Aspects of the Taxonomic Relation in the Biomedical Domain

Anita Burgun, Olivier Bodenreider
U.S. National Library of Medicine
8600 Rockville Pike Bethesda, MD 20814, USA
{burgun, olivier}@nlm.nih.gov

Abstract — Taxonomies are commonly used for organizing knowledge, particularly in biomedicine where the taxonomy of living organisms and the classification of diseases are central to the domain. The principles used to produce taxonomies are either intrinsic (properties of the partial ordering relation) or added to make knowledge more manageable (opposition of siblings and economy). The applicability of these principles in the biomedical domain is presented using the Unified Medical Language System (UMLS) and issues raised by the application of these principles are illustrated. While intrinsic principles are not challenged, we argue that the opposition of siblings brings to bear excessive constraints on a domain ontology and that the adverse effects of economy may outweigh its benefits. The two-level structure used in the UMLS is discussed.

Categories & Descriptors — I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods – *Relation Systems.* J.3 [Computer Applications]: Life and Medical Sciences.

General Terms — Theory.

Keywords — Taxonomic relation, Ontology, Biomedical domain, Unified Medical Language System.

1 Introduction

Taxonomies are useful artifacts for organizing many aspects of knowledge, much of which can be expressed mathematically with partial orders. Taxonomies are used for representing information at appropriate levels of generality and automatically making it available to more specific concepts by means of a mechanism of inheritance [18]. As components of ontologies, taxonomies can provide an organizational model of a domain (domain ontologies), or a model suitable for specific tasks (application ontologies).

The principles used to produce taxonomies are either intrinsic (properties of the partial ordering relation) or added to make knowledge more manageable (opposition of siblings and economy). In biomedicine, taxonomies such as the taxonomy of living organisms and the classification of diseases are central to the domain. However, the applicability of these principles in the biomedical domain needs to be assessed.

This study is a contribution to the Medical Ontology Research project currently developed at the U.S. National Library of Medicine [2]. The major objective of this project is to develop methods

FOIS'01, October 17-19, 2001, Ogunquit, Maine, USA.

In: Welty C, Smith B, editors. Collected papers from the Second International Conference "Formal Ontology in Information Systems": ACM Press; 2001. p. 222-233. © 2001 Association for Computing Machinery

whereby biomedical ontologies could be acquired from existing resources, as well as validated against other knowledge sources. The objective of this paper is to study how principles derived from the theory of hierarchies fit the biomedical domain. Understanding the taxonomic relation in the biomedical domain can be seen as an initial step towards acquiring and validating biomedical ontologies. As a source of biomedical knowledge, we will use in particular the Unified Medical Language System® (UMLS®), developed and maintained by the U.S. National Library of Medicine since 1990.

2 Background

In this section, we briefly present the principles underlying the production of taxonomies, particularly in the biomedical domain. An overview of the UMLS follows.

2.1 Taxonomic relation

The ability of systems to reason from taxonomies depends on the definition, identification and organization of taxonomic information [6]. Taxonomies can be examined from three different perspectives: structurally, ontologically and semantically.

From a structural perspective, the way knowledge is represented is not always formal enough for computers to reason from it. Additional structural constraints have been suggested in order to make taxonomies more usable in application contexts. One such constraint is that two sibling categories be incompatible. For example, the concepts "physical state" and "mental state" are children of "state" and incompatible. Both concepts are incompatible in the ontology because the former involves a physical object whereas the latter involves a mental object [3].

From the perspective of formal ontology, Guarino gives several examples of isa overloading. For example, "a physical object is an amount of matter" reflects a reduction of sense, since a physical object is more than just an amount of matter [10]. Guarino & Welty focus on meta-properties that help formalize constraints on the taxonomic relation. For example, "group of people" carries the meta-property +ME, which means that such entities have as a necessary identity condition that the parts of their instances must be the same. According to the rule *+ME cannot subsume properties with -ME*, "group of people" cannot subsume "organization", which is –ME, since people in organizations change [11].

From the standpoint of semantics, Brachman describes several meanings of the *isa* relation that may exist between two generic concepts in semantic networks (subset / superset, generalization / specialization, kind-of, conceptual containment, role value restriction, set / prototype) [4]. He also suggests using those semantic subcomponents as the primitives of a representation system.

In practice, taxonomic knowledge is complex and remains partially intuitive in many existing ontologies. This may lead to ruptures in knowledge representation, and thus impair the capability of reasoning from the system. For example, according to the taxonomic relationships linking the hypernyms of "fever" in WordNet® (1.6), "fever" ends up being categorized as a Psychological Feature [7] (Figure 1).

