Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Strategies for Updating Terminology Mappings and Subsets using SNOMED CT®

John Mapoles, Ph.D., Corey Smith, Jane Cook, Brian Levy, M.D.
Health Language, Inc., Aurora, Colorado 80011

Abstract:
SNOMED CT® (SCT) is a large, comprehensive medical terminology with many applications in the health care IT sector. SCT is often mapped to existing billing classifications as well as to proprietary terminologies in order to support access to and from SCT from existing applications. Subsets of SCT are used to reduce complexity and size. These subsets can vary from small sets that will be used to populate drop down lists in electronic medical record applications and larger lists that are used for reference, e.g. the SCT Non-human subset. There are costs and time limitations in maintaining mappings and subsets, particularly after each SCT release when concepts are retired and new concepts are added. There is a need for a careful strategy to identify changes, determine which changes need to be reviewed, and to rank changes so they can be reviewed systematically in order of importance. Here we outline our updating strategies for the Health Language Medical Specialty Subsets, a list of 10,000 SCT terms grouped into 45 subsets. These strategies can be used for any subset of SCT as well as for mappings created to and from SCT.

Introduction:
Electronic health records (EHR) and other healthcare IT applications rely on controlled medical terminologies to provide well defined concepts for accurate and consistent encoding of records and data mining. SNOMED CT® (SCT) provides broad coverage of all medical domains with approximately 280,000 active concepts. A great deal of attention has been focused on the models and strategies to implement SCT[1,2,3]. Most applications will use defined subsets of SCT for specific use cases rather than exposing all of SCT to all users. Subsets are collections, lists, of SCT concepts or terms. Applications using SCT will need to be able to store these lists for purposes of maintenance and delivery to EHR interfaces.

Mappings are often created between SNOMED CT and other terminologies such as billing classifications, ICD-9-CM and ICD-10, as well as local and proprietary terminologies. Mappings can be represented as a pair of codes, the source and target of the map or relationship. These mappings can be used for translation of information from one set of codes to another.

These mappings and subsets represent work that a local site or user is performing to the standard – in this case SCT. Thus, SCT is distributed from the standards body, and local users such as EHR vendors or hospitals, need to add value to their SCT version with mappings and subsets. It is critical that the local user adopts the next version of SCT in order to prevent semantic drift – multiple versions of a terminology being used that drift in meaning enough to be incompatible. In the paper, Oliver[4], et. al. discuss some of the principles of localization of terminologies and also explores the impact of migration of localized terminologies to the next version of the standard. Updating subsets and mappings as described in this paper actually only involves a small subset of the many kinds of changes that occur to the terminology. Cimino[5], et. al. discuss the types of changes that need to be considered such as refinement, name changes, code-reuse and more that impact the updating of terminologies and content based on them.

The Semantic Web work is now introducing new issues with regards to ontology versioning. Liang[6], et. al. discuss the impact of changes in ontologies to existing applications that depend on them. In this semantic web paper, a middle layer to monitor and detect changes is proposed to be used between the underlying ontologies and the dependant applications. The work we present here with subsets uses a terminology service and specialized scripts to serve as this middle layer between the standard ontology, SCT in this case, and the resulting applications – EMRs for example that depend on the subsets.

Maintenance of the mappings and subsets are costly and time dependent because the content is often in production. Maintenance can involve changes that are dictated by the applications that use the content but also because the underlying SCT data model has changed. Semiannual updates to SCT can change the SCT model to varying degrees, sometimes substantially. An SCT update involves new concepts and terms, retirement of concepts and terms, and addition and deletion of relationships. Each mapping

and subset needs to be evaluated so that only the specific groups of changes that immediately affect the content are considered. This analysis is essential for performing the update quickly and efficiently.

Health Language (HLI) provides a terminology server, or Language Engine (LE) that serves up terminologies and related content to other applications. HLI also provides modeling and mapping tools to allow users to update and localize the content in their terminology server. Health Language (HLI) has developed a set of 45 subsets, Medical Specialty Subsets (MSS), that represent "clinically friendly" terms used in medical practice. These subsets represent 10,000 entries that must be rapidly updated for production release. Changes to SCT must be filtered so that changes that impact the MSS are reviewed, without the distraction of changes that are not important to the MSS. This paper will discuss the principles used to update the subsets. These same principles are also used to update our various mapping projects which include maps between SCT, billing classifications, and proprietary terminologies.

