Modeling the UMLS Using an OODB

Huanying (Helen) Gu, Yehoshua Perl, James Geller, Michael Halper[1], Li-min Liu, James J. Cimino[2]

CIS Dept., New Jersey Institute of Technology, Newark, NJ 07102
[1]Mathematics & Computer Science Dept., Kean University, Union, NJ 07083
[2]Dept. of Medical Informatics, Columbia University, New York, NY 10032

The Unified Medical Language System combines many well established authoritative medical informatics terminologies in one system. Such a resource is very valuable to the healthcare industry. However, the UMLS is very large and complex and poses serious comprehension problems for users and maintenance personnel. Furthermore, the sets of concepts of semantic types are not semantically uniform and thus are difficult to study. We describe a method to represent two components of the UMLS, the Metathesaurus (META) and the Semantic Network, as an OODB. The resulting UMLS OODB schema is deeper and more refined than the Semantic Network. It offers semantically uniform classes, which improves support for comprehension and navigation of META. The UMLS OODB also exposes problems in the semantic type classifications.

INTRODUCTION

The Unified Medical Language System (UMLS) [1–4] designed by the National Library of Medicine (NLM) combines many well established medical informatics terminologies in a unified system. It enables electronic access to a very large compendium of medical terminologies.

However, the UMLS is large and complex, which poses serious comprehension problems. It is difficult to maintain and use the UMLS without proper comprehension. Designers, maintainers and users of the UMLS need tools to help with their work. Tools for retrieval and manipulation of the content of such a system are insufficient. Rather, they must help professionals reach a level of *comprehension* essential to performing their tasks.

In previous work [5,6], we have developed a methodology for representing Controlled Medical Terminologies (CMTs) as Object-Oriented Databases (OODBs) to provide support for comprehension of their structure and content. The comprehension support was achieved via the schema layer which gives an abstract view of the source CMT. In [7,8], we described how the schema layer of the Medical Entities Dictionary (MED) [9] helped to uncover and correct errors and inconsistencies in its content.

In this paper, we utilize an OODB representation to capture the knowledge of the Metathesaurus (META) and the Semantic Network. META is a compilation of terms and associated information from over 40 medical terminologies and classifications. The Semantic Network contains information about semantic types and the permissible relationships among these types [10–12]. Each concept in META is assigned to one or more semantic types from the Semantic Network. To handle concepts of multiple semantic types, we define compound semantics of concepts. Thus, semantic types contain semantically non-uniform set of concepts. Initially, we map all semantic types in the Semantic Network onto an OODB schema. To precisely capture the semantics of the concepts of multiple semantic types, we introduce a new kind of class, called intersection class, for their representation. Consequentially, all classes abstract semantically uniform sets of concepts. The resulting UMLS OODB schema has a deeper structure than the Semantic Network of the UMLS and supports the comprehension and navigation of META.

OODB REPRESENTATION OF THE UMLS

The Semantic Type Classes

In general, a class in an OODB schema represents a group of instances with the same properties and a common semantics. The OODB schema gives an abstract view of a database. The Semantic Network of the UMLS contains 132 semantic types, arranged in an IS-A hierarchy. Each concept of META is associated with one or more semantic types. Thus, the Semantic Network provides a high level abstract view of META. To model the UMLS as an OODB, we represent the semantic types as classes in the OODB schema, called a *semantic type class.* The IS-A links of the Semantic Network become *subclass* relationships in the OODB schema. If a concept is assigned to only one semantic type, then we make it an instance of the corresponding semantic type class. E.g., the concept **Air** is assigned to the semantic type **Substance**. After mapping, **Air** becomes an instance of the class "Substance." Thus, 357,804 concepts in META are immediately represented.

Non-Uniform Semantics of Semantic Types

However, concepts may belong to more than one semantic type. E.g., the concept **Cotton** belongs to **Substance** and **Plant**; the concept **Norepinephrine preparation** belongs to **Organism**, **Pharmacologic Substance**, **Neuroreactive Substance or Biogenic Amine**, and

Number of assigned semantic types	Number of concepts
1	357,804
2	108,905
3	9,262
4	331
5	10
6	2

Table 1: The distribution of concepts in the Semantic Network

Hormone. For more details on the distribution of concepts of multiple semantics, see Table 1.

