MEDINFO 98
B. Cesnik et al. (Eds)
Amsterdam: IOS Press

Integrating Sources for a Clinical Reference Terminology: Experience Linking SNOMED to LOINC and Drug Vocabularies

Kent A. Spackman, M.D., Ph.D.[a,b]

[a]*College of American Pathologists, Northfield, Illinois, USA*
[b]*Oregon Health Sciences University, Portland, Oregon, USA*

Abstract

Achieving the promise of higher quality, lower cost and more available health care through electronic medical records requires the support of a comprehensive clinical reference terminology. In a previous paper we described SNOMED RT (reference terminology), and the data structures and logic syntax that support the transformation of the SNOMED III nomenclature into the SNOMED RT reference terminology. In this paper, we describe an approach to linking SNOMED RT to existing nomenclatures in the area of laboratory test names (LOINC™) and therapeutic drugs (Multum's MediSource™ Drug Lexicon), in order to achieve an integrated whole that solves the problem of a clinical reference terminology.

Keywords

Nomenclature; Controlled Vocabulary; Terminology

Introduction

Builders and potential users of electronic medical records have recognized a need for common clinical terminology in order to achieve their goals [1]. In spite of numerous efforts in building terminologies, the needs of developers and users have yet to be satisfied with a single comprehensive solution.

The UMLS[2], developed by the National Library of Medicine in the U.S., provides a thesaurus that links several different types of source terminologies. It has always relied on its source vocabularies to supply terminology, the relationships between concepts in the terminology, and the intentional meaning of these concepts [3]. Thus it would not be realistic to expect the needs of clinical terminology users to be met by the UMLS if those needs are not first met by the source vocabularies.

We have previously proposed that there are many different uses for terminology, and that different efforts at building comprehensive terminology may benefit from a division of the problem not along lines of professional group or concept domain, but rather along lines of the type of use of the terminology [4]. At least three important uses of terminology are readily apparent:

1. Clinical reference terminology: a pathophysiologically-focused and coherent set of clinical concepts, with definitions of their essential characteristics and their semantic relationships, for the purpose of storing, retrieving and analyzing clinical information.
2. Natural language processing terminology: synonyms, phrase variants, spelling variants, abbreviations, stemmed forms, part-of-speech labels, and other lexically-oriented terminology for facilitation of automated natural language processing.
3. User interface terminology: pick lists, user-friendly menus, structured data entry terms, abbreviations, notational shorthand, and other terminology that facilitates the recording of clinical events.

Recognizing these different uses of terminology allows us to consider a division of effort in developing comprehensive solutions to the terminology problem. This point of view does not imply a complete separation of these terminology types, but rather suggests that it is possible for different organizations to develop, refine and use them.

Our focus with SNOMED RT is to develop a comprehensive clinical reference terminology. By focusing our efforts on this goal, we increase the likelihood that an acceptable solution will be found. There is of course significant ongoing work in the academic and vendor community to create and refine the terminological specializations required to meet the needs of natural language processing, user interfaces, and messaging standards. Our overall goal is to provide the level of support necessary to ensure compatibility with these other terminological efforts, while remaining focused on the task of building a reference terminology with comprehensive coverage of the clinical domain.

SNOMED

SNOMED International (also known as SNOMED III), is the Systematized Nomenclature of Human and Veterinary Medicine, developed over a period of more than 20 years, with the support of the College of American Pathologists [5]. It contains over 150,000 records in twelve different axes or chapters, including anatomy (topography), morphology (pathologic structure), normal and abnormal functions, symptoms and signs of disease, chemicals, drugs, enzymes and other body proteins, living organisms, physical agents, spatial relationships, occupations, social contexts, diseases/diagnoses and procedures. Some of these axes have been recognized as being very complete and rich.

Two areas of nomenclature in which SNOMED has been found to be lacking in sufficient detail include the names of laboratory tests and therapeutic drugs. Therefore, we undertook a set of studies to examine how to provide users and developers with a comprehensive solution, without duplicating work or creating unnecessary new components of SNOMED.

