An Ontology-Based Methodology for Error Correction in SNOMED-CT®

Werner Ceusters [1], Barry Smith [2]

[1] European Centre for Ontological Research, Saarbrücken, Germany

[2] Institute for Formal Ontology and Medical Information Science, Saarbrücken, Germany and Department of Philosophy, University at Buffalo, NY, USA

Abstract

Large biomedical terminologies are difficult to develop and maintain. ISO and CEN standards for good terminology practice do indeed exist; but they are seldom used, and bring difficulties of their own. Description logics have been proposed as a formal means of preventing the types of errors commonly found in biomedical terminologies; but it has been shown that such logics are of little help without an adequate supporting methodology. We argue that realist ontology is able to provide such a methodology, and we focus on SNOMED-CT®, one of the most popular and extensive biomedical terminologies available today, for examples of the types of errors that this methodology can detect. We show how, when combined with appropriate terminology standards and description logic classifiers, the methodology can also be used to correct such errors and to provide safeguards against the occurrence of similar kinds of errors in the future.

Keywords:

Biomedical terminologies, realist ontology, quality assurance, graph theory

1. Introduction

IFOMIS, the Institute for Formal Ontology and Medical Information Science in Saarbrücken, is developing a framework for ontology construction and alignment in the biomedical domain based on rigorous formal definitions and axioms [1]. This approach starts out from the idea that we need to understand the general structure and organization of a given domain (its ontology) before we start building corresponding software models or standardized vocabularies. This means above all that the basic categories and relations structuring the domain should be to the greatest possible extent defined in a logically rigorous way. The resultant definitions and axioms would then serve as constraints on how the ontology of the domain is represented in a computer [2, 3, 4, 5].

This approach differs from what has been standard practice hitherto, which conceives an ontology as a loosely structured representation of the 'knowledge' shared by domain experts. An ontology, on this conception, is built out of terms corresponding not to entities in reality but rather to the concepts in the minds of such experts (who are themselves not commonly qualified in the field of ontology construction). Increasingly, efforts are being made to resolve the formal inadequacies which result through the application of one or other description logic classifier [6]. Description logics, however, can do no more than guarantee consistent reasoning according to the terms and definitions provided to them. If the latter contain mistakes, or fall short of correspondence to the reality that they are designed to represent, then description logic will do very little to help resolve these problems.

The IFOMIS approach has now been applied successfully to the detection of errors in systems such as the UMLS Semantic Network [7], the HL7-RIM [8], and the Gene Ontology [9, 10]. In the latter case it has also contributed to efforts to provide an ontological overhaul in the context of the current OBOL reform efforts of the Open Biological Ontologies consortium. The method is also being used to find mistakes in proprietary systems such as LinkBase® [11], which was itself then used to detect mistakes in SNOMED-CT [12].

In this paper, we go beyond the algorithms outlined in [12], describing how the ideas underlying them can be made more generally applicable and at the same time implemented independently of any proprietary system. To this end, we first show how concept-based terminologies such as SNOMED-CT® can be converted into a graph-based representation which separates lexical, semantic and ontological information. We then show how the mistakes in the January 2003 and July 2003 versions of SNOMED-CT® described in [13] can be identified automatically via specific operations on such graph representations.

2. Graph representation of a terminology

The basic idea is that it is possible to search for errors in a terminology by comparing different ways in which information can be extracted from its terms, concepts, descriptions, and definitions. SNOMED-CT® for instance distinguishes clearly between "terms" and "concepts", in line with terminology standards such as ISO-704:2000. Thus the concept "breech extraction" with concept ID #177151002, is associated with three terms, each of which enjoys its own unique ID:

- "breech extraction (procedure)" (which SNOMED-CT® declares to be the "fully specified name")
- "breech extraction" (which is the "preferred term"), and
- "total breech extraction" (which is a "synonym").

In addition, SNOMED-CT® has a fixed set of relationships, such as "*is-a*", "*Method*", "*Procedure Site*", "*Finding Site*". Concepts stand in such relationships to other concepts, while terms as such are not interrelated. The relationships can be described as triples consisting of a source-concept (e.g. "breech extraction"), a relationship (e.g. "*Procedure Site*"), and a target-concept, e.g. "female genital tract". Some triples in SNOMED-CT® constitute *definitions* for the corresponding source concepts (which are then labeled “fully defined”); for other source concepts (labeled “primitive”) triples are assigned only as *descriptions*. If, while authoring the terminology, a concept C is described by using triples drawn from the triples which constitute the definition of some other concept C_1, then, after classification by a description logic classifier, C will subsume C_1 (at least theoretically, since whether this is the case for a concrete terminology depends on the computational power of the classifier used). Thus, for example, if “malignant lymphoma of breast” would be described *inter alia* by using the triples (“malignant lymphoma of breast”, “*is-a*”, “malignant neoplasm”) and (“malignant lymphoma of breast”, “*is-a*”, “tumor of breast”), and the latter two triples constitute a definition for “malignant tumor of breast”, then, after classification, “malignant tumor of breast” will subsume “malignant lymphoma of breast”.

We can now exploit for error checking and prevention purposes the information pertaining to *concepts* and *terms*, by representing the terminology as a graph with different types of edges, including edges representing the relations between the terminology on the one hand and the corresponding classs in reality on the other. (A concrete example is shown in Fig.1, which corresponds to the toy terminology of Table 1.)

