From: Hahn U, Schulz S, Cornet R, editors. ***Proceedings of the First International Workshop on Formal Biomedical Knowledge Representation*** **(KR-MED 2004), 12-20.**

Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED CT

Olivier Bodenreider [1], Barry Smith [2,3], Anand Kumar [2], Anita Burgun [4]

[1] US National Library of Medicine, Bethesda, Maryland, USA
[2] Institute for Formal Ontology and Medical Information Science, Univ. Leipzig, Germany
[3] Department of Philosophy, University at Buffalo, New York, USA
[4] Laboratoire d'Informatique Médicale, Université de Rennes I, France

Formalisms such as description logics (DL) are sometimes expected to help terminologies ensure compliance with sound ontological principles. The objective of this paper is to study the degree to which one DL-based biomedical terminology (SNOMED CT) complies with such principles. We defined seven ontological principles (for example: each class must have at least one parent, each class must differ from its parent) and examined the properties of SNOMED CT classes with respect to these principles. Our major results are: 31% of the classes have a single child; 27% have multiple parents; 51% do not exhibit any differentiae between the description of the parent and that of the child. The applications of this study to quality assurance for ontologies are discussed and suggestions are made for dealing with multiple inheritance.

INTRODUCTION

Biomedical terminologies and ontologies are increasingly taking advantage of Description Logics (DL) in representing knowledge. GALEN[1] and SNOMED Clinical Terms® (in what follows SNCT)[2] were both developed in a native DL formalism. Several other groups have worked at converting existing terminologies into terminologies with a DL formalism (UMLS® Metathesaurus® [1-3], UMLS Semantic Network [4], Gene Ontology™ [5], National Cancer Institute Thesaurus [6]). Protégé-2000's OWL plug-in now also allows developers of frame-based resources to export their ontologies into DL formalism.

The validation of an ontology by a DL-based classifier allows compliance with certain rules of classification (e.g., absence of terminological cycles) and it brings also other benefits in terms of coherence checking and query optimization [7, 8]. However, neither a DL formalism nor the use of a classifier can ensure compliance with all principles of a sound ontology [9].

The objective of this paper is to study the degree to which one DL-based biomedical terminology complies with such ontological principles. We selected SNCT as target for this evaluation because it is the most comprehensive biomedical terminology recently developed in native DL formalism. Another reason for our choice is that SNCT will soon be available as part of the UMLS[3] (at no charge for UMLS licensees in the U.S.) and is therefore likely to become widely used in medical information systems.

This paper is organized as follows. We first define a limited number of basic ontological principles with which biomedical ontologies are expected to be compliant. (These are in effect principles of good classification.) We then give a brief description of SNCT, we present the methods used to test the compliance of SNCT with these principles, and we summarize our results. Finally, we discuss the application of this method to quality assurance in ontologies and terminologies, laying special emphasis on the role of creating partitions in ontologies, and we also outline other implications of our results.

BACKGROUND

Terms, classes, and instances. We shall refer to the nodes in SNCT not as concepts but rather on the one hand as *terms* (where we are interested in the hierarchy itself, as a syntactic structure), and on the other hand as *classes* (where we are interested in the bio-

[1] http://www.opengalen.org/
[2] http://www.snomed.org/snomedct_txt.html
[3] http://umlsinfo.nlm.nih.gov/

logical entities to which these terms refer). It is classes, not concepts, which stand in *IS A*, *PART OF* and similar relations in biomedical ontologies. Classes have *instances*. In the biomedical domain, instances are generally represented in health information systems (e.g., electronic patient records) or in biomedical experiments (e.g., in the form of microarray experiments), while biomedical terminologies and ontologies are focused on classes and their relations.

Relations among classes. The possible relations of class *A* to class *B* are defined in Table 1. *A* is the root of a given taxonomy if and only if every class in the taxonomy is a child of *A*; conversely, *A* is a leaf of a given taxonomy if and only if *A* has no children.

Relation	Definition
A = *B*	*A* and *B* are the same entity (i.e., they have the same definition, and thus also the same family of instances at any given time)
A IS A *B*	1. *A* and *B* are classes and 2. all instances of *A* are instances of *B*
A is a child of *B*	1. *A* IS A *B*, 2. *A* ≠ *B*, and 3. if *A* IS A *C* and *C* IS A *B* then *A* = *C* or *C* = *B*
A and *B* are siblings	1. there is some *C* of which *A* and *B* are both children and 2. *A* ≠ *B*
A is a parent of *B*	*B* is a child of *A*
C is a differentia of *A* with respect to *B*	1. *A* IS A *B*, 2. *A* ≠ *B*, and 3. instances of *A* are marked out within the wider class *B* by the fact that they exemplify *C*

