Coping with Changing Controlled Vocabularies

James J. Cimino, M.D, Paul D. Clayton, Ph.D.
Center for Medical Informatics
Columbia University
New York, New York 10032

For the foreseeable future, controlled medical vocabularies will be in a constant state of development, expansion and refinement. Changes in controlled vocabularies must be reconciled with historical patient information which is coded using those vocabularies and stored in clinical databases. This paper explores the kinds of changes that can occur in controlled vocabularies, including adding terms (simple additions, refinements, redundancy and disambiguation), deleting terms, changing terms (major and minor name changes), and other special situations (obsolescence, discovering redundancy, and precoordination). Examples are drawn from actual changes appearing in the 1993 update to the International Classification of Diseases (ICD9-CM). The methods being used at Columbia-Presbyterian Medical Center to reconcile its Medical Entities Dictionary and its clinical database are discussed.

INTRODUCTION

One of the greatest impediments to the development of electronic medical records is the lack of a high-quality controlled medical vocabulary [1]. A number of private and public research projects are underway to help address this deficiency. An important adjunct to this research is the development of methods for coping with changes in a controlled vocabulary once it becomes relied upon.

An important use of controlled medical vocabularies is the storage of coded patient information in clinical databases. Data stored one day may be difficult to interpret the next day if the vocabulary used to encode it has changed in the interim. One method for dealing with such changes is to maintain an historical database for the vocabulary, such that the original meaning of codes can be resurrected if their storage dates are known. This strategy works well for some purposes, such as batch processing for summary reporting of archived data. However, for systems which require rapid regeneration of information from codes, such as interactive clinical record review or automated decision support, an historical coding system may be impractical. Alternatively, a change in coding could be reflected in the patient data by rewriting records in the database. But this strategy is both impractical, for a large database, and dangerous, since the original meaning of the data could be lost.

The controlled medical vocabulary used in the clinical information system at the Columbia-Presbyterian Medical Center (CPMC) is in a constant state of change. It is used successfully for storing and retrieving patient data, yet it does not rely on an historical format for maintenance. This paper describes the strategy used to retrieve historical data, coded in a changing vocabulary, without losing the meaning of the original data. Examples of vocabulary evolution will be drawn from changes in the *International Classification of Diseases, Ninth Revision, with Clinical Modifications* (ICD9-CM) [2] from the 1992 to the 1993 version.

BACKGROUND

The CPMC clinical information system is coded using the CPMC Medical Entities Dictionary (MED) [3]. The MED consists of a semantic network, based on the Unified Medical Language System (UMLS) [4], with a directional acyclic graph to provide for multiple, coexisting hierarchies. The MED grows in a monotonic manner; that is, concepts are incrementally added and, once added, cannot be removed nor have their *inherent* meaning altered. This is not to say that individual concepts may not change; however, they may only change in ways which clarify or improve their meaning *explicitly*. For example, if a term exists called "glucose test", it might be later changed to "serum glucose test" if and only if the change reflects its true meaning. If the term was previously used to code data which were actually serum test results, then the name change would be allowed. If, on the other hand, the term was used to code data which could reflect either serum or plasma test results, then the name change would be invalid. In this latter case, the original term would be left unchanged (or perhaps changed to "serum or plasma glucose tests") and two new terms

0195-4210/94/$5.00

("serum glucose test" and "plasma glucose test") would be added as descendants of the original term.

The MED currently contains over 40,300 terms drawn from a number of sources, including the UMLS, local departmental systems, and ICD9-CM. As each of these sources undergoes changes, the MED must be modified to reflect those changes. On the surface, these changes consist of name changes in existing codes, term deletions, and term additions. If the MED were only used to look up terms in current vocabularies, as one might do with the UMLS when retrieving on-line information or with ICD9-CM when filling out a coding form, simply reflecting these changes in the MED would be adequate. However, when patient data are to be encoded and stored for later retrieval and reconstitution, close attention must be paid to how alterations affect the meaning of the terms. These changes are incorporated into the MED in a systematic way, depending on the type of change involved.

RENAMING TERMS

Changes in controlled vocabularies are often detectable only because the name associated with a unique identifier differs from the name which was present in a previous version. The October 1993 update to ICD9-CM, for example, includes 47 instances of name changes. Such changes are classified as minor (no meaning change) or major (meaning change).