2.2 Taxonomies in biomedicine

Taxonomies are ubiquitous in biomedicine. A typical example is the taxonomy of living beings. Taxonomies have also been developed for decades in order to organize biomedical subdomains, where categories may be fuzzier than those referred to as natural kinds. The hierarchical relations implemented in medical classifications may be pragmatically driven. Since it has been established that the characteristics of living organisms are coded for in genes, differences in their genetic code become the means for organizing living beings in a taxonomy. Before it was possible to rely on genotypic characteristics, the creation of taxonomies used to rely upon phenotypic characteristics, i.e. external features. Part of the classification of micro-organisms is still based on external features, such as the shape of bacteria – cocci are spherical, bacilli are rod-shaped bacteria – and whether they are stained by

standard techniques or not, e.g., the Gram technique. This leads to four categories: cocci Gram positive, cocci Gram negative, bacilli Gram positive, and bacilli Gram negative. For example, bacteria of the genus Salmonella are Gram negative bacilli. In fact, this classification of micro-organisms was meant for identification purposes, leading to clusters sharing external properties, rather than for organizing micro-organisms in a taxonomy of categories reflecting their essential properties. Moreover, some classifications are driven by specific objectives that may influence their design. For example, the International Classification of Diseases, by design, provides a limited number of slots (terms for diseases), suitable for general purposes (e.g., epidemiology, evaluation of health care outcomes).

Fever
- Symptom
- • Evidence
- • • Information
- • • • Cognition / Knowledge
- • • • • Psychological feature

Figure 1 – Hypernyms of "fever" in WordNet

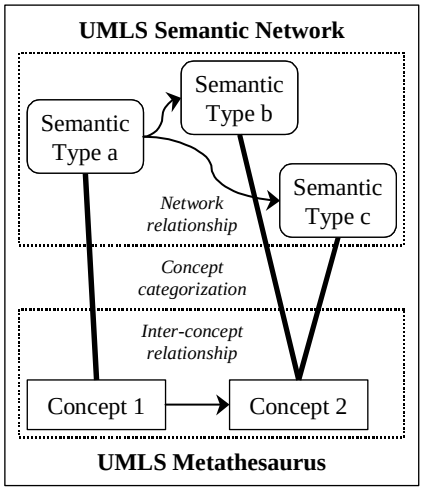

Figure 2 – A two-level structure

2.3 The Unified Medical Language System

The Unified Medical Language System (UMLS) is intended to help health professionals and researchers use biomedical information from different sources. The UMLS[1] comprises two major inter-related components: the Metathesaurus®, a huge repository of concepts, and the Semantic Network, a limited network of semantic types. The current version (2001) of the Metathesaurus integrates about 800,000 concepts from more than fifty families of vocabularies such as the International Classification of Diseases and Medical Subject Headings. While the structure of each source vocabulary is preserved, terms that are equivalent in meaning are clustered into a unique concept. Furthermore, interconcept relationships, either inherited from the source vocabularies or specifically generated, give the UMLS Metathesaurus additional semantic structure. The UMLS building process imposes no restrictions on the source vocabularies prior to integrating their terms and structure into the Metathesaurus. Therefore, hierarchical relationships in the Metathesaurus are not expected to represent homogeneous taxonomic relations, but rather to reflect the several organizational principles inherited from the source vocabularies.

The UMLS Semantic Network is a network of 134 semantic types used to categorize Metathesaurus concepts. A definition is given for each semantic type. The semantic types are organized in two high-level single-inheritance hierarchies, one for entities, one for events. The *isa* link allows nodes to inherit properties from higher-level nodes. In addition, associative relationships divided into five subcategories (physical, spatial, functional, temporal,

[1] http://umlsks.nlm.nih.gov/ (checked July 20, 2001)

conceptual relationships) are instantiated between the semantic types. They represent general high-level knowledge, such as “drugs treat diseases”. Conversely, Metathesaurus inter-concept relationships instantiate specific low-level knowledge, such as “aspirin treats fever”. When two semantic types are linked by some relationship, the relationship may hold or not for any particular pair of concepts that have been assigned to those semantic types (obviously, not every drug treats every disease). Each Metathesaurus concept is assigned to at least one semantic type from the Semantic Network, providing each concept a categorization that is independent from its relationships to other concepts (Figure 2).

One major principle used for building the UMLS is economy, i.e. to prevent unneeded categories from being represented. Applied to the construction of the Semantic Network, the Economy Principle resulted in three rules affecting not only the design of the Semantic Network but also the way Metathesaurus concepts are categorized [14]:

R1. *Assign the most specific semantic type available*. The level of granularity varies across the UMLS Semantic Network. The intent is to establish a set of semantic types, which are useful for a variety of tasks without introducing undue complexity. The most specific semantic type in the semantic type hierarchy is assigned to the concept.

R2. *Assign multiple semantic types if necessary*. Instead of creating a lattice structure, with hybrid types inheriting from two supertypes, the Semantic Network has a single inheritance tree structure. As a consequence, a Metathesaurus concept inheriting from two semantic types is assigned to both types.

