Integration of Tools for Binding Archetypes to SNOMED CT

Erik Sundvall[1], Rahil Qamar[2], Mikael Nyström[1], Mattias Forss[1],
Håkan Petersson[1], Hans Åhlfeldt[1], Alan Rector[2]
[1]Department of Biomedical Engineering Linköpings universitet, Sweden
[2]Department of Computer Science, University of Manchester, UK

Abstract: *The Archetype formalism and the associated Archetype Definition Language have been proposed as standard for specifying models of components of Electronic Healthcare Records as a means of achieving interoperability between clinical systems. This paper presents an archetype editor with support for manual or semi-automatic creation of bindings between archetypes and terminology systems. Lexical and semantic methods are applied in order to obtain automatic mapping suggestions. Information visualisation methods are also used to assist the user in exploration and selection of mappings.*

The methods and tools presented are general, but here only bindings between SNOMED CT and archetypes based on the openEHR reference model are presented in detail.

INTRODUCTION

This paper describes the integration of three applications related to archetypes and terminology systems: a) an editor for archetypes b) MOST; a system for selecting terms from SNOMED CT to be bound to archetypes and c) TermViz; a tool for visualizing and navigating terminology systems.

The 'archetype' approach to information modelling is introduced below and is followed by descriptions of the three applications and their integration.

Modelling in openEHR

The openEHR [a] foundation aims to facilitate interoperable implementations of electronic health record systems (EHRs) by developing and promoting open specifications and specifications-based implementations. The intention behind the specifications is to enable interoperability while still being flexible regarding choice of terminology systems, implementation technology, and human language translations.

The architecture of openEHR aims to scale from small desktop systems for general practitioners to distributed patient centred lifelong shared care health record systems. [1]

The openEHR architecture [1] includes a design principle called 'Ontological[b] separation', which regulates the EHR modelling (see Figure 1). The structure is divided into two main categories entitled 'ontologies[b] of information' and 'ontologies[b] of reality'.

The 'ontologies of information' contain the information models of the content in the EHR whereas the 'ontologies of reality' describe real phenomena with descriptions and classifications. The 'ontologies of information' are then divided into:

- 'Domain content models' containing formal definitions of the clinical content. They can be developed using archetypes which are designed to be easy to change when new clinical needs arise.
- 'Information representation models' which are implemented in the electronic health care systems software. They are used as a foundation for the domain content models and are designed to be stable with regards to model changes.

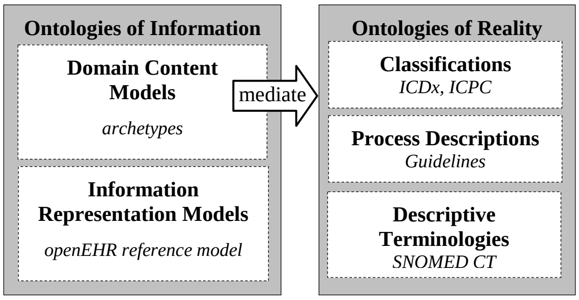


Figure 1: openEHR's ontological[b] structure. Adapted from [1]

Historically, these two parts have not been separated in clinical software. This lack of division made the EHRs harder to adapt to new clinical needs. [1]

The 'ontologies of reality' contain:

- 'classifications', like ICDx and ICPC,
- 'process descriptions', like guidelines
- 'descriptive terminologies', like SNOMED CT.

The domain content models (e.g. archetypes) can have mappings to any or all ontologies of reality. EHR extracts based on common shared archetypes are proposed as a means to exchange information between different health care providers [1]

SNOMED CT

SNOMED CT [c] is the terminology system discussed in this paper. It is a medical terminology based on concept representations that are related to

[a] http://www.openehr.org/

[b] The words 'Ontological' and 'ontologies' come from the source [1], but in this case 'models' could be equivalent.

