Proceedings of AMIA Annual Symposium 2001;:86-90.

Mapping the UMLS Semantic Network into General Ontologies

Anita Burgun, M.D, Ph.D., Olivier Bodenreider, M.D., Ph.D.
National Library of Medicine, Bethesda, Maryland
{burgun,olivier}@nlm.nih.gov

In this study, we analyzed the compatibility between an ontology of the biomedical domain (the UMLS Semantic Network) and two other ontologies: the Upper Cyc Ontology (UCO) and WordNet.
1) We manually mapped UMLS Semantic Types to UCO. One fifth of the UMLS Semantic Types had exact mapping to UCO types. UCO provides generic concepts and a structure that relies on a larger number of categories, despite its lack of depth in the biomedical domain.
2) We compared semantic classes in the UMLS and WordNet. 2% of the UMLS concepts from the Health Disorder class were present in WordNet, and compatibility between classes was 48%. WordNet, as a general language-oriented ontology is a source of lay knowledge, particularly important for consumer health applications.

INTRODUCTION

The Semantic Network in the Unified Medical Language System® (UMLS®) is a high-level representation of the biomedical domain based on Semantic Types under which all the Metathesaurus® concepts are categorized, and which "may be considered a basic ontology for that domain" [1]. Ontologies range in abstraction from very general concepts that form the foundation for knowledge representation for all domains to concepts that are restricted to specific domains [2]. Four main categories of ontologies may be described according to their coverage, and the task(s) they are designed for.

1) **General ontologies**. They represent general knowledge, independently of specific domains or tasks, with a medium level of precision.

2) **Domain ontologies**. Ontologies that are specific of a domain but independent of a task are generally called domain ontologies [3]. They should reflect underlying reality and theory of the domain.

3) **Upper-level ontologies**. Concepts related to Space, or Time are high level ones, and apply to all domains. Thus they belong to ontologies referred to as upper-level ontologies (ULOs). ULOs should be universal, i.e. they should not refer to specific domains, and every concept one needs in a specific domain can be linked to a ULO. They should be multi-purpose, i.e. they should not have been designed for specific tasks.

4) **Application ontologies**. These have restricted scopes and are driven by specific objectives (tasks). Application ontologies are more or less embedded in an application, and contain a relatively small number of concepts that should be defined in some detail, with relations and inference rules that enable reasoning with them for the intended tasks.

Part of the research in medical informatics focuses on reusability of knowledge in new applications and design of sharable ontologies. For Musen [4], the principal obstacles to knowledge sharing and reuse involve the difficulties of achieving consensus regarding what knowledge representations mean, of enumerating the context features and background knowledge required to ascribe meaning to a particular knowledge representation, and of describing knowledge independent of specific processes. Nevertheless, several attempts have been made:

- to merge general ontologies [5],
- to reuse domain ontologies for specific applications [6-7], or
- to integrate domain ontologies with large-scale ontology libraries [8].

This study is a contribution to the Medical Ontology Research project currently developed at the National Library of Medicine [9]. The major objective of this project is to develop methods whereby ontologies could be acquired from existing resources (including the UMLS), as well as validated against other knowledge sources. Our work focused on the UMLS Semantic Network, as a potential domain ontology for biomedicine. Our objective was to test the compatibility of UMLS categories with categories from some general ontologies or ULOs. The UMLS includes categories that do not specifically belong to the biomedical domain, such as *Physical Object*, or *Animal*. Similarly, general ontologies incorporate biomedical categories. Compatibility is expected:

- in Semantic Types (STs) viewed as categories (intensional definition). The intensional representation of STs had to be compared to a general ontology whose function is Knowledge Representation, i.e. which aims at describing the world, not studying language [10]. The choice of Cyc® was motivated by the fact that it provides a sufficient general grounding, while it may encompass specific views (microtheories) [11].

- in STs viewed as classes (extensional definition). The UMLS provides a two-level representation, where STs are categories under which Metathesaurus concepts are categorized. We call class, or ST extension, the set of Metathesaurus concepts that are assigned to a given ST. The extensional representation of STs was compared to WordNet® [12]. Contrary to Cyc's, the structure of WordNet is close to that of the Metathesaurus (terms, concepts, hierarchies). Classes in WordNet can be derived from hyponymic (isa) relations. Furthermore, WordNet provides a general terminology in many semantic fields, including the biomedical domain.

