Journal of the American Medical Informatics Association

JAMIA

Proceedings

Beyond the Superhighway: Exploiting the Internet with Medical Informatics

1996 AMIA ANNUAL FALL SYMPOSIUM℠

Formerly SCAMC

October 26 - 30, 1996
Sheraton Washington Hotel
Washington, DC

Edited by
James J. Cimino, M.D.

AMIA

The Official Journal of the
American Medical Informatics Association

Published by Hanley & Belfus, Inc.
ISSN 0195-4210

Preface

This *Proceedings* contains the papers and abstracts presented at the 1996 Annual Fall Symposium of the American Medical Informatics Association. Although the first meeting by this name, the Annual Fall Symposium carries on the tradition of the Symposium on Computer Applications in Medical Care (SCAMC), the premier annual medical informatics conference in the world since 1977. It is fitting that the meeting's twentieth anniversary is marked by a record number of submissions and presentations. I believe the contents of this *Proceedings* also set a record for quality, reflecting the advances in our relatively young field and the maturation of its researchers and practitioners.

This year's theme, "Beyond the Superhighway: Exploiting the Internet with Medical Informatics," is also a fitting subject for marking a milestone year. Long the domain of university researchers, the Internet and World Wide Web were fast becoming household words while the medical and medical informatics communities lagged behind. Despite routine use of the Internet by the end of the 1970s, the word "Internet" did not appear in the Medline database until 1992, when it was cited eight times. Since then, the medical informatics community has made up for lost time—the number of citations has tripled annually. By 1995, there were 206 citations that satisfied a search for "Internet," "World Wide Web," or WWW." I am happy to report that the leader that year, by far, was the SCAMC *Proceedings*, with 23 citations (The *New England Journal of Medicine* edged out *JAMIA, the Journal of the American Medical Informatics Association*, 9 to 8). This year's *Proceedings* makes an even greater contribution to the world's medical literature.

The Internet and the Web are proving to be ideal adjuncts for medical computing, breaking down the barriers to access that have prevented sharing and dissemination of systems. A survey of the past nineteen SCAMC *Proceedings* shows literally thousands of promising applications that failed to meet their full potential, undoubtedly due in part to their stand-alone, isolated nature. In this *Proceedings* you will find many examples of applications that exploit the opportunities presented by the Internet and the Web to make systems more accessible and, therefore, more usable and useful.

Conversely, medical information science has much to offer the Internet and the Web. The information overload in medicine, lamented for years, has reached new extremes and we find ourselves "drinking from a firehose" of information. Freely available Internet search engines seem to be capable of answering any question if one is willing to wade through the thousands of documents they offer. Medical informatics provides solutions; not only can we exploit the Internet for our purposes, but we can also tame it with our methods. In this *Proceedings*, you'll find examples of how information retrieval, expert systems, controlled vocabularies, clinical systems, networking, education, and cognitive science provide us with techniques that are relevant to the new challenges presented by the Internet.

Of course, medical informatics is not just about the Internet, and neither is this *Proceedings* or the symposium it represents. As in the past, you will find all areas of health care informatics represented. We defined seven tracks into which we shoehorned the diverse contributions of our authors: Training, Education, and Cognition; Policies and Standards; Expert Systems and Algorithms; Information Retrieval and Digital Libraries; Clinical Information: Storage and Use; Health Information Networks; and User Interfaces and Hospital Systems. These only hint at the great diversity to be found in the symposium. The track names are less important than the individual papers they subsume. This *Proceedings* will serve as a valuable resource for those who wish to learn who's doing what and how to build on that experience.

If anyone reads the preface of the *Proceedings* at all, they most likely do it long after the symposium is over. If you attended the symposium, you know that there was much offered that is not represented between these pages. If you weren't there, think about attending the next one. Reading the papers cannot begin to match the excitement of hearing more than 150 presentations, witnessing nearly 80 demonstrations, and interacting with close to 100 poster presenters, not to mention the elements of the meeting that are not recorded here, including the opening Plenary Session, nine track keynote lectures, 50 tutorials, 19 panels, 18 workshops, 26 "Meet the Experts" sessions, the American College of Medical Informatics Distinguished Lecture, numerous award presentations, Working Group meetings, and commercial exhibits.