Table 1 – Definition of the relations between classes A *and* B

Principles of classification. Scientific classification has evolved from Aristotle to Linnaeus to large and varied classifications of modern times. Along the way, classification principles were elaborated. One such principle, resulting from the use of a unique *fundamentum divisionis* or single classificatory principle in differentiating the species of each successive genus, is that subclasses be mutually exclusive and jointly exhaustive [10]. Some other highly general organization and classification principles – which we believe rest on a wide consensus among those working on biomedical terminologies [11, 12] – are:

- Each hierarchy must have a single root
- Each class (except for the root) must have at least one parent
- Non-leaf classes must have at least two children
- Each class must differ from each other class in its definition. In particular: each child must differ from its parent and siblings must differ from one another

Principles of subsumption. More interestingly, principles can also be derived from the study of the way subsumption is in fact treated in biomedical terminologies and ontologies. As noted by Bernauer [13], two major types of difference can be observed between a parent and its child: the introduction in the child of a new "criterion" (introduction of a *role* in DL parlance), and the *refinement* of an already existing criterion (corresponding to DL's *refinement of a role value*[4]). For example, the introduction of the role *CAUSATIVE AGENT* with value *Infectious agent* explains the subsumption relation of *Meningitis* to *Infective meningitis*. Similarly, the subsumption relation of *Infective meningitis* to *Viral meningitis* is explained by the refinement of the role value for *CAUSATIVE AGENT* since *Infectious agent* subsumes *Virus*. Such refinement can be a matter of specialization as in the previous example, where the role value for the parent is more generic than that for the child. Less frequently, partitive refinement can occur. For example, *Neuropathy* subsumes *Peripheral motor neuropathy* because the value in the parent of the role *FINDING SITE* (*Nerve structure*) includes as part the corresponding value in the child (*Peripheral motor neuron*).

The following *inheritance principle* is standardly taken for granted in work on ontologies and terminologies: if *A* is a child of *B* then all properties of *B* are also properties of *A*. As a corollary, no cycles are allowed in an *IS A* hierarchy. Additionally, one inheritance principle based on our approach to subsumption can be expressed as follows: All roles of a parent class must either be inherited by each child or refined in the child. From the perspective of the child, differentia from child to parent should uniquely result in every case either from refinement of the value of a common role or introduction of a new role

Single vs. multiple inheritance. Some of the principles presented above are the object of a large consensus (e.g., *that each class must have at least one parent* is needed if a terminology is to have a proper hierarchical structure). Others, however, still spur debate among terminology developers. This is the case in regard to the issue of single vs. multiple inheritance, i.e., of whether classes should be allowed to have more than one parent. As noted by Cimino:

[4] Also called role filler in DL parlance.

"There is some disagreement, however, as to whether concepts should be classified according to a single taxonomy (strict hierarchy) or if multiple classifications (polyhierarchy) can be allowed." While it is beyond the scope of this paper to argue for or against multiple inheritance, we will make some suggestions for dealing with this issue in the discussion.

MATERIALS

SNOMED CT was formed by the convergence of SNOMED RT and Clinical Terms Version 3 (formerly known as the Read Codes). The version used in this study (January 31, 2004) contains 269,864 classes. The first level is subdivided into eighteen classes listed in Table 2 with their frequency distribution.

Class	Freqency
Attribute	990
Body structure	30,651
Clinical finding	95,604
Context-dependent categories	3,648
Environments and geographical locations	1,619
Events	86
Observable entity	7,273
Organism	25,025
Pharmaceutical / biologic product	16,866
Physical force	198
Physical object	4,200
Procedure	46,065
Qualifier value	8,133
Social context	4,895
Special concept	177
Specimen	1,052
Staging and scales	1,097
Substance	22,266

Table 2 – The 18 first-level classes in SNOMED CT and their frequency distribution

Role	Value
CAUSATIVE AGENT	*Virus*
ONSET	*Sudden onset; Gradual onset*
SEVERITY	*Severities*
EPISODICITY	*Episodicities*
COURSE	*Courses*
ASSOCIATED MORPHOLOGY	*Inflammation*
FINDING SITE	*Meninges structure*

Table 3 – Roles present in the description of Viral meningitis

Each SNCT class has a description[5] consisting of a variable number of elements. For example, the class *Viral meningitis* has a unique identifier (58170007), two parents (*Infective meningitis* and *Viral infections of the central nervous system*), several names (*Viral meningitis*, *Abacterial meningitis*, and *Aseptic meningitis, viral*). The roles present in the description of this class are listed in Table 3.