Methods

Medical Specialty Subsets: The MSS consist of 45 subsets containing 10,000 SCT terms. The subsets contain “clinically friendly” SCT terms most often encountered in clinical practice. The subsets were constructed so that SCT concepts were specific to the level of granularity commonly needed by clinicians in that specialty practice. The top approximately 150 diagnoses and procedures applicable to each specialty were incorporated into the subsets. Claims data as well as medical domain expert knowledge was used to determine these concepts that are applicable for the subsets. Concepts with low incidence based on claims data, e.g. rabies, were not included.

Only concepts mapped to billable ICD-9-CM and CPT codes were used to construct the subset, using the College of American Pathologists SCT to ICD-9-CM cross maps and HLI SCT to CPT cross maps. This increases the possible utility of the subsets for billing purposes-

Some concepts may be too specific for one subset, but applicable in another. For example, *Acute anterior myocardial infarction* may be too specific for the Family Practice subset, but applicable in the Cardiology one.

The MSS are stored in the HLI LE® database. Members of the subsets can be viewed and managed using the HLI browser and editing tool LExScape® or the HLI Java application programming interface (API). HLI APIs are a full set of Java classes, interfaces, and methods that allow access and management of data in an HLI LE database. The APIs can be bundled into complete applications such as LExScape and other HLI management tools or they can be used to terminology enable applications that require both terminology support and other functionality not related to terminology. The APIs are collected into standard jar files and can be accessed and using standard Java programming methods.

Terms included in the MSS must meet the following requirements:

a. Subsets related to disease take terms from concepts in the SCT Clinical Findings taxonomy.
b. Subsets related to procedures take terms from concepts in the SCT Procedure taxonomy.
c. Terms must have an SCT description status of 0 on a concept with a status of 0. These are referred to as active terms. A status of 1 or greater is retired or limited in use.
d. A subset can contain only one term from each concept.
e. Concepts that contain terms in the disease subsets must have a valid relationship to ICD-9-CM as defined by the College of American Pathologists produced SCT to ICD-9-CM cross mapping. Only those concepts that have a cross map to a billable ICD-9-CM code are considered.
f. Concepts that contain terms in the procedure subsets must have a valid relationship to CPT as defined by the HLI SNOMED - CPT relationships. Only those concepts that have a cross map to a billable CPT code are considered.

SNOMED CT: The SCT release of January 31, 2007 was downloaded from the College of Amercian Pathologists and was transformed into the HLI's data structure and stored in the LE database.

Changes to SCT: Changes to SCT between releases are calculated by comparing consecutive versions of SCT using the HLI Java APIs. The SCT core vocabulary consists of three broad object types: concepts, terms (descriptions), and relationships. Although concepts and terms are never deleted from SCT they can change status from active to limited or retired. Concepts can also move between SCT hierarchies. The number of changes that occur each time SCT is updated varies widely and is distributed throughout all SCT taxonomies.

Within the scope of this project we are concerned with term changes, because the subsets are lists of SCT terms. Concept changes must also be tracked because if a term's concept moves out of a subset's target hierarchy the term can no longer be

used in the subset. Changes in the defining relationships of the concepts generally do not impact the placement of the concept in the MSS because if the concept changes in meaning completely, then it will be retired in SCT. The types of changes that are considered were:

a. Concepts or terms that became limited or retired.
b. New terms in the target hierarchy.
c. A term's concept no longer has a relationship to a reference terminology, i.e. ICD-9-CM or CPT.

These change types were then analyzed by creating custom Java scripts written to the HLI APIs. These scripts can then be re-run for each SCT update. The output of these scripts are then fed into the HLI modeling and mapping tools. Each change type is presented to the modeler in the tool as a separate project of affected concepts; thus a collection of small update projects is generated for each SCT release.

Once the update is completed, the MSS are then versioned as a set and released. Versioning collects all subset changes, including new terms, and stamps them with a version number applicable for the HLI product. Each version number is then tied to a release of SCT.

Results

The goal of any update is to isolate changes that bear directly on the data sets of interest because the SCT update is generally so large that all changes can not be reviewed in a timely fashion. Once the changes that impact the data sets are isolated they must be categorized and arranged so that the most important changes are reviewed first.

