Artificial Intelligence in Medicine (2007) **39**, 183—195

ARTIFICIAL
INTELLIGENCE
IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies

Olivier Bodenreider[a,*], **Barry Smith**[b,c], **Anand Kumar**[b], **Anita Burgun**[d]

[a] *U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
[b] *Institute for Formal Ontology and Medical Information Science, Saarland University, Germany*
[c] *Department of Philosophy, University at Buffalo, New York, USA*
[d] *EA 3888 Laboratoire d'Informatique Médicale, Université de Rennes I, France*

Received 12 October 2004; received in revised form 10 December 2006; accepted 11 December 2006

KEYWORDS
Biomedical ontologies;
SNOMED CT;
Description logics;
Ontological analysis

Summary

Objective: Formalisms based on one or other flavor of description logic (DL) are sometimes put forward as helping to ensure that terminologies and controlled vocabularies comply with sound ontological principles. The objective of this paper is to study the degree to which one DL-based biomedical terminology (SNOMED CT) does indeed comply with such principles.
Materials and methods: 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.
Results: Our major results are 31% of these 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.
Conclusions: The applications of this principles to quality assurance for ontologies are discussed and suggestions are made for dealing with the phenomenon of multiple inheritance. The advantages and limitations of our approach are also discussed.

1. Introduction

Biomedical terminologies and ontologies are increasingly taking advantage of description logic (DL)-based formalisms in representing knowledge. GALEN[1] and SNOMED Clinical Terms® (in what follows SNOMED CT)[2] were both developed in a native DL formalism. Several other groups have worked at converting existing terminologies into terminologies with a DL

* Corresponding author. Tel.: +1 301 435 3246;
fax: +1 301 480 3035.
E-mail address: olivier@nlm.nih.gov (O. Bodenreider).

[1] http://www.opengalen.org/ (accessed: 10 December 2006).
[2] http://www.snomed.org/ (accessed: 10 December 2006).

0933-3657/$ — see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2006.12.003

formalism, including the UMLS® Metathesaurus® [1–3] and Semantic Network [4], the Medical Subject Headings (MeSH) [5], the Gene Ontology™ [6] and the National Cancer Institute Thesaurus [7]. The Ontology Web Language (OWL) plug-in developed for the ontology editor Protégé 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 serves to ensure 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 [8,9]. However, neither a DL formalism nor the use of a classifier can ensure compliance with all principles of a sound ontology [10].

The objective of this paper is to study the degree to which one DL-based biomedical terminology complies with a basic set of ontological principles. We selected SNOMED CT 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 SNOMED CT is now available as part of the UMLS[3] at no charge for UMLS licensees in the U.S. It is therefore likely to become widely used in medical information systems.

The 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 SNOMED CT, we present the methods used to test the compliance of SNOMED CTwith these principles, and we summarize our results. Finally, we discuss the application of this method to quality assurance in ontologies and terminologies in general, laying special emphasis on the role of creating partitions in ontologies. The advantages and limitations of our approach are also discussed.

2. Background

2.1. Terms, classes, and instances

We shall refer to the nodes in SNOMED CT 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 biological 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 reports of biomedical experiments (e.g., in the form of microarray data), while biomedical terminologies and ontologies are focused on what is general, on classes and their relations.

2.2. Relations among classes

The possible relations of class *A* to class *B* which are relevant to our purposes here 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.

2.3. Principles of classification

Scientific classification has evolved from Aristotle to Linnaeus to the 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 [11]. Some other highly general organization and classification principles — which we believe rest on a wide consensus among those working on terminologies in biomedicine and elsewhere [12] — are:

- Each hierarchy must have a single root.
- Each class (except for the root) must have at least one parent.

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

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	*A* and *B* are classes and all instances of *A* are instances of *B*
A is a child of *B*	*A IS A B*; $A \neq B$; and if *A IS A C* and *C IS A B* then *A* = *C* or *C* = *B*
A and *B* are siblings	There is some *C* of which *A* and *B* are both children and $A \neq B$
A is a parent of *B*	*B* is a child of *A*
C is a differentia of *A* with respect to *B*	*A IS A B*; $A \neq B$; and instances of *A* are marked out within the wider class *B* by the fact that they exemplify *C*

[3] http://umlsks.nlm.nih.gov/ (accessed: 10 December 2006).

- 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.

2.4. Principles of subsumption

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, and assuming that *A* and *B* are distinct, we have the principle:

- No cycles are allowed in an *IS A* hierarchy.

Additionally, one inheritance principle based on Bernauer's 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.

  This principle can also be formulated from the perspective of the child as follows:

- 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.