R3. *Assign a less specific semantic type (supertype) if no more specific semantic type (subtype) is available*. Rather than proliferating the number of semantic types to encompass additional subcategories, concepts that cannot be categorized by any sibling semantic type are simply assigned their common supertype.

Our study investigates how principles derived from the theory of hierarchies are implemented in the UMLS. We explore the following three axes: (1) Categories, also called types, are abstract specifications whose extensions are sets of things, also called classes. In a taxonomy of types, the *isa* relationship between two types entails that the corresponding classes are in a relation of inclusion; (2) Taxonomies are based on the *isa* relation, a partial ordering relation that is reflexive, antisymmetric and transitive; (3) An additional principle commonly suggested is that siblings be organized in a system of oppositions. When principles fail to be applied, or when their application raises issues, we investigate whether discrepancies are related to the principles in their definition or in their implementation, or to the characteristics of the biomedical domain.

3 The principles, their implementation, and the biomedical domain

In this section, we examine the principles mentioned above, and their application to the representation of the biomedical domain.

3.1 Subordination of categories is equivalent to inclusion of classes

By category is meant a type, i.e. an abstraction that applies to objects. By class is meant a set of objects that are considered equivalent and fall under a category. Taxonomies are systems in which categories are related to one another by means of subordination, or, in class parlance, systems in which classes are related to one another by means of class inclusion. When a category K has subcategories K_1, K_2, … K_n, its extension, the class C_K is the union of the classes for each of its subcategories, i.e. C_{K1}, C_{K2}, … C_{Kn}. Applied to the UMLS, the higher-level Semantic Network constitutes a taxonomy of semantic types, in which each semantic type T is a category that subsumes concepts in the lower-level Metathesaurus. The set of Metathesaurus concepts that are assigned to a given semantic type T is the UMLS class C_T. The process of categorization involves UMLS editors, assisted by guidelines included in the Semantic Network (intension), and by reference to other concepts already assigned to a given semantic type (extension). In practice, however, the UMLS classes often contain far too many

concepts for the editors to examine in detail. Under rule R3 of the Economy Principle, and as illustrated in Figure 3, the class MANUFACTURED OBJECT, C_{MO}, i.e. the set of Metathesaurus concepts that are assigned the semantic type Manufactured Object, is the set of manufactured objects that cannot be assigned a subtype of Manufactured Object. Instances of C_{MO} are, for example, "45 inch calibre bullet", "magnetic tape", and "corridor". As a consequence, the class C_{MO}, extension of the category Manufactured Object contains instances that do not belong to the union of the classes for each of its subcategories, i.e. C_{MD} (Medical Device), C_{RD} (Research Device), and C_{CD} (Clinical Drug). Although Medical Device and Research Device may be thought of roles, an equivalent phenomenon would occur even if Device and Drug were the only two subcategories. In the example above, some concepts in C_{MO} (e.g., "corridor") cannot be categorized by any subtypes of Manufactured Object, which could justify the creation of an additional subtype, called, for example, Other Manufactured Objects.

