Winter 2010-11-17

Terminology Enhanced EHR: integration of archetypes and terminology, an implementation experience

Sheng Yu
Technological University Dublin, sheng.yu@tudublin.ie

Damon Berry
Technological University Dublin, damon.berry@tudublin.ie

Recommended Citation

Sheng, Y., Berry, D.: Terminology Enhanced EHR: integration of archetypes and terminology, an implementation experience. 2010.

Terminology Enhanced EHR: integration of archetypes and terminology, an implementation experience

Sheng Yu, Damon Berry

TeaPOT Research Group,

Dublin Institute of Technology

Abstract

The integration of terminology and EHR information models is an important step in the journey towards semantic interoperability. Archetypes and two-level models for EHRs provide a mechanism that not only applies constraints on clinical content but also ensures effective terminology binding. However the lack of a standardised mechanism to bind terminology to the EHR and the difficulty of systematically coding clinical content, has led to a number of possible implementation choices.

This study presents a review of the problems that may occur when working with modern terminology systems and discusses some related state of art technologies. The paper aims to share the experience of prototyping a minimum terminology integration service. A set of tools utilising medical text processing and a customised SNOMED-CT data source are the output of prototyping that enables quick processing of archetypes and automatic link suggestions to SNOMED-CT. The elaboration of prototypes of this sort can be used as components of an integration engine.

Keywords: *EHR, Archetypes, Terminology, Term binding, SNOMED-CT*

Introduction

The goal of semantic interoperability in the e-health domain is an elusive but worthwhile one that both the industry and the research community are actively pursuing. In the absence of semantic interoperability, heterogeneous systems have the potential to cause integration difficulties and possible mis-interpretation of information during data exchange. This is a recurring issue in the current e-health environment.

The Electronic Health Record is not merely implemented as a replacement of the paper records but also seeks to adopt an approach that utilises a sharable and reusable data model to promote this common understanding between users of different systems that exchange health related information. Consequently, standard development organisations such as CEN TC251 who have an interest in the EHR have been engaged in data modelling activities that aim to provide a generic and flexible data structures for recording clinical information.

Two complementary approaches that support EHR development: information models and terminology

Background

EHR standards such as EN13606 (Eichelberg 2005) have introduced information architectures that are based on *two-level models*. In the two-level approach to modelling health information, if a piece of clinical data is to be shared between co-operating systems, it must comply with a set of constraint rules that together define a dataset for that clinical concept. The first level, which is called a *Reference Model*, is carefully designed to represent a set of fundamental and general classes that are relatively common across all health specialties and scenarios.

The second level in the two-level model approach is a flexible *Archetype model* (Beale 2002) Archetypes are detailed clinical models expressed as a set of constraints that specify and organise the classes from the reference model to express specific clinical ideas (e.g. blood pressure, heart rate. They define a maximal recordable information set that can be reused by health professionals. In the recently published ISO EN13606 standard for electronic health record communication, *archetypes* are intended to be designed by clinical experts to enforce such rules. Thus a pair communicating systems will only need to know about Archetypes in order to interoperate and in fact Archetypes are designed to facilitate health data quality enforcement and interoperability.

In contrast to and concurrent with development of structural characteristics in the data model of the electronic health record, symbolic representations of the meaning and context of the clinical information are developed as “Terminology” in health care. A medical terminology is the terminology relating specifically to topics in medicine. It has many aliases such as “controlled vocabulary”, “clinical terminology” and “coding system”.

Terminology in health care is regarded to be as old as computers, because initially shorthand codes and terms were invented and designed to minimise disk space usage. For example, a textual description of “Diabetes Mellitus” can be shortened by simply using a term like “DM” or even a code that can be understood by the computer. The history of using codes pre-dates the origin of digital computers. The idea of coding lies in the use of symbolic or alphanumeric representations to refer to agreed concepts or real world objects.

Research and development work into clinical terminologies to classify a wide range of clinical phenomena has become increasingly important. The introduction of electronic health records and EHR systems opens the possibility that increased automation of clinical process can be supported by embedding terminology from terminological systems such as SNOMED-CT within e-health information systems. Given this potential, electronic health record approaches such as EN13606 are designed to work seamlessly with terminology systems.

In order to enhance the quality of communication method with commonly understood codes, software and services are implemented to ensure the correctness of coding medical data. The recent development in SNOMED-CT reflected the growth of terminology demand in healthcare. It represents a comprehensive standard vocabulary for medical term use, which is ready to be supported by EHR information models for integration. Related work in terms of encoding clinical information is enormous (Ruch et al. 2008). The axis of interest of this study is the integration of binding terminology reference to EHR information model artefacts, which are archetypes.

