SNOMED Clinical Terms: Overview of the Development Process and Project Status

Michael Q. Stearns, MD[1], Colin Price, MPhil, FRCS[2], Kent A. Spackman, MD, PhD[1,3], Amy Y. Wang, MD[1]

[1]College of American Pathologists, Northfield, IL
[2]NHS Information Authority, Aston Cross, Birmingham, UK
[3]Oregon Health Sciences University, Portland, OR

Two large health care reference terminologies, SNOMED® RT† and Clinical Terms Version 3†, are in the process of being merged to form a comprehensive new work referred to as SNOMED® Clinical Terms. The College of American Pathologists and the United Kingdom's National Health Service have entered into a collaborative agreement to develop this new work. Both organizations have extensive terminology development and maintenance experience. This paper discusses the process and status of SNOMED® CT† development and how the resources and expertise of both organizations are being used to develop this new terminological resource. The preliminary results of the merger process, including mapping, the merger of upper levels of each hierarchy, and attribute harmonization are also discussed.

INTRODUCTION

The Systematized Nomenclature of Medicine Reference Terminology (SNOMED RT) is the most recent edition in a series of SNOMED® terminologies that have been developed by the College of American Pathologists over the past 35 years.[1,2] It is a comprehensive health care terminology that contains over 120,000 interrelated health care concepts, supported by synonyms and semantic definitions. SNOMED RT is designed to serve as a common reference terminology for the aggregation and retrieval of health care data recorded by multiple organizations and individuals.

Clinical Terms Version 3 (Read Codes CTV3) was developed by the United Kingdom's National Health Service. CTV3 is the most recent edition of the Read Codes, which originated in the early 1980s as a mechanism for storing structured information about primary care encounters in individual, patient-based records in the United Kingdom.[3,4] CTV3 consists of approximately 200,000 interrelated concepts. Its development process has featured extensive quality control measures and input from clinical specialists.[5,6,7]

SNOMED RT and CTV3 have several features in common. They are both comprehensive controlled medical terminologies with reference properties that support enumerative and compositional functionality. Each provides semantic definitions for procedure and disorder concepts via concept inter-relationships that were developed and refined through manual and automated processes.[7,8]

In 1999, the College of American Pathologists (CAP) and the National Health Service (NHS) in the United Kingdom formed a strategic alliance to merge the two terminologies. Both parties agreed to share experiences and resources, allowing SNOMED CT to become a highly comprehensive terminology. Despite the similarities, the merger of two large-scale terminologies of this magnitude presents several challenges. This paper provides an overview of the merger process, technical design, anticipated content, and status of the project through early 2001.

ORGANIZATIONAL FRAMEWORK

SNOMED CT development is an open process that has involved a large number of health care organizations and professionals. The structure and content of SNOMED CT is determined by the SNOMED International Editorial Board, which advises the SNOMED International Authority. A SNOMED CT design team, supported by content and technical working groups, makes recommendations to the editorial board regarding the technical structure and clinical content.

Approximately fifty physicians, nurses, physician assistants, pharmacists, informaticists, medical technicians and other health care professionals in the U.S. and U.K. are directly involved with modeling the terminology. Specialized terminology groups focussing on the needs of specific terminology domains such as nursing, allied health care, and pharmacy have been formed, and additional working groups are being created to provide domain-specific input and to assist with quality assurance. These

1067—5027/01/$5.00

groups forward recommendations to the content working group and editorial board for consideration.

DEVELOPMENT PROCESS

The development of SNOMED CT has been divided into six stages:

Stage I. Start-up and Initiation

This first stage focused on establishing U.S. and U.K. representation on the SNOMED International Authority, SNOMED International Editorial Board and the SNOMED CT Design Team and working groups. A prototype design for mapping the semantically equivalent concepts in SNOMED RT and CTV3 was specified and a tool was developed for this purpose. The mapping process was initiated in the U.K. during this stage.

Stage II. Terminology Design

During this stage the technical working group developed several consultation documents that describe the structure and functionality of SNOMED CT. These include documents describing:

- SNOMED CT goals
- Core structure
- Subset mechanisms
- Analysis of requirements
- Proposed structure for cross-mapping tables
- Proposed history mechanisms

Following approval by the editorial board, these draft documents were made publicly available for wider consultation. Feedback from users in the U.K. and the U.S. was used to refine the core structure and subset structure documents through an open and iterative process. These documents remain available for public review and comment at the SNOMED web site.[9]

Another activity of the terminology design stage was the merger of the upper level hierarchies of SNOMED RT and CTV3 by teams of two or more editors. This allowed legacy issues and differences in meaning based on culture to be identified.

