.* There were two options: i) to create a new class *Hypertension* as a subclass of the existing class, or ii) to rename the existing class. (Renaming in SNOMED is trivial because all names are label annotations. The primary identifiers are "nonsemantic" numeric identifiers.)

To choose, we examined how the original class *Hypertensive disorder* was used elsewhere using the Protégé "usage view" as shown in figure 4. Clearly, the meaning in all cases is of the disorder "hypertension" rather than "complication of hypertension". Furthermore, the core of the existing definition of the class can be paraphrased as "disorder of cardiovascular system characterized by increased blood pressure", which likewise fits "hypertension". In addition, "Hypertension" was a synonym for the original class.

```
Usage: 'Hypertensive disorder, systemic arterial (disorder)'
   Found 9 uses of 'Hypertensive disorder, systemic arterial (disorder)'
➢   'Family history: Hypertension (situation)'
➢   'History of – hypertension (situation)'
➢   'Hypertension screening (procedure')
➢   'Hypertensive encephalopathy (disorder)'
➢   'Hypertensive heart disease (disorder)'
➢   'Hypertensive renal disease (disorder)'
➢   'Neonatal hypertension (disorder)'
➢   …
```

Figure 4: Usages of Hypertensive disorder

It was therefore decided to:

- Rename the existing class to *Hypertension (disorder)*
- Create a new class *Hypertensive complication* as described above.
- Create a new class *Hypertension AND/OR Hypertensive Complication* as a common parent of the two classes to maintain backwards compatibility with the overall shape of the hierarchy. (NB SNOMED's formalism does not include disjunction.) OPPL scripts were created for the above and the resulting hierarchy is as outlined in Figure 5.

```
▿  'Disorder of cardiovascular system (disorder)'
      ▿   Hypertension AND/OR Hypertensive complication(disorder)'
             ▿   'Hypertension (disorder)'
                        …kinds of hypertension…
             ▿   'Hypertensive complication (disorder)'
                        …kinds of hypertension complication...
```

Figure 5: Outline of revised structure for hypertension and complications.

Analysis by categorization

Even after removing the complications of hypertension, the list of direct subclasses was still too long for experts to analyse easily. To ease the load on the experts and to help them decide if the lists were correct and complete, we added additional defined classes and restrictions to reflect the apparent meaning of the names. The result is a better organized list as shown in Fig 7.

For example, systolic and diastolic hypertension[22] were previously near the beginning and end of the list, respec-

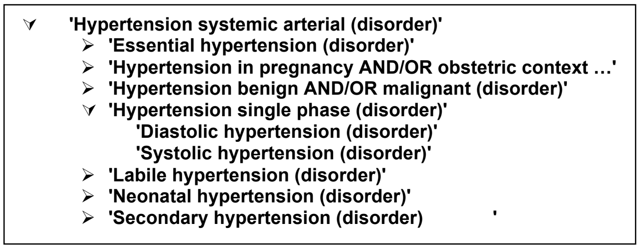
[19] NB: Comments are not yet implemented in the current release of OPPL

[20] 42752001 | Due to (attribute)

[21] 47429007 | Associated with (attribute)

[22] The first and second numbers in hypertension readings such as 140/90

tively, but are now brought together. This case is trivial, but *Hypertension in pregnancy AND/OR Obstetric context* subsumes nearly a dozen classes that were previously scattered under three headings, two with apparently identical definitions. Without bringing them together, as shown in Figure 6, it is nearly impossible to check if the lists are correct and complete or whether their logical definitions super-classes and subclasses match their intended meaning.



- 'Hypertension systemic arterial (disorder)'
  - ➢ 'Essential hypertension (disorder)'
  - ➢ 'Hypertension in pregnancy AND/OR obstetric context ...'
  - ➢ 'Hypertension benign AND/OR malignant (disorder)'
  - 'Hypertension single phase (disorder)'
    - 'Diastolic hypertension (disorder)'
    - 'Systolic hypertension (disorder)'
  - ➢ 'Labile hypertension (disorder)'
  - ➢ 'Neonatal hypertension (disorder)'
  - ➢ 'Secondary hypertension (disorder) '

Figure 6: Hypertension after categorization

(Ontologists may object to classes such as 'Hypertension benign AND/OR malignant', and these were implemented so as to be removed easily if desired. However, their utility in quality assurance is difficult to dispute.)