The 1996 symposium was the product of tremendous teamwork. This year's Program Committee labored longer and harder than any I have known, and I believe the result is self-evident. Each has given generously of his or her time and made creative intellectual contributions. I owe each of them a personal debt. The Program Committee members are Mark Tuttle, Vimla Patel, Mark

Musen, Moon Mullins, Alexa McCray, Dan Masys, Gil Kuperman, George Hripcsak, Sue Henry, Chris Chute, Randy Barrows, and Andrew Balas. The AMIA staff has also worked incredible hours to make sure all the minute details of the meeting, the programs, and the *Proceedings* are not only done, but done well. Their pride in their work has been an inspiration to me. Led by Executive Director Jeanne Nevin and Meetings Manager Vernell Henry, they are Janice Kennedy, Andria Brummitt, Monica Jonas, Sharon Jadrnak, Kristina Sims, Michelle Daniels, and Debbie Preusse. There have also been contributions from the AMIA membership, including the Working Group chairs, who helped develop special parts of the program; Ted Shortliffe, who helped with the hardest task (getting a keynote speaker); and, of course, the reviewers. I also thank each of the authors who contributed their intellectual effort to the symposium. I have had personal support from Paul Clayton and the rest of the gang at the Columbia University Department of Medical Informatics, with their unofficial contributions, such as covering for me when I was absorbed in Program Chair duties. Any personal contributions I made to the meeting were the result of training I received under previous Program Chairs Paul Clayton, Mark Frisse, Charlie Safran, Judy Ozbolt, and Reed Gardner, and AMIA's former executive director, Gail Mutnik. Acknowledgments would not be complete without recognizing the sacrifices of my family, who gave me up to AMIA for some periods of time and put up with my distractedness for others.

With this introduction, I now urge you to venture on into the Proceedings. Whether you enter via the table of contents, the author list, or the subject index, I think you will find something of value. If you are looking at this during the twentieth century, you will find relevant methods, results, and discussion. If you are looking during the twenty-first century (or beyond!), you will find an unparalleled archive of the state of the art of medical informatics. The papers and abstracts contained herein include landmark articles by current leaders of the field and the first significant efforts by tomorrow's.

James J. Cimino, M.D.
Program Chair

Utilizing OODB Schema Modeling For Vocabulary Management

Huanying (Helen) Gu[1], James J. Cimino[2], Michael Halper[3], James Geller[1], Yehoshua Perl[1]

[1]CIS Dept. & CMS NJIT, Newark, NJ 07102
[2]Dept. of Medical Informatics, Columbia University, New York, NY 10032
[3]Dept. of Math & Comp. Sci., Kean College of New Jersey, Union, NJ 07083

Comprehension of complex controlled vocabularies is often difficult. We present a method, facilitated by an object-oriented database, for depicting such a vocabulary (the Medical Entities Dictionary (MED) from the Columbia-Presbyterian Medical Center) in a schematic way which uses a sparse inheritance network of area classes. The resulting Object Oriented Health Vocabulary repository (OOHVR) allows visualization of the 43,000 MED concepts as 90 area classes. This view has provided valuable information to those responsible with maintaining the MED. As a result, the MED organization has been improved and some previously-unrecognized errors and inconsistencies have been removed. We believe that this schematic approach allows improved comprehension of the gestalt of large controlled medical vocabulary.

INTRODUCTION

Large medical vocabularies are emerging as important resources for use in medical information systems. Acceptance of these vocabularies has been slow, however. Part of this may be an inability to understand and adapt a system developed elsewhere to systems grown at home—the "not invented here" syndrome. These vocabularies also present significant maintenance challenges for their creators, especially when they grow to 10s or 100s of thousands of terms. We are exploring the use of an object-oriented database (OODB) paradigm for generating high-level vocabulary schemata, intended to enhance comprehension by users and maintainers of a large controlled medical vocabulary. We present one such schema (for the Object Oriented Health Vocabulary repository (OOHVR)) generated from the Columbia-Presbyterian Medical Center (CPMC) Medical Entities Dictionary (MED) [1] and show how the comprehension it provides has improved the MED content.