Mapping to LOINC

LOINC [6], or the Logical Observation Identifiers, Names and Codes, is an effort in terminology building that focuses on fully-specified names of laboratory tests and other clinical observations, with the goal of transmitting clinical and laboratory information in electronic messages. The approach to building LOINC was empiric, gathering the test name master files from seven large U.S. laboratories. This empiric approach resulted in very complete sets.

LOINC's model is well documented elsewhere [6]. It contains five major components for each laboratory test (or other clinical observation) name:

component (or analyte) measured

property (type of measurement)

time aspect (single point in time vs duration)

type of sample (specimen or system)

type of scale (e.g. quantitative, nominal)

A sixth component of the name, added only when necessary, is the method employed.

For example, a quantitative measurement of fat in a 72 hour collection of stool would be represented as:

FAT:MASS:72H:STL:QN

Here MASS indicates the property measured, 72H is the time aspect, STL is the abbreviation for stool, indicating the "system" or specimen source, and QN is the abbreviation for "quantitative", differentiating the value named from a nominal or semi-quantitative result.

In SNOMED RT we employ *description logic* to represent formally the essential characteristics of concepts. Description logics have been developed for precisely this purpose, and arose out of efforts in artificial intelligence and knowledge representation to reason about concepts and their inter-relationships [7]. Our syntax is a modified version of the KRSS syntax [8] developed as part of a knowledge-sharing effort [9].

Within SNOMED, there exist concepts for "Stool fat measurement", and "72 hour stool collection". We can represent these concepts in our modified description logic syntax as follows:

Stool fat measurement:

Laboratory procedure &

(measured-analyte fat)

72 hour stool collection:

Laboratory procedure &

(assoc-time-period 72H) &

(assoc-specimen stool)

The first statement indicates that a "stool fat measurement" is a laboratory procedure where the measured analyte is fat. The second statement indicates that a 72-hour stool collection is a laboratory procedure where the associated time period is 72 hours and the specimen is stool.

Because SNOMED RT provides this level of detail in defining the essential characteristics of its concepts, it is highly compatible with the LOINC model of laboratory procedures. As a result of this high level of compatibility, we decided that it would be of great benefit to users of LOINC and SNOMED to have a tightly integrated semantic linkage between the concepts in each.

We carried out two types of mapping between LOINC and SNOMED. In the first mapping, we connected fully-specified LOINC names to there corresponding (more general) concepts in SNOMED P3, the laboratory procedures chapter of the P (procedures) axis. In the second mapping, we linked LOINC analytes (substances), systems (specimen sources), and methods to their corresponding SNOMED concepts in various axes, including T (topography – for example, for polymorphonuclear neutrophils), M (morphology – for example, for lymphoblasts), F (function – for example, for enzymes such as LDH), and C (chemicals – for various drugs and chemical substances).

First mapping

We manually mapped each LOINC laboratory test (version 1H) to the corresponding SNOMED concept in the P (procedure) axis. There were 8,276 names in LOINC, and 4,710 records in the SNOMED P axis, section 3 (Laboratory Procedure or Service). Of the LOINC names, we did not map the 360 that were drug dosage names, nor 168 others that did not fit in the laboratory procedure or service category, leaving 7748 LOINC names to be mapped. Manual review of all names was undertaken; all LOINC names were at a more detailed level of specificity than the SNOMED concepts to which they were mapped, so all the linkages are parent-child, with the SNOMED concept more general and the LOINC concept more specific. The benefit to the users of SNOMED and LOINC is that there is now a coordinated set of terms, at several levels of detail and specificity, for recording the procedures performed in the laboratory as well as the names of the resultable values of those procedures. In addition, both vocabularies were strengthened because:

1. by adding new general SNOMED concepts as needed to accommodate LOINC, SNOMED's procedure axis became more complete;
2. by examining the LOINC concepts that naturally fell as children of SNOMED concepts, some LOINC duplications were discovered;
3. by examining SNOMED test concepts for which no LOINC children were found, some LOINC omissions were discovered.

Second mapping

The beginnings of this second mapping were done lexically and have been reported previously by Rocha and Huff[10]. In that previous effort, an attempt was made to lexically match each component of the fully-specified LOINC name with the name of a SNOMED III concept. Various degrees of accuracy of match were found, with many matched SNOMED codes either

too narrow or too broad relative to the LOINC concept. Their report indicates that less than 60% of the analyte components had an exact lexical match.