EHR information models represent the syntax side of the communication while the terminologies represent the semantic side. In order to achieve semantic interoperability the two have to work seamlessly to provide meaningful and reusable clinical information (Cimino 1998). One big challenge is that with the flexibility provided by EHR information models, it can be difficult to link clinical information to the appropriate code or term that should be referenced properly for reuse. This lack of integration is due to the parallel development of both information models and terminologies.

This study aims to solve the problem of integrating SNOMED-CT with archetypes in a bid to bring terminology and EHR together among many other emerging terminology binding technologies.

Terminology standards

From a terminologist's point of view, medical terminologies can be classified under many criteria.

	For nurses	**Surgical Procedures only**	**Diseases**	**Laboratory**	**Drugs**	**Billing**	**Epidemiology and Statistics**	**All medicine**
ICD	no	no	**yes**	no	no	no	**yes**	no
ICNP	**yes**	no	no	no	no	no	no	no
SNOMED	**yes**	**yes**	**yes**	**some**	**some**	no	**yes**	**yes**
CPT	no	**yes**	no	no	no	**yes**	no	no
ICD9-CM	no	no	no	no	no	**yes**	no	no
DPD	no	no	no	no	**yes**	no	no	no
LOINC	no	no	no	**yes**	no	no	no	no

Table 1- coverage of different terminologies with respect to different medical domains (Rogers 2010).

Other standards such as HL7 include codes for clinical data. These internal code sets are associated with a information model. Other similar code sets are incorporated into openEHR and part3 of the EN13606 standard. These code sets are designed to be used by systems that adopt the information model. Overlaps between these internal codes and medical terminologies listed in table 1 above have provoked integration issues (Markwell et al. 2008).

SNOMED-CT is the choice of this study because of its broad multi-purpose applicability, which mean that it can serve as a general reference terminology (SNOMED 2006). It covers a wide range of medical domains which makes it suitable for integration with EHR information models.

Terminology services and tools: review of state of art and related work

With so many, sometimes overlapping terminologies available for different purposes, a guide for implementing terminological services should be provided to encourage implementation. The following section reviews some of the key technologies in this field and related work.

There are many types of terminology applications and software to support provision and consumption of terminological resources. The authors categorise them into two broad categories: *terminology services* which are providers of the terminology; *and local terminology tools* which utilise the terminological resource and interact with the provider.

A terminology service will typically endorse tight integration with its client applications. However following successive initiatives by HL7 group and by OMG Healthcare DTF, recent effort has been made to standardise and develop a common interface for accessing terminology by healthcare applications. This work has resulted in a new emerging specification named "Common Terminology Service" (Apelon 2009b) that could greatly promote the standardisation of terminology application development. Originally designed only to work with HL7 message standards, the current state of this specification has moved on to accepting version 2 proposals which intend to be platform and implementation independent.

Commercial products such as terminology server applications are available from several venders. Products taken for investigation include an open source terminology server DTS from Apelon (Apelon 2009a), and a commercial licensed terminology service from Ocean informatics (Ocean 2008). Followed a trial evalution of these applications, the authors are of the opinion that these commercial products are mostly extensions of what is specified as a minimum terminology provider by the common terminology service standard.

Client terminology applications vary according to their intended purpose. One example application is a standalone program which queries a terminology service to resolve a code. Another is the tool used at design or configuration time where the client queries the server to give specific information. Terminology subsetters perform customisation and subsetting of big terminology sets (Ocean 2008). Others may entitle users to edit and author terminology content.

Despite the existence of healthcare applications that directly embed codes in the screens, forms, and data entries, there are also tools to map clinical information to standard codes. These tools will inevitably need to search for concepts based on text. They are categorised as mappers and used for either mapping local terms to standard terms or searching for an appropriate code based on free text. RELMA from LOINC (Regenstrief 2009) is such an example.