2.5. Single versus multiple inheritance

Some of the principles presented above enjoy a large degree of 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 versus multiple inheritance, i.e., of whether classes should be allowed to have more than one parent. As noted by Cimino [14]: "There seems to be almost universal agreement that controlled medical vocabularies should have hierarchical arrangements. [. . .] 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.

3. 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 (31 January 2004) contains 269,864 classes,[5] named by 407,510 names.[6] The first level is subdivided into 18 classes listed in Table 2 with their frequency distribution.

Each SNOMED CT class has a description[7] 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.

[4] Also called role filler in DL parlance.

[5] SNOMED CT has a total of 357,135 classes of which 269,864 are "current".

[6] Among the 957,349 names in SNOMED CT, 407,510 correspond to the 269,864 "current" classes, excluding fully specified names and keeping only names whose status is "current".

[7] 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.

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

Top-level class	Frequency
Attribute	991
Body structure	30,652
Clinical finding	95,605
Context-dependent categories	3,649
Environments and geographical locations	1,620
Events	87
Observable entity	7,274
Organism	25,026
Pharmaceutical/biologic product	16,867
Physical force	199
Physical object	4,201
Procedure	46,066
Qualifier value	8,134
Social context	4,896
Special concept	178
Specimen	1,053
Staging and scales	1,098
Substance	22,267

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 SNOMED CT calls the ''primary hierarchy'' of the class, the latter corresponding roughly to one of the top-level classes in the hierarchy. The list and frequency distribution of the primary hierarchies in SNOMED CT are presented in Table 4, along with their corresponding top-level classes. For example, the fully specified name for *Viral meningitis* is *Viral meningitis* (*disorder*).[8] This assignment to a primary hierarchy is not explicitly recognized as a property of the class in the SNOMED CT 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 SNOMED CT classes. Thus for example we use *disorder* as the category for *Viral meningitis*.

Inheritance in SNOMED CT 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 specialized value for the role *ASSOCIATED MORPHOLOGY* in the child (*Fracture*, *open*[9]) 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.

Table 3 Roles present in the description of *Viral meningitis*

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

4. Methods

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

4.1. Quantitative analysis: number of children, parents 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 18 top-level classes is tested by traversing the graph of all classes in SNOMED CT from each class upwards. We use this method for verifying **P4**. As illustrated in Fig. 1, the top-level class subsuming *Viral meningitis* is *Clinical finding*.

4.2. Qualitative analysis of differentiae

In order to verify SNOMED CT'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**.

The relationship between the values of a role shared by the parent and child classes is examined and, when the values differ, is expected to be either specialization (*IS A*) or partitive refinement (*PART*

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

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

Table 4 The list of high-level categories ("primary hierarchies") in SNOMED CTwith their frequency distribution and corresponding top-level class

Category	Frequency	Corresponding top-level class
Administrative concept	54	Qualifier value
Assessment scale	870	Staging and scales
Attribute	991	Attribute
Body structure	25,395	Body structure
Cell	603	Body structure
Cell structure	501	Body structure
Context-dependent category	3,649	Context-dependent categories
Disorder	62,301	Clinical finding
Environment	1,007	Environments and geographical locations
Environment/location	1	Environments and geographical locations
Ethnic group	254	Social context
Event	87	Events
Finding	33,304	Clinical finding
Geographic location	612	Environments and geographical locations
Inactive concept	7	Special concept
Life style	21	Social context
Morphologic abnormality	4,153	Body structure
Namespace concept	5	Special concept
Navigational concept	165	Special concept
Observable entity	7,274	Observable entity
Occupation	4,153	Social context
Organism	25,026	Organism
Person	302	Social context
Physical force	199	Physical force
Physical object	4,201	Physical object
Procedure	42,782	Procedure
Product	16,867	Pharmaceutical/biologic product
Qualifier value	8,080	Qualifier value
Regime/therapy	3,284	Procedure
Religion/philosophy	145	Social context
Social concept	21	Social context
Special concept	1	Special concept
Specimen	1,053	Specimen
Staging scale	15	Staging and scales
Substance	22,267	Substance
Tumor staging	213	Staging and scales

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**.

5. Results

5.1. Quantitative analysis: number of children, parents and roots

5.1.1. 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 Fig. 2. About 196,237 classes (73%) have no children. These classes are leaf nodes in the SNOMED CT 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*.

