

MoST: A System To Semantically Map Clinical Model Data To SNOMED-CT

Rahil Qamar, Alan Rector
Medical Informatics Group
Department of Computer Science
University of Manchester
Manchester, U.K.

{qamarr,rector}@cs.manchester.ac.uk

Abstract: Data retrieved from electronic health records may be interpreted in several ways. Terminology standards often help in guiding the interpretation so that the data retains its original semantics. Current efforts point towards encoding data to these standard terminologies to resolve the issue of interpretation. This paper discusses the Model Standardisation using Terminology Systems (MoST) methodology that uses lexical and semantic techniques to find matches between clinical model data and a chosen standard terminology. The standard adopted to code the data was SNOMED-CT.

1. INTRODUCTION

Government efforts [7][4][2] to bring about *shared care* has led to official recognition for the need to standardise data in EHRs¹. Data standards have been recognised as the basic building blocks of interoperability, allowing different information systems to securely and rapidly access health information when appropriate and where needed. In addition, an interoperable information system is vital for improving healthcare, mainly because it reduces medical errors and improves efficiency [1].

Standardising data requires accurate mapping of clinical data to a chosen standard terminology. The first step towards mapping is to find appropriate semantic matches in the terminology that correctly infers the *intended meaning* of the data. The semantics of clinical data is derived not only from their individual data values but also from the context in which they are linked together as compound clinical concepts. However, it is not easy to derive semantic equivalence between data models and terminologies as they have exclusive styles of modeling and naming concepts. Large medical terminologies such as the SNOMED-CT has over 360,000 concepts and over a million triples with a rudimentary form of classification making it difficult to perform quick semantic searches.

This paper discusses an approach to achieve concept mapping of data from a European standard clinical model i.e. *Archetypes* to the SNOMED-CT terminology, referred to as SNOMED in the paper. The Model Standardisation using Terminology Systems (MoST) application implements the mapping methodology by providing candidate SNOMED matches for an archetype term. The system then allows the

archetype modeler to select the most suitable candidates for formal mapping. An archetype modeler is a clinically qualified person who is also involved in modeling clinical data. MoST has been intentionally designed as a semi-automated process so that the opinion of the modeler can be taken into account. The purpose is to ensure that only those SNOMED codes are mapped to the data that represent its true intended meaning.

1.1 The SNOMED-CT Terminology

SNOMED was chosen for the research as the external terminology serving as the standard for the clinical model data. The main reason for its selection was the range of medical areas it covered making it a richer terminology. It also provided computable concept definitions and relationships that could help determine the semantic appropriateness to its corresponding archetype data. Another research hypothesis was that if our application could successfully cope with the magnitude and complexities of SNOMED, it would be simpler to replicate the process using smaller, more specialised terminologies like LOINC² and ICD³. The Jan 2006 data version of SNOMED was used for testing the mapping approach.

1.2 The Clinical Data Model

Clinical data models form part of health information systems and serve as a backbone to templates⁴. These templates are used by clinicians to capture data in a structured form. The data models are usually governed by some formalised reference schema, which place constraints on the data values. The schemas also govern the relationship of the particular data type with other data units forming part of the same template. For example, HL7 documents⁵ conform to the Reference Information Model (RIM). Similarly, Archetype Models, commonly referred to as Archetypes, are a European standard data model that conform to the *openEHR Reference Model*⁶.

²Logical Observation Identifiers Names and Codes

³International Classification of Diseases

⁴The term *template* in this paper is used in the capacity of a clinical data-entry form

⁵Health Informatics Standards: <http://www.hl7.org/>

⁶The *openEHR Information Model* - http://svn.openehr.org/specification/TRUNK/publishing/architecture/rm/ehr_im.pdf

