Maintaining The Integrity of a Web-Based Medical Vocabulary Glossary

Ron M. Aryel, M.D., James Cai, M.S., Henry C. Chueh, M.D., M.S.
G. Octo Barnett, M.D.
Laboratory Of Computer Science
Massachusetts General Hospital
Boston, Massachusetts

Medical vocabulary databases are vital components of electronic medical record (EMR) systems. While their performance and efficiency has been extensively explored by many authors, few have dealt with the maintenance of their semantic integrity. Clinicians' preference for an optimistic system has introduced a need for monitoring and filtering proposed additions to the vocabulary tables. We propose the use of batch processing and memo-posting as means of implementing a World-Wide-Web-based Controlled Vocabulary Glossary as a monitored optimistic system.

INTRODUCTION

Most emerging electronic medical record (EMR) systems include a controlled medical vocabulary which includes such categories as diagnoses, procedures and medications. In a solo practice such a glossary might exist on the same computer which also stores patient records; in a larger organization, such as a hospital or group practice, a centralized database, accessible to all and easy to update, is more desirable. The essential components of a vocabulary service include administrative tools, that is, tools used to build and maintain the vocabulary structure and content; browsing tools and interfaces with useful medical applications.[1] Increasingly, both user access and maintenance of EMR systems are occurring over the World-Wide Web.[2,3] The importance of vocabulary maintenance tools has been acknowledged by more than one author; terms become obsolete (for example, diagnoses are revised when researchers learn important new details about diseases), medications are added or dropped from formularies and terms are added whose synonymy is only later recognized.[4] While the recent medical informatics literature has dealt extensively with the semantic, ontological, and search efficiency issues, few authors have been concerned with the implementation of maintenance tools and the implications for the integrity of shared databases. One consequence of this neglect is the introduction of errors in the database, such as redundancies, erroneous terms and illogical associations between terms.

BACKGROUND

Conceptual approaches to the creation and maintenance of medical vocabulary databases may be characterized as optimistic or pessimistic. The optimistic model allows the database's users to add terms to, or modify terms in, the database themselves; the pessimistic model requires the consent of a gatekeeper, such as a data librarian, for this activity. The optimistic model offers users easier access and faster response at the expense of the vocabulary's semantic integrity. The pessimistic model can prevent the introduction of unwanted data but restricts administrative access to the database; its effectiveness depends on the quality of the gatekeeper. (See Figure 1)

In 1996, John Tu implemented a controlled medical vocabulary system which serves an EMR system used by two practice groups affiliated with the Massachusetts General Hospital.[5] The vocabulary itself could be located on any server deemed convenient to its users. A user (physician and/or transcriptionist) queried the database by entering an index word, which was generally a parsed word in the desired term, but could also be a synonym, abbreviation, root word or a common misspelling. Each query returned a result set of the most likely matches, ranked in order of likelihood of matching based on weighing individual words as well as their usage

1091—8280/97/$5.00

FIGURE 1: VOCABULARY ADMINISTRATION

MODEL	GATEKEEPER?	ENTRY'S INCLUSION IN DATABASE	EFFICIENCY TO GATEKEEPER	FEEDBACK TO USER
Pessimistic	Yes	Delayed	No	No
Optimistic	No	Immediate	Yes	Yes
Monitored Optimistic	Yes	Delayed	Yes	Delayed

frequency. As more letters were typed, the returned list became shorter and more specific.

The system was optimistic in nature, seen as more "user friendly" in that it allowed physicians to add terms in a timely fashion. It was intended to supplement, and its successors to ultimately replace, an earlier EMR system called COSTAR. [6] COSTAR is a pessimistic system; physicians who desire to add vocabulary terms to the COSTAR database must inform the Laboratory of Computer Science (LCS) that an addition is desired. An LCS fellow must then create the entry, utilizing a terminal emulation program.

While granting users direct read/write access to the database offers them a great deal of convenience, it also allows the database to be populated with redundant, or incorrect data due to the gatekeeper's removal. Since patient records, once entered, by law cannot be altered and since the vocabulary database is directly tied into those records, misspelled or incorrectly-employed terms must be kept and accounted for and corrected terms added to the database. Were a researcher to become interested in utilizing the EMR to collect statistics, extensive auditing of the data would be required, chiefly in the form of deciding which listed diagnoses, procedures or medications represented the same term. Educating users could help reduce, but not eliminate, mistakes.

