SEMANTIC MAPPING OF CLINICAL MODEL DATA TO BIOMEDICAL TERMINOLOGIES TO FACILITATE INTEROPERABILITY

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2008

By
Rahil Qamar
School of Computer Science

Contents

List of Tables

List of Figures

Abstract

Work in the field of recording standard, coded data in electronic health records and messages, is important to support interoperability of clinical systems and reduce medical errors caused by misinterpretation and misrepresentation of data. Interoperability of clinical systems requires the integration of standard data models such as HL7 messages or *open*EHR Archetypes with terminologies such as SNOMED-CT.

The thesis presents and evaluates, the Model Standardisation using Terminology Systems (MoST) methodology, for integrating the clinical content in data models and terminology models. The MoST system developed for the purpose, aims to *find* semantically equivalent SNOMED terminology codes to *map* to archetype data model fragments. The two key stages of MoST include, (i) term finding, and (ii) data mapping. While the *term finding* procedure is completely automated, the *data mapping* procedure is assisted by clinical experts. The research recognises the significance of human intervention in ensuring the quality of the terminology codes being mapped to the data model fragments. Ensuring the quality of the mappings, helps maintain accuracy and unambiguity of coded data. The evaluation of the MoST system shows the importance of incorporating linguistic and semantic procedures, in addition to lexical lookups, to increase the chances of finding semantic matches.

A significant contribution of the thesis is the description of the issues with current Archetype and SNOMED models with regards to the information needed to achieve effective model integration at content level. These issues were highlighted by the qualitative analysis of the evaluation. The issues point to semantic gaps in both the data and terminology models, which inhibit automated systems, such as MoST, from making intelligent inferences on the semantic appropriateness of

the content. Suggestions for resolving these issues are detailed, where appropriate.

A final contribution of the thesis is the set of guidelines that are suggested to the two modeling (Archetype and SNOMED) communities, to improve the quality of their model content. The hypothesis is that an increase in the content compatibility of the two models will increase the likelihood of the overall integration of the models, to achieve interoperability

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

Copyright

Dedication

This thesis is dedicated to my parents and husband, without whose unconditional love and support, I would not have been able to achieve so much. Thank you for being there for me always.

Acknowledgments

Writing an acknowledgment is perhaps the most difficult part of a thesis. There are several people who have provided me with wisdom, knowledge, help, and encouragement, during this doctoral programme. Please accept my deepest gratitude, irrespective of your name appearing in the acknowledgment.

I would like to begin by thanking my supervisor, Prof. Alan Rector, for guiding me through the long Ph.D. process in a way that has sharpened my skills as a researcher. I cannot thank him enough for his untiring help and support throughout my research degree, especially during the thesis write-up. A special word of thanks to him, for lifting my spirits and giving me that much needed morale boost, when things seemed to be looking gray. I am also grateful to him for providing me with opportunities, to attend various conferences and meetings, and publish papers, to widen my knowledge and research network.

A heartfelt thank you to all the members of my Medical Informatics research group for their clinical, technical, and emotional help. The fun, smiles, and laughter, brightened up an otherwise lonely process. Also, my gratitude to all members of our research department, who have provided a wonderful environment to enhance my research abilities.

The Semantic Mining project that funded my Ph.D., also provided me with the perfect platform to present my research work and ideas. A special thanks to Dr. Dipak Kalra for having unwaivering faith in my work and capabilities. Also, my deepest gratitude to the IMT group at Linköping University (LiU) in Sweden, for agreeing to collaborate, and providing me with tremendous help. I would also like to recognise the expert advise given to me by one of the pioneers of the Archetype approach, Thomas Beale. My knowledge of Archetypes would not have been complete without his guidance.

I would also like to thank my husband, Fardeen Siddiqui, who lived every moment of my Ph.D process with me. For being there through all the ups and downs, successes and failures, and accolades and criticism. His presence in my life, and pride in me, have been my greatest strengths. My strength have also been my two sisters, Dr. Hina Qamar Siddiqi, and Dr. Shiba Qamar. They complete my life with not only their love and support, but also their constructive criticism, and professional advise. I thank my entire family and all my friends for always believing in me and encouraging me. Last but not the least, thank you to my little niece, Huma, for making even the worst of days seem the best, with her innocent smile and lively chatter.

This acknowledgment cannot be complete without thanking my parents, for giving me my biggest gift besides life - education. I wish my father would have lived to see the fruits of his labour, but I know my mother is extremely proud of me. I thank them for making me who I am, and teaching me to never give up. I am grateful to both of them, eternally.

Chapter 1

Introduction

Electronic healthcare records, or EHRs, require various standards at different levels in the health informatics domain, to interoperate between the different health information systems. Some of the important standards required are:

1. *Clinical care standards* that govern the quality of care through evidence-based medicine and clinical guidelines.
2. *Workflow standards* that not only provide support to the process of patient care but also aim to improve the process, by making it more cost-effective and efficient.
3. *Data model standards* focusing on data interoperability, and finally
4. *Terminology standards* that help in the standardisation of data by providing precise, controlled, versioned and well-documented semantics to the data. The data are usually represented as 'concepts' in the terminologies.

It is essential to get all these standards working together smoothly in order to achieve interoperable EHRs. The area of work undertaken in this research lies in the cross domain of the data model standards and the terminology standards (see points 3 and 4 above). The research attempts to align the the data model with the terminology model through *data mapping*, also known as *term binding*. The aim of performing data mapping is to standardise the data/clinical content (in the data models) used for capturing patient-related information, to one or more terminologies, to achieve semantic interoperability. A terminology serves as the standard, controlled vocabulary to which the data conforms.

1.1 The task of semantic interoperability

Semantic interoperability is 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. The issue of semantic interoperability requires that all participating data conform to some standard terminology or vocabulary in order to interpret and use it uniformly in all partaking information systems. Therefore, it becomes a requirement to first standardise the data to an approved terminology that acts as a standard. In the medical domain, standardising data has a great significance, as controlling the vocabulary used to record patient data is critical to making EHRs safe to exchange and reuse. It also improves the quality and completeness of the data, which are critical to the successful functioning of health information systems. Therefore, it is imperative that standardisation is introduced at an early stage of the data aggregation process.

Data standards have been recognised as the basic building blocks of interoperability, allowing different information systems to securely and rapidly access health information when appropriate and where needed. In addition, an interoperable information system is vital for improving healthcare, mainly because it reduces medical errors and improves efficiency [1] by enabling all professionals engaged in the process of care to access the same records irrespective of its point-of-origin. Various government efforts [51][44][8] to bring about *shared care* through integrated health records, amongst other things, has led to official recognition for the need to standardise data in EHRs [1]. In addition, increased mining of data for medical research and population health statistics is also being conducted, which requires more structured and standardised storage of data for accurate results.

In the UK, the IT programme initiated by the NHS in 2002, now known as Connecting for Health (CfH), highlighted its four key deliverables (Taken from [80]):

- Delivering a robust infrastructure that encompasses security and confidentiality.
- Delivering electronic booking or appointments

[1]Electronic Health Records

- Delivering electronic transfer of prescriptions
- Delivering an integrated clinical record system.

The fourth key deliverable i.e. the integrated clinical record system, which will be the next major milestone for EHRs in the NHS, aims to deliver in essence a single NHS EHR for each patient, integrated across primary, secondary, and social care. In order to achieve this objective it will be necessary to achieve not only system-level interoperability but also semantic-level interoperability. System interoperability is a more holistic approach, as compared to semantic interoperability. It is the ability to enable different computer systems to integrate their workflows, so as to automatically process transmitted health information. The research work detailed in the thesis directs its efforts to achieving semantic interoperability, in order to help efforts by other researchers and implementers to achieve subsequent system interoperability.

1.2 The Research Agenda

Even though the data model and terminology models co-exist, they have been developed separately. Efforts to integrate them to achieve the high level objective of semantic interoperability has proved to be a difficult task [89][90]. One of the main reasons for the integration difficulty is the lack of stable and reliable software applications to quickly and efficiently map data in the two models (i.e. data models and terminologies). Manual mapping of structured data to terminologies has not been done frequently. For instance, very few HL7 messages have been mapped to terminologies. However, where mapping has been done, it has been manual and has taken several man-months. However, it is very difficult to obtain published data revealing such issues [Communication with A.L. Rector].

While there is not much evidence available to cite successful attempts at mapping structured data to terminologies, there have been a few projects that have successfully mapped concepts between two different terminology models. For instance, cross mappings of SNOMED-CT[2] concepts to ICD[3] and OPCS[4] terminology concepts have been performed but the process has been manual. Another

[2]Systematized Nomenclature of Human and Veterinary Medicine-Clinical Terms
[3]International Classification of Diseases
[4]Office of Population Censuses and Surveys

program to map legacy interface terminologies to the SNOMED terminology was performed manually but took approximately 8 man-months to complete[110]. The long period taken was due to the gaps and the short duration of time that the participating clinicians were able to contribute to the manual mapping exercise. Such long periods of study are a major limitation to quick advancement in the field of achieving model integration at the content level. In addition to the manual mapping being time-consuming, it also often leads to incomplete mappings or mappings done with out-of-date terminology versions. The aim of the research is to reduce the time taken to perform such mapping exercises significantly, by automating a major part of the manual effort required.

In order to accomplish the model integration task, clinical content from one data model (treated as the *source*), will be mapped to the content of one terminology model (regarded as the *target*). The case studies chosen for testing the integration process are the *open*EHR Archetype models [23] and SNOMED CT [73]. The *open*EHR Archetypes are the data models providing the source content, whereas SNOMED CT is the clinical terminology providing the target for mapping. The two models are explained in detail in Chapter 2.

The research agenda discussed in this section includes the *research objective*, and *question*. It also presents the *significance of the research*, and a list of *research contributions* aimed to be made to the scientific community. Finally, the *research programme* is presented, to describe the flow of the work in the thesis.

1.2.1 Research Objective

The key objectives of the research are:

1. To enable clinical data modelers to quickly and efficiently integrate data model content with terminology concepts. Manual data mapping procedures often take several months to complete. This is because they are completely dependent on clinicians to search for the ‘correct’ matches from a corpus of thousands of terminology concepts. Such manual processes use up a clinician’s valuable time in tedious, mundane tasks. This not only causes delays in the completion of mapping exercises but also results in fewer such attempts pursued. Automating search procedures to perform

term finding and manually assisted data mapping is intended to motivate several more model integration attempts in the healthcare domain.

2. To establish the importance of considering all available context information about a specific data fragment. Taking into account context information can help automated procedures in determining the semantic equivalence of a data model fragment to a terminology concept. The automated term finding procedure can also eliminate irrelevant results based on the context information available. In effect, better quality results reduce the time spent by clinical experts in identifying the most relevant results for data mapping. Such assisted data mapping reduces human effort, which is otherwise a major disadvantage in the widespread use of controlled vocabularies in the medical domain.

3. To provide a mechanism by which data model authors can create data fragments that can easily conform to controlled terminology concepts. Being able to almost immediately view the terminology concepts that can(not) conform to the fragments of a data model, will enable the authors to make suitable modifications both structurally and content-wise to increase chances of integration.

4. To highlight issues with the content and structure of data models and terminologies. Working at the content level of both data models and terminologies helps raise awareness of fine-grained semantic detail. The issues highlight the knowledge gaps in the two models that inhibit automated systems from making intelligent inferences and decisions. The issues mainly arise from the ambiguities in naming, defining, and describing the data in the two models.

5. To make recommendations for improvement in the quality of data in the two models to facilitate easy model integration. Also to provide useful criteria for the improvement of content structure in the data models, which provides the source for the data fragments to be mapped to terminology concepts. It is essential to 'correct' the source before attempting to 'correct' the target, as the degree of success of mapping, depends largely on the quality of the source data fragments/content. In other words, a key objective is to highlight the need for authoring unambiguous and usable

data models to succeed in coding their fragments accurately. Dealing with terminology issues is a problem that can be addressed once issues with the source of the data i.e. the data models are resolved [76].

6. Finally, to suggest guidelines for modeling specific categories of data models. The two most commonly used data model categories include models to record clinical 'observations' and 'procedures'. The guidelines aim to suggest the types of fragments that must be represented in any observation or procedure data model.

1.2.2 Research Aim

The main aim of the research is to develop an approach to map the archetype fragments to semantically equivalent SNOMED concepts. By achieving semantic mapping, it is intended to achieve a single, integrated model that can help achieve semantic interoperability at the individual fragment level (i.e content level). Figure 1.1 provides an overall view of the aim of the research, which is to achieve a single integrated model by semantically mapping the data model fragments to concepts from a chosen terminology. Of course, not all data model fragments will find a relevant terminology concept for mapping, nor will all the terminology concepts be included in the integrated model. Only those concepts will be included, which will provide a standard, semantic equivalent for representing a data model fragment.

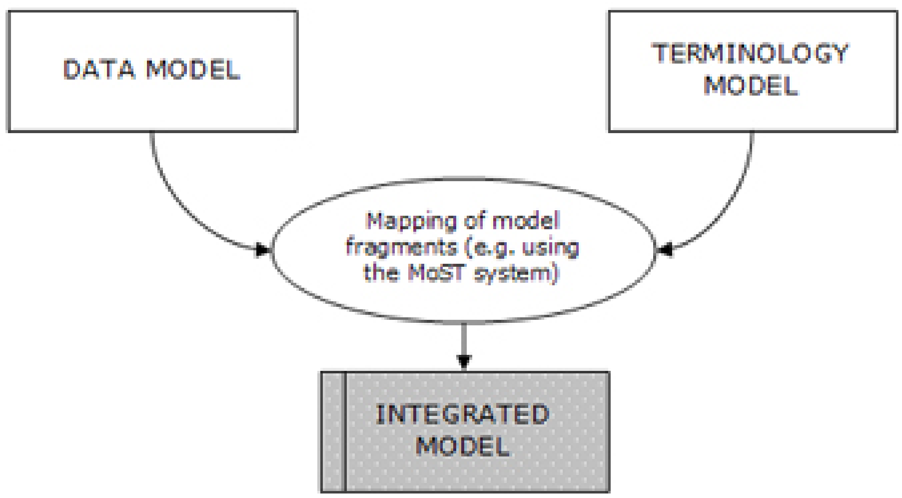


Figure 1.1: Mapping of model fragments to achieve an integrated model.

1.2.3 Research Question

This thesis forms one part of a broader research programme on semantic interoperability. It addresses the specific question of *whether a data model and terminology model can be integrated by a methodology making use of data mapping/term binding procedures.*

It also addresses the question of whether the methodology for data mapping can successfully combine automated and manual techniques, to ensure a high and reliable quality of coded data. It further seeks to identify the semantic issues in the use of the two models together, and suggests ways to resolve them.

1.2.4 Research Contributions

The research work is aimed at making scientific contributions to improve the state of informatics solutions in the health domain, especially in the field of data mapping, model integration, and semantic interoperability. Some of the intended contributions are:

- To demonstrate an effective strategy for the term finding and data mapping processes. The strategy combines lexical, linguistic, and semantic procedures to identify semantically equivalent terminology concepts.
- To share with the community strategies that failed during the research work.
- To exemplify the issues encountered when mapping the clinical content of *open*EHR Archetypes with SNOMED CT, and the solutions developed. However, to also highlight the problems that could not be resolved due to limitations in the modeling schemas and/or the principles on which the two models are based, thereby inhibiting their integration.
- To present a set of criteria and guidelines to help the communities concerned to gather momentum to improve the quality of the model content. It is intended that an increase in compatibility of the content of the two models will increase the likelihood of integration of the model content, ultimately helping to achieve semantic interoperability.

1.2.5 Significance of the Research

The significance of the research lies in the need to standardise the vocabulary used to record clinical care data so that it can be shared amongst different care settings. Aggregating data using standard vocabularies is also important for querying complete data sets stored in EHRs. Therefore, the aim of the research to develop techniques to perform quick and reliable coding of fragments to terminology concepts is considered a significant contribution to achieving the standardisation of data.

The research study aims to investigate the possibilities of semantic integration of the data and terminology models at the **clinical content/data** level. The findings made during the semantic integration process in the form of issues and their resolution, are aimed to make a significant contribution to the terminology binding and data standardisation communities. The aim is also to determine the areas of improvement, required in the form of suggested guidelines, as well as to address the semantic gaps in the two models.

The tools developed for the study intend to achieve the model integration objective using a two-tier approach. Tier One aims to perform *term finding* by developing an automated search procedure. Tier Two involves semi-automating the *data mapping* process. This includes the use of the automated procedure complemented by manual intervention. In the manual process, clinical experts evaluate the *term finding* results to determine the most semantically equivalent result for *data mapping*. The end result of the term finding and data mapping processes is intended to be the integration of the data in the two different but complementary models. The aim is not to automate the entire process but to include human expertise and knowledge, in the selection of semantically equivalent results. Participation of clinical experts in the data mapping process not only helps ensure the quality of the mapping but also helps ascertain the overall quality of the data in the two models. The latter is aimed at providing useful insight to help resolve discrepancies in the use of data in both the models. In addition, data in EHRs can only be considered ‘safe for use’ if clinical experts can examine and correct beforehand the kind of data that might be recorded in the EHRs.

1.2.6 The Research Programme

The research programme is designed to achieve the research aim of an integrated model, stated in Section 1.2.2. The programme is divided into two key stages. Stage 1 is to develop a system based on a methodology to achieve the mapping of data model fragments to terminology concepts. The two models chosen to perform data mapping are *open*EHR Archetypes, which will serve as the data model, and SNOMED CT, which will serve as the terminology to which the data model fragments will be mapped. The selection criteria forms part of the initial step of the programme shown in the flowchart in Figure 1.2. Stage 2 is to qualitatively assess the results from Stage 1 of the programme. The research programme has been detailed as a flowchart in Figure 1.2.

In Stage 1, a two-tier process is designed to achieve the mapping of the archetype fragments to SNOMED concepts, which will help with the integration of the two models at content-level. The two levels of the integration process are *term finding* and *data mapping*, as stated in Section 1.2.5. In this research, a semi-automated system called the Model Standardisation using Terminology (MoST) system, is developed to implement the two-tier process.

The three main claims made for the MoST system are that (i) semantic techniques need to supplement lexical techniques to achieve higher coverage of mappings, (ii) an automated term finding process provides a faster and more convenient method to perform mapping, as against traditional manual procedures, and finally (iii) human intervention is necessary to ensure high quality and reliability of mapped codes.

Getting a good mapping coverage for any archetype is dependent on four important factors. These are (i) the existence of unambiguous archetype fragments, (ii) the existence of unambiguous SNOMED CT concepts, (iii) good coverage of the fragment concepts in SNOMED CT, and (iv) no relevant SNOMED CT codes missed by MoST during the term finding procedure.

In accordance with the claims of the MoST system, the term finding procedure is completely automated by the MoST system, while the data mapping procedure is partially assisted by manual intervention by clinical experts. The results returned are SNOMED concepts considered by the MoST system to have some semantic equivalence to the archetype fragments. These results are evaluated by a panel of clinical experts, as part of the evaluation study. The evaluation study forms the main part of the research work, as it achieves the two main objectives. Firstly, it helps in determining the SNOMED results that are most semantically equivalent to map to the archetype fragments. The mapped codes (i.e. SNOMED concepts) help in achieving an integrated model. The first objective is a quantitative, as well as qualitative assessment, and forms the core of Stage 1 of the programme, as shown in Figure 1.2. Secondly, the observations made during the implementation process and feedback received from the clinical experts helps in analysing the issues and quality of the clinical content of the two models. As a result, a set of criteria and guidelines are suggested to the *open*EHR archetype and SNOMED communities, in particular, and modeling communities, in general, to improve the content quality to help achieve seamless model integration. This second objective is a qualitative assessment and forms the core of Stage 2 of the programme, as shown in Figure 1.2.

An issue with the research programme is the lack of a suitable gold standard to test the quality and coverage of the MoST results. In the medical domain, the ideal gold standard for judging the accuracy of data-manipulation procedures should be the true state of the patient. However, as Hoagn and Wagner recognize, this ideal is almost impossible to achieve [27]. Therefore, in the absence of any suitable predecessor(s) and the impracticality of including real patients in the study, the opinions of the clinical experts is relied upon to gauge the performance of the MoST system. The evaluation of the results by the clinicians is used as a yardstick and their opinions are regarded as the gold standard for the work. In addition to the evaluation results, an experiment is also performed as an extension to the main research work. The supplementary experiment is conducted to test whether there is any improvement in the quality and coverage of the MoST results if the data source, i.e. the archetype model fragments, are redesigned to work as a gold standard. In order to compare any improvement in the results, an existing archetype is selected. At first, the MoST system works off the original archetype fragments. The mapping coverage is recorded and retained. In the next phase, the same archetype is redesigned by clinical experts to resolve any issues to improve the content. The MoST system then repeats the process using

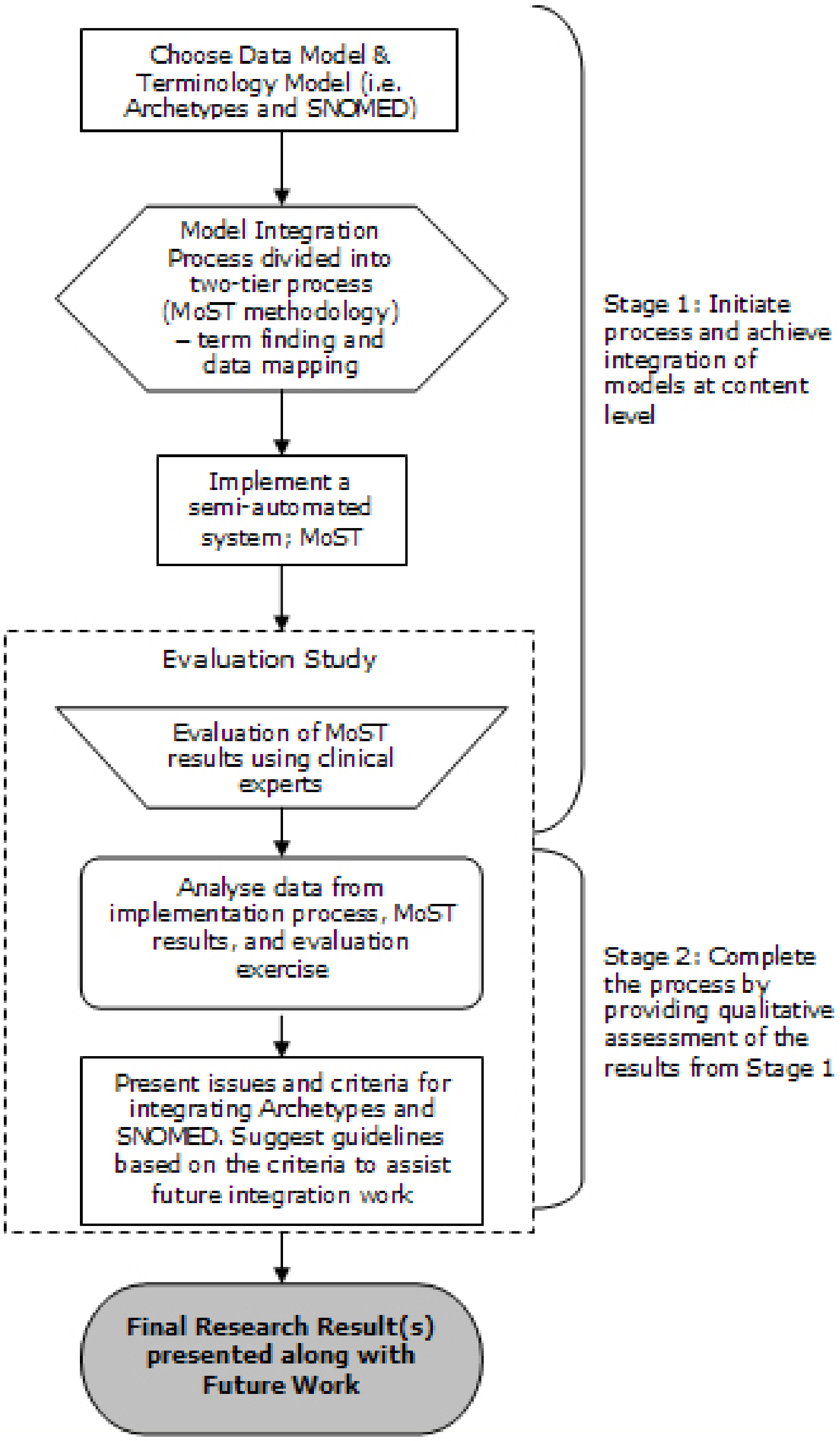


Figure 1.2: Flowchart representing the sequence of events performed as part of the research programme.

the revised archetype. The new coverage of the results is recorded and any improvement in the performance is presented. The improvements, if any, are attributed to the dependence of the MoST system on the quality of the source data (i.e. the archetype fragments). The assumption is that better quality source data, results in better quality of mapping and vice versa. Details of the experiment are discussed in Section 6.2 on Page 173.

1.3 Vocabulary and Notations

- The *open*EHR Archetype models will be referred to as 'Archetype Models/Archetypes'. They are models defining some domain concept, expressed using constraints on instance structures of an underlying reference model.
- The SNOMED CT terminology will be referred to as 'SNOMED'. It is essentially a clinical terminology that provides clinical content and expressivity for clinical documentation and reporting.
- The archetype clinical content will be referred to as 'archetype fragments' (data source), while the SNOMED clinical content will be referred to as 'SNOMED concepts/codes' (data target).
- The person who develops an archetype model will be called the 'archetype author' or the 'archetype modeler', interchangeably.
- The term 'data mapping' will also often be referred to as 'term binding' and will be used interchangeably.
- The phrase 'term finding' also refers to the 'term matching', 'data matching', or 'searching' procedures performed in the research.
- GALEN CORE, which is a terminology model will be referred to as GALEN. In addition, the GALEN terminology service will be referred to as the GALEN terminology.
- The terms 'clinical experts', 'experts', and 'evaluators' will be used interchangeably to refer to the evaluators of the MoST results. Clinical experts typically include practicing doctors, nurses, and clinicians involved in medical research.

- The MoST system developed for the research will also be referred to as MoST.
- The MoST results present in the 'log' files will be referred to as the 'pre-filtered results'. While the MoST results in the 'final' files will be referred to as the 'post-filtered' results.
- The 'mapping coverage' refers to the coverage of an archetype of finding semantically equivalent SNOMED codes. It is also referred to as the 'coverage'.

Archetype fragments will be presented in 'single quotes', whereas SNOMED concepts and codes will be presented in `typewriter font`.

1.4 Thesis Outline

Chapter 2 provides the first part of the literature review. The review is focused on clinical data models and terminologies, which are integral components of the research work. In particular, it discusses the Archetype and SNOMED models, chosen for the research. Chapter 3 provides the second half of the literature review which discusses related work that has either previously been done or is being carried out within the field of term binding, and using standard vocabularies and terminologies.

The main part of the research work is covered in four chapters. The flow of the research has been demonstrated in Figure 1.2. Chapter 3 explains the research programme and the MoST methodology and system. It provides explanation of the functionality and scope of the MoST system, which is primarily a system to automate the term finding process through the use of lexical, linguistic, and semantic techniques. It also attempts to semi-automate the mapping procedure by pre-determining those SNOMED concepts that are computationally semantically equivalent to the archetype fragments.

The evaluation of the MoST results is covered in Chapters 5 and 6. These two chapters also form part one and two of the evaluation process. The first part of the evaluation, discussed in Chapter 5, includes a quantitative analysis of the evaluation results. The second part of the evaluation, covered in Chapter 6, critically analyses the results for each of the four evaluation archetypes, by taking into account the evaluators qualitative feedback.

Chapter 7 concludes the research work by providing an overall discussion of the model integration process. It highlights the semantic issues when working together with the archetype and SNOMED models. The issues serve as the basis for the criteria to model unambiguous data to improve results. In addition, it also suggests guidelines to follow when modeling archetypes. Although the guidelines are stated with respect to archetypes, the theory can be applied to any structured data modeling formalism. Guidelines are also stated for making terminologies more logically sound and computable.

Finally, Chapter 8 concludes the thesis with a summary of the work performed and discussed in the preceding chapters. It also provides suggestions for future work to take the research forward. The chapter revisits the contributions and objectives stated in this chapter and discusses whether the claims have been successfully achieved.

1.5 List of Publications

The list of research papers published during the Ph.D. degree are:

- R. Qamar, J. Kola, and A. Rector. Unambiguous Data Modeling to Ensure Higher Accuracy Term Binding to Clinical Terminologies. In *AMIA 2007 Annual Symposium*, 2007. Chicago, U.S.A. Awaiting Publication. [76]. Available at http://www.cs.man.ac.uk/q̃amarr/papers/AMIA2007_RQamar.pdf

- R. Qamar and A. Rector. Semantic Issues in Integrating Data from Different Models to Achieve Data Interoperability. In K.Kuhn et. al (Eds), *Proceedings of Medinfo 2007*, Pages 674−678, 2007. Brisbane, Australia. [78]. Available at http://www.cs.man.ac.uk/q̃amarr/papers/Medinfo_Paper_RQamar.pdf

- R. Qamar and A. Rector. Semantic Mapping of Clinical Model Data to Biomedical Terminologies to Facilitate Data Interoperability. In *Healthcare Computing 2007 Conference*, Pages 116−26, 2007. Harrogate, U.K. [79]. Available at http://www.cs.man.ac.uk/q̃amarr/papers/ HealthcareComputing2007_Qamar.pdf

- E. Sundvall, R. Qamar, M. Nyström, M. Forss, H. Petersson, H. Ȧhlfeldt, and A. Rector. Integration of tools for binding archetypes to SNOMED CT. *Supplement in Biomed Central (BMC) Journal 2007*, 2007. [103]. Available at http://www.hiww.org/smcs2006/proceedings/12SundvallSMCS2006final.pdf

- E. Sundvall, R. Qamar, M. Nyström, M. Forss, H. Petersson, H. Ȧhlfeldt, and A. Rector. Integration of tools for binding archetypes to SNOMED CT. *SNOMED-CT and Semantic Mining Conference (SMCS)*, Pages 64–68, October 2006. Copenhagen, Denmark. [104]. Available at http://www.hiww.org/smcs2006/proceedings/12SundvallSMCS2006final.pdf

- R. Qamar and A. Rector. Most: A System to Semantically Map Clinical Model Data to SNOMED-CT. *Semantic Mining Conference on SNOMED-CT (SMCS)*, Pages 38–43, October 2006. Copenhagen, Denmark.[77]. Available at http://www.hiww.org/smcs2006/proceedings/7QamarSMCS2006final.pdf

- A. Rector, R. Qamar, and T. Marley. Binding Ontologies and Coding Systems to Electronic Health Records and Messages. *Formal Biomedical Knowledge Representation (KR-MED 2006)*, Pages 11–19, 2006. In: Bodenreider O,editor [86]. Available at http://www.cs.man.ac.uk/q̃amarr/papers/Terminology-binding-KRMED-rector-final.pdf

Chapter 2

Clinical Data Models and Terminologies

Coiera states in his book the *Guide to Medical Informatics, the Internet and Telemedicine*[33] that health informatics is as much about computers as cardiology is about stethoscopes. The tools of informatics include clinical guidelines, formal health languages, information systems, or communication systems like the Internet. However, these tools are only a means to an end, which is the delivery of the best possible healthcare.

From the start of the health informatics efforts, one of the most critical factors has been the quality of data. There have been various efforts to encourage medical practitioners to use computer systems to collect and distribute patient data, such as in UK in 1987 [80]. However, most of the quality of the data collected has been below the expected standard of consistency and completeness, requiring further efforts to improve the health systems.

One of the advances in the field of health informatics has been undertaken by the health services in UK. In September 1998, the Department of Health in UK, published the *Information for Health* document to examine the issues of developing an Information Management and Technology strategy at the local level of the National Health Services (NHS) [97]. This led to further development of Electronic Health Records (EHRs), which has now created an increasing need to integrate the various EHRs.

Clinical data models such as *open*EHR Archetypes and the HL7 Clinical Document Architecture (CDA) define the *structure* of the information to be stored in Electronic Health Records (EHRs)[91]. On the other hand, terminology models such as SNOMED CT[1] and GALEN[2][87] define the *meaning* of what is stored [91]. The clinical data models and terminology models should ideally be able to work together to enhance the quality of data being recorded, to increase data reuse across different domains. Despite this, each model formalism is developed independently of the other, with little or no knowledge of each others' requirements.

The research objective of developing a methodology to map clinical data between two models was tested with the help of a data model and a terminology model. Clinical content/fragments from the data model were mapped to appropriate clinical concepts/codes from the preferred terminology. The *open*EHR Archetype data model and SNOMED CT terminology were selected for *model integration*, to test the feasibility of the MoST[3] methodology. This chapter will provide a detailed explanation of the two main models chosen for the study i.e. the *open*EHR Archetype models and SNOMED CT. It will also state the reasons for selecting these two models over other comparable models.

2.1 Why is the medical domain particularly complex?

Kohn [62] stated in his report *To Err is Human* that health care is a decade or more behind many other high-risk industries in its attention to ensuring basic safety. This is partly due to the lack of a single designated government agency devoted to improving and monitoring safety throughout the health care delivery system.

Rector states that the problems of standardising the medical language and terminology has been a major concern in health informatics efforts for over a

[1]Systematised Nomenclature for Medicine - Clinical Terms

[2]General Architecture for Languages, Enclopaedias and Nomenclatures in Medicine

[3]Model Standardisation using Terminology System (or MoST) is the name given to the research initiative discussed in the thesis, to perform semantic mapping, to achieve interoperability.

decade [90]. Health care data has several unique characteristics that differentiate it from other industries. These include (i) data complexity and volume of data, (ii) multi-dimensional and overlapping use of data, (iii) constantly changing set of attributes and local tailoring of data, and (iv) the need to change the reference of data based on the context-of-use [90]. A well tabulated list of points, shown in Table 2.1, that clearly compare a conventional data warehouse from a clinical one (also known as the EHR), has been presented by Pedersen and Jensen [72] in their research analysis.

	Conventional	**Clinical**
Data Model	Simple	Complex
Temporal Support	Medium	Advanced
Classifications	Simple	Advanced
Continuously Valued Data	No	Yes
Dimensionally Reduced Data	No	Yes
Very Complex Data	No	Yes
Advanced Business Rules	Maybe	Yes(Protocols)
Data Mining	Maybe	Yes(Medical Research)

Table 2.1: Characteristics of Conventional versus Clinical Data Warehouses (Taken from [72]).

It can be seen from Table 2.1 that clinical data warehouses are more complex than conventional data warehouses. However, conventional warehouses can be made more complex, if required. The authors [72] aim to highlight the challenges that researchers and developers face when working in the clinical domain.

Besides the quality of data itself, it is also difficult to get clinicians to use knowledge-based systems to help guide the process of data capture and use. The primary reason for the lack of common and widespread use of knowledge systems for decision making and data capture, especially in hospitals, is the severe time pressures under which clinicians work. The success and widespread use of a system is possible only if a large number of people agree to use it for the purpose for which they are designed. Finally, financial pressures, especially in secondary care, often leads to trimming of government budgets on hardware and computer software. Subsequently, this leads to failures in consistently pursuing computerisation of several areas of the health care process [89]. Some of the limitations have progressively declined over the past decade and there is an increasing computerisation of health care systems. However, 'data' which is at the heart of all health care, continues to require improvements to improve quality and accuracy.

2.2 Problems facing interoperability standards

EHRs have primarily been developed in response to the inability of most existing systems to record the fine grained detail, required for clinical care and clinical decision making, in a structured way [85]. Coiera [33] states that if the data contained in EHRs are to be analysed they need to be accessible in some regular way. This is usually thwarted by the variations in medical terminology used by different individuals, institutions and nations. To solve the problem, large dictionaries of standardised medical terms are being created.

Terminologies and data models play an important role in building structured EHRs and achieving semantic interoperability. Despite the need for both data models and terminologies to work together to achieve interoperability, little has been done to achieve integration of the two models at the content/data level.

Although some efforts are being made to align the use of content in structured data models to one or more chosen terminologies [64], most often the data and terminology models are developed independently by separate professional groups. With respect to the research work, archetype models and SNOMED CT have been developed separately. While the data modeling techniques are dominated by I.T. professionals, terminologies are primarily developed by clinicians and other health professionals. As such there is little communication between the two groups and understanding of each others requirements and scope. For instance in a typical scenario, Archetype modelers do not develop archetypes keeping SNOMED, LOINC, or other terminologies in mind. Likewise, SNOMED developers and other terminologists are not designing their hierarchies in a manner that can be regarded as "data model friendly".

In the absence of collaboration of the two groups of professionals i.e. the informaticians and the health professionals, it will not be possible to achieve any efforts at developing safer health information systems that interoperate seamlessly at different levels of granularity. On the one hand, clinicians are not well-versed in the field of information technology and are not aware of the technical limitations of certain solutions proposed by them. On the other hand, informaticians do not have the clinical knowledge and expertise to develop systems independently that meet the requirements of the health professionals and are, at the same time, simple to use and understand.

The lack of collaboration between health informatics solution providers is slowly disappearing with big players such as the government and big health organisations entering the scene. Although there has been a noticeable change in the manner of operation of the two sides, a lot more work needs to be done to establish formal processes to ease the integration work. For instance, a typical HL7 or SNOMED meeting has several hundred attendees, which makes achieving consensus a difficult task. Therefore, future effectiveness of collaboration work is highly dependent on formal guidelines and workflows being adhered to in order to maximise output and minimise complexity and rework.

The three main standardisation committees in the US and UK viz. ANSI, ISO, and CEN [105][106], are pushing efforts to standardise different levels of work in the area. HL7 messages have achieved approval from the ANSI standards board to be used for the exchange of health data in and across health organisations. Similarly, the *open*EHR Information Model has received accreditation as a standard by the CEN committee. Archetype models are in the process of becoming a standard for the representation of data required for structured data entry and storage in EHRs. In UK, the NHS has agreed to adopt SNOMED as the standard to be used to code all data being recorded [68]. As mentioned earlier, the main aim of selecting a particular terminology as a standard is to control the vocabulary to record patient data. A controlled vocabulary helps minimise the possibility of differing interpretation of information and the possibility of errors resulting from traditional paper records.

2.3 Benefits of standardised data in health care

The single most critical aspect in any health care system is *data*. Therefore, capturing data in a precise, standardised, and reproducible manner is critical to exchanging information accurately among widely distributed and differing users [74]. If the data can be captured and stored in a consistent and unambiguous coded form, a consistent and reliable flow of information within and across organisations will soon follow.

A structured terminology covering all aspects of clinical care is required to help codify and standardise data. Standardised data will not only benefit front line clinical staff but will also boost the performance of all sectors directly and indirectly connected to the health care process. In addition to improved patient care it will also help fields such as

- Sharing and integrating information from different applications, medical records, and decision support systems
- Querying information about populations of patients held across conforming organisations
- Determining hospital performance and improvement in care through research analysis
- Helping government health agencies to determine future care plans and quality assurance based on health statistics
- Making patients and the general public more aware of their health by access to their medical records and general medical information. (Taken from [90])

2.4 Clinical Data Models

Work in the field of controlling the clinical content in free text narratives and documents to standard terminologies using NLP[4] techniques is a growing field [56][107][65][45]. Although narratives and free text documents are in existence and need to be mapped to specific terminology codes, the field is difficult due to lack of proper context. Without taking into account the context in which a particular clinical statement has been made it is difficult to rely on the results obtained. In recognition of the importance of context, efforts are being made to increase the use of structured data entry procedures, which can control the quality and content being stored in electronic health records (EHRs). In addition to

[4] Natural Language Processing

providing sufficient traceability and reduction of omissions and errors, structured information also provides a good degree of context for the data recorded.

The aim of clinical data models is to enable clinicians to enter patient-related information in a structured form. Discussion of clinical data models will focus on the *open*EHR archetype model, which has been selected for the study. The term 'Archetypes' is used in the thesis to refer to the *open*EHR Archetype Models.

2.4.1 Need for Structured Clinical Data Models

Traditionally, patient-related data was captured on paper in the form of free text narratives in clinical notes, or in prescription notes. Several projects such as MedLee[46], CLEF[32], and other Natural Language Processing (NLP) projects such as Gate[35], and MeSH[29] are working with clinical documents and narratives to extract clinically relevant information for coding to standard terminologies or for storage directly to EHRs. However, incomplete and poor quality data leading to medical errors and negligence has brought about an awareness to improve the methods used to capture and record data.

Structured patient records and messages are an upcoming field of research that have the backing of several governments [51][44][8] today. These structured models are required to cope with the size and complexity of the medical domain, as traditional models have been found to be insufficient. Structured EHRs can be developed through the use of *information models*. The information models typically represent the structure of instances of different types of information and how they relate to each other in a data repository/message. The granularity of the structure is determined by the information required to be stored or communicated [91]. The *open*EHR Information Model (referred to as the Information Model) [21], and the HL7 Reference Information Model (RIM) [53] are examples of information models that have been approved by standardisation committees for use as health messaging standards.

In addition to the information models that usually define the structure of a typical EHR, there are *data models* that are subsets of the entire information model. These data models represent a specific clinical statement or recording.

The aim is to provide a model in which clinicians can record detailed clinical concepts in a structured manner through 'predictive data entry' [85]. The *open*EHR archetype models are such an instance of structured data models that conform to the Information Model. These structured data models help guide the data-entry process, and are being viewed as a necessary replacement to the traditional methods.

2.4.2 Examples of Data Models

The two most popular and widely recognised clinical data models are the *open*EHR Archetypes and the HL7 CDA documents. These data models are usually governed by some formalised reference schema such as the *open*EHR Information Models [21] and the HL7 RIM [37]. The reference schema places constraints on the kinds of values that will be permitted for an instance of the data. It also governs the relationship of the particular data type with other data units which form part of the same subset. In the present context, the term 'subset' is used as a unit of information required to successfully complete a particular task. This unit may contain one or more entities from the original model. This section will provide a brief explanation of the *open*EHR Archetypes, and the HL7 CDA documents.

2.4.2.1 The *open*EHR Archetype Models

Beale [19] states that the name *Archetype* is introduced to denote a model defining some domain concept, expressed using constraints on instance structures of an underlying reference model. In the case of the *open*EHR Archetypes, the underlying reference model is the *open*EHR Information Model. An important feature of archetypes is the ability to specify constraints on the data being represented in the archetype models. A detailed explanation of the archetype modeling approach has been presented in Section 2.4.3, as it has been selected as the preferred data model for the research. The *open*EHR specific archetype modeling approach, along with the Information Model, is detailed in Section 2.5.

Briefly, the *open*EHR archetypes were selected because of the easy availability of published archetype examples, and their sound principles of separating the data model from the knowledge models. In addition, the *term binding* feature, which took advantage of the two-layered approach (i.e. separation of data model from knowledge model), helped perform semantic mapping of archetype fragments (i.e. clinical content) with external terminology codes, with relative ease. A comparison with the HL7 CDA documents, is also provided later in Section 2.7.2.

2.4.2.2 HL7 CDA documents

The HL7 CDA is a document markup standard that specifies the structure and semantics of a clinical document (such as a discharge summary, progress note, or procedure report) for the purpose of exchange [39]. The document can include text, images, sounds, and other multimedia content. It can be transferred within an HL7 message, and can exist independently, outside the transferring message. CDA documents are encoded in XML[5] and derive their machine processable meaning from the RIM and use the HL7 Version 3 data types. The RIM and the V3 data types provide a powerful mechanism for enabling CDAs incorporate concepts from standard coding systems such as SNOMED CT and LOINC [39]. Figure 2.4.2.2 shows a RIM[6] representation of the structured document section of the CDA object model taken from [39]. See Figure A.1 in Appendix A for the HL7 RIM model.

One of the forms in which the CDA documents can be exchanged is an HL7 message. The CDA specification does not exhaustively replicate RIM definitions, but instead refers the reader to the RIM for complete definitions. While CDA may further constrain RIM definitions, at no time will CDA definitions conflict with those in the RIM. This feature is similar to the Archetype model which conforms to the *open*EHR Information Model.

The CDA document has a header and a body. The `header` is used to set the context for the document as a whole and other meta data while the `body` contains the clinical report. The `body` model of the CDA Release 2 was heavily influenced by the CEN ENV 13606, *open*EHR, and DICOM[7][36] models, particularly in helping to determine the optimal level of abstraction and in validating the model [39][37]. A sample clinical document is shown in Figure A.2 in Appendix A.

In the 2001 publication[38] on release 1 of the CDA specification, level two

[5] Extensible Markup Language

[6] HL7 Reference Information Model

[7] Digital Imaging and Communications in Medicine

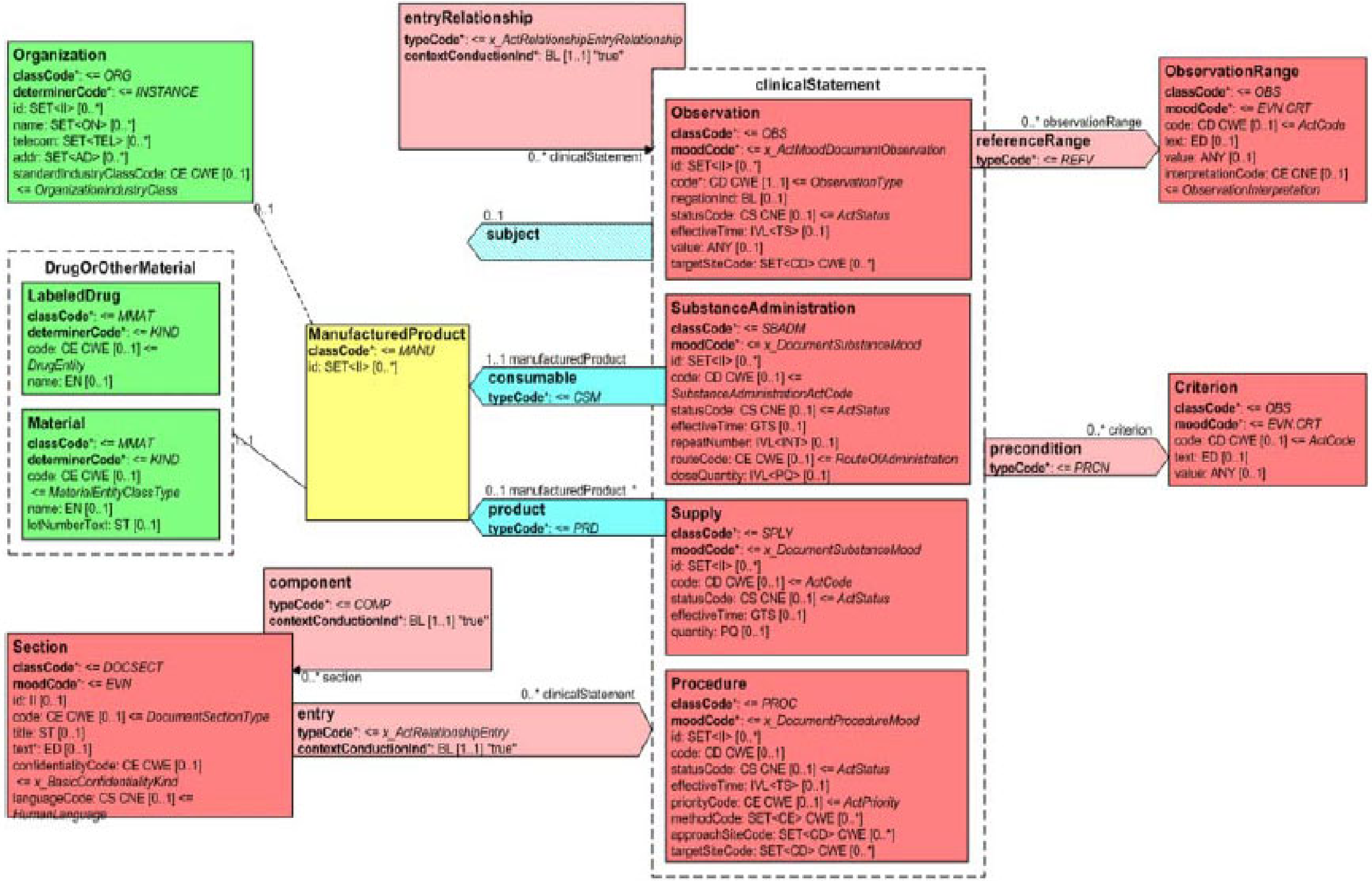


Figure 2.1: Release 2 of the CDA object model showing the connection from a document section to a portion of the CDA clinical statement.

of the architecture envisioned the use of 'templates' to constrain data in the documents. Level two of the architecture which sets the ground for the implementation of templates comes closest to the principle of Archetypes. The 2005 publication[39] of release 2 of the CDA has made a concrete proposal for the use of templates. Templates (in HL7 the term is used in the context of 'implementation guides') can be used to constrain the CDA specification within a particular implementation and provide validating rule sets that check conformance to these constraints. Like Archetypes, CDA entries can nest and can reference external objects. These external objects always occur within the context of a CDA entry and refer to content that exists outside the CDA document, such as some other image, some other procedure, or some other observation.

A common question relates to the distinction between an HL7 document and an HL7 message, and knowing which to use when. Dolin *et al.* state that [39] while there are gray zones, messages tend to be transient, trigger based, and non-persistent, whereas clinical documents have persistence, wholeness, and clinician authentication and are human readable. In this respect, Archetypes are closer in purpose and state to HL7 CDA's as against HL7 messages. Archetype repositories are maintained and can be searched to retrieve the required archetype for performing data entry and storage in EHRs.

2.4.3 History of *open*EHR Archetypes

The *open*EHR archetypes are agreed, formal, and interoperable specifications of the data structure and their inter-relationships[8]. These archetypes may be logically represented and persisted in an EHR for documenting one or more specific clinical observations, evaluations, instructions, or actions. An *open*EHR archetype conforms to the *open*EHR Information Model (referred to as the 'Information Model'). This means that the archetypes inherit the basic entities, attributes, and logical structure from the Information Model. The Information Model also specifies the set of constraints to be placed on specific data elements, along with their permitted values. The archetype approach is in the process of being standardised by the *open*EHR community for wider acceptance by the European healthcare domain.

[8]Email correspondence with Dr. Dipak Kalra, Senior Clinical Lecturer in Health Informatics, CHIME, UCL, London, U.K.

The *open*EHR foundation is an international, not-for-profit community whose objective is to enable interoperability of electronic health records to improve the provision of care. The foundation is building on international research on interoperability and is in the process of being incorporated into a European standard [10].

Precursors to the present *open*EHR archetype approach were first proposed and implemented across Europe in 1996-98, during the EU funded Synapses project[57] [58], and in 1997-99 in Australia within the Good Electronic Health Record Project [18][50]. Both approaches arose in recognition that generic information models to represent EHR data gave considerable freedom to implementers and clinicians. They could compose hierarchical representations of particular clinical record entries in potentially different ways. These "pre-archetypes" sought to standardise the way in which health record (clinical domain) concepts should be represented within generic EHR models, to meet clinical requirements and at the same time be able to interoperate [75].

2.4.3.1 Archetype Modeling Features

Archetypes include only those archetype fragments that are required to model a faithful representation of what is required to record a particular clinical statement. As such more than one archetype model may be required to record a 'complete statement'. For instance, a diabetes recording may include fragments from a blood pressure archetype, a blood sugar archetype, as well as a drug medication archetype. In the *open*EHR community, these complete statements are called *templates*. However, the term template is used differently in different communities. For instance, in the HL7 community it is used in the context of an implementation guide (see Section 2.4.2.2).

Archetypes are essentially nested hierarchies of data elements required for recording a particular clinical scenario. The main reason for having a cluster of fragments belonging to different semantic categories is due to the modeling objective. Archetypes are models-of-use and therefore include as a *nested/containment hierarchy* all those archetype fragments that are required to model a faithful representation of what is required to record a particular clinical statement. Adhering to the objective of archetypes will require inclusion of all fragments that are required irrespective of the semantic category to which they belong. Therefore, an Observation archetype could include fragments mapping to different SNOMED semantic categories such as `observable entity`, `clinical finding`, `procedure`, `body structure`, or even `physical object`, provided they are essential to be included.

Archetypes also specify pre-defined constraints on the data recorded. An archetype may be logically represented and persisted within electronic health records for documenting one or more specific clinical observations, evaluations, instructions, or actions[9].

An archetype defines and gives specifications of a data structure including optionality and multiplicity, data value constraints, and relevant bindings to natural language and terminology systems. An archetype model might contain other archetype(s), and/or might import and reuse blocks of elements occurring elsewhere in the archetype or in some other archetype. Although the archetype is designed for wide re-use, it can be specialised to include local peculiarities [24]. For instance, the *pulse* archetype is a specialisation of the *heart rate* archetype in the set of published *open*EHR Archetypes[6].

Another important feature of archetypes is their ability to translate the non-RM specific data structure, i.e. the clinical data, to more than one natural language. For example, the clinical data in the *barthel index* archetype from [6] has been translated to Dutch besides the English language. Also, the archetype modeling technique enables more than one external terminology to be bound to a single clinical data node or fragment present in the archetype. This feature is known as *term binding* in the archetype community (See Sections 2.4.3.3 and 2.4.3.4).

2.4.3.2 Purpose of Archetypes

The intended purpose of archetypes is to empower clinicians to define the content, semantics, and user interfaces (or templates) of systems independently from the Information Model. It also proposes proper integration with terminology systems

[9]Personal communication with Dr. Dipak Kalra, 2006

for different purposes. Some of the terminology systems proposed include (i) SNOMED-CT, so that reliable inferencing and decision support based on EHR data can be made possible, (ii) LOINC[10], so that traceability and sharing of laboratory data can be achieved, and (iii) ICD [11] and ICPC [12] classifications, so that reliable reimbursement, management, and public health studies can be made possible [22].

2.4.3.3 Reasons for selecting Archetypes

Archetypes were selected for the research primarily because of their "terminology separation" feature [23]. This feature, which has also been adopted by the CEN 13606 archetypes, separates the fragment expressions from the terminology used to identify (i.e. label and define) the fragment (see Figure 2.6). Local codes or identifiers are assigned to each fragment expression, which have a lexical description available separately in the `ontology` section of the archetype. Some or all of the fragments can then be legally mapped or bound to a concept from a controlled terminology. Therefore by separating the internal content representation from its lexical or controlled-coded form, the risk of depending on any terminology(ies) to complete the modeling of a clinical event is significantly reduced. As a result, several archetypes can be authored quickly and made available for use as constrained data-entry models, without being strongly coupled with external terminologies.

The *term binding* module in the `ontology` section provides a placeholder for archetype fragments to be bound to coded concepts from a controlled terminology. The term binding may be done at any stage while authoring the archetype or once the archetype has been modeled completely. Section 2.4.3.4 illustrates an example of archetype fragments mapped/bound to SNOMED codes. The intention of using archetypes is to take advantage of this "terminology separation" feature. The research aims to use this feature to perform lexical and semantic searches of all the archetype fragments to relevant SNOMED codes, once the archetype model is complete. The semantic equivalence of a SNOMED code to an archetype fragment is determined by the clinical experts, completing the *data mapping*

[10]Logical Observation Identifiers Names and Codes - http://www.regenstrief.org/loinc/
[11]International Classification of Diseases
[12]International Classification of Primary Care

process of the MoST methodology. The other reason for selecting Archetypes over other clinical data models has been detailed in Section 2.7.2.

2.4.3.4 Term Binding in Archetypes

The key objective of an archetype is to provide a suitable data-entry mechanism to the clinician, to guide and control clinical data entry. The *term binding* feature in archetypes enable fragments to be mapped to one or more terminology concepts, allowing the archetypes to comply with terminology standards. This helps in a standard and controlled method of data entry to enable reuse leading to semantic interoperability. However, at present, if a semantic match is found for an archetype fragment, it is manually mapped to external terminology systems such as SNOMED-CT, GALEN, and ICD. A descriptive syntax (or pseudo-code) of the term binding statements in archetypes to map the fragments to terminology codes, is shown in Part (a) of Figure 2.2. Part (b) of the figure displays a term binding example from the `blood pressure` archetype in the formal Archetype Definition Language (ADL) syntax. As shown in Figure 2.2(b), there are three main parts, which are required to create a term binding statement in ADL. These are (i) the internal fragment identifier, (ii) the name of the terminology model such as SNOMED, ICD, or LOINC, and (iii) the terminology concept code (or identifier). In addition, the fragment labels (i.e. internal rubrics) are available as comments (preceded by double dashes) at the end of each line in the ADL syntax, where a fragment identifier has been used. The ADL syntax is described in further detail in Sections 2.5.1.2 and 2.5.1.3.

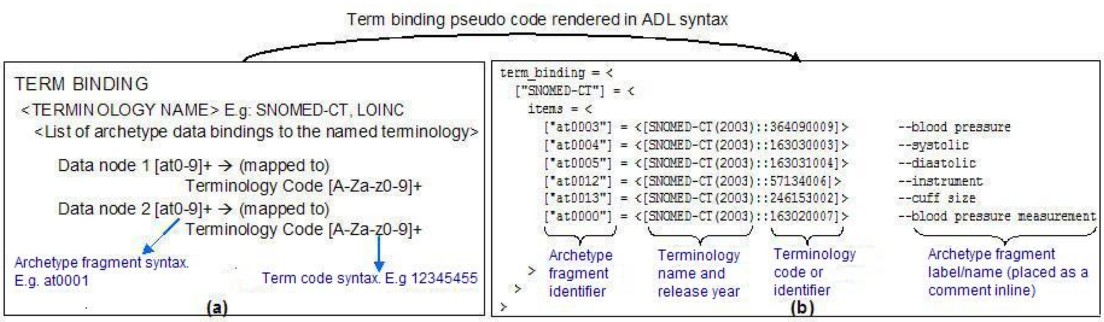


Figure 2.2: Term Binding feature in Archetypes:(a) Descriptive syntax (or pseudo code) for scripting the term binding section, (b) Sample ADL syntax of term binding in the `blood pressure` archetype.

Figure 2.2 displays an example of the *term binding* feature taken from the `blood pressure` archetype[6]. Interestingly, the blood pressure archetype is the only archetype, which has some of its internal archetype fragments bound to SNOMED codes (Last accessed [6] in June 2007). However, it can be seen in the figure that searching for relevant SNOMED matches and binding them to the archetype fragments has been performed manually with an out-dated version of the SNOMED terminology, from the year 2003. For example, Part (b) shows that the 'blood pressure measurement' fragment with the internal identifier `at0000` is bound to the SNOMED CT code `163020007`, from the 2003 version. In total, only six archetype fragments have been bound to SNOMED CT codes by the author. Since the term binding process is manual and tedious, there are very few archetypes with an active and completed *term binding* section. As stated earlier, at present only one archetype model could be found in [6] to provide an example of *term binding*. Moreover binding archetype fragments to out-dated versions of terminologies, increases the unreliability of the bound codes.

The number of archetypes written for different clinical situations are likely to increase, as are the number of clinical codes in terminologies. As such, it will become increasingly difficult to rely on human intervention to look up semantic matches for the several thousand archetype fragments, from the even larger number of terminology concepts. Techniques and softwares need to be developed to allow the archetype author(s) and clinicians to choose from an automatically generated list of candidate terminology codes, to which the archetype fragments can potentially bind/map. Chapter 4 will look more closely at the MoST methodology that was developed to achieve the automated lookup of candidate SNOMED codes, and semi-automated *data mapping/term binding*.

2.5 The *open*EHR Information Model

A two-level approach developed by *open*EHR distinguishes an Information Model from Archetypes. The Information Model represents the generic properties of health record information. It is a higher level, abstract or logical model. On the other hand, Archetypes are physical models based on the Information Model, to represent the specific characteristics required to model a specific clinical scenario. They are expressed as constraints on the Information Model, which share the features of the archetype being developed in the domain [23].

