MEDINFO 95 Proceedings. R.A. Greenes et al. (editors)

Managing Vocabulary for a Centralized Clinical System

Cimino J. J.[a], Johnson S. B.[a], Hripcsak G.[a], Hill C. L.[a], and Clayton P. D.[a]

Department of Medical Informatics, Columbia University, New York, New York 10032, USA

The clinical computing environment at Columbia-Presbyterian Medical Center is organized around a centralized database of coded patient information collected from various ancillary sources. The Medical Entities Dictionary (MED) is the central repository for the controlled vocabulary used to encode the patient data. The MED is composed of terms used in the ancillary departments and, as such, changes in the source vocabularies must be maintained in the MED. The MED also contains some basic knowledge about the terms, and sophisticated maintenance tools have been developed that take advantage of this knowledge. This paper describes the success of the knowledge-based approach by describing the techniques used in two tasks: addition of a new vocabulary and maintenance of an existing one.

1. Introduction

One common scenario for clinical information systems is the departmentalized approach: each ancillary department has its own system for generating, using and maintaining data in a specialized domain (laboratory, pharmacy, radiology, etc.). In such a scenario, there are several options for giving health care providers access to this data. At the Columbia-Presbyterian Medical Center (CPMC), a centralized clinical database has been created to collect clinical data from diverse systems and to act as a clinical information server for client applications such as results review, adverse event monitoring, clinical research, and quality assurance [1]. A critical feature in the central database is the ability to code the data and support their use by these applications. Each ancillary system uses its own approach to code its data. Simply storing these codes in the central system is inadequate, since the client applications may not know each coding system. In fact, the client applications may make use of their own coding schemes, which may not be compatible with those of the ancillary systems. Application developers can strive to map the ancillary vocabularies into those of their own systems; however, such mapping is extremely sensitive to changes in the vocabulary that may change in ancillary vocabularies as new systems are added or existing ones are modified [1, 2].

2. CPMC Approach

One solution would be to adopt a general-purpose controlled vocabulary, such as the Systematized Nomenclature of Human and Veterinary Medicine (SNOMED International) [3]. Unfortunately, the ancillary and client systems at CPMC do not use vocabularies that are readily represented in SNOMED, nor are they likely to do so. It has been necessary at CPMC, therefore, to create a centralized controlled vocabulary, the Medical Entities Dictionary (MED). The MED is a repository for the source and client vocabularies and serves to allow the integration of these vocabularies [4]. For example, the laboratory system includes several codes for specific tests measuring glucose in serum, while the clinical alerting system includes a more general term "Serum Glucose." In the MED, the more specific terms are positioned as children of the more general term. Thus, when the alerting system requests a "Serum Glucose," the MED can provide the codes for the more specific tests that actually appear in the patient database.

The MED includes features which are intended to facilitate vocabulary maintenance. One of these features is the inclusion of formal definitional knowledge about the terms. For example, the MED includes the knowledge that laboratory tests have specimens and measure substances. When a new serum glucose test (e.g., "Stat Serum Glucose") is added to the laboratory vocabulary, it is added to the MED with the additional knowledge (implicit in the name but made explicit in the MED) that the test has serum as a specimen and measures the substance glucose. This information allows it to be placed automatically with the other serum glucose tests so that it will be properly recognized by the clinical alerting system.

Since the original description, and subsequent implementation [5], of the MED design [5], significant experience

MEDINFO 95 Proceedings. R.A. Greenes et al. (editors)

has been gained with the use of knowledge-based techniques for vocabulary maintenance. The purpose of this paper is to describe that experience in two tasks: adding a new ancillary vocabulary (laboratory) and maintaining an existing one (pharmacy).

3. Adding a New Laboratory Vocabulary

The first terms to be placed in the MED in 1989 were those of the CPMC clinical laboratory system. This locally-developed system included 2,309 terms for individual tests, procedures (panels of tests), specimens, and test results. Over the subsequent five years, the laboratory vocabulary remained relatively stable, with the addition of 224 terms (less than a 10% increase).