Other issues

Systematic repair of erroneous schemas

Looking up the hierarchy from several conditions of the arteries and nerves of leg and foot, we found them to be classed not only as disorders of the lower extremity, but also of the pelvis and trunk or even abdomen. The issue was traced back to subtle conflation in the schema of branches and parts.

In SNOMED, a disorder of the part is a disorder of the whole. To achieve this, SNOMED currently uses the SEP triple schema [23] for anatomy. For each anatomical entity, there is an *S* class for the *Structure or its parts* with two children, a *P* class for its *parts* and an *E* class for its *entirety* as shown in Figure 7. Each *structure* class for a part is a subclass of the *part* class for its whole. For example, in Figure 7 the *Cusp of the aortic valve* is a descendant of *Heart Structure*, via *Heart part*, etc. so that a *Disorder of the cusp of the aortic valve* is classified as a *Disorder of a heart structure – i.e.* a *Heart disease*. This trick simulates transitive relations for classifiers that do not support them.

However, while disorders of the part are disorders of the whole, disorders of a branch are not disorders of its root – otherwise all disorders of arteries would be disorders of the root artery, the *Aorta*, etc. However, SNOMED makes branches subclasses of the *Structure* node – making them behave as parts and entailing many unintended inferences. In fact, queries with OPPL established that the *Entire* node did not even exist for any *Structure* for which there was a *Branch.*

Where naming conventions are consistent, such problems can be dealt with by an OPPL script as in Figure 8.

Unfortunately, there are other cases involving misuse of SEP triples where the naming is not explicit Indeed, the entire SEP triple schema is now redundant and requires revision.

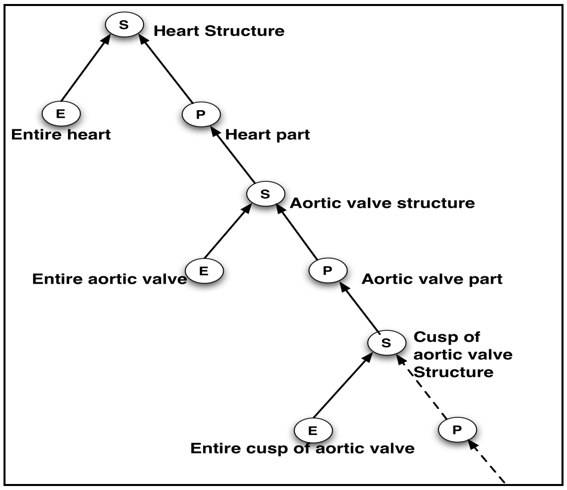


Figure 7: Correct use of SEP triples to represent structure, entire, and parts

```
?B:CLASS=MATCH((".*)[Bb]ranch(.*)"),//Select class names “…branch…”
?S:CLASS=MATCH(".*[Ss]tructure.*")  //Select class names “…structure…”
?entire:CLASS = create( B.GROUP(1)+"_Entire_"+B.GROUP(2))
                                        //Create a new Entire node from name
SELECT ASSERTED ?B SubClassOf ?S // Find all branch-structure pairs
BEGIN
   REMOVE ?B SubClassOf ?S,        //remove branch subclass structure
   ADD ?entire SubClassOf ?S        //make entire subclass of structure
   ADD  ?B SubClassOf (is_branch_of some ?entire )
                                       //assert that branch is branch of entire
END;
```

Figure 8: Simplified OPPL script to move branches to newly created “Entire” node a SEP triple[23]

Misunderstanding of semantics of attributes

In cases where the problem is a direct subclass axiom between named classes, repair is simple. If the offending link is inferred, the inference must be traced to its root. The example of the *site* of *Hypertension* has already been discussed above. In another example, when looking up the hierarchy the experts noticed that *Diabetes* was classified as a *Disease of the abdomen*, which is clearly wrong. Investigation revealed that this resulted from the axiom that the *site* of *Diabetes* was the *Endocrine pancreas*, which experts also considered wrong[24]. Since the Endocrine pancreas is part of the Abdomen,[25] the error followed.

