A Proposal for Incorporating Health Level Seven (HL7) Vocabulary in the UMLS Metathesaurus

Stanley M. Huff, MD, Intermountain Health Care, Salt Lake City, UT
W. Dean Bidgood Jr., MD, Duke University, Durham, NC
James J. Cimino, MD, Columbia University, New York, NY
W. Ed Hammond, PhD, Duke University, Durham, NC

The HL7 Vocabulary Technical Committee (TC) was organized to select and maintain the vocabulary used in HL7 messages. The goal is to make implementations of the Version 3 HL7 Standard more plug-and-play compatible. In order to make the vocabulary readily accessible to the public, HL7 is collaborating with the U.S. National Library of Medicine (NLM) to include HL7 vocabulary in the Unified Medical Language System (UMLS) Metathesaurus. This article describes a proposal for how HL7 data elements and coded values can be represented accurately in the relational tables of the UMLS Metathesaurus.

INTRODUCTION

The HL7 Standard is one of the most successful message standards for communicating clinical data between heterogeneous computer systems. However, the Version 2.3 standard does not achieve plug-and-play interoperability between communicating systems. One reason for the lack of plug-and-play is that the vocabulary used in HL7 messages has not been standardized. Each implementation of an HL7 interface can use its own codes and descriptions for coded fields in the message. Most of the work in implementing an HL7 interface is consumed in matching codes and vocabulary between the communicating systems. The work is further compounded because the matching of codes and vocabulary between systems is an on-going process, due to the dynamic nature of clinical vocabulary. Thus, the lack of a vocabulary standard coupled with a rapid rate of vocabulary change means that the expense of not having a standard continues for all time.

Version 3 of the HL7 Standard is being designed to overcome the current problems with plug-and-play interoperability. As part of the Version 3 effort, the HL7 Vocabulary TC was created in August of 1996. The mission of the HL7 Vocabulary TC is to identify, organize and maintain coded vocabulary terms used in HL7 messages. The intention of the TC is not to create new codes and descriptions, but to coordinate the use of existing vocabularies with the HL7 message structure.

The TC considered several issues as it discussed how to meet its goals and objectives in maintaining codes to be used in HL7 messages. First, there is a large set of terms and codes to be maintained. The current (version 2.3) HL7 Standard contains over 300 vocabulary tables. However, many of the tables (so-called HL7 tables) are only example tables and were not meant to be complete. Users, however, want these tables to be real tables and become frustrated in the fact that (a) the tables are incomplete and (b) there is no current way to add to these tables. Second, there was a desire to select HL7 vocabulary from existing coding systems, rather than create an entirely new coding system within HL7. Third, we wanted to reuse any existing tools for building and maintaining coding schemes since maintaining a large controlled vocabulary is non-trivial[1, 2]. Fourth, we wanted the vocabulary to be readily available to anyone implementing the HL7 Standard. Fifth, we wanted to coordinate HL7 vocabulary work with similar work going on in other standards development organizations (SDOs). We particularly wanted to coordinate with the DICOM (Digital Image Communications in Medicine), ASTM (American Society for Testing and Materials), and ASC (Accredited Standards Committee) X12 organizations.

The idea of creating a common repository for vocabulary used by the message standards groups is not new. Dean Bidgood proposed the creation of "a generic message/terminology mapping resource based on the SNOMED (Systematized Nomenclature of Medicine) DICOM Microglossary (SDM) model -- the *Terminology Resource for Message Standards* (TeRMS)."[3, 4] The SDM[5] is the current vocabulary resource used in the DICOM standard. The TeRMS database would incorporate many of the features of the SDM, but would accommodate vocabulary content from multiple message standards. The plan was to build the TeRMS database as an extension of the UMLS Metathesaurus. This shared resource would provide the basis for a common understanding

of coded-entry concepts as they were used in each of the individual message standards. The vocabulary of any given standard would be a proper subset of the proposed TeRMS message/terminology mapping resource.

Coupling the idea of a TeRMS resource with their other goals, the HL7 Vocabulary TC investigated the possibility of incorporating HL7 vocabulary in the UMLS Metathesaurus. The NLM began developing the UMLS[6] in 1986. The goal of developing the UMLS was to "improve the ability of computer programs to 'understand' the biomedical meaning of user inquiries and use this understanding to retrieve and integrate relevant machine-readable information for users." A practical outcome of the UMLS project was the development of the Metathesaurus, which is a resource that cross-references many biomedical vocabularies. The first version of the Metathesaurus was released in the fall of 1990, with yearly updates since that time. The Metathesaurus is distributed on CD-ROM and is also available via the Internet.