In general, a vocabulary service's administrative applications should be accessible only to supervisory staff familiar with its architecture and content, and only verified and sanctioned data should be entered. [7] Thus, an optimistically-designed vocabulary database should be equipped with a semantic and logical validation process which, on the one hand, will protect the database and, on the other hand, not impose significant costs on the vocabulary service. To illustrate this challenge, one may consider a process whereby a gatekeeper (generally a physician; for example, an informatics fellow) reviews requested entries case-by-case, as they arise, with the involved health-care providers. This would potentially require a significant time commitment by providers, something which would be understandably unwelcome in a busy practice or hospital setting. Moreover, this process would not be the most parsimonious to the gatekeeper. Hence, in order to make practical the human review of submitted entries, "real-time" review must be replaced by a mechanism allowing an aggregate review of data at a time appropriate to the reviewer. This may be accomplished by supplying the reviewer with a list of entries rather than by making individual requests, but this again imposes the cost of the reviewer's time in actually entering the requested items.

One very important concept is that while human reviewers must inspect potential entries, the entry process itself need not require their supervision. [7] This creates an opportunity to devise a cost-effective monitoring process. This challenge is analogous to the problem faced by commercial banks in introducing automatic teller machine (ATM) services to the public. Branch employees could enter transactions directly into account databases, but customers could not be given the same access. The ATM could accurately dispense cash for a withdrawal, but could not verify that the contents of a deposit envelope matched the transaction typed in by the customer, creating an opportunity for costly errors. The answer selected by the banks lay in the combination of two very common techniques, both originating from the earliest uses of computers: batch processing and memo-posting. Batch processing is the processing of data in a fully automated, non-interactive

fashion; memo-posting involves the recording of a transaction (whether financial or involving a change in a medical vocabulary table) in a temporary table which is intended to be reconciled with the master database's tables at a later time. This temporary table may reside on the same server as the master database, or it may reside on a client machine (for example, the ATM, or a workstation in a physician's office). [8]

Semantic and logical validation in an optimistic database model may be accomplished in a cost-effective manner by the use of batch processing. Combining on-line query capability with batch processing offers both convenience to health-care providers and a cost-efficient, secure means of updating that database. Users may use familiar on-line interfaces to propose the addition of desired vocabulary terms and relationships, and receive feedback that their requests have been logged. The proposed terms are added to the database after appropriate review and editing. Much of the editing may be performed by the batch process itself, such as rejecting or resolving misspellings, rejecting illogical parent-child relations, and detecting the presence of duplicate terms, leaving only more conceptually difficult decisions to a physician-reviewer acting as a professional data librarian. Decisions requiring a physician would include determining the proper place of a new term in a logical hierarchy (for example: Tetralogy of Fallot falls under Congenital Heart Disease). Batch database updates may be scheduled as often as desired, and processing can take place during times of little demand on the database (for example, the middle of the night or during the weekend) instead of competing for attention during peak hours and degrading service to users. The physician(s) in charge need not spend a great deal of time editing entries on-line.

IMPLEMENTATION

We sought to demonstrate the use of a gatekeeper-augmented, or "monitored," optimistic model to assure the integrity of a database by building a small controlled medical vocabulary service, available for query through the World-Wide Web. A Web-based vocabulary service consists of a Web server, a relational database, and an interface between the two to allow users to send queries to the database and receive a reply.

Hence, the Controlled Vocabulary Glossary (CVG) consists of a master vocabulary table, temporary memo-post table, an index table, and a batch transaction processor. The tables themselves may reside on any relational database, but a prototype system was built using Access (Microsoft Corp.). Users may access the database via a Web page generated by an HTML script. The Web-database interface generator chosen was Cold Fusion (Allaire Corp.).

A user completes a submittal form, transmitted by Cold Fusion, to a temporary memo-post table. The submission form allows users to specify where in the hierarchical structure they wish terms to be included; they may also defer that decision to the LCS staff. After a physician reviews all the proposed entries, he or she updates the master vocabulary table by launching a batch transaction processor written in Visual Basic (Microsoft Corp.). Visual Basic was chosen for its ease of maintenance as well as compatibility with emerging programming standards at this hospital's parent organization, but any of a number of other programming languages such as C++ or Pascal would have been adequate for this task.

The batch-processing program filters the memo-post table, then adds each term to its proper place in the master table, placing rejected data in an exception report generated for review by LCS staff. It then rebuilds the index table, redefining as needed the relationships between the terms. Once the program has finished executing, the new, validated terms are available for query; the preprocessed hierarchical index provides for a more efficient, faster query.

The integrity of the master vocabulary glossary is assured because while any authorized user may query the database (indeed, the medical vocabulary itself, not including patient records, contains no confidential information and could be made available to anyone), the only users allowed write privileges to the master tables are the batch transaction processor and the CVG data librarian.

DISCUSSION

There are some limitations inherent to this type of monitored optimistic system.

Depending on the schema design, the glossary may be unavailable during the master table's batch update; the amount of time required will vary depending on the number of terms being added and the size of the table. If 24-hour access is required, however, this limitation may be overcome by "atomizing" the entries or by representing the batch program to the computer as another "user," submitting transactions in a queue with other users. The price to be paid will be some degradation of response time to queries.