¹Electronic Health Records

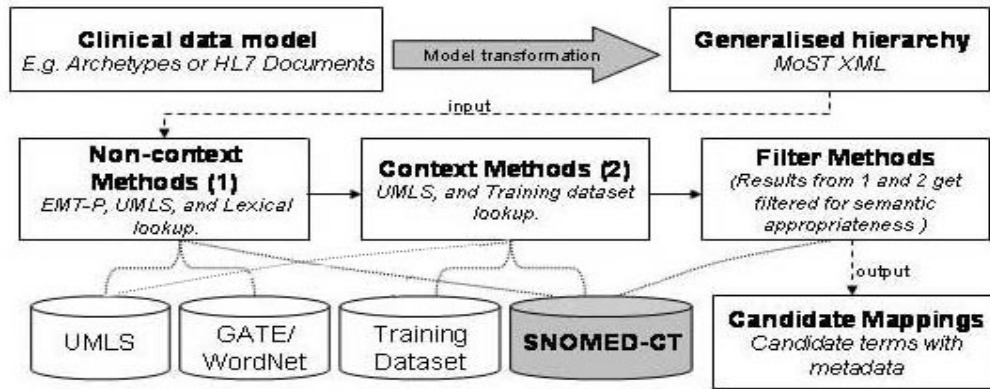


Figure 1: MoST System Methodology

1.2.1 Archetype Models

Archetype Models are computable expressions of a domain content model of medical records. The expression is in the form of structured constraint statements, inherited from the *openEHR* Reference Model [3]. The expressions are stated in the Archetype Definition Language (ADL), which is a formal programming language.

The intended purpose of archetypes is to empower clinicians to define the content, semantics and user interfaces (or templates) of systems independently from the information systems. It also proposes proper integration with terminology systems like SNOMED-CT so that reliable inferencing and decision support based on EHR data will be possible.

Archetypes were selected to test the principles of mapping clinical model data to SNOMED. The primary reason to select this model was its “terminology separation” feature. This feature separates the data expressions from the terminology used to identify the data. Each data node is assigned a unique local identifier. These identifiers have a local term description and may also have an external terminology code bound to it at a later stage. This feature reduces the risk of depending on any single terminology to complete the modeling of a clinical event. It also enables binding the clinical data to terminologies like SNOMED at a later stage, without requiring changes to the original model structure.

2. SEMANTIC MAPPING OF ARCHETYPE DATA TO SNOMED-CT

Our approach attempts to map the clinical data to a standard coded form before recording the data in the EHR. The approach is tested by mapping the data nodes to SNOMED once the entire clinical event is modeled at design-time. An extension to the approach is to map data at the time of recording, i.e. at run-time, in the event that a SNOMED code has not already been assigned or empty data nodes (free text) are later populated. Of course, not all archetype data will find a corresponding semantic match in SNOMED in which case the archetype data modelers can either rethink their modeling strategy or ignore and proceed without a SNOMED code. The main advantage of this approach is that the modeler has control over the SNOMED codes that

will be sent to the EHR. This avoids possible misrepresentation of the actual intended meaning and purpose of the original data.

The mapping methodology serves as a middleware to find semantic matches by interfacing between the Archetype and SNOMED. When working with proprietary models and schemas it is best to eliminate the details of the syntax and retain only the data hierarchy and properties. For instance, archetypes are expressed in ADL, which is not required for our purposes of achieving mappings. Therefore, the clinical content in the archetype is imported to a local XML format along with its properties and hierarchy, as shown in Figure 1.

2.1 The MoST system - Testing the approach

MoST was developed to test our mapping approach. Initially, only lexical methods were applied to non-contextual archetype data. This was followed by a thorough local and non-local context semantic search. A training dataset was also used to predict the possible SNOMED matches. Finally, the *intended meaning* was applied to all the accumulated results to filter out irrelevant matches while retaining the possible candidate matches, as shown in Figure 1.

As part of the lexical processing, the archetype terms were sent to the EMT-P⁷ service, which is an NLP system. It processes raw text entries before looking up matches in UMLS[11]. Adding this service is more beneficial when processing large medical texts but is used in MoST to exploit its NLP techniques. Some other NLP techniques that were utilised to help in constructing the search queries were word sense disambiguation using GATE⁸, term synonyms using WordNet⁹, and several term simplification methods.