All archetypes are modeled specific to an Information Model entity based on their intended use and purpose. The archetypes then inherit the properties (or attributes) of not only the immediate entity to which they belong but also their parent entities. For example, there are four main entity classes in the Information Model i.e. ENTRY, COMPOSITION, SECTION, and STRUCTURE. The COMPOSITION and STRUCTURE submodels are concerned with the administrative aspects of patient care while the SECTION submodel is concerned with the messaging protocol. For the purpose of the research, the key model of interest is the ENTRY submodel. This submodel focuses on the actual recording of clinical content during a patient Observation, Examination, Assessment, or Intervention. The objective of the ENTRY submodel helps meet the research interest, which is to capture the actual clinical scenarios presented at the time of patient care.

The Information Model document [20] states that "all information, which is created in the *clinical statement* context, is expressed in terms of ENTRY instances" in the Information Model. These instances are in effect `classes` in the UML-based Information Model. The other reason for choosing the ENTRY submodel was because its subclasses, viz. Observation, Evaluation, Instruction, and Action, model actual clinical statements, which can be semantically categorised into SNOMED categories. Figure 2.3 taken from [21] shows the Information Model with the four main ENTRY subclasses. The Action subclass is one of the latest inclusions to the model. The Observation subclass, highlighted in the diagram, is the category from which archetypes were selected for the research.

2.5.1 The *open*EHR ENTRY submodel

The ENTRY submodel in the *open*EHR Information Model has four subclasses. As shown in Figure 2.3, the four subclasses are: (i) Observation, which covers models representing clinical observations, such as body weight, and blood pressure, (ii) Evaluation, which models assessments made after a clinical observation is completed, such as risk assessment, and (iii) Instruction and Action, which typically model surgical procedures, referrals, medication, and other clinical interventions and actions taken [21]. The 'subclasses' are referred to as the

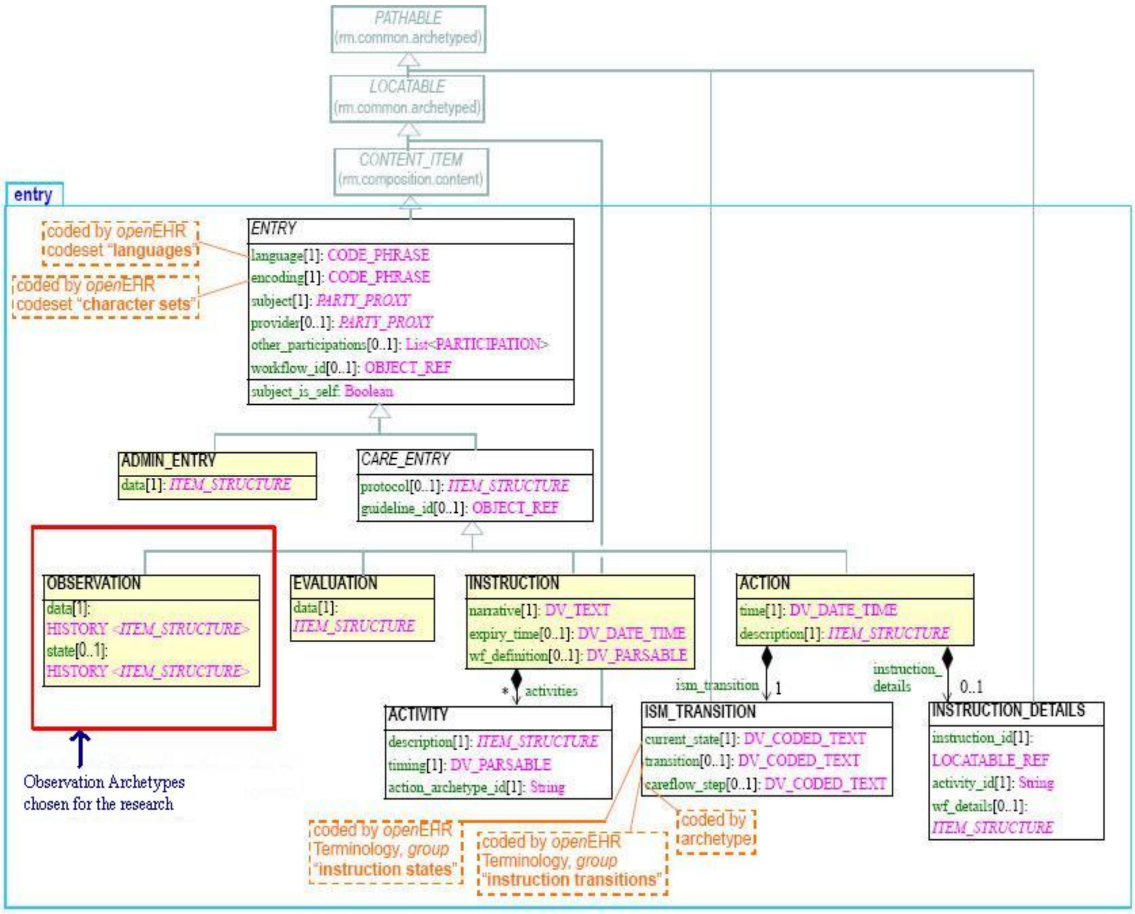


Figure 2.3: The RM.COMPOSITION.CONTENT.ENTRY sub model comprising of the Observation, Evaluation, Instruction, and Action instances. The Observation entity inherits the ENTRY attributes. In addition it has two local attributes with specified cardinality constraints (Taken from [21]).

'archetype categories' in the thesis.

The approach adopted by *open*EHR, to categorise the archetypes into the four ENTRY subclasses, helps in determining the top-level semantics to which an archetype belongs. Briefly, the four subclasses, as described in [21], are stated below.

- *Observation:* The Observation subclass is used to record the observation of any phenomenon or state of interest related to the patient. It only records information relating to the situation of the patient, not what is actually done with the patient. Among other things, observations also include pathology results, blood pressure readings, as well as the family history and social circumstances of the patient as told to the clinician.

- *Evaluation:* The Evaluation subclass, also known as the Opinion category, includes a problem/diagnosis, risk assessment based on family history, adverse reactions, goal statement, and/or recommendation of a patient care approach. For instance the problem or diagnosis would be based on a set of observed signs and symptoms in a patient, for the purpose of determining and managing treatment. The evaluation class is primarily the opinion of the clinician and is, therefore, variable.

- *Instruction and Action:* In *open*EHR, Instructions specify Actions to be performed in the future. They are more specific as compared to the Opinion category, as they are specified in sufficient detail. These instructions, specified in an Instruction chart, are to be directly enacted, for example, by the patient or a nurse, without further clinical decision-making. Instructions and actions may include all kinds of interventions such as simple medication orders, to complex multi-drug courses. The Instruction subclass includes an attribute for the narrative description of the instruction, as well as a further subclass called 'Activity'. An 'Activity' is associated with a single Instruction, such as a specific medication to be administered and its timing. However, more than one 'Activity' may be present in a single instance of an Instruction. Once an Instruction is executed, the control passes to the Action subclass.

  The Action subclass describes what was done and committed to the EHR, as a result of an Instruction. The 'ISM_Transition object' in the Action instance describes the stage at which the Instruction is being executed, while the 'Instruction_Details' object is optional and includes the workflow execution details.

2.5.1.1 The Observation subclass

For the research, only archetypes belonging to the Observation subclass were selected. Observation archetypes were selected for the study because this subclass/category had the maximum number of authored archetypes, available on the *open*EHR website [6]. In addition, these archetypes were also rich in clinical content, increasing the scope to test the degree of coverage in SNOMED.

The attributes and data types of the Observation subclass can be seen in Figure 2.3. Each attribute has a cardinality constraint, which signifies the number of times an attribute can occur in a single instance of an archetype. For example, an instance of an Observation archetype can have one and only one occurrence of the 'data' attribute recorded at some point in time. The data structure used to capture time is the *open*EHR 'History' data structure. In addition, it may or may not have a 'state' attribute but can occur only a maximum of one time, if at all.

Since the Information Model is based on object-orientation, the attributes of the parent, i.e. the ENTRY submodel, are automatically inherited by the Observation subclass. Therefore, besides the two Observation attributes (i.e. 'data' and 'state'), the subclass also inherits ENTRY attributes such as 'subject', and 'protocol'. The cardinality constraints are also inherited along with the attributes. Therefore, an instance of Observation, can have one and only one 'subject', and an optional 'protocol' which can occur only once, if at all. Typically, in an Observation archetype, the 'data', 'state', and 'protocol' attributes contain the clinical content (or archetype fragments), which are required to be mapped/bound to external terminology codes.

2.5.1.2 Structure and Syntax of Archetype Models

A typical *open*EHR archetype has a header, body, and a terminology section, which is authored in the Archetype Definition Language (ADL). Due to the complexity of the ADL syntax, archetype editors provide a simple and intuitive graphical user interface (GUI), to author the archetypes. The two more popular archetype editors are Ocean Informatics Eiffel editor [55] and Linköping University's Java editor[52]. The Java archetype editor was used for the research.

ADL is a formal language for expressing archetypes and is divided into two main syntaxes, (i) the data definition syntax of ADL or dADL, and (ii) the constraints definition syntax of ADL or cADL [26]. The *header/description* and *terminology/ontology* sections of the archetype are written in dADL, which is a syntax to express data. The main *body/definition* section of an ADL archetype

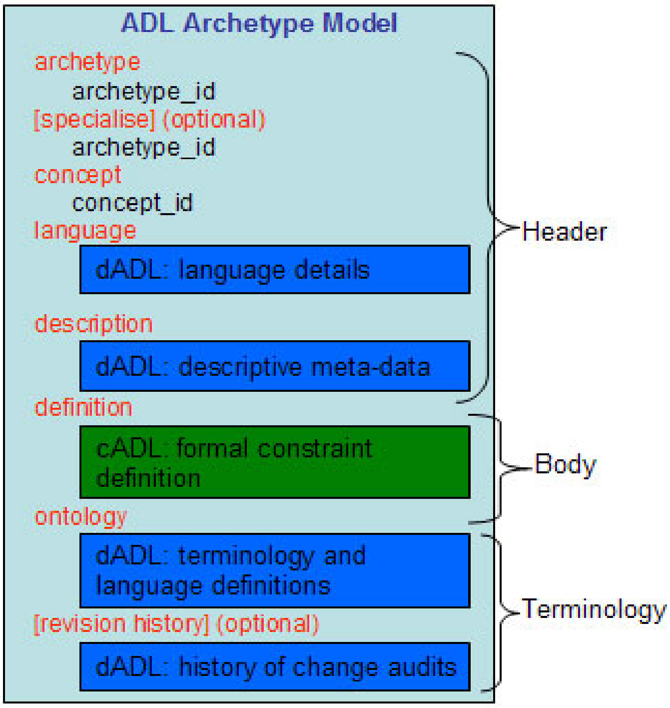


Figure 2.4: ADL Archetype Structure: (i) Description of Archetype and other meta-data in data ADL (dADL) syntax, (ii) Definition - the constraint-based data model in constraint ADL (cADL) syntax, and (iii) Ontology - the specific use of terminology and their binding to respective coding systems in dADL syntax. Other optional sections may also exist in the structure.

is written in cADL, which is a syntax to express the constraints on the data. The dADL and cADL syntaxes are demonstrated with the help of an example in subsection 2.5.1.3.

The various parts which comprise a typical archetype are shown in Figure 2.4. The `header` consists of meta-data about the archetype, including its purpose and intended use. It may also optionally state whether the archetype is a specialisation of another archetype, as seen in Figure 2.4. The `body` of the archetype is the main section, which consists of the actual attributes, constraints, and data types to be adhered to when recording a particular clinical scenario. The terminology section, which is known as the `ontology` in the *open*EHR Archetype community, consists of the "term definitions" in one or more natural languages such as English, Portugese, and others. The term definitions consist of the label/name and definition of the archetype fragments present in the `body` section of the archetype. The terminology section also consists of an optional "term binding" subsection, which maps the local archetype identifiers to external terminology concepts. The term binding feature in archetypes has been explained earlier in Section 2.4.3.4.

ADL can be used to write archetypes for any domain where formal object models describing data instances exist [25]. ADL is the primary formalism for expressing archetypes, which can also be serialised to either the XML, or HTML syntax. Previously, archetypes have been expressed as XML instance documents conforming to W3C XML schemas, as in the Good Electronic Health Record (GeHR)[18], and *open*EHR projects[26].

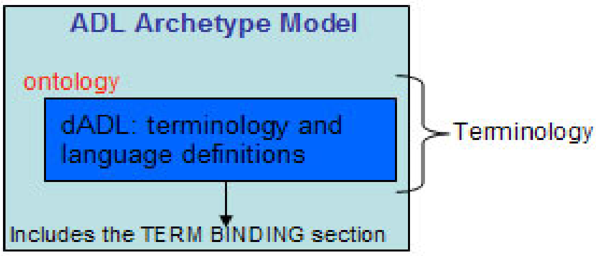


Figure 2.5: ADL Archetype Structure - Terminology/Ontology section: The *term binding* feature in the `ontology` section of a typical archetype.

The primary focus of the research application is to extract the archetype fragments along with their nested hierarchy from the `body` section of the archetype. In addition, the term definitions are also required from the terminology/ontology section, to provide the labels and definitions of the fragments, in one or more natural languages. These fragments are processed by the MoST system (see Chapter 4), to result in semantic mappings to SNOMED codes/concepts. The mapped fragments, along with the semantically equivalent SNOMED codes, are then to be stored in the "term binding" subsection of the archetype ontology, as shown in Figure 2.5. The intention is to replace the archetype fragments with standard terminology codes in EHRs, to enable uniformity in semantics and use.

2.5.1.3 Archetype Example

The three main sections of an archetype, i.e. the header, body, and terminology sections, are shown in Figure 2.6, with the help of the *autopsy* archetype in the ADL syntax. The body section, which is known as the 'definition' section in an

ADL archetype, contains the archetype fragments, which are later mapped to terminology concepts.

Figure 2.6 shows that the archetype fragments have only internal identifiers (such as at0008, and at0009) in the definition section of an archetype. However, these internal identifiers have a local name/label and a definition, which can be looked-up in the terminology/ontology section. While the internal identifiers are scripted in the cADL syntax, the term definitions are scripted in the dADL syntax[13]. The fragment labels that can be seen in the figure preceded by double dashes beside the fragment identifiers in the definition section, are provided as comments for ease of reading only. The rectangular boxes in the Figure 2.6, highlight the two main sources (i.e. internal identifiers with nested hierarchy, and term definitions) required to obtain sufficient information about the archetype fragments for use in the research.

The `definition` section is the main part of the archetype where constraints are placed on the archetype fragments used to record a particular clinical scenario. As mentioned earlier, this section is scripted in the cADL syntax. As shown in Figure 2.6, the 'internal examination' fragment is assigned the identifier `at0006`. Constraints in the form of a list of permitted body system(s) that can become the subject of an internal examination are shown in the cADL highlighted box in the figure. Therefore, the archetype block denotes that an 'Internal autopsy examination [at0006]' can be performed on one or more of the following 'Body systems [at0007]' viz. 'Cardiovascular system [at0008]', 'Respiratory system [at0009]', and 'Gastro-intestinal system [at0010]', amongst others. The labels and descriptions of the fragment identifiers at0008, at0009, and at0010, can be found in the `ontology` section of the figure. It is worth noting that Information Model specific fragments are not considered for mapping, such as at0001 (in the figure), which is an identifier for the 'History' structure. These fragments were ignored for the purpose of finding SNOMED matches, as they did not constitute clinical content.

The `header/description` section consists of meta data about the archetype. It includes general information about the archetype itself, such as the clinical

[13]See Section 2.5.1.2 for an explanation of the cADL and dADL syntax

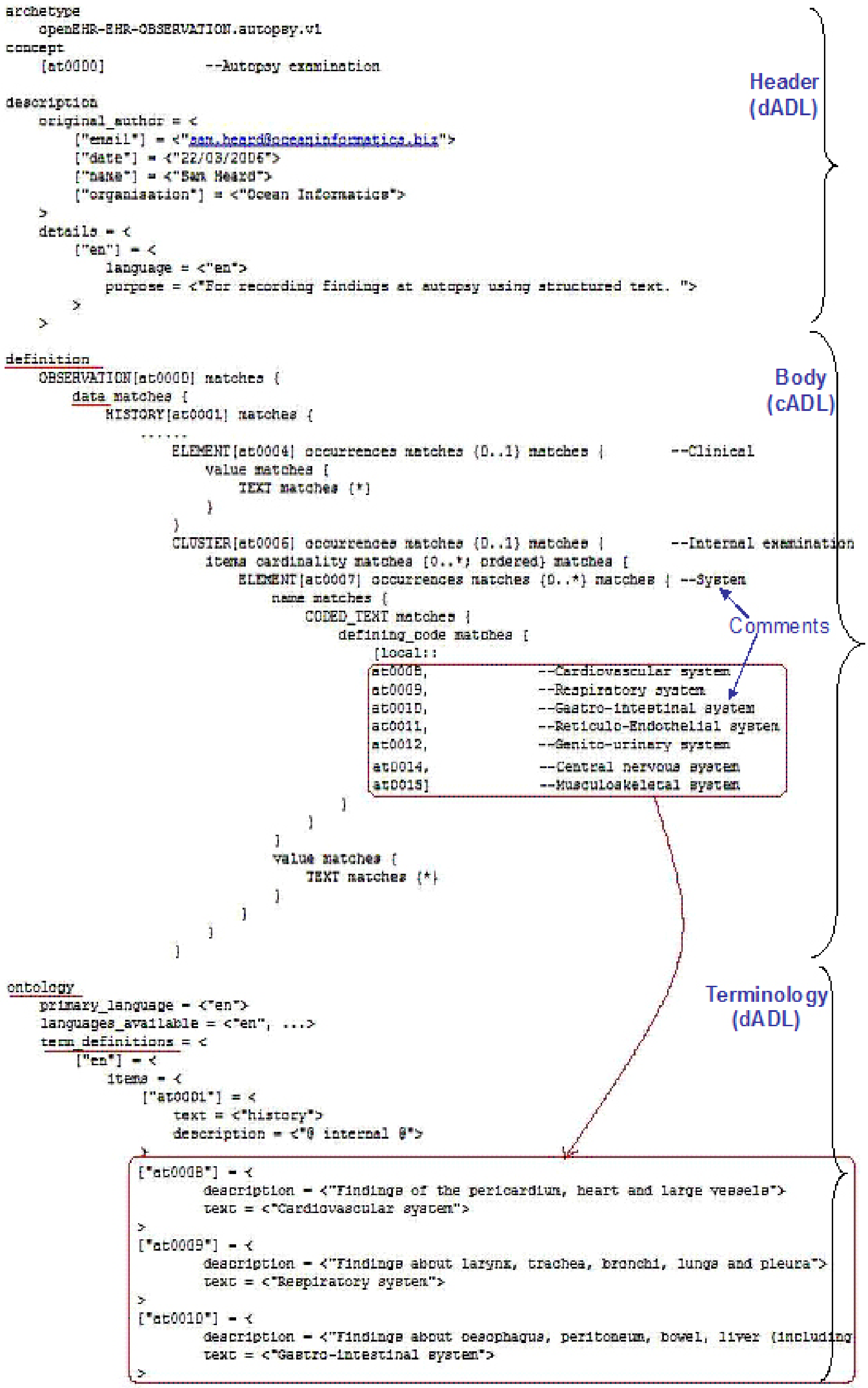


Figure 2.6: Autopsy Archetype: Sample dADL and cADL syntax in an ADL archetype. Term definitions for archetype fragment identifiers in the `definition` section are available in the `ontology` section.

scenario being modeled i.e. autopsy examination, the author of the archetype, the date of authoring, and the purpose of the archetype. It might also include whether the autopsy examination archetype is a specialisation of a more general archetype. The `description` section is scripted in the dADL syntax, as is the `ontology` section.

The `ontology` section has two main subsections viz. the *term definitions* and the *term binding* sections. The term definitions subsection provides a label and lexical description for the archetype fragment identifiers created in the `definition` section. More than one language can be used to describe the archetype fragments enabling the same archetype to be available in multiple languages. For example, all the fragments used in the barthel index archetype have been assigned labels and lexical descriptions both in the English and Danish languages.

2.6 Clinical Terminologies

A clinical terminology is a structured list of terms for use in clinical practice [5]. These terms describe the care and treatment of patients over a large area of medicine such as diseases, operations, treatments, drugs, and health care administration [5]. Rector [90] states that a clinical terminology concerns the meaning, expression, and use of concepts in statements in the medical record or other clinical information system. The use of the word 'terminology' in the thesis, refers to 'controlled vocabulary'. In this respect, the change and management of the terminology concepts is overseen by a formal committee of experts in the field. Some of the terminologies being used today in the biomedical field are SNOMED-CT, LOINC[14], ICD, and NANDA[15].

Rector [81] contends that there are four broad types of functions that require to be modeled by clinical terminologies:

1. *Conceptual:* There needs to be a formal definition, classification, and composition of concepts.
2. *Linguistic:* They need to generate and understand more complex linguistic units, including the difficulties of synonymy, metonymy, allusion, etc.

[14]Logical Observation Identifiers, Names, and Codes
[15]North American Nursing Diagnosis Association

Simple lexical labels are insufficient for representing a concept.

3. *Inferential:* It should be possible to draw conclusions about the world represented by the concepts.

4. *Pragmatic:* It should be intuitive and convenient for a human to interact with the concepts, facts, and language asserted in or inferred by the terminology.

Ideally a terminology that is capable of performing the four functions stated above would contend to being a good, computable terminology. However as a result of the research done two extensions to the above functions appear to be necessary for the computational usefulness of a terminology:

- *Unambiguous and unique:* The concepts should be unambiguous in their definition and classification to avoid conflicting inferences from being drawn. Uniqueness in representation is also required and can be achieved by using classification reasoners to identify duplicate or illogical representations.

- *Chunking hierarchies:* Due to the size of present-day terminologies, often exceeding 100,000 concepts, it is desirable to chunk the terminologies into smaller sets to increase computational performance and decrease processing time. In addition it should be in a format that enables quick lookup of the entire terminology.

2.6.1 Need for Clinical Terminologies

The need for terminologies in the field of medicine arose due to the desire to re-use clinical data. Data reuse is required either to integrate systems, to link patient records to decision support and knowledge management, or to re-use information collected in the course of patient care for management, remuneration, quality assurance or research [83]. Therefore, terminologies are primarily required by information systems to capture, process, store, use, and transfer clinical data in a standard form.

Efforts to control the vocabulary used to record clinical data for an increasing number of medical areas has led to a need to expand the term coverage for a specific domain. Cimino [31] suggests adding terms and complex expressions to the vocabulary as they are encountered rather than attempting a systematic, anticipatory solution. An alternative solution is to allow users to combine multiple terms to form a single compositional whole which will represent a required coded term. Such compositions of two or more individual terms to represent a single concept is called *post-coordination*. Single terms that can represent a concept as a whole are called *pre-coordinated* expressions. The pre-coordinated and post-coordinated expressions are often used in terminologies. These two expressions will also be used in the thesis.

One of the other requirements that is necessary for a terminology to be used in clinical information systems is to make explicit all information required to define, describe, and categorise a concept. Rector also states that it is important to explicitly state whether a concept is a kind of the other, or whether they differ so that computers making use of the terminologies can correctly infer the information and work with it [90]. Since terminologies are built primarily for computer systems rather than for the human eye, such requirements should be intrinsic features of any terminology aiming to be widely and safely used.

2.6.2 Problems with clinical terminologies and their use

Medicine has a long history of attempting to codify and classify its terminology[33]. In medicine, coding data is required to allow some degree of consistency in recording clinical data. However, as the field of medicine grows and the complexity of care increases there has been a proportionate increase in the size and complexity of clinical terminologies.

The scope of clinical information is very broad, which together with the need to express similar concepts at different levels of granularity, results in a requirement to support a huge number of concepts and recognise the relationships between them [64]. Besides, in clinical practice there are several ways of stating the same clinical expression resulting in incompatibilities between different terminologies. In addition, developing terminologies in the medical world becomes harder as new diseases, their prevention and cure, and other medical findings are constantly discovered requiring additions, modifications, and deletions to previous clinical expressions. These are some of the challenges faced by communities developing clinical terminologies. Besides the scope and complexity, there is currently no systematic means for choosing concepts from complex terminologies to conform to specific uses in a clinical data model. The problem becomes harder to resolve with so many different terminologies and no guidelines available on which one to use and when. There is also the problem of clinical data modeling communities developing their own proprietary vocabularies to guard themselves against the rapidly changing world of external terminologies. In order to maintain the reliability and quality of data, there is an urgent need to ensure that all conforming information systems access common, agreed-upon standard terminologies to record all clinical data.

2.6.3 Examples of Terminologies

There are various terminologies available today which cater to specific areas of medicine. Some examples are the clinical LOINC naming structure for laboratory test [17], ICD for disease codes [16], DICOM for imaging [17], and FMA for the human anatomy [18]. On the other hand, there are larger terminologies like SNOMED CT [73] and GALEN [87] that model more general clinical care. In addition, there is also the Unified Medical Language System (UMLS) [54], which serves as a metathesaurus of terminologies such as SNOMED CT, LOINC[19], and ICD.

Three of the widely discussed terminologies viz. LOINC, GALEN, and SNOMED CT are discussed briefly in the following subsections. The discussion provides an insight into the range of terminologies available along with their varied objectives. The SNOMED CT terminology was selected for the research, and will be explained in further detail in a separate section as well.

2.6.3.1 LOINC

The Logical Observation Identifier Names and Codes (LOINC) database provides a universal code system mainly for reporting laboratory observations. Approximately 80% of the total number of observation codes in LOINC, which equates to

[16] International Classification of Diseases
[17] Digital Imaging and Communications in Medicine
[18] Foundational Model of Anatomy
[19] Logical Observation Identifiers Names and Codes

approximately 25000 codes, relate to laboratory test observations for more than 6000 laboratory results. The aim of the LOINC database is to codify observations correctly in electronic messages stored in EHRs, as well as encourage laboratories to codify the individual tests rather than group them under the “send out lab” code [67].

The LOINC codes are one of the most widely used code-based systems in medical practice today. Large reference laboratories and federal agencies such as the CDC[20] and the Department of Veterans Affairs use LOINC. The codes are also part of the HIPAA[21] attachment proposal. Internationally, they have been adopted in Switzerland, Hong Kong, Australia, and Canada, and the German national standards organization called the Deutsches Instituts für Normung [67].

2.6.3.2 GALEN

The General Architecture for Languages Enclopaedias and Nomenclatures in Medicine technology, also known as GALEN, is designed to represent clinical information in a computable and coded format. The GALEN terminology developed as part of the GALEN Common Reference (CORE) Model, using a representation scheme called GRAIL (GALEN Representation And Integration Language), aims to be a clinical terminology that represents medical concepts that can be used by both clinicians and computers. Rector *et al.* aimed to provide a *terminology service* rather than a *terminology* through GALEN. This meant that the coding system prescribed by most terminologies is based on their functionality. The GALEN approach aimed to base the coding system on structure instead, thereby acting as an integrating force to mediate between different systems and different applications [88].

The GALEN CORE model attempted to resolve the issues of scale, scope, management, and use of large terminology code bases available in models such as UMLS, SNOMED-III, and the NHS Read Codes. The GALEN model consists of about 6000 primitive or composite categories, approximately 33000 concepts and 100 relationships. Although heavily influenced by the UMLS Semantic Network

[20] Centers for Disease Control and Prevention
[21] Health Insurance Portability and Accountability Act

[66], the GALEN ontology required finer granularity and a somewhat more general taxonomy to achieve its objectives. The coding systems focused on SNOMED International and ICD9-Clinical Model, covering most of their cardiovascular and respiratory sections, plus segments including Urology, Gastroenterology and Orthopaedic Surgery [87].

The two terminologies that have a wide enough coverage of medicine to be relevant to the research work are SNOMED CT and GALEN. The GALEN terminology had been initially selected for the research, to serve as the terminology to which archetype fragments mapped. However, attempts to work with GALEN had to be abandoned, primarily due to the narrower coverage of clinical concepts as compared to SNOMED, and the difficulty to access the available concepts embedded in the GRAIL syntax. In addition, the increasingly wide use of SNOMED meant that adoption of this terminology would result in a much larger interest in the research by the clinical terminology community, which is one of the target groups for the research contributions. Section 2.7.1 further explains the reasons for discontinuing the use of GALEN for the research. Later, SNOMED was selected as the terminology for the research.

2.6.3.3 SNOMED CT

The Systematised Nomenclature of Medicine - Clinical Terms, also referred to as SNOMED CT, aims to be a comprehensive clinical terminology that provides clinical content and expressivity for clinical documentation and reporting [34][73]. All development and management of SNOMED CT is now governed by IHTSDO[22] since April 2007, while CAP[23] continues to serve as the commissioned support organisation [2].

Briefly, SNOMED was selected for the research study because it had a large coverage of clinical concepts. Despite the large size, it was simple to access and use the concepts, and query the terminology to extract the required information. Section 2.6.4.1 will provide further details on the reasons for selecting the SNOMED terminology.

[22] International Health Terminology Standards Development Organisation
[23] College of American Pathologists

2.6.4 History of SNOMED CT

The original SNOMED work on medical nomenclature began with an initial focus on pathology (SNOP) in 1965, and progressed to include other specialities in medicine such as surgery and autopsy (SNOMED). Figure 2.7 highlights some of the major milestones achieved by SNOMED International[24], in its efforts to improve sharing health care knowledge through the use of scientifically validated terminologies [7].

Major Milestones since inception in 1965 include:

- SNOP[1] – 1965
- SNOMED – 1974
- SNOMED II (Second Edition) – 1979
- SNOMED Version 3.0 – 1993
- UMLS includes SNOMED - 1993
- LOINC codes integrated into SNOMED – 1997
- SNOMED Version 3.5 – 1998
- SNOMED RT – 2000
- SNOMED RT mapped to ICD-9-CM - 2001
- SNOMED CT (SNOMED RT + CTV3) – First release January 2002
- SNOMED CT Spanish Edition – April 2002
- Nursing codes from NIC[2] and NOC[3] integrated into SNOMED CT – August 2002
- SNOMED CT German Edition - April 2003
- Federal Health IT report recognises SNOMED CT – June 2005

1. Systematised Nomenclature of Pathology
2. Nursing Interventions Classification
3. Nursing Outcomes Classification

Figure 2.7: Major milestones achieved by SNOMED International since 1965.

Although SNOMED began as a terminology focusing on a single clinical specialisation, the current version of SNOMED CT is being widely used to index all aspects of clinical care. The terminology has been developed using the description logic (DL) Ontylog [100] to model the logical definitions of concepts. SNOMED CT concepts are placed in a subsumption i.e. 'is_a' hierarchy. Two concepts may

[24]SNOMED International is a not-for-profit organisation committed to the widespread adoption and use of SNOMED CT. Since April 2007, the control has been transferred to the International Health Terminology Standards Development Organisation (IHTSDO) located in Copenhagen, Denmark.

also be linked to each other in terms of role value maps, and defining or primitive concepts [99]. The terminology has been partially classified using Ontylog [100], and recently using FACT++ [100], to compute and infer the hierarchies. Reasoning over the hierarchy helps highlight logical errors, as well as deduce and infer relationships. At present, not many clinical terminologies can be computed using DL reasoners. GALEN is the other terminology that can be classified using FACT++ [82].

The present day SNOMED CT terminology aims to be a reference terminology for clinical data that provides a common reference point for comparison and aggregation of data about the entire health care process [99]. The terminology is the result of the merger in 2002 of NHS's [25] Clinical Terms Version 3 (formerly known as the Read Codes) in United Kingdom, with CAP's [26] SNOMED Reference Terminology (SNOMED RT) in the United States [98]. The SNOMED literature available in [98] states that in order to support ease of migration, *every* legacy Read Code is present in SNOMED CT alongside an appropriate SNOMED CT identifier/code.

The aim of merging the Read Codes and SNOMED RT was to improve and safeguard patient care by using an agreed terminology. However, SNOMED CT has greater depth and coverage of healthcare than the versions of Clinical Terms (Read Codes) and SNOMED RT that it replaces. The NHS in UK has adopted SNOMED CT as the standard computerised terminology for codifying all patient data [43]. It is intended to be used by all computers in the NHS to facilitate communication between healthcare professionals unambiguously. It is also intended to enable clinicians, researchers and patients to share and exchange healthcare and clinical knowledge worldwide [43].

The July 2006 release of SNOMED has been used for the research work. This version consists of approximately 370,000 unique concepts. The formal name given to the SNOMED concepts is the Fully Specified Name (FSN). In addition to the FSN, there is a common word/phrase called Preferred Term, and/or one or more additional terms called Synonyms. There are approximately 660,000

[25] National Health Service
[26] College of American Pathologists

Synonyms and Preferred Terms (PT) in the terminology to help expand the vocabulary used to refer to the FSN. Since the terminology has been developed using description logic it provides logical relationships between the various concepts in the terminology through a set of relationships. There are 65 unique relationships that can be broadly categorised into 4 main categories. These are the defining, qualifying, historical, and additional relationships. Examples of a few relationships used in SNOMED are `is-a`, `finding site`, `method`, and `associated morphology`. The `is-a` relationship is the most important relationship in the terminology, as it links two concepts in a logical, subsumption hierarchy. The subsumption hierarchy states the parent-child relationship between two concepts, where one concept subsumes another concept logically. In total, there exist more than 1.4 million triples in the July 2006 data release. This means that there are over 1.4 million nodes connecting two concepts together through some relationship type. These triples are known in SNOMED as the Object-Attribute-Value triples or OAV triples.

2.6.4.1 Reasons for selecting SNOMED CT

There were several reasons for selecting SNOMED CT for the research over other clinical terminologies.

- SNOMED CT is a very large terminology with over 370000 concepts and over 1.4 million OAV[27] triples (July 2006 release). It covers a large area of medicine by merging several different terminologies such as SNOMED RT, READ codes, and NANDA, making it a richer terminology to work with. The wide coverage helped in finding more SNOMED CT codes per archetype fragment, enabling a greater chance of finding a correct semantic match.

- The format in which the SNOMED CT data is delivered, is a simple text format and therefore easy to import and customise as per individual requirements. The three main data content files available in the text format were imported into a MySQL database, which helped perform simple, as well as complex querying to SNOMED CT with considerable ease.

- The SNOMED CT concept hierarchies consisting of concept definitions and

[27] Object-Attribute-Value

their relationship with other concepts in the hierarchy could be easily computed using SQL queries. The ability to compute concept hierarchies helped determine the semantic appropriateness of a SNOMED CT code with respect to the archetype fragment.

- SNOMED CT is released for use after DL reasoners have been used to classify at least part of the terminology. This made it a more reliable terminology, as compared to completely unclassified terminologies. The reliability is based on the assumption that logical errors would have been resolved during classification.

There is another research hypothesis for selecting a large terminology, such as SNOMED. The hypothesis is that if the term finding and data mapping processes, developed for the MoST system, can successfully cope with the magnitude and complexities of SNOMED CT, then it will be simpler to replicate the process with smaller, more specialised terminologies such as LOINC and DMD[28].

SNOMED CT Concept Hierarchy	
Clinical Finding	Physical force
Procedure	Event
Observable entity	Environments/geographical locations
Body structure	Situation with explicit context (formerly Context-dependent category)
Substance	Staging and scales
Physical object	Pharmaceutical/biologic product
Qualifier Value	Linkage concept
Specimen	Record artifact
Substance	Organism
Special Concept	

Table 2.2: The nineteen SNOMED CT Concept hierarchies.

2.6.4.2 The SNOMED CT Concept Hierarchies

The SNOMED content is divided into nineteen hierarchies. The root concept that subsumes (i.e. is a supertype of) these nineteen hierarchies and all the concepts below them, is the `SNOMED CT Concept`. The nineteen top-level hierarchies are shown in Table 2.2.

[28]Dictionary of Medicines and Devices

SNOMED CT Concept Hierarchy	
Concept Name	**Description**
Clinical Finding	Contains sub hierarchies of Finding and Disease. Helps document clinical disorders and examination findings. E.g.: Finding - Swelling of arm. Disease - Pneumonia.
Procedure/ Intervention	Concepts represent purposeful activities performed in the provision of care. E.g: biopsy of spine, interval appendectomy.
Observable entity	Concepts represent a question or procedure which when combined with a result constitute a finding. E.g.: medication response, auscultatory gap.
Body Structure	Concepts include both normal and abnormal structures. Abnormal structures represented as *morphological abnormalities*. E.g.: cystic vein(structure) ,vein loop(abnormality).
Physical object	Concepts include natural and man-made objects. E.g.: procedures that use devices such as catheterisation, artificial, kidney, device, latex rubber gloves
Situation with explicit context	Concepts include embedded context which changes the concept's meaning. E.g.: no family history diabetes, bacterial growth absent, birth history, outbreak.
Qualifier value	Concepts include values for SNOMED attributes that are not contained elsewhere. E.g.: unilateral, left, mild.
Substance	Concepts include biological and chemical substances such as foods, allergens, materials, and chemical constituents of all drug products. E.g: pollen, dental cement, methadone.

Table 2.3: The eight most relevant SNOMED CT Concept Hierarchies. Note: The *Context dependent category* has been changed to *Situation with explicit context* since the July 2006 release.

The nineteen hierarchies might also be considered as the semantic categories into which all the SNOMED concepts are divided. For the purpose of the research, concepts from the Special concept hierarchy are not considered. This is because this hierarchy is used for maintenance of the main SNOMED content. The hierarchy is used to record inactive, namespace, and navigational concepts, which are not used as semantic categories of the main clinical concepts.

With respect to the research work, eight hierarchies were recognised as being of high relevance. Table 2.3 lists the eight main SNOMED hierarchies/semantic categories, along with a brief description. The descriptions were taken from the official SNOMED CT website[29]. References made to, include but are not limited to, these eight SNOMED categories during the discussion of the SNOMED matches returned by the MoST system.

2.6.4.3 SNOMED CT Core Model

The SNOMED CT data is released twice every year. The main data held in the *Core Tables* is released in January and July each year, whereas the extensions (such as UK extensions) are released a few months after the main release. The Core tables consist of three main tables i.e. *Concepts*, *Descriptions*, and *Relationships*, as shown in Figure 2.8. The three core tables consist of unique `concepts`, a list of `descriptive` names for each concept, and a range of `relationships` with other concepts in the terminology. *Extensions* to the main terminology, shown as a darkly shaded box in Figure 2.8, consist of concepts developed for a particular geographic or administrative requirement that is not part of the SNOMED CT Core. For example, the US Drug Extension and the UK Administrative Extension. These extensions may be developed and maintained by IHTSDO or other organisations [14].

As shown in Figure 2.8, there are four miscellaneous features provided with every SNOMED release. These are the History mechanism, Cross Mapping mechanism, Subset mechanism, and a set of Complements. The functionality of these features is to enrich the Core table structure for terminology implementers. The *History* files help maintain a tractable log of changes made to the SNOMED data with each release. The *Cross Mappings* files contain mappings of the SNOMED codes with other coding schemes such as ICD9-CM and ICD10. The *Word Equivalents* from the *Developer's Toolkit* in the Complements feature provides a list of commonly used alternative forms of a particular clinical expression. The content of the Word Equivalent's table was appended to the MoST system's local lexicon to expand the search base (see Section 4.3.3.2; Page 108).

Another important aspect of the SNOMED Core Model are the `relationships` available to refine the definition of the SNOMED concepts. As stated earlier, there are four main characteristics of relationships: (i) Defining, (ii) Qualifying,

[29] http://www.snomedclinicalterms.co.uk/snomedct/what_is.html

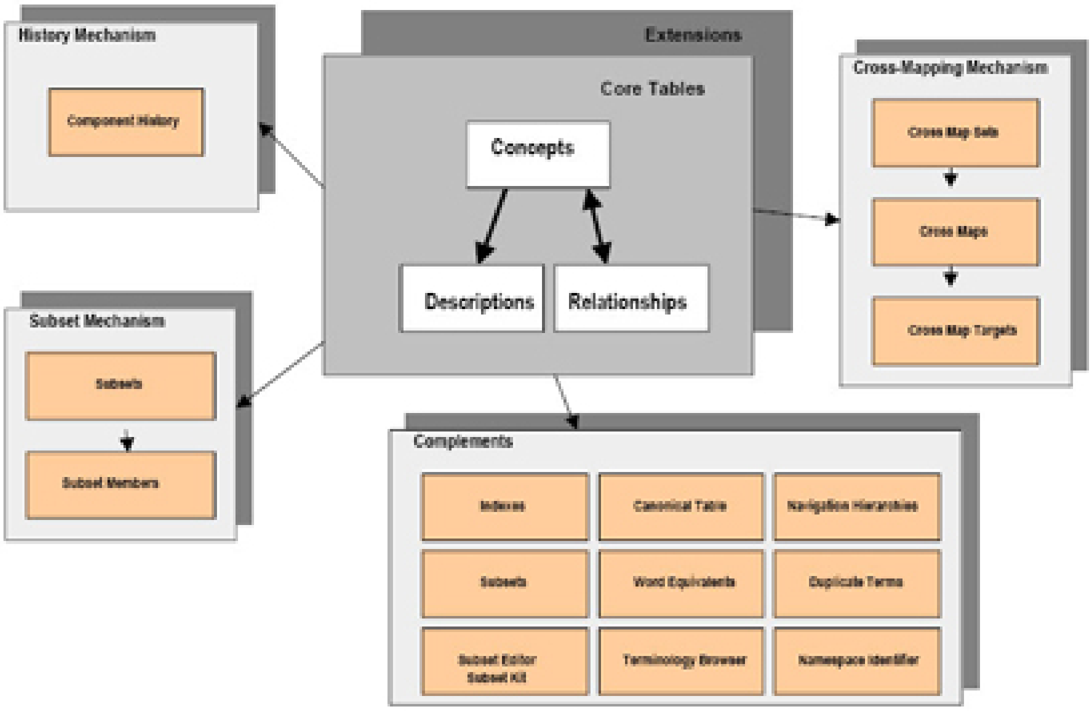


Figure 2.8: SNOMED CT File Structure: Consists of the Core Tables along with the miscellaneous data files.

(iii) Historical, and (iv) Additional. The defining, and qualifying characteristics are the two most important and relevant types, for the purpose of the research. The *defining* characteristics state things that are always true about a concept, such as 'is-a', 'method', 'finding site', and 'associated morphology'. On the other hand, *qualifying* characteristics offer options for qualifying a concept, such as 'onset', 'associated finding', 'severity', and 'episodicity'. In addition, there are six axis modifiers, which might vary the definition of a context based on the context of use (including place and time). The six modifiers are 'associated finding', 'finding context', 'associated procedure', 'procedure context', 'temporal context', and 'situation with explicit context'. Only the 'situation with explicit context' modifier is a concept of type `(situation)` and not a type of relationship such as `(attribute)`. Therefore, relationships and axis modifiers play an important role in composing different concept definitions. The two main compositions are pre-coordinated and post-coordinated concepts.

2.6.4.4 SNOMED Pre and Post Coordination

One of the features of SNOMED is its ability to provide both pre-coordinated expressions, as well as create post-coordinated expressions. These post-coordinated expressions are formed with a composition of two or more existing pre-coordinated SNOMED concepts. The ability to create post-coordinated expressions enables users of terminologies to create expressions not present in the terminology. The SNOMED CT documentation [98] defines pre and post coordination as follows:

Pre-coordination - A concept is represented using a single SNOMED code. The code semantically represents the concept in a single expression. For example, the concept `Dipstick assessment of hemoglobin concentration` is a pre-coordinated expression with the code `302781000`.

Post-coordination - A concept is represented using two or more SNOMED codes. Post-coordination is normally used to qualify a concept to refine its meaning. For example, `Emergency dipstick assessment of hemoglobin concentration` can be post-coordinated as an OAV triple, using three SNOMED codes viz. `302781000|dipstick assessment of hemoglobin concentration|` `(260870009|priority|=25876001|emergency|)`.

Post-coordinated concepts should be valid OAV triples and comply with the *refinability rules*, as specified in the SNOMED guidance document [30]. All Attributes in an OAV triple have a domain and a range. The domain specifies the hierarchy to which a specific attribute can be applied. On the other hand, the range specifies the Values which are allowed for each attribute. Therefore, any post-coordinated expression must adhere to the list of permitted domains and range, to form a legal OAV triple. For example, the domain hierarchy `procedure` can be post-coordinated to the `device` range (which is a `physical object`), using the list of permitted relationships i.e. 'procedure device', 'using device', 'using access device', and 'indirect device'.

Although the MoST methodology does not provide for post coordinated mappings, it is part of the intention for future work to address this issue through automated procedures(see Section 8.3.5; Page 214). In the research, the mappings are predominantly made to pre-coordinated SNOMED codes. However, some SNOMED matches can be used to form a post-coordinated expression of higher semantic equivalence to the archetype fragments.

Despite being a terminology with a wide coverage of clinical areas it has several duplicate and ambiguous concepts and hierarchies arising mainly due to the direct merger of the Read Codes with the SNOMED RT terminology. These shortcomings lead to issues in resolving the semantic appropriateness of some of the SNOMED codes. Further, issues were also highlighted during the evaluation by clinical experts of the MoST results, and have been discussed in Chapters 6 and 7. The SNOMED working committee, responsible for resolving ambiguities with the SNOMED content, is continually correcting and improving the content with each release. However, until a complete resolution of the ambiguities, the issues will persist and will continue to cause problems for any SNOMED mapping exercise undertaken.

2.7 Critical Comparison with Comparable Models

2.7.1 Terminology Comparison: SNOMED versus GALEN

The GALEN terminology was initially selected for the research work. However, due to certain shortcomings in terms of usability and coverage, it was replaced with the SNOMED terminology. This section describes the strengths and weaknesses of working with the two terminologies, and the reasons why SNOMED was selected over GALEN.

Implementation of the research methodology began by using GALEN as the standard terminology to code the archetype fragments. With many of the original experts on GALEN available within the research group, it was considered an advantage to work with the GALEN terminology. Besides, GALEN was comparable to SNOMED CT in the range of clinical areas it covered. The GALEN coding system focused on SNOMED-International and ICD-9-CM covering most of their cardiovascular and respiratory sections, along with segments on urology, gastroenterology and orthopaedic surgery . Also, the 'Clinical Act' section in GALEN was heavily influenced by RICHE/NUCLEUS and CEN models[87], making it a rich terminology.

As stated in Section 2.6.3.2 (Page 67), GALEN has been modeled as a terminology service rather than a simple terminology. In that sense, the GALEN terminology is based on a structure that can be used by both clinicians and computers. Terminologies such as SNOMED CT are based more on clinical functionality instead, which raises concerns on their computability. Several such computational ambiguities that arose during the research when using SNOMED CT are highlighted in Chapter 6.

One of the limitations of the GALEN terminology is the GRAIL[30] syntax in which it is made available. It is difficult to query the underlying terminology using GRAIL, as it requires significant knowledge of the language to perform any queries. The lack of documentation and support to understand and learn the language raised a significant problem. However, efforts to convert GALEN to the OWL syntax presented a solution to the problem. At the time of this research work, there was a lack of software support to load the entire GALEN OWL and perform queries on the terminology. Therefore, parts of GALEN were imported manually to the OWL format, which limited the use of the entire terminology at once.

SNOMED CT, as a comparable terminology, consisted of three core tables (see Figure 2.8) available in a simple text format. These content files were simple to load into a database to perform queries on the terminology . Queries were made using SQL[31] statements. In addition to simple queries, complex queries could also be performed to extract subsumption hierarchies to n depth level in milliseconds. It was also relatively straight forward to query the terminology to extract the concept definitions.

Although GALEN is comparable to SNOMED in the range of clinical areas included in the terminology, it is significantly smaller in size. While GALEN has approximately 33,000 unique concepts, SNOMED has approximately 370,000 unique concepts (based on the July 2006 release). Therefore, there are ten times

[30]GALEN Representation And Integration Language
[31]Structured Query Language

more concepts in SNOMED, as compared to GALEN. This means that there is a higher chance of finding one or more semantic matches for an archetype fragment in SNOMED, as compared to GALEN.

Finally, the main limitation of GALEN is the lack of active maintenance of the terminology. It is very important when using any terminology for semantic interoperability in domains, such as healthcare, that the content is constantly updated and maintained. This is important to ensure that terminology concepts being referenced by applications are always correct. It is also important to continuously expand the terminology as new requirements are identified. GALEN was found lacking in this area. In contrast, SNOMED CT has a very active working committee, which constantly reviews, updates and maintains the clinical content of the terminology. In addition, all changes made to the content are well documented, and references to *retired* and/or *inactive* concepts are always maintained. A well maintained terminology guarantees that all coded archetype fragments will always point to the same SNOMED identifier/code. This is because SNOMED identifiers are unique to a concept and are never reused to point to some other concept. Besides, being able to maintain reliable coded fragments, it is also possible to ensure using the most updated SNOMED version due to ease of use of the terminology.

As discussed in this section, the limitations of using GALEN outweighed its advantages. As a result, SNOMED CT was adopted as the preferred terminology. Previous work that had been done in the MoST system using the GALEN terminology was modified to cope with the choice of the alternative terminology i.e. SNOMED CT.

2.7.2 Data Model Comparison: Archetypes versus HL7

At the start of the research, the *open*EHR archetype models were compared to the HL7 models, to assess not only the purpose and objectives of each data model but also their ease of use, and easy availability. As a result of the assessment, the archetype models were selected as the preferred data model, for the research.

Archetypes are more easily available, and help from the experts is quicker, as the archetype community is smaller than the HL7 community. The archetype models are still evolving, which also provides the opportunity for new research initiatives to get involved in the archetype evolution process. The HL7 community is much bigger in size, and the maintenance of the models is more complex, negating the chance of any impact of individual efforts being felt by the community. These factors, along with the technical benefits of using archetypes, led to the final decision of selecting archetype models.

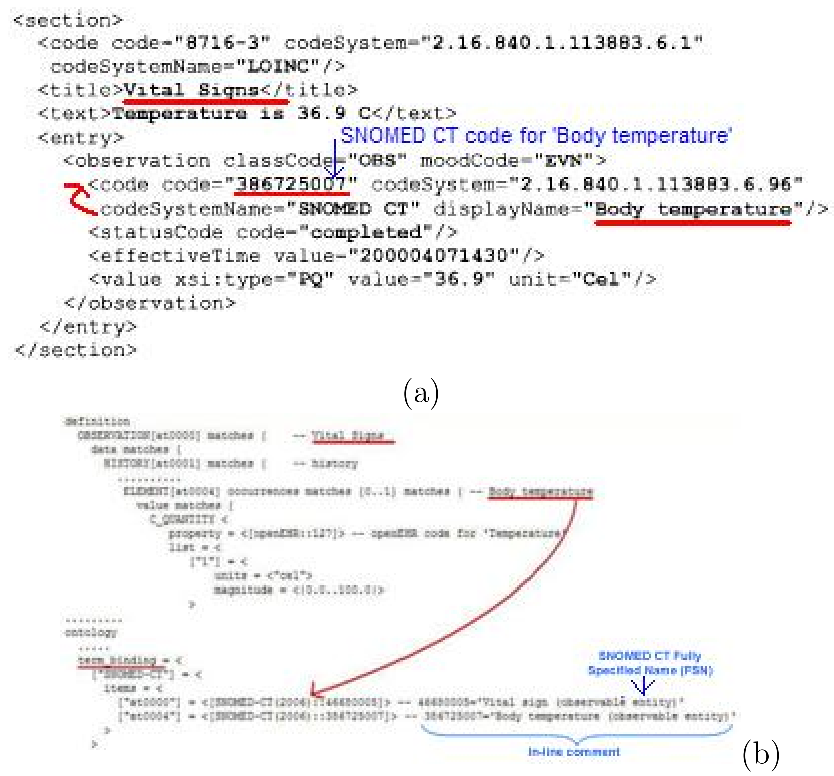


Figure 2.9: Comparison of the inclusion of terminology codes in a sample HL7 CDA document, and *open*EHR archetype. (a)Extract of a simple *observation* block in an HL7 CDA document, (b) Extract of a block from the `Vital signs` archetype.

In terms of unique features, archetypes performed better than CDAs, as they separated the core data layer from the knowledge layer, used to provide information regarding the data/fragments. The difference between CDAs and Archetypes in the way they approach the issue of binding terms to standard, controlled terminologies is shown graphically in part (a) and (b) of Figure 2.9.

A simple observation recorded in a CDA document can be seen in Figure 2.9(a). The observation block displays a body temperature measurement contained within a vital signs section. A typical observation, as seen in the XML block in the figure, has a <code> and <value> subnode. The Observation.code could refer to an external terminology code, such as the SNOMED CT code `386725007`, which identifies the concept `body temperature (observable entity)`. The attribute *moodCode* in the CDA document, has a value of "EVN" or event, which indicates that the observation has occurred. The value of the body temperature reading is stored in the Observation.value subnode.

In the CDA document in Figure 2.9(a), there is no local CDA document identifier for the rubric `body temperature`. Instead, the document directly assigns the SNOMED code `386725007` to the rubric, making the term binding tightly coupled with the SNOMED terminology. Any change in the semantically appropriate SNOMED code, or choice of the terminology itself , will require a modification to the content "hard-coded" in the data model. This shortcoming makes the model brittle and difficult to generalise by allowing more than one terminology code to bind to a single local identifier. Therefore, it can be seen that the terminology codes are embedded in the core data model of the CDA documents.

Alternatively, if the same block of the CDA document was to be modeled in an archetype it would appear as shown in Figure 2.9(b). It can be seen that the *Vital signs* archetype separates the core data model from the actual content represented in the model. Unique local identifiers (or fragment identifiers) are used in the archetype, which separately reference either local names, or terminology concepts, or both. Therefore, the internal fragment identifiers `at0000` and `at0004` can be bound to standardised codes from terminologies, such as SNOMED CT or LOINC, at any point during or after the authoring of the archetype, without any change to the core data model itself. This feature of separating the data model from the knowledge model, makes archetypes more stable and easier to bind to different external terminologies. Therefore, the ability to perform quick and easy *term binding* of archetype fragments to SNOMED CT was one of the main reasons for selecting them.

Other reasons for selecting Archetypes over HL7 CDAs were availability of examples, size and ease of use. Despite the widespread use of HL7 standards, there were no practically implemented case studies available freely. Besides, the cluster of various HL7 models, of which CDA is one such model, were too large and complex a system to be chosen for the research study. There are several notations and components in HL7 such as the RIM, RMIMs, DMIMs, CMETs, CDAs, and v3 Messages to familiarise oneself with and understand completely, before making use of them. In comparison, the archetype modeling approach is relatively simpler to understand and use. Besides, *open*EHR has several ready-to-use archetypes available on their official website [6].

Finally, the smaller size of the *open*EHR Archetype community as compared to HL7, meant more stability when implementing the data model for the research study. Too many and too frequent changes would have negatively affected the research efforts. However, this does not limit the implementation of the MoST methodology to archetypes alone. Certain customisations were required to adapt to the peculiarities of the Archetype modeling approach (discussed in Chapter 4), but this will be true when working with any other data model. In theory, the methodology is scalable and can be experimented using other models, such as HL7 CDA documents, instead.

2.7.3 Additional Vocabulary

The archetype model fragments will be referred to as 'archetype fragments', while the terminology concepts will be referred to as 'SNOMED concepts/codes'. For instance, in Figure 2.10(a), `blood gas assessment`, `arterial`, and `pH` are all archetype fragments. Likewise, in Figure 2.10(b), `acid-base balance - finding`, `pH - finding`, `hydrogen ion concentration` amongst others, are all SNOMED concepts.

In archetypes, the immediate higher-level archetype fragment of a particular fragment will be referred to as the 'enclosing fragment/enclosing archetype fragment'. For instance, in Figure 2.10(a), `Arterial` is the enclosing archetype fragment for `pH` in the Blood gas assessment archetype. Similarly, in SNOMED, the immediate higher-level SNOMED concept of a particular concept will be referred to as the 'parent concept/code'.

Figure 2.10(b) shows two different SNOMED concepts that could map to the `pH` archetype fragment. In Part (i) of Figure 2.10(b), `acid-base balance -`

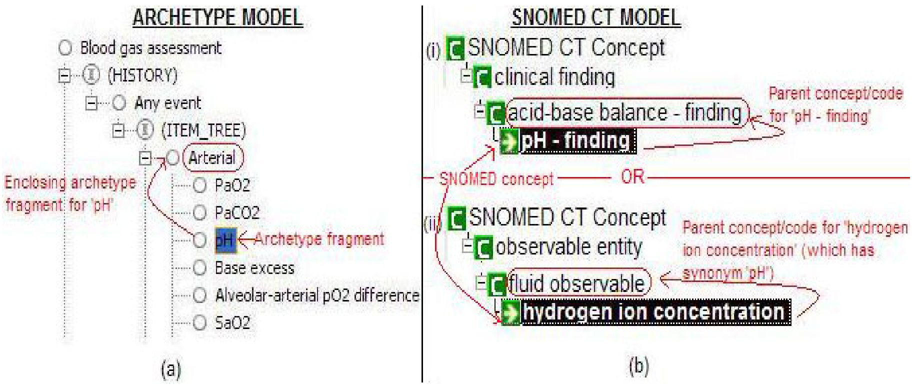


Figure 2.10: Snapshots of (a) the `pH` archetype fragment in the Blood gas assessment archetype, and (b) the `pH` concept in SNOMED (pH is the synonym for 'hydrogen ion concentration' in SNOMED).

`finding` is the parent concept of the SNOMED concept `pH - finding`. Likewise, in (ii), `fluid observable` is the parent concept of `hydrogen ion concentration`.

2.8 Summary

This chapter provided a detailed literature review of the two main models that were chosen for testing the feasibility of the MoST methodology designed to answer the research question stated in Chapter 1. The clinical data model selected for the study are the *open*EHR Archetype Models. The clinical terminology chosen is SNOMED CT. A major part of the research application includes finding semantic matches of the archetype fragments to SNOMED CT concepts.

The chapter also performed a critical comparison of the two chosen models with comparable models in the same domain. In this respect, SNOMED CT was critically compared to the GALEN terminology, and *open*EHR Archetypes were compared to HL7 CDA documents. The critique also presented the reasons why SNOMED CT and *open*EHR Archetypes were chosen over GALEN and HL7 CDA documents, respectively. The following chapter will discuss related work that has already been carried out or is being undertaken currently in the research area.

Chapter 3

Related Work

Work in the field of developing controlled vocabularies to code data uniformly in the health sector, has been progressing for nearly two decades. Sittig[96] has stated the importance of achieving a unified controlled medical vocabulary as the first of the 'grand challenges' for medical informatics. However, Rector[90] has emphasised the difficulties faced in achieving this challenge.

In relation to the work on achieving coded health data, the chapter will discuss work that has been done or is being carried out at the level of mapping fragments from different data sources to controlled medical vocabularies, and/or clinical terminologies. The four main categories of data source discussed in this chapter are (i) medical narratives and free text, (ii) databases, (iii) ontologies, and (iv) clinical data models.

The fourth data source category discussed in the chapter, i.e. clinical data models, are the main area of interest in the thesis. Work in the field of mapping data model fragments to controlled terminologies is on-going, which makes it difficult to reference any success stories. However, the Ocean Terminology Service (OTS) project recently undertaken by the Ocean Informatics group in Australia, and the TermInfo project undertaken by HL7, are major initiatives in this direction. The chapter will discuss the TermInfo project, as being the closest in principle to the thesis objective.

3.1 Mapping from Data Source to Terminologies

Work on mapping clinical content to terminology codes has become necessary to increase semantic interoperability in health care. Broadly, clinical content is extracted for mapping from four major data sources viz.:

1. medical narratives and free text,
2. databases,
3. ontologies, and
4. clinical data models

Related projects involving the use of medical narratives, databases, and ontologies, were investigated to draw lessons from previous work done in the mapping field. Projects which include the use of clinical data models were investigated to draw parallels with the research work. However, the discussion of the related work is done to compare the ways in which the research work differs.