We extended the previous work by manually examining each LOINC analyte, method, and specimen type (system), and carefully completing as many exact matches as possible to the appropriate concept in one of SNOMED's axes. We allowed composite matches, involving more than one SNOMED code. This was particularly useful for representing antibodies and antigens. For example, the LOINC analyte "ALTERNARIA TENIUS AB.IGG" is matched by two SNOMED codes: 1) F-C2450 Antibody, IgG class and 2) L-45111 Alternaria tenuis.

For the analytes, we were able to obtain exact matches for 86%. Finally, for those concepts in LOINC for which there was no exact match in SNOMED, we added appropriate concepts to SNOMED. Table 1 indicates the number of new concept additions to SNOMED arising from this process. In addition, a small number of concepts were identified in LOINC (approximately 20) that were not readily identifiable or that appeared to be problematic, and these were submitted to the LOINC committee for clarification.

Table 1 - Additions to SNOMED to complete full mapping of LOINC analyte concepts

SNOMED AXIS	# concepts added
C (chemicals & drugs)	193
F (functions, proteins)	355
M (morphology)	1
T (topography)	1
L (living organisms)	35
TOTALS	585

The resulting mapping coordinates not just the fully-specified LOINC concept, but also the various components of the LOINC name, with the rest of SNOMED, yielding a very powerful set of inter-related concepts, closely linked to each other and supported by the logical formality of SNOMED RT's description logic. As a result, we were able to create an expression in description logic for each LOINC concept, linked in a coordinated way to the description logic describing the rest of SNOMED RT. Using the autoclassification available, with a description logic engine [11], this coordinates the overall set of term hierarchies, and tightly integrates LOINC and SNOMED.

Mapping to Drug Terminology

Many have observed that SNOMED's list of therapeutic drugs lacks the detail necessary for support of drug terminology in the electronic medical record, particularly for user interface and electronic messaging applications. This is primarily because of the focus on individual generic drug *names, rather than on drug products (although there is a section on brand name products, this section has not been kept up to date and contains primarily US drug brand names).*

In each country where a standard clinical terminology is used, much additional information on drug products is needed, including brand names, product strength, route of administration, dose form (tablet, capsule, solution, etc.), cost, and packaging detail. These other types of information naturally would be sought from sources external to SNOMED; nevertheless, there remains a need to integrate drug information with the rest of the clinical terminology.

For example, we want to be able to link the code for measurement of a drug (P3-78240 Desipramine measurement) with the drug measured (C-62270 Desipramine), and in turn link this to the product prescribed. It is the link from the generic drug name to the product prescribed that can be provided by a mapping between SNOMED and an external drug database.

In the United States, the NDC codes (National Drug Codes) are used in electronic claims transactions, and definitely constitute a set of drug codes to which we want to link. However, each NDC code identifies an individual product at the finest level of detail; slight changes in packaging alone will result in a new NDC code. Users need, in addition, information about the active ingredients, dosage form, strength, route of administration, and cost of each product. The NDC codes are not distributed with a database that connects them with these pieces of information, nor to a common therapeutic content database, or to a database of therapeutic classes of drugs.

This missing information is supplied in the drug database products of several companies (in the U.S. these include First Databank, Micromedex, Multum, and others). Because Multum makes their drug lexicon, MediSource™ available freely on the Internet [12], we studied the linkage between it and SNOMED. We developed an approach that we think is compatible with both SNOMED and the drug database, while providing an integrated solution for the user.

The MediSource product is structured as a database that normalizes most of its components, permitting a clean separation between generic drug names, single component drugs, and multiple component or combination drugs and their brand names. It also provides a clean linkage between the NDC code, the brand name, and the generic components of each product.

We found that it was very straightforward to lexically match the drug names in SNOMED's C axis to the drug names in the MediSource table named multum_drug_id. This table contains 1,756 entries; of these, we found 1,337 that were single-component drugs that could potentially be mapped to SNOMED. Of these, 1,010 already exist in SNOMED (version 3.4). The additional 327 drugs, after review of primary sources, are candidates for addition to version 3.5. The completed mapping results in a table that links the 1,756 generic drug names and codes in Multum's database with the corresponding generic drug names and codes in SNOMED's C axis.