BACKGROUND

The Medical Entities Dictionary

The MED is a collection of over 43,000 concepts which denote the coded terms in use by CPMC clinical systems. Concepts are represented as frames, consisting of slots which are attributes (literal values) and relationships (pointers to other concepts). The concepts are organized into a semantic network which uses relationships to provide named (i.e., semantic) relationships between concepts (for example, a link between a laboratory concept and a specimen concept). The MED permits multiple inheritance through an IS-A hierarchy. Studies show that current users of the vocabulary at CPMC have trouble navigating through the semantic network to find desired terms [2]. One of us (JJC) is responsible for maintaining the MED, and he often finds it difficult to add terms or create links without a clear understanding of the underlying vocabulary model. Others approaching the MED have encountered similar problems [3]. Figure 1 is a small part (about 0.2%) of the MED. Four concepts (**Allen Serum Amylase Measurement, Calcification of Pericardium, CPMC Drug: Benadryl 25MG Cap,** and **Pancreatin**), all ancestors of them, the IS-A hierarchies, and semantic links between concepts are shown; for clarity, some detail has been omitted: other children of the ancestor concepts, reciprocal semantic links, names of the semantic links (only their numeric codes are shown), and literal attributes. In this example, we see some terms for laboratory tests, medications, and diagnoses.

The OOHVR Schema

The MED contains multiple inheritance and reciprocal relationships. Since the relational model cannot model the complex objects directly and does not provide the notion of inheritance, using an OODB to model MED is a natural thing. The question we faced was how to model the MED semantic network using the available constructs of an OODB schema. Previously, an object-oriented framework has been used for modeling thesauri [4,5]. A terminology editor was also built in that context as a tool for extracting relevant information from hypermedia documents [6].

One approach to modeling the OOHVR vocabulary is to view all the nodes of the network uniformly. Everything is just a concept, so we can define a single object class *Concept* and make all nodes instances of it. All attributes and relationships defined with respect to any concepts become properties of this one class. This approach is unsatisfactory for two reasons. First, the properties of different concepts can vary greatly, and these properties carry much of the semantics of the network. The nodes of the network are linked together with relationships such as *measures* and *has-site*. To assign all properties to a single class

