2010-01-01

An Investigation of Semantic Links to Archetypes in an External Clinical Terminology through the Construction of Terminological "Shadows"

Sheng Yu
Dublin Institute of Technology, sheng.yu@dit.ie

Damon Berry
Dublin Institute of Technology, damon.berry@dit.ie

Jesús Bisbal
Universitat Pompeu Fabra Departament de Tecnologies de la Informació, jesus.bisbal@upf.edu

AN INVESTIGATION OF SEMANTIC LINKS TO ARCHETYPES IN AN EXTERNAL CLINICAL TERMINOLOGY THROUGH THE CONSTRUCTION OF TERMINOLOGICAL "SHADOWS"

Sheng Yu *
*TeaPOT Research Group School of Electrical Engineering Systems DIT **
*Dublin, Ireland **

Damon Berry *
*TeaPOT Research Group School of Electrical Engineering Systems DIT**
*Dublin, Ireland **

Jesús Bisbal *
*Universitat Pompeu Fabra Departament de Tecnologies de la Informació**
*Barcelona, Spain **

ABSTRACT

The two-level model based specifications for electronic health record communication EHRcom (ISO 13606) and openEHR both support the embedding of terminological references in Archetypes. This terminological binding can be created manually by a health terminology expert during Archetype design, and the binding is assessed during Archetype evaluation. There has also been some recent work on using lexical queries to generate term sets to represent concepts in Archetypes. This work created an information construct which we call a *Terminological Shadow* that links Archetype nodes to sets of candidate concepts from a terminology system. The coding scheme used for this work is SNOMED-CT. The proposed Shadows can be used to facilitate the mapping between an Archetype information model and terminological systems. A framework, which also acts as an analysis tool, has been created to construct Shadows from Archetypes. The work also demonstrates how the framework can be used to evaluate different searching algorithms by comparing the search results to the existing bound SNOMED codes.

KEYWORDS

EHR, Archetypes, SNOMED-CT, Term binding, Semantic Interoperability

1. INTRODUCTION

1.1 Background

The progress towards semantic interoperability between health information systems promises 'common understanding' between automated or semi-automated systems which are sending and receiving health information. Towards this end, terminology experts design code sets to allow clinical users to code health information into commonly understood terms. Meanwhile the significant efforts of health information modelers have produced a rich selection of health information models for recording health information in a sharable way. If the information models and terminology can be integrated, the health informatics community will be a step closer to semantic interoperability (Markwell et al. 2008, MacIsaac et al. 2008).
Standardization of mechanisms for exchanging Electronic Health Record information between health care providers is under way and notably the use of a *two-level model* approach is gaining popularity in research and practical use.(Chen et al. 2009) The first level of the two-level model consists of a *Reference Model* that

deals with the abstract foundational building blocks of health information. The second-level is a more concrete and problem-specific metadata model which consists of domain concept descriptions called *Archetypes* (Kalra 2006, Beale 2003). Archetypes are designed by domain experts to model the health information that can be recorded or communicated. However, there is little specific guidance available for developing Archetypes in a unified style. In order to arrive at the best practice approach for sharing and reusing health information, further experience of Archetype modeling needs to be gained. In order to make Archetypes meaningful and easy to search and use, standard terminology can be used to facilitate semantic interoperability in Archetype enabled EHR systems. There is a multitude of reported research in the literature on the topic of searching for terminological concepts to encode medical text. In principle, an Archetype can represent the form that a medical document or part of the document will assume. However the cost of labor to discover the most appropriate code or codes to represent a piece of clinical information being recorded is significant (Qamar et al. 2007).
This paper proposes a structure which contains a set of concepts from a terminology that are considered to be semantically equivalent to the information in Archetypes. It consists of a tree of terminological concepts that are derived from the Archetype node tree. Each node from an Archetype is associated with one or more equivalent concepts from a clinical terminology system. The resulting structure is what we have termed a *Terminological Shadow*. This work builds a framework to process Archetypes and create Shadows based on automatic search algorithms which are reconfigurable for terminological concepts. It also verifies the auto-generated Shadows by comparing the codes within the Shadows to manually selected codes.