Minor Name Change

Minor name changes are common and (by definition) do not effect term meaning. Sometimes, change is needed to correct a spelling error. For example, in ICD9-CM code 681.10, CELULITIS was changed to CELLULITIS. In other cases, change is enacted to better reflect accepted medical terminology. For example, in code 733.1, PATHOLOGICAL FRACTURE to PATHOLOGIC FRACTURE. In still other cases, change is intended to clarify a term without changing its intended meaning. For example, code 250.11 TYPE I DIABETES MELLITUS WITH KETOACIDOSIS was changed to include the phrase NOT STATED AS UNCONTROLLED.

In cases such as these, the MED concept name is changed. Since the meaning has not changed, the meaning of the data represented by it will not be misrepresented through reconstitution using the new name (for example, by improving the spelling).

Major Name Change

In some cases, the change in a term name corresponds to a true change in its meaning. Occasionally, this might come about due to "code re-use" For example, the code 99.71 was originally assigned to the term MERCURY-ZINC PACEMAKER BATTERY but at some point between 1980 and 1992, the codes was reassigned to THERAPEUTIC PLASMAPHERESIS. More often, however, the change is due to some refinement of meaning which is, nevertheless, a change in meaning. For example, the code 354.4 changed from CAUSALGIA to CAUSALGIA OF UPPER LIMB. In cases such as these, it would be inappropriate to allow a doctor to assign the diagnosis of "causalgia" to a patient in 1992 and then, in 1993, report that the actual diagnosis was "causalgia of upper limb", since the physician might originally have meant "causalgia of lower limb", a term which was not available until 1993.

In the MED, these changes in meaning are treated as if the old terms were deleted and new terms added with the same code. The sections below describe the individual handling of a deletion and an addition.

DELETING TERMS

Terms may be deleted from a vocabulary if the creators of the terminology no longer wish to include the corresponding concept in the domain of the terminology. For example two codes were removed from ICD9-CM in 1993 (e.g., 665.14 RUPTURE OF UTERUS DURING LABOR, POSTPARTUM CONDITION OR COMPLICATION). Presumably, these codes could be reused in some later version.

The deletion of concepts poses problems for systems which have already made some use of them. For example, if a patient was noted to have a particular diagnosis on a particular date, it would be unacceptable to simply delete that fact just because the disease term was removed from the vocabulary. In many cases, however, no changes are needed to the MED. For example, if the laboratory ceases to perform a particular test, the persistence of the term in the MED is harmless - the laboratory system will simply cease to send data about that test. Meanwhile, any previous occurrences of the test remain coded in the patient databases and remain interpretable. In some cases, the term must be flagged in the MED to prevent inappropriate use. For example, in the case of deleted ICD9-CM terms, the ICD9-CM code for the "deleted" term is moved out of the "ICD9-

CODE" attribute field and into the "OLD-ICD9-CODE" attribute field. Thus, as stated above, concepts are not deleted from the MED. This approach provides a means for a data entry program to recognize that the code is no longer usable, while providing a means for interpreting previously stored occurrences.

ADDING TERMS

The periodic addition of terms to a controlled medical vocabulary is required by the evolution of the discipline of medicine. When a new concept is established, such as a new medication, disease or procedure, the creation of a new term is proper and expected. The 1993 ICD9-CM update, for example, included 160 new terms. The appropriate response to a particular concept addition depends on how the new term influences appropriate use of previously existing terms.

Simple Additions

When the new term represents a truly new concept, the proper response is simply to accept it into the vocabulary and use it when appropriate. For example, the addition of the new ICD9-CM code 704.02 TELOGEN EFFLUVIUM does not influence how any of the previously existing codes are used. Therefore, a new concept is added to the MED, corresponding to this new term.

Refinement

In many cases, one or more terms are added to allow greater levels of detail to be specified. For example, the 1992 version of ICD9-CM contained 434.0 CEREBRAL THROMBOSIS. In 1993, the codes 434.00 CEREBRAL THROMBOSIS WITHOUT MENTION OF CEREBRAL INFARCTION and 434.01 CEREBRAL THROMBOSIS WITH CEREBRAL INFARCTION were added. In cases such as these, the new terms can be added as children (in the MED hierarchy) of the existing term.

Redundancy

Sometimes, a code is added which is identical in meaning to an existing term. In ICD9-CM this often occurs in the course of adding refining terms. For example, in 1992 ICD9-CM contained 530.1 ESOPHAGITIS. In 1993, the codes 530.10 UNSPECIFIED ESOPHAGITIS, 530.11 REFLUX ESOPHAGITIS, and 530.19 OTHER ESOPHAGITIS were added. Two of these (530.11 and 530.19) were added easily as refinements. However, the new term 530.10 is synonymous with 530.1.