0195—4210/96/$5.00

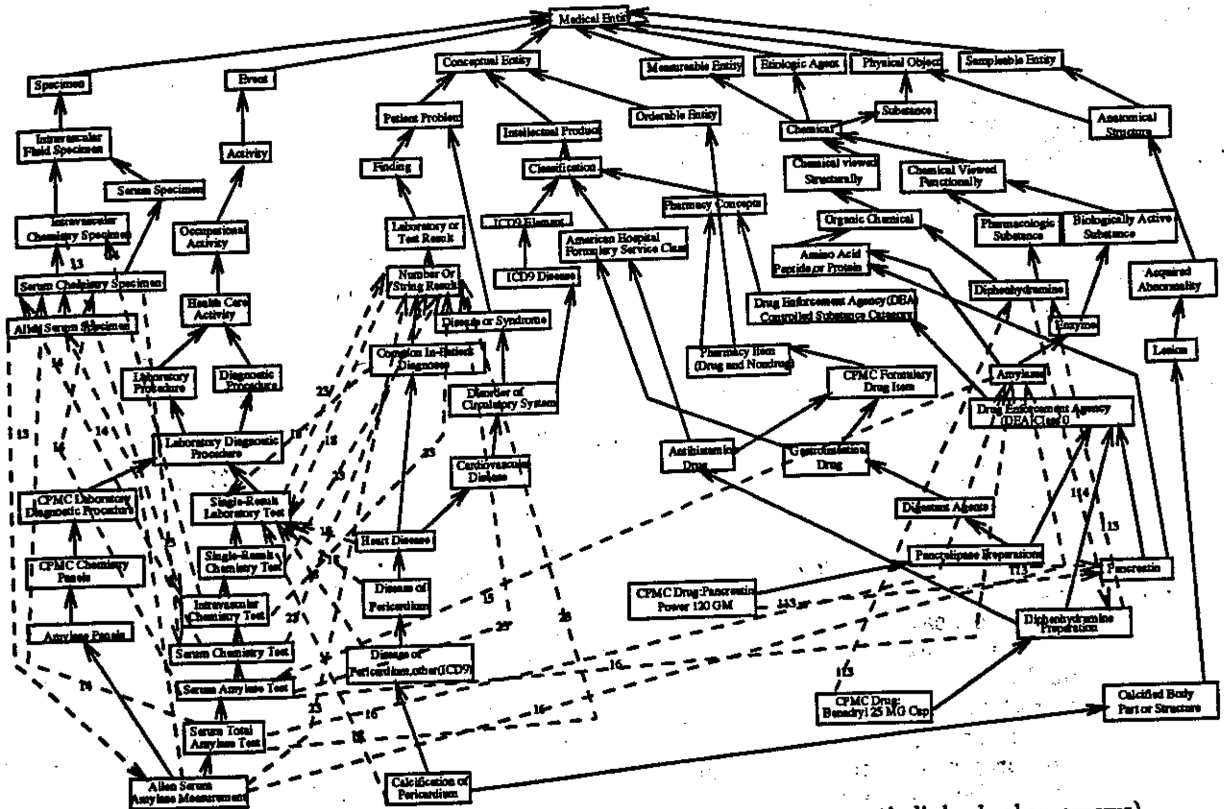

Figure 1: Sample MED content (IS-A hierarchies–solid arrows; semantic links–broken arrows)

and thus provide all concepts with them will hide the properties defined for a concept. Furthermore, it is a waste of database storage.

Second, the hierarchy supports property inheritance. E.g., **Serum Amylase Test** *is a* **Serum Chemistry Test** and inherits its properties. Defining a single class means "flattening" this hierarchy and failing to exploit a fundamental aspect of object-oriented modeling.

METHODS

Initial OOHVR Schema

Our approach to mapping the MED onto an OODB schema is based on the underlying pattern in which its properties are introduced. For each property there is a unique concept **C** where it is first introduced. This property is inherited by all the descendants of **C**.

We partition the MED into groups such that all concepts in one group have the same properties. Such a group is called an *area* and is defined more precisely as a sub-hierarchy of the MED satisfying the following conditions: (1) The sub-hierarchy has a single root, and (2) the root is the only node that introduces new properties. For 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 have the same properties are in this "Measurable Entity" area. If a concept is below **Measurable Entity** and introduces properties, then it is a root in a new area.

Each area in the MED is modeled as an object class in the schema, called an *area class*. The properties defined for an area class in the OOHVR schema are exactly those introduced by the area's root in the MED. In the case of the class *Measurable_Entity_Area*, it has the relationship *measured-by*, among others. All concepts in an area, including the root concept, become instances of the corresponding area class in the OOHVR.

Each concept in the MED is a descendant of **Medical Entity**. The root of any area is a child of a node(s) in some other area(s), except for **Medical Entity**. The root of an area has all the properties of its parents' areas plus the properties defined explicitly for it. To capture this in our model, we place each area class corresponding to a root node in a subclass relationship to the area class(es) of the root's parent(s). The subclass hierarchy induced by this process is not necessarily a tree. The area class *Medical_Entity_Area* is the root of the OOHVR's schema.