1.2 Motivation

The motivation of this work is to leverage terminology resources in Electronic Health Records to enhance the interoperability of EHR communication. In particular, this work describes a framework for creating Shadows and also for evaluating the search results by matching them to original bound codes in Archetypes.
The objectives of this work include:
1. To design a framework to search potential SNOMED concepts which are semantically equivalent to the concepts represented by Archetype nodes.
2. To store the terminologically relevant information associated each Archetype node and the resulting SNOMED concepts from the search to construct a terminological Shadow.
3. To use shadows to perform analysis of the search operations by comparing search results to existing SNOMED binding codes.
The remainder of this paper is organized as follows. The *Related Work* section introduces the problem of semantic interoperability in the health information domain. In this paper, this problem is described as how to map terminological concepts to Archetype nodes. The *Method* section describes the methodology to define a framework to create and test the Shadow and shows the main components of the framework. The *Experiment*s section shows and discusses the results of the experiments to test the framework. *Conclusion and Future Plans* are described at the last section.

2. RELATED WORK

It has been noted above, that information models and terminological models have developed in parallel. Both clinical modelers and terminologists try to cover the dense space of health information and find a way to link and aggregate health-related concepts. The resulting overlap presents a barrier to the integration of terminology and information models.(Markwell et al. 2008) The difficulties associated with integration of the two approaches have led to research in health informatics towards searching and binding terminological concepts to reduce the ambiguity in EHRs.
A semi-automatic system called MoST that searches for SNOMED codes and binds the most appropriate codes to Archetypes, has been developed in Manchester University (Sundvall et al. 2008). The MoST process involves gathering the text from an Archetype, performing related searches on a number of medical text databases. It also filters the result using natural language processing and medical text processing. The results are refined using a number of layers and rules, and finally a minimized number of matches are presented on a graphical user interface for manual selection. The Ocean Informatics Terminology Service (Ocean 2008)

consists of a terminology server and a stand-alone desktop GUI which can be used to build terminology subsets. It allows terminologists to find and build term sets which can then be integrated into EHR applications. MetaMap or SNOCat are other searching tools for auto-recognizing and mapping free text to terminological resources (Aronson 2001, Ruch et al. 2008). RELMA, (Regenstrief 2009) is a terminology searching and linking tool provided by the Logical Observation Identifiers Names and Codes (LOINC) organization to help map a local source code to the LOINC code.
Criticism of existing term mapping or binding tools focuses on the poor searching options and inadequate ranking of relevant results.(Rogers and Bodenreider 2008) Although the MoST system addressed some of these problems, the demand for an optimal way to find the most appropriate code for the intended medical meaning is still high. The problem of finding a perfect match for the search query for a SNOMED code is not straight forward.(Rector and Brandt 2008) Automation of this process is even more difficult and with little guarantee of accuracy. Customized search of terminological codes for Archetypes is needed due to the diversity of Archetype design. In order to provide better search algorithms, a test framework is needed to automate the creation of terminological bindings or links, which we present as Shadow, and test the accuracy by matching the results to manual selections.

3. METHOD

A Terminological Shadow represents potential links between an Archetype and a terminology system. It is a structure to hold semantic information about the clinical meanings in Archetypes. In the approach presented here, the Shadow contains a set of candidate SNOMED codes returned from a search query. It also contains meta-data extracted from the Archetype such as path of the node in the Archetype and the name of the reference model class concept upon which the Archetype is based.
Figure 1 gives some idea of the relationship between an Archetype and a Shadow. A small number of terms in the terminology (the black dots) will be more or less semantically equivalent to the nodes in an Archetype. In this conceptual diagram the node that represents a clinical event of measuring *blood pressure* is considered to be equivalent to the SNOMED term: *blood pressure observable entity*. Links like this, when created either automatically or manually, lead to groups of terms in a terminology to form Terminological Shadows.

