

Comprehending the Structure of a Medical Vocabulary using Object-Oriented Database Modeling*

Michael Halper¹, Huanying Gu², James J. Cimino³, James Geller², Yehoshua Perl²

¹Dept. of Mathematics & Computer Science
Kean College of New Jersey
Union, NJ 07083 USA
mhalper@turbo.kean.edu
<http://www.kean.edu/~mhalper>

²CIS Dept. & CMS
NJIT
Newark, NJ 07102 USA
{helen, geller, perl}@homer.njit.edu

³Dept. of Medical Informatics
Columbia University
New York, NY 10032 USA
James.Cimino@columbia.edu

Abstract

The purpose of this paper is to demonstrate how the transformation of a medical vocabulary based on a Semantic Network (SN) model into a vocabulary based on an Object-Oriented Database (OODB) model helps in the maintenance of the vocabulary. We describe an OODB schema which captures the overall structure of the vocabulary in a compact form and uncovers some errors and inconsistencies made in the vocabulary's original modeling. The resolution of these mistakes leads to an improved version of the SN-based vocabulary. A new OODB schema for the vocabulary is then derived based on the improved SN version. This experience demonstrates how the abstraction and modeling capabilities of OODBs can be used to enhance a user's understanding of the overarching structure of a complex medical vocabulary system. The OODB schema developed herein serves as the basis for the Object-Oriented Healthcare Vocabulary Repository (OOHVR), a medical vocabulary implemented as an ONTOS database.

1 Introduction

In this paper, we report on an experience of mapping a large controlled medical vocabulary based on a Semantic Network (SN) model [17] into an Object-Oriented Database (OODB) system [2, 11, 18]. There are two major issues addressed in this work. The first issue is the proper way of modeling a vocabulary using the OODB paradigm. The second issue is the impact of the OODB modeling on the maintenance of the original vocabulary.

*This research was (partially) done under a cooperative agreement between the National Institute of Standards and Technology Advanced Technology Program (under the HIIT contract #70NANB5H1011) and the Healthcare Open Systems and Trials, Inc. consortium. Dr. Cimino's work was supported in part by the National Library of Medicine and the IBM Corporation.

In Section 2, we describe the MED (Medical Entities Dictionary) [3, 4, 5, 6, 7], a large controlled vocabulary system designed at Columbia-Presbyterian Medical Center (CPMC). In the OOHVR (Object-Oriented Healthcare Vocabulary Repository) project, we are transforming the content of the MED into the OODB framework.

In Section 3, we present the OOHVR, an OODB that captures the entire contents of the MED. Section 3.1 describes the details of the approach used to develop the OOHVR's (OODB) schema. This approach utilizes from one side the given properties of the MED vocabulary, and from the other side the expressive power of OODB modeling. The resulting OODB schema is a very compact representation of the structure of the MED. In Section 3.2, we observe some shortcomings in this schema, and then we further extend our approach to obtain a more detailed OODB schema which fully captures the MED.

In Section 4, we examine the structure of the MED in view of the OODB schema we obtained. Section 4.1 discusses how the schema allowed us to gain further insight into the MED's content. In Section 4.2, we show how the enhanced comprehension obtained from the OODB schema led to improvements in the MED's organizational structure. We then go on in Section 4.3 to discuss how a partial OODB schema helped us uncover some inconsistencies and errors in the original modeling of the MED. Reviewing these cases suggested the way to resolve the problems found in the conceptual design. Finally, we utilize the corrected MED to re-model the partial OODB schema in order to correctly capture the structure of the MED in a compact way. Section 5 contains conclusions.

The OODB schema presented in this paper has served us in the implementation of the OOHVR within ONTOS, a commercial OODB system. The OOHVR project is part of the HIIT (Healthcare Information Infrastructure and Technology) project supported by NIST (National Institute of Standards and Technology) under the ATP (Advanced Technology Program). The work is being carried out at NJIT as part of the HOST (Healthcare Open Systems and Trials) consortium.

2 Description of the MED

The Medical Entities Dictionary (MED) is a semantic network of medical concepts which includes a concept subsumption hierarchy, a directed acyclic graph (DAG) composed of concepts connected through super-concept (and sub-concept) links. Within the network, concepts are defined through the use of hierarchical and non-hierarchical relationships. Each concept can also have attributes that contain primitive information, such as additional names (synonyms) and other codes used in various systems for the concept. The super-concept links allow for multiple parent concepts and provide the means for the inheritance of properties (i.e., attributes and relationships). The non-hierarchical links form a medical knowledge-base which can be used for the maintenance of the MED itself. For example, an automated classification tool for new concepts might operate based on the medical knowledge in the MED and the medical knowledge about the concept being added.

The MED structure was designed to allow the MED content to meet seven basic requirements:

1. Domain completeness – the structure cannot limit the number of concepts in the MED, the number of concepts which can appear at any one level in the hierarchy, the depth of the hierarchy, the number of parents a concept may have, or any other aspect of organizational complexity which may be desired.
2. Synonymy – concepts must be recognized by multiple names.
3. Nonvagueness – each concept must have a well-formed meaning.
4. Nonredundancy – no two concepts may have the same meaning.
5. Nonambiguity – each concept may have no more than one meaning.
6. Multiple classification – concepts may have more than one super-concept.
7. Consistency of views – a concept should appear the same (including having the same properties and children) no matter how the concept is arrived at in the hierarchy.

The structural characteristics of the MED as a semantic network differentiate it sharply from other controlled vocabularies, such as SNOMED [8], ICD9-CM [16], and MeSH [14]. In particular, the MED structure accommodates sophisticated editing tools for merging controlled vocabularies

and other maintenance tasks. The result is a controlled vocabulary that is well suited for supporting integration of clinical information from multiple sources and using it for multiple purposes.

The MED content includes the UMLS Semantic Network [12] and has many structural similarities to the UMLS Metathesaurus. The Metathesaurus maintains much more distinct, vocabulary-source-specific information, which is condensed in the MED into single-concept frames, while the MED makes much more use of relationships among concepts and has a single, coherent subsumption hierarchy. In addition to the UMLS Semantic Network, the MED contains 15,000 ICD9-CM terms [16], some 15,000 additional ICD9-CM “index” terms (of which 10,000 are unique concepts and 5,000 are synonyms of these concepts), the American Hospital Formulary Service classification [1], and controlled concepts used in CPMC ancillary systems such as laboratory, pharmacy, and radiology. In all, there are 43,000 concepts in the MED, of which 16,000 are linked to Metathesaurus concepts.