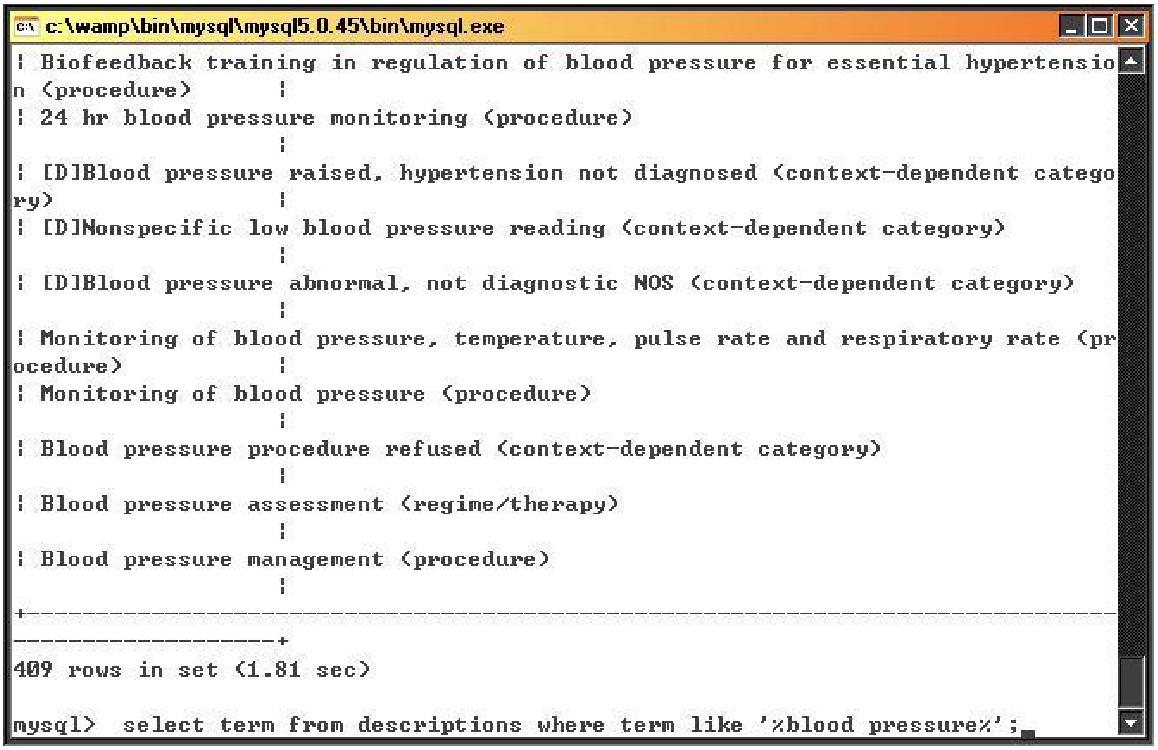
Experience from implementing a terminology service for SNOMED-CT

In order for archetypes to represent clinical concepts in a way that is commonly understood, they need to be linked to unambiguous concepts in agreed terminologies. That is because although an Archetype contains constraints on a clinical model, it only uses a set of locally defined terms and the information from the underlying reference model in order to define the constraints. Unless bindings to external terminology exist there is no restriction on representation of clinical concepts within archetypes.

To investigate the effort and tasks required to develop a basic terminology service, the authors started from the release files of a SNOMED-CT distribution. Each SNOMED-CT release provides three core tables. Their content can be briefly described to consist of SNOMED-CT concepts, descriptions, and relationships. The hierarchies of its concepts are formed mainly by the inheritance relationship between them (i.e. IS_A aka subtype relationship).

In order to be able to query this large data source it has first to be loaded into a database. A normal code resolving-query can be easily handled by a DBMS. However when it came to providing fast full text search on the database, the result was not satisfactory.

Text searching will effectively help users of SNOMED-CT to locate and finalise a term that should be used in an Electronic Health Record and where appropriate in any healthcare application. However simplicity, usability and efficiency are important to a terminology service that is provided to end-users. So usability parameters such as the search speed and result relevance ranking need to be optimised.

