

Semantic Mapping of Clinical Model Data to Biomedical Terminologies to Facilitate Data Interoperability

R Qamar¹, A Rector¹

¹ 2.89 Kilburn Building
Department of Computer Science
University of Manchester
Manchester, U.K. M13 9PL

Keywords: Semantic mapping, Data standardisation, Healthcare computing

Abstract

The lack of use of a single terminology across all health systems has led to issues of patient data interoperability. This requires data models to be independent of the terminologies when formalising data representations. The paper discusses the principles to base a middleware system to enable semantic mapping of data in the models to formal biomedical terminologies. The principles have been tested against the MoST system, which maps Archetype data to SNOMED CT codes. At first, contextual and non-contextual methods were applied using lexical and semantic procedures to obtain matches. These automated matches were then presented as candidate mappings to clinical modelers to choose from. The aim of the research is to enable clinical modelers to quickly and efficiently codify data at the time of modeling the information through automated processes. The research intends to simplify the task of mapping thereby encouraging standardising data at source by alleviating tedious manual lookups performed traditionally by clinicians.

Introduction

An unusual feature of health informatics is that biomedical terminology models and clinical data models are developed by separate groups that work independently of each other. Therefore, there are inherent differences in the principles on which each of them is based as well disparities in the semantics of representation. However, there is increasing government recognition^{1,2,3} for the need to integrate the functioning of all complementing models in health information systems to achieve data interoperability and shared care.

To achieve data interoperability it is important that data conforms to some standard, which is adopted by all conforming systems. In health care, formal biomedical terminologies provide such standards to record patient data. Patient data is increasingly being recorded using formal clinical information systems. Data models, such as Archetype Models, form part of such systems by providing structured and constrained representations of clinical recording scenarios. To facilitate data interoperability it is necessary to integrate the data in the models to standard terminologies like Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT).

The paper discusses a methodology used to perform the first level of data standardisation i.e., data mapping. The Model Standardisation Using Terminology (MoST) system was developed to test the methodology using *openEHR* Archetype Models and SNOMED CT. Context and non-context methods using lexical and semantic procedures were employed to find matches. The most appropriate matches resulting from application of filtering rules were presented as candidates to the clinical modeler for mapping. A clinical modeler is a person with medical knowledge who is engaged in the task of modeling clinical data for use in information systems. The enhancement in the precision of the results by applying semantic and filtering techniques will be demonstrated. The time, speed, and quality of the results will then be evaluated against traditional manual term matching and lexical processes.

Background

SNOMED CT

SNOMED-CT, also referred to as SNOMED in the paper, is a comprehensive terminology that provides clinical content and expressivity for clinical documentation and reporting^{4,5}. SNOMED has been developed using the description logic (DL) Ontylog⁸ to model the logical definitions of concepts. SNOMED concepts are placed in a subsumption i.e., 'is_a' hierarchy. Two concepts may

also be linked to each other in terms of role value maps, and defining or primitive concepts⁷. SNOMED has been classified using Ontylog⁸, FACT++⁹, and other reasoners, to compute and infer the hierarchies, which was the primary reason for selecting the terminology for the research. The July 2006 release of SNOMED was used for testing the mapping approach.

Archetype Models

Archetype Models, commonly referred to as Archetypes are clinical data models that conform to the *openEHR* Reference Model (RM). Archetypes are computable expressions of a domain content model of medical records. The expression is in the form of structured constraint statements, inherited from the RM¹⁰. The intended purpose of archetypes is to empower clinicians to define the content, semantics and data-entry interfaces of systems independently from the information systems. This research has used *openEHR* Archetypes as a test case.

Mapping methodology

The methodology forms part of the middleware which binds together or maps the clinical data present in the two separate and independent formalisms i.e., Archetypes and SNOMED. However, the research does not claim that all archetype data will find a corresponding SNOMED match. It only attempts to obtain as many semantic matches as possible in the shortest amount of time. The aim is to help clinical modelers to quickly and efficiently perform data mapping when building formal data models to represent different clinical observations, evaluations, and events. The words 'term' and 'data' will be used interchangeably in the paper.