In OODBs all instances of a class must have the same structure and semantics. When modeling the UMLS using an OODB, the semantics of a concept are provided by its semantic types. A concept with only one semantic type has a *simple semantics*. A concept that belongs to a set of semantic types has a *compound semantics* defined by the combination of its semantic types. Thus, we cannot assign concepts with a simple semantics to the same class as concepts with a compound semantics. The concepts of a semantic type may be semantically non-uniform. E.g., the semantic type **Experimental Model of Disease** contains 39 concepts, the concept **Radiation Injuries, Experimental** has one additional semantic type **Injury or Poisoning**; the concept **Water Deprivation** has one additional semantic type **Diagnostic Procedure**; 27 concepts have one additional semantic type **Neoplastic Process**, and the concept **Lesion, NOS** has two additional semantic types **Functional Concept** and **Sign or Symptom**, leaving 9 concepts belonging only to the semantic type **Experimental Model of Disease**. It is difficult to comprehend and use the information contained in such a semantically non-uniform semantic type.

The Intersection Classes

Each concept with multiple semantic types should be represented as an instance of one class in the schema, with the appropriate compound semantics. All concepts belonging to several semantic types should be assigned to a new kind of class, called an *intersection class*. An intersection class represents the combination of two or more semantic types. In order to create intersection classes, all concepts with multiple semantic types are partitioned into groups such that each group contains the concepts belonging to the same set of semantic types. That means, the concepts in one group have the same compound semantics. For each group, a corresponding intersection class with the corresponding compound semantics is created and the concepts in each group become its instances.

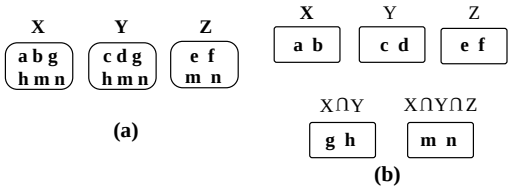

Figure 1: Three semantic types and the OODB representation

In Figure 1 (a), we show three semantic types **X**, **Y**, and **Z** and the concepts assigned to them. (To illustrate our partitioning process, the concept (instance) names are placed inside rounded-edge rectangles (rectangles) which represent semantic types (classes).) The three semantic types **X**, **Y**, **Z** are represented as three semantic type classes "X," "Y," "Z" (Figure 1 (b)). The concepts **a**, **b**, **c**, **d**, **e**, and **f** belong to only one of "X," "Y," and "Z." Since the concepts **g**, **h**, **m**, and **n** belong to more than one of **X**, **Y**, **Z**, they are removed from **X**, **Y** and **Z** and partitioned into two groups. Two intersection classes "X ∩ Y" and "X ∩ Y ∩ Z" are created to represent those two groups. The symbol "∩" indicates the intersection.

Now each one of the 476,314 concepts in META is represented as an instance of only one class. The schema consists of 1,295 classes, including 1,163 intersection classes.

Advantages of the OODB Class Representation

Uniform Semantic Classification: The concepts assigned to a semantic type are not semantically uniform since some of them belong to several semantic types. The semantic type classes and intersection classes both have semantically uniform extents, which we believe are easier to comprehend.

Reduced Average Extent Size: In the UMLS, the extents of many semantic types are very large. In the 1998 version, on average, every semantic type corresponds to about 5,000 concepts. Adding the intersection classes to the schema reduces the average size of semantic type class extents to about 2,700. The average number of concepts in each intersection class is only 100. Having classes with smaller extents simplifies the use of META.

Exposing Problems in the Current UMLS: *Omissions* In the UMLS schema, there is an intersection class "Body Part, Organ, or Organ Component ∩ Medical Device." It contains only four concepts **Dental abutments**, **Conduit with xenograft valve**, **Conduit with homograft valve**, and **Incubator.pediatric**. However, there are more medical devices, e.g., heart valves, which should be in this intersection class, and are missing

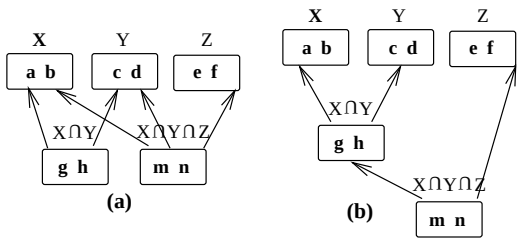

Figure 2: Two feasible solutions of adding *subclass* relationships

from META.