In the MED hierarchy most nodes do not introduce properties. We call such a hierarchy a *sparse inheritance hierarchy*. The OOHVR schema can be seen as an abstraction of the property definitions and accompanying inheritance that occur within the MED. We call this kind of schema for a sparse inheritance hierarchy a *network abstraction schema*. However, if one is still to use the concept subsumption hierarchy of the vocabulary in the other ways for which it was intended (e.g., in order to reason with respect to it), then it is mandatory that that the hierarchy appears in its entirety within the OOHVR. This is accomplished by introducing two reflexive relationships at the root area class *Medical_Entity_Area*: ***has_superconcepts*** and ***has_subconcepts***. These properties are defined as follows. In the MED, if **X** is a subconcept of **Y**, then, in the OOHVR, the object corresponding to **X** refers to **Y** with a ***has_superconcepts*** relationship; ***has_subconcepts*** is the converse. In this manner, instances in the OOHVR refer to their super- and sub-concepts. In other words, the hierarchy of concepts in the MED is represented in the OOHVR at the instance level.

Extended OOHVR Schema

One complication in the above mapping arises because of the possibility of multiple inheritance in the MED. 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 schema. This situation occurs when **X** needs the properties introduced at both *A_Area* and *B_Area*. We may even have a whole set C of concepts which have the same properties as **X**. Actually, C is exactly the intersection of the areas of A and B.

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 two separate area classes. However, this is forbidden in most OODB models. To accommodate this scenario, we define the intersection of two areas as an area, called an ***intersection area***. A class is defined for it in the OOHVR schema, even though this class does *not* introduce any new properties. Clearly such an area class will be a subclass of two other classes. The notion of intersection area can be extended to three or more unrelated areas. For example, "Diphenhydramine Prep." is the intersection of three areas "Antihistamine Drugs," "DEA Controlled Subst. Category," and "Drug Allergy Class."

For the above example, we introduce an intersection area class *C_Area* as a child of *A_Area* and *B_Area* in the extended OOHVR schema. That schema will include all intersection area classes and their subclass relationships to other area classes, which themselves may be intersection area classes. The concept **X** and the other concepts in the set C will be instances of the intersection area class *C_Area*. If **Z** is a root for C, then the corresponding intersection area class for C will naturally be denoted *Z_Area*. Otherwise, the schema designer has to arbitrarily select one of the concepts of C, as the name of the intersection area class.

RESULTS

The OOHVR schema obtained via the above described mapping is very compact in terms of the number of classes, compared to the MED. The MED's 43,000 concepts are partitioned into only 90 area classes, of which 37 were intersection classes. Instead of listing area classes in a table, graphical representation is useful for viewing and comprehending this complex schema. In Figure 2, we present the partial schema which shows 24 of the area classes, corresponding to the section of the MED shown in Figure 1. The schema is presented using the new version OOdini2 of our graphical schema editor OOdini [7]. Seven of these are intersection classes: ***Chemical***, ***Pancreatin***, ***Allen Serum Amylase Measurement (ASAM)***, ***Heart Disease***, ***Unknown and Unspecified Cause of Morbidity and Mortality***, ***Calcification of pericardium***, and ***Antihistamine Drugs***. The 24 area classes shown account for 26% of the area classes in the entire schema and represent 28001 MED concepts, or 65% of the MED. Compared to Figure 1 which is complicated and confusing , Figure 2 is much simple and easy understanding, but it still completely and correctly reflects the MED. At present, a version of the OOHVR is running as an ONTOS database [8].

Review of the schema by one of us (JJC) showed that the non-intersection areas are all appropriate and correspond to the intended design of the MED. Review of the intersection areas found that 19 of the areas are appropriate, 14 were collections of unrelated MED concepts which should be grouped into classes corresponding to the intersection areas, and 5 were outright mistakes in the MED. Examples of each are described below.

Understanding the MED Schema

The network abstraction schema provides users with a high-level view of the MED. The intersection area classes provide demonstrations of complex interactions between areas. The ***Antihistamine*** area class, for example, is a collection of 316 MED concepts grouped into 31 MED concept groups. Each concept group represents a grouping of drugs (such as antihistamines) which are descendants of the ***Pharmacy Item*** area and the ***American Hospital Formulary Service Class*** area. To a domain expert, this representation makes perfect sense, since the concepts corresponding to medications are classified in multiple ways in the MED, and inherit different attributes from each of the parent areas. Review of this area class did not result in any changes to the MED.