Initially the semantic categorisation provided by UMLS was used to add some knowledge about the archetype term. This was supplemented with the training dataset based loosely on the principles of neural networks. The dataset created connections between the *processing terms* based on the semantic similarity between the group to which the term be-

⁷Emergency Medical Text Processing

⁸<http://gate.ac.uk>

⁹<http://wordnet.princeton.edu>

longed. A third kind of semantic process was also included to equip MoST with added *intelligence*. It involved including the intended meaning of the archetype data based on SNOMED categories such as ‘observable entity’, ‘procedure’, and ‘finding’. This layer of semantic intelligence also served as the *gold standard* when filtering the SNOMED results.

At the end of the MoST processing life cycle, all the filtered results were presented as *candidate* mappings to the clinical modeler. The modeler could then verify the appropriateness of the candidates and choose only those SNOMED codes for *final* mapping that represented its intended meaning. By completely automating the service valuable human intelligence would be lost, which might make the standard coded data inconsistent and consequently less reliable.

2.1.1 Application of filtering rules

A brief depiction of the main rules in the filtering algorithm are illustrated with the help of the (a) *autopsy examination*, and (b) *blood gas assessment* archetypes. All concept ‘*is_a*’ definitions have been derived from SNOMED.

Rule 1: If one input concept subsumes another then the subsuming concept is selected.

Example: Filtering input = {Entire respiratory system, Respiratory system structure} for archetype term ‘respiratory’ in (a).

Entire respiratory system (body structure)
is_a Respiratory system structure

Output = {Respiratory system structure}

Rule 2a: If the input concepts are ‘disjoint’ with no common subsuming concept then all disjoint concepts are selected.

Example: Filtering input = {Hydrogen ion concentration, Past history of, ph+, pH measurement arterial} for archetype term ‘pH’ in (b).

Hydrogen ion concentration (observable entity)
is_a Fluid observable

Past history of (context-dependent category)
is_a Context-dependent categories

ph+ (qualifier value)
is_a skin reaction grades

pH measurement arterial (procedure)
is_a pH measurement

Output = {Hydrogen ion concentration, Past history of, ph+, pH measurement arterial}

Since ‘disjointness’ is not explicit in SNOMED it has been inferred from the names of the concepts and the absence of common children.

Rule 2b: If all input concepts have been selected in Rule 2a then the semantic categorisation of the SNOMED codes are queried against the *gold standard*. The gold standard document consisting of the intended meaning of the archetype term is only used as a guideline. Other techniques, which are beyond the scope of this paper, are also adopted to ensure no possible candidate is incorrectly eliminated.

In the ‘pH’ example used in Rule 2a, the intended meaning is a *finding* or *procedure*. Applying this rule along with the other semantic rules results in

Output = {Hydrogen ion concentration (observable entity), pH measurement arterial (procedure)}

On filtering all the results, the final set of candidate terms were presented to the clinical modeler who determined the most relevant SNOMED code to map the archetype term to.

3. EVALUATION

The MoST system was tested against nineteen Observation type archetypes¹⁰ available publicly at the *openEHR* website¹¹. Evaluation was based on comparing the precision and recall of the results obtained in the two main stages: (i) results before applying filtering i.e. *pre-filtered*, and (ii) results after filtering i.e. *post-filtered*.

Recall can be defined as the number of terms retrieved divided by the total number of terms in the index. *Precision*, on the other hand, is the number of relevant terms retrieved divided by the total number of terms retrieved [12]. The relevance of a result is finally determined by the modeler who is presented with the candidate mappings.

3.1 Pre-filtered results

The nineteen archetypes consisted of approximately 25 clinical terms each. Therefore, a total of 475 archetype terms were sent to MoST for candidate SNOMED mappings using the *1..** matching approach. This means that one archetype term could find matches to many SNOMED concepts. On an average 0 to 15 SNOMED codes were returned as matches to each archetype term resulting in a very large recall value. However, the figures have been trimmed down to represent a *1..1* mapping for ease in understanding the overall results. Table 1 demonstrates the initial recall and precision results obtained before filtering.