An example of how the MED helps with the maintenance and use of patient data can be found by looking at the concept **Serum Glucose Test (SGT)**. This concept is recognizable in virtually all clinical laboratory systems, although it may have many names: **Glucose**, **Blood Sugar**, **GLUC**, **BG**, etc. In fact, many laboratory systems will have multiple concepts for this general term, reflecting the fact that different tests are done on different machines or as parts of different panels. In the MED, **SGT** is represented as a **Single-Result Laboratory Test** which causes it to inherit relationships such as *substance_measured* and *specimen*. For **SGT**, these two properties refer to the MED concepts **Glucose** and **Serum**, respectively.

In a typical maintenance task, a lab system may add a new test, say, **Stat Blood Sugar (SBS)**. Lexical matching (as is done in the UMLS) would fail to find the relevant MED concept **SGT**. However, by indicating that **SBS** is a sub-concept of **Single-Value Lab Test**, it acquires the relationships *substance_measured* and *specimen*. The lab system may have information about

the specimen and a vocabulary editor could add information about the substance measured. Once this is done, the MED editor functions can identify **SGT** as a possible synonym for **SBS**. The user must decide if it is a synonym; if it is, its name and code are added to the information about **SGT** in the MED. Otherwise, the term may be a more specific concept, in which case it is added as a new concept, namely, as a child of **SGT**.

Typical use of patient data would include a summary of lab results, such as all serum glucose results. A program that included a hard-coded query for **SGT** results would automatically include **SBS** in the query, without the system developer knowing that a new lab test had been added. The MED would simply return a list of codes to be used for searching the clinical database. When the new term is added, its code is included along with all the pre-existing codes. Multiple use of the data is supported by the MED through the use of multiple concept levels (for example, the concept **Intravascular Glucose Test** subsumes **SGT** and its descendants and also subsumes **Plasma Glucose Test** and **Whole Blood Glucose Test**—the use of such a higher-level concept might be more relevant to the clinician). The MED’s “multiple super-concept” capability is also very useful in the manipulation of patient data. For example, while the clinician might access **SBS** through the clinically relevant concept **Intravascular Glucose Test**, someone interested in laboratory quality assurance might access it alternatively via the concept **Stat Lab Test** or **Serum Tests**, both of which are not subsumed by **Intravascular Glucose Test**. Finally, non-hierarchical information can be used in this regard as well: “Get me all the tests that measure glucose.”

3 The OOHVR

3.1 OOHVR Schema

As was noted above, the OOHVR is a controlled medical vocabulary stored in an OODB. At bottom, it can be viewed as a large database of “concept” (or dictionary entity) objects. In this

section, we describe the schema of this database and provide some insight into its development.

The OOHVR was conceived as the target of a mapping of an existing medical vocabulary into an OODB. The MED of CPMC was chosen as its source. The question we faced was how to model the MED, a semantic network of concepts and links, using the available constructs of an OODB schema. After analyzing the MED in detail, we decided that the best approach to mapping it into an OODB schema is based on the underlying pattern in which its properties are introduced. (Other modeling alternatives are discussed in [9, 13].) For each property there is a unique concept **C** for which this property is first introduced. By inheritance, this property is also defined exactly for all the descendant concepts of **C**.

The process of defining the OOHVR schema's classes began with the partitioning of the MED into groups of concepts such that all concepts in a single group have the exact same properties. Such a group of concepts is called an *area* [13]. Because of the inheritance pattern of the MED, areas turn out to be rooted sub-hierarchies of the MED's overall sub-concept/super-concept hierarchy. As an example, the concept **Measurable Entity** introduces a new relationship *measured-by* and is thus the root of a new area. All concepts below **Measurable Entity** in the hierarchy that do not introduce properties (and, therefore, by inheritance have the exact same properties as **Measurable Entity**) are in this "Measurable Entity" area. Some examples of such concepts are **Color**, **Temperature**, **Specific Gravity**, **Viscosity**, **Osmolality**, **Blood Coagulation**, and **Optical Density**. If a concept is below **Measurable Entity** and does introduce properties, then it is not in **Measurable Entity**'s area but rather in a new area of which it is the root.

After identifying the MED's areas, we defined the OOHVR schema as follows. Each area in the MED is modeled as its own object class called an *area class*. The name of the area class is the concatenation of the name of the area's root concept and "_Area." The "Measurable Entity" area described above would have the corresponding area class *Measurable_Entity_Area*. The properties

defined for the area class in the OOHVR schema are exactly those introduced by the area's root in the MED. For the class *Measurable_Entity_Area*, these would be the relationship *measured-by* and any other properties introduced by the concept **Measurable Entity**. All concepts in an area, including the root concept, become instances of the corresponding area class in the OOHVR.

The MED contains one special concept called **Medical Entity** that is the root for all concepts. In other words, each concept in the MED is a descendant of the **Medical Entity** concept. Thus, the root of any area in the MED is a child of a node(s) in some other area(s). The exception to this is the **Medical Entity** concept itself. Its area will be called the *root area*, and the corresponding area class is *Medical_Entity_Area*. It should be noted that the root of an area has all the properties of its parents' areas plus the properties that are defined explicitly for it. All the other nodes in an area, of course, have these same properties. To capture this in our model, we place each area class corresponding to a root node in an IS-A (subclass) relationship with respect to the area class(es) of its parent(s). As we have alluded to, a node may have more than one parent and the IS-A hierarchy induced by this process is not necessarily a tree, as it may exhibit multiple inheritance. The area class *Medical_Entity_Area* that corresponds to the root area of the MED serves as the root of the OOVHR's schema.

The OODB schema produced by this mapping turns out to be very compact in terms of the number of classes, particularly when one considers that the MED contains thousands of concepts. This compactness results from the fact that the total number of properties associated with concepts in the MED is 150. This implies that there are at most 150 concepts, out of the 43,000 in the entire vocabulary, where a new property is defined, and, therefore, at most 150 areas. In fact, according to the process described above, the MED's 43,000 concepts are partitioned into only 53 areas, because some concepts introduce multiple properties. Therefore, the OOHVR schema consists of only 53 area classes. Clearly, most nodes in the MED sub-concept/super-concept hierarchy do

not introduce properties. Contrast this with the IS-A hierarchy of a typical OODB schema where almost all classes define new properties.