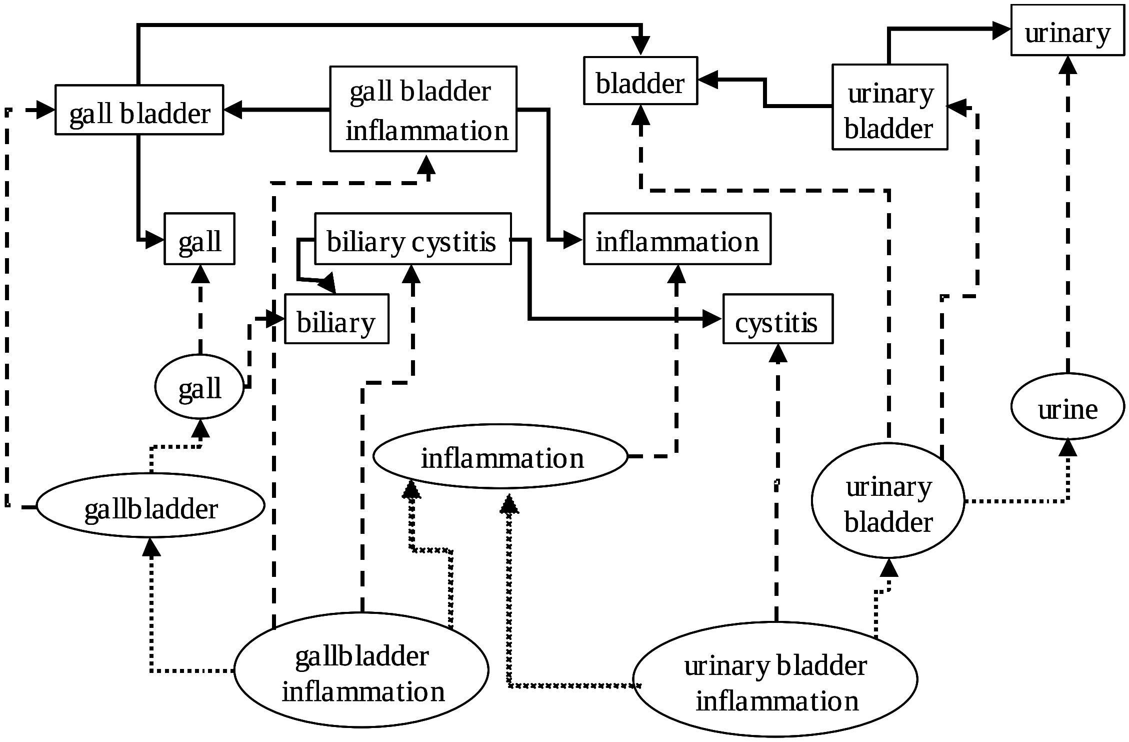

Figure 1: Three information sources for an error detection and prevention algorithm for terminologies: *lexical* (rectangles and solid arrows), *semantic* (dashed arrows) and *ontological* (ovals and light and heavy dotted arrows).

Class	**Class Description**	**Term**
GALLBLADDER INFLAMMATION	***IS-A*** INFLAMMATION **ASSOC** GALLBLADDER	gallbladder inflammation biliary cystitis
URINARY BLADDER INFLAMMATION	***IS-A*** INFLAMMATION **ASSOC** URINARY BLADDER	cystitis
GALLBLADDER	**ASSOC** GALL	gall bladder
GALL		gall biliary
INFLAMMATION		inflammation
URINARY BLADDER	**ASSOC** URINE	urinary bladder bladder
URINE		urinary

Table 1: Toy terminology corresponding to the graph representation of Fig.1

2.1. Minimal transformation algorithm

To create such a graph, which we build on the (defeasible) assumption that the concepts in the terminology do indeed correspond to classes in reality, we define three types of information:

- *Ontological information*, which concerns exclusively those classes in reality to which the concepts correspond, together with the associated relations (hereafter called class-relation-class triples);
- *Lexical information*, which concerns exclusively terms conceived as strings, together their decompositions into grammatically complete substrings;
- *Semantic information*, which concerns exclusively the relationships between the terms and the classes which are their denotata.

Graph-formation then consists of three main steps: class decomposition (C), semantic association (S), and lexical decomposition (L):

> C1: For each source concept in the terminology, we create a class node (oval in Fig. 1) if such a node does not yet exist, otherwise we flag error E1 (the same class should not be represented twice in a terminology):
>
> C2: For each class-relation-class triple employed in a definition,
>
>> C2a: if the target class does not yet exist in the graph we create a corresponding class node and flag error E2 (it should not be the case that a class appears only as a target class in a terminology, i.e. it should always have been first introduced as source class),
>>
>> C2b: if the triple in question expresses an *is-a*-relationship we create an *is-a*-arc (heavy dotted arrow) pointing from the source class node to the target class node,
>>
>> C2c: if the triple expresses a non-*is-a* relationship we create an ASSOC-arc (light dotted arrow) pointing from the source class node to the target class node;

```
S1: For each class-term link in the terminology,
    S1a: we create a lexical node for the term if such a node does not yet exist,
    S1b: we create a semantic arc (dashed arrow) from the class node representing the class from the class-term link to the lexical node;
L1: For each lexical node created,
    L1a: we create a lexical arc (solid arrow) connecting this node to all existing lexical nodes that are its immediate substrings (that is to say substrings which are not themselves properly included within some other existing substring of the lexical node in question),
    L1b: if the lexical node has still not yet been completely decomposed via application of L1a (i.e. if there are still substrings for which no lexical nodes have yet been created in the graph):
        L1b1: we create lexical nodes for each substring for which a lexical node does not yet exist;
        L1b2: create a lexical arc from the lexical node being processed to each of the lexical nodes created in the course of L1b1.
```

2.2. Advanced transformation algorithms

This minimal graph-creation algorithm can be improved in a number of ways. If, like SNOMED-CT®, the terminology to be analyzed employs the notion of a “preferred term”, then it is possible to keep track of whether or not the lexical nodes created do or do not have their origin in a term designated as preferred. This can be achieved by introducing into the graph either a distinct sort of lexical node or a distinct sort of semantic link. One could then flag an additional type of potential error whenever a lexical node representing the preferred term of a class has an incoming lexical link from a lexical node which either does not represent a preferred term for another class or has no ontological links with the first class. Thus in the example terminology of Fig. 1, it would not have been a good idea to choose “bladder” as preferred term for the class “urinary bladder” since the word “bladder” is also used in the term “gall bladder”.