From the beginning, the Metathesaurus has been maintained by Lexical Technology, Inc. (LTI). LTI has created unique and innovative software for building and maintaining each new version of the Metathesaurus.[7] Given the Metathesaurus' extensive cross references, availability via CD-ROM and the Internet, and state-of-the-art maintenance environment, the HL7 Vocabulary TC felt that incorporating HL7 vocabulary in the Metathesaurus would be of value to HL7 users, and would be a first step in creating the TeRMS resource.

DETAILS OF THE PROPOSAL

An extended example will be used to illustrate the proposed strategy for representing HL7 vocabulary in the Metathesaurus tables. Table 1 shows three HL7 fields as they are listed in Table 6 of Appendix A of the Version 2.3 HL7 Standard. Three coded HL7 fields are shown: Sex, Guarantor Sex, and Insured's Sex. Each field has an item identifier (Item), a data type (DT), and a table number (Table #). The data type of IS indicates that all three fields will contain coded values from user defined HL7 tables. The table number identifies the HL7 table where the allowed value set for the field is enumerated. In this example, all three fields refer to HL7 table 0001.

Table 2 shows the allowed values for each of the three sex fields as defined in HL7 Table 0001. The first column in the table indicates that this is a user-defined table, and the second and third columns

Table 1: An excerpt from Table 6 of Appendix A of the HL7 Standard. Three coded HL7 fields are shown, Sex, Guarantor Sex, and Insured's Sex. Each field has an item identifier (Item), a data type (DT), and a table number (Table #). The data type of IS indicates that all three fields will contain coded values from user defined HL7 tables. The table number identifies the HL7 table where the allowed value set for the field is enumerated. In this example, all three fields refer to HL7 table 0001.

Data_Element	Item	DT	Table #
Sex	00111	IS	0001
Guarantor Sex	00413	IS	0001
Insured's Sex	00468	IS	0001

Table 2: HL7 Table 0001 showing the allowed value set for the Sex field.

Type	Table	Name	Value	Description
User	0001	Sex		
	0001		F	Female
	0001		M	Male
	0001		O	Other
	0001		U	Unknown

Table 3: Example rows in the MRCON (Concept Names) table. Eight unique concepts are represented: 3 fields (Sex, Guarantor Sex, Insured's Sex), one value set (Sex Value Set), and 4 allowed values (Female, Male, Other, Unknown). Column meanings: CUI = Concept Unique Identifier, LUI = Term (Lexical Group) Unique Identifier, SUI = String Unique Identifier, STR = String. The CUIs, LUIs, and SUIs which contain X's are concepts that do not currently exist in the Metathesaurus and would need to be added in order to represent HL7 vocabulary.

CUI	LUI	SUI	STR
CX000024	L0036862	S0085417	Sex
CX000049	LX000049	SX000049	Guarantor Sex
CX000050	LX000050	SX000050	Insured's Sex
CX000025	LX000025	SX000025	Sex Value Set
C0015780	L0015780	S0040967	Female
C0024554	L0024554	S0059531	Male
CX000021	LX000021	SX000021	Other
CX000022	LX000022	SX000022	Unknown

identify the HL7 table number and table name, respectively. The Value column represents the code that would be used in HL7 messages, and the Description column contains the meaning of the code as a text string.

Three of the existing Metathesaurus tables (MRCON, MRREL, and MRSO) are used to represent the HL7 vocabulary items. Each of the three tables will be described in order.

Table 3 shows example content for the MRCON (Concept Names) table. The Metathesaurus is a concept-based vocabulary. A Concept Unique Identifier (CUI) identifies each distinct item or meaning in the thesaurus. A given concept can have one or more textual representations or strings (STR). Each unique text string is assigned a String Unique Identifier (SUI). The lexical group (LUI) provides further classification of each string in the table. See the UMLS Documentation[8] for more information.