In Figure 1 (fold out), we present the OOHVR schema obtained via the above described mapping. The schema is presented using our own OOdini graphical notation [10]. In OOdini, a class is represented as a rectangle, and a relationship, as a labeled thin arrow. We denote an IS-A (subclass) relationship as a bold arrow directed from the subclass to its superclass. An attribute appears as an ellipse connected to its class via a thin line. Due to lack of space, attribute and relationship names (which can be long) are encoded with numbers in the diagrams. For example, attribute “9” is *lab-procedure-code*; relationship “18” is *result-of-tests*.

The MED’s concept subsumption hierarchy was the foundation of the mapping into the OOHVR schema. The mapping consisted of two major aspects: (1) the identification of the “property introduction” concepts, and (2) a collapsing of the inheritance paths between these concepts. As such, the OOHVR schema is an abstraction of the property definitions and inheritance that occur within the MED, which is a semantic network. For this reason, we call the OOHVR schema a *network abstraction schema* [13].

Even though the MED’s concept subsumption hierarchy was collapsed to produce the OOVHR’s IS-A hierarchy, which completely captures the inheritance behavior of the original network, there is still a need for the concept subsumption hierarchy to appear entirely within the OOHVR. This, for example, allows one to perform subsumption-based reasoning, a common request to the vocabulary. To include the hierarchy, two reflexive relationships are defined at the root area class *Medical_Entity_Area*: *has_superconcepts* and *has_subconcepts*. In the MED, if **X** is a subconcept of **Y**, then in the OOHVR the object corresponding to **Y** is a referent of **X** with respect to the *has_superconcepts* relationship; *has_subconcepts* is the converse. In this manner, all instances in the OOHVR keep track of their super- and sub-concepts. In other words, the concept subsumption

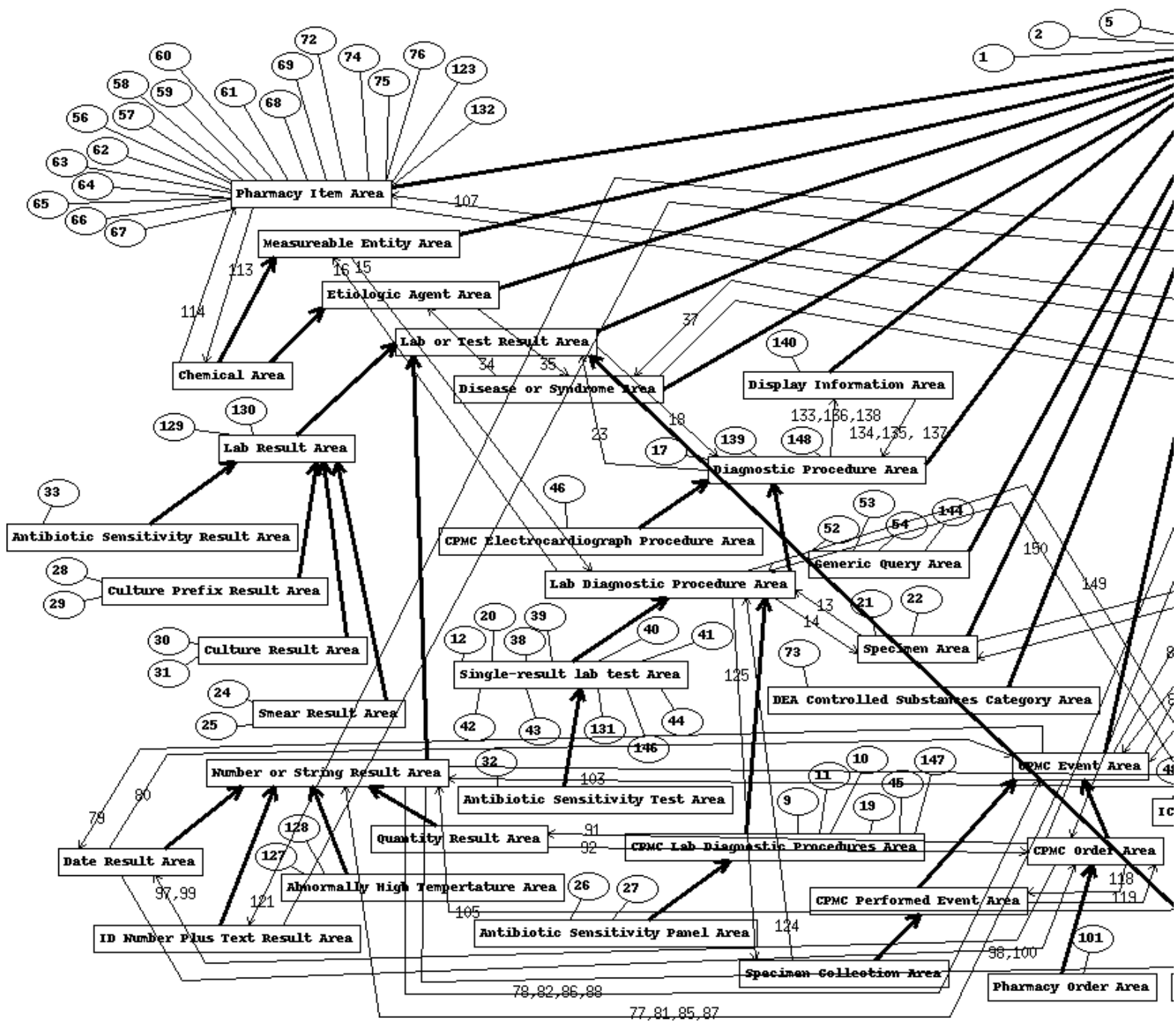


Figure 1: OOHVR Schema

hierarchy of the MED is represented in the OOHVR on the instance (object) level rather than on the schema level.