Both the above errors involved problems with the use of the *site* attribute[26] for a systemic or endocrine disease. This suggested that the use and semantics of the attribute *site* with systemic and endocrine diseases should be examined systematically. This revealed numerous potential errors and controversial classifications.

[23] NB: Comments are not yet implemented in the current version of OPPL

[24] Diabetes is a systemic or endocrine disorder. Some, but not all, is caused by, but not a kind of, disorder of the endocrine pancreas.

[25] SNOMED makes no distinction between parthood and containment.

[26] 363698007 | Finding site (attribute)

Incomplete modelling

Tracing errors to their root often uncover cases in which the SNOMED logical model is incomplete, either because there is no complete definition – *i.e.* equivalence class axiom – or because necessary conditions – *i.e.* restrictions – are missing. In SNOMED's formalism, without a complete, necessary and sufficient definition, subclasses cannot be inferred. For example, if *Heart disease* is fully defined by necessary and sufficient conditions, *i.e.* as *Any disease with site heart*[27], then all diseases with *site heart* will be inferred to be subclasses. If *Heart disease* is only partially defined by necessary conditions, *i.e.* as *Some disease with site heart*[28], then those inferences will not be made.

As a specific example, one of the first errors to come to light was that *Myocardial infarction* (heart attack) was not classified as a form of *Ischemic heart disease* (the common form of heart disease caused by poor circulation in the coronary arteries that supply the heart muscle). On investigation, it was found that none of classes *Myocardial infarction, Infarction, Ischemic heart disease*, or *Ischemia* itself were fully defined by equivalence classes. Hence, the axioms were insufficient to support the expected inference. Repair required providing full definitions via equivalence class axioms for all four classes, plus the axiom that *Infarction* was always due to *Ischemia* via a property path axiom. (Property paths are included in EL^{++}). Reclassification showed that these changes corrected not only the missed classification for *Myocardial infarction* but numerous other errors in the inferred hierarchies as well. (see [18]).

Incomplete modelling is common throughout SNOMED. A complete inspection for all cases was beyond our resources. However, cases were discovered frequently as the root cause of errors. Since cases tend to appear together, once one is found, others can often be repaired systematically.

Over-literal definitions leading to over-generalised concepts

Many medical terms have a more specialised meaning than the literal interpretation of their names. For example, "Neuropathy" is derived literally from "Disorder of Nerve". SNOMED's definition is logically equivalent. However, "neuropathy" has come to mean something closer to "dysfunction" of nerves – normally excluding tumours and injuries and, for example, swelling of the optic nerve in the eye – an important sign of increased pressure on the brain.

A clinically more serious case arises when the most common and serious form of a disorder is more specialised than the literal meaning. For example, "Subdural hemorrhage" literally means bleeding under the "dura", the covering of the brain and spinal cord. Although bleeding under the dura of the spine does occur, the vast majority of subdural hemorrhages inside the cranium where they are potentially fatal. Therefore, in normal medical usage, unless qualified by "spinal", a "subdural hemorrhage" is assumed to be "intracranial". Not to do so is a potentially life-threatening error. (This is what linguists would term an "implicature" [12].)

That the meaning, and therefore the consequences for clinical decision support systems, of "Subdural hematoma" did not imply intracranial was only discovered by carefully looking up the hierarchy and came as a shock to our experts. In fact, it was discovered that there was no class for "Intracranial subdural hematoma" *per se*. (see [18].)

A minimal correction proved possible by analogy to that for *Hypertensive disorder*. Examination of the usages elsewhere within SNOMED confirmed that the meaning "intracranial" was implicit in all but a few cases. Therefore, a new, more general, class was created for *Subdural Hemorrhages, Intracranial AND/OR Spinal*; the original class label changed to *Subdural hematoma, Intracranial*, and the remainder of the hierarchy adjusted to take care of the few exceptions that had been uncovered. The result is a structure analogous to that in Figure 5.