Table 5 Ontological principles studied in SNOMED CT

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

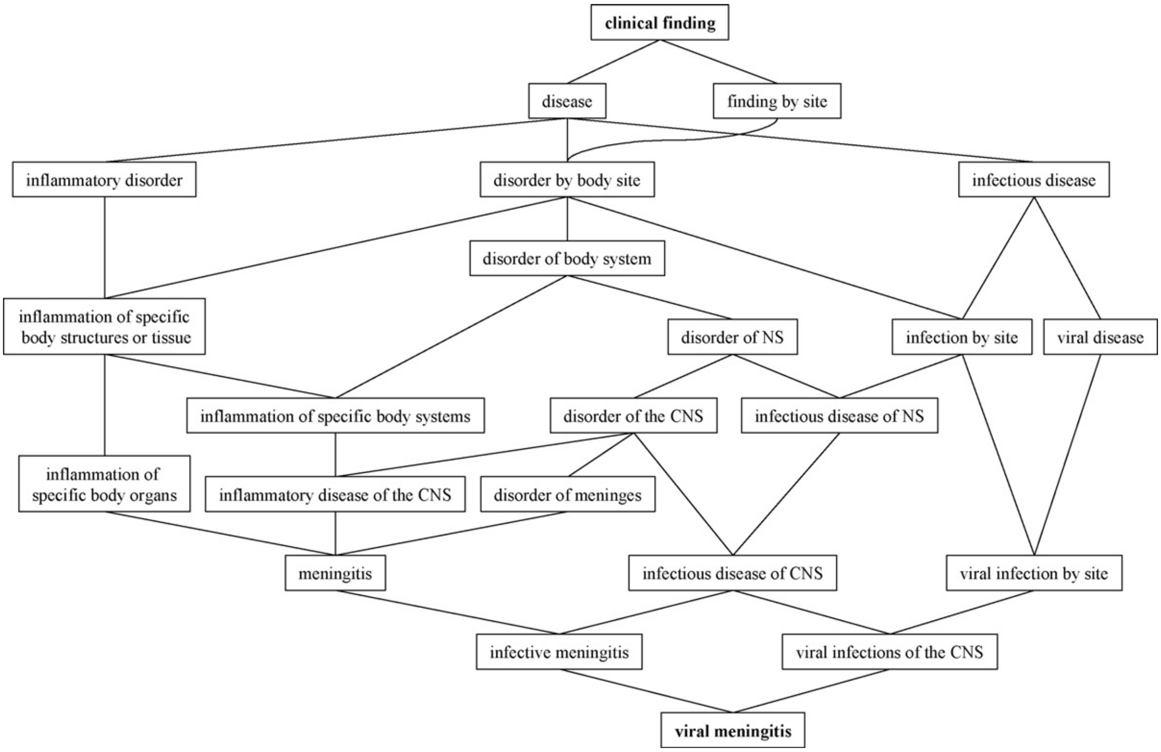

Figure 1 Ancestors of *Viral meningitis* in SNOMED CT.

Out of 73,627 classes with children, 23,174 classes (31.5%) have a single child. As shown in Table 6, this proportion is relatively constant across SNOMED CT 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*).

Eight thousand and thirty-four classes (11%) have 10 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).

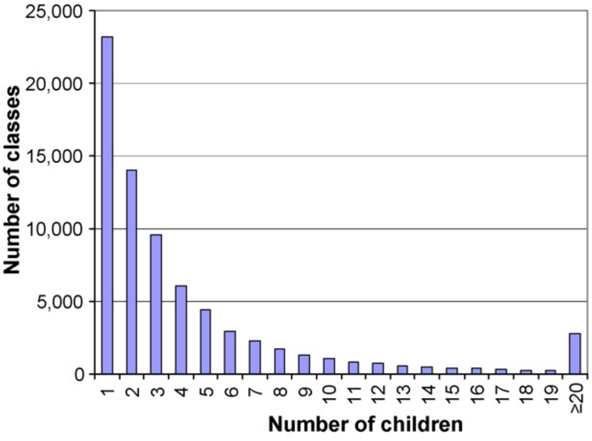

Figure 2 Distribution of the number of children.

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.

5.1.2. Number of parents

Except for the root, every class of SNOMED CT has at least one parent. The number of parents per class ranges from 1 to 13. (The three classes with 13 parents are *Anoscopy with coagulation for control of hemorrhage of mucosal lesion*, *Mandibuloacral dysostosis*, and *Entire sternocleidomastoid muscle*.) The frequency distribution of the number of parents is presented in Fig. 3. About 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.

Table 6 Distribution of the number of children and parents per class (Med: median, Max: maximum, % Mul: proportion of classes with multiple children/parents) and of the presence of differentiae between parents and children (proportion of parent—child pairs with no differentia [None], a single differentia [Single] and multiple differentiae [Mult.])