Another useful improvement involves generating class nodes for classes that are not explicitly present in the terminology but that represent potential subsumers for classes which are present. This means that two different sorts of class nodes need to be created in the graph: on the one hand those that represent classes that are fully defined (either fully defined by the terminology, as in the case of SNOMED-CT®, or fully defined by generation through the algorithm), and on the other hand those that represent non-defined classes. Some terminologies, such as SNOMED-CT®, allow also multiple definitions for the same concept, supplied through the mechanism of what SNOMED calls “role groupings” [14].

Creating subsumers can be done with varying degrees of exhaustiveness. A minimalistic approach would involve creating subsumers through combination of one *is-a*-statement taken from the class description together with one other non-*is-a* description. For the terminology of Table 1, this would mean that only two additional class nodes would need to be created, namely the fully defined versions of the class nodes for “urinary bladder inflammation” and “gall bladder inflammation”. Each would then subsume only one class, namely its corresponding primitive version. At first sight, it seems that there is very little that would be gained from this manoeuvre. As we shall see, however, whenever a newly created fully defined class subsumes only its own primitive version in a terminology that strives for a reasonable degree of domain completeness, this provides a strong reason to believe that a class internally declared as primitive should be fully defined under its description. This is to say, rather than labeling the corresponding set of triples as a mere description, it should be labeled as constituting a full definition. We agree that failure to have carried out this task in a particular version of a terminology is a matter more of incompleteness than of error. But it does have computational implications, since it means that classes whose descriptions are drawn from the triples that should have constituted the definition of the fully defined class will never be computed as being subsumed by that class.

The maximally exhaustive version of our algorithm would build a complete subsumption hierarchy by introducing an *Entity* class node that subsumes all other classes. Then, for all triples in the terminology – e.g. (“malignant tumor of breast”, “FINDING SITE”, “breast structure”) – a new class C would then to be created that is fully defined by the triples: (“C”, “*is-a*”, “Entity”) and (“C”, “FINDING SITE”, “breast structure”). Classes in the terminology that enjoy more than one non-*is-a*-description will then, when the classification part of the algorithm has been exeucted, stand in subsumption relations only to one single higher class. Logically, this maximally extended algorithm requires the arcs in the graphs to be labeled with the relationships that are used in the terminology instead of only with either the *is-a* or the non-*is-a* labels. Fig. 2 shows the maximal class node graph for the toy terminology of Table 1.

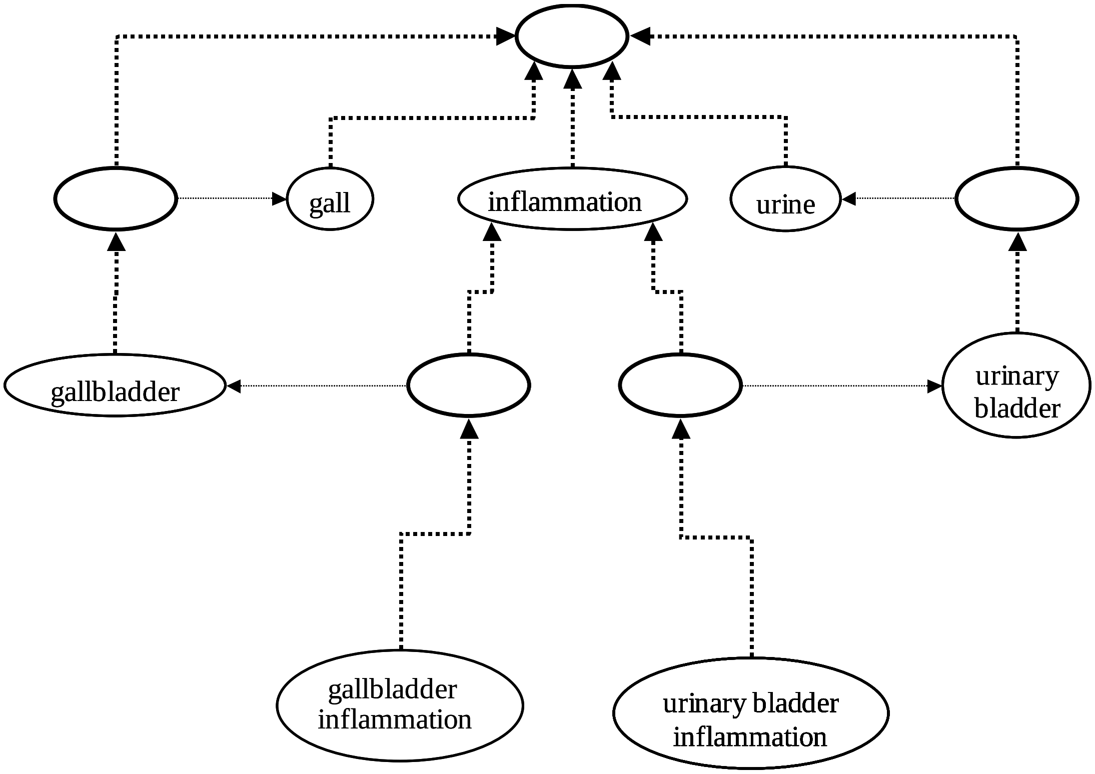

Figure 2: Maximally expanded class node graph for the terminology of Table 1.
(Thick, unnamed nodes represent fully defined classes created by the expansion algorithm.
Heavy dotted arrows represent subsumption, the other arcs association relations.)

2.3. Adding information from external resources

Still better graphs for our error-detection purposes can be produced by resorting to previous versions or releases of the same terminology. Thus linking to previous versions of the ontological part of the terminology may help us identify shifts of reference. One can then apply tools designed to establish whether such shifts have been effectuated uniformly throughout the terminology. Assistance can be sought also from repositories or technologies that have been made available independently of the terminology to be processed.