[c] Systematized Nomenclature of Medicine – Clinical Terms

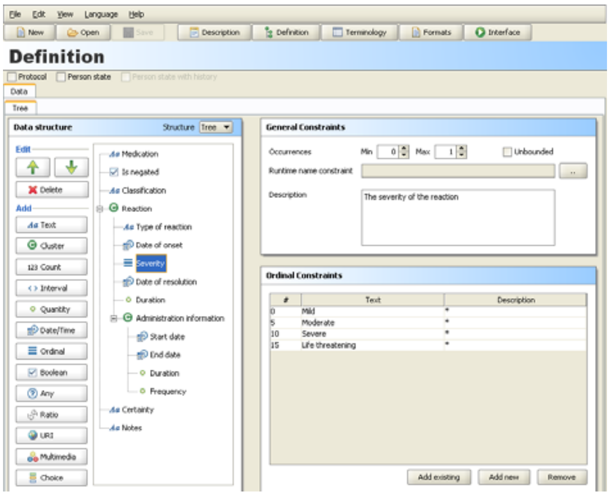
each other by different types of relationships like 'IS-A' (subtype), 'Part of' and 'Causative agent'. Each concept is associated with two or more textual descriptions [2]. The number of active core concepts in the U.K. May 2006 release are 300 000. [3]

TOOLS AND METHODS

The applications for archetype editing, semi-automatic terminology binding and terminology visualization that have been integrated are briefly described in this section.

The Archetype Editor

Authoring of archetypes is not intended to be part of the daily routine of clinicians. Instead, the goal is to develop archetypes that can be used in many different situations over a long period of time and to use them as parts of templates for clinical data entry.



Occurrences Min 0 Max 1 □ Unbounded

Runtime name constraint

Description: The severity of the reaction

#	Text	Description
0	Mild	*
5	Moderate	*
10	Severe	*
15	Life threatening	*

Figure 2: Definition view showing an archetype for adverse reactions to medications.

The purpose of the archetype editor is to let users build archetypes in an intuitive graphical environment without prior knowledge of formal representations of archetypes like the 'Archetype Definition Language' (ADL) or XML. An archetype editor that allows the user to create new archetypes and learn from previously created ones by viewing and exploration is crucial for developing good quality archetypes.

At the time of starting the development of the Java based archetype editor at Linköping University [4], there already existed an archetype editor in the openEHR community. Our new editor focused on improving terminology binding support and usability. It also removed operating system dependencies. Connections to external terminology sources like SNOMED CT and UMLS were included so that the effort required to bind terms with the help of external terminology sources compared to manually looking up codes was reduced.

The MoST System

In order to bind terms in clinical data models to terms in external terminologies we must first find appropriate matches. The Model Standardisation using Terminology (MoST) system [5] developed at the University of Manchester is a general semi-automated mapping process providing the clinical modeller with candidate mappings. The mapping manually determined to be the most suitable can then be bound to a content model entity.

The specific clinical data models selected to demonstrate the applicability of the methodology in this paper are archetypes according to the openEHR archetype model. The terminology to which they have been mapped to is SNOMED CT.

In the MoST mapping process as shown in Figure 3, archetypes are converted from ADL format to a general XML format designed to represent hierarchical data models. The clinical content of the model is then passed to the actual mapping process which executes various lexical and semantic procedures by referring to existing medical resources and SNOMED CT.

The first round of mapping includes a lexical processing of terms using the Emergency Medical Text Processing (EMT-P) service. It is a natural language processing (NLP) tool which cleans up raw text entries [6]. EMT-P then looks for matches in the Unified Medical Language System (UMLS) resources and the UMLS LVG database which consists of normalised word forms. UMLS is a large medical resource that enables the mapping of terms to controlled vocabularies in healthcare [7]. The MoST methodology makes use of the lexical