The MoST System

The methodology was tested by developing a middleware application named MoST. The two main hypotheses were that (i) semantic techniques need to supplement lexical techniques to achieve higher precision, and (ii) an automated matching process provides a faster and more convenient method to perform data mapping against traditional manual processes. These hypotheses were tested against the MoST system.

When working with data from different models it is best to eliminate the details of the model syntax and retain only the data hierarchy and properties. Therefore, the clinical content in the archetype was imported to a local XML format along with its properties and hierarchy, as shown in Figure 1. Lexical and semantic techniques were used to achieve the first stage of the mapping exercise i.e., the term matching process. This was achieved using contextual and non-contextual methods.

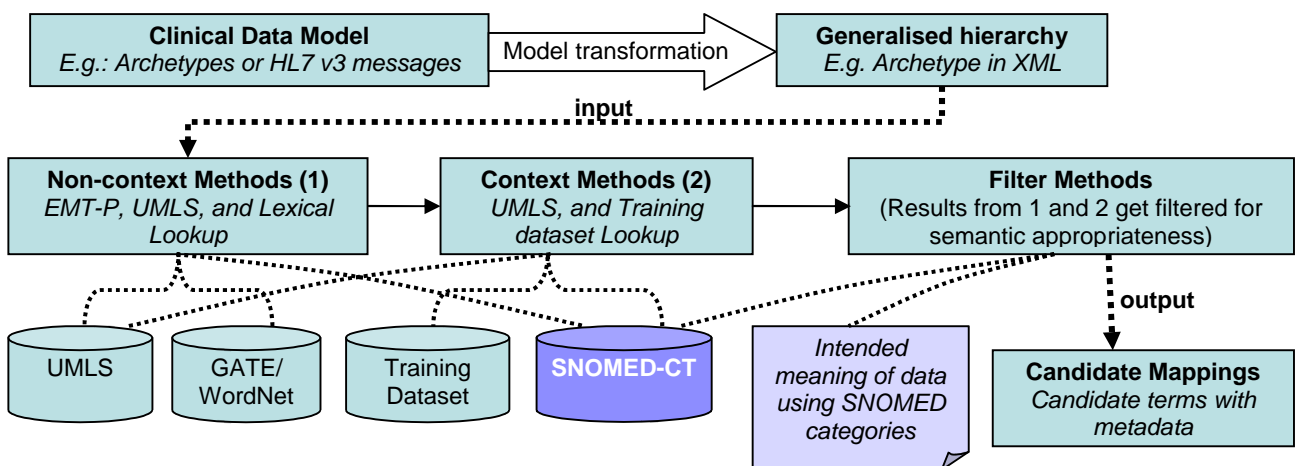


Figure 1: The MoST System Methodology

Non-context Methods

Non-contextual methods, as seen in Figure 1, included lexical searches sent (a) directly to SNOMED, (b) to the UMLS Metathesaurus, and (c) to the training dataset. The advantage of using

a resource like UMLS is that it has a large library of over a million concepts and more than 100 controlled vocabularies and classifications¹¹. The semantic groupings of these concepts were used for additional semantic information on the query data. A training data set was also used to increase the search base by including a list of clinical synonyms generated both locally as well as from the SNOMED July 2006 release.

Since MoST is not an NLP application, already established NLP systems were utilised to perform lexical processing. Initially, the archetype terms were sent to the Emergency Medical Text Processing (EMT-P) service, which processes raw text entries before looking up matches in UMLS¹². Other NLP techniques used to help in constructing the search queries were word sense disambiguation using GATE¹³, and English term synonyms using WordNet¹⁴. Local techniques applied were removal of stop words enhanced with SNOMED stop words, replacement of numeric and conjunctions with words upon unsuccessful searches, and removal or replacement of special characters and arithmetic notations.

Context methods

When sending context-based queries, it was important to consider both local and non-local context. Local context searches included the immediate parent of the archetype data being queried for a SNOMED match. Non-local context comprised of the term description available in the archetype model. This method helped the system in understanding the context in which the data was used, which aided the filtering process.