Thus the lexical part of the graph can be improved by using natural language parsers and additional third-party lexicons, especially those containing spelling variants and inflectional derivations. Parsers can also cope more adequately with the idiosyncracies of punctuation. One should, though, pay attention to possible errors in external lexicons, for these may lead to undesired effects in the graph we are constructing.

Increased error detection and prevention power can be obtained also by linking the ontological part of the graph to external ontologies covering overlapping domains. One way to do this is to use string-matching algorithms to find terms which may be co-referring (i.e. such that they designate the same classes in the two systems at issue). Such string-matching algorithms should be complemented by one or other of the available structure-comparison algorithms, which can be used to assess whether or not terms judged to be co-referring are also directly or indirectly subsumed by other terms which have already been judged to be co-referring. Logically, of course, this approach may tell us just as much about the ontology linked to as about the ontology being processed. Thus only when the ontology linked to is known to be of high quality can it be used to identify mistakes in the source ontology. Otherwise, it can at best be used to generate warnings of possible errors.

2.4. Adding top-level ontological information

Positive results can be more reliably obtained by linking to a simple top-level ontology such as BFO [15], the Basic Formal Ontology developed by IFOMIS on the basis of a long tradition of ontological research in philosophy. Linking to BFO will in turn ensure linking to associated tools for logic-based reasoning, for example about space and time and about mereological and topological relations. BFO distinguishes between *universals* (also called kinds, species, or types) and *particulars* (individuals, instances, or tokens). An example of a universal would be *urinary bladder* as studied in medical school in an anatomy course. An example of a particular would be the urinary bladder present in this particular patient before you here and now. The expression “class” used in this paper correspond to what in BFO are designated as universals – we find that either term is preferable to the more usual “concept” (for the latter is, we believe, itself a source of a number of avoidable confusions [6]). Classes (universals) are what is *general* in reality. It is a common misconception that medical terminologies deal only with classes in this sense. SNOMED-CT®, however, contains obvious references also to particular instances (such as the United States of America or the Food and Drug Administration). (The term ‘instance’, too, incidentally is often a source of confusions in terminology circles, in part because of interference from the special meaning with which this term is employed in database circles.)

Cross-cutting the distinction between universals and particulars in BFO is that between *continuants* and *occurrents* – entities which relate in different ways to time. Continuants endure through time, which is to say that they are wholly present at each moment of their existence. Examples of continuants are organisms, organs, solid tumors, cutters, chromosomes. Occurrents, on the other hand, are never fully present at any given moment in time; rather, they unfold themselves in their successive phases. Examples are processes, such as a tumor invasion or a surgery session. It is important to note that parthood relations never cross such categorial boundaries; that is, parts of continuants are always continuants and parts of occurrents are always occurrents.

A further distinction is that between *independent* entities, such as organisms and cells, that have the ability to exist without the ontological support of other entities, and *dependent* entities, such as colors and shapes, that require the existence of other entities (their bearers) in order to exist. Here, too, parthood relations never cross the boundaries between these two types of entities.

It is our experience that ontologies that do not bend over backwards to make these fundamental distinctions in consistent fashion will contain errors of a sort which are not detected by the standard knowledge representation tools for error checking. This is because such tools focus primarily on the issue of syntactical and logical consistency [16], rather than on ontological correctness.

A top-level ontology of the BFO type cannot be linked automatically to a terminology, (unless such a linkage has been established already by hand for a previous version). And while a complete manual linking may be time consuming, it has been shown nonetheless that linking even a small number of classes in a terminology to corresponding classes in a well-validated ontology such as BFO can be highly beneficial [11].

3. Error detection and prevention algorithms

We now introduce a number of algorithms for error detection and prevention designed to work on graphs generated by the simple or enhanced methods described in **2.** above.

3.1. Algorithms based on semantic distance

Our first class of algorithms works by attempting to find those terms labeled as class nodes in a graph that enjoy the closest (where possible an exact) match to a given input term. Each returned class node is accompanied by a numerical index expressing the effort exerted by the algorithms (the cost they had to pay) in order to retrieve the node by following the arcs in the graph from the input term as starting point. This index can be used as a measure of semantic distance from the input term to the retrieved class nodes. The higher its value (in a well-constructed terminology) the greater should be the semantic distance between the retrieved class node and the input term.

Algorithms of this sort can be built in many flavours, an example being the TermModelling® algorithm described in [12].

When applied to the task of quality assurance for terminologies, such algorithms can be used in two different settings. In the first, the ranking of the semantic distances of the various retrieved class nodes with respect to a given input term is assessed manually for accuracy by domain experts. If the ranking of the retrieved nodes relative to a given input term is judged by an expert to be inaccurate, then this is taken as a prima facie indication of some error in the terminology, which can rest on factors ranging from underspecification, misclassification, unresolved disambiguation (i.e. the ontology might not be aware of the different meanings of homonyms) or even plain mistakes.

As an example, the semantic distance for the retrieved class node “freeing of adhesion of muscle of hand” with respect to the query term “lysis of adhesions of fascia” must be higher than that for “lysis of adhesion of muscle”, as the second should subsume the first. This is because the graphs are constructed in such a way that a path from a class node C to a given input term (i.e. to a given lexical node) always goes via the class nodes representing C’s subsumers. Scores for subsuming classes are therefore always lower than those for subsumed classes (the distance is longer).

The drawback of this method is its need for manual verification of results. However, statistical methods can also be used to scan for unusual distributions of semantic distances. Thus one can start by examing first those class nodes for which the difference in semantic distance between the Nth and (N+1)th ranked entity is significantly greater than the mean difference over all entities; or those class nodes for which the semantic distance of the highest ranked retrieved entity for a specific query term is significantly greater than the mean semantic distance of all the highest ranked entities over all query terms.

In a second setting, we use the output for a given term of the algorithm in one or other of its variants by allowing graph traversal over lexical, semantic and ontological arcs, and compare it to the output resulting from using only ontological arcs (i.e. by taking account only of relations between classes). Different rankings of retrieved nodes can here be flagged automatically – and provide a strong indication of inconsistencies.