Background design principles are discussed, essentially from the viewpoint of the UMLS.

MATERIAL AND METHODS

The ontologies

Upper Cyc Ontology (1997 release) is a set of approximately 3,000 general concepts. *Thing* is the universal set, which is divided into *Individuals*, and *Collections*. Cyc items are hierarchically organized by means of two structuring relations:

- #$isa represents the classical subsumption, i.e. means "is an instance of". (#$isa El Col) means that El is an element of the collection Col.
- #$genls is the relation between a collection and its superordinate. (#$genls Col Sup) means that Sup is a category that is a superordinate of Col.

WordNet organizes lexical information in terms of meanings and semantic relations. Synonyms are clustered by meanings, and a set of synonyms is called a synset. The current version (1.6) contains approximately 100,000 synsets. Hyponymy relations are instantiated between synsets, according to the following definition: "A concept represented by the synset {x,x',…} is said to be a hyponym of the concept represented by the synset {y,y',…} if native speakers of English accept sentences constructed from such frames as *An x is a kind of y*" [13].

The 2001 release of the **UMLS Semantic Network** represents 134 Semantic Types that are relevant for the biomedical domain and categorize some 800,000 Metathesaurus concepts [14]. The isa link allows STs to inherit properties from higher level nodes.

Method

Mapping the UMLS Semantic Types to Upper Cyc Ontology.

Descriptions of the STs in the Cyc formalism were performed manually, using the Cyc #$isa and #$genls relationships. Additional Cyc categories are used as required to insure consistency.

The relationship between a given ST T and the closest Cyc concept U is called Similarity if T has an equivalent U, whatever its name. It is called Overlap if there is a partial overlap between T and U; in this case T and U are compatible, and have a common supertype. A classical example of overlapping categories is *Dog* and *Pet* [15].

Mapping the UMLS Semantic Types to WordNet.

The mapping of UMLS and WordNet classes is based on comparing the sets of concepts that are subsumed by a given ST in the UMLS, and the sets of hyponyms of a given synset in WordNet. We focused on two classes: ANIMAL, which is a general class, supposed to be similarly represented in both systems, and HEALTH DISORDER, which is a typical medical class. Details about the constitution of the classes are given in Table 1.

Class	WordNet	UMLS
ANIMAL	The synset *Animal* and all its hyponyms	Metathesaurus concepts assigned to the ST *Animal* or any of its subtypes
HEALTH DISORDER	The union of the following synsets and all their hyponyms: *Symptom* *Ill Health* *Disorder (sense 1)* *Mental retardation* *Mental Illness* *Defect (sense 1)* *Abnormalcy*	Metathesaurus concepts assigned to any of the STs: *Anatomical Abnormality* *Congenital Abnormality* *Acquired Abnormality* *Finding* *Sign or Symptom* *Pathologic Function* *Disease or Syndrome* *Mental or Behavioral Dysfunct.* *Neoplastic Process* *Cell or Molecular Dysfunct.* *Experimental model of Disease* *Injury or Poisoning*

Table 1 – ANIMAL and HEALTH DISORDER classes

Starting from a list of terms and concepts belonging to a given class in the UMLS, those terms were mapped to WordNet. The mapping program was based on the WordNet *wn* standard function. For each concept, it was recorded whether the WordNet term mapped to belonged to the corresponding class in WordNet. For example, "Fever", categorized as *Sign or Symptom* in the UMLS is mapped to "Fever" in WordNet, a hyponym of *Symptom*. As shown in Table 1, *Sign or Symptom* is one of the STs that define the class HEALTH DISORDER in the UMLS, and *Symptom* is one of the synsets that define the WordNet class HEALTH DISORDER. Therefore, "Fever" belongs to HEALTH DISORDER in both systems. This method provides a means for comparing what falls under a given category in the UMLS and WordNet.