Another important limitation involves the need to add new terms at the time that patient records are generated, on EMR systems which bind the controlled medical vocabulary itself to specific patient records. If a vocabulary term cannot be added immediately, the patient's record is incomplete. This problem can be addressed in three practical ways: allowing the option of "free text" entries not tied to a glossary[6]; breaking the mandatory link between a structured document and the database, that is, storing the document itself whether or not a key term within it is listed in a database, and requiring a rescanning of the document each time a query is made based on a vocabulary term; and by the use of a reconciliation program. The latter two methods can be effectively used together.

A structured record with tag-identified fields can exist in its own right without the existence of any given term in the master database. For example, if a patient's problem list contains a new unverified diagnosis, the document can still be stored; retrieval of that document by that diagnosis will not be possible, though other fields with proven terms will remain available. Once the new diagnosis is reviewed and entered into the master table at a later time, that patient's document can then be easily retrieved. This makes it possible to freely erase or write over the temporary tables holding data when they are no longer needed.

In some cases it may be desirable to maintain vocabulary tables which are not included in the master database. A reconciliation program can serve as a link between the master table and local tables existing elsewhere whose vocabulary may occasionally be needed or offer customized data for one particular health-care provider. An example is a physician's own specialized research vocabulary linked to the master database.

The reconciliation process is based on programs utilized in ATM's, allowing banking clients to see their account balances while denying them direct access to unverified sums. This is already in use in some existing systems.[3] When a desired term cannot be found in a master database, a search is performed against other tables containing as yet unverified or unsanctioned data. When found, these term or terms can be shown to the user along with their proposed relationship to other terms in the master database. Yet another methodology involves the construction of local servers with controlled vocabularies, each of which may access terms available at a shared central repository or another local site. [2]

We chose to provide the Web-based CVG with a reconciliation program, interacting with the database in the following way: When a user queries the master vocabulary table while generating a patient document, the Cold Fusion-assisted query first searches the master table. If the requested term did not yet exist there, local glossary tables (identical in structure to the temporary memo-post table) are polled to determine if the desired term exist on one of them. These tables can exist anywhere, on the server containing the master table, or on any client's computer, but their existence and address must be known by the database. If found, the reconciliation program can include the requested term in the query display, while indicating that the master table does not contain it. If a particular parent-child or synonym-term relationship were desired, the reconciliation program could show how the proposed term would fit into the master table, once verified and entered by the batch file processor.

The implementation of a monitored optimistic database query system on the World-Wide Web can offer users most of the advantages of the optimistic model; at the same time, the system's semantic and logical integrity, vital to the database's utility, are protected by a feature borrowed from the pessimistic model. A prototype CVG system is currently undergoing preliminary testing at LCS. Data will be collected with the goal of improving its performance.

With the increasing attention being paid by government and industry to controlling health-care costs, cost-effectiveness will likely

be a very important criterion for evaluating data-processing applications. As more organizations begin to rely on large vocabulary glossaries, their integrity will become more important and, paradoxically, more difficult to assure unless the means of for this integrity have been anticipated in the initial design. We have presented one practical proposal here.

This work is supported by National Library of Medicine Training Grant LM 07092.

REFERENCES:

1. Gennari JH, Oliver DE, Pratt W, Rice J, Musen MA. A Web-based architecture for a medical vocabulary server. In Gardner RM, ed. Proc 19th SCAMC. JAMIA 1995; 275-279

2. Cimino JJ, Socratous SA, Clayton PD. Internet as clinical information system: Application development using the World-Wide Web. JAMIA 1995; 2:273-284

3. Kittredge RL, et. al. Implementing a Web-based clinical information system using EMR middle layer services. In Cimino JJ, ed. Proc 20th SCAMC. JAMIA 1996; 628-632

4. Cimino JJ, Clayton PD. Coping with changing controlled vocabularies. In Ozbolt JG, ed Proc 18th SCAMC. JAMIA. 1994; 135-139

5. Chueh HC, et. al. A component-based, distributed object services architecture for a clinical workstation. In Cimino JJ, ed. Proc 20th SCAMC 638-642

6. Barnett GO, Zielstorff RD, Piggins J, et al. COSTAR - a comprehensive medical information system for ambulatory care. In Blum B, ed Proc 6th SCAMC. Computer Society Press (New York)1982.

7. Rocha RA, Huff SM, Haug PJ, Warner HR. Designing a controlled medical vocabulary server: The VOSER project. Computers and Biomedical Research 1994 (27) 472-507.

8. Private Communications, Caruthers D, Director of Engineering, Citicorp. Transaction Technology, Inc., Los Angeles, CA.; 1996