Classification Errors Intersection classes highlight some errors. E.g., **Encephalities Viruses** is the only instance of "Virus ∩ Disease or Syndrome." But it is only a virus and should not be classified as a disease. Hence, it should be only in the "Virus" semantic type class and there is no need for this intersection class. Another example is **Scotch Tape Mount** which is the only instance of "Bacterium ∩ Laboratory Procedure." However, it is not a bacterium and should be only in "Laboratory Procedure." Thus, this intersection class should not exist either.

Homonymity Intersection classes uncover some ambiguities of concepts. E.g., "Plant ∩ Disease or Syndrome" has only one instance **Toxicodendron**. However, **Toxicodendron**, known as poison ivy, refers to two different concepts, one is a plant and another is the name of a disease. In order to differentiate them, two concepts should be created such that one is in "Plant" and another is in "Disease or Syndrome". Thus, there will be no such intersection class. Let us look at another example. **Paronychia of toe** is the only instance in "Anatomical Structure ∩ Disease or Syndrome." The classification exposes two potential different concepts. One is the diseased toe which is a body part in "Anatomical Structure" and another is the disease of the toe in "Disease or Syndrome." Thus, no such intersection class is necessary. The previously mentioned concept **Cotton** is another such example.

OODB Schema Subclass Relationships

After introducing the intersection classes, we need to determine the superclasses of each intersection class. Several semantic type classes contribute their semantics to the instances of each intersection class, and we might use those classes as superclasses of the intersection class (Figure 2 (a)). Thus, by using this straightforward approach, an intersection class is one level lower than its contributing semantic type classes. There are no intersection classes which are superclasses of other intersection classes.

In OODBs the *subclass* relationships point from specific classes to general classes. By transitivity, every class is implicitly a subclass of all ancestors of its superclasses. (By ancestors we refer to classes reachable following a chain of superclass relationships.) For example, in Figure 2 (a) we see "X ∩ Y," which is a subclass of "X" and "Y," and "X ∩ Y ∩ Z" which is a subclass of "X", "Y", and "Z". For a class which is an intersection of two, the only option is to make it a subclass of those two classes. However, for the intersection of more than two, there may be more than one alternative to define the *subclass* relationships. The semantics of "X ∩ Y ∩ Z" is more specific than the semantics of "X ∩ Y." Hence, it is natural to have a subclass relationship from "X ∩ Y ∩ Z" to "X ∩ Y." Since "X ∩ Y" is a subclass of "X" and "Y," the transitivity implies that "X ∩ Y ∩ Z" is a subclass of both "X" and "Y" and thus these subclass relationships do not need to be explicit in the schema. Figure 2 (b) shows a refined modeling, where an intersection class may be a subclass of another intersection class. As a result we have intersection classes distributed into multiple levels. In order to systematically define the *subclass* relationships, we need a rule. We first need to give two definitions.

Let U be a universal set of elements and let F be a given family of sets over U (By family, we mean a set of sets). That is, F is a subset of the power set of U. We call the set of instances of a class C the extent of C, the set of concepts of a semantic type S the extent of S, and the set of all concepts of META M the extent of M.

Definition 1: Let A and B be sets in F, such that A is a subset of B. If there does not exist any set C in F such that A is a subset of C and C is a subset B, then we call A a *maximal subset* of B in F. (E.g., if $\{X, Y, Z\}$, $\{X, Y\}$, and $\{X\}$ are three sets in F, then $\{X, Y\}$ is a maximal subset of $\{X, Y, Z\}$ and $\{X\}$ is not.)

For the UMLS, U is the set of elements of META, and F is the set of the extents of all semantic types. When an intersection class is given, it is possible to identify all its *potential superclasses* for which there may exist an implied subclass relationship. For a given family of sets G which is a subset of F, the intersection I_G is the intersection of all extents in G. Furthermore, for each D such that D is a subset of G, the intersection class C_{I_D} is a potential superclass of C_{I_G}.

Definition 2: Let C_{I_G} be an intersection class corresponding to the intersection I_G. If C_{I_D} is a potential superclass of C_{I_G}, then C_{I_D} is a *minimal superclass* of C_{I_G} if D is a maximal subset of G. (Note that D may be a family of the extent of one semantic type.)

Rule: Let C_I be an intersection class in the schema. The subclass relationships in the schema are defined from C_I to all its minimal superclasses in the schema.

