Designing an Introspective, Multipurpose, Controlled Medical Vocabulary

James J. Cimino, M.D., George Hripcsak, M.D.
Stephen B. Johnson, PhD., Paul D. Clayton, PhD.

Center for Medical Informatics, Columbia University
Columbia-Presbyterian Medical Center
New York, New York 10032

The medical vocabulary used in clinical information systems must be more than a simple list of terms. We argue that such a vocabulary must have synonymy, domain completeness and multiple classifications, providing consistent views and explicit relationships, while remaining unambiguous and non-redundant. We examine the abilities of existing controlled vocabularies (ICD9-CM, SNOMED, MeSH, CMIT, CPT4, COSTAR, HELP, DXPLAIN, and UMLS) to meet these goals and propose an enhanced vocabulary structure based on a directed, acyclic semantic net. This structure provides a representation which permits introspection by the vocabulary maintenance system responsible for providing a terminology which meets our seven requirements. The vocabulary, which we refer to as the Medical Entities Dictionary, will serve a variety of applications.

Introduction

The Center for Medical Informatics of Columbia University, at the Columbia-Presbyterian Medical Center, is developing a clinical information system (CIS) which will require that we bring a number of disparate medical informatics tools into a clinical setting. Applications which will make use of the vocabulary include a clinical alerting system (based on the HELP model[1]) which will apply a standardized form of medical logic[2] to patient data, access to on-line medical literature databases, such as MEDLINE[3], and diagnostic decision-support tools, such as QMR[4] and DX*plain*[5]. These applications rely on controlled vocabularies for representing clinical data. Integrating such applications with a CIS is facilitated by the use of a controlled vocabulary for the clinical database.[6]

The introduction of a controlled vocabulary to our CIS requires that additional applications be developed for the creation and manipulation of terminology. For example, since a user will not be expected to know all of the terms which are in the vocabulary, an application will be needed which will allow the user to locate a desired term. When a desired term can not be found, a mechanism must be provided to allow its inclusion in the vocabulary. We do not intend to create, *de novo*, a vocabulary which can accommodate the needs of all of our applications. We will draw upon the terms compiled for use in other vocabularies (many of which are described below), as well as analyzing the terminology already in use at our institution.[7] However, we expect that the construction of a satisfactory vocabulary will be an iterative process. We therefore seek a vocabulary structure which will facilitate not only our clinical applications, but vocabulary maintenance and browsing tools.

This paper describes the requirements we have defined for our controlled vocabulary. We examine existing controlled vocabularies to determine how they meet, or fail to meet, each of our requirements. We then describe how the preliminary design of our controlled vocabulary is built upon the features of existing vocabularies. An important feature of our proposed vocabulary is that it can function in an introspective manner; that is, it will contain knowledge to assist in the vocabulary's growth and maintenance.

Existing Vocabularies

Before elaborating on the requirements of our vocabulary, we will briefly describe the structure and content of nine existing medical vocabularies which we reviewed during the design of our vocabulary.

The *International Classification of Diseases* was developed by the World Health Organization for collecting health statistics. The Ninth Revision was introduced in 1977 and was found to be inadequate for detailed clinical coding.[8] In response, the Department of Health and Human Services provided the *Clinical Modifications* extension which forms ICD9-CM.[9] This vocabulary has gained wide acceptance for coding clinical disorders, particularly for hospital billing purposes. It also includes terms for medical and surgical procedures, occupations and other factors influencing health status. The basic structure of ICD9-CM is a numerically coded, strict hierarchy. The ICD9-CM coding manuals also include a large number of terms which act as synonyms by referring to particular terms.

The College of American Pathologists also found ICD9 inadequate for coding clinical information and, encouraged by the success of their *Systematized Nomenclature of Pathology* (SNOP),[10] developed the *Systematized Nomenclature of Medicine* (SNOMED)[11]. The intended domain of SNOMED is all of clinical medicine. It is arranged into seven axes, with each axis being a strict, numerically coded hierarchy, as in ICD9-CM. Because the axes are mutually exclusive, the entire vocabulary is really a single strict hierarchy. Terms from different axes can also be combined to form compound terms. Like ICD9-CM, the coding manuals contain many synonyms which refer to individual or combined SNOMED terms.