RESULTS

UMLS vs. Upper Cyc Ontology

Roughly 50 Cyc categories were used for strictly covering the UMLS Semantic Network field. Approximately half of them were similar in both systems, for example, *Fish* is similar in Cyc and in the UMLS. For the others, there was overlap between the Cyc type and the UMLS ST. For example, Cyc *GeneticCondition* represents abnormal conditions that developed in a particular organism due to that organism's genetic configuration, and are often harmful, but also may be beneficial. Thus, it maps totally neither *Genetic Function* nor *Cell or Molecular Dysfunction* in the UMLS. Representation of anatomy differs, since elements of Cyc category *Animal Body Region* may be unhealthy body regions such as blisters, puncture wounds, which are *Abnormalities* in the UMLS. In Cyc, *Animal Body Part* is a subtype of *Animal Body Region* that includes both organs and body systems.

For several UMLS STs (e.g., for chemicals), there was no equivalent category in the public version of Upper Cyc Ontology.

Additional Cyc categories that had no equivalent in the UMLS were integrated in our representation of the UMLS in order to build a structure that was consistent with Cyc. They represent:

- intermediate nodes, such as *Primate* (#$genls Person Primate), (#$genls Primate Mammal)
- generic concepts, such as *SimpleRepairing* which is a supertype of *MedicalTreatmentEvent*
- additional knowledge, such as *BiologicalTaxon* which provides information about biological categories, according to the general taxonomy of living beings.

UMLS vs. WordNet

The set of 11,634 UMLS concepts from the ANIMAL class was mapped to WordNet, whose ANIMAL class contains 3,984 synsets. 2,154 UMLS concepts (19%) were found in WordNet, 73% of them in the WordNet ANIMAL class.

Examples of UMLS specific ANIMAL concepts are "Acanthamoeba", "Angiostrongylus". Examples of WordNet specific concepts are "kitty", "unicorn", "cotton ballworm", and "Mickey Mouse".

Most of the UMLS ANIMAL concepts that are categorized differently in WordNet are biological taxons. For example, "Cetacea", "Ascaridia" are categorized as animals in the UMLS while they are hyponyms of "taxonomic group" in WordNet.

The UMLS HEALTH DISORDER class contains more than 140,000 concepts, which were mapped to WordNet. 2,639 UMLS concepts (2%) were found in WordNet, and among them, 1,257 concepts (48%) belonged to the WordNet HEALTH DISORDER class. The WordNet HEALTH DISORDER class contains 1,379 synsets.

Among the HEALTH DISORDER concepts present exclusively in WordNet, 80 are plant diseases. Other specific WordNet items include "astraphobia", "crick", and "sword cut".

Among the UMLS HEALTH DISORDER concepts that are found in WordNet outside the HEALTH DISORDER class, many are hyponyms of generic concepts in WordNet, mostly referring to the process involved in the disorder. For example, in WordNet, "bronchospasm" is a hyponym of "constriction", and "abortion" is a hyponym of "termination".

Within a class, concepts may be categorized differently, even when the categories look similar. For example, *Symptom* has equivalent definitions in WordNet, where it is "any sensation or change in bodily function that is experienced by a patient and is associated with a particular disease", and in the UMLS, where *Sign or Symptom* is "an observable manifestation of a disease or condition based on clinical judgment, or a manifestation of a disease or condition which is experienced by the patient and reported as a subjective observation". This semantic similarity leads to a high proportion of concepts categorized similarly in both systems, e.g., "cyanosis", "fever". However, *Symptom* in WordNet is also a hypernym of "encephalitis", "tennis elbow", and numerous other conditions that are categorized as *Disease or Syndrome* in the UMLS.

DISCUSSION

Diversity of Ontologies

Diversity in structure. Despite current efforts of the IEEE Standard Upper Working Group, no standard ULO is available yet. In this context, other ULOs may have been analyzed, such as the Generalized Upper Model, which has three top categories: *Configuration*, *Element*, *Sequence* [16], or Sowa's T root which combines five top categories: *Abstract*, *Physical*, *Independent*, *Relative*, and *Mediating* [17]. Such diversity in the top level of ULOs illustrates the potential difficulties in generalizing our results.

Diversity in formalisms. While the UMLS makes a distinction between upper level categories in the Semantic Network and lower level concepts in the Metathesaurus, both Cyc and WordNet use a unique formalism whatever the level in the concept "hierarchy". Formalism influences the representation of top level categories, properties and roles. Properties, e.g., 'non human', are categories in some ULOs, while they are ST attributes in the UMLS.