For the MSS the following change types are considered important:

1. Invalid concepts or terms: these are concepts or terms that have been retired, that have changed to status limited SCT status value 6 - representing a classification or administrative concept, that are no longer mapped to a reference hierarchy, or that are still active but have been moved out of the target taxonomy.
2. New terms on concepts in the subset: Because terms from these concepts are already in the subset a new term on these concepts are especially interesting as a possible replacement.
3. New terms on descendents that have a relationship to ICD-9-CM or CPT: Descendents of concepts that are in the subset are of special interest because they are likely to contain terms that are more specific than the term in the subset.
4. All other new terms: These are all new terms in the target taxonomy that are not part of the type-2 and type-3 group. These are due to the addition of new concepts and the addition of new terms on existing concepts.

Changes should be reviewed in the order of importance. Limited time and resources dictate that essential changes must be repaired first.

Changes in relationships and qualifiers in SNOMED CT are not considered during the udpates process. Minor changes in relationships are not considered to change the meaning of the concept. Major changes usually cause the concept to be retired and replaced. Changes to IsA relations, and indirectly to defining non-IsA relations, are considered in change types 3 and 4. Changes to qualifying relations are not considered because these relations are not definitional

Management of the process

Changes of type-1 must be reviewed first because they break the requirements of the MSS. These terms must be removed and possibly replaced. Whenever possible the SCT historical relationships (e.g. SAME AS, MAYBE A, REPLACED BY, etc.) are used to identify an active term replacement. For example, in an earlier SCT update, the concept of *Coronary artery thrombosis* was retired and placed into the *Ambiguous concept* hierarchy. This retired concept now has a *MAYBE A* relationship to the active *Myocardial infarction* concept. The Java scripts written to identify the changes include an algorithm to locate these historical relationships and potential replacement SCT concepts. Modelers then can view these potential replacement concepts in the tool.

Once the requirements of the subset are satisfied, possible new or replacement terms can be considered. Changes of type-2 are new terms on concepts in the subsets. These concepts are the most likely to contain replacement terms for existing terms.

SCT is arranged into hierarchies. Concepts become more specific when moving from the top to the bottom of a hierarchy. Changes of type-3 leverage the SCT hierarchy. Terms that are descendents of terms in the subset are likely to be of interest. They are likely to be more specific than existing terms providing replacements or valuable

additions. Only concepts that have mapping relations to ICD-9-CM or CPT are considered in this review.

Changes of type-4 include all other new terms in the target taxonomies. These are the terms that are least likely to be of interest but all new terms must be reviewed to ensure each subset contains the most recent SCT content. Terms of type-4 may also contain new ideas for inclusion in subsets. An important part of the management of the process is that all terms of type-4 are reviewed together once and considered as a group for inclusion in all subsets. This is a very large group and reviewing only once is a valuable time saver.

The data for all of the subsets is not presented here. Of the 45 subsets, 24 were not changed at all. Table-1 presents a sample of change data for four subsets.

Subset Name	# of Members	Type-1 Changes	Type-2 Changes	Type-3 Change
Critical Care - Disease Subset	558	3	3	137
Neurology - Disease Subset	367	4	2	185
Gastroenterology - Procedure Subset	185	6	4	3
Neurosurgery - Procedure Subset	327	1	1	7

Table 1: Changes in selected subsets

Subset Name	Changes Made, Type-1 Changes	Changes Made, Type-2 Changes	Changes Made, Type-3 Changes	Changes Made, Type-4 Changes
Critical Care - Disease Subset	3	3	2	1
Neurology - Disease Subset	4	1	0	1
Gastroenterology - Procedure Subset	6	2	2	1
Neurosurgery - Procedure Subset	1	1	0	2

Table 2: Changes made to selected subsets based on type

Data from Table-2 demonstrates how this approach focused the update tasks. The review of each subset is limited to a controlled number of review tasks that are specific to the subset.

Changes of type-1 always result in a corresponding action (Tables 1 and 2). Changes of type-1 break the rules of the subsets so they must result in a change.

Changes of type-2, type-3, and type-4 are new terms. Changes of type-2 are new terms on the concepts that have a term already in a subset. Of the 14 candidates 7 were added to the subsets. The percentage of type-3 changes added to the subsets is much smaller. These are terms on descendents of concepts already in the subsets. They are added at a much lower rate because modelers consider them too specialized for the subsets.