Terms sent (1)	Codes retrieved (2)	Relevant codes (3)	Precision (3/2)	Recall (2/1)
475	425	350	82.35%	89.47%

Table 1: Pre-filtered SNOMED results with a high recall value and a low estimated precision score.

Table 1 shows that only 50 archetype terms did not find any match in SNOMED resulting in an overall high recall value of 89.47%. The recall value was high primarily due to the rich coverage of clinical concepts in SNOMED. At this stage, a manual inspection of the results was done to get an estimate of the current precision of the matches. It was estimated that on an average the precision of the present set of results without any filtering was 82.35%. Despite being a good value, the precision was not reliable at this point.

3.2 Post-filtered results

In order to get a reliable precision value by eliminating all irrelevant results, it was essential to apply the filtering

¹⁰Archetypes that model the recording of a clinical observation event.

¹¹http://oceaninformatics.biz/archetypes/index_en.html

algorithm described in section 2.1.1. Also it was important to apply the layer of semantic intelligence to the filtering process, which helped retain only those SNOMED codes sharing common semantics with the archetype term. On application of the filtering rules, the precision improved significantly.

Table 2 represents results after application of the filtering rules. Of the 425 SNOMED codes returned as matches only 385 were found to be relevant by the MoST system resulting in a precision of 90.58%. The precision improved significantly after filtering as most of the irrelevant results were eliminated resulting in a smaller but better final result set. This result set was displayed to the clinical modeler as candidate mappings.

Terms sent (1)	Codes retrieved (2)	Relevant codes (3)	Precision (3/2)	Recall (2/1)
475	425	385	90.58%	89.47%

Table 2: Post-filtered results consisting of SNOMED code matches to the Archetype clinical terms

On comparing the quality of the results obtained pre and post filtering it can be seen from Figure 2 that approximately 9.1% (385/350*100) of the results were eliminated during filtering.

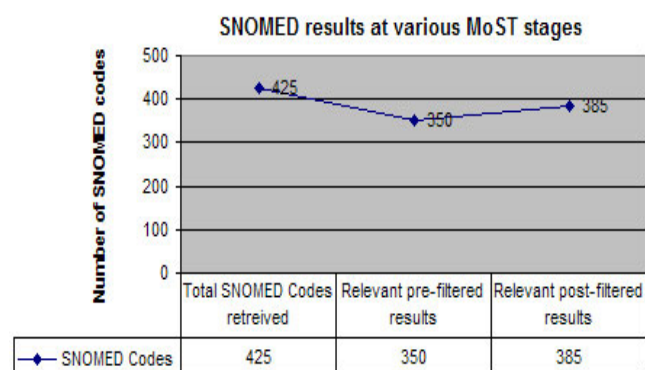


Figure 2: Comparing results obtained pre and post filtering

3.3 Discussion

- Although we talk about Archetypes and SNOMED, the principles can be applied to any other data model and terminology.
- The main purpose of demonstrating the difference in the number and quality of results obtained before and after *filtering* is to reiterate the point that middleware applications need to have more intelligence to internally eliminate as many unsuitable results as possible. This feature is critical to ensure that clinicians adopt new informatics solutions aimed at improving patient care.

4. ISSUES WITH DIFFERING MODELS

Terminologies are often developed independent of the data models that use them. They are also developed by separate

groups having little or no interaction with each other. However, if the aim of safe semantic interoperability is to be realised, a unified approach needs to be developed that brings together the strengths of the terminology model and the information model [13]. Our system aims to act as such an interfacing middleware. However, when working with these different source and target schemas several issues were encountered.