3.2 OOHVR Schema with Intersection Area Classes

One flaw in the above mapping arises because of the possibility of multiple inheritance in the MED's concept subsumption hierarchy. (Recall that it is a DAG, not a tree.) Let a concept, say, **X** that does not introduce any new properties reside in two unrelated areas, say, *A* and *B*. By unrelated areas we mean that neither *A*'s area class nor *B*'s is a descendant of the other in the OOHVR schema's hierarchy. This situation occurs when the concept **X** needs the properties introduced at both class *A_Area* and class *B_Area*. As a matter of fact, we may have a set *C* of concepts which have the same properties as concept **X**. Actually, *C* is exactly the intersection of the two areas *A* and *B*. Some examples of concepts in the intersection of two areas are **Antihistamine Drugs**, **Anti-Infective Agents**, and **Anti-Inflammatory Agents**. These three all reside in the two areas "Pharmacy Item" and "American Hospital Formulary Serv. Class."

According to the mapping described above, **X**'s membership in the two areas implies that the object corresponding to it in the OOHVR must be an instance of the two separate area classes. However, such an arrangement is forbidden in a conventional OODB model. Thus, we need to modify the mapping in order to accommodate this scenario, which, as it happens, occurs infrequently within the MED.

Our solution is to extend the notion of area and define the intersection of the two areas as an area of its own. This new kind of area is called an *intersection area*. As with all other areas, a class is defined for it in the OOHVR schema. It should be noted that this class does *not* introduce any new properties. It serves instead as the intersection of two separate branches in the OOHVR's IS-A hierarchy. Clearly, such an area class will be a subclass of two other classes. Let us note

that the notion of intersection area can be extended to encompass the intersection of three or more unrelated areas. One such area in the MED is “Chloramphenicol Prep.” which is the intersection of three areas “Antihistamine Drugs,” “DEA Controlled Subst. Category,” and “Drug Allergy Class.”

In the above example, we introduce an intersection area class C_Area as a child of the area classes A_Area and B_Area in the extended OOHVR schema. That schema will include all intersection area classes and their IS-A relationships to other area classes, which themselves may be intersection area classes. The concept \mathbf{X} and the other concepts in the set C will be instances of the intersection area class C_Area . Note that the set C may or may not have a root (i.e., a concept which is an ancestor of all other concepts in C). If, for example, \mathbf{X} is a root for C , then the corresponding intersection area class for C will naturally be denoted X_Area . Otherwise, the schema designer has to arbitrarily select one of the concepts of C , say, the first one to appear in the MED, as the name of the intersection area class.

In Figure 2, we present the IS-A hierarchy of the schema in Figure 1. All attributes and relationships have been omitted. (We call this an OOdini Level 3 display [10, 15].) Note that the suffix “_Area” has been left off area class names in order to save space. We have included this figure as a comparison to Figure 3 which presents an excerpt of the revised OOHVR schema that includes intersection area classes.

Referring to Figure 3, we see that intersection area class $Antihistamine_Drugs_Area$ is a child of two area classes $American_Hospital_Formulary_Service_Class_Area$ and $Pharmacy_Items_Area$ in the OOHVR schema. The class $Antihistamine_Drugs_Area$ itself is a parent of another intersection area class $Aminoglycoside_Preparations_Area$ which is also a child of the area class $Drug_Allergy_Class_Area$.

