

Semantic Issues in Integrating Data from Different Models to Achieve Data Interoperability

Rahil Qamar^a, Alan Rector^a

^a Medical Informatics Group, University of Manchester, Manchester, U.K.

Abstract

Matching clinical data to codes in controlled terminologies is the first step towards achieving standardisation of data for safe and accurate data interoperability. The MoST automated system was used to generate a list of candidate SNOMED CT code mappings. The paper discusses the semantic issues which arose when generating lexical and semantic matches of terms from the archetype model to relevant SNOMED codes. It also discusses some of the solutions that were developed to address the issues. The aim of the paper is to highlight the need to be flexible when integrating data from two separate models. However, the paper also stresses that the context and semantics of the data in either model should be taken into consideration at all times to increase the chances of true positives and reduce the occurrence of false negatives.

Keywords:

Medical Informatics Computing, Semantic Mapping, Term Binding, Data Interoperability.

Introduction

The field of medical informatics is growing rapidly and with it informatics solutions to improve the health care process. One of the more important issues gaining the attention of clinical and informatics experts is the need to control the vocabulary used to record patient data. Terminologies covering various aspects of medicine are being developed to bring about some standardisation of data. However, these terminologies are seldom integrated into information systems which are used to capture data at various stages of the health care process.

The paper briefly discusses a middleware application developed to map back-end data model terms to clinical terminologies. However, the main focus of the paper is to highlight the issues encountered when integrating the terminology and data models to achieve the common objective of interoperable patient data through the use of controlled vocabularies.

Archetype models, commonly referred to as archetypes, and the SNOMED CT terminology system were used to test the mapping process. Therefore, the issues discussed are focussed on these two modeling techniques.

Background

Archetype Models

In this paper, archetypes refer to the *openEHR* archetypes. Archetypes are being put forward by the *openEHR* organisation as a method for modeling clinical concepts. The models 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 RM [1]. 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 [1]. Archetypes were selected because of their feature to separate the internal model data from formal terminologies. The internal data is assigned local names which can later be bound or mapped to external terminology codes. This feature eliminates the need to make changes to the model whenever the terminology changes.

SNOMED CT Terminology

SNOMED-CT, also referred to as SNOMED in the paper, aims to be a comprehensive terminology that provides clinical content and expressivity for clinical documentation and reporting [3]. SNOMED has been developed using the description logic (DL) Ontylog [4] to allow formal representation of the meanings of concepts and their inter-relationship [5]. 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. The SNOMED hierarchy is easy to compute, 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. It has approximately 370,000 concepts and 1.5 million triples i.e. relationships of one concept with another in the terminology.

Data Integration Issues

The aim of the research is to enable health care professionals to capture information in a precise, standardised, and reproducible manner to achieve data interoperability. Data interoperability can be defined as the ability to transfer data to

and use data in any conforming system such that the original semantics of the data are retained irrespective of its point of access. Standardised data is critical to exchanging information accurately among widely distributed and differing users.

Matching clinical data to codes in controlled terminologies is the first step towards achieving standardisation of data. The research focuses on accomplishing the task of performing automated lexical and semantic matches of clinical model data to standard terminology codes using the Model Standardisation using Terminology (MoST) system. A list of candidate SNOMED codes generated by MoST are presented to the modeler who chooses from the list the most relevant codes to bind the archetype term to. However, the matching task is made difficult because of two main reasons. First, the size and complexity of the terminologies makes it difficult to search for semantic matches. Second, the ambiguity of the intended meaning of data in both the data models and terminology systems results in inconsistent matches with terminology codes. The paper will focus on the second issue and suggest some of the solutions that were developed to resolve ambiguity of purpose and use of archetype terms with respect to SNOMED.

In this paper, the term ‘modeler’ is used to denote a person with clinical knowledge who is engaged in the task of modeling clinical data for use in information systems. Four modelers were used to conduct the study and provide their feedback on the results of the mapping. The second author was also involved in the evaluation process.