3.1.1 Projects using Medical narratives and free text

3.1.1.1 RELMA

The Regenstrief LOINC Mapping Assistant (RELMA) is a mapping program, which is available with the LOINC terminology. The RELMA program assists in the mapping of local term codes to LOINC codes. It also helps in browsing the LOINC terminology [67]. RELMA enables users to specify their local terms in the mapping screen. It then returns all the LOINC codes that include the specified term. It also provides several search specifications to be added to the base term query to refine the mapping process, as shown in Figure 3.1.

3.1.1.2 MedLEE

The Medical Language Extraction and Encoding System (MedLEE) is a natural language processor that identifies clinical information in narrative reports and maps them to a controlled vocabulary [45]. Initially, MedLEE mapped radiology

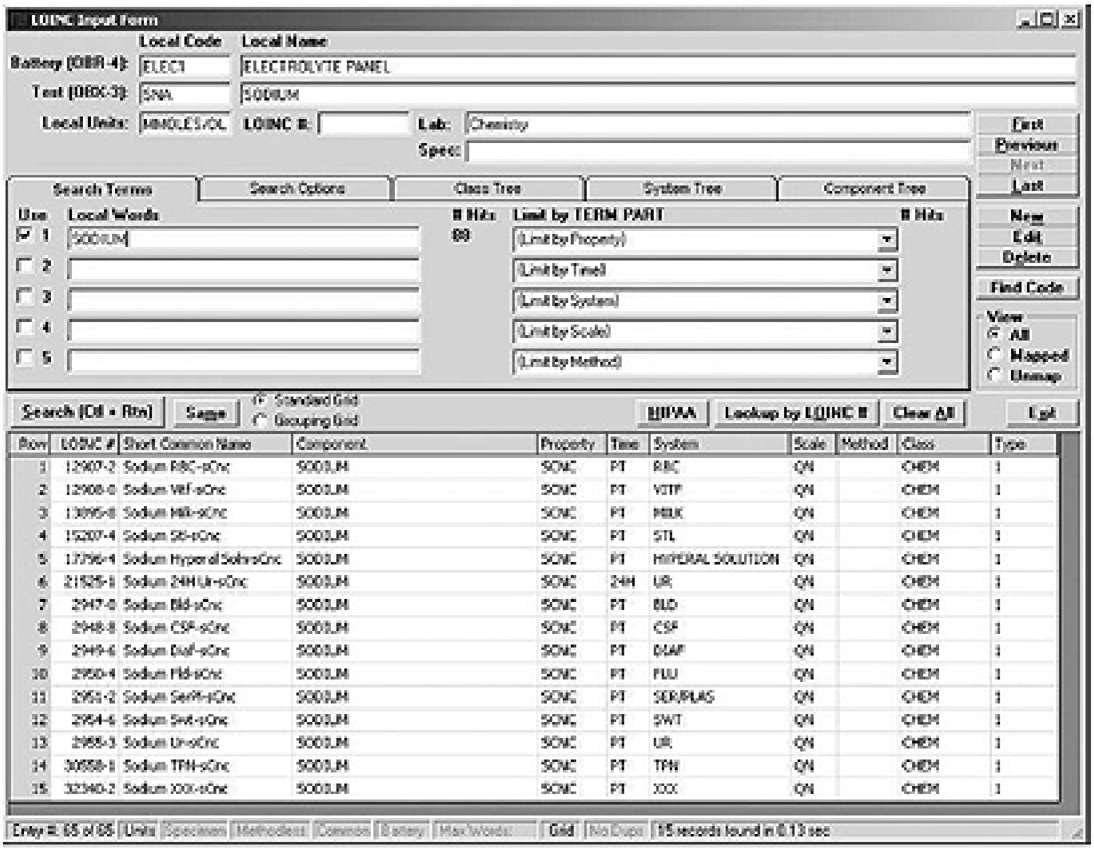


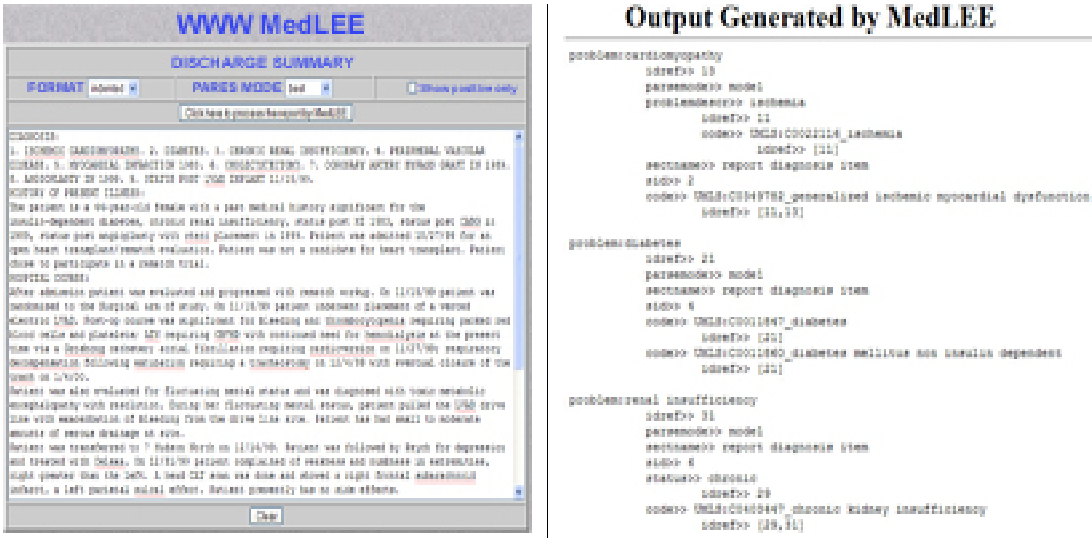
Figure 3.1: RELMA mapping table showing successful search for sodium tests that could have units = mmol/dl [67].

terms to the Medical Entities Dictionary (MED) developed at the Columbia Presbyterian Medical Center (CPMC). Ten years later, the system has been advanced to map to UMLS concepts based on structural matching using modifiers [46].

MedLEE employs various lexical and semantic rules to regularise terms identified in documents like CDA[1]. A regularised term is looked up in the UMLS knowledge source and suitable UMLS concept identifiers are returned as matches, as shown in Figure 3.2 (snapshot taken from the MedLEE website[2]). Clinical concepts from a discharge summary, shown in Part(a) of the figure, are extracted and mapped to UMLS concepts, shown in Part(b).

[1]The HL7 Clinical Document Architecture

[2]Demo available at http://zellig.cpmc.columbia.edu/medlee/demo/



(a) (b)

Figure 3.2: Demonstration of the MedLEE system: Part (a) shows a discharge summary, and (b) shows concepts mapped to UMLS CUIs (Unique Concept Identifiers).

3.1.1.3 Comparative Overview

RELMA has a good user interface with several features, thereby empowering its users to refine their search criteria to suit their requirements. However, the mapping process is rather time consuming and tedious [60]. It requires substantial user input to guide the search process. It also requires training and experience to use, to get the best performance from the system. This means that two users inputting the same query term but with different search specifications, might obtain different LOINC codes as results. Such variations might lead to mapping inconsistencies or errors [60].

The MoST system aims to reduce to the minimum, the degree of manual intervention required to perform mapping. The only user input required is at the end of the search when the final filtered SNOMED codes are to be chosen for mapping. MoST operates close to the *source* i.e. the model generating the terms for recording a clinical event. RELMA operates close to the *target* i.e. the structured terminology to which the terms are mapped.

On the other hand, MedLEE is a good medical language processor that helps obtain UMLS codes. However, if mappings are to be performed to specific terminologies such as SNOMED CT, or ICD, then an additional process will be needed subsequent to MedLEE encoding [46]. Several similar research projects [101] [94][54] have also mapped data terms to UMLS. However, in strict terms UMLS is not a terminology but a metathesaurus of terminologies and vocabularies [109]. Therefore, once a fragment has found a lexical and/or semantic match with a UMLS CUI (Unique Concept Identifier), additional work needs to be done, to extract and map to the required terminology concept(s) present in the CUI hierarchy.

At present the MoST system utilises the EMT-P service, which is similar to MedLEE. However, in the future, MoST could employ MedLEE in parallel to EMT-P. This would help compare the efficiency of performance, and quality of results, of both the systems. The better system could then be retained for future MoST processing.

3.1.2 Projects using Databases

3.1.2.1 MediClass

MediClass is a knowledge-based system that processes both free-text and coded data to automatically detect clinical events in electronic medical records (EMRs) [49]. It processes both coded and free text data from an EMR, and expresses it in the HL7 CDA data standard. If the data is already in a coded form and belongs to one of the source vocabularies supported by the UMLS Metathesaurus, the UMLS concepts are simply retrieved and returned. However, if the coded form is not supported by the UMLS Metathesaurus, the data are then mapped to UMLS concepts using various lexical and semantic techniques.

3.1.2.2 Comparative Overview

The intended purpose of MediClass to map terms from the database to UMLS concepts, irrespective of its source, is useful. However, some of the assumptions of the technique are questionable. Firstly, if a coded data term is found in the EMR, the code is assumed to be ‘correct’, and no further checks are applied to test its appropriateness. Secondly, more specific terminology codes already present in the EMR are replaced with the more general UMLS concept identifiers. In that sense, the process has adopted a preference for ‘general’ rather than ‘specific’ coding schemes. Of course, the more popular approach is to prefer ‘specific’ rather than 'general' coding schemes. In other words, most EMRs would prefer to record a code from SNOMED CT, LOINC, or ICD rather than the more general UMLS codes. A very generic map loses information and is suitable only for aggregation but not for recording information on specific diseases or patients.

3.1.3 Projects using Ontologies

3.1.3.1 BioPAX

The Biological Pathway Exchange (BioPAX) project aims to facilitate the integration and exchange of data maintained in biological pathway databases [111]. At present there are over 170 online databases storing biological pathway data. This makes it very difficult to merge the diverse database schemas to achieve integrated results from more than one database. BioPAX aims to provide a common data exchange format for pathway data that will represent the key data model elements from a wide range of popular pathway databases [111].

3.1.3.2 Anchor-PROMPT

Anchor-PROMPT is essentially an algorithm that finds semantically similar terms between two ontologies [69]. It takes as input a set of anchors i.e. start and end nodes. Based on the sub-ontology existing between the two anchors, the algorithm determines the classes that appear frequently in similar positions on similar paths [69]. The anchors are either specified manually or are determined automatically using lexical similarities.

3.1.3.3 Comparative Overview

The methodology adopted by BioPAX is very different from the MoST methodology, although the objective is similar. Attempting to map complex data, across different databases, through one common meta-ontology, runs the risk of the ontology becoming too large in size to be usable. It also runs the risk of having similar issues to other large ontologies, which it set off to resolve at the start.

On the other hand, the Anchor-PROMPT can only be used with closely related ontologies i.e. with ontologies belonging to the same domain. It is not suitable for working with ontologies belonging to different domains but using similar vocabulary. For instance, a 'clinical' ontology might have some relationship between the terms *instrument* and *hammer* when modeling the field of orthopedics. However, a 'building construction' ontology might include the same terms *instrument* and *hammer* for modeling the field of carpentry. Relying on deriving semantic similarity based on similar lexical terms alone is not a very reliable and safe method for mapping ontologies.

3.1.4 Projects using Clinical Data Models

3.1.4.1 Ocean Terminology Service

The work being done by the Ocean Terminology Service (OTS) at OceanEHR[3] is similar in intent to the research work. Besides, it is also similar in the selection of the SNOMED terminology, for the initial research implementation. OTS is a recent development initiated in 2006, and aims to connect query terms to terminology subsets [13]. No research publications detailing the internal methodology was available at the time of this writing to provide proper references. However, the Ocean Informatics product description report [13] states that OTS is aimed at providing terminology services to clinical applications. It has been designed from the outset to work with a range of terminologies including SNOMED CT, LOINC, and ICD.

The Terminology Subset Builder, which forms part of the OTS toolset, allows clinical experts to manually define subsets in the form of queries. For example, an expert can define the SNOMED subsets to be returned for a particular query. Figure 3.3 presents a screenshot of the SNOMED subset(s) constructed for the query 'Heart rhythm'. This means that for the recording of a 'heart rhythm', SNOMED codes from the subsets listed in the OTS will be returned to the user. Therefore, the subsets help in constraining the SNOMED codes chosen as a value in a clinical data entry form, rendered using data models such as archetypes.

3.1.4.2 Comparative Overview of OTS

The OTS project is still in its development stage and aims to address the issues of mapping query terms to standard terminologies such as SNOMED. The approach is similar to MoST and RELMA, although it differs in the form in which the results are returned. While RELMA returns a list of LOINC codes, OTS

[3] http://www.oceanehr.com/

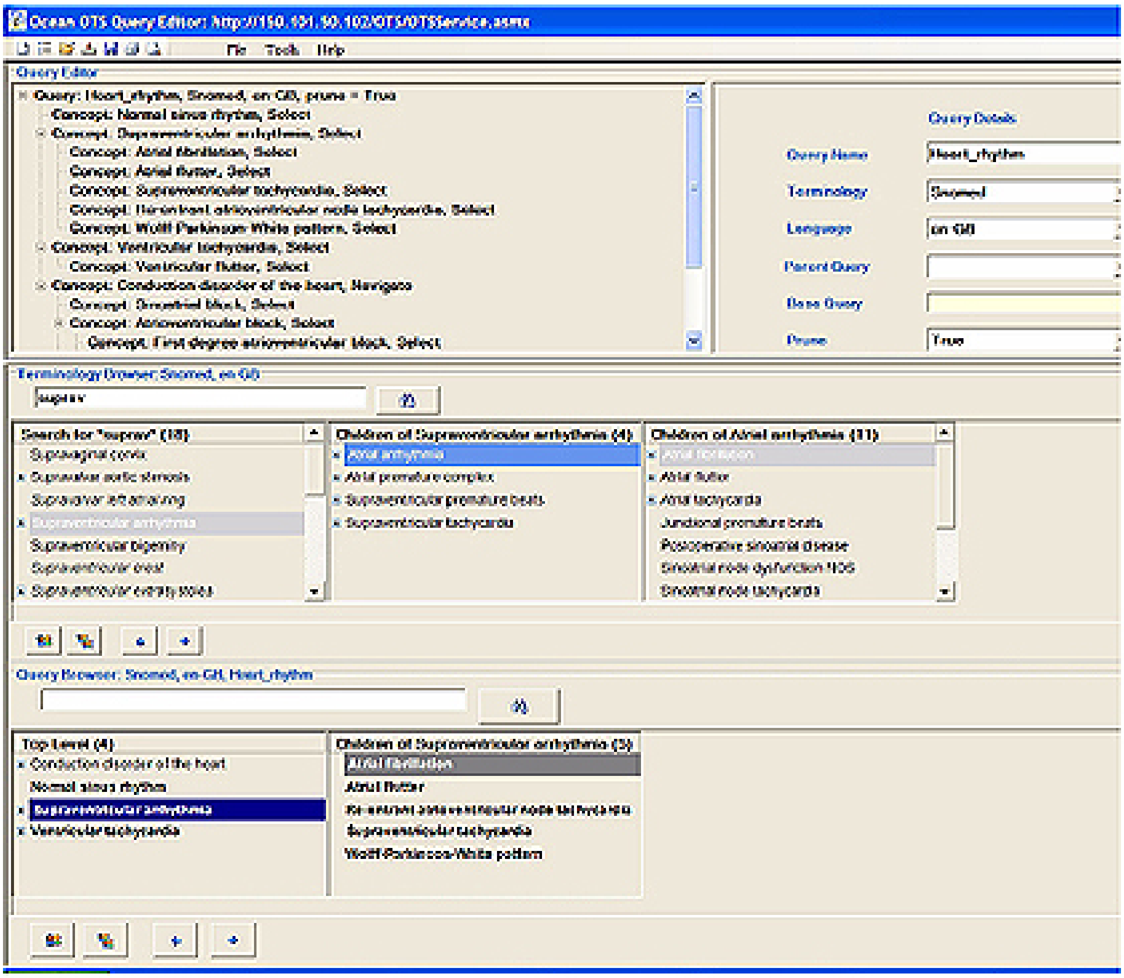


Figure 3.3: Demonstration of the Ocean Terminology Service through results returned for the query term 'heart rhythm'.

returns a subset of the SNOMED hierarchy for further browsing. However, OTS still lacks in its ability to perform more than simple lexical matches like most other search mechanisms. MoST addresses the issue by expanding the search to include synonyms, abbreviations and context information.

OTS might provide a good method to express mappings once found. However, it provides little help with finding the right codes, i.e. with *term finding*. Therefore, it has a different approach than the one adopted by the research. Further, on analysing the OTS, it is unclear whether the query terms are derived from an archetype, or are abstract query terms entered manually, and are independent of any archetype. If the latter is true then the aim of promoting the use of structured models to guide the data-entry process will be defeated. Since the terminology service (i.e. OTS) is a development initiated by the pioneers of the archetype modeling principle, it will be surprising if archetypes are not used as a data source for driving the OTS project.

3.1.4.3 HL7 Terminfo

The HL7 Terminfo special interest group has been formed to enable the clinical information contained in various HL7 Version 3 standards, to be represented using SNOMED concepts, to help achieve semantic interoperability. The objective of the Terminfo project is closest to the objectives of the research covered in the thesis. It utilises similar sources of data i.e. structured data models, for mapping to a controlled terminology. The Terminfo project relies on data from various HL7 modeling sources, which are governed by the abstract Reference Information Model (RIM) for representing health related information. The RIM specifies internal vocabularies for some structurally essential coded attributes but also supports use of external terminologies to express more detailed information. SNOMED CT is one of the external terminologies that may be used in HL7 communications. It is also the only terminology that has been tested by HL7 for establishing data compliancy rules [64].

A detailed report on the purpose, scope, and guidance on overlaps between RIM and SNOMED semantics are stated in the latest version of the Implementation Guide available for public access [64]. It is not possible to discuss details of the overlapping semantics, as it is beyond the scope of the thesis. However, a brief discussion of the various semantic equivalence that have been drawn up by the Terminfo group is presented in a tabular form.

Table 3.1 [64] displays the range of SNOMED codes that can be allowed as permissible semantic representations of the attributes in the RIM classes. For instance, the value of an HL7 Observation in 'event mood' is equivalent to a SNOMED `Clinical Finding [404684003]`, or any of its subtypes. Briefly, an 'event mood' is an Act that has happened, such as ([measurement_of]temperature = 36^{o}C). Another example is the Observation.value code with explicit context such as present or absent, past or present. This Observation value is equivalent to either a SNOMED `Clinical Finding [404684003]`, or a SNOMED `Finding`

RIM Attributes	SNOMED Representation
OBSERVATION	
Observation.code	$\leq$ 386053000 Evaluation procedure (procedure); $\leq$ 363787002 Observable entity (obs entity)
Observation.value	$\leq$ 404684003 Clinical finding (finding); $\leq$ 413350009 Finding with explicit context (situation)
PROCEDURE	
Procedure.code	$\leq$ 71388002 Procedure (procedure); $\leq$ 129125009 Procedure with explicit context (situation)
ENTITY	
Entity.code	$\leq$ 410607006 Organism (organism); $\leq$ 373873005 Pharmaceutical/biologic product (product) $\leq$ 260787004 Physical object (physical object) $\leq$ 105590001 Substance (substance) $\leq$ 123038009 Specimen (specimen) $\leq$ 308916002 Environment or geographical location (environment / location)
SUBSTANCE ADMINISTRATION	
SubstanceAdministration.code	$\leq$ 416118004 Administration (procedure);
ORGANIZER	
Organizer.code	$\leq$ 419891008 Record Artifact (record artifact); $\leq$ 386053000 Evaluation procedure (procedure)

Table 3.1: Value set constraints using SNOMED CT for some of the major classes of the Act and Entity RIM classes. NOTE: $\leq$ implies 'the class and any of its subtypes' (Taken from [64]).

`with explicit context [413350009]`, or any of their subtypes.

Following the guidelines stated by the document [64], the user can implement precisely the SNOMED codes that will 'legally' comply with the rules stated. In this manner, a true standard for interoperability can be applied to various data models for compliance with specific terminologies such as SNOMED, LOINC, and others.

3.1.4.4 Critical Overview of Terminfo

The work by the Terminfo group is by far the most well documented and clearly defined set of rules to establish semantic interoperability. However, the efforts are mainly theoretical, and there is little proof to establish the practicality of the guidelines through empirical evidence. In the absence of a strong pragmatic approach adopted by the Terminfo group, the soundness of the theory will remain debatable and open to questioning.

Despite the lack of empirical evidence, the guidelines set out a strong foundation that can be established between data models and controlled terminologies. Such sound principles of overlap between attributes in Archetypes and SNOMED CT do not exist at the time of writing of this thesis. In future, a similar joint effort between the *open*EHR and SNOMED community can increase the integration of the two models. However, care must be taken to address the issues with the SNOMED concept categorisation when setting guidelines for integrating archetypes with SNOMED. The research methodology discussed in the thesis is based on the existing state of the two models and tries to cope with the semantic gaps between the two models. It is anticipated that the effort involved in translating any such future guidelines will be minimum and will further enhance the results and performance of the MoST system. There will be fewer but compliant results returned in such a case, improving the ratio of true positives to false positives.

3.2 Other Recent Efforts

In addition to the defined areas of work explained in the previous section, there are two other relevant research efforts undertaken in the field of enhancing the quality of coded data. The first are the SNOMED subsets, developed in addition to the SNOMED CORE model. The second is the ongoing research undertaken by Rector et. al. [86], which is a logical extension to the work detailed in the thesis.

3.2.1 SNOMED CT Subsets

A SNOMED CT subset is a group of concepts, descriptions, and/or relationships chosen to be relevant for use in a given circumstance or scenario [15]. The objective of creating subsets is similar to the OTS subsets explained in Section 3.1.4.1. The SNOMED objective of creating subsets is to constrain choices, where required, to particular pre-defined categories, such as national data sets, and cancer registry data sets. Recent efforts for the NHS Connecting for Health initiative, has led to the development of the pathology and diagnostic imaging subsets [15]. These subsets are created manually by a panel of experts. Any two subsets might have overlapping use of certain SNOMED concepts or hierarchies.

The main aim is to construct a subset which will help structured clinical data entry, by ensuring that a diagnosis, finding, or procedure, is entered as such. Inclusion of subsets as an alternative format of results presented by automated systems, such as MoST, is an area to investigate further. However, an initial piece of work has already been carried out by MoST, to include primitive SNOMED subsets as results for some archetype fragments, such as 'is a kind of artery or cavity' in the *blood gases* archetype (available from [6]). The MoST results included the SNOMED concept `Arterial structure (body structure)[51114001]` and all its subtypes, as well as the SNOMED concept `Cavity (morphologic abnormality)[2483006]` and all its subtypes. Therefore, there is a logical similarity in the approach adopted by MoST and SNOMED subsets in its treatment to constrain the values that can be entered for a specific entry. The only difference is that MoST generated the results automatically before being evaluated by clinical experts, while the SNOMED subsets are created manually by experts from the offset. In addition, MoST also generated a list of single SNOMED codes as results.

3.2.2 Code Binding Interface

There are two separate pieces of research, being carried out at the University of Manchester to achieve the logical integration of different models. While the research work described in the thesis is largely involved with finding the semantically 'correct' SNOMED codes to bind to the archetype fragments, the second piece of work by Rector et. al.[86] sets out the logical foundations for binding

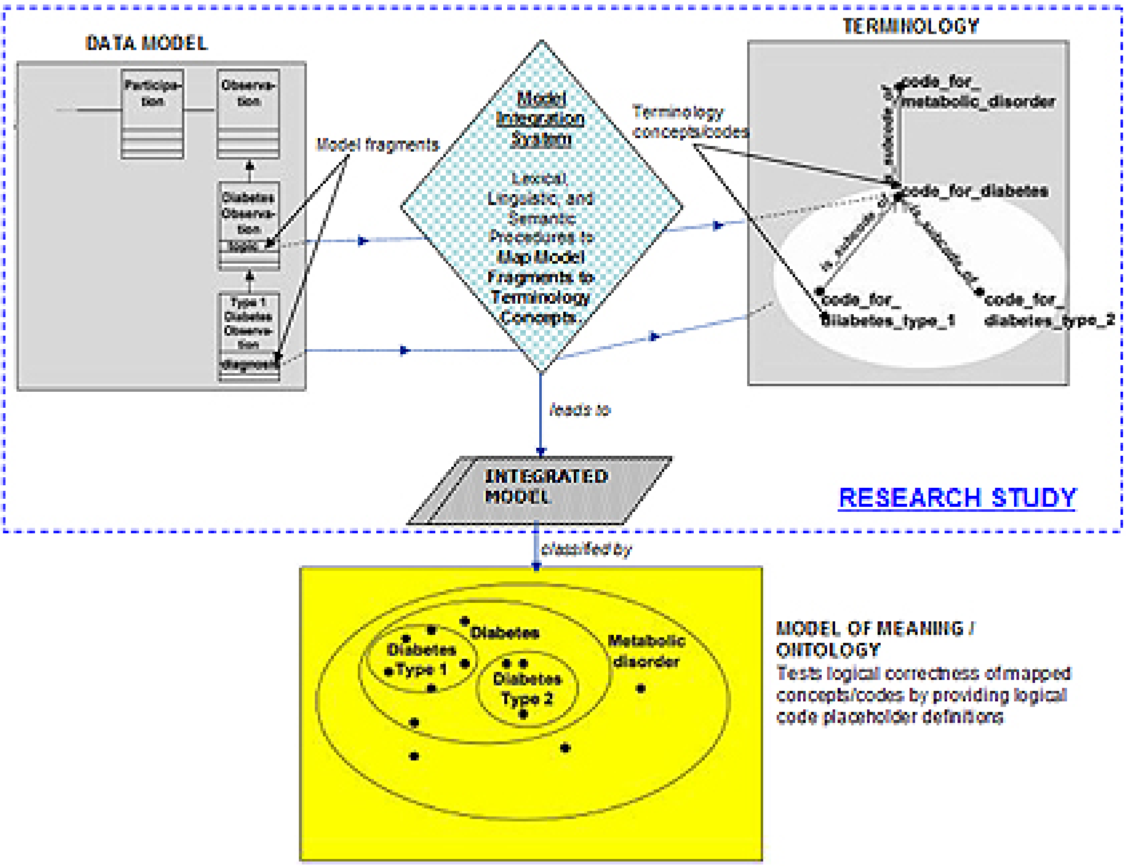


Figure 3.4: Complete area of work on integrating models and classifying them to check consistencies with the original data model. The integration work is the primary objective of the research study showed by the (blue) dotted rectangular box. The ontologies, shown in the (yellow) oval at the bottom, are classified to test for any inconsistencies in the integrated model (Adapted from [86]).

the codes when they are found. In other words, on the one hand is the research problem of finding suitable codes to map to the terms, while on the other hand is the problem of testing the logical correctness of these mapped codes.

Figure 3.4 (adapted from [86]) displays the combination of the two separate pieces of research to achieve the semantic and logical integration of data and terminology models.

The (blue) dotted rectangular box in Figure 3.4 is concerned with the first half of the research problem, and is the key area of study presented in the thesis. As shown in the diagram, the model fragments are mapped to terminology codes by a model integration system (i.e. MoST methodology) using lexical, linguistic, and semantic techniques. The result is an integrated model that controls the use of the individual fragments with respect to standardised terminology concepts/-codes.

The second half of the research problem, denoted by the (yellow) oval box at the bottom of Figure 3.4, is being carried out as a separate piece of work by Rector et. al. [86] and [92]. The research problem is being addressed by developing an ontology that acts as a meta 'model of meaning', defining the codes that can be logically included in place of the fragments, to retain the semantic and logical correctness of the original data model. The intention is to classify the ontology to indicate any inconsistencies in the integrated model. However, since the ontology work is beyond the scope of the research work presented in the thesis it will not be discussed in further detail. The combination of the two pieces of research is essential to ensure logically sound, and semantically equivalent coded data to achieve interoperability.

3.3 Summary

The chapter listed a range of projects that have been done or are being carried out to achieve semantic interoperability. The discussion was based on projects using the four main sources of data viz. (i) medical narratives and free text, (ii) databases, (iii) ontologies, and (iv) clinical data models. Work being done with the help of the fourth kind of data source (i.e. clinical data models) was closest to the research study presented in the thesis. Discussion of each of the related areas was completed with a comparative analysis with the thesis approach. The chapter also presented other relevant pieces of research work being undertaken in the area of coding data to standard terminologies.

The following chapter will introduce the MoST methodology, designed to achieve mapping of terms or *term binding* to standard terminologies. The chapter will also describe the research application (i.e. MoST system) developed to provide a practical tool to test the theoretical soundness of the methodology.

Chapter 4

The MoST Methodology and System

Chapter 2 explained the Archetype clinical data model and the SNOMED terminology model that were chosen as case studies for the research work. This chapter will discuss the *Model Standardisation using Terminology (MoST)* methodology and its supporting application, which were designed and developed to achieve the integration of the archetype fragments with SNOMED concepts/codes. The supporting application is referred to as the MoST system.

The MoST methodology was developed to provide a mechanism by which users can quickly and accurately find semantically equivalent terminology codes to bind/map to the data model fragments. Although the research focuses on the mapping of the archetype fragments to SNOMED codes, the methodology is applicable to any other complementary model(s), in theory. For instance, alternative sources for data model fragments are HL7 CDA documents, templates (i.e. data-entry forms), as well as data already stored in Electronic Health Records (EHRs). Similarly, alternative sources for terminology codes are LOINC, GALEN, and ICD-10 (see Figure 4.2).

Figure 4.1 shows a general view of the research objective, whereby the data model fragments are semantically mapped to the terminology concepts to create a single integrated model. Of course, not all data model fragments will find a relevant concept for mapping, nor will all the terminology concepts be included in the integrated model. Only those terminology concepts will be included which provide a standard, semantic code for representing a particular model fragment.

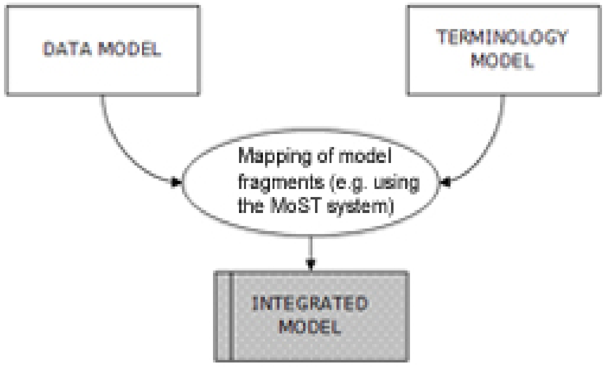


Figure 4.1: Mapping of model fragments to achieve an integrated model.

4.1 Background

The fundamental difference in the objectives and structure of the hierarchies in Archetypes and SNOMED required adjustment in the approach adopted to achieve model integration at content level. Integrated models, which address the modeling peculiarities while retaining the clinical logic addressed by each of the models, could help bridge the gap. However, at the start of the research in 2004, such integrated models were not available. Therefore, efforts were made to directly translate Archetypes to Ontologies to ease the process of mapping, by working off a common platform. Attempts were also made to achieve such translation by developing a *Headache archetype*, and remodeling it manually into an Web Ontology Language (OWL) ontology. However, the work was found to be difficult, as the archetype model conflicted with the ontology-framework of modeling. While the archetype approach focused on nested/containment hierarchies, OWL ontologies were essentially subsumption hierarchies (see Section 2.4.3.1 and 2.6.4; Pages 50-69). As a result, it was extremely difficult to develop an automated mechanism that could safely remodel a ‘nested hierarchy’ to a ‘subsumption hierarchy’, without either resulting in an OWL archetype with too many or too few concepts. Typically, the archetype required fewer data fragments to model a clinical statement, as opposed to an OWL ontology. Parallel work has since been started by Rector et. al. [86], to develop a Code Binding Interface model using OWL. This ontology has been initially tested using the HL7 and *open*EHR models (see Section 3.2.2; Page 95).

The difficulty of safely and accurately translating the ADL[1] archetype to an OWL archetype, required a change of approach. The new approach did not attempt to perform any cross-model translations but worked off the existing frameworks. The only translation that was performed was to simplify the ADL syntax used by Archetypes, to a simpler XML format. The XML structure focused on extracting only the data hierarchy and other relevant information described further in Section 4.3.3. The data present in the XML files was then used throughout the term finding and mapping processes, which not only simplified each process but also isolated the research work from rapid and frequent changes in the ADL syntax published by the *open*EHR technical community. The SNOMED data that is released as plain text files was loaded to the MySQL database, and queried using SQL[2]. Relevant SNOMED concepts were extracted, along with their concept definitions and hierarchical information. Therefore, it is advisable when working with different models to develop indigenous solutions to resolve local distinctive characteristics. The MoST methodology developed for the research helped in achieving the research objective, with a combination of indigenous and established procedures.

4.2 The MoST Methodology

The MoST methodology aims to help clinical experts to perform data mapping quickly and efficiently when building formal data models for the electronic health records. Figure 4.2 depicts the area of the research work in the (blue) dotted rectangular box labeled the 'MoST methodology'. As shown in Figure 4.2, the MoST methodology is a two-tier process to achieve an integrated model. Tier 1 is the *term finding/matching* process, which is performed by the automated MoST system. Tier 2 is the *data mapping/term binding* process, which is performed by both the MoST system, as well as by a panel of clinical experts. The experts evaluate the MoST results, and ensure that only the most semantically equivalent terminology codes are mapped to the data model fragments.

[1] Archetype Definition Language
[2] Structured Query Language

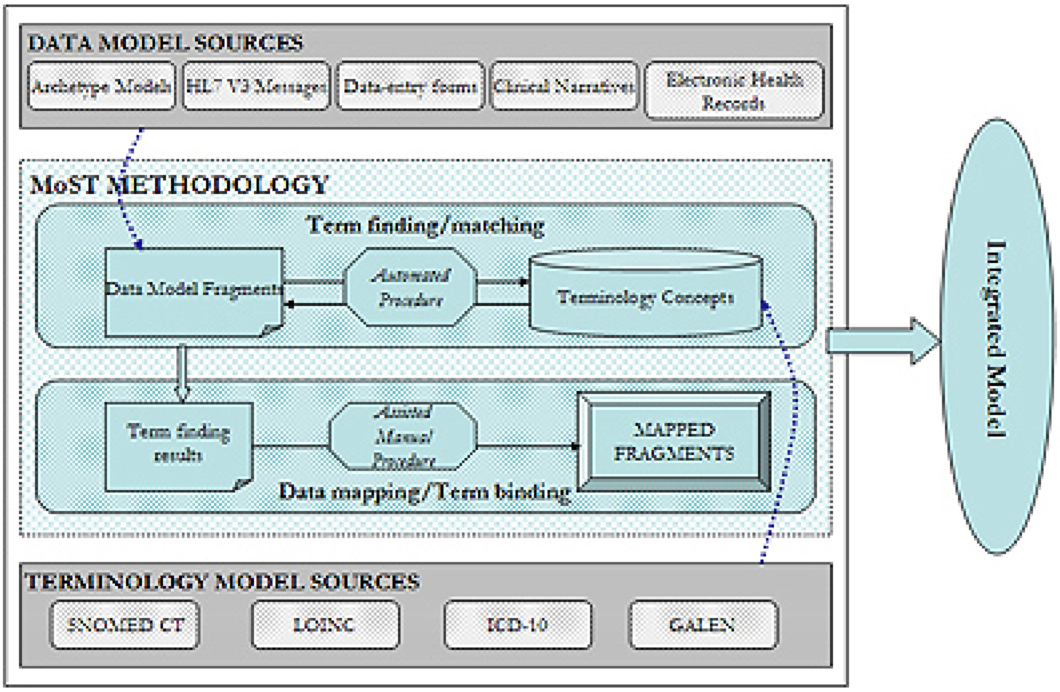


Figure 4.2: MoST Methodology depicting two-tier process for model integration: (i) Term finding, and (ii) Data mapping.

One of the research hypotheses that justifies the need for manual intervention by experts is the importance of the human eye and intelligence. The opinions of the clinical experts are invaluable in ensuring the highest quality of mapped codes, which might otherwise suffer due to complete automation. Therefore, the semi-automated MoST system proposes to take advantage of both automated procedures as well as the human intellect to achieve high quality mappings.

4.3 The MoST System: Semi-automated Term Finding and Mapping

The three main claims made for the MoST system are that (i) semantic techniques need to supplement lexical techniques to achieve higher mapping coverage results, (ii) an automated matching process provides a faster and more convenient method to perform data mapping as against traditional manual procedures, and finally (iii) human intervention is necessary to ensure high quality and reliability of mapped codes[3].

The MoST system [77][79] was integrated into the Java-based Archetype Editor developed at IMT, Linköping University (LiU) in Sweden [52]. A screenshot of the MoST system embedded in the Archetype Editor can be seen in Figure 4.9 (Page 118). This project was part of a research collaboration to extend work on the terminology binding initiative proposed by the *open*EHR Archetype community[4].

The main steps involved in the MoST system process are listed below.

Preparation Step:

Step 1: Render ADL Archetypes to XML.

Step 2: Optionally, add SNOMED semantic categories to denote 'intended meaning' to some or all of the archetype fragments.

Tier 1: Automated Term Finding

Step 1: Context Independent Search

Step 2: Context Dependent Search

Step 3: Filtering Process

Tier 2: Assisted Manual Data Mapping by clinical experts

Step 1: Determine semantic equivalence (Context dependent).

Step 2: Optionally, consider semantic categories of archetype fragments added at 'Preparation Step'.

Step 3: Manual scoring between 0-10 of SNOMED results by clinical experts to reaffirm semantic equivalence.

The Preparation Step and Tier 1 process will be explained in this chapter. An introduction to Tier 2 will also be provided although it will be extensively covered in Chapters 5 and 6. Step 1 and 2 of Tier 2 are performed by the MoST system, and are based on the same principle of searching for SNOMED codes applied to Tier 1.

[3]Two conference papers have been published to discuss the semantic mapping process achieved by the MoST system in [77][79]

[4]A joint paper between the UoM[5] and LiU have been published in both a SNOMED conference [104], as well as in a supplement to the Biomed Central Journal [103]

4.3.1 The Four Archetypes used in Development

A set of four archetypes were chosen for experimenting with the different components of the MoST system during the development stage. These archetypes were separate from the four archetypes later chosen for the evaluation exercise, and discussed in Chapter 5. The four development archetypes were (a) the Apgar archetype to represent the recordings of the Apgar index or assessment of the well-being of a newborn, (b) the Autopsy archetype to represent the findings at the time of autopsy, (c) the Barthel index archetype to record a score of dependency on help to undertake important activities of daily living, and (d) the Blood gas assessment archetype for recording the arterial or venous blood gases and respiration products.

4.3.2 The Preparation Step

When working with different data models it is best to eliminate the details of the model syntax and retain only the fragment hierarchy and properties. This simplification enables a system, such as MoST, to isolate itself from any peculiarities of the chosen data model. It also helps in extracting the main set of fragments and their hierarchies from the general data model structure, which serves as input to the term finding and mapping processes. Therefore, the archetype fragments are extracted to a local XML format along with their properties and hierarchy, shown in Figure 4.3, as the 'Generalised hierarchy' in the 'Preparation Step' module.

The second stage of the Preparation Step is to optionally add SNOMED semantic categories to denote the 'intended meaning' to some or all of the archetype fragments. The 'intended meaning' suggests the archetype author's intention of the semantics of an archetype fragment in terms of one or more SNOMED categories. This helps the MoST system to determine the semantic category(ies) from which SNOMED results would be most relevant when coding the archetype fragments. However, this process is optional, as it was found difficult to consult the original archetype author to determine their original intention when creating each archetype fragment in an archetype.

4.3.3 Tier 1: Automated Term Finding Process

On completion of the Preparation Step (see Section 4.3.2), the archetype fragment hierarchy extracted in the local XML format is sent to Tier 1 of the integration process. The *term finding* process is the first stage of the two-tier process shown in Figure 4.1. The automated procedure searches for semantically equivalent SNOMED concepts for the archetype fragments in an archetype model. The term finding process has been diagrammatically presented in Figure 4.3.

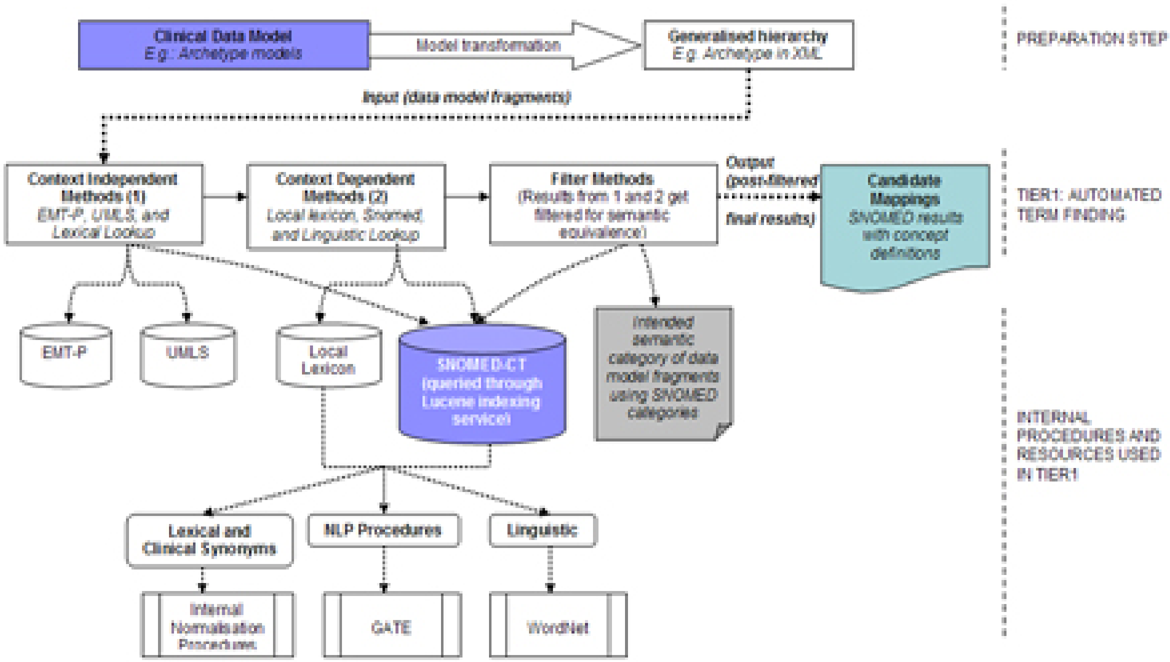


Figure 4.3: The MoST System depicting the Preparation Step, and term finding process along with the internal procedures and resources used to help in term finding.

Once the ‘Generalised hierarchy’ is obtained, the term finding process begins with the search procedure. The search procedure is divided into two main types: (1) searches in which the context is not considered (i.e. context independent), and (2) searches in which the context is considered (i.e. context dependent). As shown in Figure 4.3, various lexical and semantic techniques are applied to both types of searches (explained in detail in subsections 4.3.3.1 and 4.3.3.2). The results from both the context dependent and independent searches are then filtered to retain only semantically equivalent SNOMED codes (discussed in Subsections 4.3.3.3 and 4.3.3.4). The filtered equivalent codes are referred to as ‘candidate mappings’, shown in Figure 4.3. The candidate mapping codes, also known as the MoST results, are presented to a panel of clinical experts to establish those codes that can be finally mapped to the archetype fragments, explained Section 4.3.4 below.

4.3.3.1 Step 1: Context Independent Methods

The context independent methods take as input the archetype fragments, without taking into consideration its context information. This means that the enclosing fragments (for definition see Section 2.7.3 Page 82) and annotations, i.e. term definitions, are not taken into consideration when looking for appropriate SNOMED matches. As seen in Figure 4.3, context independent methods include lexical searches sent (a) to the Emergency Medical Text Processing (EMT-P) service [71], (b) to the UMLS Knowledge Server (UMLSKS) [12], and (c) directly to the SNOMED database using the Lucene indexing service [3]. The three services used viz. EMTP, UMLS, and Lucene will be explained later in this section.

Both EMT-P and UMLSKS are used to exploit their NLP techniques. Initially, the two services were used in isolation to determine whether one or the other could be used to return suitable SNOMED concepts. However, it was observed that each returned useful results that the other did not necessarily return. Therefore, it was decided to use both services for wider coverage of results. In addition, the SNOMED data is loaded into the Lucene indexing service to exploit its Query Parser Syntax when sending search strings. The SNOMED data is also loaded into a MySQL database to extract the concept definitions once the SNOMED concepts have been returned by Lucene.

EMT-P

The first external resource used in the context independent search is the EMT-P service. As shown in Figure 4.3, the archetype fragments were sent to the EMT-P service, which is a Natural Language Processing (NLP) system for medical terms. It processes raw text in natural language entries and is used in conjunction with UMLS [71]. Adding this service is beneficial when processing large medical texts but is used in MoST to exploit its NLP techniques. However, relying on the results returned by EMT-P alone is not sufficient. At times, its internal word sense disambiguation techniques returned UMLS CUIs (Unique Concept Identifiers) matching only a part of the original archetype fragment. Usually, such partial matches had little or no semantic significance to the original fragment. Therefore, in addition to the EMT-P service, the original archetype fragments were also sent directly to UMLSKS, to check for any additional CUIs previously not returned by EMT-P.

UMLSKS

The second external resource used in the context independent search is the UMLS Knowledge Source Server (UMLSKS). UMLSKS gives access to the UMLS Metathesaurus. The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them [109]. It is designed for use by system developers, and cross-references various 'source vocabularies', such as biomedical terminologies, thesauri, and biomedical cataloging literature. Only the SNOMED CT terminology concepts present in the UMLS are of interest to this research. Some of the other specific sources of vocabulary are ICD9, ICD10, MedDRA[6], NCI Thesaurus[7], and SNOMED 1982. The UMLS version used in the MoST system is 2006AA. The latest version available for download was 2007AB at the time of writing this thesis.

The UMLS Metathesaurus includes an extensive classification of medical terms from various vocabulary sources into common semantic type hierarchies, accessible through the UMLSKS Semantic Network application [109]. In addition to manually categorised concept hierarchies, it is also very large in size and coverage. This increases the chance of finding a semantically equivalent SNOMED code. UMLS has a large library of over a million concepts and more than 100 controlled vocabularies and classifications [102].

The reasons for using UMLSKS are three-fold. Firstly, the MoST system takes advantage of the NLP techniques employed by UMLSKS and the cross-referencing capabilities of UMLS; especially to SNOMED CT. Figure 4.4 demonstrates the use by the MoST system of the cross-referencing facility, which helps in extracting the SNOMED codes from a UMLS CUI. Secondly, querying UMLSKS returns certain useful results that might not be returned by EMT-P. Finally, the semantic categories to which the SNOMED concepts belong in UMLS are helpful in adding to the semantic intelligence of the MoST system.

Lucene and MySQL

[6]Medical Dictionary for Regulatory Activities
[7]National Cancer Institute Thesaurus

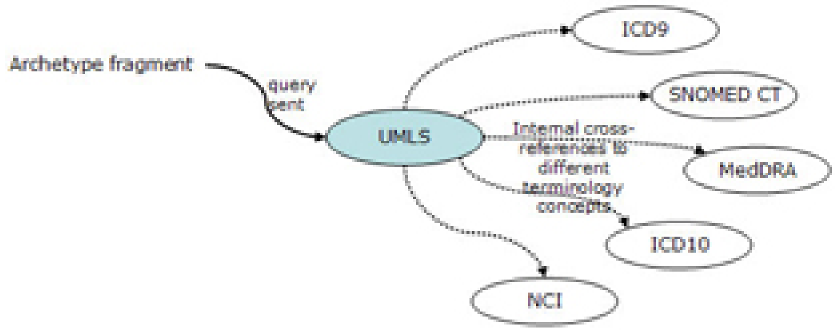


Figure 4.4: Demonstration of use by MoST system of the cross-referencing facility of UMLS to retrieve appropriate SNOMED codes for an archetype fragment.

In the last phase of the context independent search, the fragments are also queried directly to SNOMED, to look for appropriate lexical matches. At the start of the implementation, the July 2005 version of the released SNOMED data was utilised in the term finding and data mapping processes. However, with further releases, the database was updated to include the most recent version. At the time of the last system update, the July 2006 release of the SNOMED data was the latest version available. Therefore, all results obtained from SNOMED belong to the July 2006 data release (which is released along with the UK extensions in October in the UK region). The *descriptions* table in SNOMED, which contains approximately 1.3 million labels for the 370,000 unique concepts, is searched to find the top 35 unique lexical matches for each of the archetype fragments.

As mentioned at the start of Section 4.3.3.1, the Lucene indexing service was used to index the entire SNOMED *descriptions* table. Lucene provides the ability to create custom queries through its rich query language, which is interpreted by the Query Parser [3]. The query syntaxes used by the MoST system include Boolean operators (OR, AND, NOT), Wildcard Searches (?, *), and Exact Searches (+). MoST also takes advantage of the scores assigned by Lucene to the search results. The scores range between 0 and 1, where 1 indicates the highest level of similarity with the search string (i.e. the archetype fragment). For the research, Lucene results with a score above 0.7 were taken into consideration. This was because after manual analysis, it was deduced that results with a Lucene scoring below 0.7 were almost always insignificant for consideration by MoST.

On obtaining the top 35 unique SNOMED codes using the Lucene service, the SNOMED tables loaded in the MySQL database were queried to extract the concept hierarchies and definitions of the 35 concepts. Extracting the concept details helps in analysing the semantic equivalence of the SNOMED code to a specific archetype fragment. SNOMED was queried in MySQL using the Structured Query Language (SQL). Lucene and MySQL were queried once again for SNOMED codes in the context dependent search as well.

Results obtained from all the three context independent methods, i.e. EMTP, UMLS, and Lucene/MySQL, were collated and retained in a temporary results file. The contextual information of the archetype fragments were then included to expand the search base.

4.3.3.2 Step 2: Context Dependent Methods

The context dependent methods use information about the enclosing archetype fragment and term definitions in the archetype model. Taking into consideration the enclosing fragment enables MoST to determine the context in which the fragment is used in the archetype. The annotation or term definition provides detail about the term itself and its intended use. Therefore, the search base expands on inclusion of the context information increasing the chances of obtaining an exact or highly equivalent SNOMED match.

The context information sent to a series of local normalisation procedures. As shown in Figure 4.5, the normalisation procedures involve various lexical and linguistic manipulation techniques. Some of the lexical and linguistic resources have been developed locally while others are accessed from external sources. The local normalisation procedures are not used with the context independent search, as the normalisation techniques of the EMT-P and UMLSKS external resources are made use of instead.

Local Lexical and Linguistic Techniques

The first six methods i.e methods 1-6 in Figure 4.5 include local normalisation procedures. The local linguistic technique includes a search for clinical synonyms in the Local Lexicon, shown in Figure 4.3. The Local Lexicon contains a list of clinical abbreviations and acronyms collected during the development work. Synonyms and abbreviations available from the "word-equivalents" table of the SNOMED supplementary data are also appended to this Lexicon to expand the repository. In addition to the clinical terms, the local lexicon also included non-clinical expressions such as *meal time*, *frequent*, and *cannot*, which are often used to complete a clinical recording. The local lexical techniques applied are to split the query fragments to two or more queries especially when there are phrases in parentheses, and to compress the query fragments to include at first 50% and then 75% of the complete fragment. In addition, the other local lexical techniques applied are removal of stop words (i.e. list of words to exclude from a query fragment when searching through SNOMED), replacement of numerics and conjunctions with words, and removal or replacement of special characters (such as _-() etc.) and arithmetic notations (such as =+-* etc.). The stop words are taken from the list of SNOMED stop words available in the "excluded-words" table of the SNOMED supplementary data.

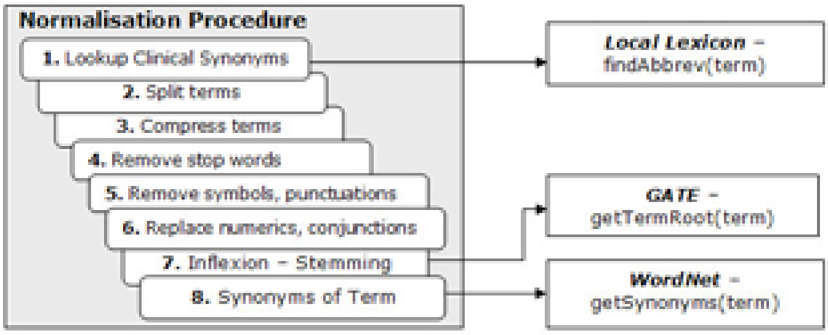


Figure 4.5: Normalisation procedures adopted for the term finding process in the MoST system. (i) Methods 1-6 are local procedures, while (ii) methods 7-8 are external procedures.

External Lexical and Linguistic Techniques

The last two methods i.e. methods 7 and 8 in Figure 4.5 include external normalisation procedures. The external lexical resource used is GATE [35]. One of the word sense disambiguation features of GATE includes detecting the inflexion of terms. This helps in reconstructing queries by extracting the *root* form of the query fragment, also known as *stemming*. For instance, the root form of the fragment 'crying' is 'cry', 'disambiguation' is 'disambiguate', and so on. The external linguistic resource used is WordNet [40], which helps in expanding the search base by including English language synonyms of the query fragment. For instance, WordNet returned the following synonyms for the fragment 'disambiguate' viz. 'clarify', 'clear up', and 'elucidate'.

The modified search queries obtained as a result of the normalisation techniques were sent to the Lucene index to get back SNOMED matches, as explained earlier in Section 4.3.3.1. The results from all the context dependent searches were appended to the same temporary results file that contained the context independent results. The syntax chosen for storing the SNOMED concepts and their definitions was the XML language. This format was selected, as it was easy to store the results in a structured form, which could be easily rendered by the Archetype Editor for graphical display.

4.3.3.3 Step 3: Filtering Process

On completion of the context independent and context dependent searches, the results were subjected to elimination through filtering. In Chapters 5 and 6, the term 'pre-filtered results' is used to refer to the log results stored in the 'log files', whereas the term 'post-filtered results' is used to refer to the final results stored in the 'final file'. The filtering process is the main part of the MoST system. Once all possible SNOMED codes had been returned, it was important to discard all the nonequivalent results before presenting them as candidate mappings to the clinical experts evaluating the results.

Initially the semantic types to which the archetype fragments belonged in the UMLS were taken into account to provide some external information. Optionally, MoST uses a second semantic source to further refine the filtering process by adding an extra layer of semantic knowledge about the fragment(s). It includes the categorisation of the archetype fragment(s), using one or more SNOMED categories to denote the intended meaning, shown as the 'intended semantic category of data model fragments using SNOMED categories' module in Figure 4.3. These SNOMED categories should ideally be specified by the author of the archetype. However, as it was impractical to repeatedly approach the archetype authors for their opinions on each of the chosen archetypes, the opinions of the experts were taken into consideration instead, during the research.

The clinically qualified experts were asked to categorise each fragment or its enclosing archetype fragment, of the four development archetypes, into one or more SNOMED categories, as shown in Figure 4.6. For example, Figure 4.6 shows the 'intended meaning' of the archetype fragment *cardiovascular system* in the Autopsy archetype using SNOMED categories. The *cardiovascular system* fragment was categorised as a *kind_of* `body structure`, `observable entity`, and/or `clinical finding`. This means that during the filtering process, MoST needed to retain only those SNOMED codes that belong to either one or all three categories depending on the rules stated by the experts. SNOMED concepts belonging to any other category could then be eliminated from the result set. Adding this extra layer of knowledge enables search mechanisms, such as MoST, to improve accuracy of performance.

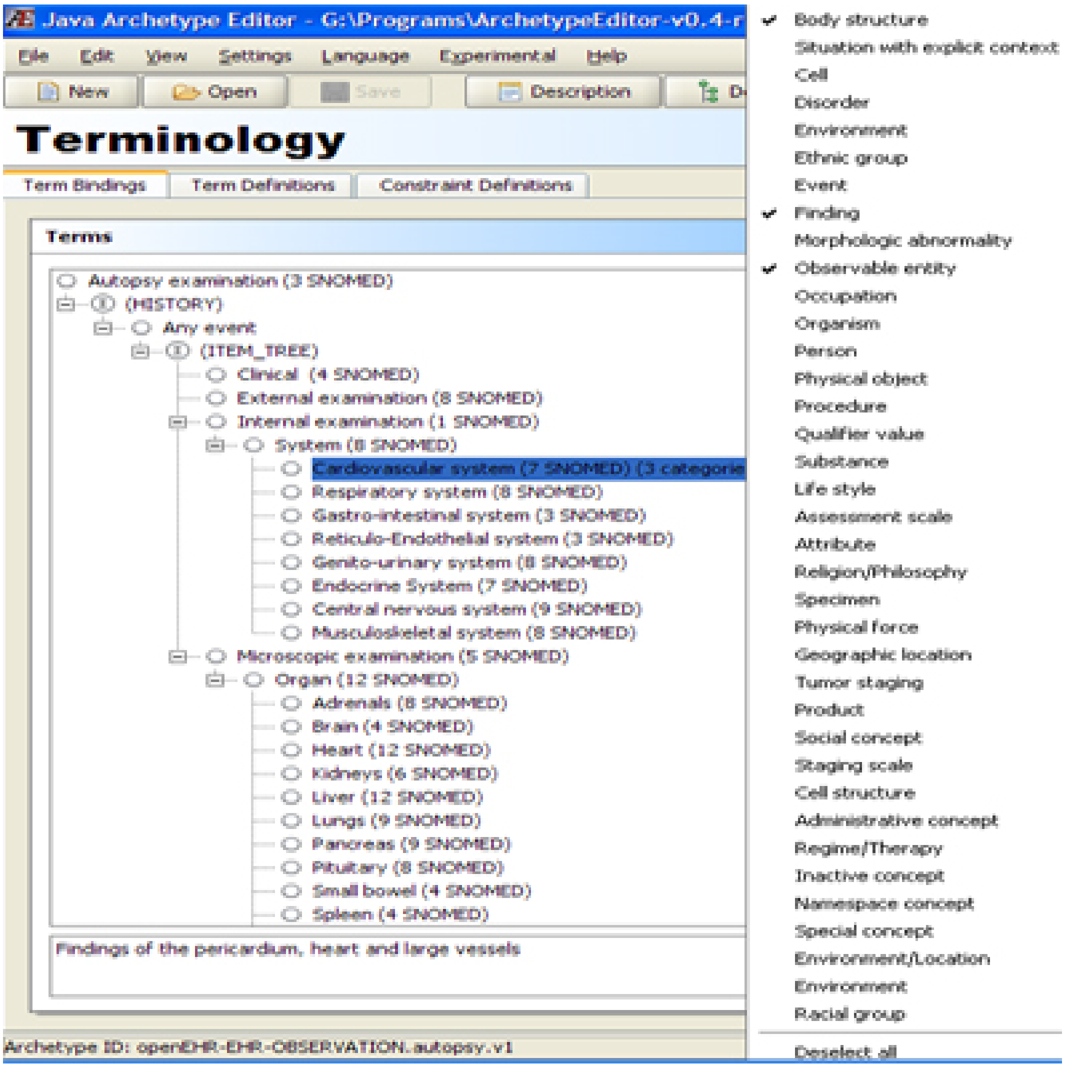


Figure 4.6: SNOMED categories assigned to the archetype fragment *cardiovascular system* to indicate intended meaning.

4.3.3.4 Step 3a: Filtering Rules

The filtering rules, which are applied to the collated results from the context independent and dependent search procedures, are based on certain Description Logic features and axioms such as subsumption and disjointness. In addition, the filtering rules are also based on the SNOMED categories assigned to the archetype fragments in the 'intended meaning' document, explained earlier in Section 4.3.3.3.

The three main rules that are applied to the results are exemplified using the (A) *autopsy examination*, and (B) *blood gas assessment* archetypes. All the examples shown below are in the format <Concept_Name (Concept_Category)[Concept_Code]>, and have been taken from the SNOMED July 2006 data release. This format in SNOMED is used to denote the official name used to represent a concept and is known as the *Fully Specified Name (FSN)*. The subsumption or *is_a* relationships of the FSN's have also been taken from the SNOMED hierarchy.

