Journal of Biomedical Informatics 42 (2009) 494–503

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Practical experience with the maintenance and auditing of a large medical ontology

David Baorto [a,*], Li Li [a], James J. Cimino [b,c]

[a] New York Presbyterian Hospital, 622 West 168th Street, VC-5, New York, NY 10032, USA
[b] Department of Biomedical Informatics, 622 W 168th Street, VC-5, New York, NY 10032, USA
[c] National Institutes of Health Clinical Center, 10 Center Drive Bethesda, Maryland 20892, USA

ARTICLE INFO

Article history:
Received 22 July 2008
Available online 12 March 2009

Keywords:
Terminology
Medical Entities Dictionary
Maintenance
Semantic network
Auditing
Vocabulary
Ontology

ABSTRACT

The Medical Entities Dictionary (MED) has served as a unified terminology at New York Presbyterian Hospital and Columbia University for more than 20 years. It was initially created to allow the clinical data from the disparate information systems (e.g., radiology, pharmacy, and multiple laboratories, etc.) to be uniquely codified for storage in a single data repository, and functions as a real time terminology server for clinical applications and decision support tools. Being conceived as a knowledge base, the MED incorporates relationships among local terms, between local terms and external standards, and additional knowledge about terms in a semantic network structure. Over the past two decades, we have sought to develop methods to maintain, audit and improve the content of the MED, such that it remains true to its original design goals. This has resulted in a complex, multi-faceted process, with both manual and automated components. In this paper, we describe this process, with examples of its effectiveness. We believe that our process provides lessons for others who seek to maintain complex, concept-oriented controlled terminologies.

1. Introduction

When medical centers create central clinical data repositories, they generally find a need for a central controlled terminology by which to code data from disparate sources (such as test results from laboratory systems and medication orders from pharmacy systems). Although the mapping of local terms to a standard terminology might offer advantages, this option has not been practical due in part to the lack of a satisfactory single standard and due in part to the lack of standards adoptions by the local data sources. Instead, developers have found the need to create a unified terminology, consisting of the merger of local terminologies, similar to the approach taken by the National Library of Medicine to unify standard terminologies into a Unified Medical Language System (UMLS) [1]. Early examples of this approach include the development of the Directory in the Computer-Stored Ambulatory Record system (COSTAR) [2] and PTXT in the HELP system [3].

Over time, some of these terminologies have evolved into ontologies, as their content has expanded to include biomedical knowledge, application knowledge, and terminologic knowledge. Notable examples include the Vocabulary Server (VOSER) used in 3 M's Health Data Dictionary [4] and the Vanderbilt Externalized General Extensions Table (VEGETABLE) [5]. The expansion and maintenance of these terminologies requires significant effort on the part of the developers, with constant vigilance towards continued maintenance of terminology quality [6]. The terminologic knowledge they contain adds to the burden of keeping the content accurate, but also provides some support for the task in the form of knowledge-based terminology maintenance.

The original plan for the clinical information system being constructed at Columbia University and the New York Presbyterian Hospital (NYPH, formerly Presbyterian Hospital) in 1988 required that a single coding system be used to encode data acquired from multiple sources, for storage in a single, coherent data repository [7]. The data sources did not use the same (or often, any) standard terminology, but no single standard terminology existed to which the source terms could be mapped. Rather than attempting to create a comprehensive controlled terminology ourselves, we sought to create a "local UMLS" that brought together the disparate controlled terminologies used by source systems into a single conceptual dictionary of medical entities that could serve as that comprehensive terminology. From the beginning, this Medical Entities Dictionary (MED) was conceived as a terminologic knowledge base that could be used to support its own maintenance and auditing [8]. As such, it has proven to be a fertile substrate for terminologic research by ourselves [9] and others [10,11]. However, the MED supports a number of important day-to-day patient care, educational, research and administrative operational activities at NYPH and Columbia [12]. Thus, the auditing of its content, like similar efforts at other medical centers, is more than an academic exercise.

* Corresponding author. Fax: +1 212 305 3302.
E-mail address: baorto@dbmi.columbia.edu (D. Baorto).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.03.005

The process by which we maintain the MED, including auditing for continual quality monitoring, has evolved over the past two decades as a result of concerted informatics research into the application of good terminology principles, together with intensive analysis of data sources and their terminologies. The purpose of this paper is to describe the requirements that shaped the maintenance process, and to describe that process itself (with special attention to auditing and error detection) that has resulted from those requirements.

2. Requirements

2.1. Terminology model

The MED was designed along the lines of the UMLS: when a term from a terminology was added to the MED, it was to be mapped to an existing concept identifier (MED Code) if an appropriate one already existed in the MED. If not, a new MED Code would be created to accommodate the term. Like the UMLS Concept Unique Identifiers (CUIs), MED Codes could correspond to multiple terms from multiple terminologies.