The content working group evaluated all existing and proposed attributes (i.e., concept interrelationship types) in SNOMED RT and CTV3. Many of the existing attributes were readily merged, although a significant number required some degree of transformation to form harmonized models. For example, there was partial overlap between how the SNOMED RT attribute "morphology" and the CTV3 attribute "pathologic process" were used. The usage of each attribute was further defined through a consensus process. Several new attributes have also been added or are in the evaluation process. For example, attributes that support definitional relationships between disorders and molecular biology concepts are being reviewed.

An additional key component of the design process is the development of a robust quality assurance plan that makes optimal use of prior experiences[7,8] and available resources.

Stage III. Production

The production stage consists of the actual merger of the content of SNOMED RT and CTV3, including the establishment of harmonized semantic definitions. It is divided into several phases.

Phase I: Description Mapping

The first step in merging the two terminologies involved identifying and creating maps between semantically equivalent concepts.[10] This process identified equivalent concepts in the two terminologies and flagged potentially ambiguous concepts in the source terminologies for additional review.

As anticipated, the process illustrated a significant number of issues related to synonymy (see results section). These were referred to as "description mapping conflicts." Resolution of these conflicts required the development of a specialized software tool. This interface allowed editors to view the original statuses of descriptions involved in a conflict and the mapping actions that led to the conflicts. It also provided the ability to resolve conflicts by placing descriptions in the appropriate SNOMED CT concepts. General clinical editors reviewed conflicts and resolved the ones they viewed as "straightforward." Editors flagged particularly challenging conflicts for review by a senior editor or domain specialist. The content working group will act upon the recommendations of domain experts to resolve any remaining conflicts.

Phase II: Concept Modeling

Following completion of the description mapping conflict resolution process, the concepts table and the descriptions table will be fully populated. At that point, each SNOMED CT concept will have concept interrelationships that may have originated from one or more sources. These include relationships from one or both source terminologies, newly assigned is-a relationships from the mapping process, or newly assigned is-a relationships from upper level hierarchy

merger work. The concepts will undergo automated processing to remove redundant or semantically less proximate superordinate relationships. Editors in the U.S. and U.K. will then review each SNOMED CT concept and refine definitional (e.g., attributes) and non-definitional (e.g., synonyms and qualifiers) properties.

Phase III: Terminology Refinements

This phase involves the final editing steps needed to prepare the content of SNOMED CT for implementation in clinical systems. These include assigning preferred and fully specified names to concepts, adding non-definitional attributes (i.e., qualifiers), developing subsets, adding new concepts to domain regions determined to be deficient, and quality control using automated processes and manual review by editors. The quality assurance of SNOMED CT will be based upon a harmonization of automated processes developed in the U.K.[7] and U.S.[8] and manual review by editors and clinical domain experts.

Stage IV: Alpha Test

The editorial board authorized an alpha test of SNOMED CT in order to prototype, test and scope the development process and to provide developers with data for testing the core table structure. The content of the alpha test is limited to six clinical domains: Orbital region procedures, Orbital region disorders, Urinary system disorders, Urinary system procedures, Respiratory system infectious disorders, and Breast neoplasms. All concepts in these domains were modeled using a specified set of editing heuristics that utilized harmonized attributes from CTV3 and SNOMED RT. System developers and interested clinicians will evaluate the alpha test data and their feedback will be used to refine SNOMED CT.

Stage V: Beta Test

At the termination of the production stage, the concepts, descriptions and relationships tables will be fully populated with data (minus a limited number of refinement settings). This unrefined data will be made available to system developers and other interested groups as a beta test of SNOMED CT. It will allow testing of the table structure populated with a data set similar to the full release content of SNOMED CT. Information from users will be used to refine the initial release version of SNOMED CT. The beta test stage may run in parallel with the refinement phase of the production stage.

Stage VI: Release Process

Following the end of all production editing activities, the content will undergo a series of automated quality control steps with necessary corrections and revisions to ensure data integrity. SNOMED CT will then be made available for download from the Internet or distributed on compact disc to licensed users.

PRELIMINARY RESULTS

Hierarchy Merger

The Root hierarchies of CTV3 and SNOMED RT were merged to form the SNOMED CT root hierarchy (Figure 1). At this level, the two terminologies are highly compatible. For example, the CTV3 root for procedures, labeled "Operations, procedures and interventions" was found to be semantically equivalent to SNOMED RT's "Procedure." At lower levels, however, there were significant differences in the structure of each corresponding hierarchy. For example, the procedure hierarchy of SNOMED RT includes a large section that organizes procedures based on method. CTV3 does not have a corresponding procedure by method hierarchy, but the same information is represented through an attribute-value pair. The current proposal is to represent method through both a "Procedure by method" grouper term and attribute-value pairs.