The MoST results for the Autopsy archetype are also highlighted during the discussion of the three filtering rules. The Autopsy archetype example helps to demonstrate the number of SNOMED codes that are retained at the end of the application of each of the three filtering rules. Table 4.1 presents a summary view of the number of SNOMED codes that were present at the start and end of the filtering process. It can be seen that from a total number of 386 SNOMED codes returned by MoST initially, only 186 codes were retained at the end of application of the three filtering rules discussed below.

Results for Autopsy archetype	
Filtering Stage	**Total num. of SCT Codes**
Before filtering	386
After filtering	186

Table 4.1: Total number of SNOMED (SCT) codes that were returned by MoST before and after the filtering process.

Rule 1: *Filtering based on subsumption* - If three or more children of the same parent concept are present in the result set, then all the child concepts are eliminated retaining only the parent. The intention of this rule is to avoid having too many specialised results while at the same time allowing for a reasonable number of them (less than three). This means that if less than three child concepts are present in the result set along with their parent, then all children are also included in the final result set along with the parent. However, if three or more child concepts are present along with their parent concept, then only the parent concept is retained in the final result set. A demonstration of Rule 1 has been explained below with the help of an example.

Example: Filtering input = {Respiratory system structure, Entire respiratory system, Lower respiratory system structure, respiratory system subdivision, respiratory tract structure} for archetype fragment 'respiratory' in archetype (A).

Entire respiratory system (body structure)[278197002]
is_a Respiratory system structure (body structure) [20139000]
is_a Entire body system (body structure) [278195005]

Output = {Respiratory system structure}

In addition to the subsumption hierarchy to which the SNOMED code `278197002` belongs, the SNOMED concepts `Lower respiratory system structure, respiratory system subdivision, respiratory tract structure` also belong to the same hierarchy with a common parent `Respiratory system structure`. Therefore, the parent code `20139000` is retained whilst eliminating all the child concepts (more than three child concepts).

A demonstration of the impact of Rule 1 on the MoST results returned for the Autopsy archetype is shown in Figure 4.7. The three charts A, B, and C in Figure 4.7 graphically presents the distribution of the total number of SNOMED codes that did and did not have any children codes in the MoST result set for the Autopsy archetype. Chart A shows that of a total number of 386 unfiltered MoST results, 112 SNOMED codes had no child codes in the result set. However, 274 SNOMED results were a combination of parent and child codes. Chart B expands on the 274 SNOMED results to display that 40 SNOMED results had at most one child code in the MoST result set, 32 SNOMED results had at most two children codes in the result set, and 24 SNOMED results had at least three or more children codes in the result set. Finally, Chart C shows the number of child codes that were present in each of the three categories graphically presented in Chart B. Therefore, there were 40 child SNOMED codes in the 1 child category, 64 children SNOMED codes in the 2 children category, and 74 children SNOMED codes in the 3 or more children category. Therefore, on application of Rule 1, the 74 children of the 3 or more children category are eliminated. Only the 24 parent SNOMED codes are retained in the post-filtered result set to avoid too many specialised/granular results.

At the end of application of Rule 1 on the Autopsy archetype, 74 SNOMED codes were eliminated, as explained with the help of Figure 4.7. Therefore, of the 386 unfiltered results, 312 (=386−74) SNOMED codes were applied to Rule 2.

Rule 2: *Filtering based on the disjoint axiom* - If the concepts in the result set are disjoint they are all selected in the post-filtered results. Since ‘disjointness’ is not asserted in SNOMED, it has been inferred from the names of the concepts and the absence of common children. The intention is to avoid accidental elimination of any concept that might have some semantic equivalence to the archetype fragment. As a consequence, the filtering rules include all disjoint results so that subsequent semantic procedures can then be applied to determine their equivalence.

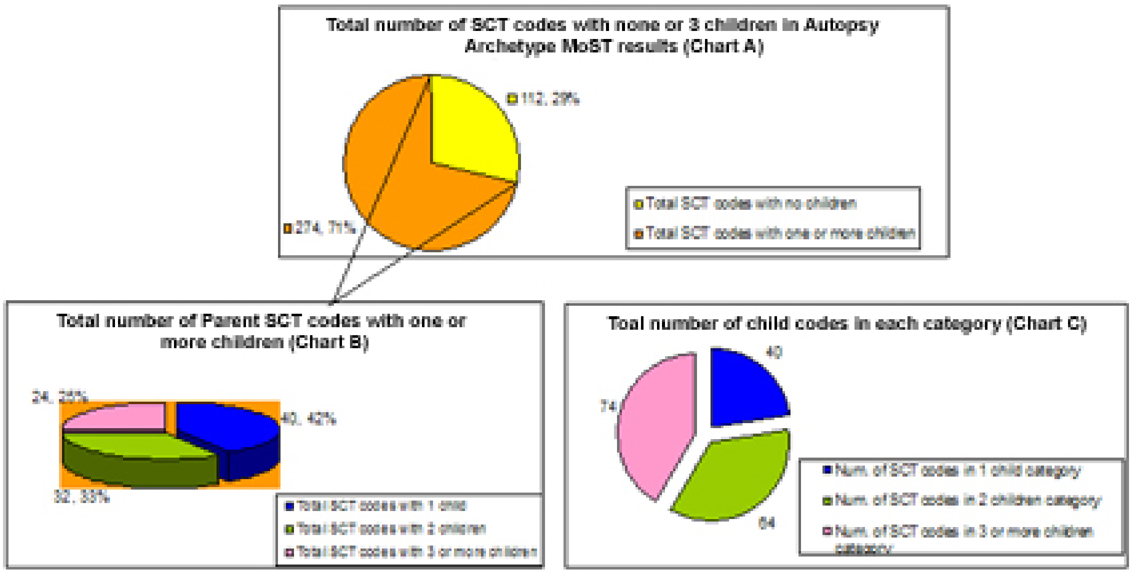


Figure 4.7: Rule 1 Stage - The three charts present the distribution of the SNOMED results returned for the Autopsy archetype before the application of the filtering rules. Chart A presents the number of SCT codes that were returned with (274 codes) and without (112 codes) child codes. Chart B presents the number of parent SCT codes that have one (40 codes), two (32 codes), and three or more (24 codes) children. Finally, Chart C presents the number of children codes in each category of Chart B.

Example: Filtering input = {Hydrogen ion concentration, Past history of, ph+, pH measurement arterial} for archetype fragment 'pH' in archetype (B).

Hydrogen ion concentration (observable entity) [27327002]
  is_a Fluid observable (observable entity) [396277003]

Past history of (situation) [4908009]
  is_a Situation with explicit context (situation) [243796009]

ph+ (qualifier value) [264723009]
  is_a skin reaction grades (qualifier value) [277284009]

pH measurement, arterial (procedure) [27051004]
  is_a pH measurement (procedure) [81065003]

Output = {Hydrogen ion concentration, Past history of, ph+, pH measurement arterial}

Since the results satisfy Rule 2, all the SNOMED results are selected at this stage of the filtering process.

In the case of the Autopsy archetype, Table 4.2 displays the impact of Rule 2 on the filtering process. The two main elimination stages in Rule 2 are the 'similarity in concept names', and 'absence of common children', which have been used as yardsticks to measure the disjointness of two SNOMED concepts. At the end of application of Rule 2, the number of SNOMED codes that are retained are 280 (=110+170), as shown in Table 4.2.

	Results for Autopsy archetype at Rule 2 Stage: Disjoint axioms			
SCT Category	**SCT codes at end of Rule 1**	**At end of *Concept name similarity* process**	**At end of *Absence of common children* process**	**Total codes remaining**
SCT codes with no child codes	112	110	N/A	110
SCT codes with child codes	200	190	170	170
	Total Remaining SCT Codes at end of Rule 2:			280

Table 4.2: Total number of SNOMED (SCT) codes that were retained at the end of application of Rule 2 for the Autopsy archetype.

Rule 3: *Filtering based on semantic knowledge* - Results which pass the filtering Rules 1 and/or 2 are subsequently applied to Rule 3. This rule applies all semantic information available about an archetype fragment when determining the semantic equivalence of a SNOMED concept. The two sources of semantic knowledge are (i) the SNOMED categories assigned to the archetype fragments in the 'intended meaning' document, and (ii) the context in which the fragment is modeled in the archetype.

The 'intended meaning' document that lists the SNOMED categories to which an archetype fragment might belong, is supplemented with the list of 'default' SNOMED categories most common to *open*EHR observable archetypes. These 'default' categories are arrived at by manual inspection of the categories to which most of the archetype fragments map. They are observable entity, procedure, finding, disorder, situation (previously known as context dependent category), and qualifier value. Therefore, in addition to filtering on the basis of the 'intended meaning' categories, SNOMED codes which belong to the 'default' category list are also included.

A final check is made by the filtering mechanism to ensure that the SNOMED concept conforms to the archetype fragment context. As explained earlier in Section 4.3.3.2, the context is determined from the enclosing archetype fragment and the term definition. For example, in the 'pH' example used in Rule 2, the intended meaning of the fragment is a *finding* or *procedure*. In addition, the 'default' list includes the *observable entity* SNOMED category, amongst others. Applying this knowledge along with the context information leads to the following output.

Output = {Hydrogen ion concentration (observable entity), pH measurement arterial (procedure)}

The `Hydrogen ion concentration (observable entity)` code `27327002` is selected although it is not an 'intended' category, as it belongs to the 'default' category. In addition, the context information also justifies its selection. The definition of the term 'pH' in archetype (B) states that it represents *the negative logarithm of the Hydrogen ion concentration in the blood* when recording the *arterial blood gas readings*.

Finally, the Autopsy archetype example demonstrates the impact of the application of Rule 3 to the remainder of the SNOMED codes. The 280 SNOMED codes retained at the end of Rule 2 are checked against their 'contextual' relevance to and 'intended meaning' of the archetype fragments. Table 4.3 displays the number of SNOMED codes that are retained at the end of each of the semantic filtering stages of Rule 3.

	Results for Autopsy archetype at Rule 3 Stage: Semantic Filtering			
SCT Category	**SCT codes at end of Rule 2**	**At end of *Contextual relevance* process**	**At end of *Intended meaning* process**	**Total codes remaining**
SCT codes with no child codes	110	95	68	68
SCT codes with child codes	170	144	118	118
	Total Remaining SCT Codes at end of Rule 3:			186

Table 4.3: Total number of SNOMED (SCT) codes that were retained at the end of application of Rule 3 for the Autopsy archetype.

At the end of the application of Rule 3, which is also the last filtering rule in the MoST system, 186 SNOMED codes are retained. The final result set conforms to the filtered SNOMED codes for the Autopsy archetype shown earlier in Table 4.1 (Page 112). These 186 SNOMED codes are presented to the clinical evaluators to determine the most semantically equivalent results that can be mapped to the archetype fragments.

The filtering process concludes the automated *term finding* process, which is the first tier of the MoST methodology, presented in Figure 4.1 above.

4.3.4 Tier 2: Assisted Manual Data Mapping Process

The second tier of the methodology entails the assisted manual evaluation of the MoST results by a panel of clinical experts. During the development of the MoST system, four development archetypes (stated in Section 4.3 above) were used to perform a preliminary experiment with the help of three clinical researchers present in the research group. Results and feedback obtained from this experiment helped in refining the actual evaluation and mapping process, which was later performed with the help of a panel of experts. The actual evaluation exercise and the MoST results are discussed in detail in Chapters 5 and 6. This section provides only an overview of the mapping process.

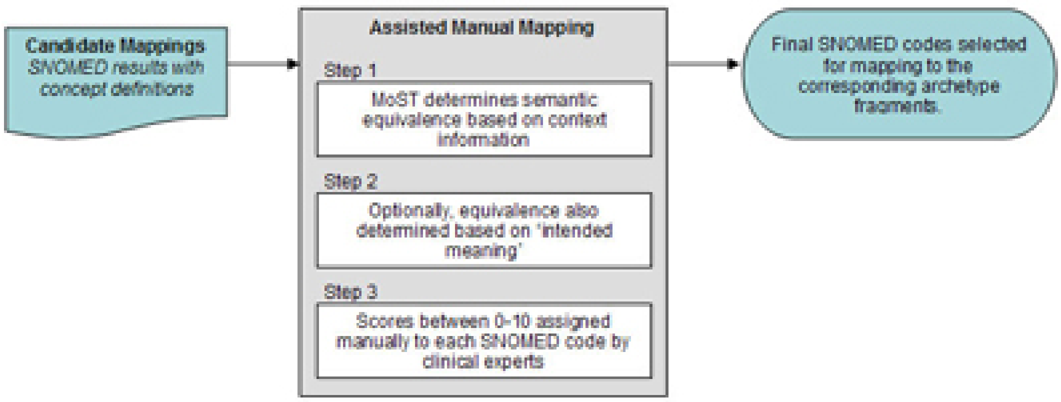


Figure 4.8: Assisted manual mapping process of the results obtained from the term finding process performed by the MoST system.

The center rectangular box in Figure 4.8 displays the three key steps of the mapping process. These are (i) determining the semantic equivalence of the SNOMED codes on the basis of the context information, (ii) optionally, determining the equivalence of SNOMED codes based on the SNOMED categories assigned to the archetype fragments, as 'intended meaning', by the experts at the 'Preparation Step' (see Section 4.3.2), and finally (iii) manually scoring the SNOMED codes on a scale of 0-10 by a panel of clinical experts.

The first two steps are performed by the MoST system and are, therefore, automated. The filtered SNOMED codes are subjected to a last stage to determine their semantic equivalence to the archetype fragments. All available context information of both the archetypes and SNOMED are taken into account once again. In addition, the SNOMED categories used to optionally categorise the archetype fragments, are also used to determine the most semantically relevant SNOMED code(s) to map to the fragments.

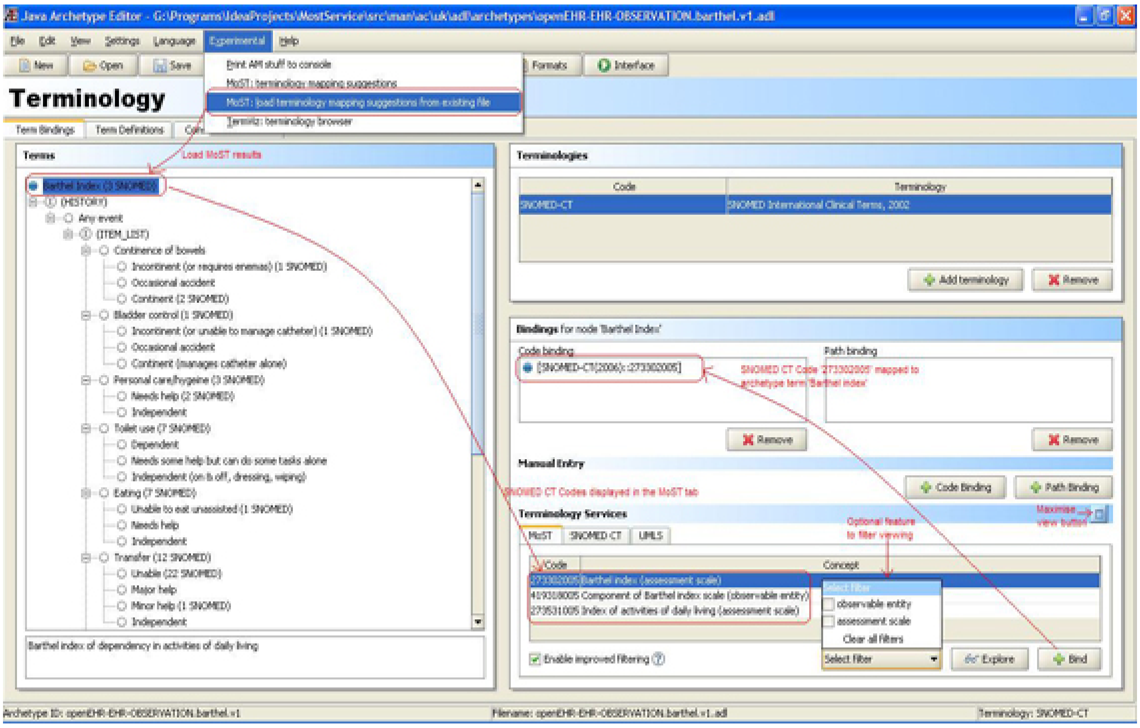


Figure 4.9: Mapped codes in the MoST interface of the LiU Archetype Editor. The SNOMED code 273302005 is mapped to the archetype fragment ‘Barthel index’.

The final step of the mapping process is to get a panel of clinical experts to individually score the SNOMED codes on the basis of their semantic equivalence to the corresponding archetype fragment. The scores range between 0 and 10; 0 being completely nonequivalent and 10 being an exact match. This means that the higher the score given to a SNOMED code, the higher its semantic equivalence to the archetype fragment. Further details on the scores given to archetype fragments have been discussed in Chapter 5.

At the end of the mapping process, the scores assigned to the SNOMED results are cumulated to determine the most equivalent codes to map/bind to the archetype fragments. The number of archetype fragments that find a SNOMED code to which it could map, indicates the *mapping coverage* of the entire archetype. In addition, the confidence the experts placed on the overall semantic equivalence of the SNOMED codes for an archetype model is indicated by the *trust score*. The mapping coverage and trust score, along with the other statistical measures used in the evaluation, are discussed in Section 5.3 on Page 125.

An example of how the mapped codes appear in the LiU Archetype Editor is shown in Figure 4.9. The screenshot displays the barthel index archetype loaded in the MoST interface. In this example, the SNOMED code `Barthel index (assessment scale [273302005]` scored the highest amongst the experts. Therefore, this code was mapped to the archetype fragment `Barthel index`.

Figure 4.9 also highlights some visual display buttons, such as 'maximise view' and 'filter view', which have been explained in Appendix B (Page 232). The mapped codes are intended to be recorded both in the archetype model, as well as the Electronic Health Record (EHR), to aid in semantic interoperability.

4.4 Summary

This chapter discussed the MoST methodology developed for the integration of models at content level. Discussion of the MoST methodology included the two-tier system viz. (i) the *term finding* process performed by the MoST system, and (ii) the assisted *data mapping* process performed with the help of clinical experts evaluating the results from the first process (i.e. term finding).

The chapter detailed the MoST system process and discussed the various components that formed part of the application. A brief introduction was also provided to the key phases of the data mapping process. Details regarding the MoST results and the evaluation exercise have been provided in Chapters 5 and 6. The following chapter 5 will present the results from the scoring system used by the experts to assess the equivalence of the results for the four evaluation archetypes. It will also present the rate of mapping coverage and the trust score achieved for each of the four archetypes.

Chapter 5

Evaluation I: Quantitative Analysis

The previous chapter described the MoST methodology for finding codes in SNOMED to bind/map to the archetype fragments. The results from the MoST system, which was developed to implement the methodology, were then evaluated both quantitatively, as well as qualitatively. This chapter presents a quantitative analysis of the evaluation. However, the thesis focuses primarily on the qualitative analysis (discussed in Chapter 6). The quantitative analysis of the results presented serious difficulties because of (a) the lack of a gold standard, and (b) the irregular structure of the data in the two models, which meant that the assumptions of established statistical procedures could not be applied completely. Nonetheless, a basic quantitative analysis was attempted, and is described in this chapter.

The quantitative results are based on the scores given to the SNOMED codes returned by MoST (i.e. the MoST system). The scores were given by a panel of clinical experts evaluating the MoST results. The evaluation criteria and methodology, along with the evaluation results are presented in this chapter, which is the second in the four-chapter series on the discussion of the research programme.

This chapter begins with a description of the evaluation criteria and methodology that was designed to perform the analysis and evaluation of the MoST system. The emphasis will be on analysing the scores assigned by the evaluators during the evaluation. These scores help determine the six most important measures i.e. the rate of true positives, false positives, true negatives, false negatives, mapping coverage, and trust score. A seventh measure viz. the inter-rater reliability has been calculated for only the histology pap and tendon babinski archetypes, due to limitations in the approach with respect to the research. Each of these measures will be defined in Section 5.3 of the chapter. The six main measures will be calculated for each of the four evaluation archetypes, and presented comparatively.

In addition to the six measures, Appendix B (see Section B.4) presents the three categories of results that were aggregated specific to the SNOMED results for each of the four archetypes. The three categories were (i) top five scoring SNOMED codes, (ii) post-coordinated SNOMED codes, and (iii) SNOMED codes missed by MoST. The evaluators will be referred to as either 'experts', 'clinical experts' or 'evaluators' in the thesis.

5.1 Evaluation Criteria

The criteria for evaluating the research work describes the need for:

1. *Qualified evaluators:* The study requires that the evaluation of the MoST results is performed by professionals from the field of health care. These health care professionals might include general practitioners, hospital doctors, nurses, or medical researchers. However, the list is not limited to the above. Essentially, qualified personnel were required, as the study was primarily based on assessing the semantic equivalence of clinical codes, which required a thorough understanding of the medical field. More details on the selection of the evaluators is available in Appendix B (see Section B.1).

2. *Sufficient number of evaluators:* Each archetype is required to be evaluated by a reasonable number of evaluators. This criterion is required to obtain an unbiased rating and feedback of the MoST results.

3. *Range of Archetypes:* In order to get a fair coverage of the potential of the MoST system, a range of archetype models are required. Therefore, when selecting the archetypes for evaluation, the right mix of large and small archetypes must be selected (i.e. many and few fragments per archetype). In addition, it is also important to ensure that archetypes, which cover both clinical as well as non-clinical fragments, are selected, to test their coverage in SNOMED. This also helps in assessing the ability of MoST to cope with any kind of information passed to it.

4. *Accessibility:* The MoST results should be easy to view by the evaluators. A pre-existing or internally developed graphical interface is required, to enable the

evaluators to view in one place the archetype as a whole, its fragments, as well as the SNOMED concepts returned by MoST. However, the interface should easily display the archetype fragment hierarchy, along with its definitions/annotations, if any. SNOMED concept definitions and their partial hierarchy is also required to be displayed, to help the expert in taking an informed decision.

5. *Quality:* A reasonable basis for determining the quality of the MoST results is required. Two ways to ascertain the quality of the results are (i) to enable the evaluator to gain access to the pre-filtered results i.e. those SNOMED codes that have been eliminated by MoST as being irrelevant/nonequivalent, and (ii) to provide the option to the evaluator to browse through the SNOMED database directly, to verify the omission by MoST of any relevant results.

6. *Suitable statistical measures:* Suitable statistical measures are required to calculate the rate of true and false positives, true and false negatives, mapping coverage, trust score, and inter-rater reliability. Since, there are several axes for measuring the evaluation results, it is anticipated that established metrics might be difficult to adopt. In such an event, the most suitable measure(s) might have to be developed internally, to reflect the most appropriate statistical findings. In addition, adequate comparative analysis of the statistical results achieved for all the evaluation archetypes must be provided.

7. *Unbiased Qualitative Review:* Based on the term finding and data mapping processes, as well as the evaluators feedback, an unbiased, summative review is also required. The review should focus on the general qualitative issues concerning archetypes, SNOMED, and MoST. Suggestions might also be made on inclusions and improvements that can be made to improve performance and results of model integration techniques.

5.2 Evaluation Methodology

The evaluation methodology divided the evaluation exercise into two main phases. Phase One required the evaluators to score the individual SNOMED codes returned by MoST. Besides being quantitative, it was also partially qualitative, as it required feedback from the experts on the issues that might have arisen during the scoring. Phase Two of the evaluation was mainly qualitative. Besides analysing the scores and determining the quality of the results, it also focused on the factors in the Archetypes and SNOMED models which limited their performance.

Four archetypes from the list of published *open*EHR Archetypes [6] were selected in Phase One of the evaluation. The list of published archetypes has seen a significant growth in the number and scope, ever since they were first accessed for evaluation in August 2006. The four archetypes will be referred to as the 'evaluation archetypes' and are listed in Subsection 5.2.1 below. The criteria for their selection is also discussed in Subsection 5.2.1.

A total of 14 evaluators took part in the evaluation exercise. Each expert was assigned a maximum of two archetypes to evaluate. A balance was struck in the selection; with one being a large and the second being a comparatively smaller archetype. The size of an archetype was determined by its number of fragments. Typically, a large archetype had more than 30 fragments. There were two main reasons for having a combination of one large and one small archetype. Firstly, *Time constraints* - The evaluation was done with busy professionals, who could devote a maximum of two hours only. Secondly, *Motivation* - It was perceived that having two large archetypes would discourage the experts to pursue with the evaluation, which was anyway a tedious exercise.

A detailed description of the evaluation setup process for conducting Phase One of the evaluation is available in Appendix B (see Section B.2). It explains the MoST interface integrated to Linköping University's Java Archetype editor[52]. It also provides screenshots of the evaluation sheets used in the exercise, with a description of the significance of the various scoring requirements.

The scores assigned by the evaluators to the SNOMED codes returned by MoST were in the range of 0-10 (i.e. Zero-to-Ten). A zero score indicated that the SNOMED result had no remote equivalence to the archetype fragment. However, it was not required to manually assign a score of zero for each nonequivalent match. All SNOMED concepts with no scores assigned by the experts were assumed to have zero equivalence. The evaluators were instructed that 'low equivalence' codes were to be scored between 0-3, 'moderate equivalence' codes were to be scored between 4-6, and 'high equivalence' codes were to be scored between 7-10. The evaluators were also instructed to assign a score of 9 or 10 to the most equivalent codes.

5.2.1 The Four Evaluation Archetypes

The four archetypes chosen for the evaluation exercise were (i) histology pap (cervical smear), (ii) visual acuity, (iii) body weight, and (iv) tendon babinski. The intended purpose of the four archetypes as stated by the author(s), along with the reasons for selecting these archetypes, are listed below:

- *Histology Pap archetype:* It was the main archetype evaluated by the clinical experts. The intended purpose of this archetype was to *record the specialised histological findings from a PAP test or cervical smear*, to help diagnose the presence/absence of cervical cancer or any other abnormalities.

  The reason for selecting this as the main archetype was due to the large number of fragments that were unlikely to find a direct SNOMED match either due to their non-specific labeling or non-clinical semantics. The lack of direct matches was intended to test the MoST threshold for finding semantically equivalent SNOMED codes despite the complexity and/or ambiguity of the archetype fragments. The largest group of 10 experts were engaged in the evaluation of the histology pap archetype.

- *Visual acuity archetype:* The second archetype that was selected for evaluation was the visual acuity archetype. Its intended purpose was to *record the functional acuity/sharpness of vision with and without the use of visual aids*, to help diagnose and evaluate any visual impairment.

  The main reason for selecting this archetype was to test the capabilities of the MoST system to deal with fractions. Majority of the fragments which represented the different visual acuity scores were of the form 'x/y'. The visual acuity archetype was more straight-forward, as compared to the histology pap archetype, thereby enabling the evaluators to test the equivalence of more than one semantically equivalent SNOMED code for the same archetype fragment. Since the evaluators had to choose between very similar codes, it made the evaluation task more difficult. However, it highlighted the importance of having well-annotated archetype fragments as well as SNOMED concepts, to help make an informed decision. The archetype was evaluated by a total of 5 evaluators.

- *Body weight archetype:* The third archetype chosen for evaluation was the body weight archetype. The author intended to utilise this archetype to *record the whole body weight of a neonate or fetus*, to diagnose the general health and well-being of a newborn. The author did not intend to record "the weight of any other entity or part of the body".

The reason for selecting this archetype was its small size, which helped to adhere to the 2 hour evaluation time limit. This archetype comprised of only 12 fragments, which were initially considered to be less ambiguous and hence less time-consuming. The archetype was evaluated by 4 experts and their feedback and scores were recorded for further analysis.

- *Tendon Babinski archetype:* The last archetype to be evaluated was the tendon babinski archetype. It was intended to *record the tendon reflexes and babinski response* as part of a neurological assessment.

Archetype fragment	**Term Definition**
+/-	Reflex possibly present - markedly reduced amplitude
+	Reflex present but reduced amplitude
++	Reflex normal
+++	Reflex increased
++++	Reflex markedly increased

Table 5.1: Symbols used to measure degree of reflexes in the *tendon babinski* archetype and the corresponding lexical annotations.

The main reason for selecting this archetype was to test the ability of the MoST system to resolve arithmetic notations used as symbols. These symbols were used to denote the degree of reflex observed in the tendons, such as +/-, +, ++, +++, and ++++. Table 5.1 presents the annotations/term definitions assigned to each of the symbols in the archetype. The MoST system replaced the symbols (+/-,++, and so on) with their definitions, whilst also considering the context (i.e. body part) in which the degree of reflex was being measured. The archetype was evaluated by 4 clinical experts.

5.3 Statistical Measures for Evaluation

Initially, two statistical measures were considered important for measuring (i) how well MoST did in terms of the clinicians (i.e. evaluators) view; using the sensitivity, specificity, and mapping coverage ratings, and (ii) how well the clinicians agreed amongst themselves. The measure of agreement amongst the clinicians was based on the confidence they placed on the SNOMED codes selected for mapping, known as the 'trust score'. In addition, the agreement was also based on the general consensus reached amongst the clinicians on the overall relevance of the SNOMED codes, known as the 'inter-rater reliability' and measured using Fleiss' Kappa.

5.3.1 Sensitivity and Specificity

The scores accumulated at the end of the first phase of the evaluation were analysed and categorised into four main types. These were the true positive results (TPR), false positive results (FPR), true negative results (TNR), and false negative results (FNR). The sensitivity and specificity ratings were determined on the basis of the TPR and FNR, respectively. A standard measure such as ROC curves [4][63] could not be used for plotting the TPR and FPR of all the four archetypes cumulatively, as the archetypes differed in their total number of fragments. The ROC curves could not be plotted for individual archetypes either as it was not possible to track the TPRs and FPRs per SNOMED code returned by MoST. The fundamentals differed, requiring only the total number of SNOMED codes per archetype in each of the four categories (TPR, FPR, TNR, and FNR). Therefore, the four categories have been used in the study to determine the conformance of the MoST results with the evaluators view. A brief description of each of them is provided below.

1. *True positive results(TPR):* These results highlight the number of SNOMED concepts that were correctly retained by MoST in the post-filtered set of results and were considered to have some degree of equivalence (score of 4 or above) by more than one evaluator. The degree of correctness was determined by the clinical experts evaluating the results.

2. *True negative results(TNR):* The true negative results signify the number of SNOMED codes that were correctly eliminated by the MoST system during the filtering process. The true negatives were calculated by subtracting the total number of false negatives present in the pre-filtered results from the total number of pre-filtered SNOMED codes. Therefore,

   True negatives = Total no. of pre-filtered SNOMED codes - Total no. of *false negatives* from pre-filtered SNOMED codes

3. *False positive results(FPR):* These results highlight the number of post-filtered codes, which were not considered to be of much semantic equivalence (score below 4) by the evaluators. Therefore,

   False positives = Total no. of post-filtered SNOMED codes - Total no. of *true positives* from post-filtered SNOMED codes

4. *False negative results(FNR):* The last axis for measurement of scores was the number of false negatives present in the pre-filtered results. These accounted for those SNOMED codes, which the evaluators considered to have some semantic significance to the archetype fragments, although MoST had rejected them during the filtering process. Therefore, the FNR accounted for those codes that MoST had incorrectly eliminated. The false negatives were calculated by totaling the number of codes that had been given an equivalence scoring of 4 and above by more than one evaluator.

Of the four category types, the three most important ones were the TPR, TNR, and FNR[1]. The true positives (TPRs) and true negatives (TNRs) indicated the rate of sensitivity and specificity, respectively. They also indicate the rate of positive and negative predictive values, which are important to the study. Definitions of sensitivity, specificity, positive predictive value, and negative predictive value, are listed below.

- ***Sensitivity:*** Sensitivity refers to how good MoST is at correctly identifying SNOMED codes that are semantically equivalent to a particular archetype fragment. In other words, sensitivity indicates the proportion of results returned by the test, relative to all the results which actually satisfy the test condition. In terms of the research, as shown in Equation 5.1, sensitivity would imply the ratio of the number of SNOMED codes agreed upon by the clinical experts to have semantic equivalence to a particular archetype fragment (i.e. TPR+FNR), to those codes which are returned by MoST in the post-filtered results(i.e. TPR). The equation to calculate sensitivity is shown below.

$$Sensitivity = \frac{TPR}{TPR + FNR} \times 100 \tag{5.1}$$

  where, TPR = True Positive Results, and
  FNR = False Negative Results

- ***Specificity:*** Specificity is concerned with how good MoST is at correctly identifying SNOMED codes that are semantically non-equivalent to a particular archetype fragment. In other words, the rate of specificity indicates the ability of a system to correctly eliminate those results that do not meet the test condition. In terms of the research programme, as shown in Equation 5.2, specificity implies the ratio of the number of SNOMED codes that were correctly identified by MoST as being semantically nonequivalent to a particular archetype fragment

[1] TPR - True Positive Results; FPR - False Positive Results; TNR - True Negative Results; FNR - False Negative Results

(i.e FPR+TNR), to the total number of pre-filtered results (i.e. TNR). This means that both the clinical experts as well as the MoST system agreed that the filtered codes had been rightfully eliminated.

$$Specificity = \frac{TNR}{FPR + TNR} \times 100 \tag{5.2}$$

where, TNR = True Negative Results, and
FPR = False Positive Results

- ***Positive Predictive Value:*** Positive predictive value refers to the chance that a positive test result will be correct [61]. That is, it looks at all the positive test results, as shown in Equation 5.3. In terms of the MoST system, it indicates the number (or rate in percentage) of total positive SNOMED codes returned by MoST, which were correctly identified by the evaluators to be semantically equivalent to a particular archetype fragment.

$$PPV = \frac{TPR}{TPR + FPR} \times 100 \tag{5.3}$$

where, PPV = Positive Predictive Value,
TPR = True Positive Results, and
FPR = False Positive Results

- ***Negative Predictive Value:*** On the other hand, the negative predictive value is concerned only with negative test results [61]. It refers to the chance that a negative test result will be correct, as shown in Equation 5.4. In terms of the MoST system, it indicates the number (or rate in percentage) of the total negative results returned by MoST, which were also identified by the evaluators to have no semantic equivalence to a particular archetype fragment.

$$NPV = \frac{TNR}{TNR + FNR} \times 100 \tag{5.4}$$

where, NPV = Negative Predictive Value,
TNR = True Negative Results, and
FNR = False Negative Results

The positive and negative predictive values are of greater interest to the research study as compared to the sensitivity and specificity ratings. The main reason is that they help in determining the level of agreement between the MoST system and the evaluators in identifying the results that both considered to have been correctly retained (i.e. true positives) and correctly eliminated (i.e. true negatives) during the filtering process.

5.3.2 Mapping Coverage

The mapping coverage rate indicates the percentage of fragments in an archetype covered in the SNOMED terminology. In other words, the coverage indicates the rate of archetype fragments that could be mapped to SNOMED codes because the codes were rated by the evaluators as highly relevant. It is calculated on the basis of the true positive (TPR) and false negative (FNR) results. However, unlike the other quantitative results discussed in the chapter, the coverage rate takes into consideration the total number of "archetype fragments" for the calculation, instead of the total number of "SNOMED codes".

Getting a good mapping coverage for any archetype is dependent on four important factors. These are (i) the existence of unambiguous archetype fragments, (ii) the existence of unambiguous SNOMED CT concepts, (iii) a good coverage of the fragment concepts in SNOMED CT, and (iv) no relevant SNOMED CT codes missed by MoST during the term finding process.

There are two important notations used when determining the coverage rate. A SNOMED code that has been given an equivalence score of 4 or above by at least two evaluators, is regarded as a 'relevant' code. Of these 'relevant' codes, those codes which achieve a total equivalence score of 50% or above, is regarded as a 'correct' code. These 'correct' SNOMED codes are assumed to most likely map to the archetype fragment during the *data mapping* process. Therefore, the two criteria to be satisfied to determine the mapping coverage are:

- *Criteria 1:* Total number of archetype fragments, for which MoST returned a 'relevant' result/code at any stage of the filtering process (i.e. true positive and false negative results).
- *Criteria 2:* Total number of archetype fragments, for which MoST returned a 'correct' result from amongst the 'relevant' results. As mentioned earlier, a 'correct' result was a result that was given a total equivalence score by all the evaluators of 50% or above.

An example of applying Criteria 1 and 2 to the results of the archetype fragment 'cervical smear' is shown in Table 5.2. The two SNOMED codes returned for the fragment are `pap smear due` and `pap smear test`. It can be seen that the concept `pap` `smear due` has been given an equivalence score of 4 by at least two of the five evaluators. Likewise, the concept `pap smear test` has been given an equivalence score of 4 and above by four of the five evaluators. Therefore, both concepts satisfy Criteria 1. This means that both the concepts are 'relevant' matches. However, the total equivalence score of the `pap smear due` concept is only 28%. Therefore, it does not satisfy Criteria 2. This means that it is a 'relevant' but not 'correct' result, thereby indicating that its chance to be selected for mapping is lowered. On the other hand, the concept `pap smear test` has a total equivalence score of 60%, thereby satisfying Criteria 2 of being a 'correct' result.

Results for archetype fragment 'cervical smear'							
SCT code	**1**	**2**	**3**	**4**	**5**	**Total Score**	**Percentage Total (%)**
Pap smear due	2	4	4	3	1	14	28%
Pap smear test	2	4	6	8	10	30	60%

Table 5.2: Example scores of the two SNOMED codes for the archetype fragment 'cervical smear', to illustrate the application of Criteria 1 and 2. Scores are between 0 and 10, and have been assigned by 5 experts.

Based on the two criteria, the mapping coverage rate is calculated as shown below.

$$Coverage = \frac{Criteria2}{Criteria1} \times 100 \tag{5.5}$$

As shown above, the coverage rate is the percentage total of archetype fragments that have a 'correct' SNOMED match from amongst the 'relevant' matches. The 'correct' codes are also referred to as the 'highly relevant' codes. These SNOMED codes are considered to most likely map to the archetype fragments. Therefore, the mapping coverage determines the degree of success achieved by MoST in finding semantically equivalent SNOMED codes to map to the archetype fragments of a particular archetype.

A suggestion was made to calculate the mapping coverage before and after filtering of the SNOMED codes. However, manual evaluation by the experts of all the pre-filtered codes for all four archetypes was not considered feasible due to time constraints. Therefore, the post-filtered results were presented to the experts for careful evaluation, while the remainder of the filtered results were presented for quick evaluation. Manual intervention at both stages of filtering (i.e. pre and post) would have defeated the initial research aim to benefit from automated systems, such as MoST, to perform the tedious manual processes. The MoST methodology assumed that any filtered SNOMED codes could be retrieved from the system logs for mapping. This was because filtered codes were also taken into consideration when calculating the coverage rate. Therefore, the term 'mapping coverage/coverage' will include the coverage of archetype fragments before and after filtering.

5.3.3 Trust Score

The 'trust score' is used to denote the trust or confidence the experts placed on the SNOMED codes returned by MoST. A higher trust score meant that the experts placed higher confidence in the overall semantic equivalence of the SNOMED codes for an archetype model, and vice versa. The trust score was based on the scores given by the experts to the True Positive SNOMED codes (TPR) from the post-filtered results. Only those TPRs were selected, which had an equivalence score of 50% or above (as in calculation of the coverage rate).

Calculation of the trust score was a two step process, as shown in Equations 5.6 and 5.7. First, the median value[2] of the scores collected for the TPRs ($\geq 50\%$) returned for each archetype fragment was calculated, as shown in Equation 5.6. Next, the overall median value was calculated from the individual median scores returned in the first step (i.e. Equation 5.6). This overall median score reflected the trust score, as shown in Equation 5.7.

$$IMS = \underbrace{Median}_{STPR \geq 50\%}(STPR) \tag{5.6}$$

where, IMS = Individual Median Score, and

STPR $\geq$ 50% = Scores of those True Positive Results, which have a total equivalence score greater than equal to 50%.

$$TS = Median(IMS_1, IMS_2, ..., IMS_n) \tag{5.7}$$

where, TS = Trust score, and

IMS_1, IMS_2, ..IMS_n = Individual Median Score for archetype fragment 1, 2,..n.

Besides indicating the level of confidence, the trust score also indicates the quality of the archetype fragments and SNOMED codes. The quality of the fragments and codes are directly proportional to the score. The term 'quality' refers to the semantic equivalence of a SNOMED code, which determines whether or not the code will be

[2] The middle value that separates the greater half from the lower half of the data set

mapped to an archetype fragment. Therefore, a higher trust score reflects higher quality of semantic equivalence between archetype fragments and SNOMED codes.

5.3.4 Inter-rater Reliability

Fleiss' kappa was originally suggested to be used for the research, as the metrics to measure the inter-rater reliability. Although the metrics is best applicable to categorical data, it was expected to provide a reasonable estimation of the agreement for the interval-based, ordinal data used in the research. However, it soon became clear that kappa was not an appropriate measure. Kappa (κ) is defined in terms of a ratio of the observed to the expected agreement [41][42].

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{5.8}$$

where, $\bar{P} - \bar{P}_e$ = degree of agreement actually achieved above chance (i.e. observed),
$1 - \bar{P}_e$ = degree of agreement that is attainable above chance (i.e. expected),

and

$0 \leq \kappa \geq 1$; κ=1 implies complete agreement.

In typical studies using kappa, the subjects observed fall roughly equally into the different categories. However, in this study, best agreement occurred when all the raters agreed that the MoST results were appropriate, so that all raters chose the same category (i.e. either one of the three equivalence score intervals: low (0-3), moderate (4-6), or high (7-10)). As a result of the peculiarity of the research, the kappa tends to zero rather than one, as it assumes a bias amongst all the raters toward a particular category. Such a bias, shifts the values of both $\bar{P} - \bar{P}_e$ and $1 - \bar{P}_e$ toward zero, as it is regarded by kappa as being an *agreement by chance* rather than an *observed agreement*. Therefore, in this situation, it became necessary to consider a subset of the overall kappa, to satisfy the research requirement. The subset of interest is the value of κ_{max} [95].

Sim and Wright state that the interpretation of kappa is assisted by also reporting the maximum value that kappa could attain for the set of data concerned, referred to as the maximum attainable kappa (κ_{max}) [95]. The κ_{max} value represents the greatest possible agreement that can be achieved amongst the raters, and is of particular research interest. Therefore, only the κ_{max} value (as a percentage) will be considered as an inter-rater reliability result in the thesis. The results obtained for the two main evaluation archetypes are discussed in Section 5.5.4.

It is seen from the study that despite kappa being an established metric for the measurement of the inter-rater reliability, there are problems with the overall measure (or kappa) with regards to the research data. This led to the adoption of the max kappa (or κ_{max}) value, which at least provided a rough measure of the maximum agreement that could be achieved amongst the evaluators of the MoST results. However, the measure should be used with caution, as it has been noted that these metrics might be more harmful than helpful depending on the data in some studies [48]. Therefore, no strong inferences should be drawn from this statistic in the thesis.

5.4 Summary of Evaluation Data Results

The summary of the data analysis performed in Phase One of the evaluation of the four evaluation archetypes is shown in Tables 5.3 and 5.4. Table 5.3 summarises the true positives and negatives, as well as the false positives and negatives obtained at the end of Phase One of each of the four archetypes. In addition, Table 5.4 displays the overall coverage rating and trust score achieved for the four evaluation archetypes.

Archetype Model	**True positives**	**False positives**	**True negatives**	**False negatives**
Histology pap	69	108	219	17
Visual acuity	81	69	85	5
Body weight	17	22	48	2
Tendon babinski	200	82	45	17

Table 5.3: Summary of true and false positives and negatives for each of the four evaluation archetypes.

	Histology pap	**Visual acuity**	**Body weight**	**Tendon babinski**
Coverage (in %)	48.6	86.9	70	90.9
Trust score (scale 0 to 1)	0.45	0.8	0.7	0.9

Table 5.4: Summary of coverage and trust score for each of the four evaluation archetypes.

5.5 Evaluation Phase 1: Comparative Quantitative Analysis

The evaluation results obtained at the end of Phase one of the exercise (i.e. the term finding and data mapping processes) is presented in this section in the form of a comparative study of the four evaluation archetypes. The intended purpose is to provide a summary view of the different results calculated based on the scores assigned by the evaluators. Individual discussion of the issues faced with each of the four archetypes will be presented in Chapter 6. In addition, Appendix B (see Section B.5) has a detailed data distribution table for each of the four evaluation archetypes. These tables display the distribution of the SNOMED codes, before and after filtering (pre and post filtered results), for each archetype fragment of the four archetypes.

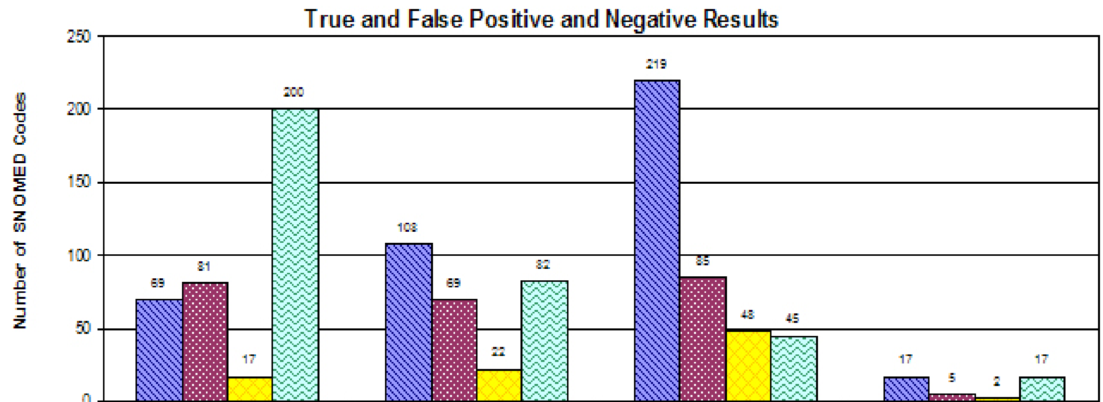
Archetype Model	**Total Arch. Frag.**	**Frag. with SCT* match**	**Pre-Filt. Results (SCT)**	**FNR** (SCT)**	**Post-Filt. Results (SCT)**	**TPR*** (SCT)**
Histology Pap	41	41	413	17	177	69
Visual Acuity	24	23	240	5	150	81
Body Weight	12	10	89	2	39	17
Tendon Babinski	45	44	344	17	282	200
			(TNR+FNR)+ (TPR+FPR)	(FNR)	(TPR+FPR)	(TPR)

Table 5.5: Distribution, for the four evaluation archetypes, of (i) total number of archetype fragments, (ii) fragments that found at least one SNOMED match, (iii) SNOMED codes eliminated during filtering (pre-filtered results) and (ii) SNOMED codes retained after filtering (post-filtered results). NOTE: * `SCT` denotes `SNOMED CT codes`; ** `FNR` denotes `False Negative Results`; *** `TPR` denotes `True Positive Results`.

In summary, Table 5.5 shows the total number of archetype fragments that were present in each of the four archetypes along with the total number of pre and post filtered results (i.e. SNOMED codes). The 'pre-filtered results' represent only the total number of SNOMED codes, while the 'post-filtered results' represent only the non-filtered/final number of SNOMED codes.

5.5.1 Comparison of Sensitivity and Specificity

A comparison of the ‘true’ and ‘false’ positive and negative results achieved for the four evaluation archetypes is shown in Figure 5.1. The figure graphically presents a comparison of the number of TPR, FPR, TNR, and FNR[3]. Based only on the proportional analysis of the true and false negative results, the visual acuity archetype seems to have performed the best, while the tendon babinski archetype seems to have performed the worst. The proportional values of the four results can be seen as the percentage view displayed in the graph table in Figure 5.1. However, we will see later in the chapter that the tendon babinski archetype in fact performed the best amongst the four evaluation archetypes, with respect to both the mapping coverage, as well as trust score (see Sections 5.5.2 and 5.5.3).



	True Positive (TPR)	False Positive (FPR)	True Negative (TNR)	False Negative
Histology Pap	69	108	219	17
Visual Acuity	81	69	85	5
Body Weight	17	22	48	2
Tendon Babinski	200	82	45	17

Figure 5.1: Comparative view of TPR, FPR, TNR, and FNR for each of the four evaluation archetypes.

On the basis of the results displayed in Figure 5.1, the sensitivity and specificity ratings of the four archetypes are shown in Table 5.6. The positive and negative predictive values are also calculated to demonstrate the rate of correct true positives and negatives.

The visual acuity archetype achieved the highest sensitivity rate of 94%. This signifies that the MoST system identified the maximum proportion of semantically equivalent SNOMED codes to map to the archetype fragments for the visual acuity

[3]TPR-True Positive Results, FPR-False Positive Results, TNR-True Negative Results, FNR-False Negative Results.

Archetype Model	**Sensitivity (%) (TP/TP+FN)**	**Specificity (%) (TN/FP+TN)**	**Positive Predictive Value(%) (TP/TP+FP)**	**Negative Predictive Value(%) (TN/FN+TN)**
Histology pap	80	66	39	92.7
Visual acuity	94	55	54	94.4
Body weight	89	68.5	43.6	96
Tendon babinski	92	35.4	71	72.6

Table 5.6: Overall view of the sensitivity and specificity ratings, as well as the positive predictive and negative predictive values, of the four evaluation archetypes.

archetype. On the other hand, Table 5.6 shows that the histology pap archetype had the lowest sensitivity rate of 80%.

With regards to the specificity rating, the body weight archetype achieved the highest rate of 68.5%. This signifies that MoST had identified the highest proportion of semantically nonequivalent SNOMED codes in the body weight archetype. Once again, the tendon babinski archetype had the lowest specificity rating of 35.4% despite a very good coverage rating and trust score (see Sections 5.5.2 and 5.5.3).

In addition to the sensitivity and specificity rates, the level of agreement between the MoST results and the evaluators was determined by calculating the positive and negative predictive values. The highest level of agreement on the positive values (i.e. positive predictive value) was achieved for the tendon babinski archetype at 71%. This meant that the evaluators agreed maximum with the SNOMED codes that MoST identified to be true positives. Similarly, the body weight archetype achieved 96%, which was the highest level of agreement on the negative values (i.e. negative predictive value). This meant that the evaluators agreed maximum with the SNOMED codes that MoST identified to be true negatives.

5.5.2 Comparison of Coverage Rating

The coverage rate was calculated using Equation 5.5 shown earlier in Section 5.3.2 (Page 129). The coverage rate determines the degree of archetype fragments in an archetype that found semantically equivalent SNOMED codes using the MoST system. The application of Equation 5.5, generated the following results of the coverage of the four evaluation archetypes. The four evaluation archetypes are histology pap, visual acuity, body weight, and tendon babinski.

1. *Histology Pap archetype:* The histology pap archetype had a total of 41 fragments, originally.

$$Criteria1 = 35fragments(total41fragments)$$
$$Criteria2 = (TPR + FNR) \geq 50\% = 16 + 1 = 17fragments$$
$$HistPapCoverage = \frac{17}{35} \times 100 = 48.6$$

2. *Visual Acuity archetpe:* The visual acuity archetype had a total of 24 fragments, originally.

$$Criteria1 = 23fragments(total24fragments)$$
$$Criteria2 = (TPR + FNR) \geq 50\% = 20 + 0 = 20fragments$$
$$VisualAcuityCoverage = \frac{20}{23} \times 100 = 86.9$$

3. *Body Weight archetype:* The body weight archetype had a total of 12 fragments, originally.

$$Criteria1 = 10fragments(total12fragments)$$
$$Criteria2 = (TPR + FNR) \geq 50\% = 6 + 1 = 7fragments$$
$$BodyWeightCoverage = \frac{7}{10} \times 100 = 70$$

4. *Tendon Babinski archetype:* The tendon babinski archetype had a total of 45 fragments, originally.

$$Criteria1 = 44fragments(total45fragments)$$
$$Criteria2 = (TPR + FNR) \geq 50\% = 34 + 6 = 40fragments$$
$$TenBabCoverage = \frac{40}{44} \times 100 = 90.9$$

As shown above, calculation of the coverage of the four archetypes shows that histology pap achieved a coverage of 48.6%, visual acuity achieved 86.9%, body weight achieved 70%, and tendon babinski achieved 90.9%. Figure 5.2 graphically presents the coverage rates achieved by each of the four archetypes. The histology pap archetype had the lowest coverage of 48.6%. The low score implies that less than half the archetype fragments that found semantically equivalent SNOMED codes.

The coverage rating also indicates the level of ambiguity at the content level of an archetype, and its compatibility with the SNOMED terminology. Therefore, a high

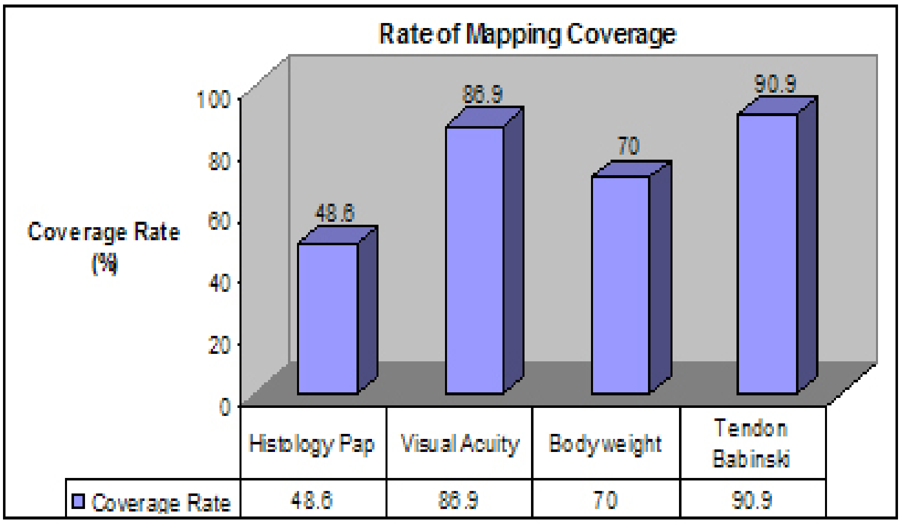


Figure 5.2: Rates of mapping coverage achieved for each of the four evaluation archetypes.

coverage rating indicates low level ambiguity with the archetype fragments, and a high level compatibility with the SNOMED concepts. On the other hand, a low coverage rating implies the presence of ambiguous fragments and low compatibility with the SNOMED concepts. It is not necessary that low compatibility implies the existence of poor quality archetype fragments alone. Low compatability might also be the result of poor coverage of a certain clinical area(s) within SNOMED.

The low coverage rating of the histology pap archetype led to an extended piece of additional research. The main aim of this research work was to test whether resolving the ambiguities present in the histology pap archetype would improve its compatibility with SNOMED, thereby improving the coverage rating. The work has been described in Section 6.2(Page 173) of Chapter 6. The work was also used as a measure to test the upper limit performance of the MoST system by setting the revised histology pap archetype as a gold standard archetype for the mapping purpose. The upper limit would indicate the highest achievable coverage rate in the event that ambiguities in the archetype fragments were reduced to the minimum. The upper limit of the coverage achieved would then highlight the degree of coverage and compatibility with the SNOMED terminology.

5.5.3 Comparison of Trust Score

The trust score was calculated on the basis of Equation 5.7. It helped to analyse the level of confidence/trust with which the evaluators agreed to map the SNOMED codes to the archetype fragments. The study places emphasis on the overall trust score achieved by the four archetypes to help determine the overall quality of not only the archetype model but the SNOMED codes. As mentioned earlier, the qualitative issues arising on evaluation of the results of each of the four archetypes has been described in Chapter 6.

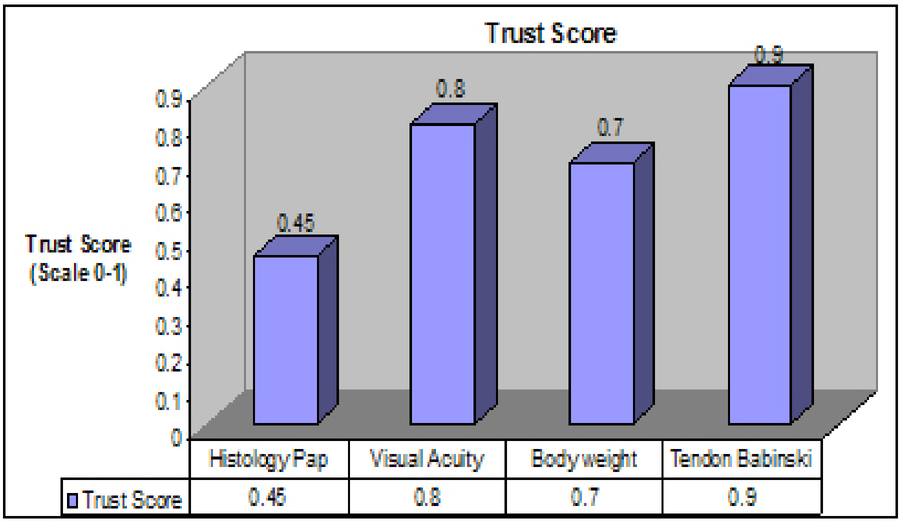


Figure 5.3: Comparative view of trust score achieved for each of the four evaluation archetypes on a scale of 0 to 1.

Figure 5.3 graphically presents the trust scores on a scale of zero to one. A quotient of zero states that the evaluators had no confidence on any of the results returned for an archetype model. On the other hand, a maximum quotient of one states that the evaluators had 100% confidence on all the results returned for an archetype model.

As shown in Figure 5.3, it can be seen that the SNOMED codes returned for the Tendon babinski archetype had the highest trust score of 0.9. This means that on an average two or more evaluators assigned a score of 9 (on a scale of 0 to 10) to the SNOMED codes they believed had high semantic equivalence to the archetype fragments. Once again, the Histology pap archetype had the lowest trust score of 0.45, implying that on an average the SNOMED codes were assigned a median score of 4.5 (on a scale of 0 to 10), even though they were thought to have some semantic equivalence to the archetype fragments.

The Visual acuity and Body weight archetypes achieved reasonably good trust scores of 0.8 and 0.7, respectively. Therefore, it can be concluded that the coverage rates are directly proportional to the trust scores. This means that the experts had more confidence in mapping the SNOMED codes to the archetype fragments when there was low ambiguity in the archetypes and higher compatibility with the SNOMED terminology.

5.5.4 Comparison of Inter-rater reliability

As explained in Section 5.3.4 above, given the limitations of the reliability of the inter-rater reliability scores using Fleiss' kappa, only the maximum attainable kappa, or κ_{max}, were considered for inclusion in the evaluation results. The reliability or agreement rate was calculated for only the worst and best performing archetypes, i.e. the histology pap and tendon babinksi archetypes, respectively.

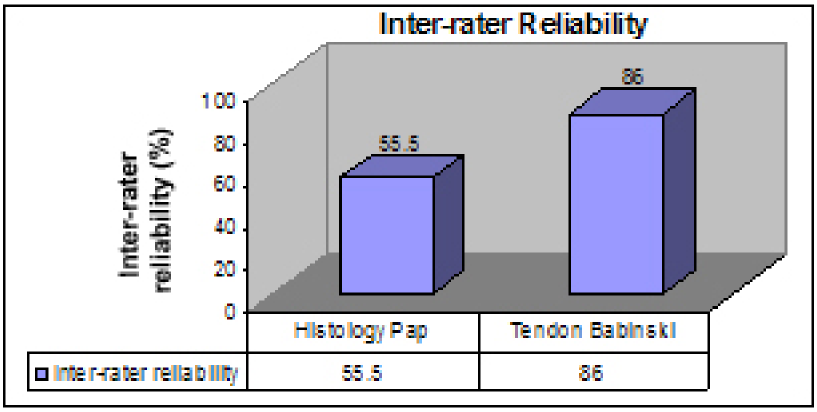


Figure 5.4: Degree of inter-rater reliability achieved for the histology pap and tendon babinski archetypes.