Benefits of this approach are readily apparent. The strengths of SNOMED as a clinical reference terminology are retained, allowing procedures, tests and diagnoses that refer to generic drug names to continue to be tightly integrated, while extending the ability of the system implementer to add the additional necessary and rich features of a fully-functional drug lexicon. The ongoing integration effort required is minimal, since the vast majority of work in maintaining a drug lexicon occurs with information other than the names of new single generic drugs.

Each terminology development organization can focus on complementary aspects of the problem. Developers and users achieve integrated solutions. In each country where different drug products and names are used, a lexicon developed for that country can be mapped into SNOMED the way that MediSource has been for primarily U.S.-source drug products.

Status and Future Directions

Ongoing work on SNOMED RT is focused on developing relationships with numerous other terminology organizations, particularly those with specialty-specific concepts which need to be integrated with SNOMED. Notable among these are the American Dental Association, the American Academy of Ophthalmology, and the DICOM (Digital Image Communications) standard development organization. These organizations have acknowledged the need to create an integrated clinical reference terminology that functions as a coordinated whole, rather than creating multiple fragmented and uncoordinated terminologies. By using SNOMED RT as the scaffolding or framework by which this integration takes place, we achieve logical and conceptual rigor. By having the individual specialty groups themselves identify and provide an initial organization of the concepts they need in a clinical terminology, we achieve depth, clinical currency, and appropriateness for the ultimate users of the reference terminology.

Conclusion

We have described an approach to coordinating SNOMED's clinical reference terminology with a nomenclature of laboratory and clinical observations (LOINC™) and with a database of drugs and drug products (Multum's MediSource™ Lexicon). We believe this approach, involving linkages between terminologies, can yield a comprehensive and sustainable solution to the reference terminology needs of organizations and developers.

Acknowledgments

The author expresses appreciation to Keith Campbell and Simon Cohn for originally suggesting the separation of reference terminology from user interface terminology. The SNOMED Editorial Board, and in particular Roger Cote, contributed to the overall concepts described here. Significant work on the LOINC mapping was contributed by Stan Huff, Roberto Rocha, and Robert Dolin.

References

[1] Board of Directors of the AMIA. Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. JAMIA 1994; 1,1-7.

[2] Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine 1993; 32:281-91.

[3] McCray AT, Hole WT. The scope and structure of the first version of the UMLS semantic network. In: Miller RA, ed. Fourteenth Annual Symposium on Computer Applications in Medical Care. Washington DC: IEEE Computer Society Press, 1990.

[4] Spackman KA, Campbell KE, Cote RA. SNOMED RT: A Reference Terminology for Health Care. In: AMIA Fall Symposium: 1997.

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

[6] Logical Observation Identifier Names and Codes (LOINC™) Users' Guide vs. 1.0, Release 1.0h. Obtained electronically at www.mcis.duke.edu /standards/termcode/loinc.htmWoods WA, Schmolze JG. The KL-ONE family. Computers and Mathematics with Applications 74:2-5, 1992.

[7] KRSS working group of the DARPA Knowledge Sharing Effort. Draft of the specification for description logics. http://www-ksl.stanford.edu/ knowledge-sharing/ papers/ index.html#dl-spec 1993; July 1993.

[8] Patil RS, et al. The DARPA knowledge sharing effort: progress report. Principles of Knowledge Representation and Reasoning: Proceedings of the Third International Conference, Cambridge, MA, Morgan Kaufmann, 1992.

[9] Rocha RA, Huff SM. Coupling vocabularies and data structures: Lessons from LOINC. In: AMIA Fall Symposium, 1996:90-94.

[10] Mays E, Weida R, Dionne R, et al. Scalable and expressive medical terminologies. AMIA Fall Symposium. Washington DC: Hanley & Belfus, Inc., 1996:259-263.

[11] MediSource Lexicon™ Drug Product and Disease Listings, Internet Version. Obtained from www.multum.com / lex.htm