- Archetype designers do not always have SNOMED in mind when modeling archetypes. The primary source of reference is the *openEHR* Reference Model.
- There are no clear guidelines for categorising an archetype to an OBSERVATION, ACT, EVALUATION, or INSTRUCTION. An important SNOMED category named ‘Procedure’ exists loosely in archetypes making it difficult when interpreting the intended meaning.
- In SNOMED, sub concepts inherit the category of its parent concept. For example, all sub concepts of the SNOMED concept *Complete blood count (procedure)* are also ‘procedure’ types. However, in archetype models although the archetype term *Complete blood picture* is an OBSERVATION type, its sub terms need not necessarily belong to the same type. For instance, it contains terms like *haemoglobin* and *red cell count* whose intended meaning, as specified by the modeler, is ‘finding’ or ‘procedure’. Therefore, when filtering in MoST, it was difficult to eliminate results based on a strict categorisation principle. It was necessary to accommodate the fundamental differences in the objectives of the archetype and SNOMED model when attempting to map terms from one to the other. The approach used to resolve this problem was to allow the modeler to indicate which semantic category(ies) they intended the archetype term to belong to.
- To add to the complexity, SNOMED’s semantics and how it differentiates between ‘Observation entities’, ‘Procedures’, and ‘Clinical Findings’ is also not always clear. For instance, the SNOMED literature states that a concept is an *Observable entity* if it can consist of a value. However, the term *pregnancy* occurs as a *function* type below the observable entity hierarchy. Clearly, a value cannot be assigned to pregnancy. It can, however, have a *present/absent* value, which would categorise it as a *Clinical Finding* instead. Such discrepancies in categorisation often leads to problems in interpreting the semantics of a concept.
- Another classic problem is the existence of conflicting source and target term categories. For example, the *autopsy examination* OBSERVATION archetype was assigned with Finding as the intended category. However, all autopsy-related codes in SNOMED belonged to the ‘Procedure’ hierarchy, e.g. *Autopsy examination*, and *Postmortem examination*.

5. RELATED WORK

Work on mapping clinical terms to structured clinical terminologies is becoming increasingly popular. Broadly, clinical terms are extracted for mapping from three major sources: (i) medical narratives and free text, (ii) databases, and (iii)

ontologies. The research presented in this paper introduces a fourth source i.e. *clinical data models*, which alleviates the need to code data after it has been recorded. Despite a different source, the work draws its principles from previous work done in the mapping field. We compare the ways in which our work differs from some other authors.

5.1 RELMA

The LOINC database comes with a mapping program called Regenstrief LOINC Mapping Assistant (RELMA) to assist in the mapping of local term codes to LOINC codes and to browse the LOINC database [9]. RELMA enables users to specify their local terms in the mapping screen. It then returns all the LOINC codes that include the specified term. It also provides several search specifications to be added to the base term query to refine the mapping process.

5.1.1 Discussion

RELMA has a good user interface but the mapping process is rather time consuming and tedious [8]. It requires a lot of user input to guide the search process. One also needs to educate themselves completely with the RELMA mapping program to ensure they get the best performance from the system. This means that two users inputting the same query term but with different search specifications might obtain different LOINC codes as results. Such variations may lead to mapping inconsistencies or errors [8]. The MoST system requires minimum intervention from the user to perform mapping. The only user input required is at the end of the search when the final filtered SNOMED codes are to be chosen for mapping. MoST operates close to the *source* i.e. the model generating the terms for recording a clinical event. RELMA operates close to the *target* i.e. the structured terminology to which the terms are mapped.

5.2 MedLEE

The Medical Language Extraction and Encoding System (MedLEE) is a natural language processor that identifies clinical information in narrative reports and maps them to a controlled vocabulary [5]. The present-state system has been advanced to map to UMLS concepts based on structural matching using modifiers [6].

5.2.1 Discussion

MedLEE is a good medical language processor that helps obtain controlled UMLS codes. However, if mappings are to be performed to other controlled terminologies such as SNOMED-CT, or ICD, some other process would be needed subsequent to MedLEE encoding [6]. At present the MoST system utilises the EMT-P service, which is similar to MedLEE. However, in future MoST could also assess its performance when using MedLEE as compared to EMT-P, and retain the service that gives better results.

5.3 Anchor-PROMPT

Anchor-PROMPT is essentially an algorithm that finds semantically similar terms between two ontologies using lexical similarity [10]. It takes as input a set of anchors i.e. start and end nodes. Based on the sub-ontology existing between the two anchors, the algorithm determines the classes that appear frequently in similar positions on similar paths [10]. The anchors are either specified manually or are determined automatically using lexical similarities.