Table-2 also demonstrates that Type-4 changes are important. They add valuable new content to the subsets. The percentage of added terms here is low as would be expected since these terms are on concepts that are not in the subsets or their descendents. There is pressure to keep the subsets tightly defined to a core group of diagnoses and procedures.

The update task for the HLI MSS has been broken into a series of specific tasks and one large task (type-4). In this release, January 2007, of SCT there were over 3000 new concepts and 3400 new terms on existing concepts as well as 6500 concepts and terms retired. Thus despite the large number of changes that occurred in SNOMED CT in this past release, only a small number of changes actually affected the subsets, limiting the amount of review required.

Discussion

Development of any data model requires time and resources as part of a pre-production effort.

Updates are post-production and under the additional constraint that updates must be finished in a timely fashion so that new production data is available as soon as possible.

We present a strategy here for the maintenance of SCT subsets based on the experience of maintaining the Medical Specialty subsets. These strategies can be adapted to any subset or mapping created with SCT or other terminologies. The basic premise is that it is not feasible to review all entries in all subsets after each new SCT release. Thus, we identified the types of changes that occur in SCT that would be more relevant to the subsets. These 'units' of change are then ranked in order of importance. Automated processes are employed whenever possible followed by manual review of the changes when necessary. The SCT hierarchies are also leveraged to allow for identification of possible new, more specific terms to be included in the subsets.

Similar strategies can be used to update mappings to and from SCT as well. As in the case for subsets, a set of requirements defines the rules used to create the mappings. Many of the same change types described above that affect subsets also may affect mappings as well. But, in addition to changes in SCT, consideration needs to be given toward similar changes in the other terminology being mapped to or from SCT. For example, a mapping between ICD-9-CM and SCT needs to be updated for changes to ICD-9-CM as well as SCT. Therefore, evaluation for other change types particular to the non-SCT terminology need to be considered. Modelers then need only review a portion of the maps that are affected by the changes. In the case of various mapping projects that HLI has performed, these reviews usually only include hundreds of concepts instead of tens of thousands.

Changes to medical domains in the real world define the requirements for medical terminologies such as SNOMED CT. New terms that are added to SNOMED CT to meet these requirements. This paper discusses how HLI manages these changes in the MSS There will always be cases where the medical world moves faster than the terminology. SNOMED CT provides a mechanism for post-coordination and extension to fill these gaps. HLI however has decided, as a rule, that the Medical Specialty Subsets only use pre-coordinated concepts. HLI does make submissions to the International Healthcare Terminology Standards Development Organization (IHTSDO) whenever a new concept is required to fill a gap between the medical domain and the terminology. Using the HLI framework MSS users can make additions and deletions to their local copy of the MSS to account for local conditions.

As SCT becomes more widely used, managing its changes and effects on content based on SCT will be critical. Using efficient processes such as those identified here will help manage SCT changes.

References

1. Richesson R., Young K., Guillette H., Tuttle M., Abbondondolo M, and Krischer J. Standard Terminology on Demand: Facilitating Distributed and Real-time Use of SNOMED CT During the Clinical Research Process. AMIA Annu Symp Proc. 2006; 2006: 1076.
2. Vikström A., Skånér Y., Strender L-E., and Nilsson G. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: Rule development and intercoder reliability in a mapping trial. BMC Med Inform Decis Mak. 2007; 7: 9. Published online 2007 May 2. doi: 10.1186/1472-6947-7-9.
3. Jacobs A., Quinn T., and Nelson S. Mapping SNOMED-CT Concepts to MeSH Concepts. AMIA Annu Symp Proc. 2006; 2006: 965.
4. Oliver DE, Shahar Y. Change Management of Shared and Local Versions of Health-Care Terminologies. Meth Inform Med 2000;39:278-290.
5. Cimino JJ. Formal Descriptions and Adaptive Mechanisms for Changes in Controlled Medical Vocabularies. Meth Inform Med 1996;35:202-210.
6. Klein M., Fensel D*., Ontology Versioning on the Semantic Web,* Proceedings of the International Semantic Web Working Symposium (SWWS), Stanford University, California, USA, July 30 -- Aug. 1, 2001.