In addition to a unique identifier, each class is assigned a unique, fully specified name consisting of a regular name suffixed (in parentheses) with a reference to what SNCT calls the "primary hierarchy" of the class, the latter corresponding roughly to one of the top-level classes in the hierarchy. For example, the fully specified name for *Viral meningitis* is *Viral meningitis (disorder)*[6]. This assignment to a primary hierarchy is not explicitly recognized as a property of the class in the SNCT representation. However, because the corresponding high-level category can be easily extracted from the fully specified name of the class, we found it useful it to use it for purposes of categorizing SNCT classes. Thus for example we will use *disorder* as the category for *Viral meningitis*. The list and frequency distribution of such categories in SNCT is presented in Table 4.

administrative concept	54	*navigational concept*	165
assessment scale	870	*observable entity*	7,274
attribute	991	*occupation*	4,153
body structure	25,395	*organism*	25,026
cell	603	*person*	302
cell structure	501	*physical force*	199
context-dependent category	3,649	*physical object*	4,201
disorder	62,301	*procedure*	42,782
environment	1,007	*product*	16,867
environment / location	1	*qualifier value*	8,080
ethnic group	254	*regime/therapy*	3,284
event	87	*religion/philosophy*	145
finding	33,304	*social concept*	21
geographic location	612	*special concept*	1
inactive concept	7	*specimen*	1,053
life style	21	*staging scale*	15
morphologic abnormality	4,153	*substance*	22,267
namespace concept	5	*tumor staging*	213

Table 4 – The list of high-level categories ("primary hierarchies") in SNOMED CT and their frequency distribution

Inheritance in SNCT is indicated by the presence of *IS A* relationships among classes. For example, the class *Fracture of calcaneus* subsumes two classes (*Closed fracture of calcaneus* and *Open fracture of calcaneus*). The difference between the descriptions of the classes *Fracture of calcaneus* and *Closed fracture of calcaneus* lies in the presence of a special-

[5] Throughout this paper, we use 'description' with the common meaning that is also standard in the DL-context, i.e., to refer to the list of properties of a given class (more precisely: of its instances), expressed by roles. In SNOMED CT parlance, however, a description corresponds to a name for a class.

[6] The primary hierarchy for *Viral meningitis* is *Clinical finding*, while the category mentioned in parentheses in the fully specified name is *disorder*.

ized value for the role *ASSOCIATED MORPHOLOGY* in the child (*Fracture, open*[7]) compared to that of the parent (*Fracture*). Also of note, the class *Fracture* subsumes *Fracture, open*. The refinement of the value of the role *ASSOCIATED MORPHOLOGY* between the two classes constitutes the differentia, while the other roles are all inherited from the parent class.

METHODS

The methods presented below were developed for testing the compliance of SNCT with the seven principles listed in Table 5.

P1	Each class must have at least one parent
P2	Non-leaf classes must have at least two children
P3	Children should have exactly one parent
P4	Each hierarchy must have a single root
P5	Each child's description must differ from its parent's description
P6	All roles of a parent class must either be inherited by each child or refined in the child
P7	Differentia from child to parent should uniquely result in every case either from refinement of the value of a common role or introduction of a new role

Table 5 – Ontological principles studied in SNCT

Quantitative analysis: Number of parents, children, and roots

By simply counting the number of parents and children for each class, we verify the degree of compliance with **P1**, **P2**, and **P3**. Additionally, the existence of a path between each class and the eighteen top-level classes is tested by traversing the graph of all classes in SNCT from each class upwards. We use this method for verifying **P4**.

Qualitative analysis of differentiae

In order to verify SNCT's compliance with **P5**, we analyze the differentiae in pairs of parent-child classes by comparing the roles and role values for each class in the pair. First, we verify that at least one role or one role value is present in the description of the child but not in that of the parent.

The second step consists in examining the roles shared by the two classes and those specific to each class. All roles of the parent are searched for in the description of the child in order to verify compliance with **P6**.

[7] Despite similarities in their names, *Fracture, open (morphologic abnormality)* and *Open fracture (disorder)* are distinct classes in SNOMED CT.

The relationship between the values of a role shared by the parent and child classes is examined and is expected to be either specialization (*IS A*) or partitive refinement (*PART OF*). The presence of roles specific to the child is also examined. The number of differentiae (i.e., the number of role values refined and of roles introduced in the child) is recorded. This step is used to verify **P7**.

RESULTS

Quantitative analysis: Number of parents, children, and roots

Number of children

The number of children per class ranges from 0 to 2532. The frequency distribution of the number of children is presented in Figure 1. 196,237 classes (73%) have no children. These classes are leaf nodes in the SNCT hierarchy. Examples of such classes include the substance *Tartrate dehydratase*, the finding *Anuria*, the organism *Trypanosoma evansi*, and the body structure *Upper left third premolar tooth*.