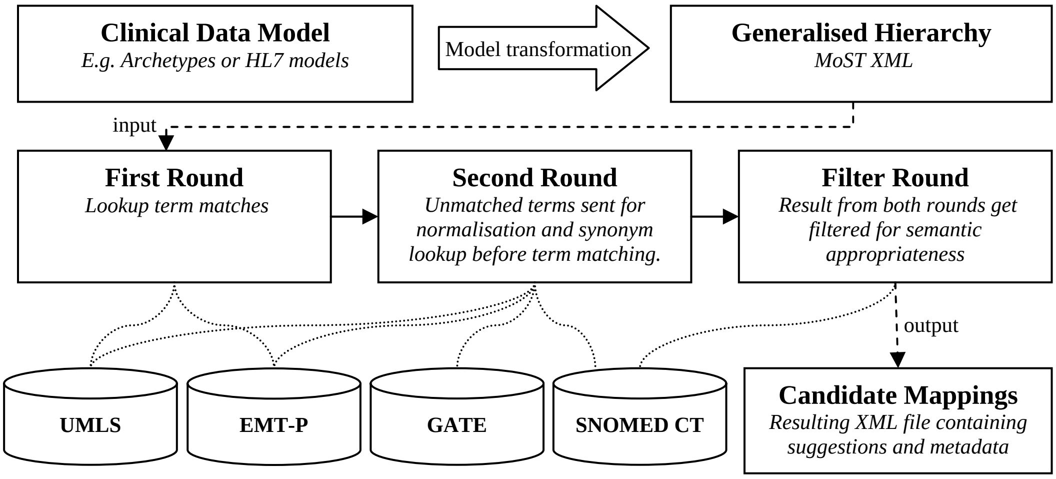

Clinical Data Model
E.g. Archetypes or HL7 models
Model transformation
Generalised Hierarchy
MoST XML
input
First Round
Lookup term matches
Second Round
Unmatched terms sent for normalisation and synonym lookup before term matching.
Filter Round
Result from both rounds get filtered for semantic appropriateness
output
UMLS
EMT-P
GATE
SNOMED CT
Candidate Mappings
Resulting XML file containing suggestions and metadata

Figure 3: MoST System Methodology

procedures of both the EMT-P tool and the UMLS resource at the same time to draw upon their individual and combined strengths to find relevant matches.

All archetype terms, irrespective of whether they have found a match in the first round, are sent to the second round for normalisation. Normalisation involves execution of a series of lexical and semantic methods and collation of results from each. Some of the methods employed include a training dataset with commonly used clinical synonyms and abbreviations and context search. An external NLP application named GATE[d] was used for stemming, based on regular expression rules developed for its Morphological Analyzer and synonym search using its WordNet[e] plugin.

At the end of both the rounds, the collated results are subjected to elimination through filtering. All filtered SNOMED CT results are presented to the clinical modeller as candidate mappings. The filtering and evaluation details are described in [5] as this is beyond the scope of this paper. Briefly, filtering comprises of two main levels. The first is exclusion of all concepts subsumed by a parent concept occurring in the result set, and inclusion of all non-occurring parent concepts if more than three child concepts are present in the result set. The second level involves inclusion of only those results whose semantic category(ies) is similar to the one specified by the clinical modeller. However, MoST provides for the possibility of a human and/or SNOMED CT categorisation error.

The candidate mappings can be viewed in simple tabular form (see Figure 6) in the editor along with the facility to further explore the relevant SNOMED CT hierarchy using the visualization technique described below.

Terminology visualization

Large terminology systems with complex intertwined structures can be hard to navigate and get acquainted with. Free-text queries are possible entries into the exploration of such systems, and the way results are presented has impact on the user's ability to grasp the overall structure of the system. Complex hierarchies like the one used in SNOMED CT, where nodes have multiple parents and several other relationship types, makes visualization challenging. A previous paper [8] presented a prototype, called *TermViz*, applying well-known methods from the fields of Information Visualization and Graph Drawing like 'focus+context' and self-organizing layouts. The user can simultaneously focus on several nodes in terminology systems and then use interactive animated graph navigation for further exploration without loosing context. 'Semantic zooming' i.e. reducing the amount of visible information, e.g. text labels far from focused nodes, is also available. This part of the tool can also be used as a stand alone SNOMED CT browser.

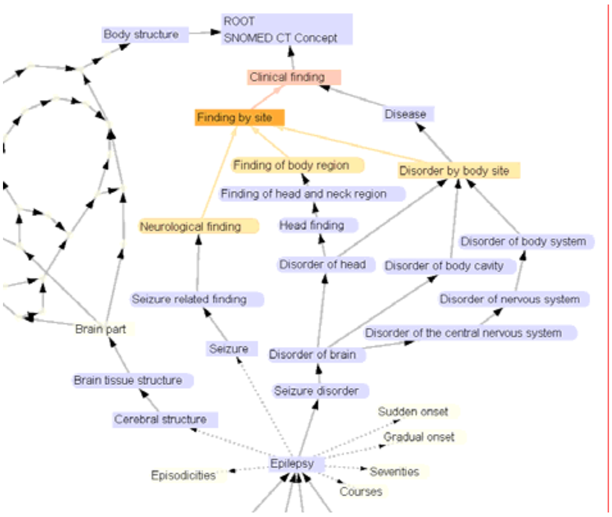
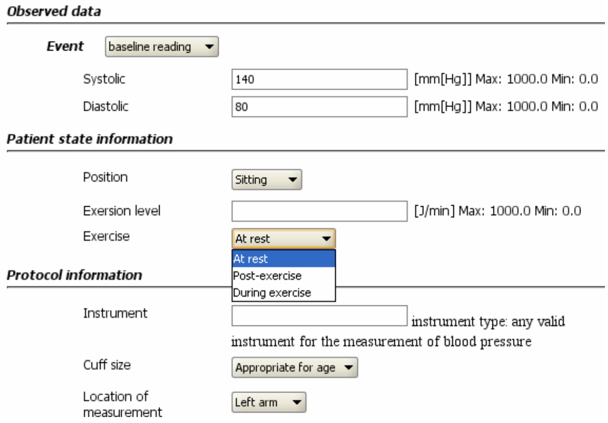