Filtering Process

The collated SNOMED results from the context and non-context search procedures were subjected to elimination based on certain DL⁶ class axioms as well as based on the intended meaning of the archetype data. The subsumption and disjoint class axiom rules were applied to the SNOMED results. Firstly, all subsumed SNOMED concepts were eliminated where the subsuming concept also occurred in the result set. An example of this rule is shown in Figure 2. However, the

subsuming concept was selected even if it was not present in the result set if several of its child concepts were present in the result set. Secondly, SNOMED does not deal with disjoint axioms. This means that it does not assert or imply a particular concept to be disjoint from another concept. Based on this property, the filtering rules included all results that did not have common parents or hierarchies.

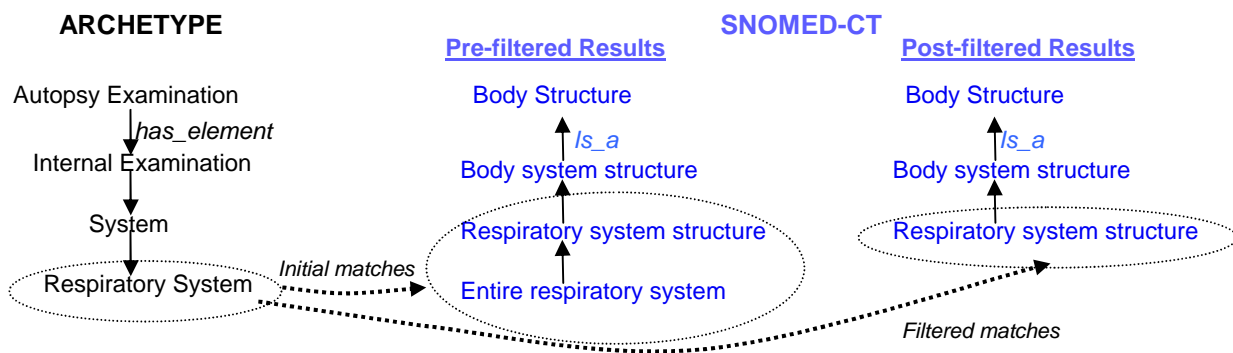


Figure 2: Filtering rule based on subsumption relationship. E.g., match for term 'Respiratory System' in autopsy examination archetype.

Finally, the SNOMED categories of the remaining results were checked against the intended meaning of the clinical modeler when creating the term in the model. SNOMED categories such as observable entity, procedure, disorder etc. were used to indicate the intended meaning. This extra layer of intelligence was useful to ascertain which results were closest in semantic similarity to the archetype term. An example of this can be seen in Figure 3, which deals with elimination of disjoint results based on their intended meaning.

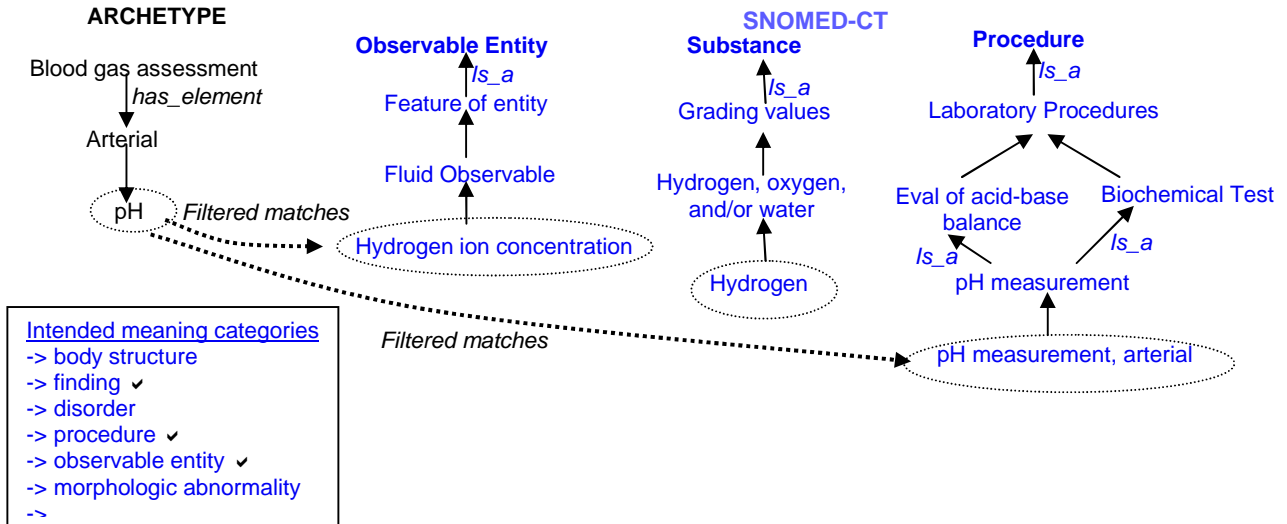


Figure 3: Filtered rule based on disjoint axiom and intended meaning. E.g., match for term 'pH' in blood gas assessment archetype using ticked SNOMED categories as intended meanings.