Category	Children			Parents			Differentiae (%)		
	Med	Max	% Mul	Med	Max	% Mul	None	Single	Mult.
Administrative concept	2	13	57.1	1	1	0.0	100.0	0.0	0.0
Assessment scale	2	724	55.0	1	1	0.0	100.0	0.0	0.0
Attribute	3	142	69.7	1	2	1.2	100.0	0.0	0.0
Body structure	2	295	53.9	1	13	45.5	46.3	29.8	23.9
Cell	3	206	75.0	1	3	16.7	71.4	21.8	6.8
Cell structure	2	98	76.1	1	4	27.5	52.8	40.8	6.4
Context-dependent category	3	150	78.7	1	2	0.1	60.9	38.6	0.5
Disorder	3	505	72.9	1	13	45.9	24.3	43.3	32.4
Environment	3	39	79.1	1	2	0.6	100.0	0.0	0.0
Environment/location	2	2	100.0	1	1	0.0	100.0	0.0	0.0
Ethnic group	3	54	84.6	1	2	1.6	100.0	0.0	0.0
Event	3	17	81.0	1	2	1.1	100.0	0.0	0.0
Finding	3	251	78.1	1	5	15.2	67.9	23.1	9.0
Geographic location	5	46	94.6	1	3	2.3	100.0	0.0	0.0
Inactive concept	6	6	100.0	1	1	0.0	100.0	0.0	0.0
Life style	3.5	6	83.3	1	1	0.0	100.0	0.0	0.0
Morphologic abnormality	3	410	70.4	1	4	30.2	99.3	0.5	0.2
Namespace concept	4	4	100.0	1	1	0.0	100.0	0.0	0.0
Navigational concept	164	164	100.0	1	1	0.0	100.0	0.0	0.0
Observable entity	2	77	73.8	1	3	4.9	99.8	0.2	0.0
Occupation	3	34	81.1	1	3	15.7	100.0	0.0	0.0
Organism	2	551	64.5	1	4	4.9	100.0	0.0	0.0
Person	2	23	83.8	1	2	23.2	100.0	0.0	0.0
Physical force	2	21	66.7	1	2	6.5	100.0	0.0	0.0
Physical object	2	118	74.3	1	4	7.0	100.0	0.0	0.0
Procedure	2	996	67.7	1	13	45.6	22.6	34.9	42.5
Product	2	2532	69.2	1	4	7.6	65.4	30.8	3.8
Qualifier value	3	359	79.6	1	3	6.9	100.0	0.0	0.0
Regime/therapy	2	51	69.1	1	7	26.0	60.9	23.6	15.6
Religion/philosophy	2	29	74.1	1	2	1.4	100.0	0.0	0.0
Social concept	2	10	71.4	1	1	0.0	100.0	0.0	0.0
Special concept	3	3	100.0	1	1	0.0	100.0	0.0	0.0
Specimen	2	82	70.3	1	4	17.2	13.8	68.0	18.1
Staging scale	6	6	100.0	1	1	0.0	100.0	0.0	0.0
Substance	2	763	64.8	1	6	13.8	100.0	0.0	0.0
Tumor staging	3	23	91.7	1	2	0.5	100.0	0.0	0.0
Total	2	2532	68.5	1	13	27.7	51.4	27.1	21.5

Overall, the proportion of classes having multiple parents, i.e., exhibiting multiple inheritance, is 27.7%. As shown in Table 6, 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—17% for *cell*, *organism*, and *substance*).

5.1.3. Number of roots

Except for the root and for the 18 top-level classes themselves, each class of SNOMED CT can be linked hierarchically to exactly one top-level class. This means that SNOMED CT consists of 18 independent hierarchies.

5.2. Qualitative analysis of differentiae

5.2.1. 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, as shown in Table 6, 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

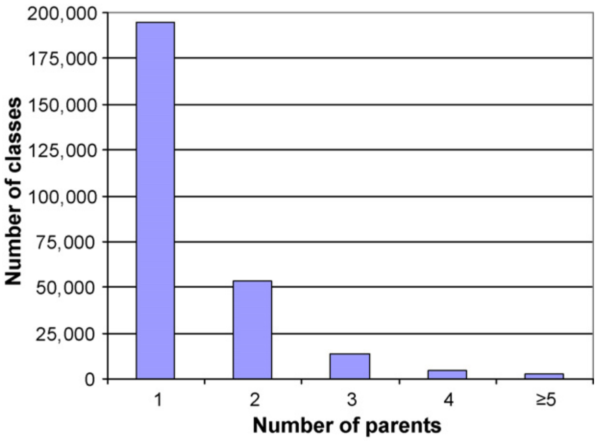

Figure 3 Distribution of the number of parents.

description ranges from 29% (*cell*) to 86% (*specimen*).