The terms used as inputs to these algorithms may come from a variety of sources, one obvious source being the very terms contained within the terminology under review. Of special relevance are shorter terms (as measured by the number of substrings (words) they contain) and terms whose constituent strings, excluding stop words and the modifiers “NOS” and “NEC”, have a high distribution.

3.2. Extended graph structure analysis algorithms

We can also use the structure of the ontological part of the graph generated using the more advanced algorithm described in *2.2.* above. Suspicious configurations are:

- the presence of only one class node subsumed by a given class node in the graph generated by the expansion algorithm;
- the presence, in a list of sibling class nodes for a given subsuming class, of only one pre-existing (non-generated) class node;
- the presence of a pre-existing class node that is subsumed by a generated class node with no further arcs going from the pre-existing class node to other nodes.

A simple algorithm working recursively through the graph is able easily to detect such configurations and to produce output for further manual verification.

Structural discrepancies can also be detected by taking into account both the lexical and ontological parts of the graph. One can, for instance, take as input a class node C1 that is linked by a semantic arc to a lexical node L1 itself linked by lexical arcs to other lexical nodes, say L1a and L1b. (L1a and L1b thus represent terms that are substrings of the term represented by L1.) One can then ask how C1 relates to those class nodes (C2, C3, …) (if any) connected by a semantic arc to the lexical nodes L1a and L1b. An expected configuration is one in which C1 has an outgoing ontological arc to C2, which in turn stands in a semantic arc to L1a. This is exemplified in Fig. 1 by the configuration around the class node “urinary bladder” in relation to the class node “urine”. An unexpected configuration, on the other hand, is illustrated by the configuration around the class node “urinary bladder inflammation”: here a corresponding lexical node – one which would decompose into the lexical nodes mirroring the ontological decomposition of the class node – is missing. The structural discrepancy is thus here an indication of a missing term, namely one built out of the strings “urinary”, “bladder” and “inflammation” arranged in some order (our graph does not contain any syntactic information, hence cannot predict the order in which these words should properly appear). The graph representation and associated algorithms are in this way able to identify not only errors but also omissions.

One can check not only for missing terms, but also for missing classes. Both the class nodes “urinary bladder” and “gall bladder” have semantic arcs to lexical nodes that stand in lexical arcs going to the lexical node “bladder”. Yet the latter is directly semantically related to the class node “urinary bladder” rather than to a class node “bladder” – i.e. to a *generic* bladder class which would subsume both “urinary bladder” and “gall bladder” in a properly constructed hierarchy. Whether “(generic) bladder” does indeed represent an acceptable class is, however, something which can be decided only on the basis of an ontologically supervised analysis of the relevant anatomical facts, and can of course not be inferred by any algorithm.

Obviously, some of the mistakes detected using graph-structure analysis will also be triggered by the semantic distance-based algorithms.

4. Mistakes in SNOMED-CT®

What follows is a brief analysis of errors found when algorithms as described above are applied to the January and July 2003 versions of SNOMED-CT®. In what follows we assign for purposes of further reference in the discussion section of this paper an identifying label of the form “Ja-#”, “Ju-#”, or “Jau-#” to each reported mistake or inconsistency. These labels indicate the presence of the corresponding error in the January, July or in both versions of the system, respectively.

4.1 Human error

Some mistakes must have their origin in inattentiveness on the part of human beings during the manual phases of the process of creating and error-checking SNOMED-CT®. The following are some of the types of errors that were found under this heading.

4.1.1. Improper assignment of is-a relationships

The class “265047004: diagnostic endoscopic examination of mediastinum NOS” is subsumed by “309830003: mediastinoscope”. Thus a procedure is classified as an instrument (Jau-1). The former is marked as “limited” (meaning: “of limited clinical value”) as it is based on a classification concept or an administrative definition. Yet SNOMED-CT® still considers entries with this status as valid for current use and as active.

Another example has a procedure wrongly subsumed by a disease. Thus the class “275240008: Lichtenstien repair of inguinal hernia” is directly subsumed by “inguinal hernia” (Jau-2).

Mistakes of this type can be identified rather easily by comparing the length of the paths that can be followed over the relevant lexical arcs as compared to the much shorter paths (i.e. of length: one arc) in the ontological portion of the graph.

A specific subtype of this sort of mistake consists of the *improper treatment of the partial/complete distinction*. We found 9 classes whose terms included the qualifier “complete” yet were subsumed by altogether 17 classes qualified as “partial”. 6 “partial” classes were, in the other direction, subsumed by 11 “complete” parents. As an example, “359940006: partial breech extraction” is subsumed by “177151002: breech extraction”, which is in turn subsumed by “237311001: complete breech delivery” (Jau-4).

The reason for these mistakes turned out to be the assignment of a term of the form “complete X” to a SNOMED-CT® class with the preferred name “X”, where “X” then also subsumes “partial X”. Mistakes of this type can be detected only when external ontological information is added to the graph – in this case information to the effect that classes qualified as “partial X” are disjoint from classes qualified as “complete X”. Note that searching for evidence of a corresponding structural mismatch only in the ontological part of the augmented graph would not have led to the desired result in relation to the versions of SNOMED-CT® here under review. This is because the latter record no corresponding differences in the class descriptions (other than the difference in immediate subsumer), and this in spite of the fact that SNOMED-CT® *has* classes for “complete” and “partial”. Rather, we need to use the more elaborate version of the graph structure analysis algorithm that traverses also the lexical parts of the graph.

Other subtypes of erroneous assignment of *is-a* relations can be classified under the heading: *improper treatment of negation.* Thus “203046000: Dupuytren’s disease of palm, nodules with no contracture” is subsumed by “51370006: contracture of palmar fascia” (Jau-3). These mistakes can be detected automatically only when the graph is built using a parser that is able to analyse in an appropriate way sentences involving negation operators.