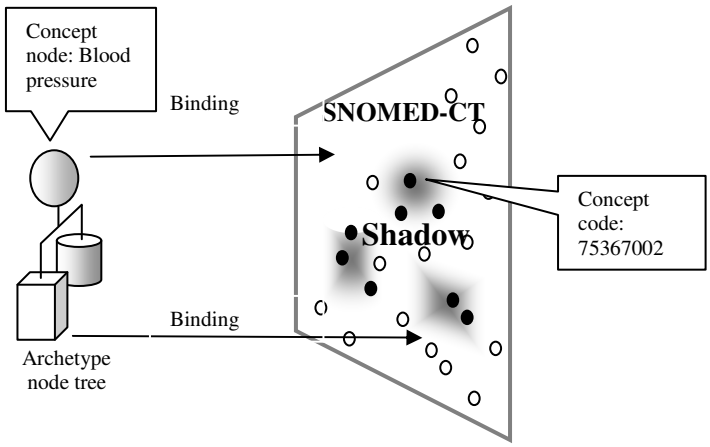

Figure 1: Archetype Shadows – projecting Archetypes into terminological systems

At the base of the framework is an algorithm to search SNOMED codes using text attributes from Archetype nodes. The 2008 release of SNOMED-CT has been used for this work. An open source full text search engine called Lucene from Apache (Gospodnetic and Hatcher 2005) is used to index the textual description entries from the SNOMED database. Over 700,000 terms in the SNOMED-CT *description* table have been indexed using Lucene and the resulting *term index* has been used for full text search. Archetypes from the NHS Connecting for Health project (NHS 2010) were selected and an ADL parser from the openEHR java reference implementation (openEHR 2007) was used to extract the relevant information from these Archetypes. To evaluate the algorithm, the set of suggested terms returned by the search were compared

against the existing binding SNOMED codes. The *recall* and *precision* (Salton and McGill 1986) of each search query was calculated for the Archetype nodes that have existing SNOMED binding codes.
Figure 2 shows the process of Shadow construction from Archetypes. Archetypes are expressed in a dedicated Archetype Definition Language (ADL) which includes support for binding of terms in the Archetype to external terminologies. This bound code will be called a *binding SNOMED code* from this point on. The ADL files, from which the shadows are to be extracted, are retrieved and parsed by the ADL parser in the following way. The framework extracts general information about the Archetype from the Archetype header of the ADL file. Next, it searches each node in the Archetype for terminologically relevant information, such as descriptions and names of nodes and term bindings. It then stores this relevant information from the shadow's object tree to be used as parameters for SNOMED search algorithms.
The framework then iterates through the nodes in the shadow. For each node, it extracts selected node information and issues a query to search the term index for SNOMED codes using a user selected search algorithm. After searching, the returned terms are added to the associated nodes alongside pre-recorded Archetype information in the shadow. The framework includes a persistence layer which can store Shadows in an RDBMS or as an XML file.