Figure 4: Visualization of a part of the SNOMED CT hierarchy. The graph can be interactively explored and expanded.

RESULTING APPLICATION

In this section the integrated application is demonstrated using the blood pressure archetype, shown in the interface view of the editor.



Observed data

Event baseline reading

Systolic 140 [mm[Hg]] Max: 1000.0 Min: 0.0

Diastolic 80 [mm[Hg]] Max: 1000.0 Min: 0.0

Patient state information

Position Sitting

Exersion level [J/min] Max: 1000.0 Min: 0.0

Exercise At rest

At rest
Post-exercise
During exercise

Protocol information

Instrument instrument type: any valid instrument for the measurement of blood pressure

Cuff size Appropriate for age

Location of measurement Left arm

Figure 5: Interface view showing a blood pressure archetype

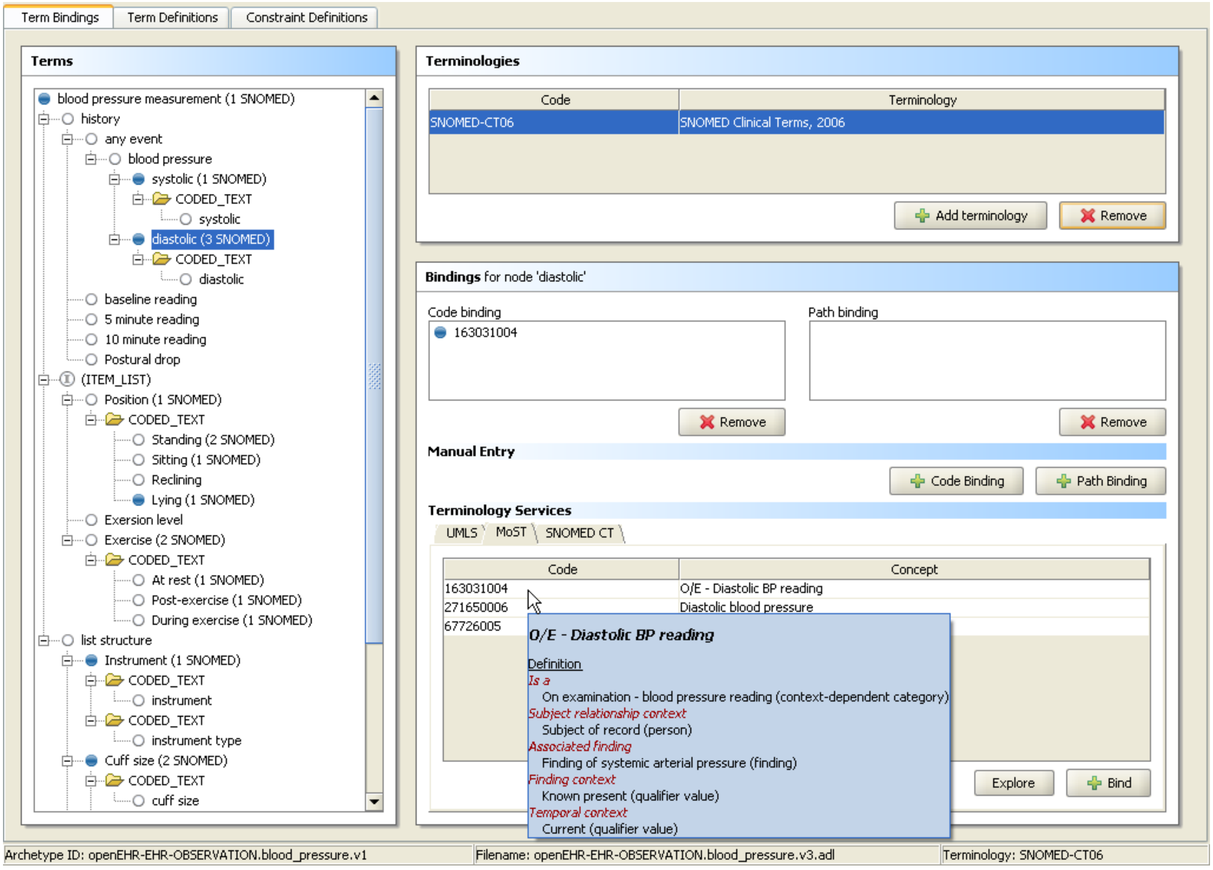
The *definition view of the editor*, see Figure 2, can be used to:

- structure and name the fields in the archetype
- mark fields as mandatory or optional
- restrict format and kind of information to be allowed in a field

In an archetype, the 'fields' described above are nodes within a tree structure. Nodes can be bound to terminologies, such as SNOMED CT, as shown in Figure 6. The archetype is sent to the remote MoST-service (accessed using a SOAP-based Web service). In the tree structure to the left are labels ending with e.g. (3 SNOMED) indicating that MoST has found three candidate mappings for the node. Upon selecting a node the suggestions are shown in the list at the bottom right of the screen.

[d] http://gate.ac.uk/

[e] http://wordnet.princeton.edu/



Code	Terminology
SNOMED-CT06	SNOMED Clinical Terms, 2006

Code binding: ● 163031004

Code	Concept
163031004	O/E - Diastolic BP reading
271650006	Diastolic blood pressure
67726005	

O/E - Diastolic BP reading

Definition

Is a

On examination - blood pressure reading (context-dependent category)

Subject relationship context

Subject of record (person)

Associated finding

Finding of systemic arterial pressure (finding)

Finding context

Known present (qualifier value)

Temporal context

Current (qualifier value)

Archetype ID: openEHR-EHR-OBSERVATION.blood_pressure.v1 | Filename: openEHR-EHR-OBSERVATION.blood_pressure.v3.adl | Terminology: SNOMED-CT06

Figure 6: Terminology binding view showing suggestions from MoST

The SNOMED CT codes can be selected and 'bound' to the archetype node. A blue dot in front of a node shows that it has been bound to one or more terms in the currently selected terminology. Holding the cursor over a candidate mapping brings up a tool tip (the blue box) showing a short definition of the term.

Free text queries for individual nodes can also be sent to UMLS or, if locally available, to a database containing SNOMED CT tables.

Results from terminology services can be explored using visualization. On clicking the Explore button (Figure 6), an interactive graph opens, as shown in Figure 4. The graph is constructed by climbing the hierarchy using the IS-A relations starting from the search results ending at the top concept. Other types of relations can also be explored by selecting any node. In addition to exploration, archetype bindings can be created from the graph view as well.

The archetype editor download, and more information can be found at http://www.imt.liu.se/mi/ehr/

DISCUSSION

Archetype based systems have only been implemented and deployed in very limited numbers yet. We believe that semantic interoperability through the archetype approach will have greater chances of success if extensive bindings to terminologies are provided. Finding the right terms to bind is a difficult task but the effort to achieve terminology bindings is reduced with the help of our methods and tools. The integrated editor eliminates the need for users to swap applications to find appropriate terminology entries. The mapping process is further assisted by the ability to get candidate mappings from MoST.

Visually relating results from the terminology services (instead of only browsing a list) assists the user in making the correct binding even if there are a large number of terms returned.

Future work

Currently, only terminology service assistance for equivalence bindings, i.e. 'this archetype node is synonymous to this SNOMED CT concept' is available in the editor. In the archetype formalism [9] this is called 'term bindings'. Archetypes also support 'constraint bindings' that would allow for more advanced bindings to terminologies using compositions of concepts and relations. The formalism for this is not well specified by openEHR as yet [f], but if it becomes expressive enough the archetype editor could:

- assist post-coordination of concepts at the time of archetype creation (e.g. the ones provided

[f] Currently the specifications only show examples like: http://terminology.org?terminology_id=SNOMED-CT;has_relation=102002;with_target=57134006 The url is intended to point to some future terminology server with a query format yet to be specified.

by MoST). From the perspective of the clinician using the archetype this could be regarded as a pre-coordination (pre-runtime).