In 1994, a new commercial laboratory system was purchased. The new system provided a number of new capabilities for the laboratories, including opportunities to expand the coding of tests and results. As a consequence, the new vocabulary consisted of 5,291 terms. Although there was some correspondence between the old and new vocabularies, there were no formal translation capabilities and, due to some changes in the modeling of tests and results in the new system, one-to-one correspondence was rare. Therefore, it was necessary to place all of the terms from the new vocabulary into the MED as entirely new entities requiring their own classification. The new vocabulary was not finalized until mid-June 1994 and had an expected go-live date of July 24, 1994. Given this time frame, attempting to add the 5,291 new terms in a coordinated manner was not possible using manual methods.

The new terms were analyzed with a combination of automated lexical and knowledge-based approaches. Lexical methods included partially matching new terms with existing terms. For example, new test names were compared to existing terms for matching words. Test names were also compared to substance terms to help create the links between tests and the substances they measure.

Knowledge-based techniques made use of links between terms to help with automated classification. For example, tests were initially placed in the MED as children of the term "Single-Result Laboratory Test." Once a test was linked to a particular specimen (using information available from the laboratory system) and to a particular substance (using lexical techniques), it was then possible to "push" it down the hierarchy by comparing it to all of the other children in this class.

In some cases, the new terms were similar enough to old terms to justify the creation of a new class to include them both. For example, the old system had only one test which involved the measurement of ceruloplasmin, so no class of "ceruloplasmin tests" was created. However, with the addition of the new "ceruloplasmin test" term, the creation of such a class was desirable. When the class detection feature of the MED Editor was applied to the class "Serum Chemistry Test," it detected the similarity between the two tests and suggested that a new class be added that would include the two tests as children (see Figure 1).

By July 4, all of the new terms had been successfully added to the MED, with complete classification approximately half accomplished by that time. As of this writing, the classification process is continuing, with every expectation of completing the task well before July 24. This accomplishment would not have been possible using manual methods for inserting the 5,291 new terms into the centralized vocabulary.

4. Maintaining the Pharmacy Vocabulary

In 1993, CPMC purchased a commercial pharmacy system. This system contributed several new vocabularies to the MED, including American Hospital Formulary System (AHFS) codes [6], allergy classes, Drug Enforcement Agency (DEA) classes, and the medications themselves. For the most part, these terminologies are stable. However, the hospital formulary is constantly changing and, since the initial 2,091 medication terms were placed in the MED, an additional 1,402 have been added. These inclusions can occur on a daily basis and, as soon as they occur, they may appear in the drug order information which is transmitted from the pharmacy system to the central patient database. Therefore, it has become imperative that the MED updates occur in a timely manner.

In response to this need, an automated maintenance program was created. This program directly accesses, via a local area network, a formulary file on the pharmacy's computer and compares the information in this file with the information in the MED. When information about a medication has changed, the program calls the MED Editor functions needed to make the update. When a new medication is encountered, the program determines a likely position in the MED hierarchy for placing the new medication, based on information such as its AHFS code, generic ingredi-

ents, allergy codes, etc. In the fully automated mode, the program then adds the new medication to the MED. In the interactive mode, it suggests a position and allows the user to agree or to browse the MED and select a new position. For example, during a recent update, the program encountered the drug "Erythromycin Estolate 25mg/ml" and suggested adding it as a child of the class "Macrolide Antibiotic Preparations." When the user examined this class in the MED (see Figure 2), a better class was found ("Erythromycin Preparations") which was "exported" to the update program as the appropriate class for the new medication term. This combination of suggested placement, browsing, and automated addition has greatly simplified the daily maintenance of the formulary changes in the MED.

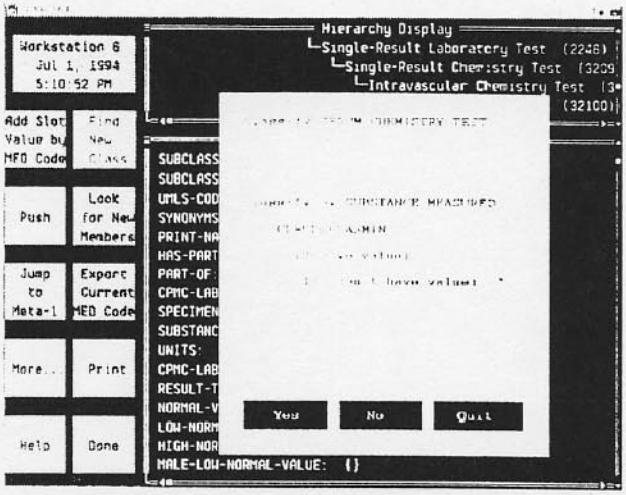

Figure 1. The MED Editor suggesting a new class: tests that measure ceruloplasmin