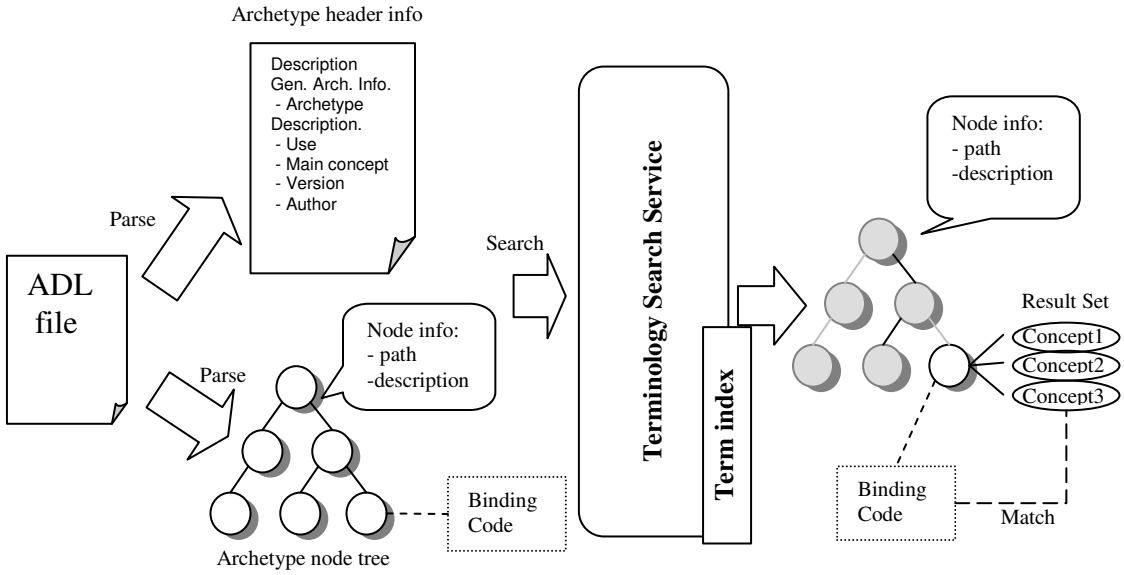

Figure 2: General process of constructing a Shadow from Archetypes

4. EXPERIMENTS

4.1 Implementation

The implementation of the framework consists of *ArchetypeCrawler*, *TermIndexer*, *TermSearcher* and *ShadowCreator* components, which collaborate in the Shadow creation process. The *ArchetypeCrawler* component parses the ADL of a set of Archetypes and gathers textual attributes. Lucene was used to implement the *TermIndexer* and *TermSearcher* components to provide a default search algorithm that takes advantage of reversed indexing and term frequency–inverse document frequency (*TF-IDF*) result ranking (Salton and McGill 1986). Each SNOMED term is regarded as a document. The *TermIndexer*, indexes each SNOMED term with the associated concept ID from the terminology system to create the term index. The *TermSearcher* component takes the name attribute of a node as a parameter for a query and automatically searches against the term index. It gathers results that are ranked according to the TF-IDF weighting scheme provided by Lucene, which is related proportionally to the frequency of occurrence of a word in a document

and is inversely related to the frequency of occurrence of the word in the corpus (Gospodnetic and Hatcher 2005). The *ArchetypeCrawler* and *TermSearcher* components feed Archetype information and search results to the *ShadowCreator* which generates Shadows.

In experiments conducted by the authors, the data set is comprised of seven Archetypes from the NHS Connecting for Health Archetype repositories. Their names are listed in table 1 column 1. These Archetypes were chosen because the ratio between bound and unbound nodes is relatively large compared to other Archetypes. The choice of these Archetypes was random in terms of the clinical content.

The implementation employs a straightforward algorithm for searching SNOMED concepts and a threshold filter is used to gather the top 10 ranked SNOMED terms. A matching procedure is carried out on the shadow to compare the codes returned by the algorithm to the existing manually assigned codes in the Archetype. The rationale for this method is to check whether a shadow contains the choice of codes selected by the expert who assigned the codes for the corresponding Archetype. This assumes that this manual assignment of codes is correct. A score of matches can be generated and it will vary for each searching and filtering algorithm.

In order to assess this approach, the Shadows of the selected Archetypes were constructed by the framework using the technique described in section 3. The resulting shadows were analyzed to produce term-matching statistics.

4.2 Results

The results of the experiment are shown in table 1 below. The second column records the "Total number of nodes in Archetype". The third "Number of existing SNOMED binding codes" column indicates how many nodes in each featured Archetype have manually assigned SNOMED binding codes. The information of these columns is gathered by parsing the given Archetype.