This rule will increase the depth of the schema by making some intersection classes subclasses of others. As McCray [13] notes, it is considered desirable to increase the depth of the Semantic Network. For example (Figure 2 (b)), the classes "X ∩ Y" and "Z" are the only two minimal superclasses of the intersection class "X ∩ Y ∩ Z." Compared with Figure 2 (a), with 5 subclass relationships, Figure 2 (b) contains only 4 subclass relationships. In [14], we defined the complexity of a schema as the ratio between the number of relationships and the number of classes. Thus, when two schemas contain the same classes, the schema with fewer relationships is simpler, and Figure 2 (b) is simpler than Figure 2 (a). Using the above rule to define subclass relationships results in a more refined schema with 2,807 subclass relationships (2,677 from intersection classes). Figure 3 shows the distribution of classes.

Figure 4 is a subschema of the resulting UMLS schema, which shows the intersection class "Organic Chemical ∩ Organophosphorus Compound ∩ Pharmacologic Substance ∩ Therapeutic or Preventive Procedure" and all its superclasses and ancestors. It contains 15 semantic type classes and 6 intersection classes distributed over 11 levels.

ADVANTAGES OF THE OODB SCHEMA REPRESENTATION

Deeper Schema

We cite from McCray and Nelson [13]: "The current scope of the (Semantic) Network is quite broad, yet the depth is fairly shallow. We expect to make future refinements and enhancements to the Network, based on actual use and experimentation." Introducing the intersection classes achieves this goal.

Uncovering Redundant Classifications

By creating intersection classes, we uncovered that 8,622 concepts in META are assigned to several semantic types which stand in parent-child or ancestor-descendant relationships. E.g., the intersection class "Organic Chemical ∩ Organophosphorus Compound" (Figure 4) has two superclasses "Organic Chemical" and "Organophosphorus Compound." However, "Organophosphorus Compound" is itself a child of "Organic Chemical." This is not in line with the intentions of the UMLS designers. In [13], when discussing the assignment of concepts to semantic types, it is stated that "In all cases the most specific semantic type available in the hierarchy is assigned to a term." If all those redundant classifications are removed from the UMLS, 77 intersection classes will disappear from the schema. A list of the above 8,622 concepts and their associated semantic types was submitted to NLM and redundant type assignments will be removed from the next version of the UMLS.

Traversal

Because the Semantic Network and META are unified into an OODB, a combined traversal of the schema and concept layers is possible. This combined traversal is faster and shorter than a traversal of META itself, since the OODB schema is much smaller than META.

Suppose that a user is searching for a concept in the UMLS and does not know its name, but he would recognize it when he encounters it. Instead of traversing META through its many levels, he can traverse the OODB schema until the proper class, say S, is identified. At this point, the user switches to the subnetwork of all the concepts of S. The traversal runs through this subnetwork until the desired concept is recognized. A traversal requires repeated scanning through lists of children and chosing one. This is easier at the schema level, since the number of subclasses of a class is typically much smaller than the number of children of a concept.

Consider a traversal to the concept **Delusion of self-accusation**. We will list a sequence of concepts with the number of children of each in (). The user needs to pick one child at every step. We traverse through **Medical Subject Headings** (15), **Diseases (MeSH Category)** (45), **Symptoms and General Pathology** (38), **Disease** (124), **Mental Disorders** (226), and **Delusions** (19), to **Delusion of self-accusation**. This path of 7 concepts requires the user to scan a total of 467 children. We will now contrast the above with a traversal using the OODB schema. We traverse from the root class "Event" (2), through "Phenomenon or Process" (3), "Natural Phenomenon or Process" (1), "Biologic Function" (2), "Pathologic Function" (3), "Disease or Syndrome" (2), "Mental or Behavioral Dysfunction" (14), to "Mental or Behavioral Dysfunction ∩ Sign or Symptom." At this class, we switch to the concept level. From **Delusions** (19), we continue to **Delusion of self-accusation** which is our goal. This traversal uses 9 classes with a total of 29 children and 2 concepts with 19 children. The total children number (29+19=48) is much smaller than the 467 from before. Thus, the combined traversal using the schema is faster.