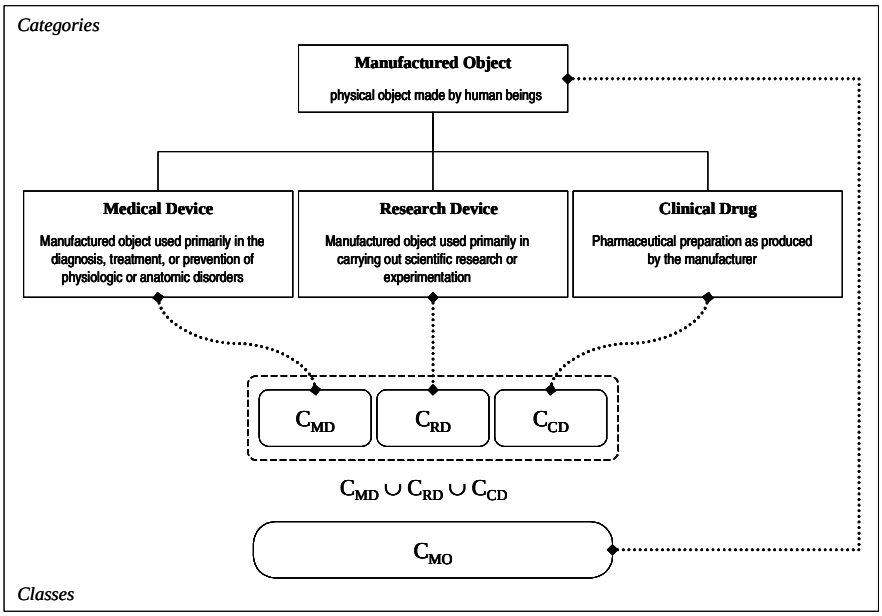

Figure 3 – Categories and classes in the UMLS

A different situation occurs in the Animal category, whose subtypes provide complete coverage of the subdomain. Therefore, the class ANIMAL is expected not to contain concepts other than those corresponding to the union of the classes for each of its subcategories. However, 41 Metathesaurus concepts are assigned the semantic type Animal. Some of them clearly correspond to roles (e.g., "pests", "domestic animals", "livestock"). Other concepts, however, correspond to a dimension orthogonal to that used to create the taxonomy. For example, transgenicity (in "transgenic animal") or gender (in "male animal") correspond to essential properties, not roles. Moreover, not only are these concepts useful and valid, but they also are licitly categorized as Animal, since the categories necessary to represent these properties are not available in the Semantic Network taxonomy.

3.2 Hierarchical relation and partial ordering relation

In this section, we examine the three properties of the partial ordering relation: reflexivity, antisymmetry and transitivity.

3.2.1 Reflexivity

Although no reflexive *isa* relationship is explicitly implemented in the UMLS Semantic Network, the *isa* relation is reflexive, and a reflexive *isa* relationship is needed for semantic processing. For example, in the Metathesaurus, ibuprofen is a non-steroidal anti-inflammatory (NSAI) substance. Both "ibuprofen" and "NSAI" are categorized with the Pharmacologic Substance semantic type. However, there is no *isa* relationship of Pharmacologic Substance to itself represented in the Semantic Network to support the *isa* relationship between the two concepts in the Metathesaurus (Figure 4).

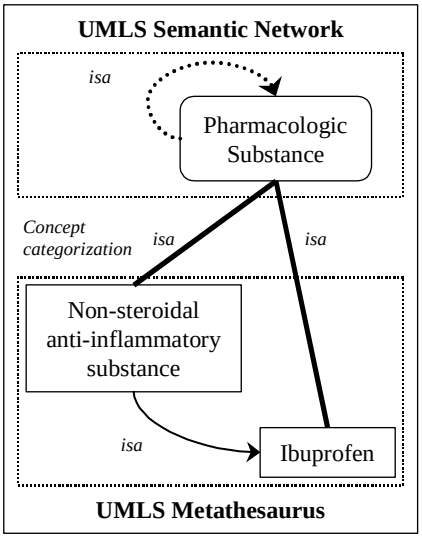

Figure 4 – Implicit reflexive hierarchical relationship in the UMLS Semantic Network

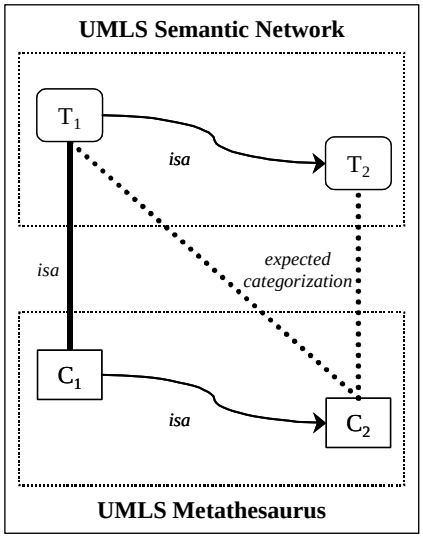

Figure 5 – Transitivity of the isa relation between semantic types and Metathesaurus concepts

3.2.2 Antisymmetry

The antisymmetry property is present throughout the hierarchy of semantic types in the UMLS Semantic Network. In the UMLS Metathesaurus, the combination of hierarchical structures is also expected to result in a directed acyclic graph. Patterns that lead to antisymmetry violations have been studied extensively in [1]. They are mostly related to the fact that, although recorded and used at the concept level, many hierarchical relationships in the Metathesaurus were defined at the term level.

3.2.3 Transitivity

The *isa* relation is found in the UMLS at three different levels: between semantic types in the Semantic Network, between concepts in the Metathesaurus, and between a concept and a semantic type through the categorization. Assuming that this *isa* relation represents the same kind of abstraction at different levels in the UMLS, transitivity is expected to apply not only between semantic types, or between Metathesaurus concepts, but also between semantic types and Metathesaurus concepts. Thus, the semantic type of any ancestor C_1 of a concept C_2 is expected to be a supertype of the semantic type of C_2 (Figure 5).