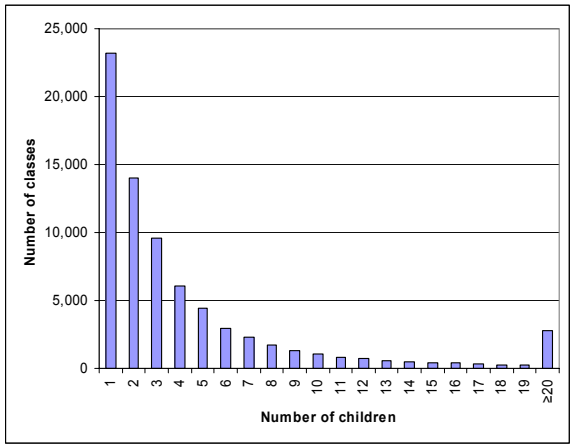

Figure 1 – Distribution of the number of children

Out of 73,627 classes with children, 23,174 classes (31.5%) have a single child. This proportion is relatively constant across SNCT categories. Examples of classes with a single child include {*Cervical secretion sample*, child: *Cervical mucus specimen*} (*specimen*), {*Deferoxamine*, child: *Deferoxamine mesylate*} (*substance*), {*Multiple polyps*, child: *Multiple adenomatous polyps*} (*morphologic abnormality*), and {*Referral to general medical service*, child: *General medical self-referral*} (*procedure*).

8,034 classes (11%) have ten children or more and 150 have more than 99 children. The median number of children is 2. Example of classes with a large number of children include *Infectious gastroenteritis*

(10 children), *Operation on heart valve* (25 children), *Sodium compound* (51 children), and *Disorder of eye proper* (100 children).

Some classes have an unusually large number of children, including *Veterinary proprietary drug AND/OR biological* (2532 children), *Biochemical test* (996 children), the substance *Oxidoreductase* (580 children), the organism *Bos taurus* (551 children), and *Congenital malformation* (505 children). Although these classes often correspond to large collections of drugs, tests, or disorders, the large number of children in these classes may point to issues such as a lack of organization or incomplete descriptions.

Number of parents

Except for the root, every class of SNCT has at least one parent. The number of parents per class ranges from 1 to 13.[8] The frequency distribution of the number of children is presented in Figure 2. 195,053 classes (72.3%) have a single parent, 53,517 classes (19.8%) have two parents, 13,969 classes (5.2%) have three, 4,692 classes (1.7%) have four, and 2,632 classes (1.0%) have five or more.

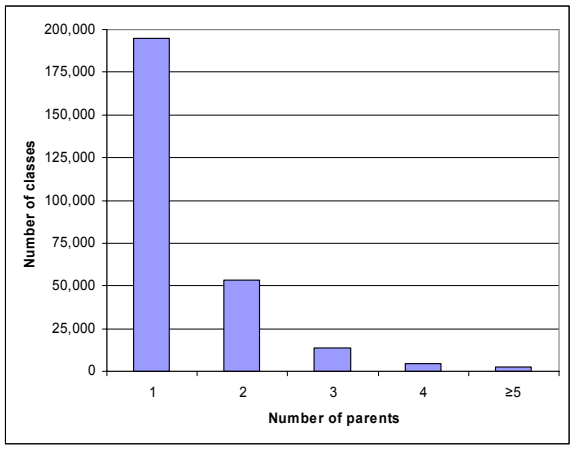

Figure 2 – Distribution of the number of parents

Overall, the proportion of classes having multiple parents, i.e., exhibiting multiple inheritance, is 27.7%. This proportion tends to be higher in some categories (e.g., around 45% for *body structure*, *disorder*, and *procedure*) and lower in others (e.g., around 5-15% for *cell*, *organism*, and *substance*).

Number of roots

Except for the root and for the eighteen top-level classes of SNCT excluded from this test, each class can be linked hierarchically to exactly one top-level class. This means that SNCT consists of eighteen independent hierarchies.

Qualitative analysis of differentiae

Existence of a differentia between parent and child

Out of the 377,681 parent-child relations examined, 193,957 (51%) do not exhibit any differentiae between the description of the parent and that of the child. However, the presence or absence of differentiae in children varies considerably across categories. In most categories – including *geographical location*, *organism*, and *substance* – no differentiae are ever mentioned. In the other categories, the proportion of children exhibiting differentiae in their description ranges from 29% (*cell*) to 86% (*specimen*).

Number and nature of differentiae

When there does exist a differentia between a child and its parent, i.e., when their descriptions are not identical, the difference in the descriptions can affect one role or multiple roles, and one or more values within each role.

Single differentia. Out of the 183,724 parent-child relations where there is at least one differentia between the child and its parent, 102,426 (56%) exhibit exactly one differentia. For example, the classes *Fracture of calcaneus* and *Open fracture of calcaneus* presented earlier differ only by the value of their common role *ASSOCIATED MORPHOLOGY*. In 60% of the cases, the differentia comes from the refinement of the value for a given role; in 40% of the cases, it comes from the introduction of a new role in the child. The example above (*Fracture of calcaneus*) illustrates the refinement (from *Fracture* to *Fracture, open*) of the role *ASSOCIATED MORPHOLOGY*. Conversely, the introduction of the role *FINDING SITE* (with value *Ear structure*) differentiates the class *Otitis* from its parent *Inflammatory disorder*.