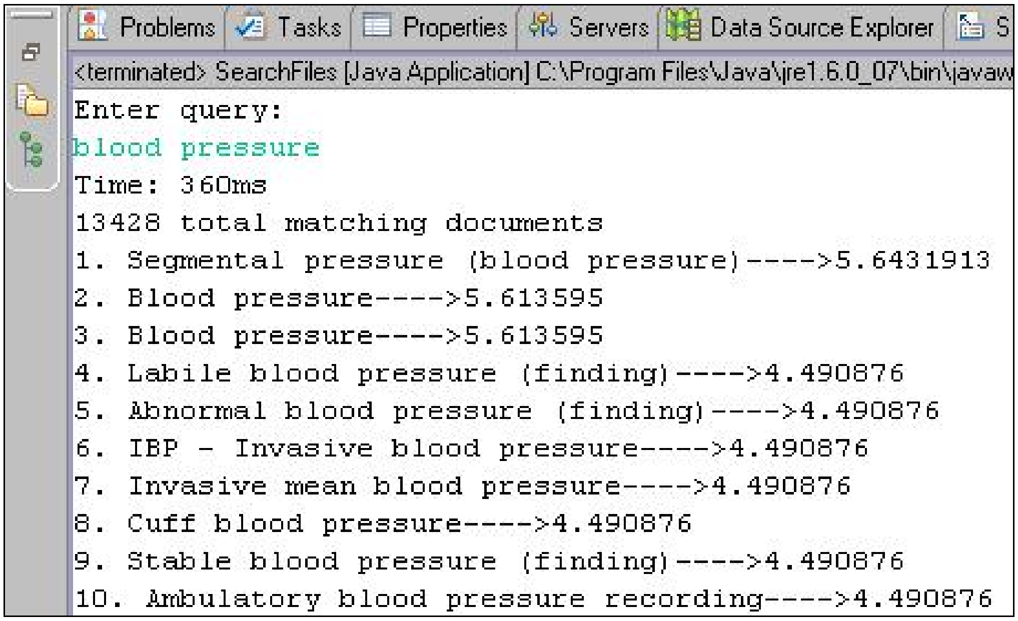


```
¦ Biofeedback training in regulation of blood pressure for essential hypertensio
n (procedure)      ¦
¦ 24 hr blood pressure monitoring (procedure)
                   ¦
¦ [D]Blood pressure raised, hypertension not diagnosed (context-dependent catego
ry)                ¦
¦ [D]Nonspecific low blood pressure reading (context-dependent category)
                   ¦
¦ [D]Blood pressure abnormal, not diagnostic NOS (context-dependent category)
                   ¦
¦ Monitoring of blood pressure, temperature, pulse rate and respiratory rate (pr
ocedure)           ¦
¦ Monitoring of blood pressure (procedure)
                   ¦
¦ Blood pressure procedure refused (context-dependent category)
                   ¦
¦ Blood pressure assessment (regime/therapy)
                   ¦
¦ Blood pressure management (procedure)
                   ¦
+-------------------------------------------------------------------------------
-------------------+
409 rows in set (1.81 sec)

mysql>  select term from descriptions where term like '%blood pressure%';
```

Figure 1 - SQL query to select concepts contain ‘blood pressure’ on SNOMED database

One adaption to improve the usability of the terminology service was to use information retrieval techniques. Lucene (Gospodnetic and Hatcher 2005) is a full text indexing and searching engine using TF-IDF (Yates and Neto 1999) which is a basic but effective term weighing scheme in information retrieval. An index built with Lucene shows great improvement of query time over a more orthodox “relational” query. Also a noticable feature of using an information retrieval based approach is that all the results returned by the query are ranked by their relevance. The scores in figure 2 are factors of relevance which are measured by Lucene.



```
<terminated> SearchFiles [Java Application] C:\Program Files\Java\jre1.6.0_07\bin\javaw
Enter query:
blood pressure
Time: 360ms
13428 total matching documents
1. Segmental pressure (blood pressure)---->5.6431913
2. Blood pressure---->5.613595
3. Blood pressure---->5.613595
4. Labile blood pressure (finding)---->4.490876
5. Abnormal blood pressure (finding)---->4.490876
6. IBP - Invasive blood pressure---->4.490876
7. Invasive mean blood pressure---->4.490876
8. Cuff blood pressure---->4.490876
9. Stable blood pressure (finding)---->4.490876
10. Ambulatory blood pressure recording---->4.490876
```

Figure 2 – The same query was issued with argument *–repeat 100* to run 100 times

Furthermore, the authors extended the tool to also process ‘raw’ archetypes which are defined by clinical experts and produce the corresponding terms. (We have termed these lists of terms terminological shadows (Yu 2010)). By using a similar technique, a prototype which parses archetypes and generates suggestions for archetype term bindings was developed. The recommended SNOMED-CT codes are automatically created and stored in an XML mapping file. The core of this system allows the clinical experts to design sharable EHR artefacts, known as Archetypes. The system enables automatic searching and suggestions for SNOMED-CT code bindings, which occur without human intervention. This promotes quality assurance when developing archetypes by embedding codes in an archetype before its release. Tools like this have have potential uses in a wide variety of clinical systems which could benefit from embedded codes. The authors have provided a live demo of this process which is available on the EHR*land* project website (EHRland 2010).