4.1.2. Improper assignment of non-is-a relationships

The class “51370006: contracture of palmar fascia” is linked by the *Finding Site* relationship to the class “64799002: plantar aponeurosis structure”. Probably as a consequence of automated classification, the latter is wrongly subsumed by “disease of foot” (reflecting the fact that “plantar aponeurosis structure” is subsumed by “structure of foot”) (Jau-5). A similar phenomenon is observed in relation to “314668006: wedge fracture of vertebra”, which is subsumed by “308758008: collapse of lumbar vertebra” (Ja-6). Although this erroneous subsumption is no longer present in the July version, the wrong association via *Finding Site*: “bone structure of lumbar vertebra” has been retained (Jau-7). Equally the class “30459002: unilateral traumatic amputation of leg with complication” is classified as an “open wound of upper limb with complications” due to an erroneous association with *Finding Site*: “upper limb structure” (Jau-8).

Errors of this kind can be detected only by adding to the graph an external ontology such as BFO. Their prevention is more difficult, since they are due simply to inattention on the part of the terminologists or ontologists working on the system.

4.2. Technology-induced mistakes

A first example of a mistake of this type has been referred to already above (Jau-5): wrong subsumption because of relationships inappropriately assigned. Other errors are probably induced by tools performig lexical or string matching. We can hardly imagine that a human being would allow “9305001: structure of labial vein” to be directly subsumed by both “vulval vein” and “structure of vein of head”. The error probably comes from an unresolved disambiguation of the word “labia” that is used for both lip (of the mouth) and vulval labia (Jau-9). Error detection is possible for this sort of case only through the exploitation of an external ontology such as BFO and an associated external reference anatomy such as the FMA [17]. Error prevention would then require the terminology authoring system to enforce corresponding class disjointness at run-time.

4.3 Shifts in meaning from SNOMED-RT® to -CT®

The meanings of some SNOMED-CT® terms have changed with respect to the corresponding terms in SNOMED-RT© even where these terms have the same numerical identifier. Above all, the adoption of [18]’s idea of SEP-triplets (structure-entire-part) led to a large shift in the meanings of nearly all anatomical terms. One might argue that in RT anatomical terms such as “heart” were never supposed to mean “entire heart”, but rather always: “heart or any part thereof”; in CT this distinction has been made explicit.

Many other terms appear also to have changed in meaning even though they have the same unique identifier in both RT and CT. A notable example is “45689001: femoral flebography” which in RT relates only to ultrasound but in CT involves in addition the use of a contrast medium (Jau-10). Changes in meaning of this kind can be detected by augmenting the graph with information deriving from the relevant previous version.

There are also changes in the meaning of terms which are less easy to detect or classify. As an example, the meaning of “leg” has changed from SNOMED-RT© to SNOMED-CT®. In RT “leg” was invariably intended to mean “lower leg”; in CT the situation is unclear. The term “34939000: amputation of leg” means in RT: “amputation of lower leg” and in CT: “amputation of any part of the lower limb, including complete amputation” (Jau-11). We observed also numerous examples of inconsistent use of “leg” within CT itself: “119675006: leg repair” refers explicitly to “lower leg structure”, while “119673004: leg reconstruction” refers explicitly to “lower limb structure” (Jau-12). Such inconsistencies become overt when differences in semantic distance are calculated when using the ontological part of the graph only and compared to results for the entire graph.

4.4. Redundant concepts

8,746 SNOMED-CT® concepts were identified in the experiments described in [13] as the seat of redundancies, which is to say: cases where no apparent difference in meaning can be detected between one concept and another one on the basis of the terms that were assigned to it. (This is in reality a severe underestimation, since our parameters for matching lexical variants were set very conservatively, sacrificing recall for precision.) These are all pairs or larger pluralities of terms among which differences in meaning could be identified neither ontologically nor linguistically. Many of them are, we believe, the result of incomplete or inadequate integration of the Read terms into SNOMED-CT®. An astonishing example is “210750005: traumatic unilateral amputation of foot with complication”, which co-exists in SNOMED-CT® with “63132009: unilateral traumatic amputation of foot with complication”. (Jau-13)

Of the same nature is the co-existence of “41191003: open fracture of head of femur” and “208539002: open fracture head, femur” (Jau-14), which fit differently into the class hierarchy but in such a way that the technology used in the development of SNOMED-CT® was unable to find the redundancy involved: the former is directly subsumed by “fracture of femur”, the latter by “fracture of neck of femur”.

Some redundancies become overt only when a larger part of the subsumption hierarchy is examined. Thus one can question to what extent “172044000: subcutaneous mastectomy for gynecomastia” is different from its immediate subsumer “59620004: mastectomy for gynecomastia” when the latter is itself immediately subsumed by “70183006: subcutaneous mastectomy” (Jau-15). All these errors are easily detectable, again, by using semantic distance based algorithms.

4.5. Missing full definitions

The graph expansion algorithm is able to detect many cases in which a SNOMED-CT® class is declared to be “primitive” where it could easily have been fully defined. The typical scenario for this type of mistake is one in which a class node introduced by the expansion algorithm subsumes precisely one class in SNOMED-CT®. Examples are “302829009: adenoma of nipple”, and “63348002: excision of benign tumor of breast”. The latter is especially surprising, given that “46116005: excision of malignant tumor of breast” is itself correctly declared “fully defined”. This again poses questions as to the appropriateness of the methodology that is applied in building SNOMED-CT®.

4.6. Mistakes due to lack of an underlying ontological and anatomical theory

4.6.1. Lack of sound mereotopology

It is difficult to imagine that a single connected object can be a proper part of two regions that are topologically disconnected. Despite this, “45684006: structure of tibial nerve” is directly subsumed by both “thigh part” and “lower leg structure”, which explicitly refer to the upper and lower parts of the lower limb, respectively (Jau-16).