Multiple differentiae. In case of multiple differentiae, the differentiae involved reflect the introduction of several roles (34%), the refinement of several values (20%), or the combination of introducing at least one role and refining at least one value (46%). For example, *Endoscopy of jejunum* differs from *Procedure on jejunum* by 1) the introduction of two roles (*METHOD*, with value *Inspection – action*, and *ACCESS INSTRUMENT*, with value *Endoscope, device*) and 2) the refinement of the role *ACCESS* (from *Surgical access values* to *Endoscopic approach – access*). Figure 3 illustrates the roles introduced and inherited for the class *Endoscopy of jejunum*. Not surprisingly, multiple differentiae are often associated with multiple inheritance. In the example above,

[8] The three classes with 13 parents are *Anoscopy with coagulation for control of hemorrhage of mucosal lesion*, *Mandibuloacral dysostosis*, and *Entire sternocleidomastoid muscle*.

the role *METHOD* is actually inherited (and refined from *Evaluation – action* to *Inspection - action*) from *Gastrointestinal investigation*, the second parent of *Endoscopy of jejunum*. The role *ACCESS INSTRUMENT*, however, is truly specific to *Endoscopy of jejunum* (i.e., not present in any of its parents).

Our analysis of differentiae reveals a number of **other potentially problematic issues**. In 7,226 cases, some role or value present in the parent is not inherited or refined in the child. For example, the role *ONSET* has two possible values in the class *Subjective visual disturbance* (*Sudden onset* and *Gradual onset*), of which *Gradual onset* is not inherited by its child class *Sudden visual loss*. The role *ONSET* is involved in roughly half of the cases where some role is specific to a parent class but eleven other roles are also involved in this phenomenon.

In 21,799 cases, although the parent and child classes share a role, the values of this role are neither identical (inherited by the child from the parent) nor such as to stand in any taxonomic relation (with the specialized value in the child) or meronomic relation (with the part in the child). For example, the class *Diabetic retinopathy* and its child *Diabetic retinal microaneurysm* share the role *FINDING SITE*, but their values for this role (*Retinal structure* and *Visual pathway structure*) do not stand in a hierarchical relation. Typically, this problem is associated with multiple inheritance. The role value which does not stand in hierarchical relation with corresponding role values in one parent most often does in one of its other parents. In the example above, *Retinal structure* is actually inherited from *Retinal microaneurysm*, the other parent of *Diabetic retinal microaneurysm*.

DISCUSSION

The work described in this paper is in the tradition of studies auditing large medical terminologies such as [14]. However, we are interested here not just in the consistency of the terminological structure but also in compliance with general classification principles. We found SNCT to be fully compliant with principles such as *each class must have at least one parent* and *each hierarchy must have a single root*. In contrast, we observed non-compliance with many other principles, the consequence of which will be presented next. We will then revisit the problem of single vs. multiple inheritance and outline a possible solution to it.

Application to quality assurance for ontologies

Non-leaf classes with a single child

The recognition by biologists of the phylum *Chordata* rests on the distinction of several subphyla: *Vertebrata* (or *Vertebrates*), *Cephalochordata*, and *Urochordata*. Compared to *Vertebrates*, the latter two might be of lesser relevance to clinical medicine. However *Vertebrates* is defined in opposition to the two other subphyla and all three should therefore be represented in a well-formed ontology of organisms. Moreover, in a world in which *Vertebrates* had only one child, the distinction between parent and child would not be made by biologists. Therefore, the presence of such cases is reason to suspect the presence of error.

The review of a limited number of classes having a single child suggests the following possible issues. One is the incompleteness of the hierarchy (e.g., *Subphylum Vertebrata* is the only subphylum recorded in SNCT for *Phylum Chordata*). Another issue is the presence of a hybrid class, resulting from the intersection of two parent classes, as the single child of at least one of the two parent classes (e.g., *Closure of abdominothoracic fistula*, hybrid child of *Closure of fistula of thorax* and *Abdomen closure*) and single child of *Closure of fistula of thorax*). Finally, the presence of redundant classes, where a parent and a child class bear no differences, can also be at the origin of single child classes. This issue is discussed in detail in the next section.

Among the 23,174 single child classes, 12,928 (56%) have a single parent and therefore do not correspond to hybrid classes. Examples of such classes can be found in virtually every category and include the procedure *Arthroscopy of toe* (single child of *Arthroscopy of foot*), the disorder *Congenital absence of lobe of liver* (single child of *Congenital absence of liver*), and the substance *Urine* (single child of *Urinary tract fluid*).