Like ICD9-CM and SNOMED, the National Library of Medicine's *Medical Subject Headings* (MeSH)[12] vocabulary has been successful in meeting the needs for which it was designed: indexing medical literature. Like the other vocabularies, it is a hierarchy. However, it has several structural properties which differ significantly from other standard nomenclatures. First, its hierarchy is not tied to a rigid coding scheme, so that any number of terms may appear at a given level. Second, the MeSH hierarchy is not strict: terms may appear in more than one position in the tree (called contexts). Third, MeSH introduces inheritance as a vocabulary feature through the use of Subheadings. These are terms which are used to provide contextual information for literature citation indexing. For example, the Subheading "Etiology" might be used in conjunction with some MeSH disease to index a citation which deals with causes of the disease. MeSH also includes a large number of "Entry Terms" which act as synonyms for MeSH terms (although the relationship between the Entry Terms and their corresponding MeSH terms varies).

The American Medical Association addressed the need for a standardized vocabulary of medicine with the publication of *Current Medical Information and Terminology* (CMIT)[13], which consists of a controlled vocabulary of diseases with a description (including synonyms) for each disease, composed of structured free text. The CMIT diseases are in a strict hierarchy, organized by organ system.

0195-4210/89/0000/0513$01.00 © 1989 SCAMC, Inc.

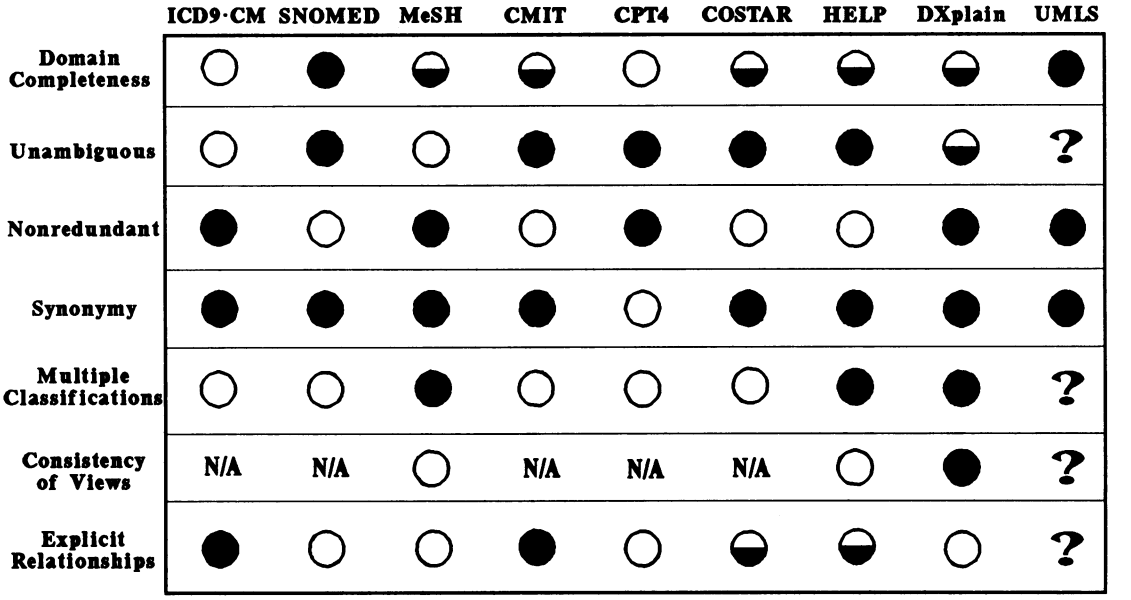

Figure 1 - Evaluation of Properties of Controlled Vocabularies - Empty circle indicates property absent; filled circle indicates property present; half-filled circle indicates property partly present; "?" appear for undefined properties of the UMLS; "N/A" indicates that multiple views are not applicable.

The American Medical Association has also published *Current Procedural Terminology* (CPT4)[14] to provide a means for encoding billable procedures. CPT4 has achieved great acceptance, as the *de facto* standard for reimbursement for procedures. The terms in CPT4 are organized in a strict hierarchy, but unlike ICD9-CM and SNOMED, the coding does not determine the formal structure.

The Computer-Stored Ambulatory Record (COSTAR), developed at the Massachusetts General Hospital, makes use of a customizable controlled vocabulary called the Directory[15] to provide a way to encode information about signs, symptoms, diseases, procedures and medications. The structure is that of a strict hierarchy, with each term assigned a five-character code plus an optional modifier. The Directory also accommodates synonyms.

HELP, a hospital information system developed at the LDS Hospital in Salt Lake City, also makes use of a customizable controlled vocabulary, called the HELP Dictionary[16], which is a strict hierarchy with byte addresses providing the coding scheme. Each level of the structure connotes information about the term. For example, medications are assigned to a Class (such as Antibiotics) and a Subclass (such as Penicillin), with specific drug names appearing at the next level. The Dictionary also includes Modifiers, which can be introduced at any level in the hierarchy and are inherited by all of the descendant terms. One example of this is the modifier, "Route", which is introduced at the data-class level for drugs. Thus, all drugs can be modified by "Route". Like MeSH, terms which are appropriate to more than one classification can appear as multiple entries in the vocabulary. A keyword index assists users in finding the desired terms when they appear in more than one context. This index also provides synonyms for facilitating vocabulary queries.