Issues with Archetype Models

Archetypes can be regarded as models of use. Archetypes based on the *openEHR* specification have four main ENTRY types i.e. Evaluation, Instruction, Action, and Observation [12]. For the research, only Observation archetypes were considered as they are the most commonly used model type with the maximum number of examples. However, there is no strict guideline used to categorise archetypes. Also all archetype terms contained in a particular archetype model do not necessarily belong to the same archetype model category. For example, the ‘autopsy’ archetype¹ belongs to an Observation type. However, the ‘cardiovascular system’ term contained in the model could belong to either the SNOMED category ‘body structure’ or ‘finding’ based on the intended meaning of the modeler. The SNOMED category ‘clinical finding’ is referred to as ‘finding’ and ‘observable entity’ is referred to as ‘observable’ in the paper.

1) *Determining the semantics of use*: The main issue encountered when looking up matches for archetype terms in SNOMED was the semantics of use. As stated earlier, there are no strict guidelines to categorise either the archetype models or the terms contained in them. The main reason being that archetype models are intended to be used as archetypal representations of a particular clinical scenario. In addition, archetype modelers do not always have SNOMED in mind when modeling clinical scenarios. For example, the recording

¹ openEHR archetypes obtained from http://svn.openehr.org/knowledge/archetypes/dev/html/index_en.html

of the apgar score of a neonate at 1, 5 and 10 minutes from birth will not only require assessment of the breathing, color, reflexes, heart rate, and muscle tone but also the total score calculated at the end of the assessment. Using SNOMED, the assessment terms can be categorised as observables or findings, among other intended semantics, while the ‘total’ could be a qualifier value or an observable with a value. Therefore, it is important when working with two different models that loose semantics are maintained initially unless stated otherwise. Too much reliance on the categorisation of terms in a particular model will result in fewer matches with terms categorised differently in another model. Strict adherence to categories can only be maintained when working within the same modeling environment as it can be assumed that the hierarchies are logically sound and classifiable.

2) *Determining the source of semantics*: Another issue that was commonly observed was determining the main source of the semantics of an archetype term. For example, in the ‘blood film’ archetype, the term ‘haemoglobin’ had a local meaning of ‘the mass concentration of haemoglobin’, shown in Figure 1. The modeler was questioned whether he would prefer a match for ‘haemoglobin’ or ‘haemoglobin concentration’. The suggestion was that a match for the later term would be closer to its intended meaning. However, based on this suggestion, the assumption of assigning more weightage to the term definition proved incorrect in the ‘autopsy’ archetype. Conversely based on the above weightage, the term ‘cardiovascular system’ defined in the model as ‘findings of the pericardium, heart and large vessels’ should have resulted in ‘findings’ of the cardiovascular system instead of the ‘body structure’ itself. However, in this case the modeler stated that the use of the term was intended to serve as a label i.e. ‘body structure’ rather than the actual values i.e. ‘findings’ to be entered during data-entry. Therefore, it is advisable not to depend on any single semantic source when looking up matches for an archetype term.

3) *Spelling errors leading to incorrect or no matches*: Finally, there was the issue of resolving spelling errors in the model. For instance, ‘haemoglobin’ using U.K. English was spelt as ‘haemoglobin’ in the ‘blood film’ archetype. Such spelling errors can give rise to incorrect or no results.

Terms **Blood Film Archetype Model (partial)**

- Complete blood picture (5 SNOMED)
 - (HISTORY)
 - any event
 - (ITEM_TREE)
 - Haemoglobin (5 SNOMED)**
 - Red cell count (RCC) (11 SNOMED)

The mass concentration of haemoglobin ← Term Definition

Terminology Services

UMLS \ MoST \ SNOMED CT \ Matching Results (SNOMED Codes)

Code	Concept
118539007	Mass concentration (property) (qualifier value)
38082009	Hemoglobin (substance)
123767004	Hemoglobinemia (disorder)
365809007	Finding of hemoglobin concentration, dipstick (finding)
302781000	Dipstick assessment of hemoglobin concentration (procedure)