Absence of difference in the description between children and parents

Beyond hierarchy, one of the major reasons for interest in DL-based systems is that they promise to make available for formal reasoning tools detailed descriptions for each class, representing through roles the defining characteristics of these classes. However, DL systems can also accommodate classes with minimal descriptions (i.e., restricted to bare subsumption links). We reviewed a small number of classes (in the domain of disorders) for which no difference was provided between the parent and the child in terms of roles or role values. The major issue brought to light by this limited analysis seems to be the incomplete-

ness of many descriptions. For example, while no difference is provided between the descriptions of *Bullous lichen planus* and *Lichen planus*, such a difference is provided for *Bullous dermatosis* (*ASSOCIATED MORPHOLOGY* with value *Blister*) and *Skin lesion*. In other cases, the representation of some characteristics seems to have been purposely omitted (e.g., *COURSE* for acute and subacute variants of diseases, although *Acute* exists as a class). Generally, morphologic distinctions seem better represented than physiological ones. Also of note, some classes represent what are in fact mere collections (e.g., *Extrapyramidal disease*). These classes are defined in extension (i.e., via a list of their subclasses) rather than in intension (i.e., via a list of characteristics). Extensional definitions are less desirable since they imply the need for more radical revisions in light of the discovery of new types of cases.

Finally, in some cases, there is actually no difference to be represented between the parent and the child class (e.g., *Closed fracture of skull without intracranial injury* vs. *Closed fracture of skull*). The issue, in this case, is the presence of two classes for representing one biomedical entity. The distinction between the two classes lies not in the biomedical entity they represent (i.e., the skull is fractured, but not open), but merely in the knowledge of the physician that intracranial injuries might be associated with such fractures. In other words, this distinction is epistemological in nature and, arguably, should not be represented in an ontology. It would be a valuable extension of the current DL in SNCT if ways could be found to do justice to operators, such as 'with' and 'without,' which play an important role in the organization of SNCT's term hierarchy. As things stand, the information conveyed by such operators is not accessible in ways which would support reasoning with terminological knowledge in medicine. This means more generally that the information conveyed by the compositional structure of SNCT's terms is at the moment not available for automatic retrieval.

Presence of roles specific to the parent class

In most of the cases we examined, the presence in a parent's description of roles not inherited by its children has to do with the representation of specialization in DL-based structures. As noted earlier, *Subjective visual disturbance* is described as having possibly a *Sudden onset* or a *Gradual onset*. However, the only valid onset for its child *Sudden visual loss* is *Sudden onset*. Therefore, *Sudden visual loss* can be seen as a specialization of *Subjective visual disturbance*. This could be represented in DL form by '∀(*HAS-ONSET Onsets*)' for *Subjective visual disturbance* and '∃(*HAS-ONSET Sudden onset*)' for *Sudden visual loss* [15].

Characterizing inheritance

The uncontrolled use of *IS A* to signify a variety of different sorts of relations (including *PART OF*, *IS AN INSTANCE OF*, and so on) results in what Guarino has called '*IS A* overload', which is often associated in turn with examples of incorrect subsumption [16]. Examples of this phenomenon in SNCT include *Both testes IS A Testis Structure*, *Deferoxamine mesylate IS A Deferoxamine*, and *Urine sediment IS A Urine*.

IS A overload, which is often associated with multiple inheritance, may be alleviated by making explicit which sort of subsumption link is involved in each specific type of case – for example by replacing *IS A* as it occurs between *Viral meningitis* and *Infective meningitis* with *IS* A_{AGENT} or as it occurs between *Viral meningitis* and *Viral infection of the central nervous system* with *IS* A_{SITE}.

The use of such explicit subsumption links also enables a large taxonomy such as SNCT to be divided into *partitions* within which taxonomic reasoning can be more reliably performed. Through a locative partition, for example, which we can think of as a window or view on reality with a specific type of focus, *Viral meningitis* would appear in its locative guise: as a *Viral infection of the central nervous system*, and inferences could be performed safely along the *IS* A_{SITE} relationship within this partition. Analogously, in a causative partition, *Viral meningitis* would be linked to *Infective meningitis* and subsumption could be performed safely along the *IS* A_{AGENT} relationship. The locative and causative partitions would then yield complementary views of different aspects of one and the same reality. This view is illustrated in Figure 4, and the underlying formal theory is presented in [17].

CONCLUSIONS

SNCT is the most comprehensive biomedical terminology recently developed in native DL formalism and is expected to play an important role in clinical information systems. Unlike thesauri built for information retrieval purposes, SNCT should enable reasoning about biomedical knowledge. We have listed some principles, mostly related to classification, and tested the degree to which SNCT complies with them. While we found SNCT to be more coherent than many other terminologies, we also found the description of many of its classes to be minimal or incomplete, with possible detrimental consequences on inheritance.

