

# Inference of a SNOMED CT Data Model

Jon Patrick<sup>1</sup>, Ming Zhang<sup>1</sup>, Donna Truran<sup>2</sup>

<sup>1</sup>*School of Information Technologies  
University of Sydney, NSW, Australia*

<sup>2</sup>*National Centre for Classification in Health  
University of Sydney, NSW, Australia*

## ABSTRACT

*This paper identifies the underlying data model inferable from the SNOMED CT data. It describes the reverse engineering of the SCT data files to create a conceptual data model of its organisation based on its data contents, not its hypothetical design model, so as to improve the means of managing and implementing the SCT data through a conventional persistent store. This also provides some potential for re-installing SCT back into an ontological store that better suits local requirements. The analysis of the resultant data model shows the SCT does not have any integrity constraints usable within a relational database model and thereby any information systems implementation will have to create its own integrity maintenance. This creates the likelihood that differences between implementers can be expected and modifications to the SCT releases can be not be assured of consistent implementation and so threatens true interoperability between information systems.*

### *Keywords:*

*SNOMED CT, information systems, interoperability*

## INTRODUCTION

The importance of terminological systems, classifications and ontologies, all with overlaps in content and purpose has significantly increased in recent years. Many different systems have been developed and implemented such as ICD-10, GALEN, UMLS and SNOMED CT (SCT). The International Classification of Diseases (ICD), maintained by the World Health Organization, is perhaps the best known classification system used in medicine. The Unified Medical Language System (UMLS), implemented by the National Library of Medicine, combines a large number of distinct terminologies into a single platform [1] [2]. The GALEN Common Reference Model (CRM) comprises elementary clinical concepts, relationships and complex concepts [3]. One study has been made to evaluate the consistency of SNOMED CT with the principles of ontologies such as subsumption and hierarchical structure according to a Description Logic (DL) perspective [4]. Also the forms of ontology have been classified and a future direction towards producing more coherent ontologies recommended [5].

Substantial effort has been made recently for a precise understanding of existing terminology systems and for applying them into operational information system applications which is necessary to assess whether a terminology system is appropriate for use in certain circumstances, or when one has to develop a new system [6]. Various terminology systems were compared from the aspect of typology and evaluated from the perspective of conceptual model and formal representation of structure [6] [7]. On the other hand, there are many real applications which discuss the issues of procedures for the integration of terminology systems and information systems. A methodology for developing and implementing the principles for maintaining the structure of large increasing terminology system was researched to provide a way to keep consistency between those systems and their applications [8]. The Veterans Health Administration in United States has evaluated the coverage of the SCT terms and applied it into their Information system [9].

## SNOMED CT DATA SYSTEM

SNOMED Clinical Terms (SCT) provides “a common language” for use by clinical specialists for clinical notes to serve for data capture, retrieval and aggregation of health data. The basic features

of SCT as an ontology can be described as a concept-based terminology, which means that each clinical idea is uniquely identified and described often with multiple descriptions. All descriptions for the same concepts are linked together. All concepts are organized into a hierarchy and, the relationships between any two concepts give information that describes characteristics of concepts.

SCT is used in a range of information systems and is often reorganised for local needs of processing and functionality. The purpose of this paper is to describe the reverse engineering of the SCT data files to create a conceptual data model of its organisation such in an EER model so as to improve the means of managing and implementing the SCT data through a conventional persistent store. This also provides some potential for re-installing SCT back into an ontological store that better suits local requirements.

An inferred data model as such will enable us to understand the extent of deviations from an underlying ideal superstructure and make decisions as to whether they form true logical extensions or are just idiosyncratic. The ultimate target is to produce a data model and an ontology that are congruent so that it can be implemented in a working information management system more readily using good data modelling principles.

## **METHODOLOGY**

The research work presented by this paper primarily focuses on constructing a conceptual data model for SCT from its own data. The method to fulfil this objective consists of three steps.

### *Data preparation*

SCT is distributed as text files transformed in some unspecified manner from their original source. This raw data needs to be restored to a framework to identify SCT's elementary data structures. Also a fast and consistent way to query and manipulate the data through the framework is needed. Hence, the raw data, consisting of the three tables *concepts*, *descriptions* and *relationships*, were placed into the relational database management system, MySQL.

### *Structure Investigation*

To create the description of a conceptual data model represented in the data for SCT, both the explicit characteristics and implicit characteristics of the data need to be identified. The explicit characteristics are detected by direct querying of the prepared relational database with various filters. Then, inferring from the explicit results, deductions can be made to decide which implicit structural aspects buried in the hierarchical structures might be discovered by programming more complex and extensive queries.

### *Data modelling*

After structural features are discovered from the SCT raw data, conceptual data modelling was commenced using the Extended Entity Relationship (EER) model. In the process of designing an EER model, the Entities and the Relationships between them have to be defined.