DXplain, a program which assists with medical diagnosis, uses two controlled vocabularies, one for diseases and one for clinical findings, with synonyms for both.[5,17] Initially, the disease vocabulary was unstructured and the findings vocabulary included only some small hierarchies (for example, all abdominal pain terms are arranged under the term "Abdominal Pain"). These vocabularies were subsequently reorganized: diseases into a hierarchy by organ system and etiology, and findings into a hierarchy by organ system, finding type.

The Unified Medical Language System (UMLS) project, proposed by the National Library of Medicine, presents ways by which the "Tower of Babel" of medical terminologies (such as those described above) can be consolidated.[18] Work has focused on how medical terminologies are represented to provide common ground upon which translations can be performed. A number of structural models are being explored for representing how medical concepts are expressed in various controlled and uncontrolled lexicons. These schemata are intended not to enumerate all possible medical terms but, rather, to provide a means for mapping among them. The UMLS is now in the second stage of development, and many of its features remain topics for research. A recent article describes the initial UMLS structure to be a thesaurus of medical concepts which will include lexical mappings to multiple vocabularies.[19]

Properties of Our Controlled Vocabulary

With the above brief overview of some existing vocabularies, we can now discuss the requirements of our vocabulary and show how various design features can help meet or defeat these requirements. One way to assess vocabulary requirements is to examine the effect of the vocabulary on queries. The goal of a query may be to find a patient's serum potassium, to obtain literature references pertinent to serum potassium, or to look for the "serum potassium" in the vocabulary. Queries must be sensitive, retrieving all appropriate information. Queries must also be specific, so that inappropriate information is not retrieved. Finally, queries must be reliable, such that a query returns an identical result, no matter how it is posed.

The sensitivity, specificity and reliability of queries that we will be using can be improved by a controlled vocabulary that has: 1) domain completeness, 2) unambiguous terms, 3) non-redundancy, 4) synonyms, 5) multiple classification of terms, 6) consistency of views, and 7) explicit relationships. We define each of these below, with an examination of how they are met by existing vocabularies. These comments should not be construed as criticisms: each vocabulary functions well for its intended purpose. However, while each has some properties which meet our requirements, none provides all of the properties required by our applications. **Figure 1** summarizes our analysis of all seven properties in each of the nine vocabularies.

Domain Completeness

It is unlikely that any controlled vocabulary can anticipate and include all possible terms that lie within its domain, whether it be all of medicine or a specialized area (such as disease or procedure names). Existing vocabularies achieve varying degrees of domain completeness, demonstrating the relative advantages of their schemes.

A common reason for lack of completeness is the effort required to achieve it. ICD9-CM, for example, has been criticized for its failure to provide consistent coverage of all areas of diseases.[20] CMIT's coverage of diseases appears reasonably good; however, deficiencies have been noted[21]. For example, diseases such as herpes proctitis, pneumococcal meningitis and candida esophagitis are absent.[22] MeSH also has certain areas (most notably procedures, symptoms and physical findings) which are inadequate for clinical information systems. This is due not to oversight but to the purpose of MeSH.[23] Similarly, CPT4 often fails to completely cover its domain (medical procedures) because of it concerns itself primarily with billable procedures. For example, "Insertion, Replacement, or Repositioning of Permanent Transvenous Electrodes Only (15 Days or More After Initial Insertion)" is a CPT4 term. Apparently, separate billing is not permitted for such procedures done less than 15 days after initial insertion, so there is no CPT4 term for it.

Another important factor limiting the completeness of some existing vocabularies is their organizational structure. In some cases, the coding scheme restricts the number of terms that can be included. In ICD9-CM, each term is assigned a numeric code which defines the term's location in the hierarchical structure. The hierarchy is limited to four levels of depth, with at most ten terms at each level. In many cases, this is insufficient to allow expression of all variations of a term, in which case there are nine variations listed, with a tenth one listed as "Other". The coding scheme of SNOMED is only slightly more permissive, allowing five levels of at most twelve terms each. As the *SNOMED Manual* itself states: "An unwelcome consequence is that some parts...are so crowded that additional terms cannot be added and that others are almost empty."[24] Even HELP, which allows approximately 256 terms at each level in the hierarchy, has proven too limiting; most notably, in radiology reporting (PDC, unpublished data). It is clear that a structure which limits the depth or breadth of vocabulary organization is to be avoided.