Adding a new concept to the MED which has the same meaning as an existing concept would introduce undesirable redundancy. Instead, the existing term is given the new code (530.10) and its preexisting code (530.1) is moved to the "OLD-ICD9-CODE" attribute field. In addition, the name of the concept is altered to reflect the new variation.

Disambiguation

If a term in a controlled vocabulary is discovered to have two or more meanings (referred to as "polysemy"), an appropriate response is to disambiguate these meanings by creating a separate term for each. No examples of such disambiguation were found in ICD9-CM updates. However, the UMLS provides several examples of disambiguation. These occur mainly because UMLS developers notice that terms with the same name in different sources may have different meanings in the different sources, or because one UMLS source vocabulary was found to have two meanings for the same term. For example, the original term "Atrium" was subsequently disambiguated into "Heart Atrium" (a body part) and "atrium <2>" (an organic chemical). An important consideration in dealing with disambiguation is to determine whether the unique identifier for the original term can be retained for use with one of the original meanings.

The appropriate response to disambiguation in the MED depends on the meaning of the term from the MED's perspective. In the above example, the MED included "Atrium", but only in its anatomic meaning. In this case, the name was changed to "Heart Atrium" and the existing concept was associated with the appropriate UMLS concept. Since the meaning of the concept was always intended to be anatomical, the name change had no effect on the meaning of the information stored in the patient database. (No concept corresponding to the chemical meaning has been added to the MED.)

There have been no situations to date in which a concept in the MED was found to have multiple meanings such that patient data corresponding to the different meanings might have been stored in the database with the same code. There is no guarantee, however, that this situation will not occur. In such an event, it may be impossible to correct the database by

determining which meaning was intended in each occurrence. Instead, some response will be needed in the MED to represent the ambiguity explicitly and prevent its recurrence.

Consider a hypothetical example in which the MED contained the concept "Paget Disease", without specifying whether it was a disease of bone or breast. To correct this situation, it would be necessary to add two new terms "Paget Disease of the Breast" and "Paget Disease of the Bone". The original term could be left with its ambiguous name; however, an alternative would be to rename the term so that the inherent ambiguity is made explicit, such as "Paget Disease, Not Specified as Bone or Breast". Of course, this is a meaningless term, from a clinical point of view; however, since it can't deleted, the new name at least makes the ambiguity explicit.

SPECIAL CASES

CPMC's experience with maintaining the MED, particularly with respect to keeping the ICD9-CM information current, has provided some additional insights into how controlled vocabularies can evolve and how changes can be dealt with in a way that maintains both the monotonicity of the vocabulary and the integrity of stored coded data. The following additional vocabulary modifications, while not found in recent ICD9-CM updates are, nevertheless, likely in the future.

Obsolescence

New knowledge often requires the addition of new terms to a vocabulary; it may also render existing terms obsolete. For example, with advances in virology and immunology, the terms "Infectious Hepatitis" and "Serum Hepatitis" were replaced by "Hepatitis A", "Hepatitis B" and "Non-A, Non-B Viral Hepatitis". This last term has, in turn, been replaced by a further collection of terms.

Although a term such as "Non-A, Non-B Viral Hepatitis" has fallen out of favor, it is not possible to remove it from vocabularies such as the MED. This is because previous patient diagnoses have been coded using the term. Even though we may now be able to differentiate a new patient's condition into Hepatitis C or E, it is generally not possible to go back and determine what a patient had in the past and recode the database. Thus, the "obsolete" concept is still valid and still has valid meaning. It must therefore be retained in the MED. The new terms can be added as refinements to the obsolete term, just as with any refining terms.

Discovering Redundancy

Redundancy is an undesirable condition in any controlled vocabulary; however there is no way to prevent it from occurring. Sometimes it occurs because synonymous terms are added without recognizing their synonymy. In other cases, the true synonymy may only be recognized through subsequent medical advances.

Consider, for example, the AIDS virus. Originally, there were two reported agents: Human T-Cell Lymphoma Virus III (HTLV-III) and Lymphadenopathy-Associated Virus (LAV). Since the original description of these agents, they have been recognized to be identical (now named Human Immunodeficiency Virus-1, or HIV-1). The easiest solution would be to discard one term and save the other, with renaming or addition of synonyms as appropriate. Unfortunately, this would render data coded with the discarded term to be uninterpretable. An alternate solution is to create a class which includes all the redundant terms. The new superclass would be the preferred form and the child terms can have a pointer to that term to indicate its preferred status. Thus, in the above example, HIV-1 would become the superclass for HTLV-III and LAV. When reconstituting the coded data, the preferred name could be obtained and, when retrieving all cases of HIV-1, including a search for all descendant terms would retrieve all appropriate instances of any of the three codes. In this way, the redundancy can at least be made transparent, if not totally eliminated.