5.3.1 Discussion

The Anchor-PROMPT can only be used with closely related ontologies i.e. with ontologies belonging to the same domain. It is not suitable for working with ontologies belonging to different domains but using similar vocabulary. For instance, a 'clinical' ontology might have some relationship between the terms *instrument* and *hammer* when modeling the field of orthopedics. However, a 'building construction' ontology might include the same terms *instrument* and *hammer* for modeling the field of carpentry. Relying on deriving semantic similarity based on similar lexical terms is not a very efficient and safe method for mapping ontologies.

6. CONCLUSION AND FUTURE WORK

The most reliable and quick method of achieving mappings of free text or data terms to controlled terminologies is to employ various lexical and semantic procedures. Systems dependent on first building a meta-ontology to start the mapping process take long to complete and are subject to changes if the target terminology structure undergoes changes. The MoST system employs the first approach i.e. using lexical, semantic, and contextual methods to find relevant SNOMED code matches for the mapping exercise.

Work is underway to expand the scope of the matches from single, atomic (pre-coordinated) codes to compositional (post-coordinated) SNOMED codes. Attempts will be made to generate post-coordinated terms only for those archetype terms that have not already found a pre-coordinated match. Not all unmatched archetype terms will find a corresponding compositional term that will comply with the post-coordinated rules specified by SNOMED. Suggestions will be offered to the clinical modeler who may put in a request to the SNOMED committee for inclusion of a particular concept definition.

The MoST system is due to be released to the *openEHR* community as a module in the Java-based Archetype editor [14]. As the system is used by clinicians modeling archetypes to map the terms to SNOMED, a repository with accepted SNOMED codes for an archetype term in a given context will be populated. A frequency chart will be generated subsequently. This will help MoST in adding yet another layer of knowledge by forward guessing an archetype term match.

7. ACKNOWLEDGMENTS

This work is supported in part by the EU Funded Semantic Mining project and the UK MRC CLEF project (G0100852).

8. REFERENCES

- [1] Connecting for Health - A Public-Private Collaborative - FACTS AND STATS. http://www.markle.org/downloadable_assets/facts_and_stats.060503.pdf. As seen in July 2006.
- [2] The european environment & health action plan 2004-2010. Technical Report Vol I, Commission of the European Communities, Brussels, June 2004.
- [3] T. Beale and S. Heard. Archetype definitions and principles. (Revision 0.6), March 2005.
- [4] N. C. for Health. Better information better health. Technical report, NHS National Programme for IT Annual Report 2004-2005, 2005.

- [5] C. Friedman, P. Alderson, and J. A. et al. A general natural-language text processor for clinical radiology. *JAMIA*, 1(2):161–74, 1994.
- [6] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *JAMIA*, pages 392–402, 2004.
- [7] HHS. 2005 hhs e-gov annual report. Technical report, U.S. Department of Health and Human Services, Viewed in July 2006 2005.
- [8] L. Lau, K. Johnson, and K. M. et al. A method for the automated mapping of laboratory results to LOINC. *Proc AMIA Symp*, pages 472–76, 2000.
- [9] C. McDonald, S. Huff, and J. S. et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49(4):624–33, 2003.
- [10] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, March 2001.
- [11] U. D. of Emergency Medicine. EMT-P user manual ver 2.1.
<http://www.med.unc.edu/emergmed/EMTP/usermanual.pdf>.
Seen in July 2006.
- [12] M. Sarr. Improving precision and recall using a spellchecker in a search engine. Royal Institute of Technology, Sweden.
- [13] SNOMED. Guidance on implementing the SNOMED clinical terms context model. Technical Report Revision 0.3, May 2004.
- [14] E. Sundvall, R. Qamar, and M. N. et al. Integration of tools for binding archetypes to SNOMED CT. *SNOMED-CT and Semantic Mining Conference*, Oct 2006. Submitted.