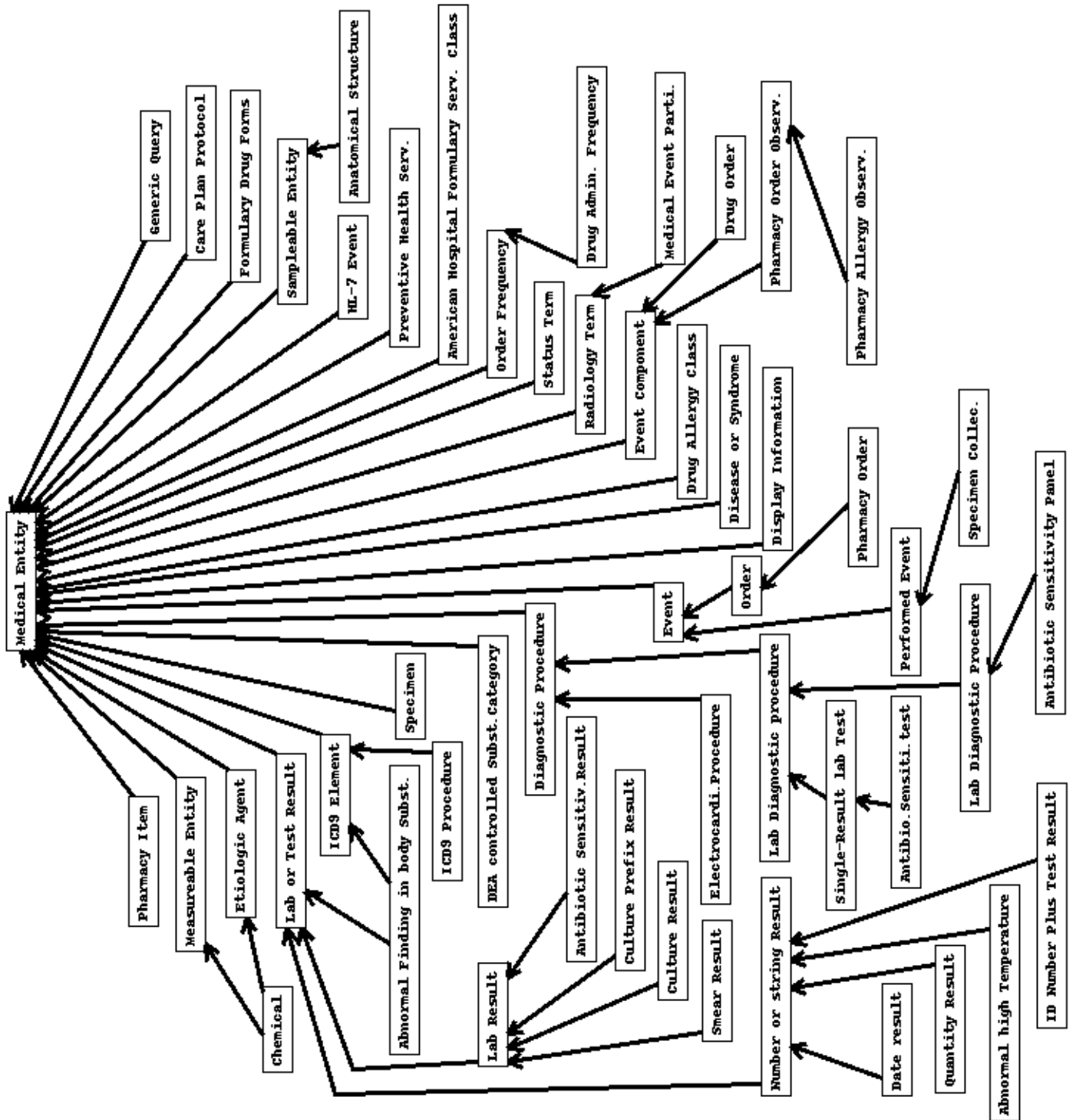


Figure 2: IS-A hierarchy of OOHVR schema

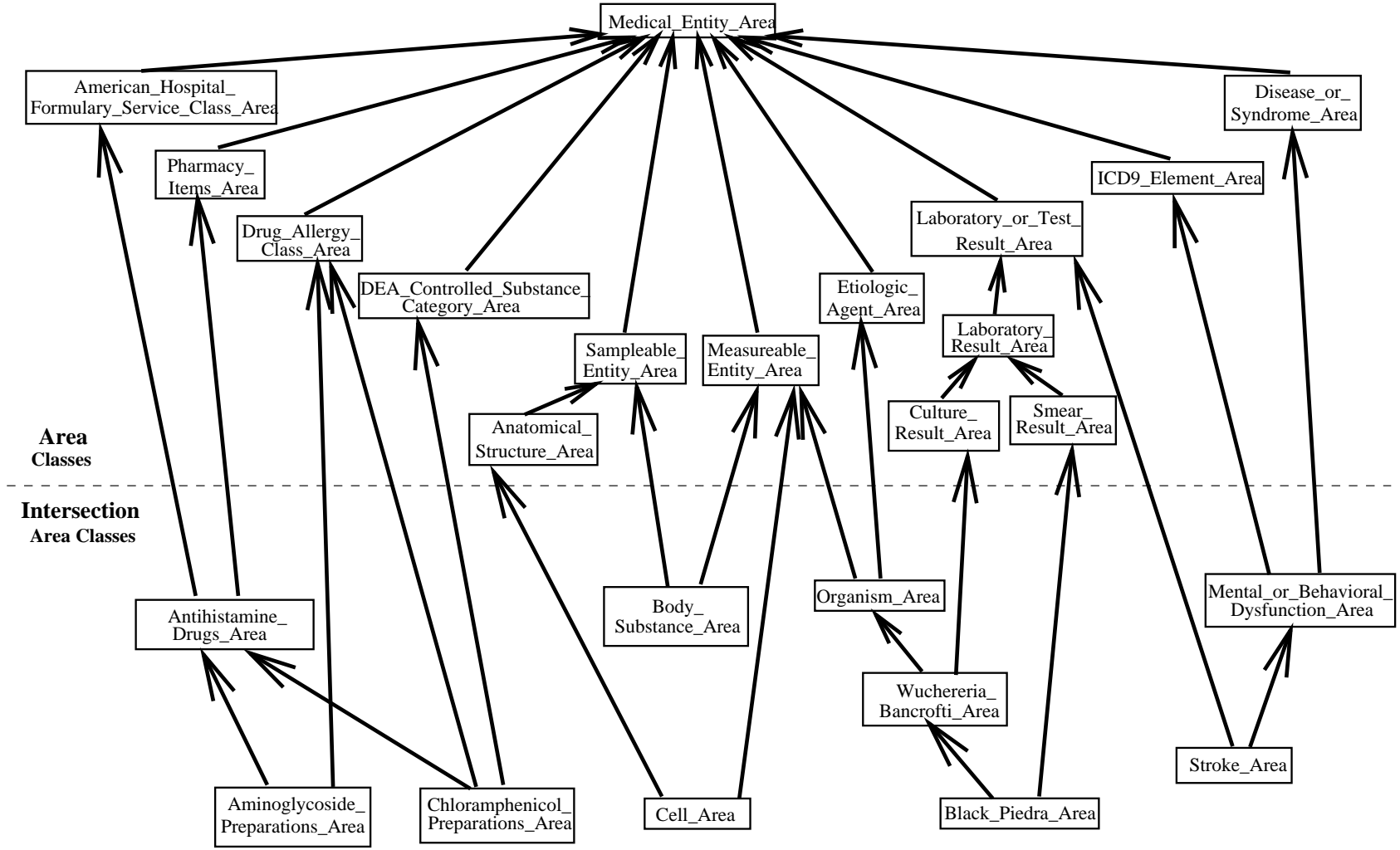


Figure 3: Excerpt from revised OOHVR schema with intersection area classes

4 Corrections to the MED based on the OOHVR Schema

The development of specialized views, such as network abstraction schemas and intersection areas, is of more than theoretical interest. The maintenance of the MED at CPMC is a complex and difficult task and no commercial tools are suitable to support this work. Browsers and editors have been developed, and continue to be developed, but providing users of the MED (both those whose job it is to maintain it and those who would build applications or knowledge bases using it) with comprehensible, comprehensive views remains difficult. Some of the challenges of maintaining and using the MED include understanding the MED schema, improving the MED organizational structure, finding and correcting inconsistencies and errors in the MED content. In particular, the latter is a crucial issue given the size and scope of the MED.

4.1 Understanding the MED Schema

With over 43,000 concepts in the MED, 88 attributes, 62 relationships, divided into 31 pairs of reciprocal relationships, 55,000 hierarchical links and 96,000 non-hierarchical links, understanding the “big picture” is difficult. When new concepts are to be added to the MED, or when someone needs to find appropriate concepts in the MED, this lack of understanding becomes immediately apparent. The situation is often worsened because those people who maintain and use the MED may not be the same people who modeled a particular domain within the MED. For example, the laboratory system at CPMC has terms¹ for individual tests (such as Serum Glucose Test) and other terms for orderable collections of tests (such as a CHEM-7, a panel of 7 individual tests). These terms are all represented as concepts in the MED with attributes appropriate to each (tests have units of measure and normal ranges, while panels have codes used for billing). The concepts are

¹We use the word “term” to refer to entities in vocabularies which are sources for the MED, while we use the word “concept” for entities in the MED. This is to emphasize the fact that the MED entities are based on meaning and, therefore, multiple external terms (from the same or different sources) may map to a single MED concept.

related to each other (**Tests** are *part-of* **Panels**) and with other terms in the MED (**Tests** *measure* **Measurable Substances**). Users of the MED are often confused about the differences between tests and panels (the latter are also called “procedures” by some and “batteries” by others). The confusion is worsened at times because individual tests can be ordered separately and therefore can take on the characteristics of both tests and panels.