Some vocabularies improve domain completeness through the use of additional structures. SNOMED allows terms to be combined, providing a virtually limitless number of complex terms. Inherited modifiers, such as MeSH Subheadings and HELP Modifiers, enhance a vocabulary by extending its expressiveness. Frequent updating can also improve completeness. MeSH is changed annually, while other vocabularies, such as HELP, COSTAR and DXplain, are changed daily in response to user requirements. The UMLS will address domain completeness by simply subsuming existing vocabularies.

Unambiguous

Vocabulary terms must not be ambiguous, defined here as referring to more than one concept (a homonym). If a term is ambiguous, then at least two disparate types of data are stored under the same term, directly affecting query specificity. "Other" terms in ICD9-CM and SNOMED represent multiple specific terms. CMIT and CPT4 prevent ambiguity through extensive, explicit disease descriptions and term names, respectively. In COSTAR, the terms are defined by the users and so the meanings should be agreed upon.

Ambiguity occurs in MeSH when terms appear in multiple contexts, with different meanings. For example, "Cardiac Output" is listed as both a "Cardiovascular Function and as a "Heart Function Test"[12]. In fact, in the latter context, the term means "cardiac output determination", but MeSH does not differentiate.

While diseases in DXplain can be shown to be unambiguous (by comparison of their descriptions), the clinical terminology is sometimes intentionally ambiguous. For example, the term "Worse In Morning" used in combination with terms such as "nausea" and "joint stiffness" to represent concepts such as "nausea worse in the morning" and "joint stiffness worse in the morning". The ambiguity can be uncovered by looking at how the term is used in the knowledge base (to evoke the diseases "Pregnancy" and "Rheumatoid Arthritis"). This ambiguity was retained intentionally because the developers felt that there was no need to clutter the vocabulary with all permutations of a term when the users would recognize the meaning of the term, given its context. This kind of ambiguity is not a problem for DXplain (it has no trouble distinguishing pregnancy from rheumatoid arthritis, so that application.

Non-redundancy

There must be no redundancy in the vocabulary. That is, there must be only one way in which each concept can be expressed. Allowing two terms to refer to the same concept will reduce query sensitivity. For example, if the terms "MI" and "Myocardial Infarction" refer to the same disease, a query for all patients with "MI" will miss the patients who are reported to have "Myocardial Infarction".

Unfortunately, some controlled vocabularies suffer redundancy due to their flexible nature. For example, the ability to combine terms in SNOMED provides multiple ways to represent the same concept. We need look no further than the favorite example of the College of American Pathologists: pulmonary tuberculosis can be expressed as either D0188 or as T2800 + M44060 + E2001 + F03003 + D0188 ("Lung" + "Granuloma" + "M. Tuberculosis" + "Fever").[25] In COSTAR, it is the strict hierarchy which allows redundancy to occur. Adding a new term requires "walking" down the structure of the Directory until the appropriate location is found. If, at any level, there is more than one correct path (that is, the term can be classified in multiple ways), one must be chosen and the other(s) ignored. There is no guarantee that the "new" term does not already exist, in a slightly different form, at the end of some path not taken. The HELP dictionary has inherent redundancy which has many useful properties for its application. For example, "Digitalis" appears under both medications and laboratory tests,[16] but proper retrieval depends on knowing about redundant contexts. MeSH avoids redundancy by allowing terms to appear in multiple contexts in the hierarchy. This approach differs from HELP because the multiple contexts are recognized as referring to the same term.

In some vocabularies, it is a lack of structure which allows redundancy to occur. In CMIT, redundant diseases appear, but it is difficult to recognize their occurrence. For example, there are entries for "Pseudotumor Cerebri" and "Papilledema, Idiopathic Intracranial Hypertension". Similarly, "Gonorrhea, Male" and "Urethritis, Acute, Gonorrheal, Male" appear[22]. The terminology used in CMIT's disease descriptions is uncontrolled and replete with redundancy.[26] In DXplain, the initial lack of structure permitted redundancy to occur. Because many of the disease names were selected from CMIT, the redundancies in CMIT were reproduced in DXplain. The classification schemes for diseases and findings in DXplain were introduced, in part, to help detect and eliminate this redundancy.[21]

Synonymy

Since we can not expect users of the vocabulary to remember all of the terms, there must be provision for including synonyms in the vocabulary. This differs from redundancy in that only one term can be used for access to data, while synonyms of that term can only be used to access the controlled term. With the exception of CPT4, all of the controlled vocabularies have synonymy of one form or another. We expect synonymy to be an important feature of our vocabulary as well.