There were, however, several important differences. First, there was no assumption that different terminologies would necessarily contain terms that were synonymous across the terminologies (that is, terms mapping to the same MED Code). In fact, the opposite was generally considered to be the case. For example, if two laboratory systems included terms for a serum glucose test, these were considered to refer to distinct entities in reality, and therefore were given unique MED Codes. Their similarity was instead captured by making each concept a child of a MED class called "Serum Glucose Test" [9].

A second departure from the UMLS model was to attempt to include in the MED formal definitional information about each term, to the extent possible and practical, expressed through semantic relationships between MED concepts. For example, each laboratory test concept was to be related to appropriate MED concepts through "Substance Measured" and "Has Specimen" relationships, while each medication concept was to be related to appropriate MED concepts through "Has Drug Form" and "Has Pharmaceutic Component" relationships.

Other ways in which the MED approach differed from the UMLS included the organization of all concepts into a single directed acyclic graph of "is-a" relationships (with "Medical Entity" as the sole top node), the assignment of unique preferred names for each MED Concept that attempted to convey the meaning of the concept (as opposed to the sometimes-telegraphic names from source terminologies), and the introduction of new concept attributes (including the potential for semantic relationships) at single points in the "is-a" hierarchy. As the MED developed, auditing methods were needed to assure adherence to all of these requirements.

2.2. Sources

As the clinical information system at Columbia grew to include new data sources, the MED needed to incorporate the relevant terminologies. Initial sources included the laboratory, radiology, pathology and billing systems. Later sources included many other systems in ancillary departments of the medical center. For the most part, systems had their own local terminologies (or set of terminologies) that were maintained in a variety of ad hoc ways, in disparate systems and formats. Applications that were developed as part of the clinical information system (such as clinician documentation and laboratory summary reporting) often had their own terminologies as well. As systems and applications were replaced, their successors often came with new terminologies that had to be added to the MED, while retaining the retired terminologies to allow proper interpretation of historical patient data.

National and international standard terminologies were not initially included in the MED, since they were not used by source systems. Over time, however, some adoption of standards began, adding to the terminology requirements of the MED.

Finally, we found that we often needed to add our own terms to the MED to support the knowledge representation requirements. Such knowledge included classification terms (such as the "Serum Glucose Test" class) and terms needed to support definitions (such as "Digoxin", to allow the proper representation of terms such as "Serum Digoxin Test" and "Digoxin 0.25 mg Tablet").

2.3. Publishing the MED

The complex requirements for developing and maintaining MED content precluded the simple approach of including the MED in the clinical information system database and editing it in that environment. Instead, we needed a more flexible, dynamic environment for editing, which led to the added requirement for publishing the MED in a way that made it available to the clinical information system and other systems as well. As this system evolved into a Web-based architecture, the need to distribute the MED to additional environments increased further.

Originally, the MED was maintained in a PC-based LISP environment, using commercial knowledge representation software. A simple table-based representation was exported that could be incorporated into the database of the clinical information system. When the MED outgrew this environment, we moved to a mainframe-based version of the product but soon the MED outgrew that as well, with a deterioration in performance. We then developed a "temporary" MUMPS-based solution that was used for over ten years as we worked to develop tools more appropriate to a modern, distributed, Unix-based environment. Although these transitions were disruptive to the maintenance processes, the same export mechanism was used by each version, so that the clinical information system continued to function without interruption.

3. Solutions

Some of the requirements described above were determined at the outset of the MED development [8]. However, many other requirements were established over the ensuing years, sometimes by natural evolution, sometimes by trial and error. With each new requirement came a need to develop maintenance methods that would assure adherence to that requirement. The result has been a collection of techniques. Some are automated, while others are manual; some are general purpose, while others are specific to a particular source terminology; and some are executed at the time of terminology updates ("instant audits") while others are applied retrospectively.

3.1. Structure

Regardless of the representational form (LISP, MUMPS, relational, etc.), the MED is conceptually a frame-based model, with string attributes and semantic relationships, represented by slots. Slots in the MED are sequential numerical attributes that hold values for concepts. Strings are held in string-valued slots, such as LAB-TEST-LONG-NAME and CERNER-FORMULARY-CODE, while semantic relationships are represented with reciprocal pairs of slots, for example, ENTITY-MEASURED and MEASURED-BY-PROCEDURE, that take MED Codes as values.

Slots are introduced at a single, appropriate point ("fathered") at any level within the hierarchy. For example, slot 61 "DRUG-