To generate the results in the fourth column *"Number of perfect matches found in the Shadow"* the framework iterates through all nodes in the Archetype. As it iterates, it passes node information in the form of name attributes for each node to the algorithm which searches SNOMED-CT for the resulting top 10 codes. That is, each textual name attribute is sent to the terminology search service on the SNOMED term index and the top 10 results are returned. This determines the number of resulting SNOMED codes, which amounts to 10 times the number of nodes in the shadow. The framework compares the manually assigned code to the members of the returned set of SNOMED codes. If one binding code is also found in the result set returned for that node the framework counts one perfect match. In the first row, the openEHR-EHR-CLUSTER.symptom.v1 Archetype, this number means that 13 existing binding codes are found in the Shadow results.

In the fifth column *"Number of nodes also hit parent or child"* the number is computed by checking whether any SNOMED codes returned by the framework happen to be the binding SNOMED code's parent code or child codes. This column provides a measure of how many nodes also 'hit' a parent or child code of a binding SNOMED code other than the existing bound ones. The sixth column *"Number of nodes only hit parent or child"* shows how many nodes did not hit the binding SNOMED codes but hit only the parent or child codes of the binding ones. The seventh column *"Number of nodes returned no match"* shows the number of nodes that its result set returned did not hit anything, thus it is judged as a failure as defined by the current search algorithm. The last two columns compute the average recall and precision of retrievals in a whole Archetype. These concepts are used in information retrieval to evaluate the quality of a single retrieval based on a query. Because the number of relevant documents is one in our case which is the binding SNOMED code, the calculation of recall will make it 1.0 and 0.1 for precision according to the following equation 1 and 2 (Salton and McGill 1986).

$$precision = \frac{RelevantDocuments \bigcap RetrievedDocuments}{RetrievedDocuments} \qquad \text{Equation 1}$$

$$recall = \frac{RelevantDocuments \bigcap RetrievedDocuments}{RelevantDocuments} \qquad \text{Equation 2}$$

An arithmetic mean is calculated to show the recall and precision of retrievals for one Archetype level for all the nodes with bindings. *n* equals the total number of bindings in equation 3 and 4:

$$\text{Average Recall} = \frac{1}{n}\sum_{i=1}^{n} \text{recall } i$$

Equation 3

$$\text{Average Precision} = \frac{1}{n}\sum_{i=1}^{n} \text{precision } i$$

Equation 4

Table 1: Results of the prototype framework tests

Archetype name	Total number of nodes in Archetype	Number of existing SNOMED binding codes	Number of perfect matches found in the Shadow	Number of nodes queries which also hit parent or child	Number of node queries only hit parent or child	Number of nodes returned no match	Average Archetype recall of SNOMED retrieval at 10 terms	Average Archetype precision of SNOMED retrieval at 10 terms
openEHR-EHR-CLUSTER.symptom.v1.adl	58	**21**	**13**	6	1	7	0.619	0.0619
openEHR-EHR-OBSERVATION.blood_pressure.v2.adl	47	**28**	**21**	2	0	7	0.75	0.075
openEHR-EHR-EVALUATION.activities_of_daily_living.v2.adl	68	**28**	**6**	2	0	3	0.214	0.0214
openEHR-EHR-CLUSTER.checklist_item-learning_disability_referral.v1	24	**15***	**14**	0	0	1	0.93	0.093
openEHR-EHR-CLUSTER.body_site.v2	13	**6**	**6**	0	0	0	1	0.1
openEHR-EHR-EVALUATION.waterlow_pressure_ulcer_prevention_score.v1.adl	81	**32**	**10**	5	9	13	0.312	0.0312
openEHR-EHR-OBSERVATION.hearing.v1.adl	33	**17**	**11**	3	0	6	0.647	0.0647

*(17 total, including 2 SNOMED codes which were part of a local extension)

5. DISCUSSION

The following sections discuss some of the problems with the technique and some features that were identified through analysis of the search results which were captured by the framework. It shows the benefit of utilizing the framework as a tool to help evaluate the quality of retrieval and identify weaknesses of search algorithm by analyzing the results.