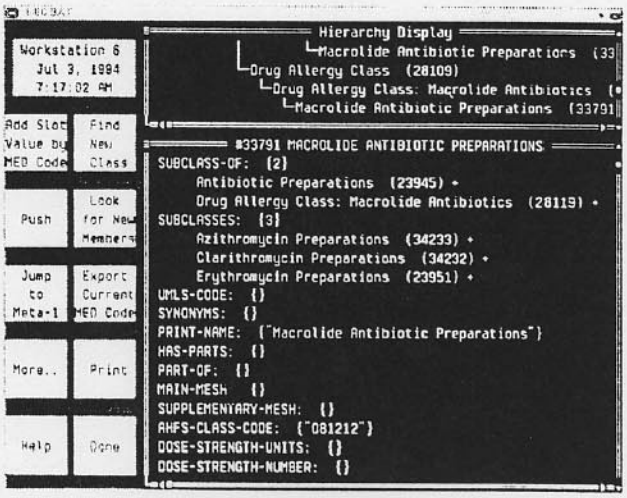

Figure 2. Browsing the MED to identify an appropriate subclass for the term Erythromycin Estolate

The update program often finds discrepancies between the formulary information and knowledge in the MED. For example, during a recent update the program generated the following message:

91054: Streptomycin 1 gm Vial in allergy class AMINOGLYCOSIDES but doesn't have code 10

This was generated because the MED had placed the drug in the Aminoglycoside class but found that the pharmacy system did not assign the corresponding allergy code. Sometimes this message occurs when a term is misclassified in the MED. Usually, however, the pharmacy system was missing important information (as in this case). Such a situation could lead to the pharmacy allowing streptomycin to be prescribed to a patient with an allergy to that

MEDINFO 95 Proceedings. R.A. Greenes et al. (editors)

drug. When such information is discovered during MED editing, it is provided to the pharmacy so that the formulary database can be corrected. The pharmacy, in turn, provides this information to the database vendor who, in turn, can correct the information provided to thousands of pharmacies throughout the United States. Thus, the knowledge-based approach can enable the MED to be better than its constituent parts.

5. Conclusions

The CPMC scenario for centralized, reusable clinical information requires the capability to centralize the myriad local ancillary-controlled vocabularies into a single resource. The work necessary to create such a resource is significant. Addition of terms to the CPMC Medical Entities Dictionary requires extra effort in order to include knowledge about the terms being added. The original argument to support this approach was that sophisticated vocabulary maintenance tools could be created that would permit automation of some of the work and yield a more consistent, higher quality vocabulary. Experience to date has substantiated this claim. New vocabularies are added much sooner than would be possible with manual methods, and the quality has been higher than that of the original source vocabularies.

6. References

[1] Clayton PD, Sideli RV, Sengupta S. Open architecture and integrated information at Columbia-Presbyterian Medical Center. *MD Computing*. 1992; 9(5):297-303.

[2] Huff SM, Craig RB, Gould BL, Castagno DL, Smilan RE. Medical data dictionary for decision support applications. In Stead WW, ed. *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care*. Washington, DC; IEEE Computer Society Press, New York, 1987:310-317.

[2] Cimino JJ, Clayton PD. Coping with changing controlled vocabularies. In Ozbolt JG, ed: *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care*. Washington, DC; November, McGraw-Hill, New York, 1994: (*in press*).

[3] Côté RA, Rothwell DJ, Beckett RS, Palotay JL, Brochu L, eds. *The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield, Illinois. College of American Pathologists, 1993.

[4] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *Journal of the American Medical Informatics Association*, 1994; 1(1):35-50.

[5] Cimino JJ, Hripcsak G, Johnson SB and Clayton PD. Designing an introspective, controlled medical vocabulary. In Kingsland LW, ed. *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*. Washington, DC; November, IEEE Computer Society Press, New York, 1989:513-518.

[6] American Society of Hospital Pharmacists. *American Hospital Formulary Service Drug Information*. Bethesda (MD), 1988.