5.2.2. 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.

5.2.2.1. 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.*

5.2.2.2. 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, as illustrated in Fig. 4, *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*). Multiple differentiae are often associated with multiple inheritance. In the example above, the role *METHOD* is actually inherited from *Gastrointestinal investigation*, the second parent of *Endoscopy of jejunum*, and its value refined from *Evaluation—action* to *Inspection—action*. The role *ACCESS INSTRUMENT,* however, is truly specific to *Endoscopy of jejunum* (i.e., not present in any of its parents).

5.2.2.3. 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* — called a qualifier in SNOMED CT — is

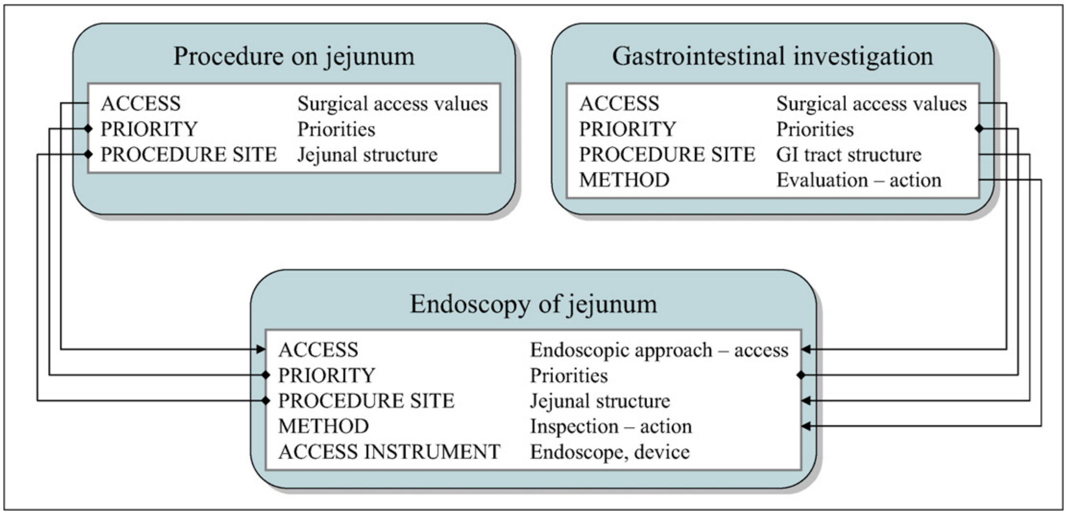
Procedure on jejunum
ACCESS Surgical access values
PRIORITY Priorities
PROCEDURE SITE Jejunal structure
Gastrointestinal investigation
ACCESS Surgical access values
PRIORITY Priorities
PROCEDURE SITE GI tract structure
METHOD Evaluation – action
Endoscopy of jejunum
ACCESS Endoscopic approach – access
PRIORITY Priorities
PROCEDURE SITE Jejunal structure
METHOD Inspection – action
ACCESS INSTRUMENT Endoscope, device

Figure 4 Inheritance of role values for *Endoscopy of jejunum*.

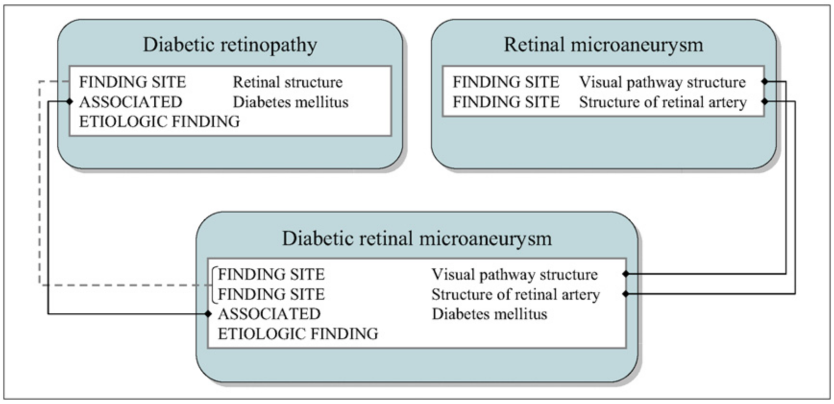
Diabetic retinopathy
FINDING SITE Retinal structure
ASSOCIATED ETIOLOGIC FINDING Diabetes mellitus
Retinal microaneurysm
FINDING SITE Visual pathway structure
FINDING SITE Structure of retinal artery
Diabetic retinal microaneurysm
FINDING SITE Visual pathway structure
FINDING SITE Structure of retinal artery
ASSOCIATED ETIOLOGIC FINDING Diabetes mellitus