Some ability to provide users with a manageable high-level view of the MED is needed. The network abstraction schema, as shown in Figure 1, provides such a view. By reducing the MED hierarchy almost 1000-fold, one can immediately see what the important areas (or area classes in the OOHVR schema) are and what their attributes and relationships are. Someone looking to add a new concept to the MED can easily review these areas and narrow them down to a handful of interesting areas. Once narrowed down, the user can then review the attributes and relationships of specific areas to determine the appropriate area for the new concept. For example, a user faced with the task of adding a new laboratory panel to the MED can quickly see (from Figure 1) that the relevant area is likely to be found as a descendant of the “Lab Diagnostic Procedure” area. Under this area, there are four candidate areas. Reviewing the names allows the user to exclude the “Antibiotic Sensitivity Test” area and “Antibiotic Sensitivity Panel” area, leaving the “Single Result Lab Test” area and “CPMC Lab Diagnostic Procedure” area. Reviewing the attributes of these areas reveals that the latter has an attribute encoded as “9” and having the name *lab-procedure-code*; since the concept to be entered is known by the user to have such a code, this area is clearly the appropriate one. Thus, the network abstraction schema can be shown to provide a valuable gestalt of the MED complexity.

The intersection area classes are a helpful addition to the network abstraction schema. They provide the specific examples of how such complex interactions occur between areas. Let us refer to Figure 4 which shows an excerpt of the revised OOHVR schema containing intersection area

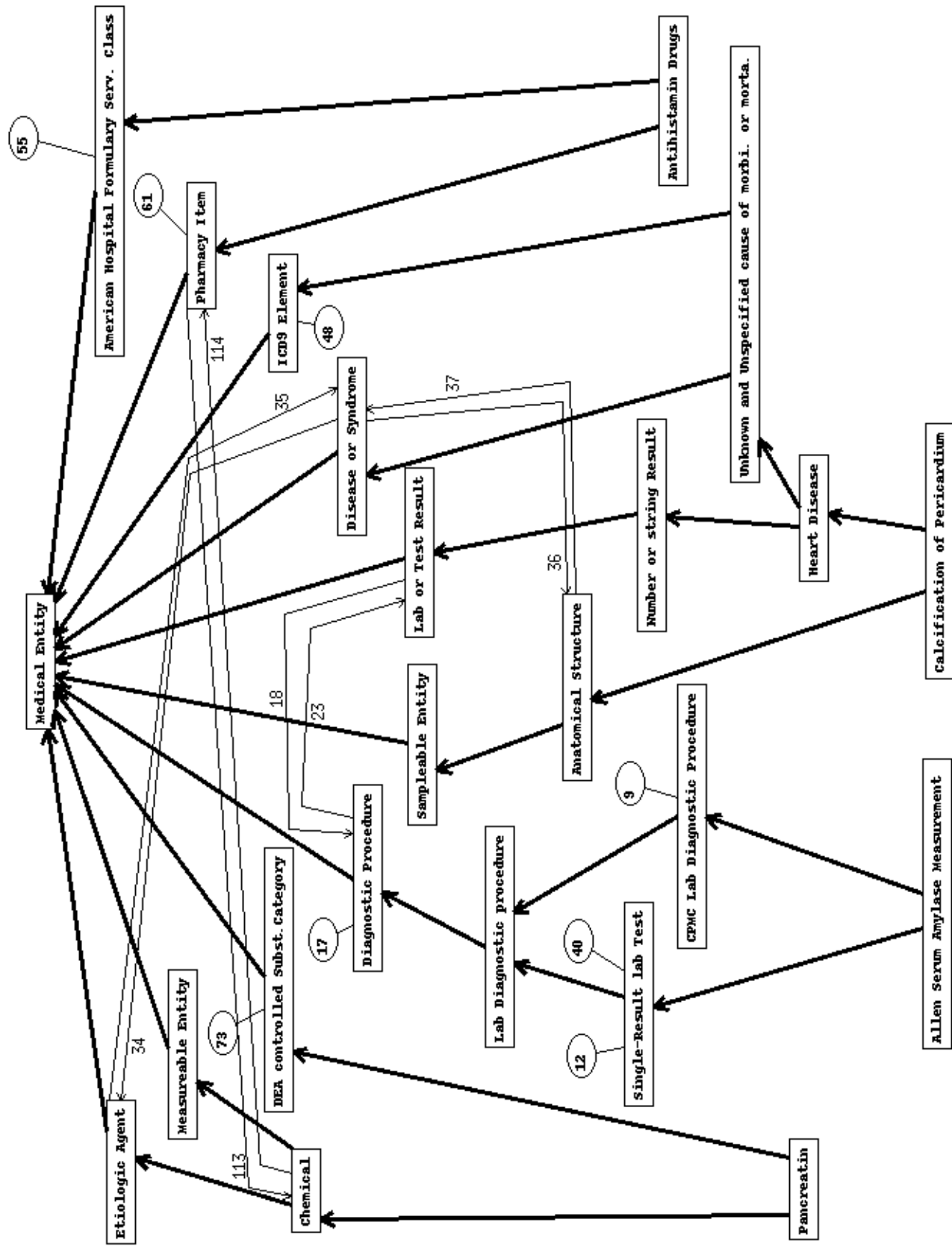


Figure 4: Portion of OOHVR schema including intersection area classes

classes. Note that this partial schema includes some properties. In the figure, we see that the “Antihistamine” area is a descendant of both the “Pharmacy Item” area and the “American Hospital Formulary Service Class” area. This is perfectly appropriate since the concepts corresponding to medications are classified in multiple ways in the MED, and inherit different attributes from each of the parent areas.