The histology pap archetype achieved a maximum agreement of only 55.5% (κ_{max} value of 0.555). Its overall kappa score was a low 14.5% (κ value of 0.145). Similarly, the tendon babinski archetype achieved a maximum agreement of 86% (κ_{max} value of 0.863), despite an overall low kappa of 15% (κ value of 0.151). See Section B.3 of Appendix B for a detailed account of the Fleiss' kappa and the measures calculated for the two archetypes.

The max kappa scores indicate that there was a high level of agreement of 86% amongst the evaluators for the MoST results returned for the tendon babinski archetype. However, the poor agreement of 55.5% for the histology pap archetype, indicates that the evaluators could not agree on the MoST results returned for the archetype. The semantic issues highlighted with the archetype and SNOMED models, to conceptualise the cervical smear (i.e. histology pap) recordings, is explained in Section 6.1.1 (Page 144).

5.6 Evaluation Phase 2: Comparative Data Analysis

Phase Two of the evaluation entailed a comparative data analysis of the SNOMED results returned by MoST for each of the four evaluation archetypes. The three main categories for the data analysis were to determine (i) the top five scoring SNOMED codes, (ii) SNOMED codes that required post-coordination, and (iii) SNOMED codes that were not returned by MoST at any stage of the *term finding* process.

Since the quantitative results were used mainly to perform the qualitative analysis, details of Phase 2 of the evaluation are relegated to Section B.4 of Appendix B.

5.7 Summary

This chapter performed a detailed quantitative analysis of the evaluation results obtained for the four archetypes selected for the study. In summary, the results returned by MoST had a higher rate of true positives as against false negatives for all the four archetypes. This highlighted the capabilities of the archetypes to find suitable SNOMED concepts for mapping/binding. The histology pap archetype had the lowest coverage of 48.6%, as well as a low trust score of 0.45. The inter-rater reliability calculated using Fleiss' kappa in Section 5.3.4, showed a low rate of agreement of 55.5% for the histology pap archetype. However, the tendon babinski archetype had a high level of agreement of 86% (see Section 5.5.4). However, the inter-rater reliability scores did not necessarily reflect the true pattern of agreement observed amongst the evaluators, and is, therefore, discarded as being of not much statistical value for further discussion in the thesis.

The tendon babinski archetype was the least ambiguous, as it achieved the highest coverage rate of 90.9% and trust score of 0.9. The visual acuity archetype achieved a coverage of 86.9% and a trust score of 0.8. Finally, the body weight archetype obtained a coverage of 70% and trust score of 0.7.

Despite the discussion of the quantitative measures, the nature of the study makes the analysis difficult. The main reason is the lack of a gold standard, together with the lack of consensus on what a gold standard would be. However, a small experiment was conducted to assess any change in the results "if" a gold standard archetype did exist. The histology pap archetype was revised to resolve its ambiguities, to bring it close to being a gold standard archetype. New results were obtained from the MoST system and compared with the existing histology pap archetype results. This experiment has been explained in Chapter 6.

The following chapter will detail the qualitative issues that were presented for each of the four evaluation archetypes. The issues highlighted in the SNOMED content with respect to the four archetypes will also be detailed. The discussion on the issues in archetypes and SNOMED will also aim at improving the specific content in both the modeling formalisms.

Chapter 6

Evaluation II: Qualitative Analysis

The qualitative feedback and discussion based on results from Chapter 5, proved to be the most significant aspect of the evaluation exercise and subsequently an important finding of the research programme. It is the third in the four-chapter series describing the core research work. The chapter focuses on the archetype modeling and SNOMED issues specific to the four evaluation archetypes. The discussion will begin with highlighting the issues with each of the four archetypes and progress to discussing issues with the coverage of the fragments in the SNOMED terminology. Issues and limitations with the MoST system will also be discussed.

The main issues[1] concerning the use of archetypes in the integration process were:

- ambiguity in the naming, classification, and annotation of the archetype fragments, and
- insufficient separation of meta data from the core data.

The terminology model posed issues with regards to:

- conflicting categorisation of closely equivalent concepts with the archetype fragments, and
- problems with synonyms, ambiguous concept definitions, and incongruent hierarchies.

[1]A research paper highlighting the archetype and SNOMED issues has been published in the Medinfo conference proceedings[78].

6.1 Issues with the Evaluation Archetypes and SNOMED

The four archetypes chosen for the evaluation (see Section 5.2.1; Page 123), as well as the four development archetypes (see Section 4.3.1; Page 103) used for the preliminary experiments, brought forward problems arising from model ambiguity. Since archetypes are the data source, it was considered critical to first address the issues related to archetypes. The central idea is to highlight the need for authoring unambiguous and usable archetypes, to have any success at coding their fragments unambiguously. Dealing with the SNOMED issues is a problem that can be resolved once issues with the source of the data, i.e. the archetypes, are resolved. Although the discussion is specific to archetypes and SNOMED, the issues are generic to any data and terminology models, used particularly in the health domain.

It is observed that archetypes are mainly limited in their potential to becoming high quality data models, not because of the technical aspects, but because of the clinical content (i.e. fragments). In other words, the problem does not arise due to the UML data structures that define the entities, sub-entities, and data types but due to the clinical content that forms the set of fragments in the archetypes. Therefore, the discussion of the issues is based on the set of fragments and their structure in the archetypes, rather than the UML data structures themselves. The aim is to take a step toward making the clinical content (i.e. clinical-side) as sound as the data structures (i.e. IT-side). The discussion hopes to establish formal quality assurance measures in authoring, and assessing archetypes. The discussion also hopes to explore the possibility of binding archetype fragments to SNOMED codes or other terminology codes, not only at author-time (covered by this research) but also at run-time (during real-time data entry). However, run-time binding will require a controlled number of SNOMED codes presented to the clinicians, to increase the speed and efficiency of binding in time-constrained healthcare environments. It is impractical to expect clinicians to mine through several hundred lexically similar terminology codes each time a binding/mapping is required.

6.1.1 The Histology Pap Archetype

The existing histology pap archetype taken from the official *open*EHR website[6] is shown in Figure 6.1. Discussion of the MoST results and evaluation scores have been previously covered in Chapter 5 (Page 134). This section discusses the issues arising from the ambiguity in the histology pap archetype and its coverage in SNOMED [2].

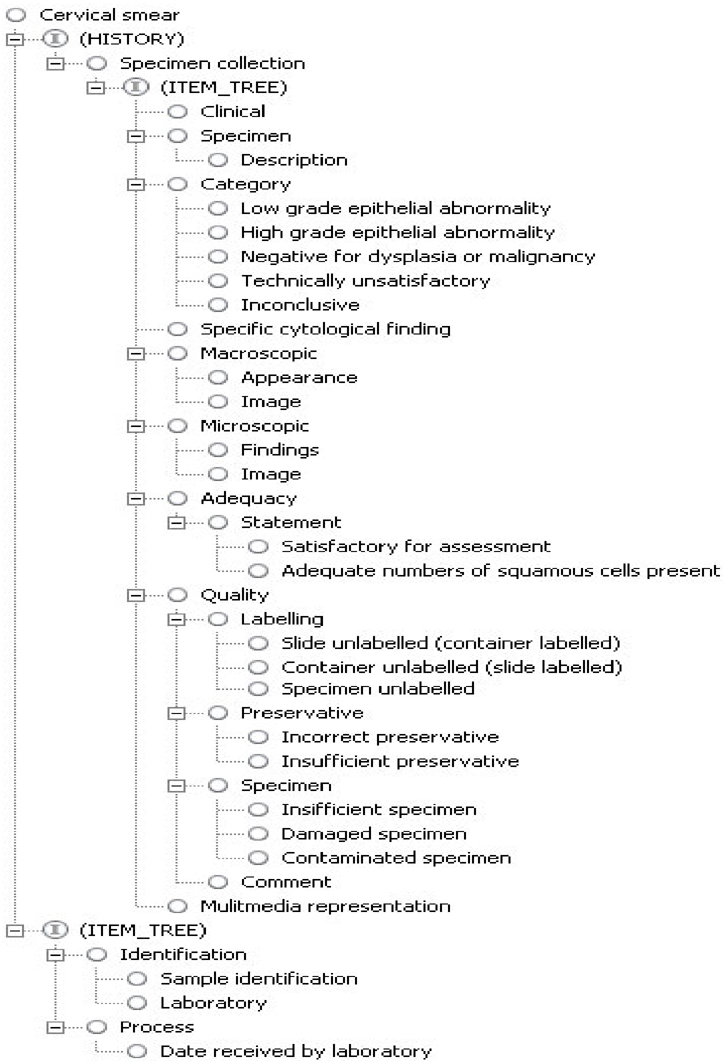


Figure 6.1: View of the Histology pap archetype from the `Terminology` section of the Archetype Editor.

6.1.1.1 Issues with the Histology Pap archetype

There were two central concerns with the histology pap archetype. Firstly, will a clinician using this archetype in a data entry process be able to clearly understand its intent i.e. what kind of information is the archetype trying to capture? Secondly, can the archetype in its present state be safely used to map to the SNOMED model without

[2] A research paper addressing the ambiguities in the histology pap archetype and resolving them through revision of the original archetype has been published in the AMIA Symposium proceedings of 2007[76].

resulting in conflict or error of the intended purpose and semantics? Based on these two main concerns, the following issues were raised:

1. *Ambiguous categorisation of top level archetype fragments:* Although the histology pap archetype belongs to the *open*EHR `observation` type, it is intended to record *cytological findings of a pap test.* In addition, the children of the root fragment 'cervical smear' are used to represent different aspects of a test, which would categorise it as a procedure instead. For example, the children include 'specimen collection', 'quality (and all its children)', 'identification (and all its children)', and 'process date'. It would have been clearer for the clinical experts to have worked with a `procedure` archetype instead, although it is not clear whether the archetype would then belong to an *open*EHR `action` or `instruction` type (as there is no `procedure` type in the *open*EHR ENTRY submodel).

2. *Conflicting categorisation with SNOMED:* The intended purpose of the archetype led to different opinions amongst evaluators regarding the basic SNOMED category to which the archetype belonged. Some experts preferred to map the root fragment 'histology pap' to a SNOMED observable (although no appropriate codes were available in this category in SNOMED), while most preferred SNOMED codes from the 'finding' and 'procedure' categories, as shown in Table 6.1.

Archetype Fragment	**Appropriate matches in SNOMED CT**
Histology pap	1. Papanicolaou smear test (procedure) [119252009]
	2. Cervical cytology test (procedure) [416107004]
	3. Cytopathology procedure, preparation of smear, genital source (procedure) [90226004]
	4. Cervical smear result (finding) [269957009]
	5. Cervical cytology finding (finding) [302796001]
Low grade epithelial abnormality	1. Low-grade squamous intraepithelial lesion (morphologic abnormality) [112662005]
	2. Low grade histologic differentiation (finding) [399415002]
	3. Low grade (qualifier value) [349915008]. Note: To be post-coordinated.

Table 6.1: Sample matches in SNOMED CT for archetype fragments. Matches include the SNOMED category and code.

3. *Ambiguously named fragments:* In more than one place the labels of the archetype fragments are inadequate to convey their intended meaning. The modeler (i.e. archetype author) provides the intended meaning and use of a fragment in the

`term definitions` section of the archetype. For example, the fragment 'clinical' is defined as 'clinical notes sent with request' in the archetype. It can be seen in Figure 6.1 that the enclosing fragment of 'clinical' is 'specimen collection'. The enclosing fragment makes it difficult to infer the intent of the modeler. A clinician might interpret 'clinical' to refer to the clinical specimen collection process, especially since there is a sibling fragment 'specimen' with 'description' as its child. The ambiguity could be avoided by using an explicit label such as 'clinical notes'. The other poorly labeled fragments in the archetype are 'macroscopic' and its child 'appearance', 'microscopic', 'process', 'identification', 'comment', 'laboratory', and 'preservative' (see Figure 6.1).

The majority of the 10 evaluators concluded that they were unable to interpret the intention of the modeler by simply looking at the labeling of the fragments. At times, the definition provided was also inadequate to resolve the issue of semantics. Besides aiding humans to make sound judgments on the semantic equivalence of the terminology codes, better defined and/or labeled fragments will also make it simpler for automated systems, such as MoST ,to get back better matches from the terminology.

4. *Similar or duplicate labeling of fragments with inadequate definitions to resolve ambiguity:* At times, the labels assigned to the fragments were either duplicated or very similar to other labels in the archetype. Although the context differed, it was difficult to precisely ascertain their intended use, as the term definitions lacked clarity. For example, 'specimen' was used twice in the same archetype. In the first instance, it meant to record 'details of specimen'. Later it was used to record the 'problem with the specimen'. Also there were other uses of 'specimen' in fragments such as 'specimen collection', 'specimen unlabeled', 'insufficient specimen', and 'damaged specimen'. Although the fragments were placed in different sub-hierarchies, the context itself was not comprehensible, to clarify its existence in the archetype. Another fragment was 'image' used for macroscopic and microscopic findings.

5. *Insufficient separation of general data from core clinical recording information:* The archetype did not sufficiently separate fragments that were generic to the recording of a cervical smear from fragments that were more specific to the cervical recording. Since most of the evaluators agreed that the archetype represents the recording of a procedure, the discussion will be based on procedure types. Fragments such as 'specimen collection', 'clinical', 'category', 'adequacy', 'satisfactory assessment', 'quality', 'labeling', 'preservative', and 'comment' are

common to any procedure and are therefore generic to the archetype. These fragments need to be placed separately from data that is specific to the recording of cervical smear such as 'low grade' and 'high grade' epithelial abnormality, 'negative for dysplasia or malignancy', and 'specific cytological finding'. In fact, since archetypes can reuse whole or parts of other archetypes, the generic data could be modeled in a separate archetype and reused in this archetype. Even if the generic data is included with the more specific clinical data in the same archetype, it should be clearly separated to reduce ambiguity and enhance simplicity of use.

6. *Adequate and appropriate use of post-coordination:* At times the archetype fragments were needlessly split into fine-grained representations of a concept (or post-coordinated expressions) to the extent that their semantics were lost in the process. For example, the fragments 'description' and its enclosing fragment 'specimen' had the same definition 'details of the specimen'. Individually they did not make much sense to the expert trying to infer the semantics of both fragments. However, when combined together to form a composite i.e. pre-coordinated expression, 'specimen description', the intent was clearer. Well-defined fragments would help systems like MoST to search for appropriate matches in SNOMED. Table 6.2 below shows the SNOMED codes returned by MoST for the current archetype representation. Had the fragments been pre-coordinated the expert might have selected only the code `115267000` (code shown in Table 6.2). It can be seen in Table 6.2 that the child fragment 'description' returns a better match with SNOMED as against its parent fragment 'specimen'. This is because the child fragment took into account the context in which it had been used.

Archetype Fragment	**SNOMED CT Match**
Specimen	1. Form and specimen details different (finding) [281322007]
Description	1. Description of specimen character (procedure) [115597007] 2. Specimen description (procedure) [115267000]

Table 6.2: SNOMED codes returned for the post-coordinated fragments in the Archetype Model.

Similarly, at times decomposing a concept into two or more fragments (i.e. post-coordinating the fragments) to convey the correct semantics was more suited than pre-coordinating them (i.e having one large code to represent a single concept). For example, 'low grade epithelial abnormality' might have been better represented in the form of post-coordinated fragments shown in Table 6.3. Similarly, coding the 'high grade' fragment to any of the three post-coordinated forms, as expressed in Table 6.3, might provide greater clarity and a better mapping.

Archetype Fragment	Post-coordinated SNOMED CT Match
Low grade epithelial abnormality	1. Epithelial cell abnormality (morphologic abnormality) [373886001] + Low grade (qualifier value) [349915008]
	2. Histological grade finding (finding) [373372005] + Low grade (qualifier value) [349915008]
	3. Squamous intraepithelial lesion (morphologic abnormality) [24253004] + Low grade (qualifier value) [349915008]

Table 6.3: Suggested post-coordination of the archetype fragment with respect to SNOMED CT.

6.1.1.2 SNOMED Issues with Histology Pap Archetype

Despite the wide coverage of clinical areas by SNOMED CT, problems still persist with its content. The problems relate to their definitions and subsumption relationships in the terminology, as well as their categorisation and naming. Despite being based on Description Logic (DL), the entire terminology is not classified using DL reasoners. While parts of the hierarchy are inferred by the reasoner, other parts are asserted manually[3].

A major problem when working with the histology pap archetype was the lack of adequate coverage of its concepts in SNOMED. Therefore, this section will focus on the concepts that are not present in SNOMED with respect to the findings from a cervical smear test (core intent of the histology pap archetype). Coverage of the concepts in SNOMED listed below were considered important by the experts. Most of the concepts not covered by SNOMED were common to clinical tests or procedures.

- `Technically unsatisfactory` - This concept was thought to be useful for inclusion to represent technical problems in any kind of test performed.
- `Appearance` - Although there is a SNOMED concept 'On examination-appearance (finding)' or 'o/e - appearance', it refers to the appearance of a person during physical examination, rather than the appearance of a specimen during a

[3]Personal Communication, Kent Spackman, 2007. Spackman, M.D., Oregon Health Sciences University, is the Interim Chief Terminologist, and one of the IHTSDO Executive Officers

histopathology examination. Also, a synonym of the concept 'Macroscopic specimen observation (finding)' in SNOMED refers to the general appearance of the specimen. This SNOMED code was selected by the evaluators for the enclosing fragment 'Macroscopic' due to its inappropriate post-coordination by the modeler. However, there was a general consensus that SNOMED should include a code for 'macroscopic appearance', which conceptualises the appearance of the tissue specimen.

- `Adequate numbers of squamous cells present` - A general concept can be created in SNOMED to conceptualise 'adequate number of cells present' and then post-coordinate it with 'Squamous epithelial cell (cell) [80554009]'. An alternative could be to create a pre-coordinated concept 'Adequate numbers of squamous cells present' localising it to Pap smear.

  Some of the other concepts suggested to be included in SNOMED were 'slide', 'container', and 'specimen' unlabelled instead of the generic `Sample unlabelled (finding) [125159002]`. In addition, modifiers on 'specimen' such as 'insufficient', 'damaged', and 'contaminated' could be included to allow post-coordination. Finally, a concept for 'preservation of specimen' enabling post-coordination, along with modifiers to highlight problems with the preservation of the specimen such as 'insufficient', and 'incorrect' could also be included in SNOMED.

6.1.2 The Tendon Babinski Archetype

The tendon babinski archetype is the second archetype that raised issues during the evaluation exercise. It is also the largest archetype amongst the four evaluation archetypes, and is intended to *record tendon reflexes and babinski responses during routine neurological examination*. A snapshot of the fragment hierarchy is shown in Figure 6.2.

Tendon reflex (or T-reflex) is a feedback mechanism that controls increasing muscle tension by causing muscle relaxation before tension force becomes so great it might lead to damage of the muscle[4]. In medicine (neurology), the Babinski reflex (or Babinski sign) is a reflex that can identify disease of the spinal cord and brain. It is more properly called the plantar reflex, as Babinski's sign in reality only refers to the pathological form[5]. Babinski's reflex occurs when the great toe flexes toward the top of the foot

[4]Taken from http://en.wikipedia.org/wiki/Tendon_reflex
[5]Taken from http://en.wikipedia.org/wiki/Plantar_reflex

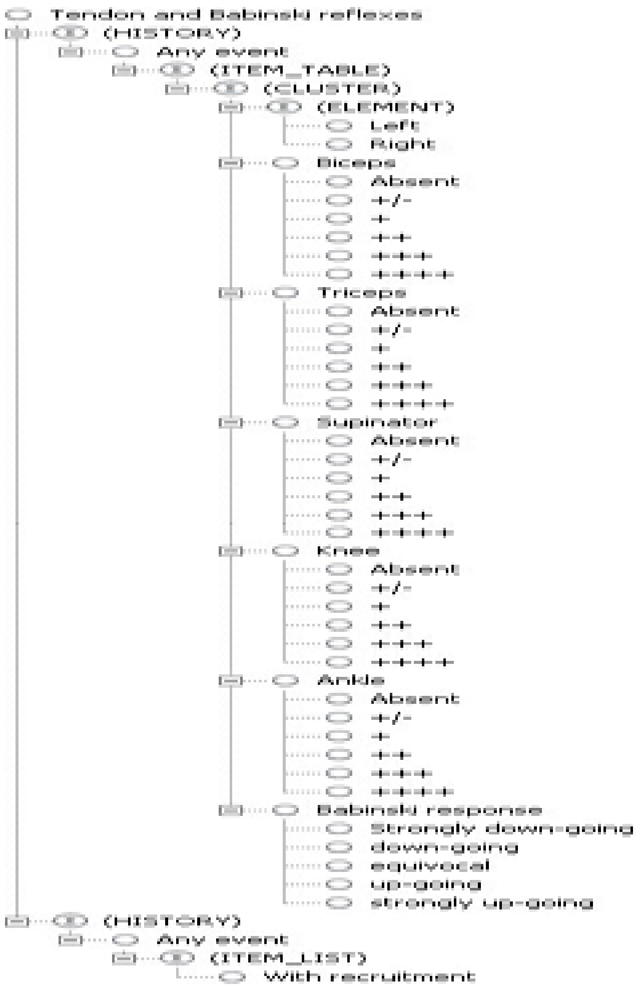


Figure 6.2: Tendon Babinski archetype. View from the `Terminology` section of the Archetype Editor.

and the other toes fan out after the sole of the foot has been firmly stroked. This is normal in younger children, but abnormal after the age of two[6].

6.1.2.1 Issues with the Tendon Babinski archetype

The main issues with the tendon babinski archetype were raised by the experts before they established the most appropriate pattern of selecting the SNOMED codes. A pattern needs to be established because the tendon babinski archetype has the same fragment labels replicated throughout the *tendon* sub-hierarchy of the archetype. However, they are used in different contexts with the help of different enclosing fragments. For instance, in Figure 6.2, the fragment `++`, described as 'Reflex Normal', has been reused five times, although in five separate contexts such as 'biceps reflex normal', and

[6]http://www.nlm.nih.gov/medlineplus/ency/article/003294.htm

'triceps reflex normal'. Therefore, once a pattern of equivalent SNOMED codes was established by the evaluators, the mapping process became a mechanical task.

In addition to the issues raised prior to establishing a pattern of equivalent SNOMED codes, there were also issues with the labeling of the fragments in the *babinski* section of the archetype. The issues with the archetype are described below.

1. *Renaming root archetype fragment:* Majority of the evaluators suggested that the word *babinski* in the archetype should be replaced with *plantar*, as the latter is more commonly used in medical practice. SNOMED CT agrees with this view, as no results are returned for the fragment 'babinski response'. However, the terminology has several concepts to cover the scope of plantar reflexes and response. Some of the concepts are `Plantar reflex (observable entity)[91365008]`, `o/e - plantar response (finding)[163852006]`, and `Plantar reflex finding (finding)` *[366553001]*, amongst several others.

   The labeling/naming of core clinical concepts appears to be a matter of geographic preference. The suggestion is to provide some of the more popularly known names for such concepts, even as synonyms, to increase the scope of finding matches in large terminologies.

2. *Ambiguity with fragments representing degree of reflexes:* Table 6.4 shows the six fragments used repeatedly in the archetype to record the degree of reflex response in different body sites. The semantics of the fragments that the evaluators had maximum difficulty understanding were '+/-', '+++', and '++++'. For instance, the fragment '+/-' was used in the archetype to record the 'possibility of some reflex being present although there was a significantly reduced amplitude'. Although the annotation appears to be unambiguous at first, it was uncertain whether to lay more emphasis on 'some reflex was possibly present' or 'reflex amplitude was markedly reduced'. It was necessary to determine the part of the term definition which was more important, as there were no precoordinated concepts in SNOMED covering both the aspects. The closest lexical match found in SNOMED was the concept `Interference pattern showing reduced amplitude (finding)[251523007]`. However, this concept was a specialised finding of an electromyogram interference pattern rather than an observation of a tendon reflex. Therefore, it was subsequently filtered by MoST, as a semantically non-equivalent match.

Archetype Frag.	Term Definition
Absent	Reflex not evident
+/-	Reflex possibly present - markedly reduced amplitude
+	Reflex present but reduced amplitude
++	Reflex normal
+++	Reflex increased
++++	Reflex markedly increased

Table 6.4: Symbols used to measure degree of reflexes in the *tendon babinski* archetype and their corresponding definitions/annotations.

Therefore, to safely and unambiguously map archetype fragments to any terminology, it is important that the labeling of the fragments and their annotations/definitions are more compatible with terminology systems. However, the labeling might not always be compatible in which case sufficient annotation should be provided to enable unambiguous interpretation of the semantics.

3. *Differences with SNOMED grading system:* There was also the issue in the number of reflex grades on a scale that were available in archetypes and SNOMED. As shown in Figure 6.4, the archetype had six grades for measuring the tendon reflexes. On the other hand, SNOMED had only four corresponding grades viz. absent, present, normal, and increased. While there was no equivalent code for '+/-', there was no obvious SNOMED match for the fragment '++++' either. The lack of a direct SNOMED match led to a dispute with the fragment '+++', which had a very similar definition (see Figure 6.4). The three SNOMED results that were close matches were `reflex brisk`, `reflex exaggerated`, or `hyperreflexia`. However, it was difficult to ascertain which archetype fragment could be mapped to either of the three SNOMED codes. For instance, there was ambiguity on mapping the archetype fragments '+++' and '++++' modeled to record the triceps reflexes, to the most equivalent of the three SNOMED codes viz. `Triceps reflex brisk (finding)[299827007]`, `o/e - triceps reflex exaggerated (finding) [163821005]`, or `Hyperreflexia (finding) [86854008]`. Although no absolute consensus could be reached, majority of the experts selected the code `299827007` for the fragment '+++', and the codes `163821005` and `86854008` for the fragment '++++'.

It is not possible to resolve issues when there is a mismatch in the number of grades used in two different models. Such issues can only be resolved by either modifying the number and/or type of grades used in either the archetype or SNOMED model.

Archetype Fragment	**Term Definition**	**SNOMED Concept**
Left	Reflexes on the left side of the body	Structure of left half of body (body structure)[31156008]
		Left (qualifier value) [7771000]
Right	Reflexes on the right side of the body	Structure of right half of body (body structure)[85421007]
		Right (qualifier value) [24028007]

Table 6.5: Archetype fragments in Tendon Babinski archetype with lack of consensus amongst evaluators due to ambiguity in the archetype.

4. *Problem resolving laterality of findings:* The qualifier fragments 'left' and 'right' were used in the tendon babinski archetype to record the laterality of measurement of the tendon reflexes. Table 6.5 shows the SNOMED codes that were considered to be most equivalent to the two fragments. However, it was not clear to the evaluators what the intended use of the lateralisation fragments were with respect to the reflex fragments. In addition to the lack of clarity in the archetype, SNOMED did not have laterality-specific reflex concepts, which accentuated the ambiguity in the archetype. Table 6.5 shows the two SNOMED codes that were given similar scores by the four evaluators indicating a lack of consensus. In addition, the evaluators stated that they were unsure which of the two codes would be a more appropriate mapping to the archetype fragment. They commented that it was unclear whether the fragments 'left' and 'right' were used as laterality headings below which all observations/findings were to be modeled, or whether the fragments were intended to refine each of the observations/findings individually. The SNOMED concepts `Structure of left/right half of body` covered the laterality of an entire side of a body, making them very general concepts. Alternatively, the SNOMED concepts `left/right` were incomplete concepts, which required post-coordination with a more specific concept. This resulted in lower, distributed scores with no consensus on either of the codes.

5. *Inappropriate labeling of fragments in the archetype:* As mentioned earlier in

Section 6.1.1.1, inappropriately labeled archetype fragments lead to issues in resolving the semantics of the fragments, and their subsequent mapping to terminology codes. A list of inappropriately labeled fragments in the tendon babinski archetype are listed in Table 6.6 along with their term definitions.

POST-COORDINATED FRAGMENTS	
Archetype Frag.	**Term Definition**
Strongly down-going	Babinski response is strongly toward the sole of the foot
Down-going	Babinski response is toward the sole of the foot
Up-going	Babinski response is toward the dorsum of the foot
Strongly up-going	Babinski response is strongly toward the dorsum of the foot
With recruitment	Using muscle recruitment elsewhere in the body

Table 6.6: Ambiguously labeled fragments in the Tendon Babinski archetype.

As seen in Table 6.6, the poorly labeled fragments are ‘strongly down-going’, ‘down-going’, ‘up-going’, ‘strongly up-going’, and ‘with recruitment’. These labels do not in themselves convey the intended semantics and use of the fragments. In addition to the labels, the term definitions also do not provide useful information about the intent of the fragments. For example, the definition of the fragment ‘down-going’ is ‘babinski response is toward the sole of the foot’. The use of ambiguous words such as ‘toward’ and ‘elsewhere’ in the definitions cause problems when looking up matches in SNOMED. For instance, SNOMED conceptualises site-related codes specific to the target body site or region. Therefore, given that SNOMED provides better body site-related concepts, increases the need for archetypes to be more precise with the annotations for body site-related fragments. In the absence of semantic clarity, coupled with the lack of SNOMED codes available to provide either pre-coordinated or post-coordinated mappings (see Appendix B, Table B.12), majority of the evaluators did not select any of the results presented to them by MoST. If they did select a SNOMED code, they gave it a very low score.

Despite some of the issues encountered by the evaluators, in general the tendon babinski archetype was fairly accurate and easy to use. There was also a very high rate of coverage achieved for mapping (see Section 5.5.2; Page 136). The major issues were with the SNOMED hierarchies conceptualising the reflex responses in the human body.

6.1.2.2 SNOMED Issues with Tendon babinski archetype

While assessing the SNOMED results for the tendon babinski archetype, the evaluators took into consideration the concept definitions in the terminology and the hierarchies to which they belonged, along with their position in the hierarchy. The single most critical observation made was related to the SNOMED hierarchies for the tendon and plantar reflexes. Several instances of duplication of concepts and hierarchies, as well as redundancy were identified. The discussion has been divided into the observations made for (i) the tendon reflexes, and (ii) the babinski/plantar reflexes. The fragment `plantar` will be used instead of `babinski` to conform to the SNOMED usage of the concept.

1. *Issues with tendon reflexes hierarchy:* The body structures with tendons listed in the archetype include the biceps, triceps, knee, ankle, and supinator. In medical terms, a tendon (or sinew) is a tough band of fibrous connective tissue that connects muscle to bone or muscle to muscle and is built to withstand tension [7]. Discrepancies were found in the SNOMED terminology with the following concepts:

    - *Problems with synonyms:* The term *radial* is also used as a synonym for *supinator* although it is explicitly stated in the SNOMED hierarchy only for the concept `Radial reflex absent (finding)[274605005]`. It is important, if SNOMED is to work successfully with automated search tools such as MoST, that it indicates explicitly in its model that the terms radial and supinator are synonyms of each other, and not two completely unrelated concepts.
    - *Incongruent hierarchies:* There were two SNOMED concepts worth considering amongst the results returned by MoST for the reflex degree fragment '++' i.e. 'reflex normal'. These were the `reflex normal` and `o/e - reflex normal` codes. Figure 6.3 displays the normal reflex codes for the triceps body structure reflexes. The definitions of the two concepts infers a subsumption relationship between the two. It can be seen from Figure 6.3 that the concept `Triceps reflex normal (finding)[299826003]` subsumes the concept `o/e - triceps reflex normal (finding)[163816007]`.

[7] Definition taken from http://en.wikipedia.org/wiki/Tendon, in April 2007

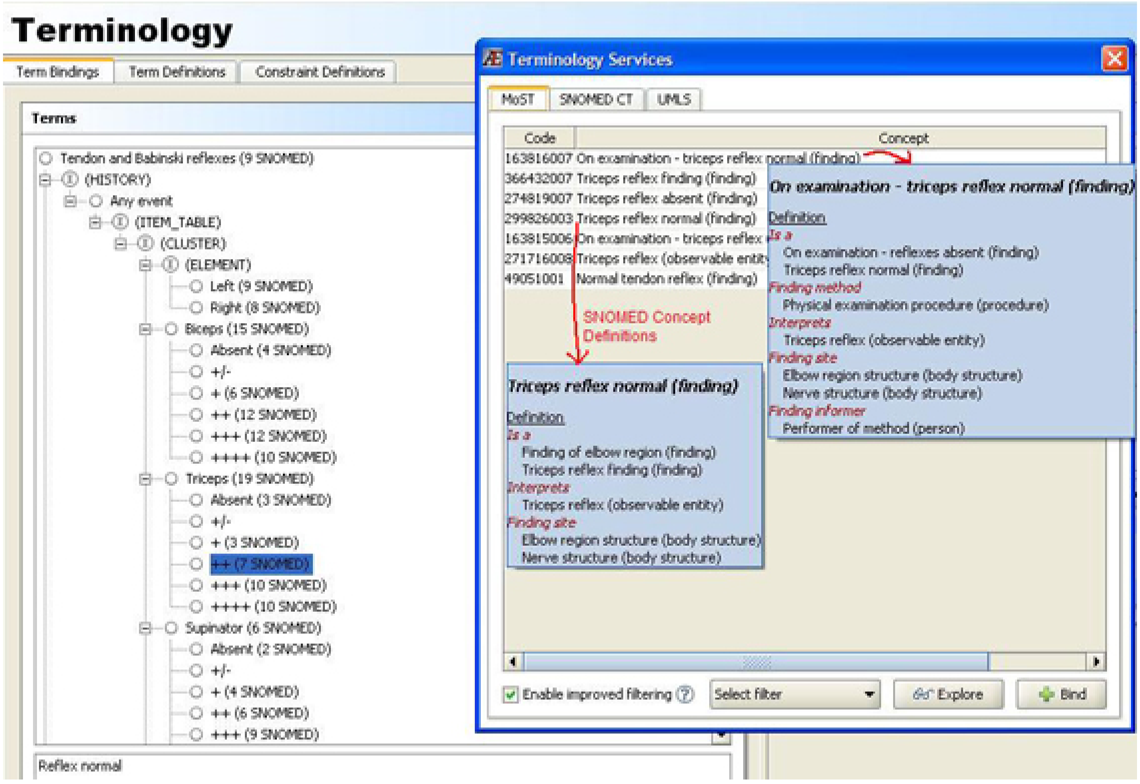


Figure 6.3: Normal reflex codes in SNOMED CT for conceptualising triceps reflexes. Displays *subsumption* relationship between the two concepts.

The subsumption relationship between the `reflex normal` and `o/e - reflex normal` codes holds true not only for the triceps body structure reflexes but also for biceps, knee, and ankle. However, the same relationship does not hold true for the supinator body structure reflexes. Under the supinator hierarchy, the two codes are `Supinator reflex normal (finding)[299838007]` and `o/e - radial reflex normal (finding)[163832005]` as siblings, as shown in Figure 6.4. As stated earlier, the term *radial* is also used as a synonym for *supinator*.

As can be seen in Figure 6.4, the two supinator concepts appear as siblings, rather than as parent-child concepts. This is in conflict with the relationship modeled for the other tendon reflexes i.e. triceps, biceps, knee, and ankle. Therefore, the hierarchies are incongruent and need correcting.

- *Inconsistently formulated definitions:* The definitions provided in SNOMED for the `o/e reflex normal` concepts are inconsistent across the terminology. In a DL context this would mean that the definition are “unsatisfiable”.

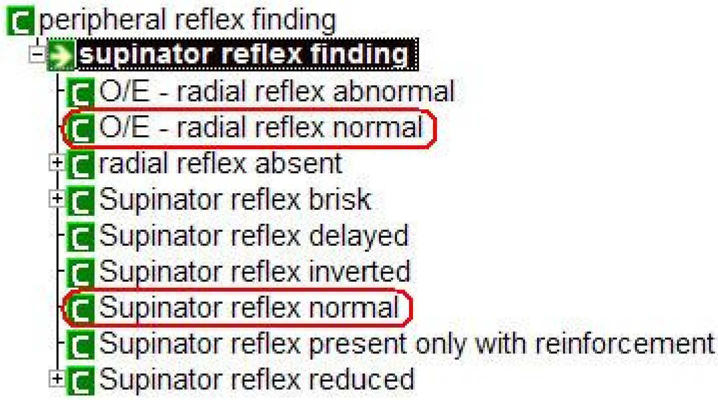


Figure 6.4: Normal reflex codes in SNOMED CT for conceptualising supinator reflexes. Displays *sibling* relationship between the two concepts.

The problem is not that each definition in itself is inconsistent but that similar concepts are not defined according to a consistent pattern. Figure 6.5 shows that at least three variations exist in SNOMED for the same objective of coding normal reflexes. While triceps, knee, and ankle have similar definitions for the `o/e reflex normal` concept, where it is subsumed by both the `o/e - reflexes absent (finding)[163785007]` and variants of the `normal reflex` codes; the same does not hold true for the biceps and supinator codes.

Part (b) of Figure 6.5 shows a more specialised subsumption relationship of the concept `o/e - biceps reflex normal (finding)[163824002]`. The concept `o/e - biceps reflex normal` is subsumed by `biceps reflex normal (finding)[163824002]` and `o/e - biceps reflex (finding)[163823008]`. The latter concept is further subsumed by the concept code `163785007`, which has been used in the definitions for triceps, knee, and ankle `o/e reflex normal` codes. It is unclear why the more specialised subsumption hierarchy, as used in biceps, could not be used for the remainder tendon variants, as shown in Parts (a) and (c).

In Part (c) there is a third variation in the definition, where the concept `o/e radial reflex normal(finding)[163832005]` is subsumed by `Supinator reflex finding[366452008]` instead of by `Supinator reflex normal[299838007]`. Therefore, there is a lack of a principled approach toward defining similar objective concepts in SNOMED.

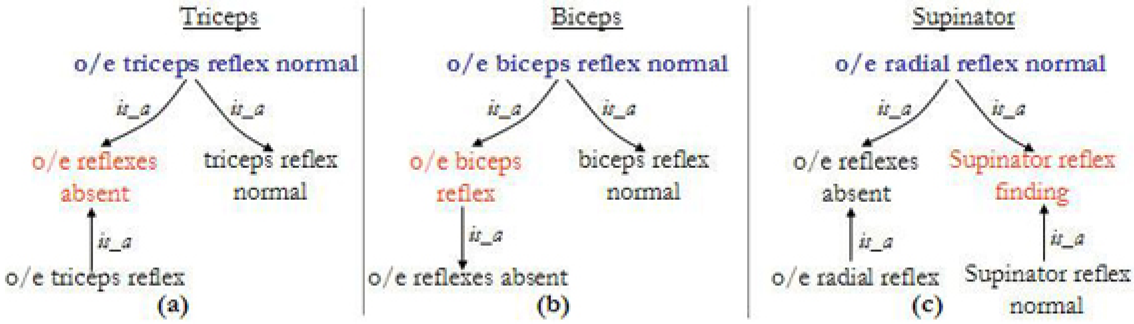


Figure 6.5: Inconsistently formulated SNOMED concept definitions of three equivalent concepts to model normal tendon reflexes.

- *Debatable hierarchy of concepts:* In continuance with the discussion on the normal reflex hierarchy, it is debatable as to whether general `‘o/e reflex’` codes such as `o/e - triceps reflex[163815006]` should be partially subsumed by `o/e - reflexes absent(finding)[163785007]`, as shown in Figure 6.6. The figure shows some of the child concepts of the `reflexes absent` code. It is unclear why the general `o/e - triceps reflex[163815006]` code and the normality concept `o/e - triceps reflex normal[163816007]` should be partially subsumed by a concept representing the *absence* of reflex, which is a very specific finding or situation. Figure 6.6 also shows the ‘ankle’ and general ‘reflex’ concepts, which are subsumed by the ‘reflex absent’ code, thereby rendering this entire sub-hierarchy unsound.

Figure 6.6: Partial view of the *O/E - reflexes absent* hierarchy with debatable child concepts.

Upon revisiting the `triceps reflex finding[366432007]` hierarchy, partially shown in Figure 6.7, there was additional ambiguity as to why the `o/e - triceps reflex normal` concept was defined to partially be subsumed by the `o/e - reflexes absent` concept (shown in Part (b) of Figure 6.7). The ambiguity was accentuated when there separately existed a `o/e - triceps reflex absent[163820006]` concept (shown in Part (a) of Figure 6.7), which should have ideally been related to the `reflexes absent` code instead.

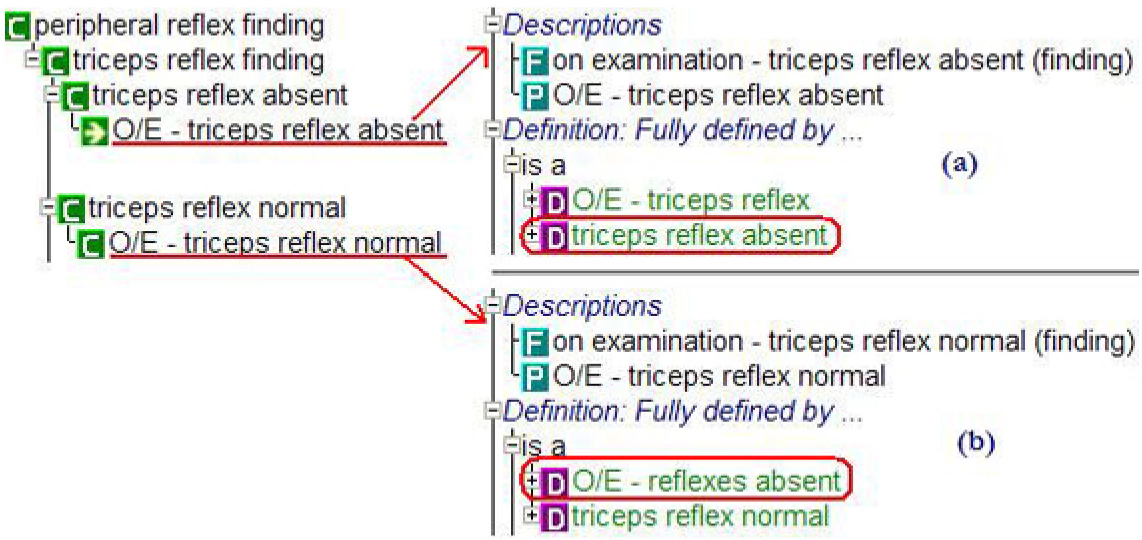


Figure 6.7: Questionable subsumption of *o/e reflex* concepts by the *o/e reflexes absent* concept. View of the triceps concepts as an example.

It would be reasonable to assume that a `normal reflex` code would depend on the definition of normality. However, it appears from the partial view of the SNOMED hierarchy in Figure 6.6 that not only have the normality codes but also other observations of the degree of tendon reflexes, been modeled as specialisations of the `absence of reflexes`.

Although the discussion has focused mainly on the `normal reflex` codes the problem has been seen to exist in other parts of the subhierarchy as well. A single example was chosen for easy understanding of the problem. In addition to the issues discussed above, a similar pattern of inconsistent subsumption hierarchies could also be observed for the `reflex exaggerated` codes for triceps, biceps, knee, ankle, and supinator/radial, as well as the general `hyperreflexia` code. The SNOMED hierarchy pertaining to the tendon reflexes is riddled with ambiguities and inconsistencies which need to be addressed by the SNOMED community.

2. *Issues with babinski reflexes hierarchy:* The fragment 'babinski' was used in the archetype rather than the name 'plantar'. However, since SNOMED uses the latter as a rubric to name its concepts related to observations and measurements of the reflexes of the great toe, the discussion will retain the name *plantar*. The issues that were encountered with the hierarchy specific to plantar were:

   - *Unclear interchangeable names:* The plantar reflexes in SNOMED were modeled as both *plantar reflex* as well as *plantar response*. The difference between the terms `reflex` and `response` have neither been asserted in SNOMED nor can be inferred from the definitions. The only subsumption hierarchy available in SNOMED, shown in Figure 6.8, is insufficient and does not help in resolving the ambiguous use of the two rubrics interchangeably. Therefore, it is difficult to justify the conceptualisation of certain concepts under the sub-hierarchy of `reflex`, while others are classified under `response`.

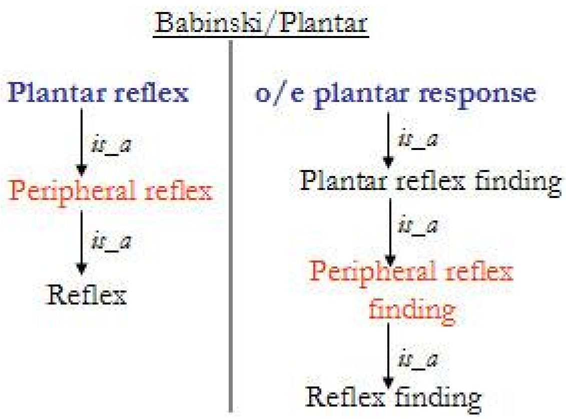


Figure 6.8: Subsumption hierarchies of *plantar reflex* and *o/e plantar response*.

     Apart from the interchangeable use of the terms, in view of the discussion in the previous header on issues with the tendon hierarchy, the `o/e - plantar response (finding)[163852006]` should have been a specialisation of a *plantar response* or a *plantar reflex* concept rather than a free-standing concept directly subsumed by `Plantar reflex finding (finding)[366553001]`. Therefore, synonyms should be explicitly stated in the terminologies when adopting the approach of using clinical name variants.

   - *Equivalent concepts not treated such:* The two SNOMED concepts `Plantar grasp reflex finding (finding)[366221009]` and `Plantar reflex finding`

(finding)[366553001] have similar definitions, as can be seen in Figure 6.9.

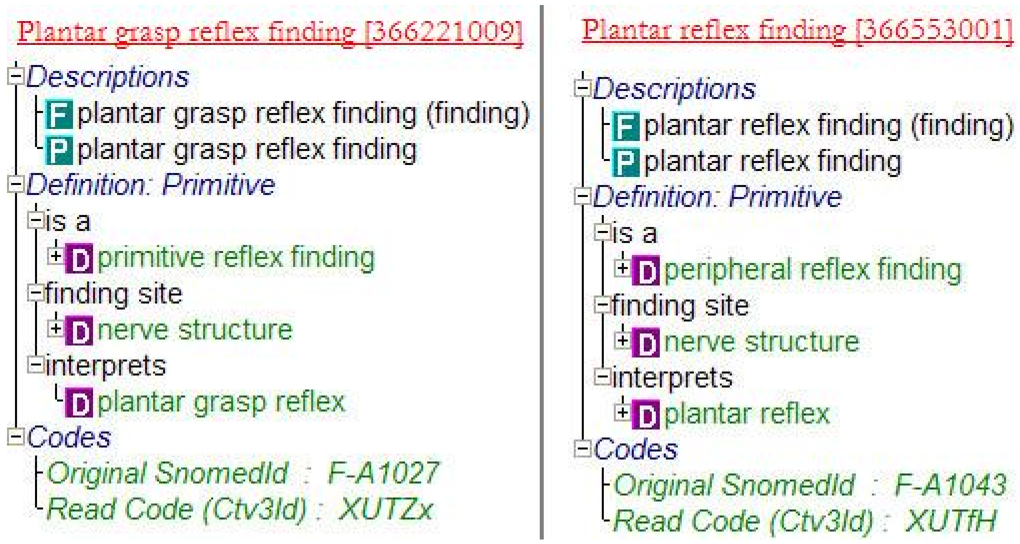


Figure 6.9: Similar definitions of the two concepts in SNOMED make them equivalent concepts.

The hypothesis for claiming the two concepts to be equivalent is because of the "open world reasoning" adopted by Description Logic, on which SNOMED is based. On this basis, in the absence of sibling concepts being explicitly stated as being disjoint, they might be assumed to be equivalent, if the definitions do not provide clear differentiation otherwise. Therefore, the lack of disjoints in SNOMED might lead to errors. On further analysing the parent concepts, as shown in Figure 6.10, there appears to be no logical difference between the two hierarchies. The parents of the concepts shown in Parts (a) and (b) of the Figure have a common parent `Reflex finding (finding)[106146005]`. In addition, (a) and (b) shows that the parent concepts *interpret* the `primitive reflex` and `peripheral reflex` codes, respectively, which are both kinds_of `Reflex (observable entity) [87572000]`.

Based on the similarities in the definitions and hierarchies of the `plantar grasp reflex finding` and `plantar reflex finding`, it would be more appropriate to regard the two as equivalents, and include one as the synonym or child of the other. It is essential to reduce the occurrence of such duplicate hierarchies to not only improve the quality of the terminology

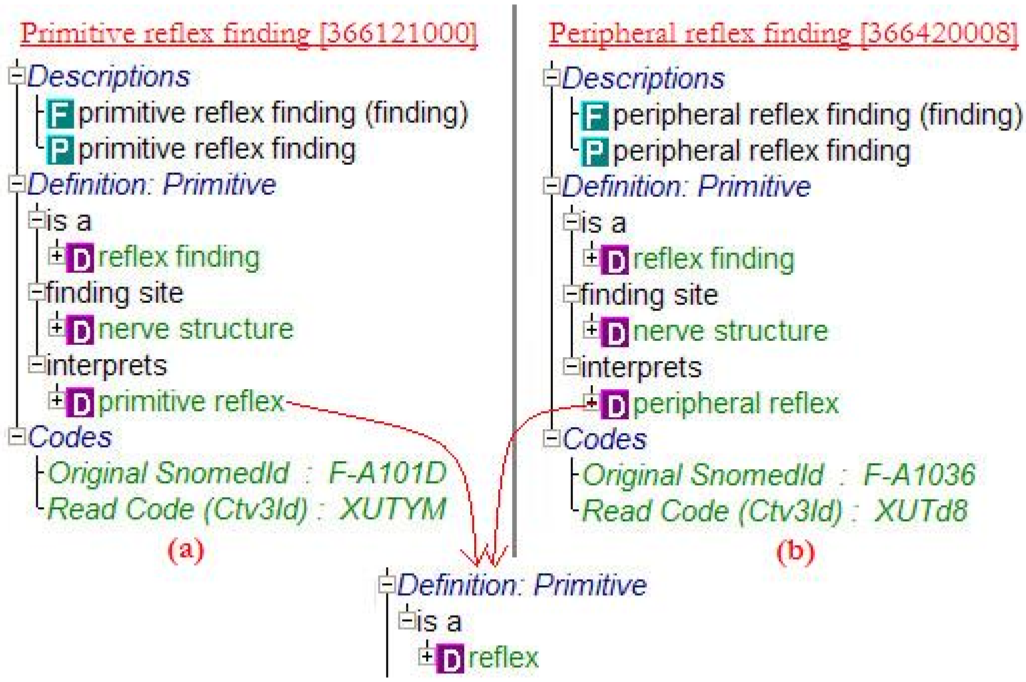


Figure 6.10: Analysis of the specific-to-general (child-to-parent) hierarchy of the *plantar reflex* and *response* concepts in SNOMED.

structures but also to aid automated search applications to retrieve better results.

- *Suggestion for more general concepts to record plantar reflexes:* The SNOMED hierarchy includes approximately 35 concepts related to various plantar reflexes and responses. However, in relation to the plantar fragments modeled in the tendon babinski archetype, there are no suitable matches available in SNOMED. The terminology mainly addresses the absence, presence, and laterality of the reflexes/responses. Concepts concerning more general direction of reflexes are not included in the terminology such as 'plantar response toward sole of foot', and 'plantar response toward dorsum of foot' used in the archetype.

  It might be worth reconsidering whether inclusion of such general concepts would necessarily benefit the data modeling community, which provides the backbone to data-entry forms to help clinicians record patient information. In addition, it might be worthwhile for the terminology community to consider the inclusion of pre-coordinated or post-coordinated expressions to help code the general direction of plantar reflexes.

6.1.3 The Visual Acuity Archetype

The visual acuity archetype was the third largest archetype amongst the four archetypes chosen for the evaluation study. The results returned by MoST for the visual acuity archetype achieved very high equivalence scores from the evaluators. Details of the evaluation results have been discussed in Chapter 5 (see Page 134). Despite the high scores, there were certain fragments in the archetype, which caused ambiguity amongst the experts, both at the level of the archetype as well as at the level of the SNOMED terminology. A snapshot of the fragments in the visual acuity archetype is shown in Figure 6.11.

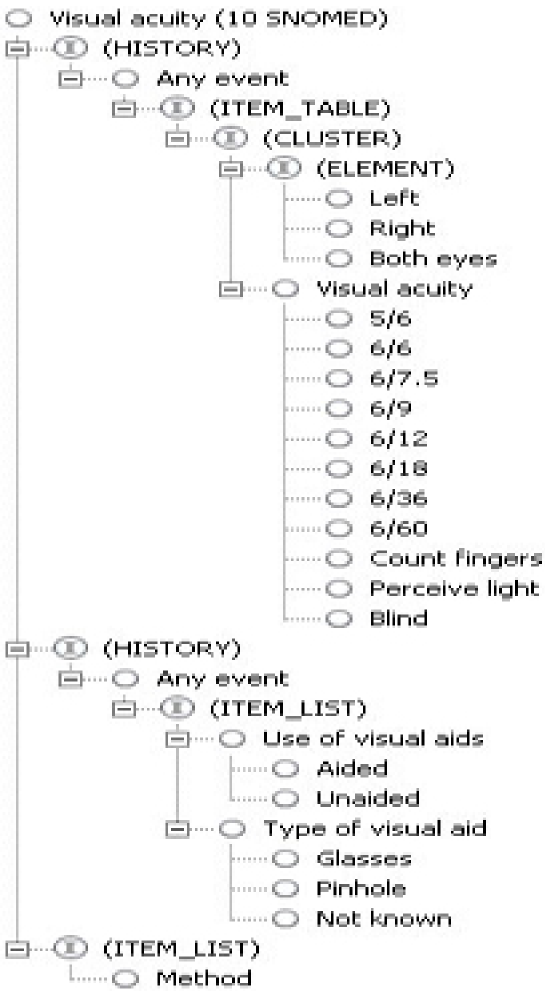


Figure 6.11: Archetype fragments represented in the Visual Acuity archetype.

6.1.3.1 Issues with the Visual acuity archetype

In general there were fewer issues with this archetype, as compared to the previous two archetypes. The issues with the visual acuity archetype mainly concerned the semantics of use of some of the fragments in the archetype. The main issues highlighted by the evaluators were:

1. *Reuse of labels in the same archetype:* Although reusing the same label for different fragments within an archetype is not inappropriate, the problem arises when proper context and annotations are not included for justification. As seen earlier, the tendon and babinski archetype had reused the reflex degree symbols such as '+', and '++' with no ambiguity. However, the visual acuity archetype reused the label 'visual acuity' without proper context or annotation, as can be seen in Table 6.7. The parent fragment in the archetype, which is also labeled 'visual acuity', does not provide proper context or a suitable definition, to justify the existence of two similarly named fragments in the same archetype.

Archetype Frag.	**Term Definition**	**Parent Frag.**
Visual acuity	The functional acuity of vision, aided and unaided	ROOT FRAGMENT
Visual acuity	Measurement of visual acuity	Any event

Table 6.7: Reuse of label 'visual acuity' in the Visual Acuity archetype with inadequate context to justify the existence of duplicate fragment labels.

   In view of the ambiguous reuse of labels, the evaluators proposed a restructuring of the archetype to eliminate redundancy. Even if it is a case of unnecessary duplication of fragments in the same archetype, rather than reuse of labels, the duplicate fragments must be removed to contain redundancy. Reduction of ambiguity and redundancy in an archetype will improve the overall quality of the archetype, and facilitate improved data mapping (See Section 6.2 for a further experiment performed to this effect).

2. *Issues with laterality:* The archetype addresses the issues of laterality, as can be seen in Figure 6.11. However, it is unclear from the archetype whether the recording of the acuity scores such as '6/6', '6/12' and so on, will require to be associated with any specific location (left, right, both eyes) or not.

   The results returned by MoST, as shown in Figure 6.12, include SNOMED codes such as `o/e - visual acuity L-eye =6/6 (finding)[163951003]` and `o/e -`

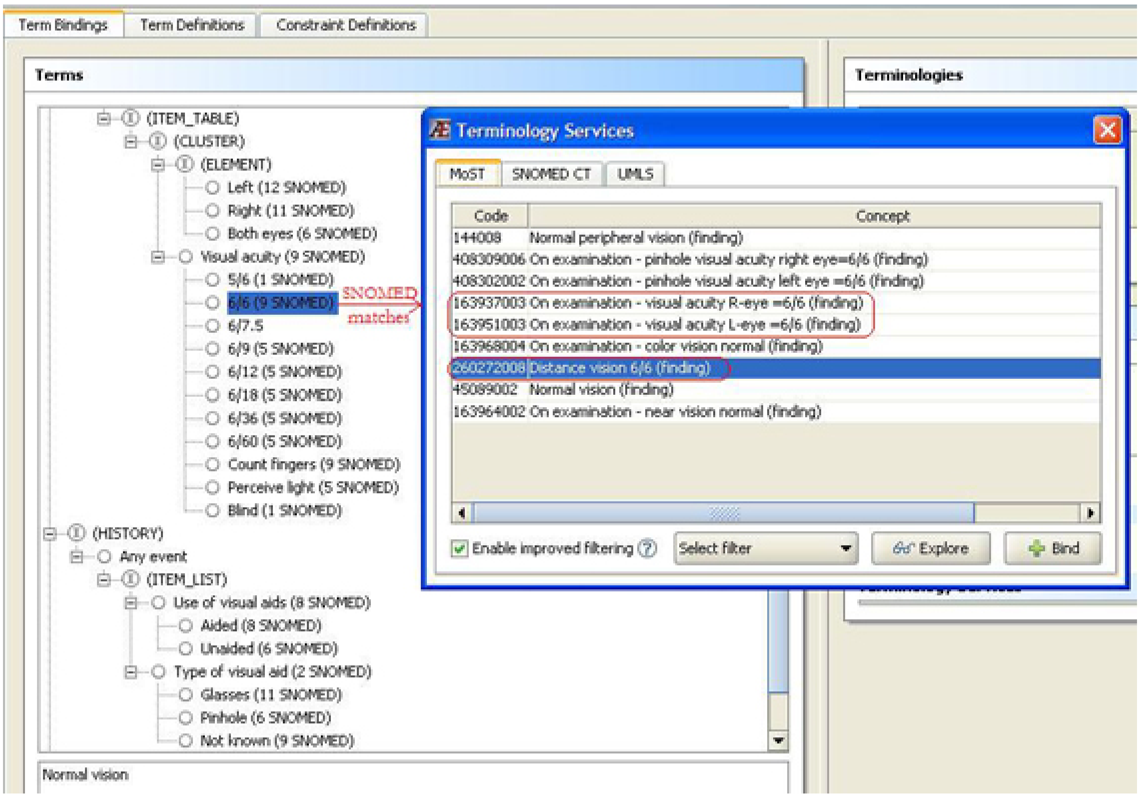


Figure 6.12: SNOMED results returned by MoST for the archetype fragment ‘6/6’. The SNOMED matches include laterality-specific codes.

`visual acuity R-eye =6/6 (finding)[163937003]`. These codes provide the possibility to map the archetype fragment ‘6/6’ along with laterality information i.e. left or right eye. However, the archetype model does not indicate which eye the score will be associated with. In the absence of further information, the evaluators selected the concept `Distance vision 6/6 (finding)[260272008]` to allow mapping to a more general SNOMED concept. It was assumed that the laterality would be post-coordinated to the finding at a later stage, if at all.

Laterality issues also existed for non-fractional fragments such as ‘count fingers’, ‘perceive light’, and ‘blind’. It was unclear whether laterality of the eye was required to be taken into account or a general result would be sufficient. Once again, in the absence of further information, the general concepts were given a higher equivalence score.

3. *Suggested renaming of fragments:* The fragment ‘blind’ in the archetype was suggested to be renamed to ‘does not perceive light’ or ‘no perception of light’ to enable it to map to the SNOMED concept `Visual acuity, no light perception`

`(finding)[63063006]`. The code `63063006` was considered to have a higher equivalence to the code `277675000`, which represented the concept `Blind (finding)`.