Figure 5 Inheritance of role values for *Diabetic retinal microaneurysm* (partial representation).

involved in roughly half of the cases where some role is specific to a parent class but 11 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, as illustrated in Fig. 5, the class *Diabetic retinopathy* and its child *Diabetic retinal microaneurysm* share the role *FINDING SITE*, but their values for this role (*Retinal structure* for the parent and *Visual pathway structure* and *Structure of retinal artery* for the child) 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* and *Structure of retinal artery* are actually inherited from *Retinal microaneurysm*, the other parent of *Diabetic retinal microaneurysm*.

6. Discussion

The work described in this paper is in the tradition of studies auditing large medical terminologies such as [15]. SNOMED CT itself has recently been investigated for inconsistencies and related types of errors [10,16]. However, we are interested here not in errors and inconsistencies in general but rather, more positively, in the question of compliance of the terminological structure with general classification principles. We found SNOMED CT 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, and we will present the consequences of such non-compliance together with a discussion of the advantages and limitations of our approach. Finally, we will revisit the problem of single versus multiple inheritance and outline a possible solution thereto.

6.1. Application to quality assurance for ontologies

6.1.1. 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 classes with just one child is reason to suspect the presence of error.

The review of a limited number of such classes suggests the following possible issues. One is the incompleteness of the hierarchy (e.g., *Subphylum Vertebrata* is the only subphylum recorded in SNOMED CT for *Phylum Chordata*). Another issue is the presence of hybrid classes, resulting from the intersection of two parent classes and appearing as the single child of at least one of these (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 the phenomenon 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 *Periappendicitis* (single child of *Atypical appendicitis*), and the substance *Urine* (single child of *Urinary tract fluid*).

Except when they are the product of hybrid classes, classes with a single child should be reviewed. For example, the classes *Congenital absence of lobe of liver* and its parent *Congenital absence of liver* do not look suspicious at first sight. However, knowing that *Congenital absence of lobe of liver* is the single child of *Congenital absence of liver* raises the question of a possible confusion between a total absence of the liver and an absence of liver whose degree on the partial/total axis is not specified. If *Congenital absence of liver* is treated as a total absence of liver (hypothesis 1), it cannot subsume the absence of a lobe of liver (partial absence). Therefore the subsumption link is inaccurate. Conversely, if *Congenital absence of liver* is treated as unspecified absence of liver (hypothesis 2), the degree of the absence — total or partial — is expected to be reflected in its children, and having only one child makes the description incomplete. In this particular case, SNOMED CT lists *Congenital absence of liver, total* as a synonym for *Congenital absence of liver* (hypothesis 1). Therefore, *Congenital absence of liver* cannot subsume *Congenital absence of lobe of liver.*

6.1.2. 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 detailed descriptions for each class available for use by formal reasoning tools, representing through roles the class' defining characteristics. 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 is the incompleteness 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 there exists a class *Courses* whose children include *Acute* and *Subacute*). 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). Such extensional definitions are less desirable for a number of reasons, including: (1) they imply an unsatisfactory heterogeneity in the classification; (2) they imply missing information, which is not available, e.g., for automatic information extraction and which also implies obstacles to correct coding (*why* are these subclasses grouped together in this way); (3) they imply the need for revisions with each discovery of new types of cases.

Finally, in some cases, there is actually no difference between the parent and the child class (e.g., *Closed fracture of skull without intracranial injury* versus *Closed fracture of skull*). The issue, in this case, is the presence of two terms naming two distinct classes in SNOMED CT for one and the same entity in reality. The distinction lies not on the side of the biomedical entities these terms represent (i.e., the skull is fractured, but not open), but rather merely in the associated knowledge on the part 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 [17]. It would be a valuable extension of the current DL in SNOMED CT if ways could be found to do justice to operators, such as 'with' and 'without,' which are characteristic of such epistemologically motivated admixtures and which play an important role in the organization of SNOMED CT'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 that in this respect, too, much of the information conveyed by the compositional structure of SNOMED CT's terms is at the moment not available for automatic retrieval.

6.1.3. 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 being such that it can have either 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* [18].

6.2. Advantages and limitations

The principles presented in this study are simple. Assessing the degree to which SNOMED CT complies with these principles can be easily implemented. Although a description logic (DL) was used in its development, SNOMED CT is not distributed through the UMLS in a way which would allow users to perform automatic classification by appealing to the DL structure. Instead, SNOMED CT classes appear as regular Metathesaurus concepts. Source transparency in the UMLS allows users to extract SNOMED CT information in the form of triples for relations, e.g., (*Viral meningitis, IS A, Infective meningitis*). Although we investigated a terminology developed in a DL environment, our method did not rely on any DL-specific feature. Therefore, it would be applicable not only to other DL-based terminologies, but also to terminologies whose relations are represented as triples, provided that the description of the classes is sufficiently rich.