5.1 Analysis of the experimental results

To support and validate the framework, it was necessary to provide an acceptable search algorithm. A key objective of the experimental work was therefore to compare the returned result and the manually selected SNOMED binding codes. The quantitative evaluation of the search results reflects the quality of the automated SNOMED code searching algorithm that was used in this work. Table 1 shows that the Average Archetype recall at 10 terms retrieved varied widely from 21% to 100%. This variation is because the terms used in some queries produced low ranked query results which caused their exclusion from the top 10. Further analysis needs to be performed to investigate this effect in general, but the discussion below shows an illustrative example.

Where the algorithm failed to retrieve the exact SNOMED binding codes, it sometimes retrieved the parent or child codes of the binding code. This suggests that certain parent or child codes could be considered as alternative binding candidates for this node. Also through inspection of failed retrievals, it is worth noting

that the leaf nodes which contain the constraint of coded item are likely to be a qualifier of the corresponding parent node. In this case, the qualifier name will sometimes be relatively generic in order to be human readable, while its binding SNOMED code will be specific. This may lead to misinterpretation by an algorithm. An example of a failed retrieval that occurred in openEHR-EHR-OBSERVATION.hearing.v1.adl nodes is illustrated in the ADL fragment shown below:

```
ELEMENT[at0018] occurrences matches {0..1} matches {  -- Rinne Test
    value matches {
        DV_CODED_TEXT matches {
        defining code matches {
                [local::
                at0019,  -- Negative
                at0020]  -- Positive
                }
```

After parsing the above ADL and searching for codes, the framework successfully retrieved the SNOMED code corresponding to node [at0018]. However, it did not find correct code for nodes [at0019] and [at0020]. The context in the example implies they are *Rinne's* Test *negative* and *positive*. But the text used to search, 'Negative; Positive', is insufficient to retrieve this intended link between the concepts. Instead more general SNOMED concepts expressing negative and positive were returned by the search. From observation of the results it appears that other failures may be due to the lack of a filtering process. The intended code is often outside the set of top 10 returned results but they could have been members of that set if irrelevant results are removed.

5.2 Analysis of the search algorithm

The implemented prototype uses the Lucene indexing algorithm to index the SNOMED term text field. One shortcoming of the approach presented here is that there is no filtering on the result. The addition of filtering would present a shorter and more accurate list of response terms to each query. Examples of SNOMED features that could be used to develop filters include synonyms, length of term and preferred categories.
Another limitation of this prototype is that using Lucene and TF-IDF alone as medical text searching tool may lead to incorrect term suggestions. A typical SNOMED term is usually too short for indexing as a text document i.e. the number of individual words in SNOMED terms is small. However, it is worth emphasizing that the framework and not the algorithm is the main focus of this work. The framework intends to provide a testing platform which facilitates multiple algorithms and multiple shadow results rather than a single optimized and mature searching algorithm. Also an effective evaluation framework is required to check the accuracy of the returned results where no existing SNOMED codes can be used as the benchmark.

6. CONCLUSIONS AND FUTURE WORK

Terminological shadows can be used to represent possible correspondences between information modeling artifacts such as Archetypes and clinical terms. In the authors' view, these correspondences can facilitate better agreement between detailed clinical models and clinical terminology. The authors have described the implementation and initial experiences with a framework which creates terminological shadows, using information from Archetypes. This study proposed, implemented and tested a framework to create Terminological Shadows of Archetypes. The implemented framework successfully demonstrated the ability to use information retrieval measurement techniques to test the effectiveness of terminological search algorithms. It has therefore been shown that the framework can be used to evaluate searching tools for terminologies such as SNOMED. Better searching and filtering algorithms can be inserted to produce improved result sets. The framework requires further work in relation to the classification of nodes in order to differentiate nodes whose purpose is not for recording clinical information, for example to act as information model compositional meta-data such as 'Items'. Further work is also needed to make the framework utilize the ontology aspects of the terminological system,

1. Reference-Model-aware and domain-specific terminological filtering is important to create useful Shadows. In the work reported here, the domain-specific filtering corresponds to concept-awareness.

2. An extended test base is needed to provide further verification, so more Archetype repositories and different types of Archetypes are needed.