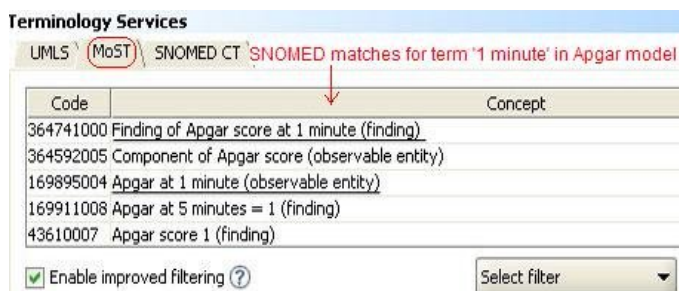
Enable improved filtering ?

Figure 1: Haemoglobin-related issues in Blood film archetype

Issues with SNOMED CT

SNOMED can be regarded as a model of meaning. It is a reference terminology for clinical data that provides a common reference point for comparison and aggregation of data about the entire health care process [5]. The multi-axial hierarchy and concept definitions are easily computable to determine the category in which the concept belongs as well as its definition in the model.

1) *Discrepancies in categorisation:* Although SNOMED has a well-defined list of categories; it is not always clear what the basis is for differentiating a concept from being an observable, procedure, or finding. Concepts in the ‘observable’ hierarchy represent a question or procedure which can produce an answer or a result [6]. On the other hand, concepts in the ‘finding’ hierarchy represent the result of a clinical observation, assessment or judgement, and include both normal and abnormal clinical states. For example, ‘colour of nail’ is an observable whereas ‘gray nails’ is a finding [6]. However, the problem arises when discrepancies occur in the categorisation of certain concepts. For instance, the term ‘pregnancy’ occurs as a sub type of ‘urogenital function’ in the observable hierarchy. Clearly, a value cannot be assigned to pregnancy. However, it can have a present/absent or positive/negative value, which would then categorise it as a ‘finding’. Such discrepancies in categorisation often lead to problems in correctly interpreting the semantics of a concept.



Code	Concept
364741000	Finding of Apgar score at 1 minute (finding)
364592005	Component of Apgar score (observable entity)
169895004	Apgar at 1 minute (observable entity)
169911008	Apgar at 5 minutes = 1 (finding)
43610007	Apgar score 1 (finding)

Figure 2: SNOMED matches obtained for term ‘1 minute’ in the Apgar archetype model

2) *Multiple representations and categorisation of similar concepts:* Another feature of SNOMED is that it allows composition or post coordination i.e. the ability to combine two or more existing concepts in the terminology to represent new meanings [5]. However, this very feature results in multiple representations of the same concept, some of which may contend with the category definitions provided in the SNOMED documentation. For instance, an ‘apgar score at 1 minute’ can be represented as an ‘apgar at 1 minute (observable)’ with a value (say 0), or with the help of a pre-existing concept ‘apgar at 1 minute = 0 (finding)’. Interestingly, SNOMED has a concept ‘apgar at 1 minute’, which is a ‘finding’ as well. Its fully specified name (FSN) is ‘finding of apgar score at 1 minute’, as shown in Figure 2. Therefore, two concepts with similar names i.e. ‘apgar at 1 minute’ belong to two different categories i.e. observable and finding. Also, Figure 3 shows that the concept ‘finding of apgar score’ has a synonym ‘observation of apgar score’, which leads to concerns about the differentiation between an observation and a finding as stated by the SNOMED community. It may be easy for the human eyes to differentiate

between the two and perform either post-coordination or simple semantic mapping. However, it becomes difficult to train a computer application to follow any strict rules for drawing inferences based on the documentation available. A similar example of ambiguity arising from close similarity of observables with findings is the presence of two very similar concepts e.g. ‘Finding of reflex hearing response (finding) is_a Audiological test finding (finding)’ and ‘Reflex hearing response (observable entity) is_a Audiological test feature (observable entity)’ in the July 2006 edition of SNOMED. In terms of aiding computation, it might be helpful to insert disjoint axioms in the hierarchy to enable applications mining through the large corpus of SNOMED data to differentiate between disjoint or non-similar concepts.