Compliance with the seven principles investigated in this study is no guarantee of complete ontological soundness. Non-compliance with the principles should be interpreted rather as indicative of possible problems and so used to trigger the review of the classes and relations involved by the editors of the terminology in the way described in Ref. [19].

In some cases, there is an indication of an error that is as best tenuous, e.g., when a relation is in compliance with one principle, but violates another principle. In the example presented earlier in the discussion, except for the fact that *Congenital absence of lobe of liver* is the single child of *Congenital absence of liver*, our method provides no indication that the latter represents a total absence and can therefore not subsume the former, which represents a partial absence of the liver. The values for the roles *ASSOCIATED MORPHOLOGY* and *FINDING SITE* in *Congenital absence of lobe of liver* do refine that of the corresponding roles in *Congenital absence of liver*. The only indication of a possible problem is given by the fact that *Congenital absence of lobe of liver* is the single child of *Congenital absence of liver*. Similarly, the existence of multiple differentiae between *Endoscopy of jejunum* and *Gastrointestinal investigation* (Fig. 4) — namely the refinement of both *ACCESS* and *PROCEDURE SITE* roles — should raise the possibility of a missing intermediary class or a missing subsumption link. For example, although the duodenum and the jejunum are adjacent segments of the small intestine, *Duodenoscopy* is linked to *Gastrointestinal investigation* through three intermediary classes (*Enteroscopy, Endoscopy of intestine*, *Gastrointestinal tract endoscopy*), while the link is direct for *Endoscopy of jejunum*. A careful review of these classes and their relations is required to identify issues such as inaccurate subsumption links and missing intermediary classes. In the two examples above, the review could have been prompted by failure to comply with the principle that no class should have a single child or because of the presence of several differentiae between a parent and its child.

Conversely, some of our principles may be too strict and may benefit from relaxation in some circumstances. More precisely, they may be refined in order to exploit implicit information. The principle of single differentia between a child and its parent, for example, rests on the assumption that roles are independent, which is not always the case. Although not explicitly related, the roles *ACCESS* (*Endoscopic approach—access*) and *ACCESS INSTRUMENT* (*Endoscope, device*) are indeed not independent. This explains in part why, as illustrated in Fig. 4, there are several differentiae related to endoscope between *Endoscopy of jejunum* and *Gastrointestinal investigation*: the introduction of *ACCESS INSTRUMENT* with value *Endoscope, device* accompanies the refinement of the value of *ACCESS* from *Surgical access values* to *Endoscopic approach—access*.

6.3. 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* overloading', which is often associated in turn with examples of incorrect subsumption [20]. Examples of this phenomenon in SNOMED CT include *Both testes IS A Testis Structure*, *Deferoxamine mesylate IS A Deferoxamine*, and *Urine sediment IS A Urine*.

IS A overloading, 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} and 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 SNOMED CT to be divided into

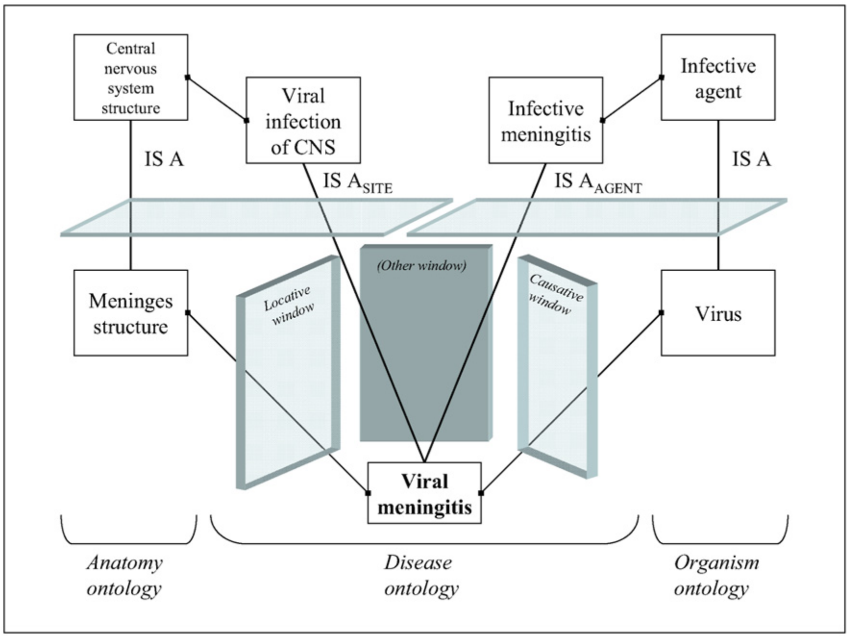

Figure 6 Two views (locative and causative) on *Viral meningitis*.

partitions within and between which taxonomic reasoning can be more reliably performed [8].