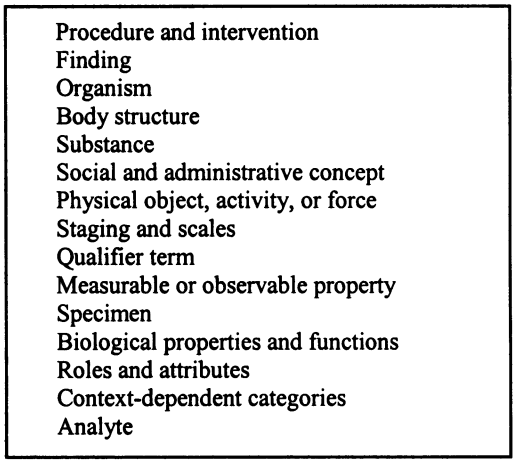

Figure 1. SNOMED CT Root Hierarchy

The hierarchy merger process supplemented the description mapping process by allowing for additional cross validation of synonymy. It also highlighted challenging areas of both terminologies that are currently under consideration, such as the representation of abnormal findings and disorders and issues related to context (e.g., header terms) such as "Family history of diabetes."

During the merger of the procedure and disorder axes, the content working group identified the need to harmonize the CTV3 and SNOMED RT anatomy definitions for systems, tracts and other body structures. For example the definitions of the anatomic concept "Urinary system" are not equivalent in CTV3 and SNOMED RT, as only CTV3 includes the prostate gland as part of the urinary system. After careful consideration, the content working group proposed excluding the prostate from the definition of "Urinary system." Decisions of this nature are subject to revision based on feedback from users.

Description mapping

Table 1 provides a summary of the results for the mapping for descriptions that were identified as having either a semantic match or an "is-a" relationship to concepts in the opposite terminology. This process is described in greater detail elsewhere.[10]

Type of map	CTV3 to SNOMED RT	SNOMED RT to CTV3	Total Maps (US and UK)
Semantic	29,016	47,731	76,747
Is-a relationship	99,532	93,142	192,674
Total maps	128,548	140,873	269,421

Table 1: Preliminary results of description mapping

The semantic match rate for maps from CTV3 descriptions to SNOMED RT concepts was approximately 23%. The corresponding rate for identifying semantically equivalent SNOMED RT descriptions to CTV3 concepts was approximately 34%.

Analysis of the mapping data has illustrated several types of mapping conflicts. For example, during the mapping process editors identified that the term "Operation" in CTV3 was placed as a synonym of "Procedure." In SNOMED RT, "Operation" is treated as a synonym of "Surgical procedure". The content working group evaluated this discrepancy and has defined a reproducible distinction between the concepts "Procedure" and "Surgical procedure" and recommended that we list "Operation" as a synonym of "Surgical procedure." This has led to a systematic resolution of this type of conflict, so that descriptions that contain the phrase "Procedure on X" are no longer merged into the same concept as "Operative procedure on X."

Alpha test

The alpha test has resulted in merger of a set of focused domains of the terminology, combined anatomic definitions, refinements in the usage of harmonized attributes, and development of SNOMED CT editing heuristics. It has also allowed us to address a series of technical issues related to the transfer of data between the U.S. and U.K development environments. The effect of algorithmic classification using a description logic classifier on the hierarchical structure and attribute-value pairs is being evaluated. This will allow a direct comparison of manual editing to editing supplemented by autoclassification. This information will be used to refine the role of autoclassification in the development process of SNOMED CT.

DISCUSSION

We based the technical design of SNOMED CT on our collective experiences and feedback from system developers and other end-users. The decision to provide additional functionality in the form of subsets and extensions was also based upon feedback from users. Subsets of SNOMED CT will allow users to specify a set of concepts, descriptions or relationships that may be useful to a particular organization or language. Extensions will allow organizations to create internal concepts for individual needs that retain the full functionality of SNOMED CT. The alpha and beta tests will allow us to refine our development process and refine the technical structure and content of the terminology based on feedback from users.

The description mapping process was designed to minimize the number of potentially ambiguous or redundant concepts in SNOMED CT. Editors identified equivalent concepts through a cross-validated process. The existing synonyms in each original SNOMED RT and CTV3 concept were reviewed by an editor from the opposite source terminology. Preliminary data at this writing suggests a 20-30% concept overlap between concepts in SNOMED RT and CTV3. Based on these figures, we estimate that SNOMED CT will contain over 300,000 concepts and over 450,000 descriptions.

Harmonization of the upper levels of the hierarchies has highlighted a number of scientific and cultural differences in terminology components at an early stage. Developers have thus been able to review and come to agreement on a variety of issues that influence broad areas of the terminology, such as anatomy models, synonym usage, preferred names, levels of pre-coordination, and issues pertaining to hierarchy navigation. These differences have not presented a barrier to the development of SNOMED CT, but rather have encouraged the reexamination of

challenging terminology issues. We view the upper level hierarchy merger process as a valuable exercise that complements the description mapping and concept modeling phases of this project.