Consequences of not tracing errors to their roots

Early on, it was noticed that some disorders and injuries of skin were classified as *Disorder of soft tissue*[29] while others were not. It was established that the subclass axiom had been omitted between *Skin and subcutaneous tissue*[30] and *Soft tissues*[31]**.** This affected the classification of every disorder of skin or subcutaneous tissue of which there are thousands in the full SNOMED release. For some of these, there were asserted subclass axioms between the disorder and *Disorder of soft tissues*. We assume these had been inserted manually in response to errors noticed by authors, without tracing those errors to their root – a practice one of our experts dubbed "helter-skelter modelling" and that is akin to "kluges" in software.

Once the root error was repaired, all of these axioms became redundant and could be removed using a script. This is important, since the anatomy model may change in future. If all classification is inferred from the anatomy model, then disorders will be reclassified correspondingly.

Lack of normalisation

The hierarchies for diseases related to head injury, skull fractures, and intracranial bleeding are too large to describe in full in a brief paper. A complete solution requires more radical reanalysis. The root of the problem is that the different axes – whether there is a hemorrhage and its type, whether there is a skull fracture and its type, whether there has been loss of consciousness or not – need to be separated and recombined with suitable definitions, *i.e.* they require normalisation [19]. One solution is suggested in [17] using a formalism including negation. However, ex-

[27] "Heart disease equivalentClass (Disease & site some Heart)"

[28] "Heart disease SubClassOf Disease Heart disease SubClassOf site some Heart".

[29] 19660004 | Disorder of soft tissue (disorder)

[30] 27856007 | Skin AND subcutaneous tissue structure (body structure)

[31] 87784001 | Soft tissues (body structure)

periments indicate that approximations can be achieved using EL^{++} in conjunction with some pre-filtering.

DISCUSSION

What is an error? A satisfactory repair?

Any method of quality assurance ultimately depends on the definition of an error. Ultimately, what constitutes an error must be decided by domain experts. It is tempting to focus on the correctness of individual definitions and axioms. However, in a description logic based system using inference, this is inadequate. *A set of definitions and axioms, however individually plausible, contains an error if it leads to an erroneous inference, as judged by domain experts.*

Some errors are clear-cut. Diabetes is not a disease of the abdomen; arteries of the ankle are not in the pelvis. Others are more controversial, *e.g.* the usage of "soft tissues varies between specialties, communities, and over time.

In some cases the semantics of the "attributes" (properties/relations) have to be treated carefully to get the desired inferences – *e.g. site* and *part* do not correspond to generic notions of location and parthood. They are effectively defined by the inferences that follow from the rule that "disorders of the part are disorders of the whole".

This was a study of opportunity and relied on the judgment of a limited number of domain experts. We have conducted a pilot evaluation of inter-rater reliability using participants in the ICD-11 revision process. It suggested few major disagreements, but a wider and more rigorous study is required. The study was deliberately conducted independently of the SNOMED organisation. However, the repairs suggested here require discussion with them and will be submitted via their usual mechanisms.

Technical issues

Importance of interactive classification

Using description logics brings major benefits. It is difficult to see how very large systems such as SNOMED could be managed otherwise. However, the real meaning of the ontology is in the classified form. If authors cannot see the effects of their changes quickly, then effective working is virtually impossible. If classification requires much more than a minute, then the methods described here become increasingly time consuming and tedious.

Use and limits scripting

Insofar as possible, we would like to make changes by scripts so that can easily be applied to new versions or alternative modules. OPPL is rapidly maturing and has become our tool of choice for manipulating OWL. Its key advantages are its ability to deal with mixed lexical and semantic criteria, both asserted and as inferred, and its independence of any specific OWL classifier. It includes the construct FAIL for negation as failure, without which scripts such as that in Figure 2 would not work.