In practice, however, inconsistencies may be caused by the fact that Metathesaurus concepts are clusters of terms, which makes it difficult to distinguish among generic concepts and prototypical forms. The Metathesaurus provides several examples of confusion between the

generic concept represented by a term X and the typical instance of the class, also referred to with X. This phenomenon is extremely frequent in the biomedical domain, where many qualifiers are implicit in medical terms. For example, "hip dislocation" is represented as a synonym of "acquired hip dislocation", because the most frequent form for hip dislocation is traumatic (i.e., acquired, as opposed to congenital). "acquired hip dislocation", the typical form, is clustered together with "hip dislocation", the generic concept. Therefore, in the Metathesaurus, "congenital hip dislocation" is a child of "[acquired] hip dislocation". The consequences of this phenomenon in terms of categorization are illustrated in Figure 6: "hip dislocation", considered by default an "acquired hip dislocation", is assigned the semantic type Injury or Poisoning; "congenital hip dislocation", although a child of "[acquired] hip dislocation", is assigned the semantic type Congenital abnormality; and Congenital Abnormality is thus expected to be a subtype of Injury or Poisoning, but only non-taxonomic relations are stipulated between these two semantic types (*has-result*, *complicates*).

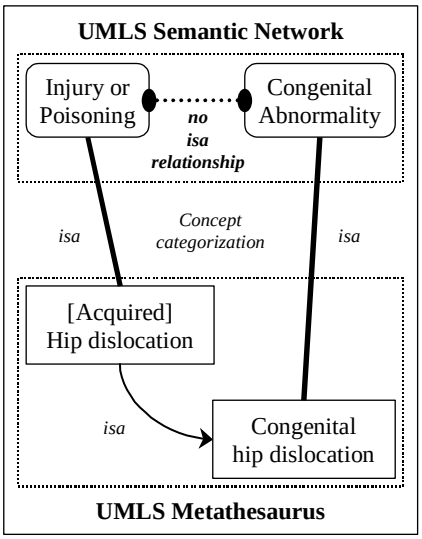

Figure 6 – Generic concept vs. typical form

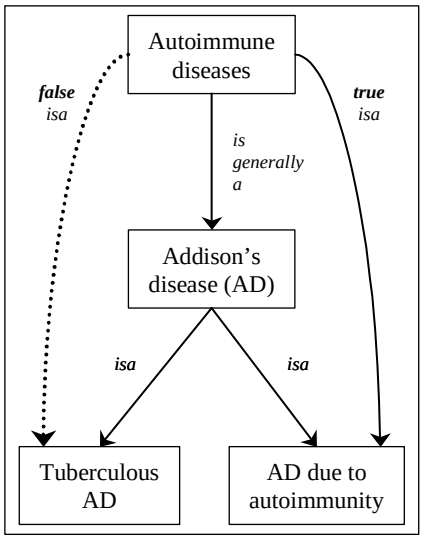

Figure 7 – False predicate produced by using the is-generally-a relation

A somewhat different problem occurs when the taxonomic relation is used to represent empirical knowledge. In such cases, the *isa* relationship may mean *is-generally-a*. For example, according to the Metathesaurus, "Addison's disease isa autoimmune disease", which, nowadays, is true in many cases, but not in all cases. Thus, despite transitivity, even if "tuberculous Addison's disease" is an "Addison's disease", the predicate "tuberculous Addison's disease isa autoimmune disease" is false (Figure 7).

Opposition of siblings

The opposition principle is derived from the representation of a hierarchy as a system of differences. A category is differentiated from its immediate parent and its siblings by some differentia, while all share a common genus. In the resulting tree, siblings are organized in a system of opposition, each child being opposed to the other children of the same type. For example, the first subdivision of the UMLS Semantic Network opposes Physical Object and Conceptual Entity. As we mentioned in the introduction, differentiation between biomedical concepts may be based on external features. And also, the criteria used for identifying differentia cannot always be defined with precision. For example, the differentiation of

elementary skin lesions is based on descriptive, imprecise criteria (Table 1). The representation of macules is based on a prototype, erythematous macula, which blanches when pressed, while purpura does not[2]. However, what is true for the prototype is not true for some other kinds of macules, such as hyperpigmented macules, which do not blanch when pressed.
Differentiation between elementary lesions is based on empirical features that are quite vague, or at least variable. For example, one reference[3] indicates that a papule is less than 1 cm in diameter (if greater, it is referred to as a plaque), while another[4] mentions .5 cm. Even if both references use a precise criterion to define the size of the lesion, the public predicate "large" remains vague. Finally, the existence of hybrid concepts, such as maculo-papule or vesiculo-pustule, makes the differentiation between elementary lesions even more difficult. As shown in this example, the opposition principle does not always seem applicable in the biomedical domain. More generally, some concepts must rely on probabilistic approaches for their definition, due to biological variability (e.g., "delayed puberty", defined as an unusually late sexual maturity). As a consequence, formally, concepts such as normal and pathologic are better represented on a scale rather than through opposition.
Using a Unique Semantic Axis Principle for taxonomic relationships is a possible way to enforce the opposition principle [3]. Many existing taxonomies in medicine, however, do not rely on this principle. Moreover, when applied in some classifications, this principle fails to represent the necessary complexity of the domain. The International Classification of Diseases attempts to build a unique tree. For each node, children are opposed, using a unique semantic criterion. Some diseases, however, end up being represented more than once in the tree. For example, "pulmonary tuberculosis" is both a "pulmonary disease" (due to tuberculous bacillus) and an "infectious disease" (located in the lung). This dual representation is clearly identified by means of cross-reference relationships.