4.6.2. Omission of obvious relationships

Certainly no large terminology can be expected to be complete. However, one can wonder why “248182008: cracked lips” *is-a* “301346001: finding of appearance of lip” but “80281008: cleft lip” *is-a* “disease” and has no relation at all to “finding of appearance of lip” (Jau-17). Such omissions have the consequence that many sound inferences cannot be made. As another example: “181452004: entire uterus” *part-of* “362235009: entire female internal genitalia”, which itself is *part-of* “362236005: entire female genitourinary system”. This means, however, that SNOMED-CT® does not allow the inference to “181452004: entire uterus” *part-of* “181440006: female genital tract”, since the latter has no relationships with “female internal genitalia”, and nor will it allow inferences e.g. to the effect that pregnancy involves the uterus (Jau-18).

Mistakes of this kind can be found, again, only by resorting to additional ontological and anatomical information.

5. Discussion

The importance of our analyses here turns on the fact that so little has been published on the use of sound formal methods for the evaluation of biomedical terminologies and ontologies. In particular, there are few rigorous studies of the mistakes that have been found in the successive versions of SNOMED created.in the last decades

In [19], Ceusters *et al.* analysed the procedure axis of SNOMED International (1998) from the perspective of controlled language principles for the construction of controlled vocabularies, thereby identifying several sources of confusion and ambiguity, including:

- inappropriate use of synonymy (e.g. inconsistently used in preferred terms of "*ear drum*" and "*tympanic membrane*"),
- misleading use of homonymy ("*ventricle*" used both for "*cardiac ventricle*" and "*cerebral ventricle*"),
- incomprehensible concatenation of noun clusters, for example in the term:

  open treatment of craniofacial separation, Lefort III type with wiring and/or local fixation, complicated, fixation by head cap, halo device, multiple surgical approaches, internal fixation, and/or wiring of teeth
- attenuated or ambiguous dependency of modifiers (e.g. in: "*epiphyseal arrest by stapling, combined, proximal and distal tibia and fibula and distal femur*").

It was accordingly argued that term-formation in SNOMED could benefit from the use of a controlled language to make the meaning of terms clearer.

Bodenreider *et al.* used lexical techniques to study the (in)consistent use of modifiers such as "bilateral"/"unilateral", and "congenital"/"acquired" in SNOMED International [20]. Every occurrence of "bilateral X" or "congenital X" would indeed call for a "unilateral X" and "acquired X" respectively, but this requirement was met in very few cases.

A more formal representation of SNOMED class descriptions was at that time not available. Such a formal representation did become available with SNOMED-RT (2000), which however at the same time opened up the possibilities for new types of mistakes. In [21], Campbell reports having found only 0.6% “editorial mistakes” in the portion of SNOMED-RT that he analysed. Whether this surprisingly low figure is accurate is hard to assess. The actual sample size is not given in the paper, but based on his report to the effect that 128 clinical statements from the University of Nebraska Medical Center Lexicon (UNMC Lexicon) were analysed, and that single statements translated into an average of 2.61 SNOMED-RT canonical representation triples, the absolute upper bound of SNOMED-RT statements verified must have been 334. Even given that the 128 statements were a representatively selected sample of 1% from the UNMC Lexicon, it is hard to defend the thesis that these 334 SNOMED triples – constituting less than 0.28% of RT as then constituted, were also representative of SNOMED as a whole. In constrast to Campbell’s positive statement of his results, Elkin *et al.* concluded that “*The current implementation of SNOMED-RT does not have the depth of semantics necessary to arrive at comparable data or to algorithmically map to classifications such as ICD-9-CM*” [22].

Some more information is now available on the quality assurance measures applied in the construction of SNOMED-CT®. There is SNOMED-CT®’s technical reference [23], and Nash’s paper on the problem of erroneous synonymy introduced by merging SNOMED-RT with Read Clinical Terms [24].

The latter describe the process used for developing SNOMED-CT®, but our thesis that this process is still inadequate receives support from Bodenreider *et al.*, who performed a quantitative analysis of SNOMED-CT®, assessing its conformance to a number of principles of good practice in classification [25]. The methodology applied was not suited to the finding of mistakes, but quite sensitive in detecting missing information. As an example, 51% of the assigned parent-child relationships were found to lack differentiating criteria, so that the semantic difference between child and parent remains for these cases unexplained. Furthermore, 31.5% of classes with children have only one child, which suggests for each such class that either at least one child term is missing (from which the available term would then be differentiated), or that there is no semantic difference between parent and child.

We noticed quality improvements in the July as opposed to the January version, as the examples Jau-7 and Ja-6 demonstrate: the wrong subsumption relation with "308758008: collapse of lumbar vertebra" has been removed (though the basic human-introduced mistake was not corrected). Every medical terminology is however of necessity constantly evolving, so that it would be unreasonable to expect perfection or completion from any given release. Through the application of algorithms such as those described above, however, those problematic features which represent practical obstacles to reasoning and to reliable coding are at least to some degree alleviated.

6. Conclusion

The general moral of this paper is that certain characteristic families of mistakes in biomedical terminologies could be prevented through the use of stronger logical and ontological theories implemented in powerful ontology authoring tools. Imposing restrictions to the effect that entities of disjoint top-level categories should not stand in subsumption relations would prevent mistakes like Jau-1 and Jau-2. Enforcement of a framework for the proper treatment of mereotopological relations (incorporating theories of completeness and incompleteness, separation and connectedness, etc.) would prevent Jau-4 and Jau-16 and lead to the flagging of cases like Jau-8 and Jau-9 for possible error. Enforcement of logical relations would prevent cases like Jau-3.