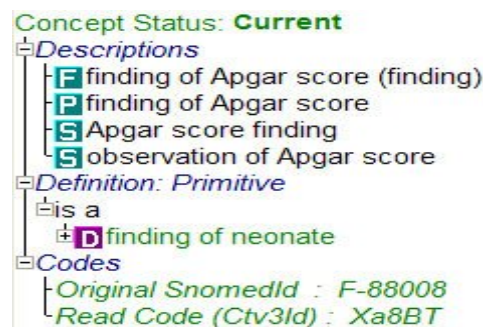


Figure 3: List of SNOMED terms belonging to the concept ‘finding of apgar score’ in the July 2006 SNOMED release.

Issues arising from mapping

In the previous two sections we discussed some of the issues present in Archetypes and SNOMED. The issues highlighted in this paper mainly concern disambiguating the semantics of concepts with respect to their categorisation in both models. This issue has been found to be one of the most difficult to resolve when automating the process of finding matches of archetype terms to semantically similar SNOMED concepts[2][10][13].

1) *Grouping SNOMED categories to improve matches:* In relation to finding matches for terms of Observation type archetypes, three main SNOMED categories were identified to which several archetype terms could generally belong to. These categories were ‘observable’, ‘procedure’, and ‘finding’. The assessment was made by manually inspecting the most common category (ies) to which relevant archetype term matches in SNOMED belonged. The relevance of a match was determined by the modeler. However, this assumption is not intended to mean that matches from other SNOMED categories were irrelevant or not possible. The initial hypothesis had to be modified such that archetype terms from an Observation type archetype model may not necessarily belong to a SNOMED ‘observable’. ‘Finding’ was the next closest category a SNOMED match could belong to followed by ‘Procedure’.

2) *Providing detailed SNOMED concept information to make informed decision:* When discussing the issues with archetype models, we exemplified how the archetype term and its definition helped in determining the intended use of the term. In the ‘blood film’ archetype, the archetype term ‘haemoglobin’ found more than one SNOMED match. Among

the SNOMED matches obtained, two of the results worth considering were ‘hemoglobin’ categorised as a ‘substance’ and ‘hemoglobin concentration’ categorised as a ‘finding’ in SNOMED. At first, the modelers on general lexical lookup of the results were of the opinion that the second match was more relevant. However on examining their SNOMED definitions and FSNs, it became obvious that the ‘haemoglobin concentration’ matches were synonyms for two different FSNs, namely, ‘Dipstick assessment of hemoglobin concentration (procedure)’, and ‘Finding of hemoglobin concentration, dipstick (finding)’, as shown in Figure 1. Both the matches were associated with the device ‘dipstick’, which was not necessarily the intended method to determine the haemoglobin concentration. Hence both the ‘haemoglobin concentration’ results were later rejected by the modelers as suitable mappings for the archetype term ‘haemoglobin’. Therefore, providing the modeler with more information about the SNOMED results rather than a simple, non-informative lexical list helped in reducing incorrect mappings.

3) *Including archetype context:* Archetype models follow an object-oriented style of modeling, which means that each element in the model conforms to some entity or the entity’s attribute in the Reference Model. The elements and their values in archetypes are represented in a post coordinated or composite manner. For instance, in the ‘apgar score’ archetype, the element ‘breathing’ is constrained by the values ‘no effort’, ‘moderate effort’, and ‘crying’, as shown in Figure 4. In a pre coordinated form the same concepts could be represented as ‘no effort breathing’, ‘moderate effort to breathe’, and ‘crying or breathing normally’. Therefore, when looking up matches for archetype terms in SNOMED it is necessary to include the context in which the particular term has been used in the model. Well-stated terms can often lead to pre coordinated matches in SNOMED. However, certain terms which are verbose or ambiguous can be unhelpful in finding sensible SNOMED matches despite considering its context of use and stated meaning. Some examples are the terms ‘Grimace and cough/sneeze during airways suction’ as a constraint value of the element term ‘reflex response’ in the apgar archetype shown in Figure 4, and ‘Needs help but can do about half unaided’ as a constraint value of the element term ‘Dressing/undressing’ in the barthel index archetype. The presence of such terms in archetypes often leads to irrelevant or no matches. However, these searches expend valuable processing time to simplify the term.