The eight rows in the MRCON table represent concepts related to the three sex fields as defined in the HL7 Standard. First, each of the three sex fields (Sex, Guarantor Sex, and Insured's Sex) is defined as a unique concept. Second, the Sex Value Set is defined as a unique concept. Finally, each value in the value set (Female, Male, Other, Unknown) is defined as a concept. These eight concepts represent the distinct meanings that will be subsequently referenced in the remaining two Metathesaurus tables.

The MRREL (Related Concepts) table is used to define relationships that exist between two concepts. The first concept (CUI) is related to a second concept (CUI2) by the relationship (REL). The two types of relationships represented in this example are VS (has-value-set) and RN (related-narrower-than). The Text column contains a text name for CUI2. Text is not actually a part of the

Table 4: Example rows in the MRREL (Related Concept) table. The Text column is not actually a part of this table, but has been added for readability. Column meanings: CUI = Concept Unique Identifier, REL = Relationship, CUI2 = Second Concept Unique Identifier, RELA = Relationship Attribute, SAB = Source Abbreviation. The CUIs in the table can be translated to a text representation by reference to the MRCON table. The meaning of values in the REL column: VS = has-value-set, and RN = related-narrower. The MRREL table establishes the association between HL7 fields and their value sets, and also establishes the associations between a value set and its individual members.

CUI	REL	CUI2	Text	RELA	SAB
CX000024	VS	CX000025	Sex Value Set		HL7
CX000049	VS	CX000025	Sex Value Set		HL7
CX000050	VS	CX000025	Sex Value Set		HL7
CX000025	RN	C0015780	Female	member-of	HL7
CX000025	RN	C0024554	Male	member-of	HL7
CX000025	RN	CX000021	Other	member-of	HL7
CX000025	RN	CX000022	Unknown	member-of	HL7

Table 5: Example rows in the MRSO (Sources) table. The MRSO table provides a mapping from the CUIs back to the identifiers used in the HL7 Standard. Column meanings: TTY = term type, SCD = Source Code. The meaning of values in the TTY column, CA = Coded Attribute, VS = Value Set, and CV = Coded Value.

CUI	LUI	SUI	Text	SAB	TTY	SCD
CX000024	L0036862	S0085417	Sex	HL7	CA	00111
CX000049	LX000049	SX000049	Guarantor Sex	HL7	CA	00413
CX000050	LX000050	SX000050	Insured's Sex	HL7	CA	00468
CX000025	LX000025	SX000025	Sex Value Set	HL7	VS	0001
C0015780	L0015780	S0040967	Female	HL7	CV	0001.F
C0024554	L0024554	S0059531	Male	HL7	CV	0001.M
CX000021	LX000021	SX000021	Other	HL7	CV	0001.O

CX000022	LX000022	SX000022	Unknown	HL7	CV	0001.U

MRREL table, but it is included to make the example easier to read. The RELA column, when appropriate, contains a specific subtype of the REL relationship. For this example, the only RELA relationship used is "member-of". The source of the relationship information is represented in the source abbreviation (SAB) column.

The seven rows in the MRREL table represent relationships between the HL7 concepts previously defined in the MRCON table. The CUIs in the table can be translated to a text representation by reference to the MRCON table. The first three rows in the table represent relationships between an HL7 field and its value set. For example, the meaning of first row is, "the *Sex* field has-value-set *Sex Value Set*." The second and third rows define similar relationships between Guarantor Sex and Sex Value Set, and between Insured's Sex and Sex Value Set. The remaining four rows in the table define relationships between Sex Value Set and its allowed values. For example, the meaning of the fourth row is "*Female* is a member-of the *Sex Value Set*." The remaining rows define similar relationships for Male, Other, and Unknown. Thus, the rows in the MRREL table establish the association between an HL7 field and its value set, and they also establish the associations between a value set and its individual members.

The MRSO (Sources) table is used to store names and identifiers for concepts as they appear in the original (source) coding system. The CUI, LUI, SUI, and SAB columns have the same meanings as previously described for the MRCON and MRREL tables. The Text column is again included only for readability. The source code (SCD) column is the name of a given concept in its source vocabulary. The term type (TTY) column indicates the type or kind of code or element that is named in the SCD column. The meanings of the values in the TTY column are as follows: CA means Coded Attribute (i.e. a coded field), VS means Value Set, and CV means Coded Value.