Issues in reconciling diverse ontologies. Our experience in this project suggests that alignment

algorithms devoted to automated mapping between ontologies (such as [18]) have to address two major semantic issues:

- Categories that have similar names in different ontologies may have distinct meanings. For example, *Entity*, *Body Part*, *Body Region* do not have the same meaning in Cyc and in the UMLS.
- Moreover, two categories may have similar intensions while neither their respective classes are identical nor is one class totally included in the other, as illustrated by the extensions of *Symptom* in WordNet and in the UMLS.

Respective contributions

Each ontology brings not only its own perspective on the world but also, practically, different pieces of knowledge (Fig 1). In the biomedical domain, mapping between domain ontologies and general ontologies consists of mapping between specific medical meaning and general meaning.

Generic concepts. General meaning refers to generic concepts, such as *Path* or *Simple Repairing*. ULOs rely on conceptualizations, which are supposed to be the same, whatever the domain they apply to [19]. *Path* applies to travel, but it is also applicable to the circulatory system. *Simple Repairing* virtually applies to repairing processes in every domain (mechanism, medical procedures, etc). Similarly, there is no reason for anatomy not to reuse general categories representing spatial objects (line, surface, etc). Therefore, our approach of mapping between domain ontologies and general ontologies may be of interest for cross-validation of basic general categories (discrepancies between general categories as defined in a general ontology and used in a domain ontology may be revealed by the mapping process), the latter providing domain ontologies with a means for reasoning from basic general theories of the world.

Common sense knowledge. General meaning also refers to common sense knowledge – in the form of folk representation of the biomedical domain [20]. Common-sense knowledge sometimes differs from expert knowledge, and some ontologies are based on lay representations while others reflect scientific theories of a domain. For example, in WordNet, epilepsy is "a disorder of the central nervous system characterized by loss of consciousness and convulsions". For health professionals, however, this definition only refers to one clinical form of epilepsy. Another aspect of common sense knowledge is represented by lay concepts provided by general language-oriented ontologies. For example, "kissing disease" in WordNet and "infectious mononucleosis" refer to the same disease while providing different representations. Therefore, our approach of mapping between ontologies representing expert knowledge and ontologies capturing common-sense knowledge may be helpful for acquiring the knowledge needed for consumer health oriented applications such as MEDLINE*plus* and *ClinicalTrials.gov*.

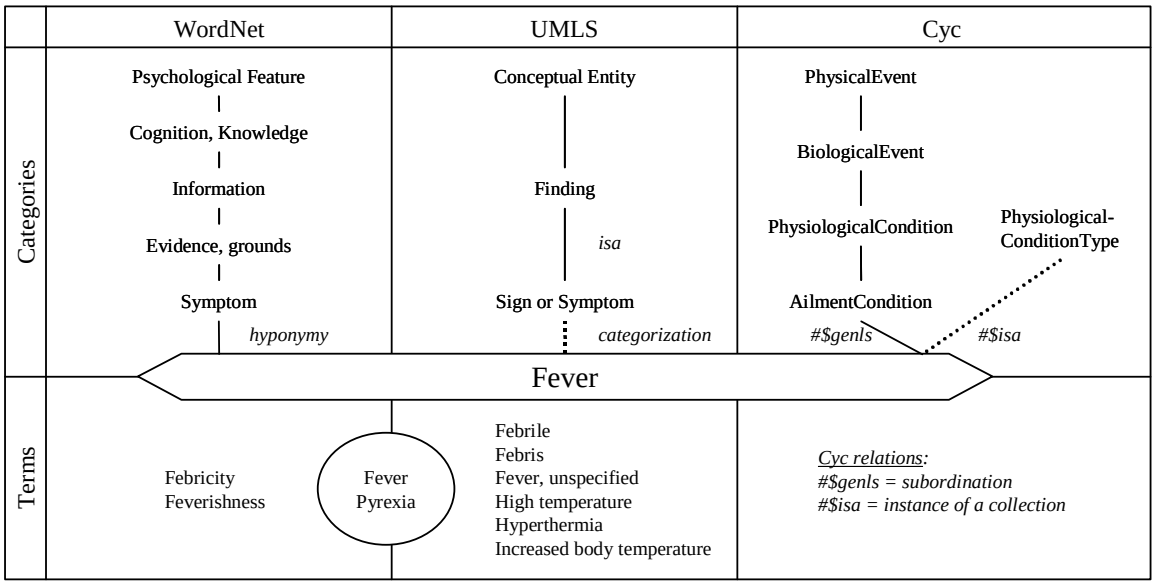

Figure 1 – Fever in WordNet, Cyc and the UMLS: categorization and associated linguistic phenomena.