Future work

As reported earlier, the algorithm in this system is replaceable. Thus a configurable algorithm can be applied for term binding suggesting adaption to a particular clinical scenario. Future plans for development of the system include possible plug-ins for the LinkEHR (Maldonado et al. 2009) archetype editor to support terminology integration at design-time. Also a terminology service tailored for EN13606 will contribute to the outputs of the EHRland project.

Discussion

There are a wide range of search assistive tools available besides Lucene. For example, powerful lexical tools designed specifically for medical text can be obtained freely from NLM (Allen C. Browne 2000). With the aid of domain specific i.e medical text processing toolkit, one likely improvement is the domain relevance in the search for appropriate concepts from SNOMED-CT. In fact, large and sophisticated programs have been written to deliver the task of mapping free text clinical notes to codes. This automatic process utlises Natural Language Processing (NLP) [ref] technology and a lot of effort was spent on researching text structure. The difference between these existing tools and the one used in this study is that an integration engine may take the underlying EHR information model into account, or in this case, archetypes. EHRs are digital entries which comply to the model designed to record clinical data. thus the process of binding codes should be altered from NLP based tools. However issues exist in such processes in relation to assessing the relevance of codes found by an automatic search with minimal human intervention.

A significant result of this study shows that the implemented SNOMED-CT search tool is both faster and more accurate than the conventional database approach. The second part of implementation, archetype integration engine shows the ease of suggesting bindings between SNOMED-CT terms and an expert-designed sharable EHR artefact. A general contribution could be to ensure that codes are embedded in EHR systems before communication happens. The same approach can also be used to map local codes to SNOMED-CT or other terminology.

Conclusion

In order to support a fully working and consistent EHR environment, a set of fundamental services need to be established. This study assessed the necessary foundation technologies for accessing terminology resources. The implementation of a SNOMED-CT based terminology integration engine is both a proof-of-concept prototype of such a service and a step further to merge EHR health information models with medical terminology. An extension of this work could contribute to many clincial scenarios which require terminology binding or code embedding. The process of merging EHR information models and terminology is also promoted by international standardisation organisations.

References:

Allen C. Browne, A. T. M., Suresh Srinivasan (2000) The SPECIALIST Lexicon *NLM, Bethesda, MD,*

Apelon, Inc. (2009a) Apelon Distributed Terminology System (DTS) DTS Quick Start Guide.

Apelon, Inc., Mayo Clinic/Foundation (2009b) HL7 Common Terminology Services 2 Service Functional Model (SFM).

Beale, T. (2002) Archetypes: Constraint-based domain models for future-proof information systems.

EHRland (2010) A technological assessment of the five-part EN13606 standard. Dublin.

Eichelberg, M. a. A., Thomas and Riesmeier, Jorg and Dogac, Asuman and Laleci, Gokce B. (2005) A survey and analysis of Electronic Healthcare Record standards. *ACM Computing Surveys,* Vol. 37, No. 4**,** pp. 277--315.

Gospodnetic, O.and Hatcher, E. (2005) *Lucene in action: a guide to the Java search engine*.

Maldonado, J. A., et al. (2009) LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics. *International Journal of Medical Informatics,* Vol. 78, No. 8**,** pp. 559-570.

Markwell, D., et al. (2008) Representing clinical information using SNOMED Clinical Terms with different structural information models. *KR-MED 2008.*

Ocean (2008) Ocean Informatics Product Catalog. 2008 ed., Ocean Informatics.

Regenstrief (2009) RELMA Users' Manual Version 4.2.

Ruch, P., et al. (2008) Automatic medical encoding with SNOMED categories. *BMC Med Inform Decis Mak,* Vol. 8 Suppl 1, No.**,** pp. S6.

Yates, R.and Neto, B. (1999) Modern information retrieval. *ACM P*.

Yu, S. B., Damon; and Bisbal, Jesús (2010) An Investigation of Semantic Links to Archetypes in an External Clinical Terminology through the Construction of Terminological "Shadows". *IADIS.* Freiburg, Germany, IADIS.org.

SNOMED CT User Guide, IHTSDO Copenhagen, Denmark 2007 http://www.ihtsdo.org/ fileadmin/user_upload/Docs_01/Technical_Docs/snomed_ct_user_guide.pdf

Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4–5):394–403.

Jeremy Rogers, What kinds of Medical Terminologies are there? viewed 15 Oct 2010, http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology.htm