The first row in the table indicates that in the HL7 Standard the Sex field is identified by the number 00111. The fourth row indicates that the Sex Value Set is identified in HL7 as table 0001. For individual values, the table number is concatenated with the code so that values in the SCD column give a unique translation for all HL7 table entries. Thus, the MRSO table provides a mapping from the Metathesaurus identifiers (CUIs) back to the identifiers used in the HL7 Standard.

DISCUSSION/CONCLUSION

The proposed strategy for including HL7 codes and data elements in the Metathesaurus can be accomplished nicely within the existing Metathesaurus table structure. Only the addition of a few new relationships and term types are required. The NLM is planning to do a test load of a subset of the Version 2.3 HL7 vocabulary in the near future.

Placing HL7 terms and codes in the Metathesaurus will allow easy cross-referencing between HL7 terms and existing vocabularies, like SNOMED International and LOINC (Logical Observation Identifier Names and Codes)[9-11]. It also creates linkages to other aspects of the UMLS Knowledge Sources such as bibliographic references, the Semantic Network, the Specialist Lexicon, definitions, and co-occurrence data. These linkages have the potential to make data in HL7 messages much more useful to the parties that send and receive clinical messages. Furthermore, the goals of the TeRMS initiative will be realized if other SDO's also use the Metathesaurus as a repository for their codes. Having all of the codes in a common database will allow comparisons that can lead to greater consistency and interoperability among the various message developers.

While many vocabulary needs of HL7 are met by the proposed design, there are some issues that require further discussion. One issue is the problem of updates. HL7 vocabulary, being based in clinical medicine, is quite dynamic, and will change at least with each version of the standard. We have discussed ways that versioning of the HL7 vocabulary within the Metathesaurus could be accomplished, but this design is not yet finalized. A closely related issue is frequency of releases. HL7 would like to be able to distribute updates more frequently than the current yearly schedule supported by the Metathesaurus. A third issue is the need for distributed maintenance. Maintaining the HL7 vocabulary is very time consuming. It would be ideal if the work could be distributed among many individuals who could maintain their part of the vocabulary via the Internet. Discussions are underway that may lead to the collaborative development of an Internet accessible maintenance site. Finally, procedures need to be worked out so that there is tighter coordination between HL7 and the vocabulary development organizations. It is only by integrating the structure of a message with the vocabulary sent in the message that unambiguous

information exchange between systems can be achieved.

References

1. 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;1:35-50.
2. Campbell KE, Cohn SP, Chute CG, Shortliffe EF, Rennels G. Scalable methodologies for distributed development of logic-based convergent medical terminology. In: Chute CG, ed. *IMIA WG6*. Jacksonville, FL; 1997.
3. NEMA PS 3 Supplement 23. Digital Imaging and Communications in Medicine (DICOM), Supplement 23: Structured Reporting. Rosslyn, VA: The National Electrical Manufacturers Association; 1997.
4. NEMA PS 3 Supplement 15. Digital Imaging and Communications in Medicine (DICOM), Supplement 15: Visible Light Image for Endoscopy, Microscopy, and Photography. Rosslyn, VA: The National Electrical Manufacturers Association; 1997.
5. Bidgood WDJ. The SNOMED DICOM Microglossary: A Controlled Terminology Resource for DICOM Coded Entry Data Elements. In: Chute CG, ed. *IMIA Working Group 6*. Jacksonville, FL; 1997.
6. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods of Information in Medicine*. 1993;32:281-91.
7. Suarez-Munist ON, Tuttle MS, Olson NE, et al. MEME II supports the cooperative management of terminology. *AMIA Fall Symposium*: Hanley & Belfus; 1996:84-8.
8. Côté RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L. The Systematized Nomenclature of Human and Veterinary Medicine - SNOMED International. Northfield, IL: College of American Pathologists; 1993.
9. United States National Center for Health Statistics. International Classification of Diseases, Ninth Revision, with Clinical Modifications. Washington, D.C.: United States National Center for Health Statistics; 1980.
10. Logical Observation Identifier Names and Codes (LOINC™) Committee. Logical Observation Identifier Names and Codes (LOINC™) Users' Guide vs. 1.0 - Release 1.0j. Indianapolis, IN: Regenstrief Institute; 1995.
11. National Library of Medicine. UMLS® Knowledge Sources. Fourth experimental ed. Bethesda, MD: National Library of Medicine; 1993:157.