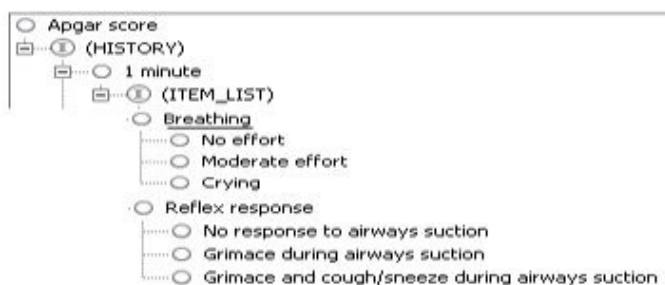


Figure 4: Apgar score archetype with a post-coordinated ‘Breathing’ term and verbose ‘Reflex response’ values.

Methodology to address issues

The MoST system was developed to perform automated

matches of archetype terms to SNOMED. In the process of obtaining lexical and semantic matches, several of the issues discussed above were addressed and computable solutions were developed.

The entire SNOMED content was imported into a MySQL database and various SQL queries were run at different stages to extract the required information. The Archetype Models were parsed from their local Archetype Definition Language (ADL) format to a simpler XML format which retained only the class containment (element-value pair) hierarchy along with its data types and some other useful information. Language specific information was discarded as it was not required for the matching process. Each term (element or value) was sent to the SNOMED database to look for matches and extract their definitions. Details of the MoST matching process has been discussed in [11].

The spellings of archetype terms were checked against the GSpell spell checker provided by the National Library of Medicine (NLM). However, GSpell performed erratically when encountered with numerics such as ‘apgar 1 minute’, prepositions, and certain words such as ‘width’ or more specific archetype data labels such as ‘DateTime’. A list of SNOMED and local stop words i.e. words to eliminate from a query terms to prevent too many results or superfluous spell checks, were used to improve results.

MoST utilised already established Natural Language Processing (NLP) systems 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 UMLS [7]. Other NLP techniques used to help in constructing the search queries were word sense disambiguation using GATE [8], and English term synonyms using WordNet [9]. 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. 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 classifications [10]. The semantic groupings of these concepts were used for additional semantic information on the query terms. 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.

Resolution of Issues

1) *Resolving issues of semantics of use:* In order to address the issue of range of permissible categories to which matches could belong to, rules regarding inclusion of SNOMED categories such as observable, finding, and procedure were introduced. Results that were lexically or semantically similar to the archetype term were returned as results whether or not they had been clearly stated as intended categories by the modeler at the outset. It was also noted that certain other SNOMED categories such as situation (earlier known as context dependent category), qualifier value, attribute, and disorder could also be some other categories results from

which may be worth including in the final result set, if appropriate.

2) *Resolving issues of source of semantics*: The issue of ambiguity of whether the term itself or its definition and context in the model are to be weighted higher was resolved by including all result variants. This may lead to a higher rate of false positives in some cases but will reduce the number of false negatives, which is more important. Therefore, both 'haemoglobin', as well as 'haemoglobin concentration' matches was included as results if it met other filtering criteria, as shown in Figure 1. Similarly, the archetype term 'total' used to record the total apgar score resulted in both SNOMED concepts 'total (qualifier value)', and 'total apgar score (observable)'.