4 Discussion

This study involves two major aspects that require comment. First, the taxonomic relation is examined in the particular context of the biomedical domain whose characteristics may have an influence on it. Second, we challenge some of the principles on which this relation is based. In addition, the two-level structure used in the UMLS is discussed.

4.1 Domain characteristics

The biomedical domain is characterized by the combination of the following three features: it is a very broad domain, whose concepts cannot always be defined with precision, and where taxonomic relationships sometimes reflect *ex datis* knowledge rather than *ex principiis*.
With some 800,000 concepts, the UMLS offers a reasonable coverage of the biomedical domain. However, the integration of a new terminology into the UMLS results in the creation of new concepts, not only new terms. Drugs and genes are typical examples of ever growing areas. The progress of medical science affects not only the number of items being represented but also the taxonomy used for representing them. In contrast to application ontologies that are constrained for specific tasks and to domain ontologies representing a more limited view or a smaller area, it is understandable that an ontology of the biomedical domain may show a certain lack of consistency throughout the domain.
Moreover, representing the biomedical domain also means dealing with vague concepts. The notion of unsharpenable vagueness is discussed by Collins & Varzi [5]. In biomedicine, along with many others, an example of a vague concept is "pain". It has vague boundaries, as illustrated by the fact that there are two distinct concepts, "abdominal pain" and "abdominal

[2] When a glass slide is placed over a lesion border and pressure is applied, the lesion loses color if it is a macule, and remains colored if a purpura
[3] http://www2.medsch.wisc.edu/derm (checked July 20, 2001)
[4] http://www.ftnotebook.com (checked July 20, 2001)

discomfort", for representing meanings whose relative position in a hierarchy is not obvious. Not surprisingly, in the UMLS Metathesaurus, the two concepts stand in a circular hierarchical relationship. In addition, "pain" is typical of private language as defined by Wittgenstein, since it refers to what can be known only to the speaker, i.e. to objects that are his immediate, private sensations [17]. It also refers to predicates that cannot be contested since they are associated not only with objects (e.g. instances of pain) but also with subjects (originators). Moreover, multiple predicates may contribute to the definition of a concept such as "pain". Vagueness may also be related to the frequent use of ostensive definitions in some areas of biomedical knowledge, in particular semiology (the study of symptoms and signs). Using ostension to give the meaning of a concept does not result in a definition, and thus does not allow for accurate discrimination with other concepts. For example, pallor, one sign of anemia, is easier shown than defined. Analogously, applied to the organization of a domain, these mechanisms result in an *ex datis* representation, while definitions would allow for an *ex principiis* one.

The taxonomic relation represents predicates that are always true. In practice, however, the taxonomic relation is often used to represent predicates that are generally true. For example, "Addison Disease isa Endocrine Disease" is always true, while "Addison Disease isa Autoimmune Disease" is not, only reflecting the most frequent etiology nowadays. This phenomenon can be compared to the use of a generic term for representing prototypical knowledge, as mentioned above.

4.2 Validity of principles

While basic properties of taxonomic relations must be imposed without restriction to create hierarchies in the biomedical domain, the validity of additional principles, such as opposition of siblings or economy is arguable.

4.2.1 Opposition of siblings

In the context of an application ontology representing knowledge for Natural Language Processing in the limited domain of cardiac catheterization, Bouaud advocates the principle of sibling opposition as a way to ensure unambiguous representations [3]. Rector, on the other hand, argues that this principle is not generally applicable to a broader domain and suggests that orthogonal representations be used instead [16]. Also addressing issues concerning the opposition of siblings, Jones & Paton give examples from neurobiology and formalize this issue in terms of sortal predicates. We argue that this issue is more general in biomedicine and cannot be easily resolved. Their example, cited in [12], presents three subcategories of remote signalling cells: "endocrine cell", "paracrine cell" and "nerve cell". They point out that the most accurate representation for "neuroendocrine cell", a cell having the properties of both "nerve cell" and "endocrine cell", is as the common subtype of "endocrine cell" and "nerve cell". In this representation, however, "nerve cell" and "endocrine cell" are no longer opposable since they have a common subtype. They suggest that the original representation may be wrong, i.e. that "endocrine cell" and "paracrine cell" correspond to functional descriptions while only "nerve cell" represents an essence. According to them, endocrine describes the behavior of the cell rather than its structure, which is not sufficient for the type to have the identity property. We argue that, although its name suggests a role, "endocrine cell" is more than just a functional concept. In fact, endocrine cells are specialized cells with structural features that allow them to secrete hormones. Our representation is compatible with the properties defined for roles by Pustejovsky. The introduction of functional types "generates a functional description for an entity without of course creating a new entity in the world" [15]. Back to our example, both "endocrine cell" and "nerve cell" do carry identity and must be represented in the taxonomy. Beyond this example, our point here is that, in the biomedical domain, outside a limited, constrained domain, opposing siblings is not a valid principle.