Improving MED Organizational Structure

Certain single-result lab tests in the CPMC laboratory system can, at times, be ordered separately as single-test diagnostic procedures. The

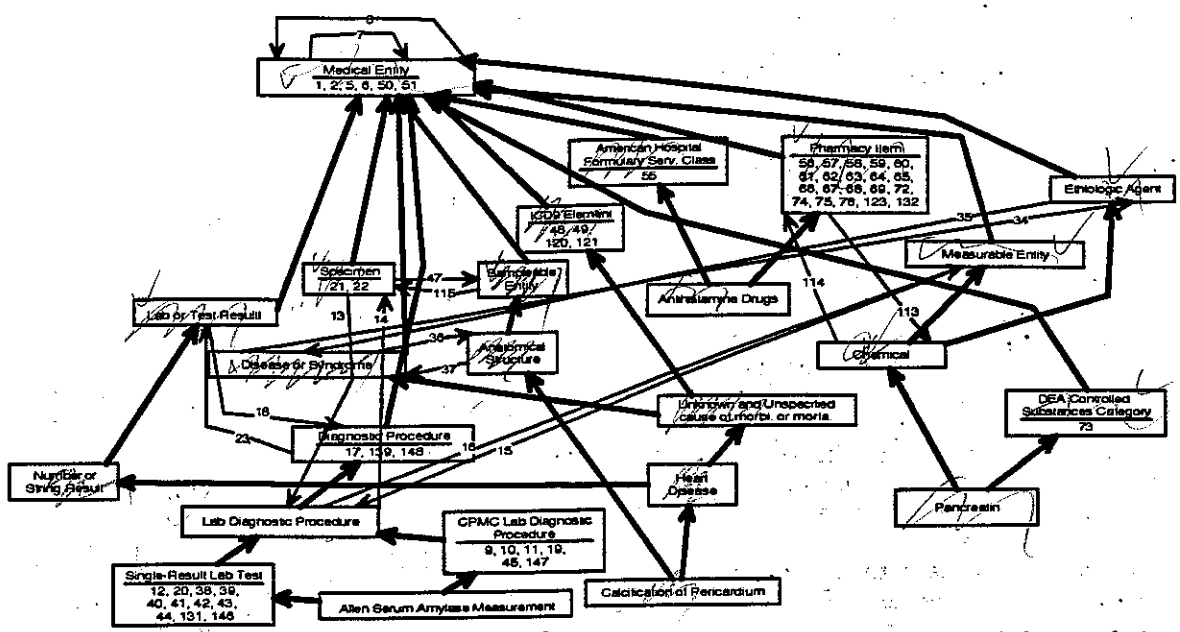

Figure 2: Partial OOHVR Schema showing the area classes which account for the Figure 1; subclasses relationships are shown with heavy arrows and relationships are numbered and shown with thin arrows; numbers inside the boxes represent attributes.

concepts in the MED which correspond to these tests therefore have attributes of both tests and procedures. The schema view grouped these tests into the *ASAM* area, under the areas *Single-Result Lab Test* and *CPMC Lab Diagnostic Procedure*. However, in the MED, there was no single concept which is the parent of these particular tests. We therefore created a new concept in the MED called **Orderable Tests** as a child of both *Single-Result Lab Test* and *CPMC Lab Diagnostic Procedure*; we then linked all the tests in the *ASAM* area as children of **Orderable Tests**. When the schema was redrawn (not shown), the *ASAM* area took on the new name *Orderable Tests*, since that concept in the MED is now the root of all the other concepts in the area class.

Finding Inconsistencies in MED Content

The intersection area *Calcification of the Pericardium* area contains all concepts which are both heart diseases and anatomical structures (40 in all). Until we saw this view of the MED, we did not realize that the same concepts were listed as both diseases and anatomical structures. This was not consistent with the original design of the MED in which disease could be linked to body parts as the "site of disease" but could not themselves be body parts. Discovery of this intersection class led directly to a study of these 40 terms and their reclassification as body parts or diseases, as deemed appropriate by an outside domain expert. As a result, when the schema was redrawn, there was no longer any intersection area below *Heart Disease* and *Anatomical Structure*.