However, certain issues such as training the application about the guidelines to follow in eliminating results from certain SNOMED categories if the terms belonged to Observation archetypes could not be implemented in a strict form. The reason was primarily the ambiguity of term categorisation when working with two separate models that differ in their fundamental objectives. While archetype models include all terms required to record a particular clinical scenario, SNOMED models the logical position of a concept in the given domain and how it relates to other concepts in the hierarchy. All these related concepts may or may not be used by an archetype model to represent a particular clinical record.

Discussion

All the results obtained at the end of the matching process were presented to the modelers who then chose the most relevant i.e. semantically similar SNOMED result code(s) to map the archetype term to. However, not all archetype terms found relevant SNOMED codes. The MoST system thus helped the modeler to codify the local terms to a standardised terminology code. 87.4% of the SNOMED codes were found to be relevant at the end of the first phase of the evaluation involving 300 archetype terms. Details of the MoST system and the evaluation can be obtained from [11], as it is beyond the scope of this paper.

With more and more archetypes being bound to SNOMED codes, the data entry process will become more standardised. Data stored in Electronic Health Records (EHRs) will begin to conform to a single reference terminology. This will increase the ease, speed, and accuracy with which data from the EHRs will be interpreted and used by health organisations irrespective of its place of capture and storage. Interoperable data will help the interoperability of clinical information systems, ultimately leading to the safe use of data and reduction in medical errors originating from incorrect data.

Conclusion

The paper focused on the semantic issues of archetype models and SNOMED, as well as the issues encountered when automating the task of matching terms from the archetype model to SNOMED terminology codes. Although specific models were chosen to test the methodology i.e. archetypes

and SNOMED CT, the approach can be tested using other similar models. However, due to the disparities in the objectives and use of the different models, care needs to be taken to address the issues arising from the disparities. Suitable solutions need to be developed to address these issues at two levels of granularity. The first level is the local level where a knowledge layer needs to be added to resolve local issues, which has been discussed in this paper. However, a second, higher level of resolution needs to be adopted when the issues concern the fundamental principles on which the models and their content are based. This will require raising a change request to each of the modeling communities.

It is important to raise issues if any real work on integrating data to achieve quick and accurate data interoperability is to be achieved. This paper hopes to start a discussion on more issues encountered by other such similar integration efforts. The work will be helpful to organisations such as the NHS in the UK who have proposed to use SNOMED CT as the standard medical terminology, and openEHR Archetypes as one of the clinical data modeling techniques.

Acknowledgement

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

References

1. Beale T, Heard S. Archetype definitions and principles. (Revision 0.6), March 2005.
2. Bodenreider O, Burgun A, Linking the Gene Ontology to other biological ontologies, *Proc. ISMB2005 SIG meeting on Bio-ontologies*, 17-18, 2005
3. Price C, Spackman K. SNOMED clinical terms. *BJHC & IM-British Journal of Healthcare Computing & Information Management*, 2000, 17(3):27-31.
4. 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, *AMIA Symp 2002*, 712-6.
5. Spackman K, Campbell K, Cote R. *SNOMED RT: A Reference Terminology for Health Care*
6. SNOMED Clinical Terms User Guide Jan 2006 Release.
7. UNC Department of Emergency Medicine. EMT-P User Manual ver 2.1. <http://www.med.unc.edu/emergmed/EMTP/usermanual.pdf>. Accessed July 2006.
8. Cunningham H, Maynard D, Bontcheva K, Tablan V, Ursu C. Developing language processing components with GATE version 3 (a User Guide). University of Sheffield, 2005.
9. Fellbaum C. *WordNet: An Electronic Lexical Database*. MIT Press, 1998. M.A.
10. Sun J, Sun Y. A system for automated lexical mapping. *JAMIA*, 2005, 334-43.
11. Qamar R, Rector A. Semantic Mapping of Clinical Model Data to Biomedical Terminologies to Facilitate Data Interoperability, *HealthCare Computing Conference 2007*.
12. Beale T, Heard S, Kalra D, Lloyd D (2006), *The openEHR EHR Information Model (Revision 5.0)*.
13. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing, *JAMIA*, 392-402, 2004.