Multiple Classification

A vocabulary can be a simple collection of all the possible terms; however, this can be extremely unwieldy for both retrieval and maintenance. All of the existing vocabularies use some hierarchical classification scheme. A *strict* hierarchy does not allow a term to belong to more than one class. Thus disease classification can not be done both by etiology and organ system. In ICD9-CM, "Meningococcal Carditis" is classified under "Infectious and Parasitic Diseases", while "Pneumococcal Myocarditis" is under "Diseases of the Cardiovascular System". SNOMED shares this problem, with "Pneumococcal Pneumonia" under "Diseases Caused by Bacteria", "Clinical Pneumonia" under "Diseases of the Respiratory System", and "Staphylococcal Pneumonia" is classified, not as a "Disease", but as a "Morphology". Clearly, the strict hierarchy is too inflexible.[27]

Consistency of Views

It should be noted that use of multiple classification can create a problem. By phrasing a query differently, one can create several paths to get to the same term. In MeSH, which allows multiple classification, we might or might not find that "Aspirin" is under "Salicylates", depending upon which tree address we examine. The flexibility of the MeSH tree permits different classification schemes and, therefore, different views of the terminology, but the structure does not guarantee that the different views will produce consistent results. Similarly, terms in HELP which appear in multiple places may have different Modifiers, depending on the context. Discrepancy between views is generally considered undesirable.[28] DXplain's structure requires terms to have the same descendants in all contexts.

Explicit Relationships

All of the vocabularies discussed here include inter-term relationships, the most common of which are represented by the hierarchical structure. Usually, the hierarchy is used to connote class-subclass relationships. We refer to these as IS-A links, as in "Clinical Pneumonia IS-A Disease of the Respiratory System". In ICD9-CM and CMIT, there are only IS-A relationships between terms. Other vocabularies often have some ambiguity to the meaning of parent-child relationships. In SNOMED, the child may be related to the parent in many ways, such as: child IS-A parent, child IS-PART-OF parent, child IS-MADE-OF parent, child CAUSES parent, child IS-IN parent, and child IS-FOR parent.[29] In MeSH, the parent-child relationships have meanings such as IS-A, PART-OF, ASSOCIATED-WITH, and EQUIVALENT-TO.[30] In some cases, a term may be the parent of another term in one context and its child in another (see "Cycasin" and "Methylazoxymethanol Acetate"). When the kind of relationship between terms is unclear, query function may suffer. If we wish to query a clinical database to find a patient's pulmonary diseases, we would be interested in all terms which have an IS-A relationship to "Pulmonary Disease" but not terms which have a CAUSES relationship. HELP and COSTAR define the different relationship types in their hierarchies.

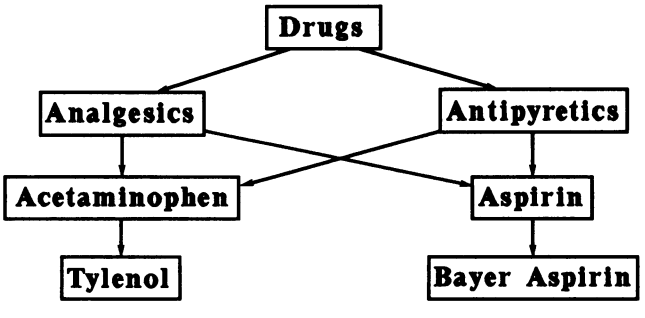

Figure 2 - Directed acyclic graph with consistent views

Design Features of Our Vocabulary

The vocabularies that have discussed employ various structures and procedures for representing the terminology used for their respective applications. While each strategy is successful for its intended purpose, our evaluation reveals that none will satisfy all of our requirements. However, it is possible to extract from each vocabulary those design features which support our functional requirements and make use of them in the construction of our own vocabulary. Like several of the vocabularies, we shall eschew a rigid coding scheme and make allowance for the inclusion of synonyms. However, we believe that some of our requirements can only be met through the use of more sophisticated features which have been derived, rather than extracted, from these vocabularies.

Vocabulary Structure

To allow us to maintain consistent views of the terminology, we can use neither a strict hierarchy nor even the flexible hierarchy of MeSH or HELP. Instead, we will need a structure in which terms can occur in multiple classes but are guaranteed to have only one set of descendants. The arrangement shown in **Figure 2** has these properties. This structure goes by various names, such as "tangled hierarchy", "semi-lattice", "poset" (for "partially ordered set") and "directed acyclic graph" (DAG). We choose DAG as the name which best describes the asymmetry of the parent-child relationship (directed), the stipulation that classes may not include themselves (acyclic) and the presence of multiple connections (graph).