Planned future extensions of the work include the following:

1. Implementation of a Reference-Model-aware function is planned which includes two major EHR standards: The EHRcom reference model and the openEHR reference model. The algorithm is expected to expand the query by employing reference model based context information such as combination of node information from target nodes and their parent nodes. It is intended that the algorithm will also take account of the reference model class type of each node (if it is a data point).
2. SNOMED-concept-aware feature: In future, more content is planned to be indexed from SNOMED and to facilitate this, hybrid index-relational database querying is proposed so that more efficient and complex searching can be achieved
3. A filtering module needs to be completed by specifying a set of common filtering rules that are extensible.
4. The authors envisage other uses of a Shadow which were not implemented in this work. Potentially SNOMED codes and reference model information in the shadow can be used to match and find similar Archetypes written in different styles and based on different reference models.

ACKNOWLEDGEMENT

This work is part of the EHRland project and was jointly funded by TSR Strand 1 Funding of the Council of Directors of the Institutes of Technology, Ireland and by the Health Information Quality Authority of Ireland.

REFERENCES

International Organization for Standardization, 2008, TC215 *Health informatics -- Electronic health record communication -- Archetype interchange specification*, ISO 13606-2:2008.

Aronson, A. R., 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of AMIA Symposium 2001,* pp. 17-21.

Beale, T., 2003. Archetypes and the EHR. *Studies in Health Technology and Informatics,* Vol. 96, pp. 238-44.

Chen, R., et al., 2009. Archetype-based conversion of EHR content models: pilot experience with a regional EHR system. *BMC Medical Informatics Decision Making,* Vol. 9, pp. 33.

Gospodnetic, O.and Hatcher, E., 2005. *Lucene in action: a guide to the Java search engine*. Manning, Greenwich, CT,USA.

Kalra, D., 2006. Electronic health record standards. *IMIA Yearbook of Medical Informatics,* Vol. 2006, pp. 136-144.

MacIsaac, P., et al., 2008. Essential SNOMED: Simplifying SNOMED CT and Supporting Integration with Health Information Models. *Proceedings of KR-MED 2008* Phoenix, Arizona, USA.

Markwell, D., et al., 2008. Representing clinical information using SNOMED Clinical Terms with different structural information models. *Proceedings of KR-MED 2008* Phoenix, Arizona, USA.

NHS, 2010. *Archetypes available from NHS Connecting for Health*, England, viewed 10 February 2010, <https://svn.connectingforhealth.nhs.uk/svn/public/nhscontentmodels/BRANCHES/Lorenzo_3.5/ContentRelease-3.0/cm/Archetypes/gen/html/index_en.html>.

Ocean Informatics, 2008. *Ocean Informatics Product Catalog*. 2008 ed., Ocean Informatics.

openEHR, 2007. *The openEHR Java Reference Implementation Project*, Australia, viewed 10 February 2010, < http://www.openehr.org/projects/java.html>.

Qamar, R., et al., 2007. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. *Proceedings of AMIA Annual Symposium,* pp. 608-13.

Rector, A. L.and Brandt, S., 2008. Why do it the hard way? The case for an expressive description logic for SNOMED. *Journal of American Medical Informatics Association,* Vol. 15, No. 6**,** pp. 744-51.

Regenstrief, 2009. RELMA Users' Manual Version 4.2.

Rogers, J.and Bodenreider, O., 2008. SNOMED CT: Browsing the browsers. *Proceedings of KR-MED 2008* Phoenix, Arizona, USA.

Ruch, P., et al., 2008. Automatic medical encoding with SNOMED categories. *BMC Medical Informatics Decision Making,* Vol. 8 Suppl 1, pp. S6.

Salton, G.and McGill, M. J., 1986. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY, USA.

Sundvall, E., et al., 2008. Integration of tools for binding Archetypes to SNOMED CT. *BMC Medical Informatics Decision Making,* Vol. 8 Suppl 1, pp. S7.