## **EXPERIMENTS AND RESULTS**

### *Explicit characteristics - Concepts, Descriptions and relationships*

The core contents of SCT are three tables; concepts, terms and relationships. The extended entity relationships model is shown in the figure 1. Each concept in SCT has a unique identifier and a fully specified name describing the nature of the concept. Any concept can have more than one

term, (ranging from 2 to 44) which are thereby synonyms. There are 366,179 concepts and 993,421 terms in SCT. This study only uses active concepts.

The Relationships table specifies the relationships between concepts, and the type of each relationship. Currently, 1,462,546 instances are recorded and each of those instances has a unique identifier. One relationship entry in the table consists of a source concept and a target concept as well as the relationship type. Analyses show there are 62 different relationship types existing in relationships table. A parameter in the relationships table is the characteristic type which indicates one of four roles of the relationship of target concept to source concept of which only two are relevant here. The first is that the target concept is used to define source concepts. For example, “burn of elbow” has the relationship “finding site” with “elbow structure”. The second role is that the target concept gives optional information for qualifying the source concept. For example, the concept “appendicitis” can have the relationship “onset” with “sudden onset” and “gradual concept”. The “sudden” and “gradual” actually describe the “onset” status for this disease. So the relationship type “onset” is applied to qualify the concept. . Some relationship types have more than one characteristic type according to their role in relating the target and source concepts.

The “IS\_A” relationship type is a special one which organizes all concepts into a hierarchical structure of sub- and super-type concepts. A separate study of this type of relationship has been performed. When all concepts are traced by the “IS\_A” relationship, they have the same root, “SNOMED CT CONCEPT”.

#### *Implicit characteristics - Classification Principles*

A classification is an arrangement of objects or concepts based on their essential characteristics into groups of concepts, called classes organised in a hierarchical manner. However, the SCT is not clearly a classification due to multiple inheritances in its hierarchical structure. For instance, one concept that has more than two parents will allocate this concept into several classes which is not permitted in a normal classification scheme. Therefore, in the process of investigating the classification principles of SCT, we firstly identified leaf nodes .

Top category	Frequency	Top category	Frequency
Physical force	200	Context-dependent categories	6836
Special concept	261	Observable entity	7568
Specimen	1044	Qualifier value	8266
Staging and scales	1108	Pharmaceutical / biologic product	19648
Linkage concept	1129	Substance	23022
Events	1642	Organism	26134
Environments and geographical locations	1666	Body structure	31760
Physical object	4355	Procedure	52741
Social context	5188	Clinical finding	111866

Table 1: Concept distribution across the top categories

The root concept “SNOMED CT CONCEPT” has 18 direct children. Those children are called “top categories”. If SCT uses a classification principle, all the concepts should belong to one category only. The route for each concept in the hierarchical structure and its top category were identified. The results show (Table 1) that each concept belongs to one category only, and more than 30% of concepts are classified into the “clinical finding” top category.

### *Inferred relationship pattern*

The above results reveal the interaction of explicit and implicit characteristics of the SCT's ontology and that more implicit features are able to be discovered. The relationship table records any relationship link between two concepts. The format of each record is that one source concept has a relationship with another target concept. For example, the concept "breast cancer" has the relationship "finding site" with another concept "breast structure". In classification principles, "breast cancer" belongs to top category of *clinical finding* and the "breast structure" is a concept in the *body structure* superclass. As all concepts are classified into 18 top categories and there are 62 relationship types between all concepts, we suppose that there is restricted use of specific relationship types between any two top categories. The foregoing example represents that the "clinical finding" has a relationship of "finding site" with "body structure". This can be treated as a record with a pattern form "top category – relationship type – top category". With 18 top categories and 62 relationship types, the number of different patterns is 18x18x62, 20088. Analysis of the data tables shows that of the 20,088 relationship patterns, only 96 patterns have instances in the relationship tables. Furthermore 18 patterns are "IS\_A" relationships. Hence only 78 patterns are useful for analysis of the constraints of the relationships between the top categories.

### **DATA MODELLING**

The inferred 78 relationship patterns between top categories and relationship types can be modelled to generate a conceptual model for SCT's data contents. To construct an EER data model, the top 18 categories are defined to be entities and the 78 valid patterns are the relationship between those entities. Therefore, for example, the relationships between the entity class "clinical finding" and other entity classes can be determined from the valid patterns shown in the table 2. From this information a part of the model can be generated and is shown in figure 1.

Source concept	Relationship type	Target concept	Number of instance
Clinical finding	Causative agent	Pharmaceutical / biologic product	7275
Clinical finding	Course	Qualifier value	65036
Clinical finding	Due to	Clinical finding	1242
Clinical finding	Episodicity	Qualifier value	64714
Clinical finding	Finding site	Body structure	85603
....			