Term Definitions

We seek a representation which will allow us to detect redundancy and ambiguity using automated tools. We believe this can be accomplished through the inclusion of a definition for each term. Duplicate definitions should identify redundant terms, while incomplete definitions should indicate ambiguous terms. Of course, for this detection to be automatic, the definitions must be in some well-defined format and, perhaps, use controlled terminology. Our approach will be through the use of attributes: each term is assigned one or more attributes, where each attribute consists of the name of the attribute, and a value, which is restricted to a particular data type (integer, character string, or data class). This kind of structure is commonly known as a frame[31], and has been used to represent semantic descriptions by UMLS researchers.[32,33,34]

Explicit Relationships

The relationships between terms in our vocabulary fall into two broad types. The first type is the traditional class-subclass, or IS-A relationship which provides the organization for the DAG. The attributes which are used in the term definitions are inherited according to class membership. For example, all members of the "Drug" class might inherit the attribute "Dosage", while all members of the class "Disease" might inherit the property "Organ System".

The second type of relationship between our terms is nonhierarchical, similar to the complex terms allowed in SNOMED. For example, there is no SNOMED term for "Nephritis", but we may combine the SNOMED terms "Inflammation" and "Kidney" to represent it.[35] In effect, the coding indicates that "Nephritis" has the attribute "Morphology" with the value "Inflammation" and the attribute "Topography" with the value "Kidney". Similarly, many of the attributes in our term descriptions will take values from the controlled vocabulary itself; that is, data classes are drawn from the vocabulary classes. Each term may have relationships to many other terms (depending on the richness and complexity of its definition), where each of these relationships conveys some semantic meaning. This structure is commonly referred to as a semantic network[36].

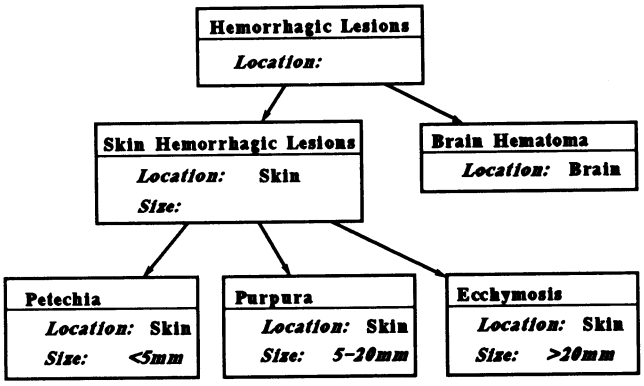

Figure 3 - Hierarchy of Terms with Definitions

Structures, Definitions and Relationships

In **Figure 3**, we see a fragment of a classification hierarchy for "Lesions". Inside each box is part of the definition of the term, in the form of attributes and values. All terms here have inherited the attribute "Location" from the parent term "Hemorrhagic Lesion" (which has inherited it from its parent). As we move downward to more specific terms, the attribute values can take on new values. We can see that the definition of each term makes it unique, at least for the small vocabulary shown. If we attempted to add a new member of the class "Hemorrhagic Lesion", such as "Bruise", and specified "Skin" as the location, it is a simple matter to recognize that the new term is either a synonym of "Skin Hemorrhagic Lesion" or a subclass of it, with some yet-to-be-defined nuance. In the former case, we would indicate that "Bruise" is a synonym, not a new term. In the latter case, we add the attribute "Size" to the term's definition and then request a value. If "Size" is specified as "30mm" we recognize that "Bruise" is either a subclass of "Ecchymosis" or its synonym.

The relationships shown as arrows in **Figure 3** represent the IS-A links of the classification hierarchy (e.g., "Purpura IS-A Skin Hemorrhagic Lesion"). It is likely that the terms shown here have other parents. For example, "Skin Hemorrhagic Lesion" IS-A "Skin Lesion" (which might explain how the attribute "Location" got its value). It should be noted that the DAG structure accommodates multiple parents, differentiating it from a strict hierarchy. Additional, non-hierarchical inter-term relationships become apparent when we consider that the values of the "Location" attribute for each term are taken from the class of anatomical terms (i.e., "Brain" and "Skin").

Discussion

To summarize, we find that the design features we wish to include are: 1) appropriate terms from all applicable terminologies (such as the UMLS) for domain completeness, 2) synonyms, 3) a nonrestrictive coding scheme to avoid ambiguity and allow domain completeness, 4) a DAG structure for multiple classifications with consistent views 4) formal term definitions to help detect ambiguity and redundancy and 5) a semantic network for explicit relationships.