- constrain allowed post-coordinations at runtime, like ‘allow any sub-concept of the SNOMED body position concept but not body position itself’ instead of enumerating a list like in Figure 6.

A powerful constraint binding formalism should allow inclusion and exclusion of arbitrary subsets.

The granularity and the degree of compositionality of an archetype also affect the terminology bindings and types of term-coordination possible, for example the difference in the modelling of ‘Position’ (enumerated options) and ‘Exersion Level’ (free text) in Figure 6.

The term binding problem between two independent models (here archetypes and SNOMED CT) and the logical control of post-coordination offer challenging tasks. Many coordination variations may in the end mean the same thing, e.g. a post-coordination may be equivalent to an existing pre-coordination or another post-coordination. Logical contradictions also have to be checked for and avoided. The possibility to combine SNOMED CT concepts from different categories further increases the logical complexity of the problem, e.g. combinations like an observable entity (tumour stage), a body structure (structure of thyroid) and a context-dependent category (family history of).

Caution is needed if we want to, for instance, interpret the bindings to do automated reasoning. Formal methods addressing these problems are being researched by one of the authors (Rector). We believe that automated support for formal logical control of terminology bindings and post-coordination in tools like the archetype editor and EHR systems must be added in order to handle the logical complexity described above.

Since the tools discussed in this paper have been developed on the principles of general applicability, it is expected that other terminology systems such as GALEN[g] or, FMA[h] (Foundational Model of Anatomy) can serve as a second use-case. HL7[i] V3 models are quite similar in purpose to Archetype Models and may also be investigated for demonstrating the mapping methodology.

The integrated editor will be publicly released and freely available as ‘Open Source’. Feedback and future user-based evaluation results can be used for further improvements. How well and easily archetype based clinical models can be mapped to terminology systems is beyond the scope of this paper but such future studies might be helped by this integrated tool.

ACKNOWLEDGEMENTS

This work was performed in the framework of the EU-funded Network of Excellence entitled Semantic Interoperability and Data Mining in Medicine.

REFERENCES

1. Beale T, Heard S, Kalra D, Lloyd D, editors. Architecture Overview, revision 1.0.1. London, Great Britain: The openEHR foundation; 2006 http://svn.openehr.org/specification/BRANCHES/Release-1.1-candidate/publishing/architecture/overview.pdf [cited 2006 September 7]
2. College of American Pathologists, SNOMED International. SNOMED Clinical Terms Technical Reference Guide. January 2006 Release. Northfield (Illinois, USA): College of American Pathologists; 2006.
3. College of American Pathologists, SNOMED International. SNOMED Clinical Terms United Kingdom Edition. May 2006 Release. Northfield (Illinois, USA): College of American Pathologists; 2006.
4. Forss M, Hjalmarsson J. Utveckling av en arketypeditor: Ett verktyg för modellering av struktur i elektroniska patientjournaler [Development of an archetype editor: A tool for modelling structure in electronic health records]. Master Thesis, 2006

   http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-6205
5. Qamar R, Rector A.L. MoST: A System to Semantically Map Clinical Model Data.. Accepted to the Semantic Mining Conference on SNOMED CT, 1-3 Oct 2006, Copenhagen, Denmark
6. Emergency Medical Text Processor (EMT-P). http://www.med.unc.edu/wrkunits/2depts/emergmed/EMTP/about.html, [cited January 2006].
7. UMLSKS, User’s Guide – Introduction, http://umlsks.nlm.nih.gov/, January 2006.
8. Sundvall E, Nyström M, Petersson H, Åhlfeldt H. Interactive Visualization and Navigation of Complex Terminology Systems, Exemplified by SNOMED CT. Accepted to MIE2006 27-30 Aug, Maastricht, The Netherlands
9. Beale T, Heard S. Archetype Definition Language Version 1 (ADL), Revision: 1.4 [Section 7.5]

   http://svn.openehr.org/specification/BRANCHES/Release-1.1-candidate/publishing/architecture/am/adl.pdf [cited May 2006].

[g] http://www.opengalen.org/

[h] http://fma.biostr.washington.edu/

[i] http://www.hl7.org/