4.2 Improving MED Organizational Structure

Over the past seven years, the MED content has grown at an average rate of 500 concepts per month. Much of this growth has been the result of work by a variety of individuals and sometimes through the use of automated mechanisms for adding concepts. As a result, it is not surprising that inconsistencies and out-right errors have crept into the MED. When two people share the task of maintaining a content domain but have slightly different organizational philosophies (e.g., “lumpers” versus “splitters”), it is easy for concepts to be characterized differently depending on who added them to the MED. The network abstraction schema provides a way for two people to share the same view of the concepts they are dealing with and to identify differences in their views. Finally, it provides ways for the MED organization to be made simpler for both parties to follow.

One example of this relates to the aforementioned fact that individual tests can be ordered separately and therefore have attributes of both tests and panels. The “Allen Serum Amylase Measurement” area, shown in Figure 4, consists of the various tests that all have the properties of panels. Once we were able to see the intersection areas, it became clear that the MED could be simplified by making explicit the fact that many such tests exist. We did so by creating a concept called **Orderable Test**. This concept was added as a descendant of both **Single-Result Lab Test** and **CPMC Laboratory Diagnostic Procedure**; all orderable tests (such as **Allen Serum Amylase Measurement Test**) were then made descendants of this concept. Since this concept

now subsumes the set of all concepts with the union of attributes from the two parent concepts, **Orderable Test** now becomes the root node in the new “Orderable Test” area (Figure 5).

From the above, we derived a general rule for dealing with intersection areas in the OOHVR schema. As we described in Section 3.3, we pick an arbitrary concept to name the area class in the case that the intersection area has no root. Instead, we will now try to create a new parent concept for all the concepts in the area. This new area root will then be used to name the area class.

4.3 Finding Inconsistencies and Errors in the MED Content

Given the ambiguity that often occurs in medical terminology, it is easy for the MED to contain a concept with a name that has multiple meanings. Since the inception of the MED model [5], it was thought that such ambiguity could be detected through automated means. The use of intersection areas has provided such a method.

In one example, we could see from the schema in Figure 4 that the “Calcification of the Pericardium” area contains all concepts which are both heart diseases and anatomical structures. It seemed strange that the same concept would be both a disease and an anatomical structure at the same time. Thus, one or the other of the parent-child links should be removed from the MED. Upon closer inspection of the “Calcification of the Pericardium” area, we found that there were many such “Calcification of the X” concepts in the MED, all of which are included as descendants of **Calcification of Body Part**. This concept is a child of **Body Part**, and both are in the “Anatomical Structure” area. **Calcification of Body Part** has 40 of its children (and three grandchildren) that are also classified as diseases, while other children are not. In order to improve consistency for the way such terms are classified, we removed the links between these “calcification” concepts and their “disease” parents. So, for example, the link between **Calcification of the Pericardium** and **Heart Disease** was removed (see Figure 5). Doing this, caused **Calcification**

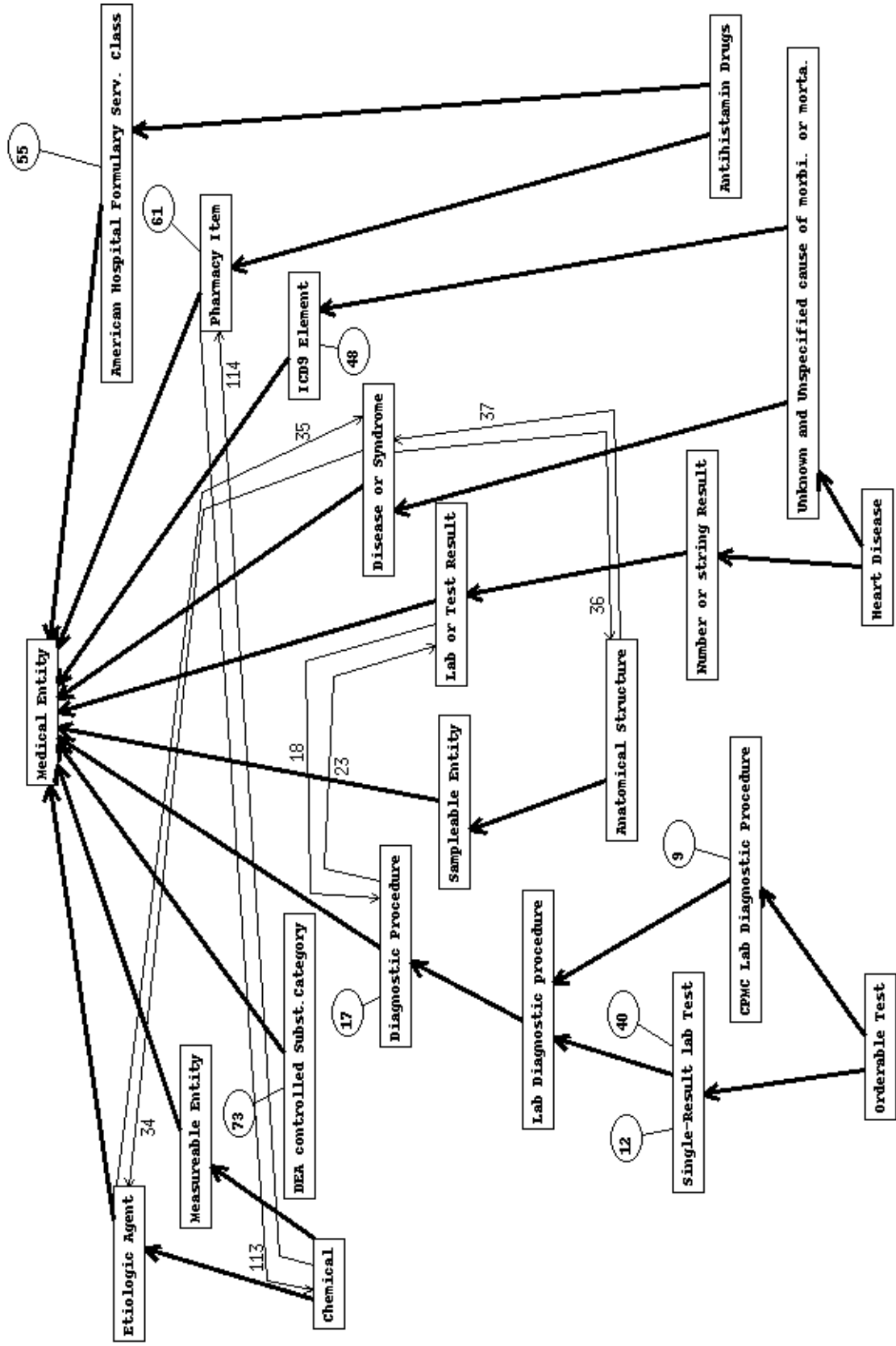


Figure 5: Improved version of schema from Figure 4

of the Pericardium to be in a single area, namely, the “Anatomical Structure” area. Therefore, it no longer defined an intersection area of its own. For this reason, the corresponding area class does not appear in Figure 5.