It has been our experience that maintenance of controlled vocabularies is enormously difficult, particularly in providing domain coverage and synonymy, while preventing redundancy and ambiguity. We have therefore imbued our vocabulary with features that will facilitate its own maintenance. In effect, it will be a vocabulary knowledge base which, like other knowledge bases, can be made introspective to identify possible internal inconsistencies and improve performance.[21,37,38] For example, through internal comparisons it should be possible to detect synonymy by looking for two canonical terms with similar descriptions.[25,33,34,39] When we consider the structural and procedural components of the vocabulary together, we say that it is "introspective", in that is able to look within itself to make statements about its composition. There are many ways to implement introspection; the "Bruise" example serves to illustrate how the knowledge contained in the vocabulary might drive a rule-based expert system to assist in vocabulary maintenance.

The fact that our vocabulary includes knowledge about its contents makes it more than simple a terminology; it acts as a dictionary for medical terminology. For this reason, we refer to it as the Medical Entities Dictionary (MED). The initial representation of the dictionary is most consistent with an object-oriented model.[40] However, the information contained in the MED can also be expressed through the use of an entity-relation model[41], as is used in relational databases. This allows us to make use of the MED in a wide variety of applications, not the least of which is the clinical database. Clinical data which has been encoded with our controlled vocabulary will be useful for many different applications in our CIS, such as literature retrieval, hospital billing, decision support and the creation of research databases.

We are often reminded that medical knowledge has grown to the point where we require the assistance of computers to manage it. One response has been the construction of controlled vocabularies to facilitate this process. We are now at the point where the vocabularies themselves have reached unmanageable proportions and must again call on computers for help. We believe that our approach provides one kind of representation which can be used to manage a large medical vocabulary.

References

1. Pryor TA, Gardner RM, Clayton PD, Warner HR: The HELP system, *J. of Med. Sys.*; 1983; 7(2):87-102.

2. Clayton PD, Hripcsak G, Pryor TA, Cimino JJ, Wigertz OB, Bair T: A Proposed Format for Sharing Decision-Making Knowledge, in Kingsland LC, ed.: *Proc. 13th SCAMC*; Wash., DC; Nov, 1989.

3. Lindberg DA, Schoolman HM: The National Library of Medicine and medical informatics, *West. J. Med.*; 1986; 145(6):786-90.

4. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD: The INTERNIST I/ Quick Medical Reference Project - Status Report, *West. J. Med.*; Dec, 1986; 145:816-822.

5. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP: DXplain. An evolving diagnostic decision-support system, *JAMA*; July 3, 1987; 258(1):67-74.

6. Linnarsson R, Wigetz O: The Data Dictionary - A Controlled Vocabulary for Integrating Clinical Databases and Medical Knowledge Bases, *Meth. Inform. Med.*; 1988; 28(2):78-85.

7. Johnson SB, Gottfried M: Sublanguage analysis as a basis for a controlled medical vocabulary, in Kingsland LC, ed.: *Proc. 13th SCAMC*; Wash., DC; Nov, 1989.

8. Slee VN: The International Classification of Diseases: Ninth Revision (ICD9), *Ann. Int. Med.*; 1978; 88(3):424-6.

9. United States National Center for Health Statistics: *International Classification of Diseases, Ninth Revision, with Clinical Manifestations*; Wash., DC; 1980.

10. College of American Pathologists: *Systematized Nomenclature of Pathology*, First Ed., Chicago; 1965.

11. Côté RA, ed.: *Systematized Nomenclature of Medicine*, Second Edition, College of American Pathologists, Skokie, IL; 1982.

12. National Library of Medicine, Library Operations: *Medical Subject Headings*; Bethesda, MD, 1989.

13. Gordon BL, ed.: Current Medical Information and Terminology, Fourth Ed., AMA, Chicago; 1971.

14. Clauser SB, Fanta CM, Finkel AJ, eds.: *Current Procedural Terminology, Fourth Edition - CPT4.* AMA, Chicago; 1984.

15. Beaman PD, Justice NS, Barnett GO: A medical information system and data language for ambulatory practice, *Computer*; November, 1979:9-17.

16. Pryor TA: The HELP medical record system, *MD Computing*, 1988; 5(5):22-33.

17. Hupp JA, Cimino JJ, Hoffer EP, Lowe HJ, Barnett GO: DXplain - A Computer Based Diagnostic Knowledge Base, in Salamon R, Blum B, Jorgensen M, eds.: *MEDINFO 86*; Wash., DC; Oct, 1986:117-21.