Description logics provide a formalism suitable for representing many features of a variety of different domains – including the biomedical domain – in a

way that can support automatic reasoning and information retrieval. In and of themselves, however, DLs do not systematically ensure compliance with the principles of classification required if reasoning is to be performed accurately. More than the use of any formalism, we believe that compliance with sound ontological principles is what guarantees the accuracy of reasoning.

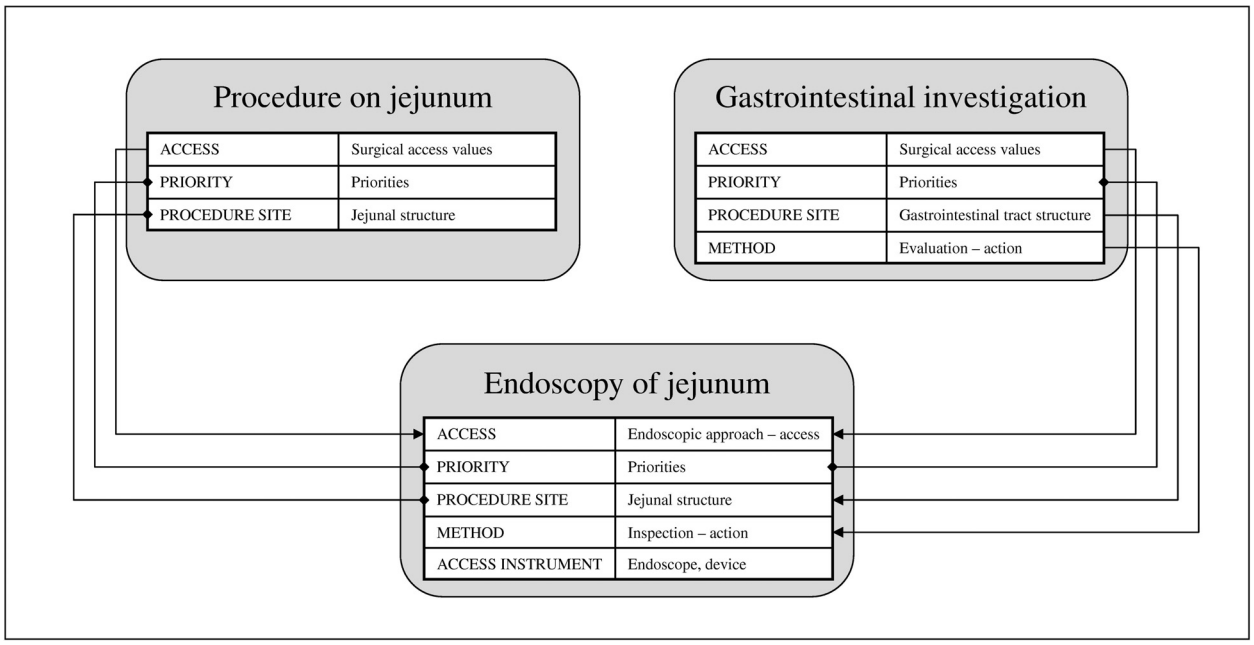

Figure 3 – Inheritance of role values for Endoscopy of jejunum.