CONCLUSIONS

The size and complexity of the UMLS make it difficult to maintain and use. To overcome this problem, we have developed a methodology for representing META and the Semantic Network of the UMLS as a unified OODB. The resulting UMLS OODB schema enhances the Semantic Network by adding more layers and providing more classification refinment than available in the Semantic Network. The UMLS OODB schema also supports a fast two level traversal of META, and the comprehension of META, by partitioning it into semanti-

Level	# of semantic type classes	# of intersection classes	Total
1	1	0	1
2	2	0	2
3	4	0	4
4	20	0	20
5	41	56	97
6	23	203	226
7	23	163	186
8	17	186	203
9	2	234	236
10	0	212	212
11	0	89	89
12	0	16	16
13	0	3	3
14	0	1	1

Figure 3: Number of classes in each level of the UMLS schema using intersection classes

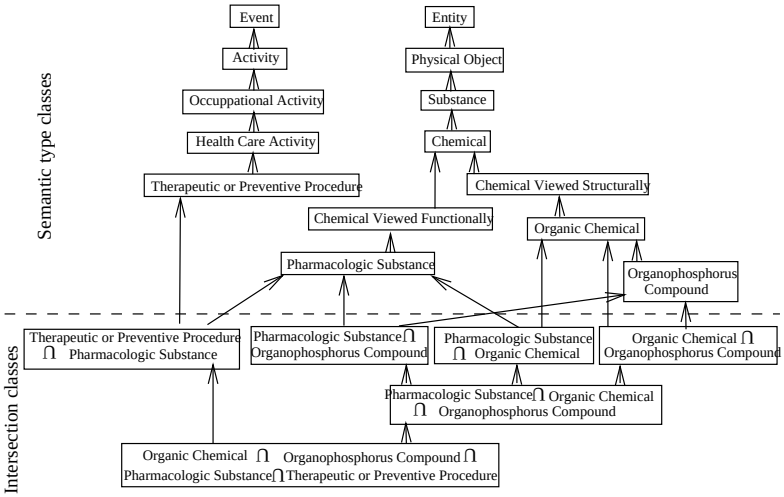

Figure 4: A subschema of the UMLS schema

cally uniform classes. The latter, in turn, leads to the recognition of errors in semantic type classifications which should be corrected.

Acknowledgments

We thank Dr. Alexa McCray from NLM, whose question inspired us to study the inherent advantages of introducing intersection classes into the OODB modeling of the UMLS. This research was done under a cooperative agreement between the NIST ATP (under the HIIT contract #70NANB5H1011) and HOST, Inc. consortium.

References

1. US Dept. of Health and Human Services, NIH, National Library of Medicine. *Unified Medical Language System*, 1998.
2. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11, 1998.
3. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods of Information in Medicine*, 32:281–291, 1993.
4. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In *Proc. Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475–480, Washington, DC, November 1989.
5. Liu L, Halper M, Gu H, Geller J, Perl Y. Modeling a vocabulary in an object-oriented database. In Barker K, Özsu MT, editors, *CIKM-96, Proc. 5th Int'l Conference on Information and Knowledge Management*, pages 179–188, Rockville, MD, 1996.
6. Liu L, Halper M, Geller J, Perl Y. Controlled vocabularies in OODBs: Modeling issues and implementation. *Distributed and Parallel Databases*, 7(1):37–65, January 1999.
7. Gu H, Cimino JJ, Halper M, Geller J, Perl Y. Utilizing OODB schema modeling for vocabulary management. In Cimino JJ, editor, *Proc. '96 AMIA Annual Fall Symposium*, pages 274–278, Washington,DC, October 1996.
8. Gu H, Halper M, Geller J, Perl Y. Benefits of an OODB representation for controlled medical terminologies. To appear in *JAMIA*, July/Aug. 1999.
9. Cimino JJ, Clayton PD, Hripcsak G, Johnson S. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA*, 1(1):35–50, 1994.
10. McCray AT. Representing biomedical knowledge in the UMLS semantic network. In Broering NC, editor, *High-performance medical libraries: advances in information management for the virtual era.*, pages 45–55. Meckler, Westport, CT, 1993.
11. McCray AT. UMLS semantic network. In *Proceedings of the Thirtennth Annual SCAMC*, pages 503–507, 1989.
12. McCray AT, Hole WT. The scope and structure of the first version of the UMLS semantic network. In *Proceedings of the Fourteenth Annual SCAMC*, pages 126–130, 1990.
13. McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Methods of Information in Medicine*, 34:193–201, 1995.
14. Gu H, Perl Y, Geller J, Halper M, Singh M. A methodology for partitioning a vocabulary hierarchy into trees. *Artificial Intelligence in Medicine*, 15(1):77–98, 1999.