4. *Intended use of fragments:* Most of the evaluators had issues determining the intended use of some of the archetype fragments. The fragments that majority of the evaluators considered had ambiguous intent are listed in Table 6.8.

Archetype Fragment	**Term Definition**	**Parent Frag.**
Use of visual aids	Aids applied to vision	Any event
Aided	Visual function aided with corrective lenses	Use of visual aids
Unaided	Vision is not aided	Use of visual aids
Method	Description of the method used	Visual acuity
Glasses	Optical lenses to correct visual acuity	Type of visual aid

Table 6.8: List of archetype fragments and their definitions in the Visual Acuity archetype. Definitions did not clearly describe the intended use.

The main query the evaluators had with the fragments 'aided' and 'unaided' was whether they were used in the capacity of boolean values to suggest whether a visual aid had been used during the observation (i.e. True/False). In such an event, the parent fragment 'use of visual aids' would best be mapped to the SNOMED concept `Visual aid (procedure)[266729006]` with qualifier values for `Unaided (qualifier value)[255367001]` and `aided` [Note: There is no SNOMED concept present for the term 'aided'].

Archetype Frag.	**SNOMED Concept Name**	**SNOMED Concept Id (code)**	**SNOMED Concept Category**
Glasses	Eye glasses, device	50121007	Physical object
	On examination - visual acuity with glasses	408333000	Finding
	Contact lenses, device	57368009	Physical object

Table 6.9: Range of SNOMED concepts returned for the archetype fragment 'glasses'.

Similarly, the use of the fragment 'method' also remained ambiguous. As shown in Table 6.8, the definition of the fragment is *description of the method used*. The fragment 'method' in this case could be used interchangeably with the 'type of visual aid' used to measure the acuity level. Therefore, it was unclear how the additional use of a fragment 'method' would enhance the clinical recording of the visual acuity of the patient. The suggestion was to use either the fragment 'type of visual aid' or 'method', with preference to the use of the former.

Finally, the definition 'type of visual aid' of the fragment 'glasses', did not conform with the context in which it was used. The definition indicated the *optical lenses used to correct visual acuity*. The definition led to confusion on which of the three SNOMED concepts shown in Table 6.9 would be most appropriate for mapping to the fragment. At first glance, the evaluators selected the code `50121007`, as a physical device was considered to be the intended categorisation of the archetype fragment. However, the definition led to ambiguities whether a `finding` code was required such as `408333000` or whether a contact lenses physical device such as `57368009` would be more appropriate. At the end, the concept `Eye glasses, device` got the highest equivalence score of 56% amongst the three contending concepts. The score would have been higher if the ambiguity did not exist. The same issue did not exist for the other fragments modeled as types of visual aid, as the definitions clearly stated that a code belonging to the physical device category was required.

6.1.3.2 SNOMED Issues with Visual acuity archetype

Some of the issues encountered with the coverage in SNOMED CT of the visual acuity concepts are listed below.

1. *Ambiguous hierarchy:* The hierarchies in SNOMED pertaining to the modeling of acuity scores such as '6/6', '6/12', and others raised issues of logical ambiguity. For instance, as shown in Part(a) of Figure 6.13, the distance vision of 6/6 is placed in three separate sub-hierarchies of the concept `Visual acuity finding (finding) [260246004]`. Since the concepts `o/e - visual acuity R-eye =6/6[163937003]` and `o/e - visual acuity L-eye =6/6 [163951003]` belong to a *distance vision finding* of a specific eye (left/right), it was considered logical to place the laterality-specific scores as children of `Distance vision finding [260269001]` instead of as sibling concepts. However, it can be seen in Part(b) of Figure 6.13, that the same logic has been applied in SNOMED for the observable entity concepts `Distance visual acuity - left eye [386716001]` and `Distance visual acuity - right eye [386714003]`.

Therefore, given that the correct logic was applied to the `observable` sub-hierarchy, it is unclear why SNOMED terminologists could not have adopted the same logic for the `finding` sub-hierarchy.

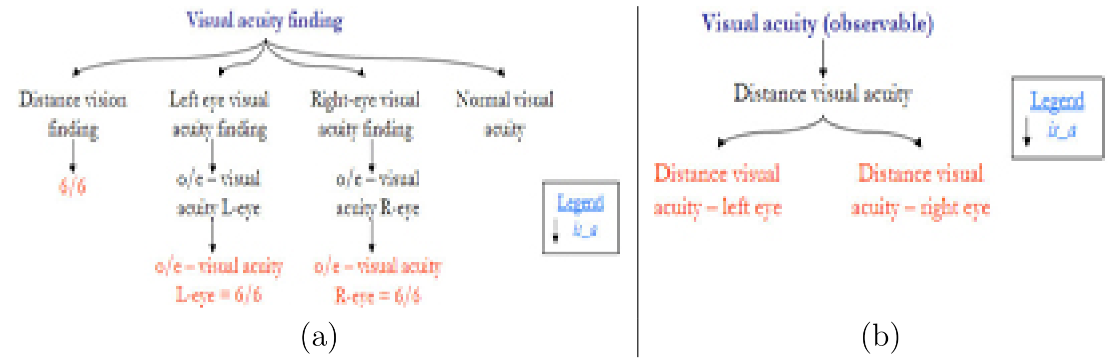


Figure 6.13: SNOMED hierarchies for *visual acuity* concepts categorised as (a) findings, and (b) observable entities.

2. *Improvement of definition:* The definition of `Normal visual acuity (finding) [82132006]` in the hierarchy has been shown in Figure 6.14. It is clinically known that the acuity vision of the eye is considered normal if the score recorded for an eye or both the eyes is 6/6. In that sense, the definition is incomplete, as can be seen in the figure. However, since SNOMED has defined the concept as a *primitive*, it has already identified the need to add further concepts to *define* it more precisely. The suggestion is to add some sort of relationship to the concept definition, such as with the `Distance vision 6/6 (finding) [260272008]` concept or a laterality-specific 6/6 concept, to make the normality of vision measurable.

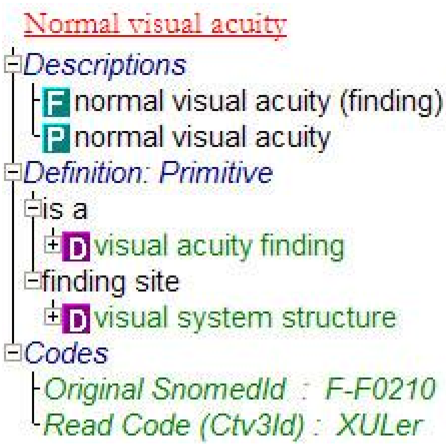


Figure 6.14: SNOMED definition for concept *normal visual acuity*.

3. *Missing codes:* There exists a code `Unaided (qualifier value)[255367001]` in SNOMED to qualify a concept such as `Visual aid (procedure) [266729006]`. However, an `aided` qualifier concept does not exist in SNOMED. It is recommended to include this concept in SNOMED to enable post-coordination, where appropriate.

   Another concept missing in SNOMED is visual acuity in *both eyes*. There exist concepts for defining visual acuity of the left and right eye but no concepts for 'both eyes'. General SNOMED codes are required to be added, such as `Both eyes visual acuity finding (finding)`, `Visual acuity - both eyes (observable entity)`, `Distance visual acuity - both eyes (observable entity)`, and/or `Near visual acuity - both eyes (observable entity)`. Specific scores observed for the visual acuity of each eye (left eye and right eye) can then be mapped to existing SNOMED concepts.

4. *Possibly missing synonyms:* It was found that although the code `266729006` conceptualises the `visual aid (procedure)`, a synonym for *optical aid* does not exist in the concept group. A near concept `Optical aid used (attribute) [246142008]` exists in the hierarchy as an `(unapproved attribute)`. The suggestion is to add a new concept *optical aid* as a synonym of `visual aid` to enrich the terminology further.

6.1.4 The Body Weight Archetype

The body weight archetype was the smallest archetype and, as a result, generated the least number of reasons for ambiguity. This means that the fewer issues raised were due to the fewer archetype fragments causing ambiguity, rather than because the fragments were generally of better quality. The main reason for the ambiguity was the lack of compatibility between the archetype and SNOMED, to model concepts related to the body weight of a person, neonate or fetus. The incompatibility resulted in very few pre-coordinated SNOMED matches. Details of the evaluation results are discussed in Chapter 5 (Page 134). Briefly, the mapping coverage was 70% and the trust score was 0.7. A snapshot of the archetype is shown in Figure 6.15.

The issues highlighted in the body weight are very few, as compared to the remaining three archetypes. However, as stated earlier, the body weight archetype was comparatively smaller in size, thereby generating fewer comments from the evaluators.

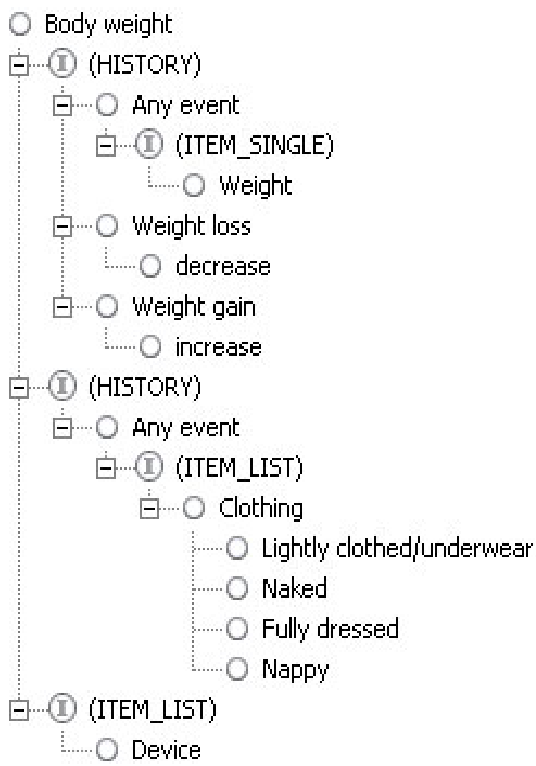


Figure 6.15: Archetype fragments represented in the Body Weight archetype.

6.1.4.1 Issues with the Body weight archetype

1. *Inadequate labeling of fragments:* The fragments used in the archetype to record the amount of clothing worn by the patient at the time of recording the body weight were inadequately labeled and annotated, as shown in Table 6.10.

Archetype Frag.	**Term Definition**	**Parent Frag.**
Lightly clothed/ underwear	The amount of clothes worn at the time of weighing	Clothing
Naked	Without clothes	-do-
Fully dressed	Clothing which may add significantly to weight - such as shoes and normal clothing	-do-
Nappy	A baby with nappy on - adds significant weight	-do-

Table 6.10: List of inadequately labeled archetype fragments and their definitions in the Body weight archetype.

As shown in Table 6.10, it was unclear from the fragments and definitions whether the author of the archetype intended to record a `(finding)`, indicating whether the patient was with or without clothing, or whether the intention was to record the actual weight of the patient along with the weight of the clothing, if any. The evaluators considered the latter to be a more appropriate use of the fragments although it was not explicitly stated by the archetype. However, there were no pre-coordinated codes in SNOMED to map to the archetype fragments. Table B.10 in Appendix B shows the post-coordinated expressions suggested by the evaluators, which could be used to code the fragments.

6.1.4.2 SNOMED Issues with Body weight archetype

The issues that concerned the coverage of body weight related concepts in SNOMED CT were:

1. *Insufficient definitions:* The top four concepts that were chosen by the evaluators to map to the archetype fragment 'body weight' are shown in Table 6.11. They were `usual body weight (observable)`, `body weight measure (observable)`, `weight and bodymass assessment procedures (procedure)`, and `body weight (observable)`. However, the SNOMED concept definitions were insufficient to determine the semantic equivalence with the 'body weight' fragment, with the definition *total body weight - a surrogate for naked body weight.* The SNOMED definitions have also been shown in Table 6.11. The lack of semantic clarity led to lack of consensus on the semantic equivalence of the codes. However, of the four SNOMED results, the `Usual body weight (observable entity) [363809009]` concept got the highest equivalence score of 85%.

SNOMED Concept	**Concept Definition**
Usual body weight(observable entity)[363809009]	*is_a* Reference weight[248350002]
Body weight measure (obs. entity)[363808001]	*is_a* Body weight characteristic [363804004]
Weight and bodymass assessment procedures (procedure) [225171007]	*is_a* Body measurement[54709006] *method* Measurement - action
Body weight (observable entity) [27113001]	*is_a* Body weight measure[363808001]

Table 6.11: Insufficient definitions available in SNOMED for *body weight* concepts.

2. *Missing concepts:* The SNOMED terminology has several codes to conceptualise the notion of *body weight* and the *actual weight in centiles.* However, it lacks concepts such as 'weight of body wearing nappy', or 'weight of body wearing

clothing', which could have provided pre-coordinated matches to some of the archetype fragments. Alternatively, SNOMED could have provided partial codes such as 'weight of nappy', or 'weight of clothing' that could have been post-coordinated with codes 363809009, 363808001, or 27113001 (for a complete list of alternatives, see Table 6.11).

6.2 Revising the Histology pap archetype as a gold standard

As stated in Chapter 1 (see Section 1.2.6;Page 32), in the absence of any gold standard to judge the performance of the MoST system an extra piece of research was conducted outside the original intended research ambit, and has since been published [76]. The objective was to test how well the MoST system performed by reducing the ambiguity of the archetype fragments, and by aligning them closer to the SNOMED style of modeling cervical smear concepts. The revised histology pap archetype was authored with the help of a clinical researcher in the department, although it would have been ideal to have authored the archetype using an independent panel of experts. This was not possible due to lack of resources and time. Despite the shortcoming, all efforts were made to replicate the scenario had an independent expert been used for the revision task. The revised archetype served as a gold standard, and provided the upper limit that MoST could achieve in terms of percentage of mapping coverage and the trust score.

The experiment was conducted with the view that, ideally, if data models and terminologies were to integrate at the content level, the two models would need to conform to the requirements of each other. The research took the task of conforming the archetype fragments to the SNOMED concepts, as it was easier to author new and modify existing archetypes, than make any changes to SNOMED CT, which is illegal without the required license(s) and permission(s) possible only by sending change requests to IHTSDO.

Since the mapping coverage of the histology pap archetype was the lowest at 48.6% amongst the four evaluation archetypes, this archetype was chosen for improvement and re-testing. Due to the issues with the histology pap archetype, it was revised to address a majority of the issues. The experiment was then performed to determine whether remodeling the histology pap archetype would result in any improvement in the evaluation results, such as the rate of (mapping) coverage and the trust score. In

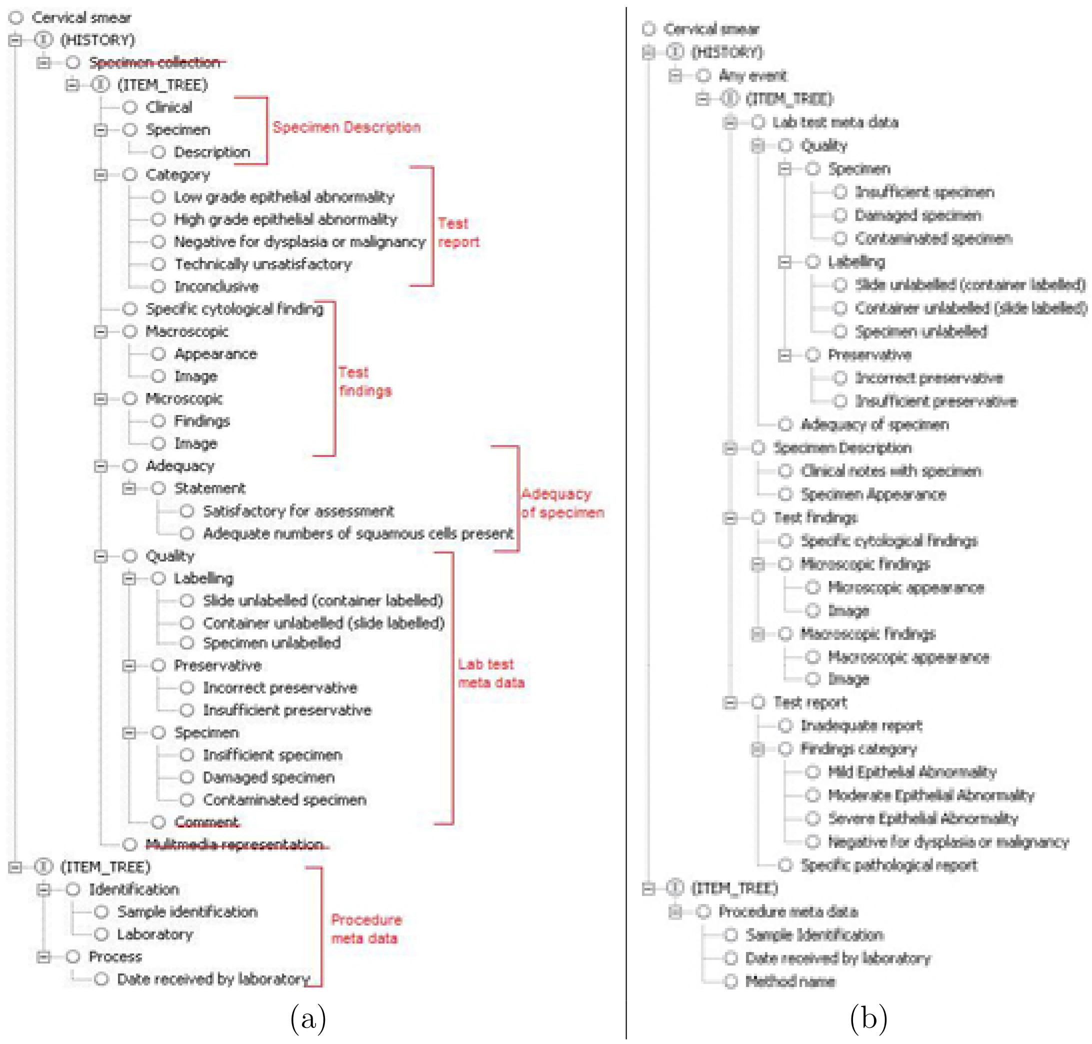


Figure 6.16: Histology Pap archetype: (a) Changes made to original archetype, (b) Revised archetype.

addition, this experiment was performed with the histology pap archetype because it had the most variable set of fragments (i.e. general and specific) compared to the remaining three archetypes.

The shortcomings in the original histology pap archetype were taken into account at the time of revision. Figure 6.16(a) graphically presents the broad level changes that were made to the original archetype. Part (b) of Figure 6.16 displays the revised version of the original histology pap archetype.

All data pertaining to a particular clinical scenario were grouped under a common heading to enable better context information. It also enabled better classification of the fragments, by placing them into common categories, making the archetype more intuitive, not only to the user but also to the MoST system. Some of the groupings used were 'procedure generic data', 'lab test generic data', 'test findings', and 'test report'. The existing archetype fragments were then either reorganised to place them into the most appropriate category or renamed before the reorganisation. In addition, clearer annotations were also provided as `term definitions`, to increase the chances of a match with a SNOMED code. The generic data common to any procedure was included in the same archetype, to maintain some level of consistency with the original.

6.2.1 Summary of changes to Histology Pap archetype

- The archetype was specified as describing elements of a procedure and its related findings.
- Generic data related to any procedure/test were separated from the rest of the archetype, specific to recordings of the PAP smear.
- The archetype hierarchy was changed to reflect a general pattern applicable to any clinical procedure or test, with some data specific to the PAP smear procedure.
- The labels and definitions of the archetype fragments were modified to reflect more clearly their intent. Suggestions from the evaluators were taken into consideration when resolving ambiguous fragments. For example, the fragment 'Clinical' was modified to 'Clinical Notes' to describe the 'Clinical history sent with the specimen'. A tabulated comparison of some of the changes made when revising the archetype are stated in Table 6.12.

Original archetype	**Revised archetype**
Low grade epithelial abnormality	Mild epithelial abnormality
High grade epithelial abnormality	Moderate epithelial abn.
	Severe epithelial abn.
Macroscopic	Macroscopic findings
Appearance	Macroscopic appearance

Table 6.12: Sample of changes made to the original archetype in the revised version of the histology pap archetype.

6.2.2 Evaluation of revised Histology Pap Archetype

The revised archetype was sent to the MoST system and the results were evaluated by three clinical experts. The *coverage* of the fragments increased from the original 48.6%

Archetype Frag.	**Appropriate matches in SNOMED CT**
Mild epithelial abnormality	Cervical intraepithelial neoplasia grade 1 (disorder) [285836003]
	Mild epithelial dysplasia (morphologic abnormality) [33288004]
Clinical notes with specimen	Clinical history and observation findings (finding) [250171008]
	Clinical history/examination observable (observable entity) [363788007]
Microscopic findings	Microscopic specimen observation (finding) [395538009]

Table 6.13: Improved matches in SNOMED CT for archetype fragments in the revised histology pap archetype.

to 80.55% (29/36*100). In addition, the *trust score* given by the experts increased from a low of 0.45 to 0.8. The high ratings were primarily due to the unambiguous fragments and structure in the archetype, as well as because of much clearer term definitions, which clearly stated the intended use of the fragment. Some of the matches returned are shown in Table 6.13.

Despite the revision of the histology pap archetype to resolve the issues that had been faced with the original archetype, there was a ceiling on the evaluation results that could be achieved. Despite the reduction in the ambiguity of the archetype, the rate of coverage increased only to a maximum of 80.55%. A complete 100% could not be achieved, as the result was not entirely dependent on the quality of the archetype alone. Two other important deciding factors were (i) the coverage of the archetype fragments in SNOMED, and (ii) the ability of MoST to pick out "all" the semantically equivalent SNOMED codes in the *term finding* process. However, it is not possible to presume that there will be no "missed opportunities" by an automated system. Neither is it possible to always find 100% coverage of all fragments from a particular archetype in SNOMED.

The intention of the exercise was to obtain as many equivalent codes as possible by continually improving the search mechanism until the desired stability and quality of results was achieved. The MoST system was developed to the extent that a substantial analysis of the results could be performed providing sufficient insight into the issues involved in attempting to integrate content. Bearing this in mind, the increased rating of 80.55% provided the upper limit of the performance of MoST with respect to the histology pap archetype. Contrary to the results obtained in this section after revision, the original tendon babinski archetype achieved a coverage of 90.9% indicating both good quality archetype fragments and corresponding SNOMED concepts.

6.3 Summary

This chapter discussed the issues raised during the mapping of the archetype fragments of the four evaluation archetypes to SNOMED concepts. In summary, there were issues with the quality of the archetype fragments and their coverage in SNOMED CT. In addition, there were also issues with the quality of the SNOMED hierarchies and concept definitions.

The coverage of the archetype fragments in SNOMED CT was best for the *tendon and babinski* archetype. The coverage was also high for the *visual acuity* archetype. The *body weight* archetype generated the least number of issues due to its size and number of fragments.

The most problematic archetype was the *histology pap* archetype, as majority of the archetype fragments were poorly labeled or contextualised. Therefore, as an additional part of the research, an experiment was performed to improve the quality of the histology pap archetype to test any improvement in the mapping results. Revisions to the archetype included an improvement in the labeling, annotations, and archetype hierarchy. Despite an improvement in the new evaluation results, there was an upper limit on the mapping performance. The highest coverage that could be achieved given the revised histology pap archetype was 80.55%.

In conclusion, the concerns were primarily about the archetypes and SNOMED rather than the procedures used by MoST. The evaluators made a number of suggestions for changes in the archetype and SNOMED models, which have been discussed in this chapter. They were positive about the MoST system itself and its potential to aid as a quality-assurance tool besides a mapping tool.

The following chapter will focus once again on issues with archetypes and SNOMED, but in a general framework. The discussion will not be specific to the four evaluation archetypes, although references might be made and exemplars might be selected from any of the archetypes discussed in the thesis.

Chapter 7

General Issues, Criteria and Guidelines

This chapter is the fourth and last in the series on the discussion of the research programme, which began with Chapter 4. The previous chapter discussed the specific issues that arose in evaluating MoST using each of the four evaluation archetypes with SNOMED. This chapter will provide a general framework for discussing the issues of mapping archetype fragments to SNOMED concepts. The chapter will also articulate the criteria, guidelines and suggestions for improving compatibility between the two models (i.e. archetypes and SNOMED). The issues will focus on the difficulties arising when *integrating* or using two separate models.

The criteria formulated in this chapter have been based on the issues that arose during the model integration process. In addition, the guidelines/recommendations stated toward the end of the chapter are intended to meet the criteria for good modeling. The main purpose of this part of the research work is to stress the importance of following strict guidelines when modeling clinical content in both data models and terminologies in order to reduce the number of issues arising from model integration efforts. Although the guidelines have been suggested for *open*EHR Archetypes, they should be applicable to any other data model. The guidelines stated for terminologies are aimed at improving their computability. Although the guidelines are applicable to any terminology, references are made specific to the SNOMED model. Collaboration work with a clinical expert to critique the "entire" SNOMED content would be desirable in future. However, a "partial" discussion of the SNOMED content has been provided in this chapter. The discussion is based on recommendations for inclusion of certain concepts in SNOMED (based on the summary of observations stated in Chapter 6).

7.1 Integration Issues and Modeling Criteria

The introduction chapter had stated that the main problem arising in the attempts to conform all recorded clinical data to a single terminology standard was the difficulty in conforming to the different naming and structuring conventions. In addition, there were also the differences in the modeling objectives and the methods of representing data in the two models (i.e. Archetypes and SNOMED). The evaluation chapters 5 and 6 presented a practical account of the problems in trying to resolve these differences to achieve the common objective of controlling the vocabulary used to record clinical data.

The issues stated in this chapter presents the criteria for authoring good archetypes and SNOMED models. The discussion will begin with the issues arising from integrating the archetype and SNOMED models. The primary reason for the problems faced when integrating the two models is because they are developed and maintained by separate groups with different objectives in mind. This often leads to differences in the naming, structuring, and categorisation of clinical content. The following subsection will provide further detail on the integration issues. This section will then progress to a discussion of the issues with Archetypes and SNOMED, individually.

7.1.1 Issues Integrating the Archetype and SNOMED Model

The model integration process discussed in the study included two main stages: (i) the term finding process, performed with the automated MoST system, and (ii) the data mapping process, assisted by the MoST system but performed manually by the clinical experts. However, there were several issues that were brought to light during the two stages to achieve the integration objective.

One of the main issues in integrating the content of archetypes and SNOMED was concerned with 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 when mapping data to controlled vocabularies and terminologies [28] [102] [46].

This section discusses the issues that arose when integrating the clinical content in the archetype and SNOMED models. Briefly, the issues that form the basis for the model integration criteria include (i) difference in the modeling techniques, (ii) categorisation of data to reflect their intended semantics, (iii) importance of semantic results over lexical results, and (iv) importance of context information.

7.1.1.1 Different Modeling Techniques

The fundamental difference in archetypes and SNOMED arises due to the difference in their modeling techniques. Archetype models represent only fragments required to record a particular clinical statement. Therefore, the archetypes include all those fragments necessary for an archetypical record, irrespective of their logical dissimilarity. This means that data belonging to different semantic categories might be included in the same archetype in a *containment hierarchy*, to achieve the objective. The SNOMED terminology is not based on the same principle.

The SNOMED terminology is an ontology[1], which includes all those concepts that relate to each other logically within a particular domain. Therefore, any concept that does not fit logically in a given domain is not included in the terminology. In addition, the *subsumption hierarchies* in the SNOMED model place all concepts belonging to the same semantic category in the same hierarchy. As a result, it is important to bear in mind the difference in the archetype and SNOMED hierarchies and structure when dealing with the issue of integration.

7.1.1.2 Data categorisation to reflect intended semantics

Since archetypes and SNOMED are developed and maintained by separate groups, the data is often categorised into different semantic types. The difference in categorisation is mainly because archetype authors do not have SNOMED in mind when modeling the archetype fragments, and vice versa. This often leads to problems in resolving the correct categories that the data/clinical content should correspond to in both models, to have mutual agreement of semantics.

At present, there only exists a top-level semantic type category to which an entire archetype model belongs. Individual fragments in the archetypes are not assigned semantic types, such as finding, procedure, body structure, and observation. Sections 7.1.2.1 and 7.1.2.2 provide further details on the categorisation discussion. Therefore, in the absence of specific categories at the fragment level indicating the exact semantic type(s), matches returned by MoST belonging to an agreed range of SNOMED categories were accepted as being semantically appropriate. For instance, in the archetypes belonging to the *open*EHR Observation subclass, three main SNOMED categories were identified, to which most archetype fragments found semantic matches. These SNOMED categories were `observable entity`, `procedure`, and `clinical finding`. The assessment was made by manually inspecting the most

[1]An ontology is a specification of a conceptualisation [47]

common category(ies) to which relevant archetype fragments mapped in SNOMED. In addition, the initial hypothesis that the top-level category of the archetype should extend to its individual fragments was also discarded. Therefore, it was not required that archetype fragments from an Observation type archetype should necessarily map to a SNOMED `observable`. SNOMED concepts from the `finding` category were the next closest match, followed by results from the `procedure` category. The relevance of a match was determined partially by the MoST system and partially by the clinical experts evaluating the results. However, this assumption does not intend to limit the matches to these SNOMED categories alone. Matches from other SNOMED categories might also be semantically equivalent depending on the intended use of an archetype fragment.

A repository specifying the 'intended meaning' of the individual archetype fragments using the 19 SNOMED category names was generated for each of the archetypes used in the research. Figure 4.3 (on Page 104) shows that the filtering stage of the MoST process utilises the 'intended meaning' repository. In addition to the concepts from the three main SNOMED categories i.e. observable entity, finding, and procedure, which usually map to the observable archetype fragments, there were certain extensions to the list. It was also noted that SNOMED concepts from certain other categories such as `situation` (earlier known as 'context dependent category'), `qualifier value`, `attribute`, and `disorder` also often found semantically equivalent matches with the archetype fragments. The set of SNOMED categories which served as a range for the mapping process in the MoST system differed from the hypothetical comparables between the *open*EHR subclasses and SNOMED categories shown in Figure 7.2.

7.1.1.3 Semantic results as against lexical results

The model integration process relies on resolving the semantics of the data/clinical content in the two models. Therefore, any procedure developed to achieve integration of data between two or more models should take into consideration all context and non-context information available about a particular concept. Section 7.1.2.3 lists the various sources available in archetypes, to derive the semantics of an archetype fragment. Since SNOMED did not have more than one source for determining the semantics, the discussion relates to archetypes. The only source of semantics in SNOMED was the concept definition provided and the category to which it belonged in the terminology.

Section 7.1.2.3 discussed how the archetype fragment and its definition helped in determining the intended use of the fragment with the help of an example. In the *blood* *film* archetype, the archetype fragment '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 clinical experts on general lexical lookup of the results in CliniClue were of the opinion that the second match was more relevant. However, on further examining their SNOMED definitions and FSNs returned by MoST, it became obvious that the 'haemoglobin concentration' matches were synonyms for two different FSNs, viz. `Dipstick assessment of hemoglobin concentration (procedure) [302781000]`, and `Finding of hemoglobin concentration, dipstick (finding) [365809007]`, shown in Figure 7.4. Both the matches were associated with the device 'dipstick', which was not necessarily the intended method to determine the haemoglobin concentration. The absence of non-device haemoglobin concentration concepts might be an error of exclusion in SNOMED, which requires correction. Similarly, the archetype fragment 'total' used to record the total apgar score resulted in MoST returning two different SNOMED concepts 'total (qualifier value) [255619001]', and 'total apgar score (observable) [249228009]' to avoid the possibility of eliminating a relevant result.

The issue of ambiguity, as to whether the term itself or its definition and context in the archetype were to be weighted higher, was resolved by including all result variants. This led to a higher rate of false positives in some cases but reduced the number of false negatives, which was considered more important. Therefore, both 'haemoglobin', as well as 'haemoglobin concentration' matches were included as results if they met other filtering criteria. On evaluation of the results, both the 'heamoglobin concentration' results were rejected by the experts as inappropriate mappings for the archetype fragment 'haemoglobin'. Therefore, by providing the experts with more information about the SNOMED results rather than a simple, non-informative lexical list, a reduction of incorrect mappings can be achieved.

7.1.1.4 Context information

Archetype modeling is based on object-orientation by associating the various elements in an archetype to some entity or attribute in the *openEHR* Information Model. The fragments in an archetype are present as either elements or values. As a result of the data encapsulation feature of archetypes, the values of an element do not convey the whole semantics without reference to its enclosing element. This data encapsulation feature results in a post-coordinated style of modeling. Therefore, the fragments (i.e. archetype elements or values) have conceptual significance only when considered in context rather than in isolation. It is necessary for term finding and mapping procedures, such as MoST, to take into consideration the context information to increase the mapping coverage.

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 7.1. 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 fragments in SNOMED, it was important to include the context information. This helped in obtaining close pre-coordinated matches in SNOMED.

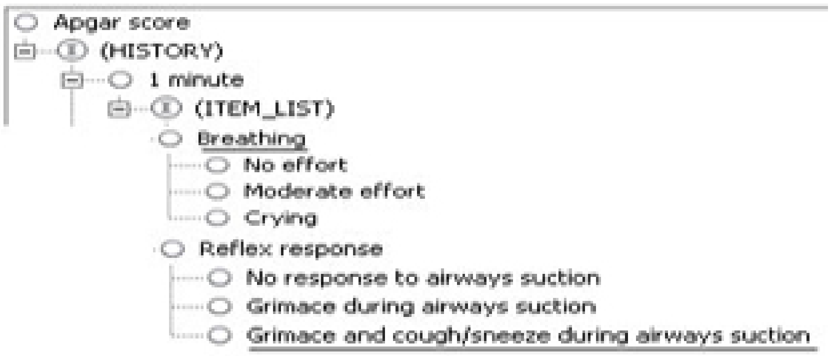


Figure 7.1: Context information taken into consideration for child values of fragments 'Breathing' and 'Reflex response' in the *Apgar score* archetype.

The tendon babinski archetype is also a good example of the effects of considering the context information of a fragment. Section 6.1.2 of Chapter 6 (Page 150), provided a detailed critique of the archetype. An example of the reuse of the same set of fragments, nested under different enclosing fragments, include the fragments representing the degree of reflex observed or found in a particular tendon region. The reuse of fragments in an archetype demonstrates the importance of considering the context of use. The SNOMED codes returned by MoST for the fragment 'Absent' as a value for the 'Biceps' reflex, differed from the fragment 'Absent' used as a value for the presence of reflex in the 'Triceps' tendon. For instance, one of the results returned for the biceps reflex absent was `On examination - biceps reflex absent (finding) [163828004]`. Similarly, a result returned for the triceps reflex absent was `On examination - triceps reflex absent (finding) [163820006]`.

However, certain fragments which are verbose or ambiguous can be unhelpful in finding sensible SNOMED matches despite considering the context. Some examples are the fragments 'Grimace and cough/sneeze during airways suction', as a constraint value of the enclosing fragment 'reflex response', in the *Apgar score* archetype shown in Figure 7.1, and 'Needs help but can do about half unaided', as a constraint value of the enclosing fragment 'Dressing/undressing', in the *Barthel index* archetype. The presence of such fragments in archetypes often leads to irrelevant or no matches. However, these searches expend valuable processing time to simplify the fragment.

7.1.2 Issues with Archetype Models

When critiquing archetypes with respect to mapping their fragments to controlled vocabularies and terminologies, it is important to focus on the *fragment/content layer* rather than the *schema/data structure layer* of the archetype. The same principle also applies to other modeling formalisms such as HL7 V3 Messages, CDA documents, or UML models. Therefore, the discussion in this section will focus on the 'content layer' specifications rather than the 'schematic' specifications of archetypes. In addition, the discussion will be with reference to the SNOMED CT terminology, as it was one of the case studies for the research programme. However, SNOMED could be replaced with any other terminology such as LOINC, ICD, and others, for achieving model integration.

The thesis draws on the fundamental principles of electronic record models stated by Rector *et al.* [93] [85]. The authors state that a model of the electronic record must be a faithful representation of what the clinicians have heard, seen, thought, and done [93]. While the thesis builds on these principles, it also presents a practical viewpoint of the essential criteria that make a data model usable, reusable, and faithful to the process of aiding in patient care. Since the criteria are based on the issues, the issues themselves represent the criteria that are required to help improve the modeling framework for integration purposes.

The main criteria (presented as issues) for good archetype content modeling presented in this section are (i) unambiguous categorisation of archetypes, (ii) proper separation of specific and general content, (iii) sufficient context information, and (iv) compatible naming conventions.

7.1.2.1 Categorisation of Archetypes

The approach adopted by *open*EHR to categorise the clinical statement models into the four ENTRY subclasses has proven to be particularly useful when attempting to work with SNOMED CT at the semantic level. Briefly, the four main ENTRY subclasses (from now on referred to as *archetype categories*) are Observation, Evaluation, Instruction, and Action. For the research, only Observation archetypes have been considered, as they are the most commonly used archetype category with the maximum number of examples. Descriptions of the four archetype categories have been stated earlier in Section 2.5.1 of Chapter 2 (Page 55). Despite definitions available in *open*EHR for the criteria that each of the four ENTRY subclasses aim to satisfy, there is no strict guideline available to author an archetype into one of the four categories. In addition, there is no quality assurance in place to detect mis-categorisation of archetypes and their subsequent archetype fragments. Intended meanings of the archetype fragments can be strongly controlled with proper categorisation of the archetype models, which can be very helpful to integration procedures.

Some of the SNOMED categories are useful parallels to the archetype ENTRY subclasses. As explained in Section 7.1.1.2, these include Observable entity, Procedure, Clinical finding, and Situation with Explicit Context (previously known as 'context dependent category'). These SNOMED categories appeared to be near semantic approximations to the archetype categories, but are not limited to this list. Brief descriptions of some of the more relevant SNOMED categories are available in Table 2.3(Page 73). Figure 7.2(a) presents a chart of the preferred semantic approximations between the archetype and SNOMED categories. However, it is not possible to restrict the approximations to only the three SNOMED categories - observable entity, finding, and procedure. During the MoST development and evaluation exercises it was found that there were more than one SNOMED category that could be placed in parallel to an archetype category. Part(b) of Figure 7.2 depicts a realistic approximation between the archetype and SNOMED categories. These approximations are hypothetical and have been arrived at on the basis of the definitions provided in the documentations, as well as the observations made. They are intended to provide an indication of the most common mappings that can be achieved at a category level.

Resolving issues with semantic categories

As stated earlier, there is no strict guideline used to categorise archetypes. The lack of formal guidelines presents a two-fold issue. At first, is the issue of categorising the archetype model itself. For example, it is unclear in which ENTRY subclass (Observation, Evaluation, Instruction, or Action) to place a clinical entry archetype. Secondly, the issue of categorisation also extends down to the individual archetype fragments in a particular archetype. All fragments present in an archetype model do not necessarily belong to the category to which the archetype model itself belongs. For example, the *autopsy* archetype belongs to the Observation (ENTRY) subclass. The purpose

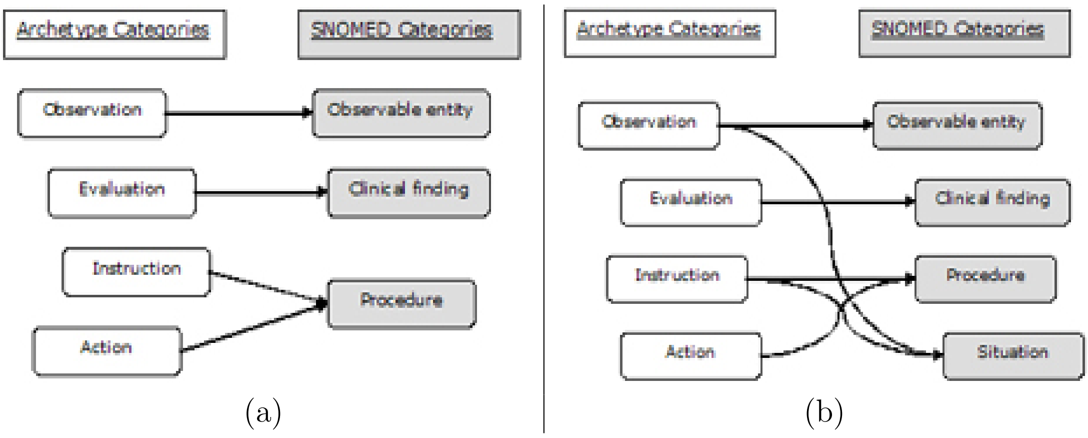


Figure 7.2: Parallels drawn between the Archetype and SNOMED CT Categories - (a) Preferred approximations, (b) Realistic approximations. **Situation* was known as *Context dependent category* in the earlier versions of SNOMED.

of the archetype is to represent the findings at the time of the autopsy. However, the 'cardiovascular system' fragment present in the autopsy archetype is not necessarily an Observable as well. The intended categorisation of the archetype author was to map the 'cardiovascular system' fragment to either the SNOMED category `body structure` or `clinical finding` rather than the SNOMED category `observable entity`.

Another example that highlights the use of fragments from different semantic categories comes from the *Apgar score* archetype. The purpose of the archetype is to represent the recordings of the Apgar index or assessment of the well-being of a newborn. The recording of the apgar score of a neonate at 1, 5 and 10 minutes from birth requires assessment of the breathing, color, reflexes, heart rate, and muscle tone. Scores from 0-2 are given for each of the assessment criteria and the total score is then calculated at the end of the entire assessment. Using SNOMED, the assessment criteria can be categorised as `observables` or `findings`, among other intended semantics, while the fragment 'total' can be either a `qualifier value` or an `observable` with a value [2]. Therefore, once again it is not necessary that all archetype fragments in an Observable subclass will necessarily map to a SNOMED Observable entity concept.

Due to the existence of archetype fragments mapping to different SNOMED semantic categories, it was not possible to constrain the mapping to the high-level

[2]The analysis has been drawn on the basis of the SNOMED data release of July 2006. There has been some revision in the coding of apgar concepts in the later release but this has not been taken into account. The reason is that the implementation and evaluation stages had been completed at the time of the 2007 data release, and no further modification tasks were undertaken, so as to proceed to writing-up the thesis.

archetype model category alone. For example, all fragments belonging to an Observation archetype were not constrained to map strictly to only the `observable entity` SNOMED category. Forcing such rules would only restrict and limit the range of SNOMED codes that could otherwise semantically satisfy the intended meaning and use of the archetype fragment. Several examples of such cross-category mapping are available in Chapter 6 and Appendix B. Therefore, it is important when working with two different models to take into account that loose semantics should be maintained to a certain extent, unless stated otherwise. Too much reliance on the categorisation of fragments in a particular model will result in fewer matches as semantically similar terms may be categorised differently in another model. Strict adherence to categories can only be maintained when working within the same model, where it can be assumed that the hierarchies have already been classified to be logically sound.

7.1.2.2 Specific and General Content

Rector *et al.* have stated in their paper on the frameworks for a medical record[85] that there are two levels of a data model viz. (i) the observational level where all such statements that pertain directly to the patient, and record what a clinician has heard, seen, or done are represented (i.e. specific content), and (ii) the meta level which is independent of the observation level and consists of statements indicating the role of the observations in the clinical dialogue, the decision making process, and the process by which the data was entered in the record (i.e. generic content).

Any good data model clearly separates data that is directly relevant and necessary to the recording of a particular clinical statement or scenario, i.e specific to the recording, from data that is optional and provides extra information about the clinical recording, i.e. is generic to any recording process. This separation is in addition to the separation of structure from content often known in business modeling as separation of the 'data structure' layer from the 'knowledge layer', which is also an essential criterion for a good model.

The research experience highlighted the importance of good content structure for mapping purposes. In particular, the histology pap archetype presented several problems during the MoST procedure. The poor separation of specific data from general statements also resulted in a low rate of mapping coverage and trust score. However, the extended piece of research work (i.e. revision of the original histology pap archetype), discussed in Section 6.2(Page 173), showed the effects of adopting and applying good archetype modeling techniques. The reorganisation of the archetype fragments in the revised histology pap archetype, to correctly separate the two data levels, helped improve the overall quality and intent of the archetype. In addition, it also significantly helped improve the evaluation results. Briefly, the mapping coverage increased from a low of 48.6% to 80.55%, whereas the trust score increased from 0.45 to 0.8 (for detail results see Section 5.5 and Section 6.2). Therefore, separating the *specific* content from the *general* clinical content, helped improve the effectiveness of integrating the histology pap archetype with SNOMED CT.

7.1.2.3 Context Information

There are two sources from which the context for a specific archetype fragment can be determined: (i) from the enclosing archetype fragment, as well as (ii) from the annotation defining the intended use and meaning of the fragment (known as 'term definitions' in Archetypes). Archetypes provide both levels of context in the modeling framework. During the model integration process, the MoST system took into consideration both levels of context to determine the semantics of the fragments for filtering and mapping purposes.

Resolving issues about semantics based on context source

When taking into consideration the context information of a fragment, it was difficult for the MoST system to determine which source of context should be given a higher weighting, to establish semantics. Once again, the issue arose due to lack of formal guidelines in the *open*EHR archetype formalism specifying the rules for establishing the preferred source of context. For example, in the *Apgar score* archetype, the enclosing fragments were the preferred source of context. This was the case with most of the archetypes, as the fragments are often placed in a post-coordinated expression. For instance, the archetype fragment 'breathing' is constrained by the values 'no effort', 'moderate effort', and 'crying' in the archetype, as shown in Figure 7.3. In a pre-coordinated form, the same concepts could have been represented as 'no effort breathing', 'moderate effort to breathe', and 'crying or breathing normally'. Therefore, when looking up matches in SNOMED for archetype fragments, it is necessary to include the context in which the particular fragment has been used in the archetype.

Another example taken from the *blood film* archetype demonstrates the use of the second source of context i.e. the term definitions. The fragment 'haemoglobin' in the archetype has a local term definition of "the mass concentration of haemoglobin", as shown in Figure 7.4. The author of the archetype was queried whether he would prefer a match for 'haemoglobin' or 'haemoglobin concentration'. The suggestion was that a

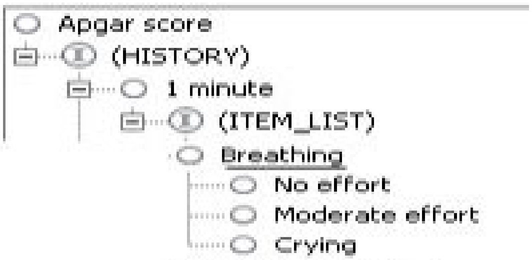


Figure 7.3: *Apgar score* archetype with the fragment 'breathing' providing context to its child terms. The enclosing fragment serves as a vital source of context information in archetype models.

match for the latter term would be closer to its intended meaning.

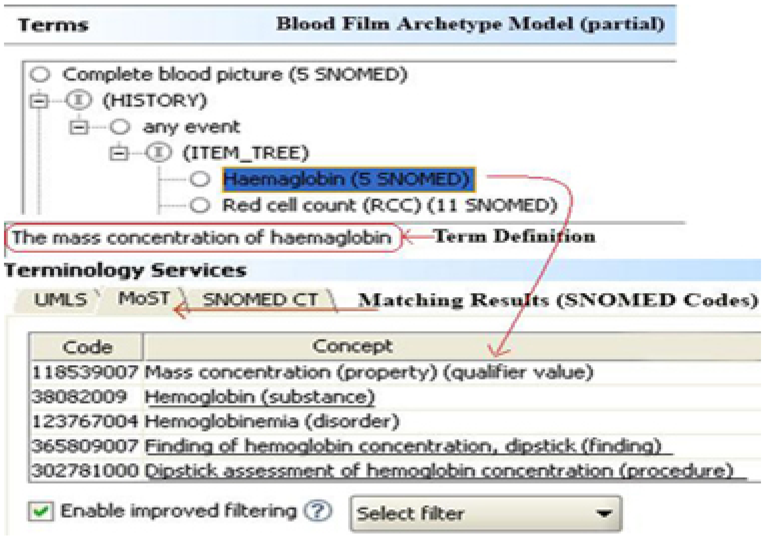


Figure 7.4: Haemoglobin-related issues in the *blood film* archetype.

In some other archetypes, assigning more weight to the term definition proved incorrect in the case of the *autopsy* archetype. Based on the above weighting, the fragment 'cardiovascular system' defined in the archetype as "findings of the pericardium, heart and large vessels" should have led to results based on the term definition. This would have provided mappings with suitable concepts from the `findings` SNOMED category. However, in this case the label chosen to represent the fragment i.e. 'cardiovascular system' provided the required semantics. This meant that a third source to establish the semantics i.e. the label of the fragment, and not the context information, was used to establish an equivalent mapping, which in this case was to a `body structure` SNOMED category.

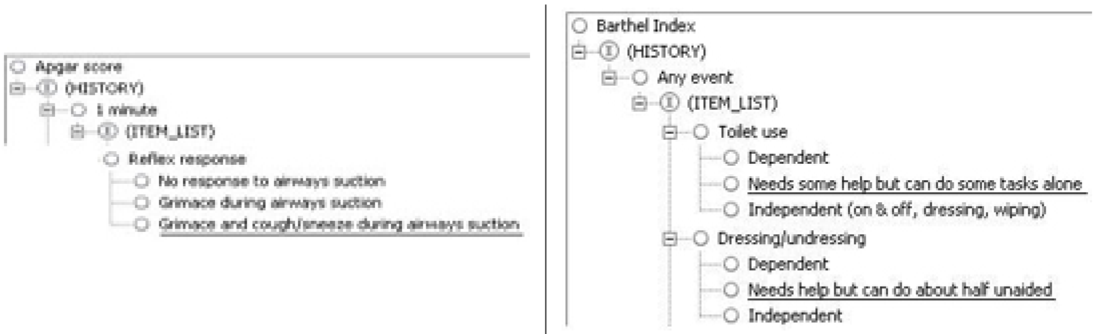
The examples aim to highlight the issues arising from the prevalence of the two sources of context, in addition to the labels of the archetype fragments, all of which provide semantic information to guide the filtering and mapping processes. The MoST system took into consideration both the sources of semantics, as well as the labels of the fragments, to assess the appropriateness of the SNOMED concepts. However, human intervention was considered crucial in confirming that the concepts returned by MoST were true semantic equivalents of the archetype fragments, before any data mapping was performed. Therefore, the study advocates that it is not advisable to depend on any single source for establishing the semantics of a fragment, as often the information is present in more than one place. Exclusion of any source may lead to unreliable and incorrect results.

7.1.2.4 Naming Conventions

The labels assigned to fragments in any model are vital to understand the concept that is being represented. However, the natural language selected to label a fragment could vary according to geographic location and local preference. The labeling of a fragment helps in providing a primary starting point to determine its semantics and its location in the overall hierarchy. Secondary information, such as term definitions, provides additional information and context to the fragment. Therefore, it is essential that suitable labels are assigned to the fragments to significantly resolve the problems faced by applications inferring fragment semantics. With the increasing need to make data interoperable in health care, the research revealed that conforming to the naming conventions of a particular terminology (to which the data is to be integrated with), helps in increasing the chance of making two different modeling formalisms interoperable. A look at the labeling and annotations assigned to the fragments in the revised histology pap archetype (see Section 6.2; Page 173), reflects a practical demonstration to resolve the problems arising due to poor labeling.

Resolving labeling and naming issues

Well labeled archetype fragments can often lead to exact or close matches with similarly labeled SNOMED concepts. However, certain archetype fragments, which are verbose or ambiguously labeled, can be unhelpful in finding sensible SNOMED matches despite taking into consideration the context information. Examples of poorly labeled archetype fragments from the *Apgar score* and *Barthel index* archetypes are shown in Figure 7.5. The presence of such fragments in archetypes often leads to irrelevant or no matches while expending valuable processing time to simplify the label.



(a) (b)

Figure 7.5: Verbose labels assigned to archetype fragments in (a) *Apgar score* archetype, and (b) *Barthel index* archetype leading to issues in finding suitable SNOMED matches.

Besides the issue of following proper naming conventions in archetypes, there also exists the issue of spelling errors. Although it may appear trivial for discussion, spelling errors in health records and supporting frameworks leads to wastage of valuable resources, as well as misinterpretation and misrepresentation of data, which might cause critical failures of systems and decision making processes.

When working with archetypes and SNOMED, the issue of resolving spelling errors and English language spelling variants utilised significant time and resources to resolve. For instance, 'haem**o**globin' using U.K. English, was spelt as 'haem**a**globin' in the *blood film* archetype. Such spelling errors can give rise to incorrect or no results. In addition, the same fragment sometimes has two different spelling variants in SNOMED. For example, the US variant was 'h**e**moglobin' while the UK variant was 'h**ae**moglobin'. Therefore, the MoST system took into account all spelling variants, as archetypes and SNOMED often use different variants of the same English language. Exclusion of variants may lead to exclusion of relevant results during searching and filtering of relevant terminology concepts.

7.1.3 Issues with the SNOMED CT Terminology Model

As stated in Chapter 2 (see Section 2.6.4; Page 69), SNOMED CT is the result of the merger of the SNOMED RT terminology and Read Codes. The multi-axial hierarchy and concept definitions in SNOMED CT, are easy to compute to determine the terminology subset and category to which a concept belongs. However, the initial merger of SNOMED RT and Read Codes has led to anomalies with the content and structure of the terminology model. The straight merger has resulted in duplicate concepts, insufficient and ambiguous definitions, as well as incongruent hierarchies. Despite being based on Description Logic, the full extent of the DL capabilities has not yet been exploited to resolve the issues present in the terminology.

Discussion of the issues with the SNOMED terminology will focus on the concept hierarchies, their categorisation and duplicate definitions. Specific focus on the content is deliberate, as the key area of research work is with the content of different models. The SNOMED July 2006 and 2007 releases have been checked to ensure the relevance of the discussion.

7.1.3.1 Concept Categorisation

SNOMED concepts are organized as subsumption or *is-a* hierarchies. Each subsumption hierarchy belongs to one of the 19 top-level categories, which defines the semantic type of a particular concept in the terminology. Section 2.6.4.2 defines some of the SNOMED categories that were important to the mapping between the archetype fragments and the SNOMED concepts. Although, theoretically, 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 during modeling of a new concept.

The SNOMED documentation states that concepts in the `observable` hierarchy represent a question or procedure which can produce an answer or a result [34]. On the other hand, concepts in the `finding` hierarchy represent the result of a clinical observation, assessment or judgment, and includes both normal and abnormal clinical states. For example, `colour of nail` is an observable whereas `gray nails` is a finding [34]. 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, which means it cannot be an observable. However, it can have a present/absent or positive/negative value, which would then categorise it as a `finding` thereby conflicting with the `observable entity` category assigned to it in the terminology.

Another issue that is a result of the straight merger of the SNOMED RT and Read Codes, is the existence of concepts in incorrect hierarchies/categories. For example, the SNOMED concept `Cervix tenderness absent (situation) [289816007]` belongs to the ‘situation’ category, as a subtype of `clinical finding absent (situation) [373572006]`. The justification for categorising the concept `289816007` as a situation is that the qualifier 'absent' modifies the concept 'cervix tenderness' (NOTE: this concept is not present in SNOMED). Therefore, stating that a cervix tenderness is absent is a 'situation' that might change, rather than a finding. However, the same logic has not been applied to several other concepts, such as `Cervical excitation absent (finding) [289820006]`, which is a subtype of `finding of cervical excitation (finding) [366314006]`. If the previous logic was to be applied, then the concept `289820006` should also belong to the 'situation' category, rather than the 'finding' category. A similar case of conflict can be seen with the two concepts `Vaginal tenderness (finding) [289594004]`, and `Vagina non-tender (situation) [289595003]`. Therefore, until all such 'finding' codes are moved to the 'situation' hierarchy, and reclassified to ensure the absence of any logical inconsistencies, working with SNOMED will continue to raise new issues.

There is also the issue of whether to include the SNOMED Preferred Term(s)(PT) and Synonyms(SYN) of a particular concept when identifying relevant archetype fragment matches. Besides the PT and SYN, SNOMED also has a main concept name, called the Fully Specified Name (FSN)[3] The approach adopted by MoST is to consider the FSN, PT, as well as SYN when determining the appropriateness of a SNOMED result. However, the inclusion of the preferred term and synonym labels gives rise to the issue of ambiguous labeling. For example, the label 'Red blood cell count' is used as a Synonym in the `finding` category, as well as a Preferred Term in the `procedure` category. However, the FSN of the concept in the `finding` category is `Finding of red blood cell count (finding) [365625004]` and in the `procedure` category is `Red blood cell count (procedure) [14089001]`, respectively. Such discrepancies in labeling SNOMED concepts very similarly often leads to problems in correctly interpreting the semantics of a concept, even though they belong to different categories.

The view of the study is that although some form of categorisation has been used when placing concepts in separate hierarchies, there are no rigorous computational resources that have been applied to check whether the concept categories strictly adhere to the guidelines laid down by the terminology. This often leads to incorrectly categorised concepts, which in turn leads to ambiguity in resolving semantics for mapping purposes.

[3]Fully Specified Name is the official name assigned to a particular concept in SNOMED CT.

7.1.3.2 Internal use of category names

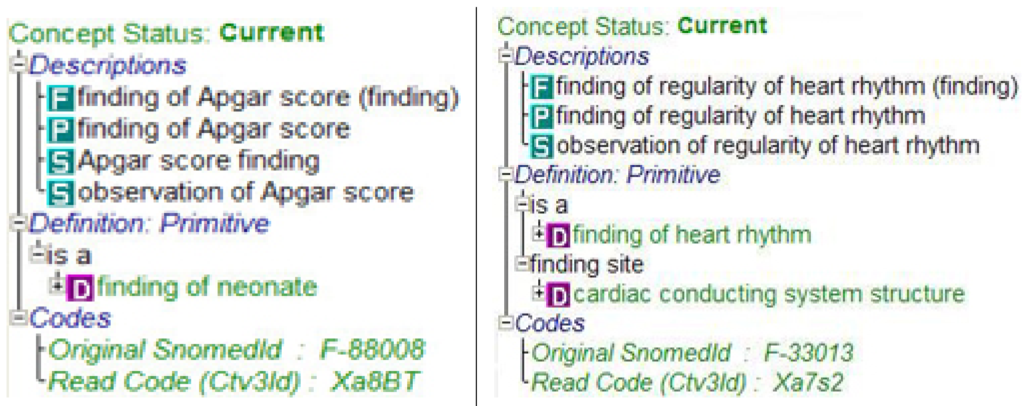
SNOMED frequently reuses the names given to the 19 top-level categories in labeling concepts in the terminology. As discussed earlier in Section 7.1.2.4, it is critical to label concepts in a way that correctly conveys its meaning. Often reusing the category name to which the concept belongs is a useful way to reflect the semantics. However, SNOMED has several instances where concept labels contain the name of other categories causing ambiguity with resolving the semantics of the concept. This usually happens when labeling the Synonyms or Preferred Terms of a SNOMED concept. As stated in Section 7.1.3.1 above, certain issues with improper or ambiguous labeling of concepts only arises when taking into consideration the PT and SYN[4], as in the case of the MoST system. If the PT and SYN are taken into consideration by an automated system then the labeling ambiguities highlight the weak implementation of categorisation guidelines in the terminology as a whole.

Some examples of improper use of category names in concept labels are shown in Figure 7.6, with the help of the CliniClue browser. Part (a) shows that the concept `finding of apgar score (finding) [302083008]` has a synonym 'observation of apgar score'. The use of the word *observation* in providing an alternative label for the apgar score `finding` concept, leads to semantic ambiguities. Similarly, Part (b) shows that the word *observation* has been used once again in the synonym for the concept 'Finding of regularity of heart rhythm (finding) [301114007]'. Repeated interchangeable use of the words *observation* and *finding*, which are also category names with distinct definitions, leads to concerns about the differentiation between an observation and a finding, as stated in the SNOMED documentation. It may be easy for the human eyes to differentiate between the two and resolve the semantics and intended use of the words. However, based on the evidence of use available in the terminology, it becomes difficult to train a computer application to follow any strict rules for drawing inferences from the documentation available. However, this problem would not exist if SNOMED synonyms were not taken into consideration for ascertaining the semantics.