However, the are two sorts of limitations: a) SNOMED's irregularities – neither SNOMED's naming conventions nor modelling style is sufficiently regular to allow all cases to be dealt automatically; and b) OPPL's limitations in dealing with collections and extra-logical constructs, which we hope will be resolved in later releases.

Checking for unintended effects and Unit testing

Changing the axioms to correct one inferred error always brings the possibility of introducing others. Ideally a full "diff" with the previous version of the classified OWL model would be performed. Since no adequate tools were found that scaled to size of model required, a combination of queries in OPPL and checks of the usages of the changed concepts in Protégé provided reasonable assurance that unwanted inferences had not been added nor intended inferences removed.

A major advantage of OPPL is that it facilitates implementing these checks as "unit tests". Unit testing is now standard in software engineering to ensure that errors once identified do not recur and that guidelines are adhered to. Experience in this and earlier projects [20] suggests that such testing is as important for OWL as for software.

Conclusion

This experience suggests that quality assurance of the content of even a very large description logic based terminology such as SNOMED-CT is possible and practical. Our approach starts from experts' intuitions and exploits lexical and semantic methods combined with strategies to break the problem down into manageable chunks – modularization, looking up hierarchies first rather than down, and progressively categorizing unstructured lists.

It is to be hoped that in dealing with the module derived from the Core Problem List Subset, most systematic errors will be uncovered. In general, changes made in modules will be propagated by the classifier to all of SNOMED – *e.g.* correction of the schemas for branches, corrections to the *sites* of systemic diseases, etc. On the other hand, no such approach can be guaranteed exhaustive, and these methods need to be complemented by systematic approaches such as that of Bodenreider [3]. What can be said is that, of the five major body systems investigated – cardiovascular, respiratory, endocrine, gastro-intestinal, and head-injuries – these methods enabled us to find fundamental problems that were unacceptable to collaborating experts in all but one, the respiratory system.

The next step is to test the approach on other subsets of SNOMED and then on SNOMED as a whole, starting from other key clinical concepts critical to other applications. Following that, the goal should be to address other large biomedical ontologies, *e.g.* the National Cancer Institute Thesaurus [10] and ontologies in molecular biology.

Few of the methods used here are specific to SNOMED, although it is easier to pinpoint the root cause of inferences in EL^{++} than in more expressive languages. The success in explaining and repairing many long-standing errors in SNOMED suggests that a collaboration between domain and description logic experts, mediated by staff with some knowledge of both disciplines, can rapidly improve the quality of even very large terminologies that have defied the efforts of either group alone.

ACKNOWLEDGMENTS

This work supported in part by Siemens Health Solutions, and the team working on the Manchester-Siemens collaboration is gratefully acknowledged. Particular thanks to Michael Lawley at his team at Australian E-Health Research Centre for SNOROCKET and their prompt and helpful support. The input of the WHO ICD-11 revision team is also gratefully acknowledged. We also thank IHTSDO for making available the stated form of SNOMED-CT and the Perl script to convert it to OWL.

REFERENCES

[1] Andrews, J. E., Richesson, R. L., and Krischer, J. 2007. Variation of SNOMED CT coding of clinical research concepts among coding experts. J American Medical Informatics Association. 14, 4, 497-506.

[2] Baader, F., Brandt, S., and Lutz, C. 2005. Pushing the EL Envelope. Proc IJCAI-05. 364–369.

[3] Bodenreider, O., Smith, B., Kumar, A., and Burgun, A. 2007. Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. Artificial intelligence in medicine. 39, 3, 183-195.

[4] Campbell, K. E., Tuttle, M. S., and Spackman, K. A. 1998. A "lexically-suggested logical closure" metric for medical terminology maturity. Proc AMIA 1988). 785-789.

[5] Ceusters, W., Smith, B., Kumar, A., and Dhaen, C. 2004. Mistakes in medical ontologies: Where do they come from and how can they be detected? Studies in Health Technology and Informatics. 145-164.

[6] Ceusters, W., Smith, B., Kumar, A., and Dhaen, C. 2004. Ontology-based error detection in SNOMED-CT. Proc MEDINFO. 2004, 482-6.