Another example of an error discovered through use of the schema was 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 schema 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 as a chemical and as a medication (as shown in Figure 1). We corrected this error by removing the IS-A link between **Pancreatin** and **DEA Class 0**. Since **Pancreatin** was the only concept in the MED with attributes of chemicals and medications, the *Pancreatin* Area had only one concept. After the correction, the area no longer existed, since **Pancreatin** was now included in the *Chemical* Area.

DISCUSSION

The development of sparse inheritance networks and intersection areas, is of more than theoretical interest. The maintenance of the MED is a complex and difficult task and no commercial tools are suitable to support it. Browsers and editors have been developed, and continue to be developed, but providing users of the MED (both MED maintainers and application builders) with comprehensible, comprehensive views remains difficult. Others using complex controlled vocabularies will undoubtedly face similar difficulties. Some of the challenges of maintaining and using the MED include understanding the MED schema, improving the MED organization, and 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. With over 43,000 concepts in the MED, 88 attributes, 62 relationships, 55,000 hierarchical links and 96,000 non-hierarchical links, understanding the "big picture" is difficult. By reducing the MED hierarchy almost 500-fold, one can immediately see the important areas 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 candidate areas. Then the user can review the attributes and relationships of those areas to determine the appropriate area for the new concept. Thus, the OOHVR schema can be shown to provide a valuable gestalt of the MED complexity.

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. For example, the difference between individual laboratory tests (such as **Serum Glucose Test**) and procedures (such as a **CHEM-7**, a panel of 7 individual tests) often confuses users of the MED [2,3]. 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. The above correction to the MED, based on the schema, simplifies this situation with the creation of the **Orderable Tests** concept.

Over the past seven years, the MED has grown by about 500 concepts per month. This growth has been the result of the work of several individuals and sometimes of automated mechanisms. Hence it is not surprising that inconsistencies and outright errors have crept in. 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. The OOHVR schema provides a way for two people to share the same view of the MED and to identify differences in their views.

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 [9], it was thought that such ambiguity could be detected through automated means. The use of intersection areas has provided such a method.

CONCLUSIONS

The maintenance of a large controlled vocabulary is a complex and difficult task. The complexity stems from the need to comprehend a complex body of knowledge consisting of many concepts and links. An example of such a medical vocabulary is the MED built at CPMC. In this paper, we describe an approach we have devised, using an OODB schema which captures the structure of the MED in a compact way. Experience shows that this approach improves comprehension of the MED and facilitates corrections of inconsistencies and errors.

Acknowledgments

This research was 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 and CMS. Dr. Cimino's work was supported in part by the National Library of Medicine and the IBM Corporation.

References

1. 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.
2. Hripcsak G, Allen B, Cimino JJ, Lee R. Access to data: comparing AccessMed to Query by review. *JAMIA*, 3(4):288–299, 1996.
3. Kannry JL, Wright L, Shifman M, Silverstein S, Miller PL. Portability issues for a structured clinical vocabulary: mapping from yale to the columbia medical entities dictionary. *JAMIA*, 3:66–78, 1996.
4. Fischer DH. Consistency rules and triggers for thesauri. *Int. Classif.*, 18(4):212–225, 1991.
5. Fischer DH. Consistency rules and triggers for multilingual terminology. In *Proc. TKE93, Terminology and Knowledge Engineering*, pages 333–342, 1993.
6. Möhr W, Rostek L. TEDI: An object-oriented terminology editor. In *Proc. TKE'93, Terminology and Knowledge Engineering*, pages 363–374, 1993.
7. Halper M, Geller J, Perl Y, Neuhold EJ. 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.
8. Liu L, Halper M, Gu H, Geller J, Perl Y. Modeling a vocabulary in an object-oriented database. In *Proceedings of the Fifth International CIKM Conference*, 1996.
9. Cimino JJ, Hripcsak G, Johnson S, Clayton PD. 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.