Table 2: Some valid SCT concept-relationship patterns for "clinical finding".

The work proceeded to create an EER diagram for all entities and relationships resulting in a diagram that needs to be reproduced at an A3 scale to be readable. The data model reveals a number of telling truths that create issues for further implementations. Firstly, using the general rules for logical database design there would be 18 tables for the entities and 78 tables for the concept-relationship patterns, a total of 86 in all. This is to be compared to SCT's own model of 3 tables. Secondly, it is apparent that there are no integrity constraints between any relationships, in the database sense of the term, that is, there are no optionality/mandatory constraints and all cardinality is many-to-many, creating a completely unconstrained EER model. Hence in principle database management technology can offer none of its usual features for managing the integrity of the data. This lack of constraint is what SCT capitalises on in its distribution strategy in releasing 3 tables but it makes for great difficulty in managing the data for re-use in an operational information system. In fact, these results mean that any modification made to the data by any third party could always claim to be consistent with the distribution of the SCT data.

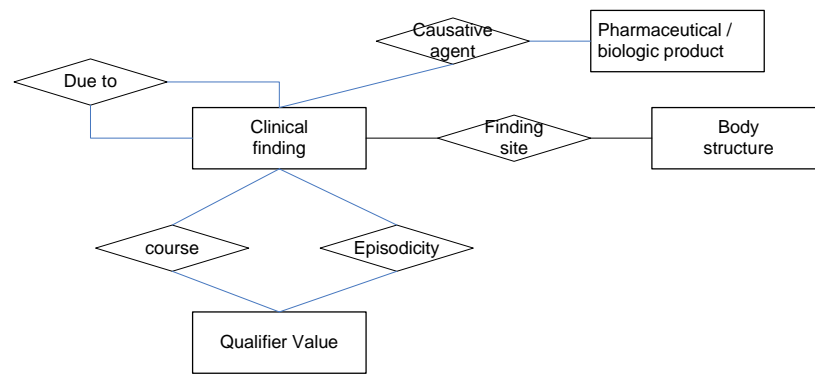


Figure 1: Part of the EER model for clinical finding.

## CONCLUSIONS

A conceptual data model of SCT has been inferred from both explicit and implicit ontology characteristics. An EER model of the features in the data has been developed. It uses the main concept classes in SCT and the internal connections between those classes. The advantage of constructing such a model is that it delivers an overview as well as providing a direct view into the interior of SCT's data configuration. From the perspective of database design, the model is a useful representation to develop an alternative approach for storing the terminology system in a robust repository rather than in the loose spreadsheet-like format. At the moment this investigative reverse engineering is incomplete without investigating structures deeper in the hierarchy and without effective representation of multiple parents and multiple relationships between concepts. Further investigation of SCT approaches to role-group relationships and post co-ordination is also necessary and planned. The analysis of the resultant data model shows that SCT does not have any integrity constraints usable within a relational database model and thereby any information systems implementation will have to create its own integrity maintenance. This creates the likelihood that differences between implementers can be expected, and modifications to the SCT releases cannot be assured of consistent implementation and so threaten true interoperability between information systems. These early findings also indicate that considerable investigation is required to delineate the boundaries between terminology models and systems, and information models, and this will be essential pre-requisite knowledge to build and maintain interoperability.

## REFERENCES

1. Humphreys BL, Lindberg DA, Schoolman HM, Barnett. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc.* 1998 Jan-Feb; 5(1): 1-11.
2. UMLS <http://www.nlm.nih.gov/research/umls/> (2006)
3. OpenGALEN. <http://www.opengalen.org>. (2006)
4. Olivier Bodenreider, Barry Smith, Anand Kumar, Anita Burgun: Investigating subsumption in DL-based terminologies: A Case Study in SNOMED CT. *KR-MED 2004*: 12-20
5. Barry Smith, Werner Ceusters. "Ontology as the Core Discipline of Biomedical Informatics: Legacies of the Past and Recommendations for the Future Direction of Research", in: Dodig Crnkovic Gordana, Stuart Susan (eds.): *Computing, Philosophy, and Cognitive Science*, Cambridge Scholars Press, Cambridge, forthcoming
6. de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. *Methods Inf Med* 2000 Mar; 39(1):16-21.
7. de Keizer NF, Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. *Methods Inf Med.* 2000 Mar; 39(1):22-9.
8. Fielding, James M., Simon, Jonathan, Ceusters, Werner and Smith, Barry 2004 "Ontological Theory for Ontological Engineering: Biomedical Systems Information Integration", *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning (KR2004)*, Whistler, BC, 2-5 June 2004, 114-120.
9. Penz JF, Brown SH, Carter JS, Elkin PL, Nguyen VN, Sims SA, **Lincoln MJ**. Evaluation of SNOMED-CT Coverage of Veterans Health Administration Terms. *Medinfo.* 2004;2004:540-544.