In another example, we examined the “Pancreatin” intersection area. In the MED, we have determined that medications (such as those classified by their DEA Controlled Substance category) would have *pharmaceutic-components* which are chemicals but that the medications would not themselves *be* chemicals. The intersection area graph clearly shows that **Pancreatin** violates this rule. On closer inspection, we found that the concept **Pancreatin Preparations** was properly classified as a medication and that it was linked appropriately to the concept **Pancreatin**. However, the concept **Pancreatin** was classified not only as a chemical (which allowed it to have the *pharmaceutic-component-of* relationship to **Pancreatin Preparations**), but it was also classified as a medication. Once we saw this error, we were able to correct it quite easily by removing the link between **Pancreatin** and its **DEA Controlled Substance** parent. Since **Pancreatin** was the only concept in the MED to have attributes of both chemicals and medications, the “Pancreatin” Area had only one concept prior to the correction. After the correction, the area no longer existed, since the concept **Pancreatin** was now included in the “Chemical” Area (Figure 5).

5 Summary

The maintenance of a large controlled vocabulary (e.g., the MED of CPMC) is a complex and difficult task. The difficulty arises from the need to comprehend the extensive network of concepts and semantic links that forms the vocabulary. In this paper, we have reported on an experience of bringing OODB modeling to bear on the task of comprehending and maintaining a medical vocabulary. In particular, we have described a technique for mapping an existing semantic network-based vocabulary into an equivalent OODB-based vocabulary, which we call the OOHVR. The

schema of the OOHVR turns out to be a very compact representation with respect to the size and scope of the original vocabulary. Because of this, it offers insight into the overall structure of the vocabulary and greatly aids in its comprehension and maintenance. In fact, we described how the schema was utilized to uncover and correct some errors and inconsistencies that had existed in the MED. The OOHVR has been implemented as an ONTOS database and is currently up and running at NJIT.

References

- [1] American Society of Hospital Pharmacists, Bethesda, MD. *American Hospital Formulary Service Drug Information*. Updated annually.
- [2] E. Bertino and L. Martino. *Object-Oriented Database Systems Concepts and Architectures*. Addison-Wesley Publishing Company, 1993.
- [3] J. Cimino and G. Barnett. Automated translation between medical terminologies using semantic definitions. *MD Comput.*, 7:104–109, 1990.
- [4] J. Cimino, P. Clayton, G. Hripcsak, and S. Johnson. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA*, 1(1):35–50, 1994.
- [5] J. Cimino, G. Hripcsak, S. Johnson, and P. Clayton. Designing an introspective, controlled medical vocabulary. In *Kingsland LC, ed. Proceedings of the Thirteenth Annual SCAMC*, pages 513–518, Washington, DC, 1989. IEEE Computer Society Press.
- [6] J. Cimino, G. Hripcsak, S. Johnson, C. Friedman, and P. Clayton. Prototyping a vocabulary management system in an object-oriented environment. In *Proceedings of the International Medical Informatics Association Working Conference on Software Engineering in Medical Informatics*, pages 429–439, Amsterdam, The Netherlands, 1991. North-Holland Press.
- [7] J. Cimino, G. Hripcsak, S. Johnson, C. Friedman, D. Fink, and P. Clayton. UMLS as knowledge base - a rule-based expert system approach to controlled medical vocabulary management. In *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*, pages 175–179, CA, 1990. IEEE Computer Society Press.
- [8] College of American Pathologists, Skokie, IL. *Systematized Nomenclature of Medicine. Second edition*, 1982.
- [9] H. Gu, J. Cimino, M. Halper, J. Geller, and Y. Perl. Utilizing OODB schema modeling for vocabulary management. To appear in Proc. 1996 AMIA Annual Fall Symposium.
- [10] M. Halper, J. Geller, Y. Perl, and E. J. Neuhold. A graphical schema representation for object-oriented database. In R. Cooper, editor, *Workshop on Interfaces in Database Systems (IDS-92)*, pages 282–307. Springer Verlag, London, 1993.

- [11] W. Kim and F. H. Lochovsky, editors. *Object-Oriented Concepts, Databases, and Applications*. ACM Press, New York, NY, 1989.
- [12] D. Lindberg and B. Humphreys. Toward a unified medical language system. In *Proceedings of the Seventh International Congress*, pages 23–31, Berlin, Germany, 1987.
- [13] L. Liu, M. Halper, H. Gu, J. Geller, and Y. Perl. Modeling a vocabulary in an object-oriented database. To appear in CIKM-96, 1996.
- [14] National Library of Medicine, Bethesda, MD. *Medical Subject Headings*. Updated annually.
- [15] Y. Perl, J. Geller, and H. Gu. Identifying a forest hierarchy in an OODB specialization hierarchy satisfying disciplined modeling. In *Proceedings of the 1st IFCIS International Conference on Interoperable and Cooperative Systems*, pages 182–195, Brussels, Belgium, 1996.
- [16] United States National Center for Health Statistic, Washington, DC. *International Classification of Diseases, Ninth Revision, with Clinical Modifications*, 1980.
- [17] W. A. Woods. What’s in a link: Foundations for semantic networks. In D. G. Bobrow and A. M. Collins, editors, *Representation and Understanding*, pages 35–82. Academic Press, New York, NY, 1975.
- [18] S. B. Zdonik and D. Maier, editors. *Readings in Object-Oriented Database Systems*. Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1990.