Meta-knowledge. Ontologies often make use of meta-knowledge for representing not the world itself but models of the world. The biological taxonomy, which is the classification of living organisms in kingdoms, species, etc, provides an organizational model, but does not describe the world. Such meta-knowledge may be represented by categories or not. For example, while Cyc and WordNet have categories for biological taxons, the UMLS has none.
When represented, meta-knowledge categories may be clearly identified as such, for example in Cyc, *Biological Taxon* is a *Biological-Taxon-Type*, which is a child of *Conventional-Classification-Type*, whose other children also represent meta-knowledge.
Choices made for representing meta-knowledge strongly influence the capability of a system to infer from the subsumption relation. For example, as represented in Fig. 1, we can infer from Cyc representation restricted to the #$genls relation, that *Fever* is a *BiologicalEvent*. Cyc representation prevents meta-categories from being considered as superordinates since they are linked by #$isa relation. When no mechanism is provided for distinguishing between meta-level categories from other categories, this can result in inaccurate inferences. For example, according to the taxonomic relationships, *Fever* ends up being categorized as a *Psychological Feature* in WordNet, and as a *Conceptual Entity* in the UMLS (Fig. 1).

This study raised interesting issues about the UMLS Semantic Network and its role as an ontology of the biomedical domain. Further research is underway, focusing on specific aspects of the compatibility between ontologies (e.g., comparing definitions), as well as clarifying the use of the isa relation in the biomedical domain.

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 NLM. The authors thank Cycorp and the University of Princeton for making their products available for research.

References

1. McCray AT, Nelson SJ. The representation of meaning in the UMLS, Meth Inform Med, 1995, 34: 193-201
2. Chandrasekaran B, Josephson JR, Benjamins VR. What are ontologies, and why do we need them? IEEE Intelligent Systems, 1999, Jan/Feb, 20-26
3. Gennari JH, Tu SW, Rothenfluh TE, Musen MA. Mapping domains to methods in support of reuse. Int J Hum-Comput St, 1994, 41: 399-424
4. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res, 1992, 25:435-467
5. Kiryakov AK, Simov KI. Mapping of EuroWordnet Top Ontology into Upper Cyc Ontology. 2000, Proc KAW 2000
6. Volot F, Zweigenbaum P, Bachimont B, et al. Structuration and acquisition of medical knowledge. Using UMLS in the conceptual graph formalism. Proc Annu Symp Comput Appl Med Care 1993;:710-4
7. Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS Semantic Network. Proc AMIA Symp 1999;:181-5
8. Gangemi A, Pisanelli DM, Steve G. An overview of the ONIONS project: applying ontologies to the integration of medical terminologies Data and Knowledge Engineering, 1999, 31 (2): 183-220
9. Bodenreider O. Medical Ontology Research, Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications, 17 May 2001
10. Davis R, Shrobe H, Szolovits P. What is a Knowledge Representation? AI Magazine, 1993, Spring, 17-33
11. Cyc Public Ontoloy. Available at http://www.cyc.com/ (July 2001)
12. Fellbaum C (edt). WordNet an electronic lexical database. MIT Press, 1998
13. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller K. Introduction to WordNet. Available at http://www.cogsci.princeton.edu/ (July 2001)
14. UMLS Knowledge Sources. (11th ed.) Bethesda (MD): National Library of Medicine, 2001.
15. Cruse DA. Lexical Semantics. Cambridge University Press, 1986.
16. GUM upper level Available at http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html (July 2001)
17. Sowa JF. Knowledge Representation. Brooks Cole, 2000.
18. Noy NF, Musen MA. An algorithm for merging and aligning ontologies: automation and tool support. Proc 16th Conference on Artificial Intelligence (AAAI-99)
19. Guarino N, Giaretta P. Ontologies and Knowledge Bases. Towards a terminological clarification. In NJI Mars (edt) Towards very large knowledge bases, IOS Press, Amsterdam, 1995
20. Smith B. Formal ontology, common sense and cognitive science. Int J Hum-Comput St, 1995, 43: 641-667