7.1.3.3 Disjoint Axioms

Conflict arising from similar concepts belonging to different hierarchies can be resolved with the help of disjoint axioms. The disjoint axioms specify whether two concepts are compatible or not. By definition, two classes are said to be disjoint if they cannot have any instances in common [70]. However, the absence of disjoint axioms in SNOMED does not help in disambiguating hierarchy-related conflicts. The last available published

[4]Preferred Term and Synonym names of a SNOMED concept



(a) (b)

Figure 7.6: Improper use of category names for labeling concepts in the SNOMED concepts (a) *finding of apgar score* and (b) *finding of regularity of heart rhythm*.

material, dated February 2006, on the issue of including disjoint value sets states that "there is still no reproducible solution due to the lack of classifier availability to manage disjoint sets. The issue continues to remain active in the Concept Model Working Group (CMWG) but holds a low priority status" [11]. There is no recent publication available to suggest any change in the current status. However, email correspondences dated June 2007, confirm that the issue still remains unresolved although interest from several user groups has kept the issue under active discussion but with a low priority status in the CMWG.

The absence of disjoint axioms in SNOMED has proved to be a shortcoming when developing computational procedures to differentiate between the SNOMED concepts returned as matches by MoST. An example mentioned in Chapter 6 (Figure 6.9) is repeated here, to demonstrate the ambiguity caused by the lack of disjointness in SNOMED.

In the tendon babinski archetype, there are two SNOMED concepts `Plantar grasp reflex finding (finding) [366221009]` and `Plantar reflex finding (finding) [366553001]`, which have similar definitions, as can be seen in Figure 7.7. The hypothesis for claiming the two concepts to be equivalent is because of the "open world reasoning" adopted by Description Logic, on which SNOMED is based. Accordingly, in the absence of sibling concepts being explicitly stated as being disjoint, they might be assumed to be equivalent, if the definitions do not provide clear differentiation otherwise. Therefore, the lack of disjoints in SNOMED might lead to errors. On further analysing the parent concepts, as shown in Figure 7.8, there appears to be no logical

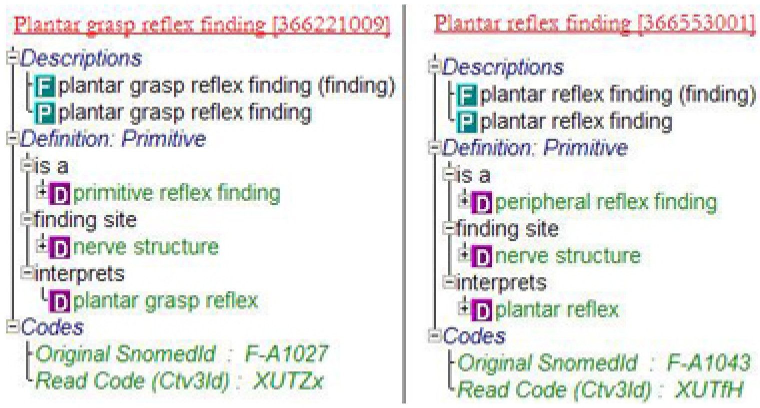


Figure 7.7: Lack of disjointness in SNOMED makes the two concepts with similar definitions equivalent.

difference between the two hierarchies. The parents of the concepts shown in Parts (a) and (b) of the Figure have a common parent `Reflex finding (finding)[106146005]`. In addition, (a) and (b) shows that the parent concepts *interpret* the `primitive reflex` and `peripheral reflex` codes, which are both kinds_of `Reflex (observable entity)[87572000]`.

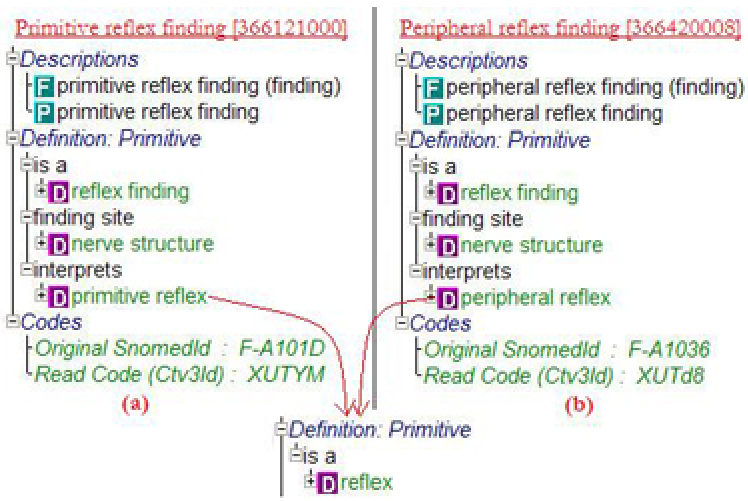


Figure 7.8: Analysis of the specific-to-general (child-to-parent) hierarchy of the *plantar reflex* and *response* concepts in SNOMED.

Based on the similarities of the definitions and hierarchies of the `plantar grasp reflex finding` and `plantar reflex finding`, it would be more appropriate to regard the two as equivalents and include one as the synonym or child (i.e. specialisation) of the other. It is essential to reduce the occurrence of such duplicate hierarchies to not only improve the quality of the terminology structures but also to aid computation processes of automated search applications to retrieve better results.

7.1.3.4 Post-coordinated concepts

The issue of the absence of disjoint axioms in SNOMED is accentuated by the presence of several general/abstract concepts that can be combined with other concepts for post-coordination. These compositional or post-coordinated expressions are more specialised than their individual sub-concepts (i.e. individual pre-coordinated concepts). The ability to post-coordinate concepts leads to the occurrence of multiple concepts belonging to different category hierarchies but representing the same semantics as certain pre-coordinated SNOMED concepts. This causes additional issues of differentiating the correct semantic categories of pre and post coordinated concepts.

Issues arising from the inability to determine the compatibility between concepts in conflicting categories, increased by the ability to form post-coordinated expressions, is explained with the help of Figure 7.9. The snapshot shows that the term '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-coordinated concept `apgar at 1 minute = 0 (finding)`. Interestingly, SNOMED has a concept 'apgar at 1 minute', which is a `finding` as well. Its FSN is 'finding of apgar score at 1 minute', as shown in Figure 7.9. Therefore, two concepts with similar names i.e. 'apgar at 1 minute' belong to two different categories i.e. observable and finding.

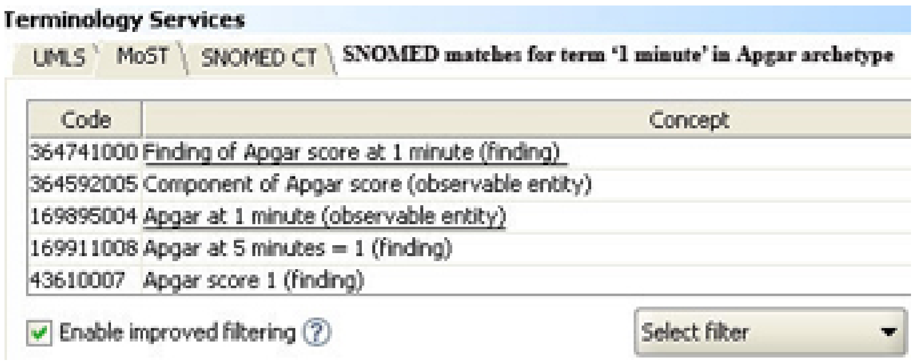


Figure 7.9: SNOMED results returned by MoST for archetype fragment '1 minute' in *Apgar score* archetype. Issues with absence of explicit disjoint axioms in presence of possibility to post-coordinate.

A similar example of ambiguity arising from close proximity of 'observables' with 'findings', is the presence of two very similar concepts, '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. Several similar examples exist in the SNOMED terminology. In terms of aiding computation, it might be helpful to add disjoint value sets in the hierarchy, to enable applications mining through the large corpus of SNOMED data to differentiate between concepts.

7.2 Data Modeling Guidelines

The term *guidelines* generally refers to pronouncements that support or recommend but do not mandate specific approaches or actions. As such, guidelines differ from *standards*, as they are considered mandatory and may be accompanied by an enforcement mechanism [16]. There are three basic questions that a data model should be able to answer, which can also be used to guide what data should be present in a typical model [89]. They are:

1. What information is the doctor likely to need to record about the current situation?

2. What one 'screenful' of information does the doctor need currently, and how can it be presented so as to be assimilated easily?

3. Is there anything which the doctor is proposing or omitting sufficiently worrying to merit a warning?

Although the guidelines stated in the section are intended to be general, they have a specific reference to the SNOMED-style of data modeling. The guidelines are also stated specifically for archetypes but can be modified for use by other data modeling formalisms. Likewise, the SNOMED terminology might be replaced with any other terminology standard such as LOINC, ICD9, or ICD10.

The guidelines are suggestions to help structure archetype fragments in a manner so as to increase the scope of integration of the fragments with SNOMED concepts. The list is not exhaustive and is not mandatory to be followed by the *open*EHR community. The recommendations are largely based on the evaluation study that was carried out as part of the research work.

At first, the section will list a general set of guidelines to model any archetype. The proposed guidelines will then be made specific to modeling the 'observation' and 'procedure' archetypes, as these two categories are most relevant to the research. The guidelines state the features that an archetype belonging to a certain category should ideally model during data representation. The 'clinical finding' category, although commonly used for mapping purposes, was not included in the topic on guidelines. This was because 'observations' and 'procedures' both result in some finding, which makes the characteristics of 'clinical finding' dependent on the characteristics of its causative category(ies) (i.e. observations and procedures).

7.2.1 General Archetype Modeling Guidelines

Listed below are the five key guidelines applicable to any archetype, in general. These guidelines can also be modified to apply to any other data modeling formalism.

- *Provide unambiguous labels and descriptive annotations:* The main starting point for resolving the semantics of any concept is the name/label assigned to it. In addition, any annotation or definition supplied to the concept, helps in gaining a better understanding of its semantics.

- *Group data into common sub-headings:* Common sub-headings should be used to group data into similar semantic types in the model hierarchy. This provides greater clarity with regards to the intended use of the data concept in the model as a whole. It also enables manual or automated processes deployed, to work with greater understanding of the context of the data. The groupings would reflect the common usage of the archetype fragments, rather than the subsumption groupings used by ontology-based models such as SNOMED.

- *Separate specific content from generic data:* Separation of data that is necessary and required for correctly representing a clinical phenomena or statement, from optional information, improves the overall quality of the data model. It also clearly indicates to the user of the model, the information that has to be included during data processing and that which can be excluded, if required.

- *Avoid use of IT-jargon with clinical content:* It is often possible when authoring a data model embedded on the principles of computer science, to use IT-jargon within the data hierarchy. Terms such as 'collection', 'list', or 'set' are data structure types, used in formal programming languages. Model authors should avoid the use of such terms to avoid ambiguity. In case certain clinical terms require the use of these labels, sufficient context and annotation should be provided to clearly specify the semantics of the concept.

- *Author general and specialised archetypes separately to create right level of abstraction:* Generic data applicable to all archetypes of the same clinical or semantic type should ideally be contained in a general model. In addition to the general archetypes containing general archetype fragments there could be more specialised archetypes representing specific clinical statements. These specialised archetypes could then take advantage of the *reuse of nodes* feature of archetypes, and import parts or the entire general model archetype fragments (see Section 2.4.3.1 of Chapter 2). For instance, general pathology or observation archetypes could be authored initially with archetype fragments such as date and time of observation, method used, specimen details, and so on. Subsequently more specialised archetypes, such as histology pap or visual acuity, could be authored reusing the generic archetype fragments in the general archetypes.

7.2.2 Observation Archetype Model Guidelines

Clinical observations include visual assessments, history taking, non-procedural physical examinations, or other information given by the patient or carer, and/or questions asked by the clinician to help with the assessment and subsequent treatment of a particular clinical phenomenon or phenomena. In this respect, listed below are the guidelines for modeling a clinical observation, which have been influenced by reports from [16] and [59].

- *Question-style data:* Clinical observations usually involve a question-answer style of modeling. All data that can lead to results/values pertaining to the immediate cause of concern in the patient care cycle should be included in the specific model as 'data elements'. Questions are mainly asked by the clinician or nurse during history taking or a non-procedural physical examination, to which the patient or the carer responds. Descriptive information about the questions are recommended to be included as annotations in the model.

- *Observation results:* The questions asked during the consultation might lead to responses in the nature of start of symptoms, duration of present state of health, symptoms observed, medication taken, and allergies, amongst others. These responses should be recorded as qualitative 'data values'. In addition, the health professional might have their own clinical or social opinion regarding the patient's health, based on the responses received from the patient, as well as professional experience. These responses need to be clearly stated as observable values in the model, and should not be confused with findings from a procedural intervention.

- *Finding results:* During the process of conducting any physical examination, certain initial findings might be diagnosed by the health professional. These findings are non-conclusive i.e. they might be proved incorrect on performing a pathology or other diagnostic procedure. However, these findings are essential for initiating the treatment process, if appropriate, and should therefore be recorded in the model. However, as the findings are not directly related to the question-answer style specific data, they might be included as additional information in the specific hierarchy as 'Observation Findings'.

- *General information:* In addition to the observations and findings recorded in the specific model, optional but useful general information, should also be included as 'generic data'. This data might include the date and time of recording the information, the method used, the state of the patient during any physical examination, and instruments used, if any. The importance of separating specific content from generic information has been stated earlier in Section 7.2.1.

7.2.3 Procedure Archetype Model Guidelines

The category *clinical procedures/procedures* is not explicitly stated in the *open*EHR Information Model. However, in theory, the `Intervention` and `Action` *open*EHR ENTRY subclasses reflect similar objectives to the SNOMED *procedure* category. Despite the existence of equivalent categories in archetypes, they are not often used to model procedure-related data. For instance, the *Histology pap*, *Autopsy* and *Imaging* archetypes should have ideally been categorised as procedure archetypes, rather than as observation archetypes. The reason being that the results of these model concepts are based on tests performed on body samples or images collected through invasive methods. These archetypes would be rightly categorised as observations had the results been obtained during the course of a medical examination using non-invasive methods.

When evaluating the histology pap archetype, several important observations were made to help distinguish the recording of a procedure from a general observation. The elements that were found common in almost all procedure type archetypes, which can also be considered guidelines for future reference are listed below.

- *Specimen information:* Procedures often result in a specimen, which might be accompanied by a description and annotation(s), to provide more information related to the specimen. A specimen might also have a sample identification number associated with it.

- *Finding results:* In general, a procedure results in Findings, which then requires to be grouped under some category. Depending on the procedure, these findings could be 'Macroscopic' or 'Microscopic', and are usually documented as a part of the test results. These are often quantitative reports rather than interpretive results. The findings could be recorded as 'Reports' with sufficient annotation in a procedure model.

- *Lab reports:* Interpretive lab reports of Findings may be divided as 'Conclusive' and 'Inconclusive'. Conclusive reports are generally reports in standard forms corresponding to the particular test. For instance, in the *Histology pap* archetype, the conclusive reports are often in terms of the degree of epithelial abnormality. The reports from other procedure types may differ depending on the qualitative interpretation of the results, in use locally.

- *Reports generic data:* In addition to all these aspects related to a procedure, there is also a need to capture information that is generic to the data but directly relevant to the test. This information can be modeled into 'generic information' pertinent to a lab report or test and is conceptually distinct from the actual report of the procedure.

- *General information:* A procedure will most likely require some general information i.e. generic data to complete the model. Such data could include the date when the procedure was performed, instruments used, method adopted, etc. as stated earlier in Section 7.2.1. Separating the general information from the specific data will enhance clarity and reuse of different nodes in the same model or across different models.

7.3 Terminology Guidelines

The terminology modeling guidelines will not be based primarily on the clinical content, as the research is based on the computational application of the terminologies. In addition, the reason for not focusing primarily on the terminology content is because the content is dependent on the scope and domain of the terminology. Therefore, it is not possible to prescribe what content can(not) be present in any terminology. Hence, the guidelines will be aimed at improving the ability of terminologies to work with computer applications instead.

Discussion on the guidelines is general and can be applied to any terminology model. However, the inspiration for the work has been obtained from the work done with SNOMED. Therefore, several references are made to the SNOMED model. This section later proposes a list of SNOMED concepts that were found missing when performing the evaluation work. These concepts were considered relevant for inclusion in the SNOMED model.

7.3.1 General Terminology Modeling Guidelines

The guidelines to follow when developing terminologies with a view to enable easy computation are listed below.

- *Logical versus lexical terminologies:* Terminologies that are based on formal logics, such as SNOMED CT and GALEN, have logical, defined hierarchies, which makes it easier for intelligent computing. On the other hand, lexical terminologies or lexicons, such as Gene Ontology and LOINC, are usually flat lists of concepts, which might have only basic subsumption hierarchies, if at all. The lexicons provide little or no concept definitions to help systems compute their semantic relevance with other concepts. Therefore, it is important to develop terminologies that are logically sound and interpretable by computer systems.

- *Simple data format and syntax:* In order to work with terminology concepts, it is necessary that the concepts are available in a format and syntax that is simple to use. For instance, the GALEN terminology is based on formal logic, making it desirable for use by systems requiring terminology models. However, the GRAIL syntax [84][82] in which the GALEN concepts are available, is difficult to use by systems not familiar with the syntax. Alternatively, the data of terminologies, such as SNOMED, are available in a simple 'text' format, which can either be used directly, or be easily imported into any choice of database(s) and/or indexer(s) [3]. The ability to manipulate data effortlessly, makes such terminologies more desirable by systems requiring the use of terminology models.

- *Inferred versus asserted hierarchies:* Efforts should be made to adopt a more rigorous reasoning-based modeling approach for terminologies. Description Logic(DL) [83] reasoners such as FaCT++ [108] must be used to classify terminologies to determine any inconsistencies or errors in the hierarchies. Terminologists must avoid the use of *asserted* hierarchies and increase the use of *inferred* hierarchies. Inferred hierarchies, which have been deduced by a DL reasoner, are less likely to result in logical inconsistencies amongst concepts. For instance, part of the SNOMED hierarchy is inferred although it suffers from the existence of asserted hierarchies (Discussions with K. Spackman, and A.L. Rector). With the adoption of SNOMED in various national health informatics strategies [51][44], it is desirable that further versions of SNOMED data are fully classified before release.

- *Use of annotations:* Annotations provide useful explanations of the use and semantics of a particular concept. These annotations are in addition to the concept definitions that must be provided to formally define a particular concept. The presence of annotations helps the user of a terminology to understand the intent of the terminologist when concept definitions do not provide adequate semantics. At present, none of the widely used and available terminologies have annotated concepts in their models.

- *Inclusion of disjoint axioms:* In terminologies, the use of disjoint axioms helps disambiguate concepts belonging to different hierarchies by asserting whether they do(not) belong to the same 'type', as their most common parent. The absence of disjoint axioms in SNOMED does not help in disambiguating hierarchy-related conflicts. These conflicts usually arise from systems trying to determine the compatibility of two similar concepts belonging to different hierarchies. As stated earlier in Section 7.1.3.3, the issue of implementing disjoint axioms in SNOMED still remains unresolved. However, there is interest from several user groups that has kept the issue under active discussion but with a low priority status in the CMWG.

7.3.2 Concepts inclusion in SNOMED

This section proposes a list of SNOMED concepts that were considered relevant for inclusion in the SNOMED Core Model. Briefly, the Core Model consists of three core data tables i.e. the Concept, Descriptions, and Relationships tables. The three core tables consist of unique `concepts`, a list of `descriptive` names for each concept, and a range of `relationships` with which a concept can bind to other concepts in the terminology (see Section 2.6.4.3;Page 74).

The list of missing SNOMED concepts is mainly limited to the archetypes used during the research study. The blood film archetype was chosen for the MoST experiments stated in Section 4.3.1(Page 103). The four evaluation archetypes were the histology pap, tendon babinski, visual acuity, and body weight archetypes (see Section 5.2.1;Page 123). However, the list of SNOMED concepts to be included is not exhaustive and conclusive. A policy is required to deal with the missing concepts.

- *With respect to the blood film archetype:* Inclusion of a general, non-device 'haemoglobin concentration' concept. At present, the two concepts available in SNOMED are both device specific. These are `Dipstick assessment of` `hemoglobin concentration (procedure) [302781000]`, and `Finding of hemoglobin concentration, dipstick (finding) [365809007]` (for details see Section 7.1.1.3).

- *With respect to histology pap archetype:* The concepts that were considered relevant for inclusion were specific to the recordings of cervical smear, as well as general to recordings of any procedure. These were (i) 'technically unsatisfactory' for technical problems in any kind of test performed, (ii) 'macroscopic appearance' for tissue specimen, as well 'histopathology specimen appearance' for general appearance of a specimen in a histopathology examination, (iii) 'adequate number of squamous cells present' localising it to `cervical smear`, or a general concept 'adequate number of cells present', (iv) 'specimen slide' or 'slide' on which a slide is examined, as well as 'slide unlabeled', (v) 'specimen container' or 'container' to store a specimen, as well as 'container unlabeled', (vi) 'specimen unlabeled' for unlabeled specimens, which might be localised to cervical smear specimens, (vii) concept modifiers for `specimen` viz. 'specimen insufficient', 'specimen damaged', and 'specimen contaminated', (viii) a general concept 'preservation of specimen', (ix) concept modifiers for preservation of specimen viz. 'specimen preservation insufficient', and 'specimen preservation incorrect' (for details see Section 6.1.1.2;Page 149).

- *With respect to tendon babinski archetype:* General plantar reflex concepts to indicate direction of reflexes such as 'plantar response toward sole of foot', and 'plantar response toward dorsum of foot' (for details see Section 6.1.2.2;Page 156).

- *With respect to visual acuity archetype:* The concepts missing in SNOMED were mostly general with respect to the visual acuity archetype. These were (i) the concept qualifier 'aided' for the visual aid procedure, (ii) 'visual acuity - both eyes', as there were concepts available for the left and right eye, and (iii) inclusion of 'optical aid' as a synonym of the concept `visual aid` (for details see Section 6.1.3.2;Page 168).

- *With respect to body weight archetype:* A set of two pre-coordinated and post-coordinated codes were suggested for inclusion to represent the same concept. These were (i) Pre-coordinated: 'weight of body wearing nappy', and 'weight of body wearing clothing' to record the true body weight of an infant, (ii) Post-coordinated: 'weight of nappy', and 'weight of clothing' to be post-coordinated with `Body weight (observable entity) [27113001]` or `Body weight measure (observable entity) [363808001]` , amongst others (see Table 6.11;Page 172) (for details see Section 6.1.4.2;Page 172).

- *General concepts inclusion:* A concept for the absence of cervical tenderness exists in SNOMED viz. `Cervix tenderness absent (situation) [289816007]`. However, there are no concepts for either 'cervix tenderness', or 'cervix tenderness present'. There are several other similar concepts missing in SNOMED, such as the non-existence of codes for 'foveal/macula reflex', and 'foveal/macula reflex present', despite the existence of the concept `absent foveal reflex (finding) [247144003]`.

7.4 Summary

This chapter highlighted the general issues observed in both Archetypes and SNOMED CT. These issues served as the basis for building a list of criteria to improve (i) the content of archetypes, (ii) the content of SNOMED, and (iii) the chances of succeeding in the integration of the content of the two models (i.e. archetypes and SNOMED).

A set of guidelines have been drawn for (i) modeling archetypes, in general, and (ii) modeling specific category archetypes. The two main semantic categories for which guidelines were listed were 'observation' and 'procedure'. These two semantic categories were most relevant to the research. The guidelines were suggestions rather than a mandatory standard to be followed by the *open*EHR community. Although the data model guidelines were specific to Archetypes, they could be modified to adapt to other data models as well.

Finally, a set of guidelines to be adopted for good terminology modeling with respect to making them more computable were also stated. However, the guidelines for data models and terminologies have not been tested, and could be a part of future initiatives in this field. These guidelines could also be included in efforts by the HL7 Terminfo group. A list of concepts that were considered relevant for inclusion in the SNOMED Core Model were also discussed. However, the author does not claim that the list is exhaustive and/or conclusive. This chapter concludes the research work that was carried out to achieve the objective of model integration.

Chapter 8

Future Work and Conclusion

The MoST methodology discussed in the thesis was successful in performing model integration using the *term finding* and *data mapping* procedures. However, the success of the coverage and quality of the MoST results depended on the quality of the archetype and SNOMED models. Evaluating the quality of the two models brought forward issues that required to be resolved by MoST, as well as issues that required to be resolved by the archetype and SNOMED communities.

8.1 Summary of Research Work

The research question that was asked at the start of the thesis was *whether a data model and terminology model can be integrated by a methodology using data mapping/term binding*. The question has been successfully answered during the evaluation exercise detailed in Chapters 5, and 6. It has been shown that it is possible to integrate data in two models to a high rate of coverage and quality, provided sufficient contextual and semantic information can be extracted from each of the models to ascertain equivalence. It is also necessary for models to use data unambiguously through logical hierarchies, annotations, and naming conventions, to increase the chances of automated systems, such as MoST, to compute semantic matches.

The MoST system has provided a practical implementation of the MoST methodology, explained in Chapter 4. The MoST system has established that it is not sufficient to perform lexical lookups alone to perform mapping of data to terminology models. In addition, it is important to employ linguistic and semantic procedures to expand the search base to increase the chances of finding a relevant match. In MoST, the linguistic procedures were performed with the help of the term root extraction feature of GATE

[35], and the English synonyms finder of WordNet [40]. The semantic procedures were performed with the help of a local lexicon of clinical synonyms, abbreviations, and commonly used alternative names. In addition, context information consisting of term definitions/annotations, and parent and root archetype fragments were also employed to expand the search base.

The research objective to enable data modelers to quickly and efficiently integrate their data content to terminology concepts has also been achieved. A paper presented at the Health Computing Conference on the research work [79] stated that a similar mapping program performed manually took approximately 8 man-months to complete[110]. The main reasons were the gaps and short duration of time that the participating clinicians were able to contribute to the manual mapping exercise. Such long periods of study are a major limitation to quick advancement in the field of achieving model integration at the content level. Consequently, this affects interoperability and specification of interoperability standards in the medical community. The MoST system took over the majority of the tedious and lengthy process of searching for appropriate SNOMED concepts by automating the procedures. A typical archetype consisting of 30 fragments took MoST approximately 5 minutes to complete the *term finding* and partial *data mapping* processes. In addition, it took an evaluator cross-checking the results returned by MoST approximately 1 hour to complete the remainder of the *data mapping* process. Even with a large archetype consisting of 45-50 fragments, the time taken to complete the entire MoST procedure was typically 1.5-2 hours. Therefore, with an average execution and evaluation time of 1.5 hours, the MoST system took 99.86% less time, as compared to a typical 8 man-month mapping project (an 8 man-month equates to an average of 1100 hours). The significant reduction in time and effort confirms the claim made in Chapter 1 (see Section 1.2; Page 26).

Automating the search procedure has brought forward several SNOMED concepts that clinical researchers working with SNOMED were unaware of, for example the absence of any non-device specific concept for 'haemoglobin concentration' (see Section 7.1.1.3; Page 181). Therefore, not only is the integration process several times faster but it also holds the potential to exploit the entire scope of clinical concepts covered by SNOMED. In addition, it also helps modelers author their archetypes (or other data models) with a better view of how the data will be represented and coded in SNOMED (or other terminologies). This enables immediate correction of the labeling, annotation, and/or hierarchical structuring of fragments in archetypes, bringing them in close alignment with the SNOMED concepts, and simplifying the model integration process.

Finally, the study also proposes criteria for modeling good quality data in both Archetypes and SNOMED. The criteria are based on the issues that arose during the research study. The aim is to help future development projects in the field of model integration, by sharing the knowledge gained during the design, implementation, testing, evaluation, and analysis of the work. Certain guidelines have also been suggested in the thesis to help the archetype modelers to author better 'observation' and 'procedure' archetypes, in particular. The thesis also illustrates the guidelines for authoring a good archetype model, in general. A guideline-based modeling approach is aimed at structuring the clinical content with the intent of achieving good quality matches to semantically equivalent SNOMED concepts.

8.2 Implications for Archetypes and SNOMED CT

The research presented in the thesis initially began with the assumption that the model integration exercise (i.e. term finding and data mapping) would help in highlighting the shortcomings of the SNOMED terminology model. The initial aim was to critique the SNOMED formalism and suggest a list of inclusions to the SNOMED committee. The second assumption was that the Archetype model would be taken 'as is', requiring no modifications or enhancements. However, these initial assumptions proved incorrect. While SNOMED did require some changes, it was not the only model that contained limitations to achieve the integration objective. The archetype models were also found lacking at various levels with regards to information needed to achieve effective term finding and data mapping.

The evaluation results presented in Section 5.5 (Page 134) demonstrated the impact of the quality of the data in the archetype and SNOMED models on the evaluation results. The three results of research significance are the mapping coverage rate[1], the trust score[2], and the inter-rater reliability[3].

[1]The 'mapping coverage' rate indicates the percentage of fragments in an archetype, which were covered in the SNOMED terminology (See Section 5.3.2; Page 129).

[2]The 'trust score' is used to denote the trust or confidence, experts placed on the SNOMED codes returned by MoST for an archetype (See Section 5.3.3; Page 131).

[3]The 'inter-rater reliability' indicates the extent of agreement between the raters (See Section 5.3.4; page 132).

The tendon babinski archetype achieved the highest coverage of 90.9%. This indicated that 90.9% of the archetype fragments found a highly relevant SNOMED match, at the end of the MoST process. The tendon archetype also achieved the highest trust score of 0.9. This means that on an average at least two evaluators assigned a score of 9 (on a scale of 0 to 10) to the SNOMED codes they believed had high semantic equivalence to the archetype fragments. Similarly, the tendon babinski archetype also achieved the highest inter-rater reliability of 86%, indicating that there was a high level consensus amongst the evaluators. The high results were primarily achieved because of the unambiguous data present in both the archetype and SNOMED models for the concepts related to tendon and babinski reflexes. On the other hand, the high degree of ambiguity in the histology pap archetype and corresponding SNOMED matches led to the lowest rate of coverage of 48.6%, trust score of 0.45, and inter-rater reliability of 55%. The contrast between the results for these two archetypes, highlights the importance of the quality of data in succeeding in the model integration objective.

The quantitative and qualitative analysis of the evaluation results discussed in Chapters 5 and 6, respectively, shifted the focus of the study from being unidirectional (i.e. critiquing SNOMED only) to being bidirectional (i.e. critiquing both SNOMED and Archetypes). Chapter 7 concluded the study with a general critical overview of Archetypes and SNOMED models with respect to their clinical content. It also made suggestions for improving the quality of the data along with a guideline on modeling 'observation' and 'procedure' type archetypes. It was observed that the straight merger of SNOMED RT and READ code concepts to form SNOMED CT had led to shortcomings in the present version of the SNOMED terminology. Often, it was difficult for the MoST system to clearly disambiguate similarly named and defined concepts only on the basis of their existence in different category hierarchies. Another observation made regarding the SNOMED content was that it had a better coverage of the pathology and other medical domains, as compared to coverage of the nursing and social care domains of health care. General issues with the SNOMED model have been detailed in Section 7.1.3 (Page 191).

At the end of the work it can be stated that it is essential to set out clear rules for naming and utilising clinical content in the model that provides the source of the data (such as archetype fragments in archetypes). If the data quality at source is more closely aligned to the target terminology, it will become much simpler to achieve better mapping results. A higher mapping coverage indicates better integration of the content of the data model with the terminology model.

8.3 Future Work

Despite the extensive work carried out in the research and receiving good feedback from the data standards and interoperability working groups, there are several areas of work that can be improved or added to enhance the quality of the results.

8.3.1 *Specific-to-general* approach

At present, the context information works on a *general-to-specific* basis. This means that the results are refined as the process moves down the data hierarchy in the archetype model. SNOMED matches for the child fragments are determined on the basis of the enclosing archetype fragment. Therefore, although the matches are further specialised to take into account the specific child fragment, all efforts are made to keep the results in proper context. An alternative approach can also be attempted, which could be based on the *specific-to-general* approach. This would include determination of a more general SNOMED match for the enclosing fragment based on the first most common SNOMED hierarchy to which the child fragments belong. For instance, in the visual acuity archetype, the child fragments of the enclosing fragment 'visual acuity' could help in determining the most appropriate SNOMED match for this enclosing fragment.

As shown in Figure 8.1, the child fragment '6/6' was found to be semantically equivalent to the SNOMED concept `Distance vision 6/6 (finding) [260272008]`. Also, the child fragment '6/18' found a match with the SNOMED concept `Distance vision 6/18 (finding) [260275005]`. The other visual acuity findings found similar SNOMED matches from the subtree of `Distance vision finding (finding) [260269001]`. Therefore, using the *specific-to-general* approach it can be deduced that the enclosing archetype fragment 'visual acuity' should be mapped to the SNOMED concept `Distance vision finding (finding) [260269001]`. Likewise, the fragment 'blind' in the visual acuity archetype could be replaced with 'no light perception' to conform to the SNOMED subtree of `Distance vision finding`.

8.3.2 Inclusion of ancestor archetype fragments

At present, the *term finding* procedure adopted by the MoST system only includes the immediate enclosing fragment of a query archetype fragment. At times, it also includes the root fragment of the archetype model to help guide the selection and elimination processes. The present rule is based on the assumption that inclusion of ancestor archetype fragments will make the results too general and, often, not of any contextual

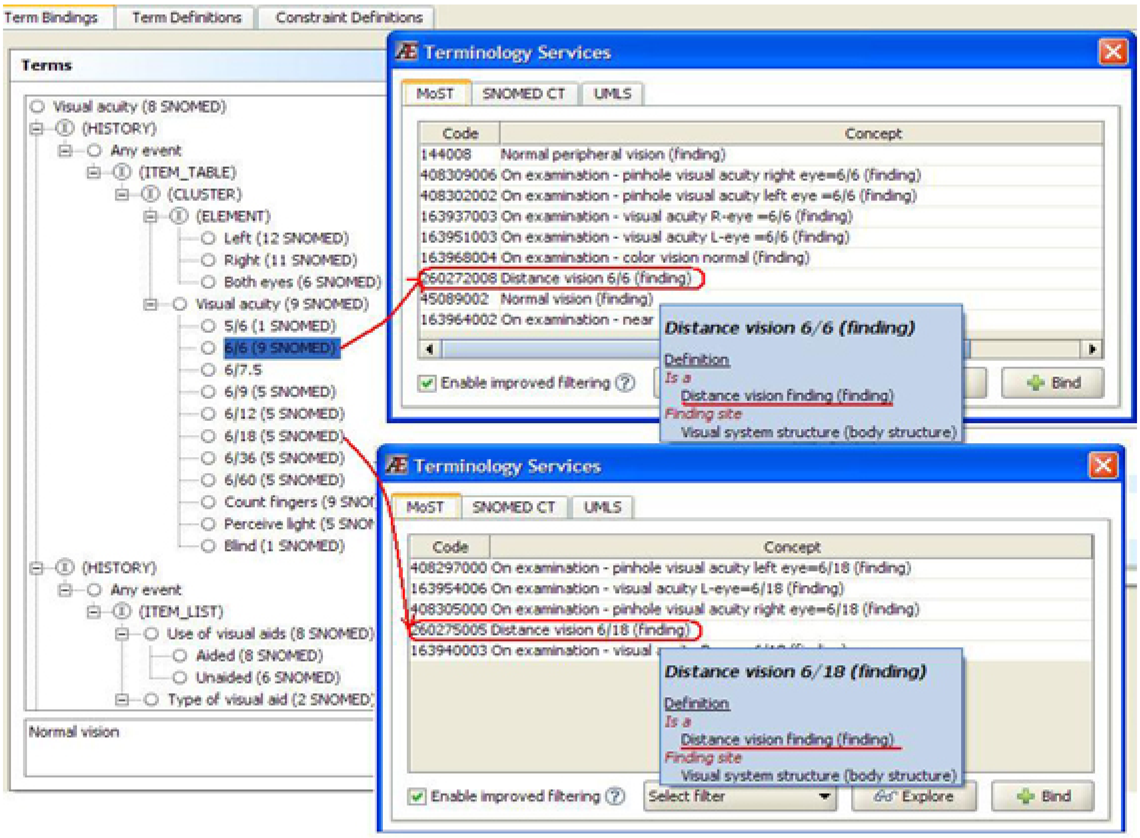


Figure 8.1: Applying the *specific-to-general* approach on the visual acuity archetype.

relevance to the query fragment. This is because archetype models are containment hierarchies (see Section 2.4.3.1; Page 50) that include only those fragments that are required to represent the recording of a particular clinical statement. As such they are not logical, subsumption hierarchies, as is the case with SNOMED CT. However, it is possible to include the ancestor fragments as well if in future any special requirement arises.

In the event of inclusion of ancestor fragments to determine the context of a query archetype fragment, it might be considered necessary to use complete archetype paths. The archetype paths consist of the entire path to a specific fragment and are essential to reference especially in the case of re-used blocks of archetypes within a model. Use of archetype paths rather than 'hard-coded' references to the ancestor fragments, will help the MoST system to cope with any changes made to the archetype at a later stage.

For instance, a prototype of an archetype block is shown in Figure 8.2. It is possible in the archetype modeling language i.e in ADL for the new enclosing fragment at0002

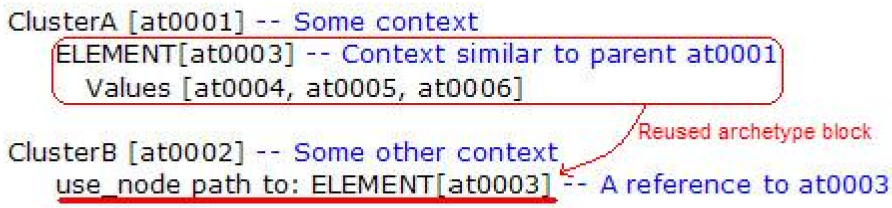


Figure 8.2: Prototype of the internal reuse of an archetype block in an archetype model.

to reference the current ancestor fragment at0001, which is the original enclosing fragment of at0003. However, at present, MoST will not take into consideration the ancestor fragment at0001 when performing the term finding and/or data mapping procedures. The only scenario when an exception will be made will be if at0001 is the root fragment.

The reuse of archetype blocks in different contexts, presented as comments in Figure 8.2 (comments are preceeded by −−), raises the question whether it will be appropriate to reference original enclosing fragments (at0001) of reused blocks (at0003) even though the context of at0001 may differ from at0002. In such an event, inclusion of ancestor fragments, with no direct relevance to the query fragment, might lead to impaired or irrelevant results. It will be best to consider the option of including ancestor fragments depending on the nature of individual archetype models.

8.3.3 Inclusion of SNOMED grandchildren concepts

In the event that a child as well as a grandchild of a SNOMED concept are present in the result set then the parent concept might be assumed to be a good candidate match and included in the final results. At present, only the existence of two or more child concepts of a particular SNOMED concept is taken into consideration to determine appropriateness. Therefore, the likelihood of including a SNOMED grandchild might occur when 'filtering' non-equivalent SNOMED codes at the end of the term finding process in the MoST system.

8.3.4 Filtering based on SNOMED Concept Model

The filtering mechanism of the MoST system can be made more strict and restrictive by adhering to the rules of the SNOMED Concept Model specified in the SNOMED documentation [14]. The rules specify the permissible use of concepts belonging to a specific semantic category with concepts from another semantic category. For instance, an `observable entity` concept cannot be mapped to a `body structure` concept according to the SNOMED documentation. Although clinically sound, this would require the categorisation of archetype fragments into specific SNOMED categories to successfully apply the rules. Although the intention is to ensure that the archetype model is compatible with SNOMED, there exists the risk of being too rigid about applying the SNOMED rules to Archetypes. The main reason is that traditionally the archetype clinical content (i.e. the fragment) is not designed to conform to SNOMED standards. This may lead to too few and incompatible results being returned by restricting the categories from which codes might be returned. If the data mapping is performed by someone other than the original author of the archetype, the chances of categorising an archetype fragment incorrectly will become higher leading to irrelevant matches, making the system brittle. However, if the task is performed by the author, it will help in ensuring that the best codes are returned as matches reducing the "noise" (i.e the rate of false positives) in the system.

8.3.5 Post-coordinated results

The initial intention was to suggest post-coordinated matches, i.e. a composition of two or more SNOMED concepts, to represent a semantically equivalent match for an archetype fragment that had not found a pre-coordinated match in SNOMED. The evaluator could then select the most relevant post-coordinated match and either map it to the archetype fragment 'as is' or add it to a change request form to be submitted to the SNOMED committee for inclusion as a pre-coordinated concept. However, the post-coordination module could not be implemented due to lack of compatibility in the style and scope of modeling data in the two models (i.e. archetypes and SNOMED). One of the other reasons that this attempt failed is that archetype models predominantly use the post-coordinated style of modeling data by default. This means that the archetype typically consists of a composite data hierarchy, as seen in the barthel index archetype in Figure 8.3.

It was found difficult to find appropriate post-coordinated SNOMED matches for archetype fragments that had already been used in a post-coordinated form in the archetype. It can be seen in the Figure 8.3 that the child fragments of the fragment 'transfer', presented independent values such as 'unable', 'major help', etc., which are general and can be reused in a different context elsewhere in the archetype. It would be more helpful to use fragments such as 'unable to sit for transfer, lifting device required', 'can sit but require people to help with transfer' etc. for post-coordination. However, it is highly unlikely that such fragments will be able to find a semantically relevant

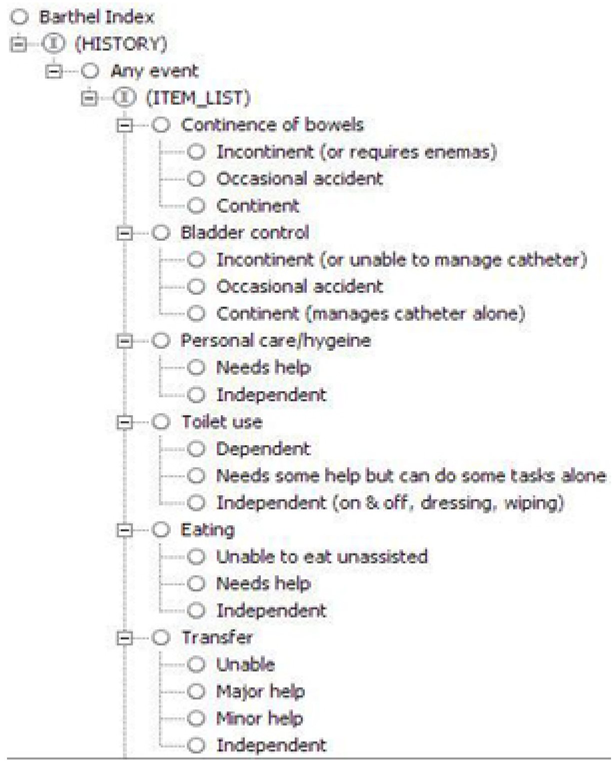


Figure 8.3: Post-coordinated style of archetype fragment modeling as seen in the barthel index archetype.

post-coordinated match either.

The three key contributions that post-coordinated entries can help provide for refining archetype models is to present the author with (a) suggestions that the archetype model cannot use the present fragment as it is not present in SNOMED, (b) suggest a mechanism for submitting the unmatched fragment to SNOMED for inclusion, and (c) generate a post-coordinated match for the fragment, if possible. In addition to the rules specified by SNOMED on post-coordination [14], certain internal rules might also be added to cope with the loose categorisation of data present in Archetypes. Some of the range of Archetype and SNOMED categories that can be treated as extensions to the SNOMED rules are:

Archetype Observation + SNOMED Procedure ⇒ LEGAL
Archetype Instruction + SNOMED Observable ⇒ LEGAL
SNOMED Observation + SNOMED Procedure ⇒ ILLEGAL

For example, the above range indicates that an Observation archetype 'autopsy' could legally use a SNOMED Procedure code `autopsy examination` in its model. This, of course, would be illegal in SNOMED (according to rule 3 above). Therefore, extensions to the rules might be required to allow different categories in SNOMED to merge with archetypes. However, this would not mean that when generating post-coordinated expressions, a composition of SNOMED observation and procedure codes will be used in the same block.

More research and testing is required in the field of post-coordination to ensure that only logically sound composite matches are returned by systems such as MoST. Else it might result in EHRs being populated with incorrect data codes, thereby defeating the purpose of making data safe and interoperable to reduce medical errors and improve health care.

8.3.6 Extending testing to HL7 CDA

At present the research hypothesis has been tested on Archetypes and SNOMED. However, the MoST methodology is scalable and can be applied to other data models and terminologies. Therefore, future work might involve extending the work to test the methodology using HL7 CDA documents and SNOMED or some other terminology such as LOINC. Formal rules already exist in HL7 for integrating its content with SNOMED [64]. In future, these rules can be applied in practice and tested using the MoST system. The results might highlight the degree of compatibility between the SNOMED and HL7 documents, which can then be comparatively analysed against archetypes. However, the comparative study will only succeed under controlled situations by providing similar test platforms.

8.4 Conclusion

In conclusion, the MoST methodology demonstrates the possibilities and difficulties to integrate the data models with terminology models. The difficulties in integration arise primarily due to the semantic gaps in the content of the two models. The thesis highlights the difficulties, and suggests solutions and guidelines to improve the content, to enable seamless integration of the two models.

The thesis stresses that in order to achieve the overall objective of semantic interoperability, it is imperative that both data and terminology models are developed with the aim of being able to integrate their clinical content. It is important that both modeling communities are not only aware of each others existence but also work closely with each other to ensure that conformance is built into the systems from conception stage. These conformance or compatibility rules should be extended to all other stages of the modeling process i.e. at design time, data integration time, as well as at run-time. It is only then that true interoperability will be achieved, making it possible to build safer health care systems. Reliable and high quality data in these systems will improve the functioning of all health care units heavily dependent on data, reducing medical errors and ultimately providing safer and better patient care.

Bibliography

[1] Connecting for Health - A Public-Private Collaborative - FACTS AND STATS. http://www.markle.org/downloadable_assets/facts_and_stats_060503.pdf. As seen in July 2006.

[2] International health terminology standards development organisation. Accessed Oct 2007.

[3] Lucene java. http://lucene.apache.org/java. Accessed June 2007.

[4] The magnificent ROC (receiver operating characteristic curve). http://www.anaesthetist.com/mnm/stats/roc/index.htm. Accessed in May 2007.

[5] NHS Connecting for Health website. http://www.connectingforhealth.nhs.uk/technical/standards/snomed/. Accessed January 2007.

[6] Published *open*EHR archetypes. http://svn.openehr.org/knowledge/archetypes/dev/html/index_en.html. Accessed June 2007.

[7] SNOMED website. http://www.snomed.org/.

[8] The european environment & health action plan 2004-2010. Technical Report Vol I, Commission of the European Communities, Brussels, June 2004.

[9] Cliniclue browser, 2006. Accessed August 2007.

[10] Introducing *open*EHR. The *open*EHR website, Jan 2006. Rev 1.1.

[11] Snomed international content steering group. Approved Minutes of Meeting, February 2006.

[12] User's guide introduction, January 2006.

[13] Ocean terminology toolset: Oceaninformatics product description, July 2007. Page 16-18.

[14] SNOMED clinical terms technical reference guide. Technical report, College of American Pathologists, 2007. January 2007 Release.

[15] SNOMED CT - the language of the NHS care records service - a guide for NHS staff in england. NHS Connecting for Health website, 2007. Accessed October 2007.

[16] A. P. Association. Criteria for evaluating treatment guidelines. *American Psychologist*, 57:1052–59, 2002.

[17] S. Bakken, J. Cimino, and R. H. *et al.* Evaluation of the clinical LOINC (logical observation identifiers, names, and codes) semantic structure as a terminology model for standardized assessment measures. *JAMIA*, 7(6):529–38, Nov-Dec 2000.

[18] T. Beale. The GEHR software architecture for a reliable EHR. *Toward an Electronic Health Record Europe '99*, pages 328–339, November 1999.

[19] T. Beale. Archetypes constraint-based domain models for future-proof information systems. *In: Eleventh OOPSLA Workshop on Behavioral Semantics: Serving the Customer*, pages 16–32, November 2002. Seattle, Washington, USA, November 4, 2002.

[20] T. Beale, D. S. Heard, D. D. Kalra, and D. Lloyd. The *open*EHR EHR information model. (Revision 5.0), April 2006.

[21] T. Beale, D. S. Heard, D. D. Kalra, and D. Lloyd. The *open*EHR EHR information model. (Revision 5.1.0), April 2007.

[22] T. Beale, D. S. Heard, D. D. Kalra, D. Lloyd, and D. P. Schloeffel. Introducing *open*EHR. Technical Report 1.1, The *open*EHR Foundation, 2005. Issued on 2 Jan 2006.

[23] T. Beale and S. Heard. The *open*ehr Archetype System. Technical Report Revision 0.3.1, The *open*EHR Foundation, 2003.

[24] T. Beale and S. Heard. Archetype definitions and principles. (Revision 0.6), March 2005.

[25] T. Beale and S. Heard. Archetype definition language (ADL). *http://svn.openehr.org/specification/TRUNK/publishing/architecture/am/adl.pdf*, (Revision 1.3.1), 2006. The *open*EHR Foundation.

[26] T. Beale and S. Heard. Archetype definition language version 1 ADL. *http://svn.openehr.org/specification/BRANCHES/Release-1.1-candidate/publishing/architecture/am/adl.pdf*, (Revision 1.4.0), 2006. The *open*EHR Foundation.

[27] E. S. Berner, R. K. Kasiraman, and F. Yu. Data quality in the outpatient setting: Impact on clinical decision support systems. *AMIA Annual Symposium Proceedings*, pages 41–45, 2005.

[28] O. Bodenreider and A. Burgun. Linking the gene ontology to other biological ontologies. *ISMB2005 SIG meeting on Bio-ontologies*, pages 17–18, 2005.

[29] O. Bodenreider, S. J. Nelson, W. T. Hole, and H. F. Chang. Beyond synonymy: Exploiting the UMLS semantics in mapping vocabularies. *In: Chute, Christopher G., editor. A Paradigm Shift in Health Care Information Systems: Clinical Infrastuctures for the 21st Century. AMIA Proceedings*, pages 815–19, 1998. Orlando, FL. Philadelphia: Hanley & Belfus, Inc.

[30] E. Cheetham. *SNOMED CT Post-coordination Rules.* National Programme for Information Technology, ver 0.985 edition, May 2005. Draft Guidance Document.

[31] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, November 1998.

[32] CLEF. CLEF - integrating information for the clinical e-scientist. *http://www.clinical-escience.org/*, 2005. A U.K. Medical Research Council funded project.

[33] E. Coiera. *Guide to Medical Informatics, the Internet and Telemedicine.* Chapman and Hall Medical, 1997.

[34] College of American Pathologists. *SNOMED Clinical Terms - User Guide*, January Release 2006.

[35] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu. *Developing Language Processing Components with GATE Version 3 (a User Guide).* University of Sheffield, January 2005.

[36] DICOM. DICOM standards links. *http://medical.nema.org/.* Last accessed Jan 2007.

[37] R. Dolin and L. A. *et al.* An update on HL7's XML-based document representation standards. In *Proc AMIA Symp*, November 2000.

[38] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J. E. Mattison. The HL7 Clinical Document Architecture. *Journal of the American Medical Informatics Association (JAMIA)*, 8(6):552–69, Nov-Dec 2001.

[39] R. H. Dolin, L. Alschuler, S. Boyer, C. Beebe, F. M. Behlen, and A. S. Paul V. Biron. HL7 Clinical Document Architecture, release 2. *JAMIA*, 13(1):30–39, 2005.

[40] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, 1998. M.A.

[41] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychol. Bulletin*, 76:378–82, 1971.

[42] J. L. Fleiss. The measurement of interrater agreement. *Statistical Methods for Rates and Proportions*, Second Edition:212–304, 1981. John Wiley & Sons Inc., New York.

[43] N. C. for Health. SNOMED CT - the language of the NHS care records service - a guide for NHS staff in england. http://www.connectingforhealth.nhs.uk/systemsandservices/data/standards/. Accessed in April 2007.

[44] N. C. for Health. Better information better health. Technical report, NHS National Programme for IT Annual Report 2004-2005, 2005.

[45] C. Friedman, P. Alderson, and J. A. *et al.* A general natural-language text processor for clinical radiology. *JAMIA*, 1(2):161–74, 1994.

[46] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *JAMIA*, pages 392–402, 2004.

[47] T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. *Stanford Knowledge Systems Laboratory*, 1993.

[48] K. Gwet. Computing inter-rater reliability with the SAS system. *Statistical Methods for Inter-Rater Reliability Assessment*, 3, 2002.

[49] B. Hazlehurst, H. Frost, D. Sittig, and V. Stevens. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *JAMIA*, 12(5):517–29, Sep/Oct 2005.

[50] S. Heard. The good electronic health record: trial of a standard healthcare record in australian general practice. *Br J Healthcare Comput Info Manage*, 17(1):24–26, 2000.

[51] HHS. 2005 hhs e-gov annual report. Technical report, U.S. Department of Health and Human Services, Viewed in July 2006 2005.

[52] J. Hjalmarsson and M. Forss. Development of an archetype editor: A tool for modelling structure in electronic health records. Masters Thesis: http://www.diva-portal.org/liu/undergraduate/abstract.xsql?dbid=6205, LiU, Sweden. Accessed February 2007.

[53] HL7. Health informatics standards - HL7. *http://www.hl7.org/*, 2007. Accessed Feb 2007.

[54] B. L. Humphreys and D. A. B. Lindberg. The UMLS project: Making the conceptual connection between users and the information they need. *Medical Library Association*, 81(2):170, 1993.

[55] O. Informatics. Ocean Archetype Editor. *http://www.oceaninformatics.biz/archetype_editor/help/ht_start.htm*.

[56] S. B. Johnson. A semantic lexicon for medical language processing. *JAMIA*, 6(3):205–18, May-Jun 1999.

[57] D. Kalra. Synapses ODP Information Viewpoint. *EU Telematics Application Programme.https://www.cs.tcd.ie/synapses/public/ deliverables/part3.pdf*, July 1998. The Synapses Project: Final Deliverable.

[58] D. Kalra. Clinical foundations and information architecture for the implementation of a federated health record service. *PhD Thesis. Univ. London. http://www.ehr.chime.ucl.ac.uk/docs/Kalra*, July 2005.

[59] J. L. Lane and R. P. Gottlieb. Structured clinical observations: A method to teach clinical skills with limited time and financial resources. *American Academy of Pediatrics*, 105:973–77, 2000.

[60] L. M. Lau, K. Johnson, K. Monson, S. H. Lam, and S. M. Huff. A method for the automated mapping of laboratory results to LOINC. *Proc AMIA Symp*, pages 472–76, 2000.

[61] T.-W. Loong. Understanding sensitivity and specificity with the right side of the brain. *British Medical Journal(BMJ)*, 327:716–19, 2003.

[62] K. LT, C. JM, and D. MS. To err is human: building a safer health system. *National Academy Press*, 2000. Washington, D.C.

[63] M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. *Workshop on Learning from Imbalanced Data Sets II, ICML*, 2003. Washington DC.

[64] E. D. Markwell. Using SNOMED CT in HL&7 version 3: Implementation guide. Technical Report Release 1.2, HL7 Terminfo Special Interest Group, 2006.

[65] D. Maynard. Benchmarking ontology-based annotation tools for semantic web. *UK e-Science Programme All Hands Meeting (AHM2005) Workshop on "Text Mining, e-Research and Grid-enabled Language Technology"*, September 2005.

[66] A. T. McCray and W. T. Hole. The scope and structure of the first version of the umls semantic net. *Proceedings of the Fourtheenth Symposium on Computer Applications in Primary Care*, pages 126–130, 1990. Washington DC, IEEE Computer Society.

[67] C. McDonald, S. Huff, and J. S. *et al.* LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49(4):624–33, 2003.

[68] NHS. NHS connecting for health: SNOMED clinical terms. http://www.connectingforhealth.nhs.uk/terminology/snomed, Accessed March 2007.

[69] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880*, March 2001.

[70] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. Technical report, Stanford University Publications, 2001.

[71] U. D. of Emergency Medicine. EMT-P user manual ver 2.1. http://www.med.unc.edu/emergmed/EMTP/usermanual.pdf. Seen in July 2006.

[72] T. B. Pedersen and C. S. Jensen. Research issues in clinical data warehousing. In I. C. Society, editor, *Proceedings of the Tenth International Conference on Scientific and Statistical Database Management*, pages 43–52, July 1998. Capri, Italy.

[73] C. Price and K. Spackman. SNOMED clinical terms. *BJHC & IM-British Journal of Healthcare Computing & Information Management*, 17(3):27–31, 2000.

[74] M. C. Proceedings, editor. *Health Care Terminology for the Electronic Era*, volume 81. Mayo Clinic College of Medicine, June 2006. 729-31.

[75] R. Qamar. Determining semantic interoperability between archetypes and clinical ontologies. Submitted to KCap Proceedings 2005.

[76] R. Qamar, J. Kola, and A. Rector. Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. *AMIA 2007 Annual Symposium*, November 2007. Chicago, U.S.A. Awaiting Publication.

[77] R. Qamar and A. Rector. Most: A system to semantically map clinical model data to snomed-ct. *Semantic Mining Conference on SNOMED-CT (SMCS)*, pages 38–43, October 2006. Copenhagen, Denmark.

[78] R. Qamar and A. Rector. Semantic issues in integrating data from different models to achieve data interoperability. In K. *et al.* (Eds), editor, *Proceedings of MedInfo 2007*, pages 674–678. IOS publishing, August 2007. Brisbane, Australia.

[79] R. Qamar and A. Rector. Semantic mapping of clinical model data to biomedical terminologies to facilitate data interoperability. In *Healthcare Computing 2007 Conference*, pages 116–26, March 2007. Harrogate, U.K.

[80] B. Quinn. Integrated care record systems. The Future of IT within the NHS http://msdforphysicians.co.uk. Accessed Oct 2003.

[81] A. Rector. Thesauri and formal classifications: Terminologies for people and machines. *Methods of Information in Medicine*, 37.4-5:501–09, 1998.

[82] A. Rector. Analysis of propagation along transitive roles: Formalisation of the galen experience with medical ontologies. *Proc. of DL 2002. CEUR - http://ceur-ws.org*, 2002.

[83] A. Rector. *The Description Logic Handbook: Medical Informatics.* Cambridge University Press, 2003.

[84] A. Rector, A. Gangemi, E. Galeazzi, A. Glowinski, and A. Rossi-Mori. The GALEN CORE model schemata for anatomy: Towards a re-usable application-independent model of medical concepts. *Proceedings of Medical Informatics Europe*, pages 186–89, 1994. Lisbon.

[85] A. Rector, W. Nowlan, S. Kay, C. Goble, and T. Howkins. A framework for modelling the electronic medical record. *Methods of Information in Medicine*, 32(2):109–119, 1993.

[86] A. Rector, R. Qamar, and T. Marley. Binding ontologies & coding systems to electronic health records and messages. *Formal Biomedical Knowledge Representation (KR-MED 2006)*, pages 11–19, 2006. In: Bodenreider O, editor.

[87] A. Rector, J. Rogers, and P. Pole. The GALEN high level ontology. *Medical Informatics Europe (MIE)*, pages 174–78, 1996. Copenhagen.

[88] A. Rector, W. Solomon, W. Nowlan, and T. Rush. A terminology server for medical language and medical information systems. *Proceedings of IMIA WG6*, May 1994. Geneva.

[89] A. L. Rector. Helping with a humanly impossible task: Integrating knowledge based systems into clinical care. In J. Hannu and L. S. (eds.), editors, *Proceedings of the Second Scandinavian Conference on Artificial Intelligence SCAI-89*, pages 453–66, 1989. Amsterdam: IOS Press.

[90] A. L. Rector. Clinical terminology: Why is it so hard? *Methods of Information in Medicine*, 38(4):239–252, 1999.