[7] Chiang, M. F., Hwang, J. C., Yu, A. C., Casper, D. S., and Cimino, J. J. 2006. Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers. Proc AMIA 2006. 131-135.

[8] Elkin, P. L., Ruggieri, A. P., Brown, S. H., Buntrock, J., Bauer, B. A., Wahner-Roedler, D., Litin, S. C., Beinborn, J., Bailey, K. R., and Bergstrom, L. 2001. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. Proc AMIA 2001. 159-164.

[9] Fung, K. W., McDonald, C., and Srinivasan, S. 2010. The UMLS-CORE project: a study of the problem list terminologies used in large healthcare institutions. J Am Med Inform Assoc. 17, 675-680.

[10] Goldbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., and Parsia, B. 2004. The National Cancer Institute's thesaurus and ontology. J Web Semantics. 1, 1, 32-36.

[11] Grau, B. C., Horrocks, I., Kazakov, Y., and Sattler, U. 2008. Modular reuse of ontologies: Theory and practice. J Artificial Intelligence Research. 31, 1, 273-318.

[12] Grice, H. P. 1957. Meaning. Phil Rev. 66, 377-388.

[13] Horridge, M., Parsia, B., and Sattler, U. 2010. Justification oriented proofs in OWL. International Semantic Web Conference (ISWC 2010). 354-369.

[14] Iannone, L., Aranguren, M. E., Rector, A., and Stevens, R. 2008. Augmenting the expressivity of the ontology pre-processor language. OWL Experiences and Directions (OWLEd 2008).

[15] Iannone, L., Rector, A., and R, S. 2009. Embedding knowledge patterns into OWL. European Semantic Web Conference (ESWC 2009). 218-232.

[16] Lawley, M. J. Exploiting fast classification of SNOMED CT for query and integration of health data. KR-MED 2008. 8-14.

[17] Rector, A. and Brandt, S. 2008. Why do it the hard way? The case for an expressive ontological schemas for SNOMED. J Am Med Inform Assoc. 15, 744-751.

[18] Rector, A., Brandt, S., and Schneider, T. 2011. Getting the foot out of the pelvis: Modelling problems affecting use of SNOMED-CT hierarchies in practical applications. JAMIA 18, (in press).

[19] Rector, A. 2003. Modularisation of domain ontologies Implemented in description logics and related formalisms including OWL. Proc KCAP 2003. 121-128.

[20] Rogers, J., Roberts, A., Solomon, D., van der Haring, E., Wroe, C., Zanstra, P., and Rector, A. 2001. GALEN Ten years on: Tasks and supporting tools. Proc Medinfo 2001. 256-260.

[21] Schulz, S., Suntisrivaraporn, B., and Baader, F. 2007. SNOMED CT's Problem List: Ontologists' and logicians' therapy suggestions. Proc Medinfo 2007. 802-806.

[22] Schulz, S., Suntisrivaraporn, B., Baader, F., and Boeker, M. 2009. SNOMED reaching its adolescence: Ontologists' and logicians' health check. International Journal of Medical Informatics. 78, S86-S94.

[23] Schulz, S., Hahn, U., and Romacker, M. 2000. Modeling anatomical spatial relations with description logics. AMIA Fall Symposium (AMIA-2000). 799-783.

[24] Spackman, K. A. and Reynoso, G. 2004. Examining SNOMED from the perspective of formal ontological principles: Some preliminary analysis and observations. KR-MED. 81-87.

[25] Vikström, A., Skånér, Y., Strender, L. E., and Nilsson, G. H. 2007. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. BMC Medical Informatics and Decision Making. 7, 1, 9.

[26] Wang, A. Y., Sable, J. H., and Spackman, K. A. 2002. The SNOMED clinical terms development process: refinement and analysis of content. AMIA Fall Symposium. 845.

[27] Wang, Y., Halper, M., Min, H., Perl, Y., Chen, Y., and Spackman, K. A. 2007. Structural methodologies for auditing SNOMED. J Biomed Informatics. 40, 5, 561-581.