18. Lindberg DAB, Humphreys BL: Toward a Unified Medical Language, *Proc MIE 87*, Rome, September, 1987.

19. Tuttle MS, Blois MS, Erlbaum MS, Nelson SJ, Sherertz DD: Toward a bio-medical thesaurus: building the foundation of the UMLS, in Greenes RA, ed.: *Proc. 12th SCAMC*; Wash., DC; Nov, 1988:191-5.

20. McMahon LF, Smits HL: Can Medicare Prospective Payment Survive the ICD9-CM Disease Classification System?, *Ann. Int. Med.*; 1986; 104(4):562-6.

21. Packer MS, Cimino JJ, Barnett GO, Kim R, Hoffer EP, Zatz S: Updating the DXplain Database, in Greenes RA, ed.: *Proc. 12th SCAMC*; Wash., DC; Nov, 1988:96-100.

22. American Medical Computing, Ltd.: AMA/NET Diseases Information Base, AMA, 1988.

23. Bachrach CA, Charen T: Selection of MEDLINE contents, the development of its contents, and the indexing process, *Med. Inform.*; 1978; 3(3):237-54.

24. Sharpe WD, ed.: Introduction to the topography field, in Côté RA, ed.: *Systematized Nomenclature of Medicine Introduction*, 2nd Ed., Coll. Amer. Path., Skokie, IL; 1982:3.

25. Wingert F: Reduction of redundancy in a categorized nomenclature, in Côté RA, Protti DJ, Scherrer JR, eds.: *Role of Informatics in Health Data Coding and Classification Systems*; Elsevier, 1985:191-202.

26. Blois MS, Tuttle MS, Sherertz DD: RECONSIDER: A program for generating differential diagnoses, in Heffernan HG, ed.: *Proc. 5th SCMAC*; Wash., DC; Nov, 1981:3263-8.

27. Dunham G, Henson D, Pacak M:Three solutions to problems of categorized medical terminology, *Meth. Inform. Med.*; 1984, 23(2):87-95.

28. Date CJ: Chapter 1 - Basic Concepts, in *An Introduction to Database Systems*, Volume 1, Third Ed.; Addison-Wesley, Reading, MA; 1982:1-32.

29. Graitson M: SNOMED as a knowledge base for a natural language understanding program, in Côté RA, Protti DJ, Scherrer JR, eds.: *Role of Informatics in Health Data Coding and Classification Systems*; Elsevier, 1985:179-189.

30. Hollander D, Greenes RA: Identification of semantic relations between child and parent MeSH terms in the MeSH tree structures: preliminary report on first-pass analysis, Decision Systems Group, Brigham and Women's Hospital, Bos; 1988.

31. Minsky M: A framework for representing knowledge, in Winston P, ed.: *The Psychology of Computer Vision*, McGraw-Hill, New York; 1975:211-77.

32. Cimino JJ, Barnett GO: Automated Translation Between Medical Terminologies Using Semantic Definitions, in Hammond WE, ed.: *Proc. AAMSI 89 Congress*, May, 1989; 113-117.

33. Masarie F.E., Cimino J.J., Giuse N.B., Miller R.A. Mapping Between Controlled Vocabularies: QMR and DXplain. Task Report, UMLS, National Library of Medicine; April 14, 1988.

34. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA: Conceptual modeling for the Unified Medical Language System, in Greenes RA, ed.: *Proc. 12th SCAMC*; Wash., DC; Nov, 1988:148-51.

35. Wingert F: An indexing system for SNOMED, *Meth. Inform. Med.*; 1986; 25:22-30.

36. Woods W: What's in a link: foundations for semantic networks", in Bobrow B, Collins A, eds.: *Representation and Understanding*, Academic Press, New York; 1975:35-82.

37. Cimino JJ, Barnett GO, Hoffer EP, Packer MS, Hupp JA: Building the DXplain Knowledge Base, *Proc. 23rd Meet. Assoc. Advance. Med. Instr.*; Wash., DC; May, 1988:24.

38. Koton P: A medical reasoning program that improves with experience, in Greenes RA, ed.: *Proc. 12th SCAMC*; Wash., DC; Nov, 1988:32-7.

39. Gabrielli ER: Interface problems between medicine and computers; in Cohen GS, ed.: *Proc. 8th SCAMC*; Wash., DC; Nov, 1984:93-5.

40. Stefik M, Bobrow D: Object-oriented programming: themes and variations, *AI Mag.*, Jan, 1986:40-62.

41. Chen P: The entity-relation model - towards a unified view of data, *ACM Trans. - Database Systems*, 1976;1(1): 9-36.