Candidate Mappings

On application of the various lexical and semantic procedures employed in the context, non-context, and filtering methods, a set of semantic matches were generated. These SNOMED matches represented candidates for possible mapping to the archetype terms. The automation process ended at this stage enabling the clinical modeler to select the best semantic matches to map and codify the archetype data to. This empowered the clinical expert to pre-determine which terminology codes would represent the data before being sent to the EHRs. Such a semi-automated mapping process provides a quicker and more reliable standardisation of data, as the clinical modeler is consulted with the final results. Mapped data are stored in archetype repositories and are available for further use in real-time patient data entry. Therefore, the data is easily interoperable within conforming information systems.

Evaluation

The two main hypotheses on which the semantic mapping methodology was based have been mentioned earlier in 'The MoST System' section. Ten different archetype models each consisting of approximately 30 archetype terms i.e., 300 terms were selected. Each archetype model was sent individually to the MoST system to perform only lexical matching initially followed by semantic matching.

Precision is the number of relevant terms retrieved divided by the total number of terms retrieved¹⁵. Although lexical matches for 200 of the 300 terms were found, the precision was a low 55.7%. This because the clinical modelers on manual inspection found only 110 terms to be relevant. On introducing semantic techniques and applying filtering rules to achieve closer semantic matches, the precision was a high 87.4%. 236 results were found to be relevant of the 270 results returned by MoST. The relevance of a result was determined by the modeler when presented with the candidate mappings. On an average the matching process of each archetype model consisting of approximately 30 terms took 60 seconds. A similar manual exercise to find semantic matches was done with a smaller sample of two archetype models i.e., 60 terms. It took approximately 15 man hours to complete the list of candidate matches. The list was small as each archetype term had fewer number of SNOMED candidates per term. However, the precision was a very high 97% when evaluated for relevance.

There is a significant trade off between time and precision when comparing an automated and manual mapping process. A similar manual mapping program¹⁶ took approximately 8 months to complete. In order to achieve the goal of making data interoperable it is necessary to choose a middle path, which balances both time and efficiency. The clinical modelers agreed that it was simpler to perform a quick manual lookup of SNOMED for a lesser number of unmatched

archetype terms, using the free text search in MoST, rather than a complete search for all archetype terms.

Issues

There were several issues encountered when resolving conflicting semantics of archetype terms and SNOMED concepts. E.g., the “autopsy examination” belonged to an observation archetype. The clinical modeler assigned it an intended meaning of ‘observation’ or ‘finding’. However, SNOMED had categorised it as a ‘procedure’, which upon manual examination was considered to be an appropriate categorisation given the context. Other similar issues were encountered due to the basic differences in the modelling strategies of the two models. However, it is beyond the scope of this paper to discuss them.

Related Work

The Meta Map Transfer (MMTx) Program¹⁷ developed by NLM is a similar tool to EMT-P used in the research. Similarly, MedLEE¹⁸ 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 encoding¹⁸ as well as MMTx. A future work could be to evaluate the difference in performance and precision between EMT-P, MMTx, and MedLEE, and choosing the best one. RELMA¹⁹ developed by LOINC helps in mapping local terms to LOINC codes. However, it requires a lot of user input to guide the search process. 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 errors²⁰. On the contrary, the MoST system requires minimum intervention from the user to perform mapping. The user only needs to provide the intended meaning of the data nodes, which is optional, and select from the resulting candidates to perform final mapping.

Conclusion

In theory, the principles on which the semantic mapping methodology is based can be applied to other clinical data models and terminologies. For e.g., clinical data in HL7 v3 messages should be able to map to SNOMED, GALEN²¹, or any other terminology. However, a pre-requisite is to work with a terminology that can be classified using a DL reasoner such as RACER²², FACT++, etc. This helps in ensuring that the hierarchies are based on sound logic and the class axioms can be used for computation in applications. Also, it is important to have the context in which a data node is used in the clinical model and preferably its intended meaning. These features significantly enhance performance and precision of the mapping results.