The goal of the description mapping phase of this project was similar to the objectives of the Unified Medical Language System (UMLS) Metathesaurus.[11] Both efforts merged synonymous concepts from source vocabularies into a single concept. The SNOMED CT development effort differed from the UMLS effort in that clinical editors from the source terminologies performed the mapping. The goals of the SNOMED CT mapping process were to preserve the intensional meaning of each source concept and to allow the over 30 SNOMED CT editors involved with mapping to identify scientific and cultural differences in the two terminologies. This information was used to establish editorial policies.

The overall SNOMED CT project also differs from other large scale terminology development efforts such as Galen[12], MeSH, CTV3,[3,4] and SNOMED RT[2,8] in that these terminologies were in general created from a single source or through the contributions of individuals or groups. In contrast, SNOMED CT will be the result of a merger of two established large-scale terminologies and the combined knowledge and resources of two medical organizations with extensive terminology development experience. SNOMED CT also differs from these terminologies in its breadth of clinical coverage, as it will feature over 300,000 controlled clinical health care concepts.

CONCLUSIONS

The process of merging two large-scale and independently developed terminologies presents several challenges. In this paper we identify similarities and differences between SNOMED RT and CTV3 and discuss our approach to the merger process.

We have attempted to structure the development process in a logical and orderly manner that optimizes the resources of both organizations. Our progress to date has demonstrated that large-scale terminology convergence is feasible, and combining the knowledge and experiences of two groups adds value to the final product. We also conclude that this process is enhanced by an open approach that involves potential users at an early stage.

Remaining challenges to the completion of SNOMED CT include harmonization of definitions, quality control, the development of detailed implementation guidance, and creating educational programs that will facilitate the use of SNOMED CT in the United States, the United Kingdom and other countries.

Acknowledgements

The authors acknowledge the significant contributions from members of the SNOMED International Editorial Board, SNOMED CT technical and content working groups, Kaiser Permanente CMT project, NHS Clinical Coding Center staff, and the American Academy of Ophthalmology, including Anne Casey, John Mason, Marcell Pooke, James Campbell, James Robb, Nick Booth, Tim Bentley, James Barrett, David Robinson, Karen Kudla, Joshua Schraeder, Corey Smith, Peter Benton, Phil Brown, Keith Campbell, Robert Dolin, Paul Frosdick, David Markwell, Shelly Nash, Jeremiah Sable, Marjorie Rallins, Julia Kim, John Fedack, Diane Aschman, and Debra Konicek.

REFERENCES

1. Spackman KA et al, eds. SNOMED RT. College of American Pathologists, Northfield, IL, Nov 2000.
2. Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Symp 1997; 640-4.
3. O'Neil M, Payne C, Read J. Read Codes Version 3: a user led terminology. Methods Inf Med 1995; 34(1-2): 187-92.
4. Schulz EB, Barrett JW, Brown PJB, Price C. The Read Codes: Evolving a clinical vocabulary to support the electronic patient record. In Conference Proceedings: Toward an Electronic Health Record Europe. Newton; CAEHR 1996; 131-40.
5. Price C, Bentley TE, Brown PJB, Schulz EG, O'Neil MJ. Anatomical characteristics of surgical procedures in the Read Thesaurus. In Cimino JJ (Ed). Proceedings of the 1996 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus; 1996: 110-4.
6. Brown PJ, O'Neil M, Price C. Semantic definition of disorders in version 3 of the Read Codes. *Methods Inf Med* 1998;37(4-5):415-9.
7. Schulz E, Barrett J, Price C. Semantic quality through semantic definition: Refining the Read Codes through internal consistency. Proc AMIA- Symp 1997; 615-9.
8. Levy D, Dolin R, Mattison J, Spackman K, Campbell K. Computer-facilitated collaboration: experiences building SNOMED-RT. Proc AMIA Symp 1998; 870-804.
9. SNOMED Clinical Terms technical specifications. College of American Pathologists, 2000. Available from: http://www.snomed.org.
10. Wang AY, Barrett JW, Bentley T, Price C, Markwell, D, Spackman KA, Stearns MQ. Mapping Between SNOMED RT and Clinical Terms Version 3: A Key Component of the SNOMED CT Development Process. Proc AMIA Annual Fall Symp 2001. In press.
11. Hole W.T., Srinivasan S. Discovering missed synonymy in a large concept-oriented metathesaurus. Proc AMIA Symp 2000; 354-358.
12. Rector AL, Nowlan WA. The Galen Project. Comput Methods Programs Biomed 1994;45(1-2):75-8.