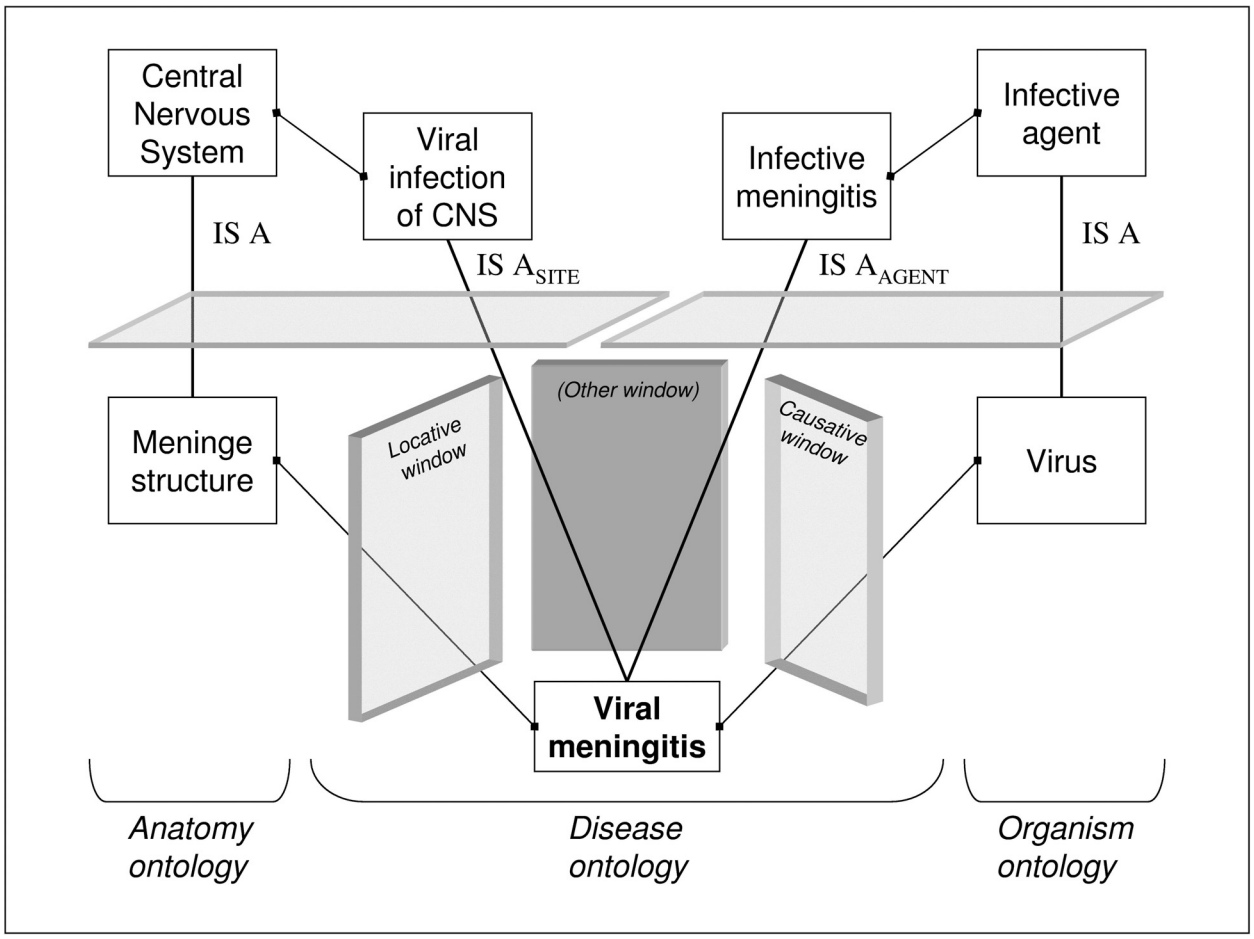

Figure 4 – Two views (locative and causative) on Viral meningitis.

Acknowledgements

Smith and Kumar are supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation.

References

1. Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. *Proc AMIA Symp* 1998:810-4.
2. Cornet R, Abu-Hanna A. Usability of expressive description logics--a case study in UMLS. *Proc AMIA Symp* 2002:180-4.
3. Hahn U, Schulz S. Towards a broad-coverage biomedical ontology based on description logics. *Pac Symp Biocomput* 2003:577-88.
4. Kashyap V, Borgida A. Representing the UMLS Semantic Network using OWL: (Or "What's in a Semantic Web link?"). In: Fensel D, Sycara K, Mylopoulos J, editors. *The SemanticWeb - ISWC 2003*. Heidelberg: Springer-Verlag; 2003. p. 1-16.
5. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput* 2003:624-35.
6. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. *Journal of Web Semantics* 2003;1(1).
7. Horrocks I, Rector A, Goble C. A Description Logic based schema for the classification of medical data. In: Baader F, Buchheit M, Jeusfeld MA, Nutt W, editors. *Proceedings of the 3rd Workshop KRDB'96*; 1996. p. 24-28.
8. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000;16(2):184-5.
9. Ceusters W, Smith B, Flanagan J. Ontology and medical terminology: Why Description Logics

are not enough. *Proceedings of TEPR 2003 - Towards an Electronic Patient Record. San Antonio, Texas, May 10-14, 2003* 2003:(CD-ROM publication).

10. Marradi A. Classification, Typology, Taxonomy. *Quality & Quantity* 1990;24(2):129-157.
11. Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. In: Westerståhl D, editor. *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain, 2003*: Elsevier-North-Holland; 2004. p. (to appear).
12. Michael J, Mejino JL, Jr., Rosse C. The role of definitions in biomedical concept representation. *Proc AMIA Symp* 2001:463-7.
13. Bernauer J. Subsumption principles underlying medical concept systems and their formal reconstruction. *Proc Annu Symp Comput Appl Med Care* 1994:140-4.
14. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *J Am Med Inform Assoc* 1998;5(1):41-51.
15. Rector A. Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. *Pac Symp Biocomput* 2004:226-237.
16. Guarino N. Some ontological principles for designing upper level lexical resources. In: Rubio A, Gallardo N, Castro R, Tejada A, editors. *Proceedings of First International Conference on Language Resources and Evaluation. ELRA - European Language Resources Association, Granada, Spain*; 1998. p. 527-534.
17. Bittner T, Smith B. A theory of Granular Partitions. In: Duckham M, Goodchild MF, Worboys MF, editors. *Foundations of Geographic Information Science*. London: Taylor & Francis; 2003. p. 117-151.