Precoordination

One troublesome way which controlled vocabularies evolve is by adding more specific, "precoordinated" terms. For example, coding Type I Diabetes with Hyperosmolarity used to require two "atomic" codes: 250.8 DIABETES WITH OTHER SPECIFIED MANIFESTATION and 276.0 HYPEROSMOLALITY AND/OR HYPERNATREMIA. ICD9-CM now provides a single convenient code (250.23 TYPE I DIABETES WITH HYPEROSMOLARITY). Querying a database which has stored patient diagnoses both ways may produce undesirable results, unless the dual representation is recognized. However, reliably detecting such situations is problematic.

The CPMC clinical database does not yet provide a simple way to query for data which might be coded

as atomic, precoordinated or both. However, the MED does provide a facility for assisting with such a process. The UMLS Semantic Net provides for concepts to be related through a has-part/part-of relation. Thus, the MED is capable of including the information that 250.23 TYPE I DIABETES WITH HYPEROSMOLARITY has parts TYPE I DIABETES and HYPEROSMOLARITY. As a result, terms can be given a "molecular" appearance, such that a single precoordinated term can be disassembled into its constituent atomic elements. Similarly, when given a set of atomic terms, the MED can be searched to locate a corresponding precoordination, if one exists. It should therefore be possible to design a database retrieval routine which can take advantage of this feature of the MED in order to cope with multiple codings of the same information.

DISCUSSION

Controlled vocabulary evolution is a fact of life for clinical system developers. Unless care is taken, patient data stored in a compact, useful coded form may become obsolete if they become uninterpretable due to vocabulary changes subsequent to their storage. Rather than wait for change and then attempt to retrofit old data to new codes, it will be imperative for developers of electronic medical record systems to anticipate the types of changes which may occur. This paper proposes an initial formal framework by which vocabulary changes can be classified and addressed.

CPMC has expended considerable effort to develop automated vocabulary maintenance methods. These methods have proved useful for applying changes in source vocabularies to the content of the MED. However, the right method can only be applied when the type of change is well understood. At present, no method exists which can automatically decide the type of change and the appropriate response (e.g., to differentiate between a minor and major name change). Instead, manual review by domain experts is needed. One reason is that vocabulary changes usually do not include information about the reason for the change. Such additional information could enhance vocabulary management, particularly if the information were to be included in a structured, machine-readable format. For example, if each disease term included references to involved body parts, then it might be possible to distinguish a minor name change (say, from PATHOLOGICAL FRACTURE to PATHOLOGIC FRACTURE, where both terms would have a reference to BONE) from a major name change (e.g., the change from CAUSALGIA to CAUSALGIA OF UPPER LIMB, which would entail the addition of new body location information).

The CPMC approach offers a means to retrieve patient information based on concepts of interest. It does not guarantee that retrieval can be done by ICD9-CM code, nor that the original form of the code (e.g., a misspelling) can be reconstructed. Old codes are kept in the MED, but the date of changes are not, nor could the correct old code be determined if the concept's code had been changed more than once. However, if such a reconstruction were needed, it could be handled by including the ICD9-CM code with the MED code in the clinical database at the time of storage or by reviewing the log files of MED changes which will identify when and how concept information was modified.

Additional research in vocabulary maintenance is needed and the lessons learned must be fed back to the developers of controlled vocabularies, such as ICD9-CM. For example, the repercussions of seemingly arbitrary actions such as major name changes and term deletions need to be clarified so that developers will be aware of the needs of the users and users will better understand the intentions of the developers. This paper attempts to define a taxonomy for describing types of vocabulary changes and offers one set of approaches for dealing with them.

Acknowledgments

This work was supported in part the IBM Corporation and the National Library of Medicine.

References

1. United States. General Accounting Office. *Automated Medical Records: Leadership Needed to Expedite Standards Development.* Washington, D.C.: USGAO/IMTEC-93-17; April 1993.
2. United States National Center for Health Statistics. *International Classification of Diseases, Ninth Revision, with Clinical Modifications,* Washington, DC; 1980.
3. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB: Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA*, 1994; 1(1):35-50.
4. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inform Med.* 1993; 32(4):281-291.