[91] A. L. Rector. The interface between information, terminology, and inference models. In *Tenth World Conference on Medical and Health Informatics: MEDINFO*. London, England, 2001.

[92] A. L. Rector. Whats in a code? towards a formal account of the relation of ontologies and coding systems. *Medinfo Conference*, 2007. Brisbane, Australia. Awaiting Publication.

[93] A. L. Rector, W. Nolan, and S.Kay. Foundations for an electronic medical record. *Methods of Information in Medicine*, 30:179–86, 1991.

[94] D. Sherertz, M. Tuttle, M. Blois, and M. Erlbaum. Intervocabulary mapping within the UMLS: the role of lexical mapping. *Proc Annual Symposium on Computer Applications of Medical Care*, pages 201–06, 1988.

[95] J. Sim and C. Wright. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85:257–68, 2005.

[96] D. F. Sittig. Grand challenges in medical informatics. *Jorunal of the American Medical Informatics Association*, 1:412–13, 1994.

[97] L. Smith and H. Preston. Information management and technology strategy in healthcare: local timescales and national requirements. *Information Research*, 5(3), April 2000.

[98] SNOMED. Guidance on implementing the SNOMED clinical terms context model. Technical Report Revision 0.3, May 2004.

[99] K. Spackman, K. Campbell, , and R. Cote. Snomed rt: A reference terminology for health care. *J. of the American Medical Informatics Association*, pages 640–44, 1997. Fall Symposium Supplement.

[100] K. Spackman, R. Dionne, E. Mays, and J. Weis. Role grouping as an extension to the description logic of ontylog, motivated by concept modeling in snomed. *Proc AMIA Symp*, pages 712–6, 2002.

[101] S. Srinivasan, T. Rindflesch, and W. H. *et al.* Finding UMLS metathesaurus concepts in MEDLINE. In *Proc. AMIA Annual Symp.*, page 727731, 2002.

[102] J. Sun and Y. Sun. A system for automated lexical mapping. *JAMIA*, pages 334–43, 2005.

[103] E. Sundvall, R. Qamar, and M. N. *et al.* Integration of tools for binding archetypes to SNOMED CT. *Supplement in Biomed Central (BMC) Journal 2007 (Submitted)*, pages 64–68, Oct 2006.

[104] E. Sundvall, R. Qamar, and M. N. *et al.* Integration of tools for binding archetypes to SNOMED CT. *SNOMED-CT and Semantic Mining Conference*, pages 64–68, Oct 2006.

[105] I. TC215. International organisation for standardisation - home page. *http://www.iso.org/iso/*, 2004.

[106] C. TC251. Cen/tc 251 - european standardization of health informatics. *http://www.centc251.org/*, 2004. Website.

[107] S. Travers D.A., Haas. Using nurses natural language entries to build a concept-oriented terminology for patients chief complaints in the emergency department. *Journal of Biomedical Informatics.*, 36:260–270, 2003.

[108] D. Tsarkov and I. Horrocks. FaCT++ description logic reasoner: System description. In *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer, 2006.

[109] UMLS. Umls metathesaurus. *http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html*, 2006. Accessed October 2007.

[110] G. Wade and S. T. Rosenbloom. Experiences mapping a legacy interface terminology to snomed ct. *Semantic Mining Conference on SNOMED CT (SMCS)*, October 2006. Copenhagen, Denmark.

[111] B. Workgroup. BioPAX biological pathways exchange language level 1 ver 1.4 documentation. Technical report, BioPAX Workgroup, April 27 2005.

Appendix A

EXTRA TABLES AND FIGURES

A.1 HL7 Reference Information Model

Health Level Seven is an ANSI[1]-accredited Standards Developing Organization (SDO) operating in the healthcare arena covering clinical and administrative data. Its mission is to create standards for the exchange, management and integration of electronic healthcare information [53]. All the HL specifications are unified by shared reference models of the health care and technical domains. In HL7, the Reference Information Model (RIM) is a shared model between all the domains and as such is the model from which all domains create their messages [53]. Besides creating health care messaging standards such as the version 3(V3) Messages, it is also involved in developing standards for the representation of clinical documents. The document standards form part of the HL7 Clinical Document Architecture (CDA) and include, for instance, standards for discharge summaries and progress notes.

The RIM is a large UML-based pictorial representation of the clinical data, and identifies the life cycle of events that a message or groups of related messages will carry [53]. Figure A.1 presents the UML model of Version 1.18 of the RIM.

A.2 HL7 Clinical Document Architecture

Chapter 2 provides a detailed explanation of the CDA. Figure A.2 provides a sample view of two typical HL7 clinical documents based on the CDA.

[1] American National Standards Institute

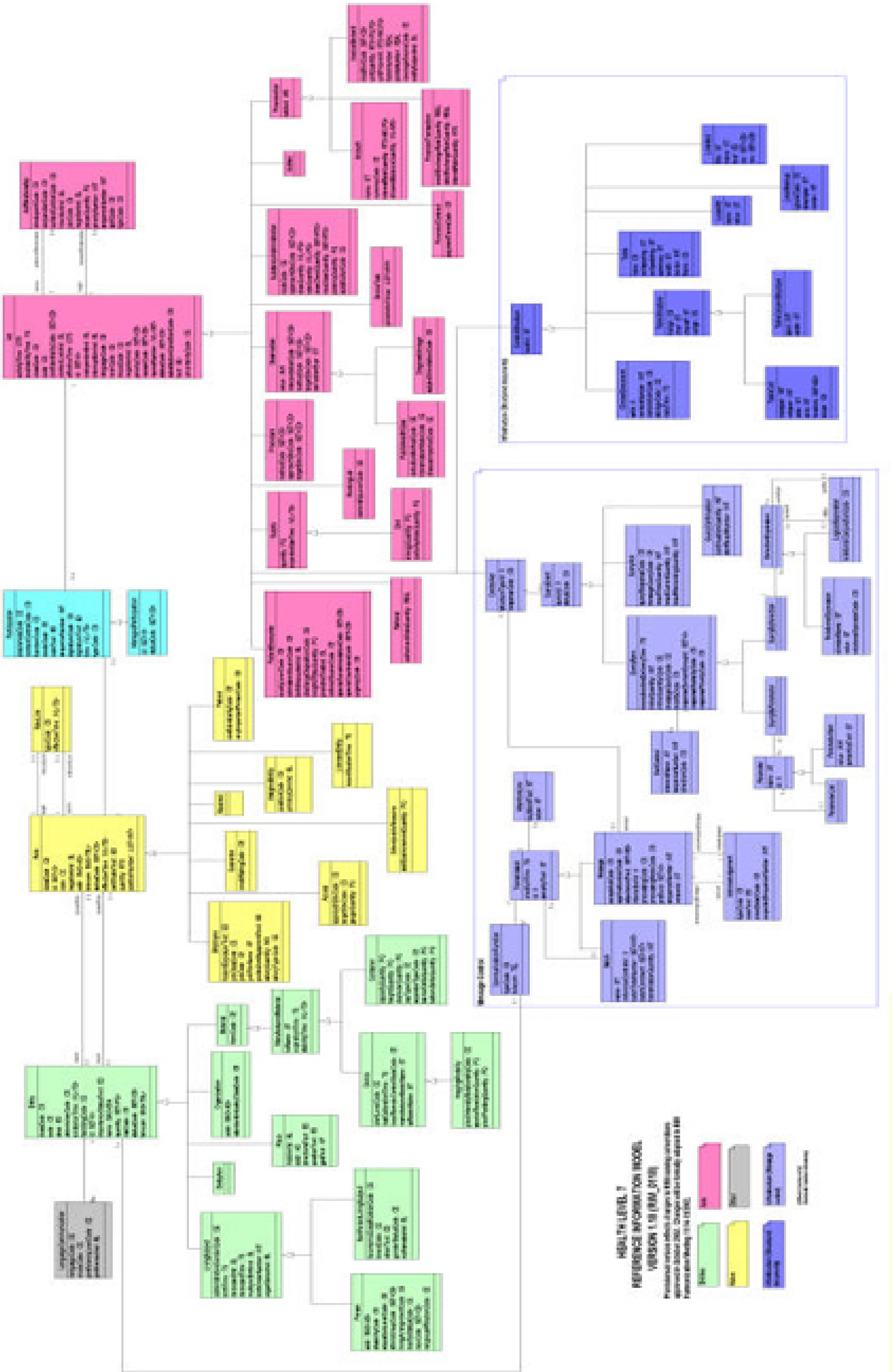
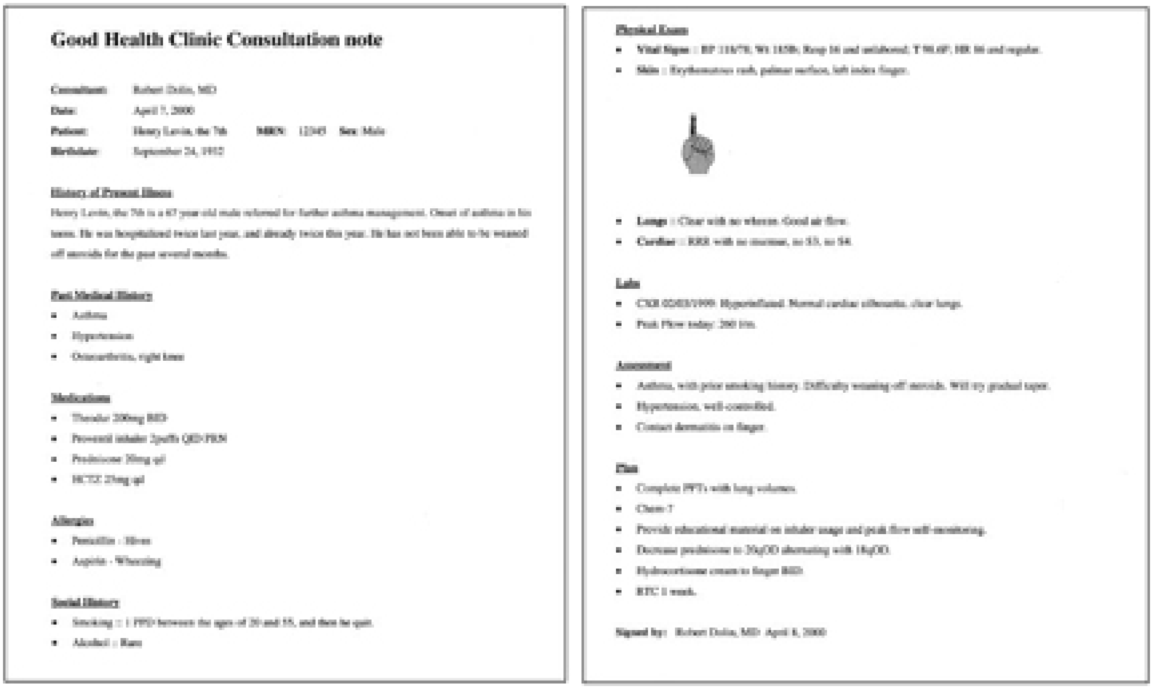


Figure A.1: Health Level Seven - Reference Information Model, Version 1.18



Good Health Clinic Consultation note

Consultant: Robert Dolin, MD
Date: April 7, 2000
Patient: Henry Levin, the 7th **MRN:** 12345 **Sex:** Male
Birthdate: September 24, 1932

History of Present Illness
Henry Levin, the 7th is a 67 year old male referred for further asthma management. Onset of asthma in his teens. He was hospitalized twice last year, and already twice this year. He has not been able to be weaned off steroids for the past several months.

Past Medical History
- Asthma
- Hypertension
- Osteoarthritis, right knee

Medications
- Theodur 200mg BID
- Proventil inhaler 2puffs QID PRN
- Prednisone 20mg qd
- HCTZ 25mg qd

Allergies
- Penicillin - Hives
- Aspirin - Wheezing

Social History
- Smoking :: 1 PPD between the ages of 20 and 55, and then he quit.
- Alcohol :: Rare

Physical Exam
- **Vital Signs** :: BP 118/70, Wt 185lb, Resp 16 and unlabored, T 98.6F, HR 86 and regular.
- **Skin** :: Erythematous rash, palmar surface, left index finger.
- **Lungs** :: Clear with no wheezes. Good air flow.
- **Cardiac** :: RRR with no murmur, no S3, no S4.

Labs
- CXR 02/03/1999: Hyperinflated. Normal cardiac silhouette, clear lungs.
- Peak Flow today: 260 l/m.

Assessment
- Asthma, with prior smoking history. Difficulty weaning off steroids. Will try gradual taper.
- Hypertension, well-controlled.
- Contact dermatitis on finger.

Plan
- Complete PFTs with lung volumes.
- Chem-7
- Provide educational material on inhaler usage and peak flow self-monitoring.
- Decrease prednisone to 20qOD alternating with 18qOD.
- Hydrocortisone cream to finger BID.
- RTC 1 week.

Signed by: Robert Dolin, MD April 8, 2000

Figure A.2: Sample Clinical Document (taken from Dolin's paper on HL7 CDA in 2001[38])

Appendix B

EVALUATION DETAILS

B.1 Description of Evaluators

The evaluators chosen for the study were professionals with clinical knowledge. Therefore, the evaluators mainly consisted of practising doctors, nurses, and clinical researchers. However, the study did not restrict the evaluators to belong to any particular clinical specialisation. The only two requirements were that (i) an evaluator should have sufficient knowledge to understand the clinical semantics of both the archetype fragments and their corresponding SNOMED codes returned as appropriate matches by the MoST system, and (ii) (s)he should be able to take clinically informed decisions on the equivalency of the SNOMED code(s) given the context in which the particular archetype fragment was used in the archetype. All the evaluators were connected either directly or indirectly to the UK National Health Service (NHS), which was helpful for the evaluation exercise, as they had some knowledge of using terminology codes (mainly Read Codes) to record patient data.

Requests for volunteers to evaluate the research application were sent to clinicians in the Medical Informatics Group at the University of Manchester (UoM) and University College of London (UCL), Manchester Children's University Hospitals NHS (MCUH), the Primary Health Care Specialist Group (PHCSG), and the UK Health Informatics Society (HiS). Voluntary evaluators from the MCUH, PHSCG, and HiS groups were mainly General Practitioners (GPs) and hospital clinical staff.

Im looking for volunteers to evaluate the system Ive developed to prove my research hypothesis. The main aim of the research application is to return a set of terminology codes, which the application assesses to be a close match to the input clinical data terms. The clinical data terms are extracted from a clinical data model (examples of clinical data models are HL7 messages and openEHR Archetypes).

For this research study, the clinical data models from which terms have been extracted are the openEHR Archetypes and the terminology system which is used to map each term to is SNOMED CT.

SNOMED CT is a controlled terminology system which covers a wide range of clinical specializations. SNOMED CT is the result of the merger of SNOMED RT (United States) and READ codes (UK) in 2002. The openEHR Archetype Model formalism is an approach for modeling data and is in the process of being standardized by the openEHR community in Europe. Archetypes are essentially nested hierarchies of data elements required for recording a particular clinical scenario (observation, assessment, intervention, etc.). A few examples of the archetype models published by openEHR are models for recording blood pressure, visual acuity, autopsy, barthel index, tendon and babinski reflexes, and so on.

The application focuses on identifying the SNOMED codes which semantically match to clinical data held in data models formulated using the openEHR Archetypes. The evaluation process will entail the evaluator to assess the SNOMED results obtained for the set of archetype terms present in an archetype model. The model will be pre-selected to avoid any waste of time. The assessment will include your views on whether the SNOMED result(s) returned for an archetype term is/are relevant (true positives), not relevant (false positives), or ambiguous w.r.t. the archetype model or SNOMED CT. Since several filtering rules have been applied to eliminate inappropriate results at various stages of the matching process logs are maintained containing the eliminated results. You will then need to browse through the log files (with the help of a GUI) to perform a further assessment of whether the eliminated results were correctly (true negatives) or incorrectly (false negatives) eliminated. An evaluation sheet will be provided to help note down your assessment.

The evaluation will last approximately two hours and will include first learning about the system and then critiquing the SNOMED matches obtained.

Figure B.1: Evaluation objective circulated amongst various groups to look for volunteers for the evaluation of the research work.

B.2 First Phase Evaluation Process

The first phase of the evaluation was a three-step process which included (i) the preparation stage, (ii) the execution stage, and finally (iii) the critiquing stage. Once they were familiarised with the graphical interface they were left to independently provide their scores and feedbacks for all the data sets in the archetype models assigned to them for evaluation.

Preparation Process

The evaluator was familiarised with Linköping University's Java Archetype editor[52], henceforth referred to as the Archetype Editor or Editor. It was important for the evaluator to browse through the archetype model in the Editor before performing the evaluation so that they could refer back to the model to clarify any queries they had with the model itself. The other reason for introducing them to this Archetype editor was because the MoST system was integrated with the editor as a plug-in application. As a plug-in application it was possible for a person to invoke the MoST application as a web service from the editor itself to get back a list of appropriate SNOMED codes per archetype fragment. Details of the system have been explained earlier in Chapter 3.

Once the evaluator was made familiar with the Archetype editor, (s)he was introduced to the functioning of the MoST system in the editor. Figure B.2 shows a snapshot of the MoST feature available in the Archetype editor. The evaluators were informed of the steps involved to load the MoST files related to the archetype model being evaluated at the time. For demonstration purpose, the *barthel index* archetype was chosen. The barthel index archetype is a nursing archetype that aims at determining the dependency of a patient in performing activities of daily living. The index serves as a measure to score a person's dependency often in nursing homes. This archetype was not reused later on during the evaluation exercise.

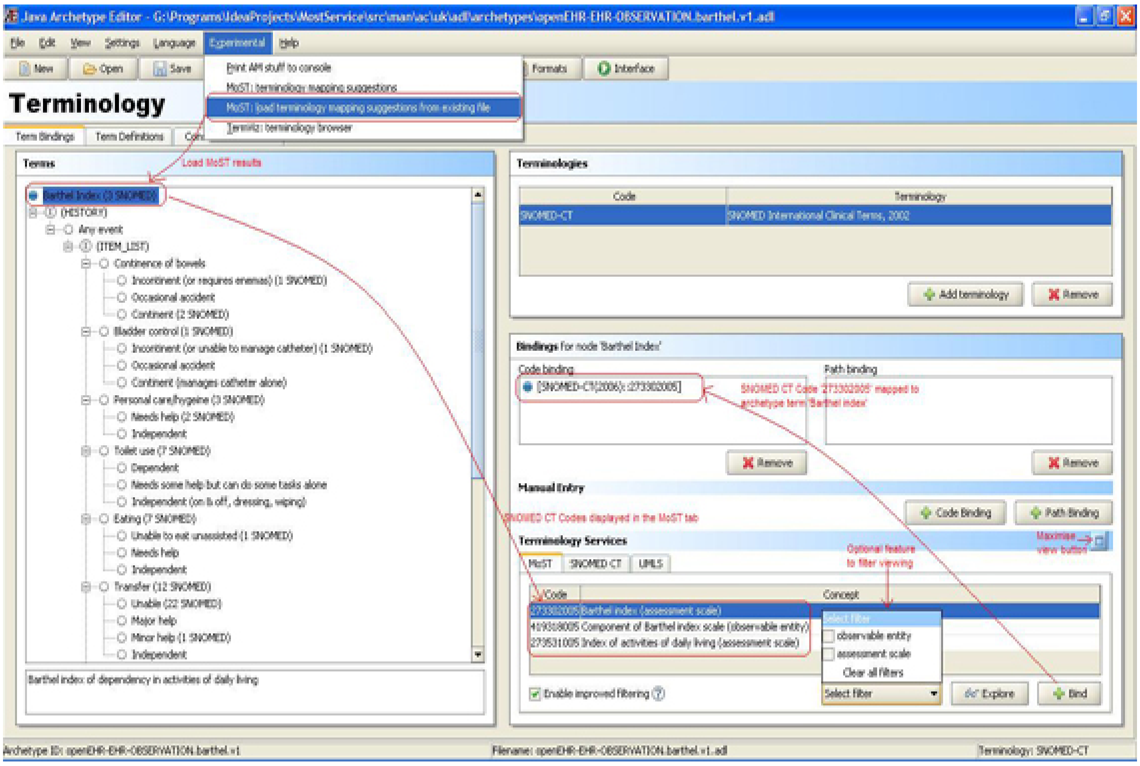


Figure B.2: The MoST System integrated as a plug-in to the Archetype Editor: The *barthel index* model with a view of the MoST menu and SNOMED CT results.

The evaluator was made aware that the results they were about to evaluate had been pre-run by the MoST system. The only reason for pre-running the archetypes was to save time during the evaluation process. With a limited time of 2 hours, it was considered more important to get the results from the MoST system evaluated rather than make the evaluator wait for the results to be generated by MoST dynamically.

In addition, the evaluators were also made familiar with the functionality of the different interface buttons and tabs related to the MoST interface. For example, the `maximise`, `filter`, and `bind` buttons, as shown in Figure B.2. The `maximise` and `filter` buttons addressed Human Computer Interaction (HCI) issues by enabling the

Editor to be more user friendly.

- The `maximise` button displayed all the SNOMED CT results returned by MoST in one large pop-up window removing the need to scroll through the default small window to browse through the results.

- The `filter` button helped the evaluator to view only those results that belonged to a particular SNOMED category(ies).

The use of these buttons was optional and non-critical to the performance or objective of the research application.

- The `bind` button is an important part of the Editor and enabled the evaluator to map one or more SNOMED codes that were considered equivalent semantic matches to an archetype fragment.

In Figure B.2, the SNOMED code `273302005` was mapped to the archetype fragment **`barthel index`**. Although the evaluator was not expected to bind SNOMED codes to every archetype fragment they were made familiar with the intented use of the SNOMED results they were evaluating.

Execution Process

On successfully familiarising the evaluator with the Editor and the MoST interface, the first archetype model to be evaluated was loaded into the Editor to illustrate the evaluation requirements. For instance, the *histology pap* model was loaded into the Editor. The evaluator was allowed to experiment with the **`Definition`** view of the model, which contained the data model structure. The pre-generated results from the MoST system were then loaded into the **`Terminology`** view of the model to setup the system to begin evaluation.

The evaluator was informed that (s)he was expected to record on an *evaluation sheet* ALL the SNOMED codes (s)he thought were most semantically equivalent to the archetype fragment. A sample of the evaluation sheet can be seen in Figure B.2. In addition to the final results returned by MoST, the evaluator was also required to review those codes that were filtered out by the MoST system to determine the existence of any false negatives i.e. incorrectly eliminated results. The log file, containing the filtered codes, could be loaded into the Editor and viewed in the same manner as the final result files, as both files had a similar XML schema that the Editor recognised.

The evaluator was provided with an *evaluation log sheet* to help record all false negatives along with an equivalency score between 0 and 10. A sample of the evaluation log sheet can be seen in Figures B.2.

The evaluator was asked to complete the remaining four columns on the evaluation sheet as shown in Figure B.2. They were:

- `Final Result Matches (SNOMED TERMS)`: The evaluator was required to record those SNOMED codes that (s)he found matched the archetype fragment both contextually (taking account of its parent in the model hierarchy) and semantically (with the help of the description provided by the original author of the model), if any. They could also provide any comments in this column.

- `Equivalence (Nonequivalent/Somewhat Equivalent/Equivalent)`: Each SNOMED code listed in the previous column was required to given an equivalency score between 0 and 10. Although the evaluators were not required to score those SNOMED codes which they considered to have zero equivalency, some evaluators preferred assigning scores to each SNOMED code returned by MoST. A zero score indicated that the SNOMED result had no remote equivalence to the archetype fragment. It was not required to manually assign a score of zero for each nonequivalent match, as all those SNOMED concepts with no scores assigned by the experts were automatically considered to have zero equivalency. On an average 'nonequivalent' codes were given a score between 0-3, 'somewhat equivalent' codes were scored between 4-6, and 'equivalent' codes were scored between 7-10. The most equivalent codes were often given a scoring of 9 or 10 by the evaluators.

- `Ambiguous (AM)`: At times, it was difficult for the evaluator to correctly infer the semantics of an archetype fragment given the context in which it was used in the model, and/or its definition provided by the model author. In such an event, the evaluator could mark the archetype model (AM) as being 'ambiguous' and provide any comments in the column, if they so desired. The ambiguity in the model may or may not have led to some or all of the SNOMED codes to have been scored for equivalency by the evaluator.

- `Ambiguous (SCT)` Finally, the evaluator might have found that although the SNOMED code rubric looked like an equivalent match to an archetype fragment at first glance, on closer examination of its definition in the SNOMED terminology model, the definition either caused the code to become nonequivalent or ambiguous. In such an event, the evaluator was required to state in this column

ARCHETYPE MODEL: Histology Pap

Evaluators Name

Date

SCORE : 0 -10 (NOTE: 0 means 'Completely different' and 10 means 'Most Equivalent')

ARCH TERM	ARCH PARENT	Final Result Matches (SNOMED TERMS)	Equivaelnce (Irrelevant/Somewhat/ Equivalent)	Ambiguous (AM)	Ambiguous (SCT)
Cervical Smear	<ROOT>				
Specimen Collection	Cervical Smear				
Clinical	Specimen Collection				
Specimen	-do-				
Description	Specimen				
Category	Specimen Collection				
ARCH TERM	**ARCH PARENT**	**Final Result Matches (SNOMED TERMS)**	**Equivaelnce (Irrelevant/Somewhat/ Equivalent)**	**Ambiguous (AM)**	**Ambiguous (SCT)**
Low grade epithelial abnormality	Category				
High grade epithelial abnormality	-do-				
Negative for dysplasia or malignancy	-do-				

Figure B.3: Evaluation sheet to score the results for the *histology pap* archetype.

that it was difficult to score a SNOMED code because of its ambiguous definition and/or placement in the concept hierarchy.

At the end of explaining the key requirements to the evaluator they were allowed to begin the evaluation process.

Critique Process

The critiquing process involved the actual critiquing and scoring of the results by the evaluator. They scored the final results and the log results one at a time. At first, they were required to determine the most equivalent SNOMED code(s) from the list of results present in the final XML file. On completing the evaluation of the final results, they were then required to browse through the log file to check for the existence of any SNOMED code(s) that had been incorrectly eliminated by the MoST system during the filtering rounds. The same method of making notes in the four columns, as stated in the `execution process` in Section B.2, applied to the log files as well. In addition, the evaluator was required to tick the column 'False Negative', as shown in Figure B.2, to highlight that a high equivalence scoring code had been found to have been incorrectly eliminated by MoST.

B.3 Phase 1 of Evaluation: Fleiss' Kappa Inter-rater Reliability

Cohen proposed the first Kappa statistic in 1960 to evaluate the extent of agreement between raters [48]. However, Cohen's kappa measures the agreement only between two raters. In order to address the issue of including more than two raters, Fleiss extended the inter-rater reliability assessment to the case of multiple raters with multiple response categories. He referred to this measure as the generalised Kappa statistic but is closer in measurement to Scott's Π statistic [48]. As with Cohen's kappa, Fleiss' kappa also takes into account the agreement occurring by chance, which makes it more reliable than simple percent agreement calculation. The kappa, κ, can be defined as [41][42]

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{B.1}$$

where, $1 - \bar{P}_e$ = degree of agreement that is attainable above chance,
$\bar{P} - \bar{P}_e$ = degree of agreement actually achieved above chance, and
$0 \leq \kappa \geq 1$; κ=1 implies complete agreement.

LOG FILE - Histology Pap

Evaluators Name
Date

SCORE : 0 -10 (NOTE: 0 means 'Most Irrelevant' and 10 means 'Most Relevant')

ARCH TERM	ARCH PARENT	Log Result Matches (SNOMED TERMS)	False Negative (Tick)	Relevance Score	Ambiguous (AM)	Ambiguous (SCT)
Cervical Smear	<ROOT>					
Specimen Collection	Cervical Smear					
Clinical	Specimen Collection					
Specimen	-do-					
Description	Specimen					
Category	Specimen Collection					
Low grade epithelial abnormality	Category					
High grade epithelial abnormality	-do-					

ARCH TERM	ARCH PARENT	Log Result Matches (SNOMED TERMS)	(Tick)	Relevance Score	Ambiguous (AM)	Ambiguous (SCT)
Negative for dysplasia or malignancy	-do-					
Technically unsatisfactory	-do-					
Inconclusive	-do-					
Specific cytological finding	Specimen Collection					

Figure B.4: Evaluation Sheet for the log file results of the *histology pap* archetype. Evaluation of the results in the log file help determine the existence of false negatives.

Fleiss' kappa has three axis of measurement i.e. number of raters (n), number of categories (N), and subjects of study (k). In the evaluation exercise, 'n' were the total number of experts evaluating an archetype, 'N' denoted the SNOMED codes returned by MoST, and 'k' denoted the scores given to each of the SNOMED codes on a scale of 0 to 10. It has been stated in a study by Sim and Wright [95] that the kappa will be higher when there are fewer categories. Table B.1 presents the interpretation of the kappa scores.

κ	**Interpretation**
$\prec 0$	No agreement
0 - 0.19	Poor agreement
0.20 - 0.39	Fair agreement
0.40 - 0.59	Moderate agreement
0.60 - 0.79	Substantial agreement
0.80 - 1.00	Almost perfect agreement

Table B.1: Interpretation of the kappa (κ) score range.

The inter-rater reliability was calculated for only the main archetype that was used in the evaluation i.e. the histology pap archetype. The histology pap archetype had the largest number of experts (i.e. ten experts) involved in its evaluation providing a more dependable scoring pattern. The other three archetypes have fewer evaluators, which could lead to a biased metrics. The two main reasons for limiting the calculation of the inter-rater reliability metrics to only one archetype were:

- *Large and variable number of MoST results:* Generally, there are a fixed and controlled number of categories, which are scored by the raters in common statistical studies. This was not the case in the evaluation exercise. There were a large and variable number of SNOMED results not only for each archetype model but also for each archetype fragment. Taking into consideration all the four evaluation archetypes, there were 122 archetype fragments that required semantic mappings. A total of 648 SNOMED codes were returned for the 122 fragments as post-filtered results. In addition to the post-filtered results, there were 1086 pre-filtered SNOMED codes that also needed to be scored to check for the presence of any false negatives.

  The number of SNOMED codes returned by MoST requiring evaluation were 1734 in total. It was considered highly impractical to carry out a code-by-code assessment of each of the 1734 SNOMED results returned. Besides, it was suspected that the assumption of assigning a score of zero to all those codes that had not been rated by the evaluators would lead to skewed measurements of inter-rater reliability formulated using Fleiss' Kappa. Therefore, it was decided that only a sample of SNOMED results from the histology pap archetype will be chosen to determine the inter-rater reliability. The sample would be chosen such that the rate would provide a fair indication of the overall inter-rater reliability, even though not all SNOMED codes were included for calculation.

- *No provision for considering post-coordinated MoST results:* Sometimes a score was given by an expert to a SNOMED code on the condition that it needed to be combined with another SNOMED code for semantic completeness. This meant that by itself the same SNOMED code would result in a zero score or a much lower scoring. However, this could not be accounted for when using Fleiss' kappa. Due to the insufficiency of the metrics to take into consideration dependent/partial scores, the inter-rater reliability score was not completely reliable for the study.

The two main limiting factors stated above and peculiar to the study necessitated multiple kappa calculations for the same histology pap archetype. Three separate types of calculations were used to analyse whether there was any significant change in the inter-rater reliability, based on the choice of SNOMED codes and scoring pattern.

In Types 1 and 2, the variables used were:
N = Sample of 40 SNOMED codes returned by MoST,
n = 10 clinical experts rating the SNOMED codes, and
k = Scores from 0 to 10, which implies k=11.

In Type 3, the variables used were:
N = Sample of 40 SNOMED codes returned by MoST,
n = 10 clinical experts rating the SNOMED codes, and
k = Scores grouped as 0-3 (low equivalence), 4-6 (moderate equivalence), and 7-10 (high equivalence), which implies k=3.

Inter-rater reliability for Histology Pap archetype

The three main types of kappa scores (as discussed in the previous section) calculated for the histology pap archetype were:

1. *Type 1 - Sample 'inclusive' of post-coordinated SNOMED codes:* In the first instance, the inter-rater reliability was determined by including all SNOMED

results that had been given a 'moderate' to 'high' equivalence score[1]. At times, the scoring had been given by the experts on the basis that the SNOMED code(s) required post-coordination with other SNOMED code(s). A sample of 40 codes were selected from a total of 177 post-filtered results in the histology pap archetype.

The kappa(κ) score reflecting the inter-rater reliability inclusive of the post-coordinated SNOMED codes is shown below:

$$\bar{P} = 0.336$$
$$\bar{P}_e = 0.265$$
$$\kappa = \tfrac{0.336-0.265}{1-0.265} = 0.0965$$

With reference to Table B.1, the kappa score of 0.0965 shows poor agreement amongst the 10 experts evaluating the archetype.

2. *Type 2 - Sample 'exclusive' of post-coordinated SNOMED codes:* In order to alleviate the oddities of including SNOMED codes that required post-coordination and their possible effects on the kappa score, a fresh sample was considered. This sample also consisted of 40 SNOMED codes but none of them had been stated as requiring post-coordination by any expert. Majority of the 40 codes were taken from the sample used in Type 1 above. The change in the kappa score can be seen below.

$$\bar{P} = 0.369$$
$$\bar{P}_e = 0.309$$
$$\kappa = \tfrac{0.369-0.309}{1-0.309} = 0.0868$$

Once again, the kappa score of 0.0868 shows poor agreement. This indicates that even for the SNOMED codes that were pre-coordinated matches to the archetype fragments, the level of agreement on their semantic equivalence was poor. As mentioned earlier, Sim and Wright[95] stated that a low kappa will be the result of a large number of categories. To test whether this was true, Type 3 grouped the scores to result in only 3 categories from the original 10 categories used in Types 1 and 2.

3. *Type 3 - Sample based on 'grouping' of SNOMED scores exclusive of post-coordinated*

[1] Low equivalence: Scores from 0-3; Moderate equivalence: Scores from 4-6; High equivalence: Scores from 7-10

codes: The final category for calculating the kappa score was to group the scores on the basis of their degree of equivalence i.e. low, moderate, and high. The same sample of 40 SNOMED codes chosen in Type 2 (i.e. 'exclusive' of post-coordinated SNOMED codes) was used in the Type 3. The only difference was the number of categories (i.e. `k`) considered. As stated earlier in this section, the scores given to the SNOMED codes were grouped into only three categories in Type 3, which means that k=3. The kappa score achieved was:

$$\bar{P} = 0.512$$
$$\bar{P}_e = 0.429$$
$$\kappa = \frac{0.512-0.429}{1-0.429} = 0.145$$

Although the kappa score of 0.145 is relatively higher than the kappa achieved in Types 1 and 2, it still highlights poor agreement amongst the experts/raters. This contradicted the claim by Sim and Wright[2] in this case, as the kappa remained low despite fewer categories in Type 3. It is not likely that the kappa would be higher even if all 177 post-filtered SNOMED codes in the histology pap archetype are selected.

Based on the kappa achieved using all three scoring patterns, it can be concluded that there was poor agreement amongst the 10 experts on the semantic equivalence of the SNOMED codes. A general observation made when aggregating the scores to calculate the kappa was that majority of the experts were divided in their opinion on scoring a code as either of 'low' or 'high' equivalence. For instance, Table B.2 shows the results from Type 3. It shows that the 40 SNOMED codes scored a total of 222 points in the category 0-3 (i.e. low equivalence). The next highest total score of 131 was obtained for the 'high' equivalence category (scoring between 7-10). The 'moderate' equivalence scoring category achieved the lowest total of 47 points. The total scores implied that with a nearly 50-50 split in the opinions of the experts, no consensus could be reached on the equivalence of the MoST results (i.e. SNOMED codes) for the histology pap archetype.

Despite the low kappa, the maximum attainable kappa (i.e. κ_{max}) for the histology pap archetype using Type 3 was 0.555 or 55.5%. As mentioned in Chapter 5 (see Section 5.3.4; Page 132), the κ_{max} value is to be considered, given the research peculiarities.

[2]Sim and Wright stated in [95] that the kappa will be higher when there are fewer categories.

	Histology Pap Archetype		
	[Type 3: Sample of 40 SNOMED Codes]		
	Low Equiv. (0-3)	**Moderate Equiv. (0-3)**	**High Equiv. (0-3)**
Total Score	222	47	131

Table B.2: The total score obtained by the 40 sample SNOMED codes, in each equivalence category in Type 3, for calculation of kappa of histology pap archetype.

Inter-rater reliability for Tendon Babinski archetype

A similar exercise for calculating Fleiss' kappa to determine the inter-rater reliability was performed for the Tendon Babinski archetype. However, only the Type 3 calculation (i.e. grouping of scores) was carried out. Approximately, 86% of the 40 code sample had a moderate-to-high equivalence rate (scores between 7-10), given by the 4 experts. The maximum attainable kappa (κ_{max}) for the archetype was 0.863 or 86%. However, despite the high rate of agreement amongst the experts (indicated by κ_{max}), the overall kappa score was only 0.15 indicating poor agreement.

$$\bar{P} = 0.792$$
$$\bar{P}_e = 0.755$$
$$\kappa = \frac{0.792-0.755}{1-0.755} = 0.151$$

	Tendon Babinski Archetype		
	[Type 3: Sample of 40 SNOMED Codes]		
	Low Equiv. (0-3)	**Moderate Equiv. (0-3)**	**High Equiv. (0-3)**
Total Score	6	16	138

Table B.3: The total score obtained by the 40 sample SNOMED codes, in each equivalence category in Type 3, for calculation of kappa of tendon babinski archetype.

As a result of the distorted results for the inter-rater reliability, kappa for the remaining three evaluation archetypes can be performed but is subject to the limitations stated at the start of this section.

B.4 Phase 2 of Evaluation: Comparative Data Analysis

Phase Two of the evaluation entailed a comparative data analysis of the SNOMED results returned by MoST. The three main categories for the data analysis were to determine (i) the top five scoring SNOMED codes, (ii) SNOMED codes that required post-coordination, and (iii) SNOMED codes that were not returned by MoST at any stage of the *term finding* process.

Top Five Scoring SNOMED Codes

The top five SNOMED codes that achieved the highest equivalence scores for each of the four evaluation archetypes were analysed. Results obtained for each archetype are listed below.

- *Histology Pap Archetype:* As stated earlier in Section B.3, the histology pap archetype had a low inter-rater reliability amongst the 10 experts evaluating the archetype. This meant that it was difficult to get consensus on any single SNOMED code from all 10 evaluators. In addition, the low mapping coverage of 48.6% and low trust score of 0.45 also reflected the poor quality of the archetype fragments and coverage of the SNOMED codes for the archetype as a whole. The top five SNOMED codes that achieved maximum consensus amongst the evaluators are listed in Table B.4.

Archetype Fragment	**SNOMED Concept Name**	**SNOMED Concept Id (code)**	**SNOMED Concept Category**	**Equivalence Score (%)**
Macroscopic	Macroscopic specimen observation	250429001	Finding	74
Microscopic	Microscopic specimen observation	395538009	Finding	72
Date recd. by laboratory	Date sample received in laboratory	281271004	Observable entity	67
Cervical Smear	Papanicolaou smear test	119252009	Procedure	63
Negative for dysplasia or malignancy	Negative for intraepithelial lesion or malignancy	373887005	Finding	63
Sample identification	Sample identification number	372274003	Observable entity	61

Table B.4: Top five scoring semantically equivalent SNOMED codes in the Histology Pap archetype.

- *Visual Acuity Archetype:* The high coverage rating of 86.9% and trust score of 0.8 indicates a high quality of SNOMED codes returned for the visual acuity archetype. A pattern was observed in the proportional relationship between the inter-rater reliability and the coverage rating. Therefore, based on the mapping coverage and trust score, it was ascertained that the inter-rater reliability would be 'substantial' or higher (see Table B.1). The top five SNOMED codes that achieved maximum consensus amongst the evaluators are displayed in Table B.5.

Archetype Fragment	**SNOMED Concept Name**	**SNOMED Concept Id (code)**	**SNOMED Concept Category**	**Equivalence Score (%)**
Left	Visual acuity - left eye	386708005	Observable entity	94
Right	Visual acuity - right eye	386709002	Observable entity	94
Visual Acuity	Visual acuity testing	16830007	Procedure	90
Perceive light	Perceives light only	260296003	Finding	88
Blind	Blind	277675000	Finding	86

Table B.5: Top five scoring semantically equivalent SNOMED codes in the Visual Acuity archetype.

Archetype Fragment	**SNOMED Concept Name**	**SNOMED Concept Id (code)**	**SNOMED Concept Category**	**Equivalence Score (%)**
Weight loss	Weight loss (amount)	363806002	Observable entity	90
Weight gain	Weight gain (amount)	363805003	Observable entity	90
Body weight	Usual body weight	363809009	Observable entity	85
Clothing	Garments	272180002	Physical object	57.5
Naked	Undressed	248160001	Finding	57.5

Table B.6: Top five scoring semantically equivalent SNOMED codes in the Body Weight archetype.

- *Body Weight Archetype:* The body weight archetype had a coverage rate of 70% and a trust score of 0.7, indicating a moderate quality of and coverage of the SNOMED codes for the archetype. Once again, it was asserted that the inter-rater reliability would most likely be 'moderate' or higher (see Table B.1). The top five SNOMED codes that attained the highest cumulative scores are listed in Table B.6.

- *Tendon Babinski Archetype:* Finally, the tendon babinski archetype achieved the highest coverage rate of 90.9% and trust score of 0.9. This indicated not only the high quality of the archetype and the coverage of SNOMED CT but also reflected the high level of inter-rater reliability.

Relevancy Score of SNOMED Codes = 100%		
Archetype Fragment	**Enclosing Fragment**	**SNOMED Concept**
Absent	Triceps	On examination - triceps reflex absent (finding)[163820006]
Reflex normal	Supinator	Supinator reflex normal (finding)[299838007]
Absent	Knee	On examination - knee reflex absent (finding)[163794001]
Reflex normal	Knee	On examination - knee reflex normal (finding)[163790005]
Absent	Ankle	On examination - ankle reflex absent (finding)[163811002]
Reflex normal	Ankle	On examination - ankle reflex normal (finding)[163807008]

Table B.7: The six SNOMED codes in the Tendon Babinski archetype which achieved 100% equivalency i.e. were exact matches.

In theory, the highest level of consensus that can be achieved by a SNOMED code is when the code is given a maximum equivalence score of 10 by all the evaluators. In other words, if a SNOMED code is found to be an exact match for an archetype fragment, the evaluator would assign it a score of 10. Therefore, in the case of the tendon babinski archetype, a code could achieve 100% equivalence if all four evaluators gave it a score of 10 each. The tendon babinski archetype was the only archetype that contained SNOMED codes that were given a 100% equivalence rating by all the evaluators. Table B.7 shows the six SNOMED codes that achieved 100%equivalence i.e. were exact matches.

The scores for fragments ‘absent’ and ‘reflex normal’ for biceps and triceps did not achieve 100% equivalency despite having very high scores. The reason was that the evaluators were not confident in giving the maximum score of 10 at the start of the evaluation. However, as the evaluation progressed they grew more confident and gave more accurate scores. Due to the large number of codes that achieved a total equivalence score above 90%, the top 10 codes have been displayed in Table B.8. These results are besides those codes which achieved a 100%

Relevancy Score of SNOMED Codes = 100%			
Archetype Fragment	**Enclosing Fragment**	**SNOMED Concept**	**Equivalence Score (%)**
Absent	Biceps	On examination - biceps reflex absent (finding)[163823008]	96.7
Reflex normal	Triceps	Triceps reflex normal (finding)[163816007]	96.7
Absent	Knee	Knee reflex absent (finding)[274817009]	96.7
Reflex present	Knee	Knee reflex reduced (finding)[299872004]	96.7
Reflex normal	Biceps	On examination - biceps reflex normal (finding)[163824002]	93.3
Triceps	Any event	On examination - triceps reflex (finding)[163815006]	93.3
Reflex normal	Knee	Knee reflex normal (finding)[299873009]	93.3
Absent	Ankle	Ankle reflex absent (finding)[274818004]	93.3
Reflex normal	Ankle	Ankle reflex normal (finding)[299879008]	93.3
Equivocal	Babinski response	Equivocal plantar response (finding)[246589002]	93.3

Table B.8: Top ten scoring SNOMED codes in the Tendon Babinski archetype.

equivalence rating. The table format has been modified slightly to accomodate the enclosing fragment of the archetype fragment, as there was replication of the degrees to measure the reflexes in different anatomical locations of the human body.

Table B.8 shows that there was a high level of consensus amongst the evaluators once a pattern was established on the kind of codes to select for a particular degree of reflex. There were often conflicting codes in SNOMED for the same fragment, such as `On examination - knee reflex absent (finding)`[163794001] and `Knee reflex absent (finding)`[274817009], making it difficult for the evaluators to determine the most appropriate code for mapping to the archetype fragment 'Absent'. A detailed review of the issues with the SNOMED hierarchy concerning reflex measurement codes has been dealt with in Chapter 6.

Post-Coordinated SNOMED Codes

In addition to the pre-coordinated SNOMED codes that could be directly mapped to the archetype fragments, there were some codes that required post-coordination. This meant, that a single code was insufficient to represent the true semantics of the archetype fragment. The experts suggested a combination of at least two codes, and based the scoring on the resultant post-coordinated concept.

- *Histology Pap Archetype:* There were 13 archetype fragments that required post-coordination. Four post-coordinated results have been shown in Table B.9. Table B.9 suggests that most of the '(qualifier value)' codes were considered insufficient to be selected on their own as a pre-coordinated match. A code from one of the main SNOMED categories was required to disambiguate the mapping.
- *Visual Acuity Archetype:* No post-coordination was identified by any of the five experts evaluating the MoST results for the visual acuity archetype. The high coverage rate of 86.9% indicates that SNOMED has a good coverage of codes for measuring and quanitfying visual acuity. Therefore, there are sufficient pre-coordinated SNOMED codes negating the need for any post-coordination of visual acuity concepts.
- *Body Weight Archetype:* Of a total of 10 body weight archetype fragments that found a MoST result, at least two experts agreed that 5 fragments required post-coordination. The SNOMED codes returned for all five fragments needed to be qualified by the SNOMED code `Weighing patient (procedure) [39857003]`.

POST-COORDINATED TERMS		
Archetype frag.	**SNOMED code 1**	**SNOMED code 2**
Low grade epithelial abnorm.	Low grade (qualifier value) [349915008]	Epithelial cell abnorm. (morphologic abnormality) [373886001]
Inconclusive	Inconclusive (qualifier value) [419984006]	Cervical cytology finding (finding)[302796001]
Contaminated specimen	Contaminated (qualifier value) [62604006]	Specimen obscured by foreign material (finding)[54192004]
Findings	Morphologic finding (finding)[72724002]	Microscopic specimen obs. (finding)[395538009]

Table B.9: Post-coordinated terms: Archetype fragments which required to be mapped to a composition of SNOMED code 1 and 2 in the Histology Pap archetype.

As shown in Table B.10, the qualifying code `39857003` suggested that SNOMED code 1 (2nd column in Table B.10) related to the procedure of weighing a patient.

POST-COORDINATED TERMS		
Archetype fragment	**SNOMED code 1**	**SNOMED code 2**
Clothing	Garments (physical object) [272180002]	Weighing patient (procedure)[39857003]
Lightly clothed/ underwear	Undergarment (physical object)[228164001]	-do-
Naked	Undressed (finding) [248160001]	-do-
Fully dressed	Garments (physical object) [272180002]	-do-
Nappy	Diaper, device (physical object)[52065008]	-do-

Table B.10: Post-coordinated terms: Archetype fragments, which required to be mapped to a composition of SNOMED code 1 and 2 in the Body Weight archetype.

- *Tendon Babinski Archetype:* The tendon babinski archetype achieved the highest coverage of 90.9% amongst the four evaluation archetypes. However, a few SNOMED codes required post-coordination before being mapped to their corresponding archetype fragments. Most of these post-coordinated concepts were in addition to the pre-coordinated SNOMED codes returned by MoST.

The two archetype fragments that caused ambiguity amongst the evaluators

POST-COORDINATED TERMS		
Archetype fragment	**SNOMED code 1**	**SNOMED code 2**
+++ (Reflex increased)	Triceps reflex (observable entity) [271716008]	Hyperreflexia (finding) [86854008]
++++ (Reflex markedly increased)	Triceps reflex (observable entity) [271716008]	Hyperreflexia (finding) [86854008]

Table B.11: The two tendon reflex measures that can be post-coordinated in the Tendon Babinski archetype.

were *+++ (reflex increased)* and *++++ (reflex markedly increased)*. For instance, the evaluators selected the same SNOMED code `Triceps reflex brisk (finding)[299827007]` for mapping to both the fragments to record the triceps reflexes. The evaluators suggested that a SNOMED code `Hyperreflexia (finding)[86854008]` present in the pre-filtered MoST results could be used to post-coordinate either of the two ambiguous fragments to increase mapping coverage. Table B.11 shows the post-coordination of the two archetype fragments.

POST-COORDINATED TERMS			
Archetype Fragment	**Term Definition**	**Enclosing Fragment**	**SNOMED Concept**
Strongly down-going	Babinski response is strongly toward the sole of the foot	Babinski response	Structure of sole of foot (body structure)[57999000] + **?**
Down-going	Babinski response is toward the sole of the foot	-do-	-do-
Up-going	Babinski response is toward the dorsum of the foot	-do-	Structure of dorsum of foot (body structure)[2402003] + **?**
Strongly up-going	Babinski response is strongly toward the dorsum of the foot	-do-	-do-

Table B.12: Archetype fragments requiring post-coordination in the Tendon Babinski archetype. NOTE: '`?`' indicates the absence of any appropriate SNOMED codes to use for post-coordination.

Although the same post-coordinated suggestions were made for both the fragments, higher scores were given for post-coordinating the fragment *++++ (reflex markedly increased)*. It was unclear to the evaluators what was the intent of the author of the archetype when using the labels 'increased' and 'markedly increased'. In the absence of a clearer definition, they agreed that the label 'markedly increased' was semantically closer to the SNOMED label `hyperreflexia`.

In addition to the post-coordination of the triceps reflex fragments shown in Table B.11, the 'babinski response' fragments also required post-coordination. As shown in Table B.12, the four babinski response fragments were authored to record the various reflexes found in the sole and dorsum of the foot. However, no corresponding pre-coordinated SNOMED codes could be found to map to these fragments. The closest match found in SNOMED was `Plantar reflex finding (finding)[366553001]`, which was a synonym for 'babinski reflex finding'. However, it was too general for the archetype fragments shown in Table B.12. The more specific codes available in SNOMED with respect to plantar reflexes, did not conform to the intended meaning of these fragments. Due to the lack of any suitable compositional codes, a question mark appears beside the partial `(body structure)` codes displayed in Table B.12.

Missed SNOMED Codes

Finally, it was not always the case that all relevant SNOMED codes for the archetype fragments were found by the MoST system. There were two main reasons for missed matches: (i) insufficiency of the archetype fragment labels, definitions, and enclosing fragments, and (ii) limitations in the search techniques used by MoST. The first cause for missed codes (i.e. insufficiency of the archetype) could not be corrected without revising the entire archetype. However, the second cause (i.e. limitations of MoST) was rectified by improving the processes used when searching for relevant SNOMED codes. The results discussed in the thesis are for the SNOMED codes returned by MoST for the original archetypes. Revision of the archetypes to improve the quality of the MoST results was performed only for the histology pap archetype, and has been demonstrated in Section 6.2 (Page 6.2).

A SNOMED code was considered to have been missed by the MoST system, if at least two experts mentioned it during evaluation. Based on the feedback received from a majority of the experts, the histology pap and tendon babinski archetypes were found to be lacking in some SNOMED codes.

- *Histology Pap Archetype:* Five SNOMED codes were not returned by MoST at any point during the search process but were thought to be of relevance by the

MISSED CODES	
Archetype frag.	**SNOMED code**
Cervical Smear	1.Cervical cytology finding (finding)[302796001] 2.Cervical smear result(finding)[269957009]
Specimen	1.Cervical smear sample(specimen)[276446009] 2.Specimen from uterine cervix specimen(specimen)[119395005] [119395005]
Comment	1.Narrative comments on pathology specimen (obs. entity) [409770001]

Table B.13: Missed Codes: SNOMED codes not returned by MoST at any stage of the *term finding* process for the histology pap archetype.

evaluators. These codes were looked up directly in SNOMED using the CliniClue Browser [9]. The codes have been displayed in Table B.13.

DISPUTE TERMS		
Archetype frag.	**SNOMED matches**	**Total score**
Specific cytological findings	1.Cytologic (qualifier value)[40413002] 2.Abnormal cytological findings in CSF (finding) [274686008] 3.Abnorm cytolog finding specimens from female genital organs (finding)[274688009]	10 9 9
Comment	1.Provider comment report(record artifact)[371541002] 2.Thought commentary(finding)[307078008]	11 10
Image	1.Information from images (attribute) [405671001] 2.Image analysis (procedure)[24587005]	15 15

Table B.14: Dispute terms: Archetype fragments with closely rated SNOMED matches displaying lack of consensus.

The histology pap archetype was particularly ambiguous. In addition to the missed codes arising mainly due to the insufficiency of the archetype, there was also dispute on a few SNOMED codes. In other words, there was lack of consensus on the relevancy of certain SNOMED codes for a few archetype fragments, which led to approximately equal scores. As shown in Table B.14, no consensus could be reached by the ten evaluators on three archetype fragments in the histology pap archetype.

- *Tendon Babinski Archetype:* Feedback received from the evaluators highlighted that eight SNOMED codes were not returned by MoST at any time during the

MISSED CODES		
Archetype fragment	**Enclosing fragment**	**SNOMED code**
Reflex present	Triceps	Triceps reflex reduced (finding)[299825004]
Reflex markedly increased	Triceps	On examination - triceps reflex exaggerated (finding)[163821005]
Supinator	Any event	Supinator reflex finding (finding)[366452008]
Knee	Any event	Knee reflex finding (finding)[366536008]
Ankle	Any event	Ankle reflex finding (finding)[366547001]
Reflex present	Ankle	Ankle reflex reduced (finding)[299878000]
Reflex markedly increased	Ankle	On examination - ankle reflex exaggerated (finding)[163812009]
Equivocal	Babinski response	Plantar grasp reflex equivocal (finding)[299767003]

Table B.15: Missed Codes: SNOMED codes that were not returned by MoST at any stage of the *term finding* process in the Tendon Babinski archetype.

term finding process. Again, the missing codes were revealed on manual lookup in the CliniClue Browser. The codes have been displayed in Table B.15.

B.5 Evaluation Results Sheet for the Archetypes

Histology Pap Archetype				
Archetype Fragment	**Total SNOMED Results (log)**	**Total SNOMED Results (final)**	**No. of terms relevant (final)**	**No. of terms relevant (log)**
	(TNR+FNR)	**(TPR+FPR)**	**(TPR)**	**(FNR)**
Cervical Smear	15	15	5	0
Specimen Collection	21	11	3	0
Clinical	10	7	1	0
Specimen	9	3	2	0
Description	8	2	3	1
Category	23	11	2	1
Low grade e.a.	12	6	3	0
High grade e.a.	9	2	2	0
Neg. dysplasia	10	1	1	2
Tech. unsatisfactory	12	0	0	1
Inconclusive	2	1	1	0
Specific cytological finding	16	5	1	3
Macroscopic	9	4	2	1
Appearance	3	3	2	0
Image	21	16	4	0
Microscopic	2	2	2	0
Findings	12	11	4	1
Image	21	16	4	0
Adequacy	9	7	2	0
Statement	15	6	2	0
Satisfactory for assessment	16	1	1	1
Adequate nos. of squamous cells	15	0	0	1
Quality	3	2	1	0
Labeling	14	3	1	0
Slide unlabeled	2	1	1	0
Container unlabeled	6	1	1	0
Specimen unlabeled	7	1	1	0
Preservative	7	1	1	0
Incorrect preservative	13	0	0	0
Insufficient preservative	10	1	0	1
Specimen	2	0	0	1
Insufficient specimen	6	3	2	0
Damaged spec.	6	1	0	0
Contaminated spec.	4	1	1	1
Comment	5	3	3	1
Multimedia rep.	1	0	0	0
Identification	22	12	2	1
Sample id	7	4	2	0
Laboratory	10	6	3	0
Process	12	5	2	0
Date recd. by lab	6	2	1	0

Table B.16: Detailed data distribution of results for the Histology Pap archetype model.

Visual Acuity Archetype				
Archetype Fragment	**Total SNOMED Results (log)**	**Total SNOMED Results (final)**	**No. of terms relevant (final)**	**No. of terms relevant (log)**
	(TNR+FNR)	**(TPR+FPR)**	**(TPR)**	**(FNR)**
Visual acuity	18	8	4	3
Left	13	12	5	1
Right	13	11	6	1
Both eyes	18	6	1	4
Visual acuity	11	9	5	2
5/6	3	1	0	1
6/6	15	9	6	0
6/7.5	0	0	0	0
6/9	6	5	5	0
6/12	6	5	5	0
6/18	6	5	5	0
6/36	6	5	5	0
6/60	6	5	5	0
Count fingers	10	9	6	1
Perceive light	5	5	5	0
Blind	5	1	1	0
Use of visual aid	11	8	1	0
Aided	18	8	1	2
Unaided	7	6	1	0
Type of visual aid	6	2	2	4
Glasses	20	11	3	1
Pinhole	12	6	3	5
Not known	13	9	3	5
Method	12	4	3	0

Table B.17: Detailed data distribution of results for the Visual Acuity archetype model.

Body Weight Archetype				
Archetype Fragment	**Total SNOMED Results (log)**	**Total SNOMED Results (final)**	**No. of terms relevant (final)**	**No. of terms relevant (log)**
	(TNR+FNR)	**(TPR+FPR)**	**(TPR)**	**(FNR)**
Body weight	9	4	3	0
Weight	25	10	3	0
Weight loss	13	9	2	0
Decrease	0	0	0	0
Weight gain	12	9	3	0
Increase	0	0	0	0
Clothing	3	2	1	0
Lightly clothed/ underwear	4	1	1	0
Naked	7	2	1	0
Fully dressed	8	0	0	1
Nappy	5	1	1	0
Device	3	1	1	1

Table B.18: Detailed data distribution of results for the Body weight archetype model.

Tendon Babinski Archetype				
Archetype Fragment	**Total SNOMED Results (log)**	**Total SNOMED Results (final)**	**No. of terms relevant (final)**	**No. of terms relevant (log)**
	(TNR+FNR)	**(TPR+FPR)**	**(TPR)**	**(FNR)**
Tendon and Babinski reflex	15	9	1	3
Left	11	9	2	0
Right	11	8	2	0
Biceps	23	15	5	1
Absent	4	4	2	0
+/-	3	0	0	1
+	6	6	3	0
++	14	12	4	1
+++	13	12	3	0
++++	10	10	1	0
Triceps	30	19	6	1
Absent	3	3	3	0
+/-	3	0	0	0
+	3	3	2	0
++	7	7	4	0
+++	11	10	2	1
++++	10	10	2	0
Supinator	7	6	3	1
Absent	2	2	2	0
+/-	3	0	0	0
+	4	4	3	0
++	7	6	2	0
+++	11	9	1	1
++++	10	9	1	0
Knee	18	15	1	1
Absent	4	4	3	0
+/-	3	0	0	0
+	4	4	3	0
++	7	7	3	0
+++	11	10	1	1
++++	10	10	2	0
Ankle	17	14	2	0
Absent	4	4	3	0
+/-	3	0	0	0
+	3	3	2	0
++	8	8	3	0
+++	11	10	1	1
++++	10	10	2	0
Babinski response	0	0	0	0
Strongly down-going	3	0	0	1
down-going	3	0	0	1
equivocal	7	7	2	0
up-going	9	0	0	1
Strongly up-going	3	0	0	1
With recruitment	5	0	0	0

Table B.19: Detailed data distribution of results for the Tendon babinski archetype model.