In our view, SNOMED-CT®'s major problem is its failure to pay careful attention to even very simple ontological distinctions, for example between continuant and occurrent entities (i.e. between those entities, such as objects, qualities, conditions, functions – which continue to exist identically through time – and those entities, such as processes and events – which unfold themselves in successive temporal phases). When *procedures* are classified as *instruments* or as *diseases*, then this reflects a conflation of high-level ontological categories that an adequate terminology system should have ways to prevent automatically, ideally through the incorporation of formal-ontological theories of mereology and topology and of an adequate and thoroughly validated reference anatomy such as the FMA. In addition, paying careful attention to formal-ontological distinctions would result in a more accurate treatment of foundational relations such as *is-a* and *part-of* than is possible when the latter are dealt with merely intuitively because they are left formally unanalyzed. Finally, SNOMED-CT® should incorporate a clear opposition between *ontological* notions such as object, process, organism function, and *epistemological* notions such as concept, finding, test result, etc. [26].

As is argued in [27], an approach along these lines can also serve more rigorous but also more intuitive and thus more reliably applicable principles of manual curation than those employed thus far in systems like SNOMED-CT®.

Without doubt, a tremendous effort went into developing SNOMED-CT®, and no other system with similarly broad coverage and attention to detail has a comparable level of formality or richness of content. This makes SNOMED-CT® the first system to be considered when one is in need of medical terminology services. We therefore consider it to be the duty of SNOMED's curators to make themselves constantly apprised of ways to improve its quality, and we hope that the algorithms described above can be of service in this regard.

Acknowledgements

Work on this paper was supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation, by the European Union Network of

Excellence on Semantic Datamining, and by the project “Forms of Life”, sponsored by the Volkswagen Foundation.

References

[1] Smith B, Ceusters W. Towards industrial strength philosophy: how analytical ontology can help medical informatics. Interdisciplinary Science Reviews, 2003; 28: 106-11.

[2] Smith B. Mereotopology: a theory of parts and boundaries, Data and Knowledge Engineering 1996; 20: 287-301.

[3] Smith B, Varzi AC. Fiat and bona fide boundaries, Proc COSIT-97, Berlin: Springer. 1997: 103-119.

[4] Buekens F, Ceusters W, De Moor G. The explanatory role of events in causal and temporal reasoning in medicine. Met Inform Med 1993; 32: 274-278.

[5] Ceusters W, Buekens F, De Moor G, Waagmeester A. The distinction between linguistic and conceptual semantics in medical terminology and its implications for NLP-based knowledge acquisition. Met Inform Med 1998; 37(4/5): 327-33.

[6] Smith B: Beyond concepts: Ontology as reality representation. Proceedings of FOIS 2004. International Conference on Formal Ontology and Information Systems, Turin, 4-6 November 2004.

[7] Kumar A, Schulze-Kremer S, Smith B. Revising the UMLS Semantic Network, In M. Fieschi, E. Coiera and Y-C.J. Li (eds) Proceedings of the 11th World Congress on Medical Informatics. Electronic Publication.

[8] Vizenor L. Actions in health care organizations: An Ontological Analysis. Medinfo 2004 ;2004();1403-10.

[9] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. AMIA Annu Symp Proc. 2003:609-13.

[10] Kumar A, Smith B. The Unified Medical Language System and the Gene Ontology, KI 2003: Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence 2821), 2003; 135–148.

[11] Casella dos Santos M, Dhaen C, Fielding M, Ceusters W. Philosophical scrutiny for run-time support of application ontology development. In: Varzi AC, Vieu L (eds) Formal Ontology in Information Systems, Proceedings of the Third International Conference FOIS 2004; 2004, 342-52.

[12] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: Where do they come from and how can they be detected? in Pisanelli DM (ed) Ontologies

in Medicine. Proceedings of the Workshop on Medical Ontologies, Rome October 2003. IOS Press, Studies in Health Technology and Informatics, vol 102, 2004;145-164.

[13] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT(R). Medinfo 2004 ;2004();482-6.

[14] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. Proc AMIA Symp. 2002;:712-6.

[15] Grenon P, Smith B, Goldberg L. Biodynamic ontology: Applying BFO in the biomedical domain, in Pisanelli DM (ed) Ontologies in Medicine. Proceedings of the Workshop on Medical Ontologies, Rome October 2003. IOS Press, Studies in Health Technology and Informatics, vol 102, 2004; 20-38.

[16] Ceusters W, Smith B. Ontology and medical terminology: Why descriptions logics are not enough. TEPR 2003 (electronic publication): http://ontology.buffalo.edu/-medo/TEPR2003.pdf

[17] Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform. 2003 Dec;36(6):478-500.

[18] Hahn U, Schulz S, Romacker M: Part-whole reasoning: a case study in medical ontology engineering. IEEE Intelligent Systems and Their Applications vol 14 nr 5, 1999: 59-67.

[19] Ceusters W, Steurs F, Zanstra P, Van Der Haring E, Rogers J. From a time standard for medical informatics to a controlled language for health. Int J Med Inform. 1998 Feb;48(1-3):85-101.

[20] Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. Int J Med Inf. 2002 Dec;67(1-3):85-95.

[21] Campbell JR. Semantic features of an enterprise interface terminology for SNOMED RT. Medinfo. 2001;10(Pt 1):82-5.

[22] Elkin PL, Harris M, Ogren PV, Buntrock ID, Brown SH, Solbrig HR, Chute CG. Semantic augmentation of Description Logic based terminologies. Addendum to Proceedings of IMIA-WG6, Medical Concept and Language Representation, Phoenix, 1999;70-81.

[23] College of American Pathologists. Snomed Clinical Terms® Technical Reference Guide, July 2003 release.

[24] Nash SK. Nonsynonymous synonyms: correcting and improving SNOMED CT. AMIA Annu Symp Proc. 2003;:949.

[25] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A case study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, editors. Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004); 2004. p. 12-20.

[26] Bodenreider O, Smith B, and Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology, in Achille Varzi and Laure Vieu (eds.), Formal Ontology and Information Systems. Proceedings of the Third International Conference (FOIS 2004), Amsterdam: IOS Press, 2004, 185–195.

[27] Smith, B. and Rosse, C. The role of foundational relations in the alignment of biomedical ontologies. In Proc MedInfo, San Francisco, CA. 2004;:444-448.