4.2.2 Economy Principle

In the UMLS, the Economy Principle is applied to the Semantic Network with consequences for the categorization of Metathesaurus concepts. By limiting the number of categories, the Economy Principle is expected (1) to reduce the complexity of the domain, making it more manageable, and (2) to maximize the contrast between categories, making it easier to predict under which category a given item falls. Although economy in the UMLS and parsimony in ontologies may appear similar in their goals, i.e. to prevent unneeded categories from being represented, the Economy Principle does not require that *all* necessary categories be represented. However, as we mentioned earlier, when subtypes fail to be represented, there is no longer an equivalence between inclusion of classes and subordination of categories. One simplistic solution to this problem consists of defining, as needed, an additional subclass, regrouping all the items that are an instance of the superclass but are not an instance of any of the other subclasses. Such an approach is commonly implemented in many coding systems in biomedicine. In the International Classification of Diseases, for example, each major disease slot has a subdivision for diseases that cannot be classified in other slots. Creating classes by reference to sibling categories does not result in intensional definition. The definition for the resulting categories is necessarily extensional, context-dependent and unstable.

The Economy Principle may have other infelicitous consequences, as illustrated in Figure 8. A vascular dementia is a disease with both mental and somatic features. Logically, it should be categorized with the common subtype of Mental Disease and Somatic Disease. As mentioned in the introduction, rule R2 of the Economy Principle prevents hybrid subtypes from being created in the Semantic Network, and prescribes a multiple categorization instead. Thus, "vascular dementia" is expected to be assigned to both Mental Disease and Somatic Disease. However, since the only subtype available in the Semantic Network for Disease is Mental Disease, "vascular dementia" ends up being categorized directly as Disease, which is the only way its somatic features can be represented. As a detrimental consequence, based on its categorization in the Semantic Network, "vascular dementia" appears not different from, for example, "diabetes mellitus", a typical somatic disease. Moreover, the extension of Mental Disease does not contain "vascular dementia", thus conflicting with its intension.

Table 1 – Definition criteria for skin lesions

	elevated	*blanches when pressed*	*contains clear fluid*	*contains pus*	*always large*
macule	no	yes	no	no	no
purpura	no	no	no	no	no
patch	no	yes	no	no	yes
papule	yes	no	no	no	no
plaque	yes	no	no	no	yes
vesicle	yes	no	yes	no	no
bulle	yes	no	yes	no	yes
pustule	yes	no	no	yes	no

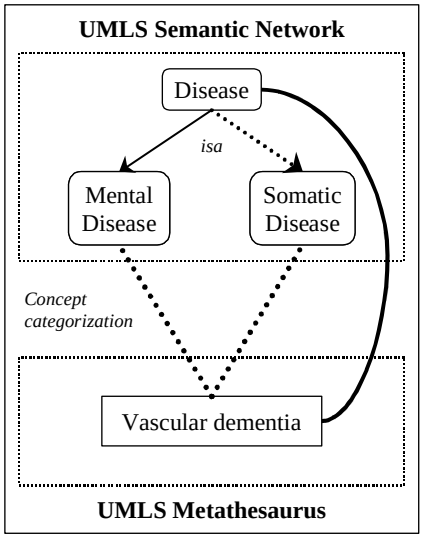

Figure 8 – Inaccurate categorization due to missing subcategories

4.3 Advantages and limits of a two-level structure

Most of the ontologies that have been developed rely on a unique structure. In contrast, a two-level structure made of, on the one hand, a small number of semantic types and on the other a huge collection of concepts, has been developed in the UMLS. A two-level structure may be justified when considering operational requirements. A two-level approach allows for organizing a small, stable, high-level taxonomy for subsequent use in reasoning activities. On the other hand, it allows for classifying the huge amount of lower-level concepts so that the most specific applicable knowledge can be inherited from the upper-level taxonomy.

An example of use of the two-level structure is given in [13]. Relationships defined between semantic types in the UMLS Semantic Network are used to infer the possible semantics of the relationship between concepts in the UMLS Metathesaurus. Inferring lower-level knowledge from a higher-level structure may be an alternative to storing all the properties explicitly where they apply. In practice, however, the inferred relationship is reported to be ambiguous in one third of the cases (e.g., a chemical can either cause or treat a disease), which constitutes an important limitation of this method.