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 Fig. 6, and the underlying formal theory is presented in Ref. [21].

7. Conclusions

SNOMED CT is the most comprehensive biomedical terminology recently developed in native DL formalism and it is expected to play an important role in clinical information systems in the future. Unlike thesauri built for information retrieval purposes, SNOMED CT should enable reasoning about biomedical classes and relations of a sort which can support intelligent information retrieval of biomedical information. We have listed some principles, mostly related to classification, and tested the degree to which SNOMED CT complies therewith. While SNOMED CT appears 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 for inheritance.

Description logics provide formalisms suitable for representing many features of a variety of different domains — including the biomedical domain — in ways 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.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). Smith and Kumar are supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation, by the EU FP6 Network of Excellence "Semantic Datamining" and by the Volkswagen Foundation project "Forms of Life".

References

[1] Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS methathesaurus. In: Chute CG, editor. Proceedings of the AMIA symposium. 1998. p. 810—4.

[2] Cornet R, Abu-Hanna A. Usability of expressive description logics-a case study in UMLS. In: Kohane IS, editor. Proceedings of the AMIA symposium. 2002. p. 180—4.

[3] Hahn U, Schulz S. Towards a broad-coverage biomedical ontology based on description logics. In: Altman RB, Dunker AK, Hunter L, Jung TA, Klein TE, editors. Pac Symp Biocomput. 2003.p. 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, editor. The SemanticWeb—ISWC 2003, vol. LNCS 2870. Heidelberg: Springer-Verlag; 2003. p. 1—16.

[5] Soualmia L, Golbreich C, Darmoni S. Representing the MeSH in OWL: towards a semi-automatic migration. In: Hahn U, editor. Proceedings of the KR 2004 workshop on formal biomedical knowledge representation (KR-MED 2004), vol. 102 (CEUR-WS, 2004). p. 81—87 http://ceur-ws.org/Vol-102/soualmia.pdf (accessed: December 10, 2006).

[6] Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML + OIL. In: Altman RB, Dunker AK, Hunter L, Jung TA, Klein TE, editors. Pac Symp Biocomput. 2003.p. 624—35.

[7] Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. J Web Seman 2003;1 http://www.websemantics-journal.org/volume1/issue1/Golbecketal2003/index.html.

[8] 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, vol. 4 (CEUR-WS, 1996), p. 24—8. http://ceur-ws.org/Vol-4/ (accessed: December 10, 2006).

[9] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16:184—5.

[10] 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).

[11] Marradi A. Classification, typology, taxonomy. Qual Quantity 1990;24:129—57.

[12] Michael J, Mejino Jr JL, Rosse C. The role of definitions in biomedical concept representation. In: Proceedings of the AMIA symposium; 2001. p. 463—7.

[13] Bernauer J. Subsumption principles underlying medical concept systems and their formal reconstruction. In: Proc Annu Symp Comput Appl Med Care. 1994. p. 140—4.

[14] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37:394—403.

[15] Cimino JJ. Auditing the unified medical language system with semantic methods. J Am Med Inform Assoc 1998;5: 41—51.

[16] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. In: Fieschi M, Coiera E, Li Y-C, editors. Proceedings of MEDINFO 2004. 2004. p. 482—6.

[17] Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: a case study in medical terminology. In: Varzi AC, Vieu L, editors. Proceedings of the third international conference on formal ontology in information systems (FOIS 2004), vol. 114. 2004. p. 185—95.

[18] Rector A. Defaults, context, and knowledge: alternatives for OWL-indexed knowledge bases. In: Altman RB, Dunker AK, Hunter L, Jung TA, Klein TE, editors. Pac Symp Biocomput. 2004.p. 226—37.

[19] dos Santos MC, Dhaen C, Fielding M, Ceusters W. Philosophical scrutiny for run-time support of application ontology development. In: Varzi AC, Vieu L, editors. Proceedings of the 3rd International Conference on Formal Ontology in Information Systems (FOIS 2004), vol. 114 (IOS Press, 2004), p. 342—52.

[20] 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. 1998. p. 527—34.

[21] 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—51.