The mapping methodology discussed in the paper was successfully tested and evaluated against the MoST system. Although the paper has concentrated on pre-coordinated or single, pre-defined SNOMED concept matches there is also the possibility of creating post-coordinated terms. They are essentially a composition of two or more pre-coordinated codes. However, archetype terms are not very suitable for generating post-coordinated codes, as they are mostly post-coordinated at the time of modeling. It is important for dynamic mapping systems to build on the advantages of both lexical and semantic procedures. It is not sufficient to rely on the traditional lexical lookups alone to get a reliable set of codified data. In addition, manual mapping processes need to be replaced with intelligent semi or automated systems to take over most of the tedious search tasks. This is necessary if informatics solutions are to be readily adopted by clinicians.

Acknowledgement

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

References

1. The European environment & health action plan 2004-2010. Technical report Vol I, Commission of the European Communities, Brussels, June 2004.
2. NHS Connecting for Health. Better information better health. Technical report, NHS National Programme for IT Annual Report 2004-2005, 2005.
3. 2005 HHS E-Gov Annual Report. Technical report, U.S. Department of Health and Human Services, Accessed Sept 2006.
4. College of American Pathologists. SNOMED Clinical Terms - User Guide, January Release 2006.
5. Price C, Spackman K. SNOMED clinical terms. BJHC & IM-British Journal of Healthcare Computing & Information Management, 2000, 17(3):27–31.
6. Brachman R J, McGuinness D L, Patel-Schneider P F, Resnick L A, Living with CLASSIC: When and how to use a KL-ONE-like language, Principles of Semantic Networks: Explorations in the Representation of Knowledge, Editor J. F. Sowa, 1991, Chapter 14, 401-456.
7. Spackman K A, Campbell K E, Cote R A, SNOMED RT: A reference terminology for health care. JAMIA, Fall Symposium Supplement, 1997, 640-44.
8. Spackman K A, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontolog, motivated by concept modeling in SNOMED, Proc AMIA Symp, 2002, 712-6.
9. Tsarkov D, Horrocks I. FaCT++ description logic reasoner: System description. In Proc. IJCAR 2006, 2006, volume 4130 of *Lecture Notes in Artificial Intelligence*, 292-297, Springer.
10. Beale T, Heard S. Archetype definitions and principles. (Revision 0.6), March 2005.
11. Sun J, Sun Y. A system for automated lexical mapping. JAMIA, 2005, 334–43.
12. UNC Department of Emergency Medicine. EMT-P User Manual ver 2.1. <http://www.med.unc.edu/emergmed/EMTP/usermanual.pdf>. Accessed July 2006.

13. Cunningham H, Maynard D, Bontcheva K, Tablan V, Ursu C. Developing language processing components with GATE version 3 (a User Guide). University of Sheffield, January 2005.
14. Fellbaum C. WordNet: An Electronic Lexical Database. MIT Press, 1998. M.A.
15. Sarr M. Improving precision and recall using a spellchecker in a search engine. Royal Institute of Technology, Sweden.
16. Wade G, Rosenbloom S T, Experiences Mapping a Legacy Interface Terminology to SNOMED CT, SMCS 2006.
17. A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In Proc AMIA, 2001, 17–21.
18. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. JAMIA, 2004, 392–402.
19. McDonald C, Huff S, Suico J G. et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. Clinical Chemistry, 2003, 49(4):624–33.
20. Lau L, Johnson K, Monson K, Lam S H, Huff S M. A method for the automated mapping of laboratory results to LOINC. Proc AMIA Symp, 2000, 472–76.
21. Rector A, Rogers J, Pole P. The GALEN high level ontology. Medical Informatics Europe (Part A), 1996, 174–78.
22. Haarslev V, Moller R. Racer: An owl reasoning agent for the semantic web. In Proc. of the IWAPS of Web-based Support Systems, in conjunction with 2003 IEEE/WIC International Conference on Web Intelligence, Halifax Canada, 2003, 91—95.