Some attempts have been made to represent the two components of the UMLS as a homogeneous system. Gu, for example, represented Metathesaurus concepts as instances of classes derived from the semantic types [9]. Using this interpretation of the relationship between a higher-level item and a lower-level item, it is not possible to obtain a unified taxonomy by merging the two levels. Although structurally homogeneous, the resulting structure, combining *isa* and *is-an-instance-of* relations, remains semantically heterogeneous. In ONIONS, by contrast, the relation between a concept and a semantic type is interpreted as an *isa* relation [8]. The semantics of the relation between the two levels of the structure is interpreted differently in these two studies, either as *isa* or *is-an-instance-of*. In fact, most Metathesaurus concepts are subtypes of their semantic type (e.g., “Salmonella” is a kind of Bacterium), while some are instances (e.g., “American Medical Association” is an instance of Professional Society).

5 Conclusion

The principles used to produce taxonomies are either intrinsic or added to make knowledge more manageable. We studied the applicability of these principles in the biomedical domain using the UMLS and pointed out many issues raised by the application of these principles. While intrinsic principles are not challenged, we argue that the opposition of siblings brings to bear excessive constraints on a domain ontology and that the adverse effects of economy may overweigh its benefits. Despite some limitations, the two-level structure used in the UMLS represents a simple way to broadly classify a huge amount of biomedical concepts.

6 Acknowledgments

This research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

7 References

[1] Bodenreider, O. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. *Proc AMIA Symp*. (to appear).

[2] Bodenreider, O. Medical Ontology Research (Report to the Board of Scientific Counselors), Lister Hill National Center for Biomedical Communications, Bethesda, Maryland, 2001.

[3] Bouaud, J., Bachimont, B., Charlet, J. and Zweigenbaum, P., Acquisition and structuring of an ontology within conceptual graphs. in *ICCS'94 Workshop on Knowledge Acquisition using Conceptual Graph Theory*, (University of Maryland, College Park, MD, 1994), 1-25.

[4] Brachman, R.J. What Is-a Is and Isn't - an Analysis of Taxonomic Links in Semantic Networks. *Computer*, *16* (10). 30-36.

[5] Collins, J. and Varzi, A. Unsharpenable Vagueness. *Philosophical Topics*. (to appear).

[6] Fall, A. Reasoning with taxonomies *School of Computing Science*, Simon Fraser University, 1996.

[7] Fellbaum, C. (ed.), *WordNet: An electronic lexical database*. MIT Press, Cambridge, Massachusets, 1999.

[8] Gangemi, A., Pisanelli, D.M. and Steve, G. An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies. *Data & Knowledge Engineering*, *31* (2). 183-220.

[9] Gu, H., Perl, Y., Geller, J., Halper, M., Liu, L.M. and Cimino, J.J. Representing the UMLS as an object-oriented database: modeling issues and advantages. *J Am Med Inform Assoc*, *7* (1). 66-80.

[10] Guarino, N. The role of identity conditions in ontology design. *Spatial Information Theory*, *1661*. 221-234.

[11] Guarino, N. and Welty, C., Ontological Analysis of Taxonomic Relationships. in *ER-2000: The 19th International Conference on Conceptual Modeling*, (2000), Springer-Verlag, 210-224.

[12] Jones, D.M. and Paton, R.C. Toward principles for the representation of hierarchical knowledge in formal ontologies. *Data & Knowledge Engineering*, *31* (2). 99-113.

[13] McCray, A.T. and Bodenreider, O. A conceptual framework for the biomedical domain. in Sung, M. and Green, R. eds. *Semantics of Relationships*, Kluwer, 2001, (to appear).

[14] McCray, A.T. and Nelson, S.J. The representation of meaning in the UMLS. *Methods Inf Med*, *34* (1-2). 193-201.

[15] Pustejovsky, J. Type Construction and the Logic of Concepts. in Bouillon, P. and Busa, F. eds. *The Syntax of Word Meaning*, Cambridge University Press, 2001.

[16] Rector, A.L., Zanstra, P.E., Solomon, W.D., Rogers, J.E., Baud, R., Ceusters, W., Claassen, A.M.W., Kirby, J., Rodrigues, J., Rossi Mori, A., Van der Haring, E.J. and Wagner, J. Reconciling user's needs and formal requirements: issues in developing a reusable ontology for medicine. *IEEE Transactions on Information Technology in Biomedicine*, *2* (4). 229-241.

[17] Wittgenstein, L. *Philosophical investigations*. Blackwell, Cambridge, Massachusets, 1997.

[18] Woods, W.A. Understanding subsumption and taxonomy: A framework for progress. in Sowa, J.F. ed. *Principles of Semantic Networks*, Morgan Kaufmann, San Mateo, CA, 1991, 45-94.