2011-12-17

Clinical Coverage of an Archetype Repository Over SNOMED-CT.

Sheng Yu
Sheng Yu, sheng.yu@tudublin.ie

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

Jesus Bisbal
Universitat Pompeu Fabra, Barcelona

Recommended Citation

Yu, S., Berry, D., Bisbal, J.: Clinical coverage of an archetype repository over SNOMED-CT. *Journal of Biomedical Informatics*, accepted December 2011. Elsevier. doi:10.1016/j.jbi.2011.12.001

Journal of Biomedical Informatics 45 (2012) 408–418

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Clinical coverage of an archetype repository over SNOMED-CT

Sheng Yu [a,*], Damon Berry [a], Jesus Bisbal [b]

[a] *Dublin Institute of Technology, School Elect. Eng. Systems, Dublin, Ireland*
[b] *Universitat Pompeu Fabra, Department of ICT, Barcelona, Spain*

ARTICLE INFO

Article history:
Received 18 June 2011
Accepted 10 December 2011
Available online 17 December 2011

Keywords:
Clinical archetypes
Conceptual modelling
Ontologies
Terminology systems
SNOMED-CT
Term-binding

ABSTRACT

Clinical archetypes provide a means for health professionals to design what should be communicated as part of an Electronic Health Record (EHR). An ever-growing number of archetype definitions follow this health information modelling approach, and this international archetype resource will eventually cover a large number of clinical concepts. On the other hand, clinical terminology systems that can be referenced by archetypes also have a wide coverage over many types of health-care information.

No existing work measures the clinical content coverage of archetypes using terminology systems as a metric. Archetype authors require guidance to identify under-covered clinical areas that may need to be the focus of further modelling effort according to this paradigm.

This paper develops a first map of SNOMED-CT concepts covered by archetypes in a repository by creating a so-called terminological *Shadow*. This is achieved by mapping appropriate SNOMED-CT concepts from all nodes that contain archetype terms, finding the top two category levels of the mapped concepts in the SNOMED-CT hierarchy, and calculating the coverage of each category. A quantitative study of the results compares the coverage of different categories to identify relatively under-covered as well as well-covered areas. The results show that the coverage of the well-known National Health Service (NHS) Connecting for Health (CfH) archetype repository on all categories of SNOMED-CT is not equally balanced. Categories worth investigating emerged at different points on the coverage spectrum, including well-covered categories such as *Attributes*, *Qualifier value*, under-covered categories such as *Microorganism*, *Kingdom animalia*, and categories that are not covered at all such as *Cardiovascular drug (product).*

1. Introduction

The specifications released by the openEHR foundation and as part of the CEN/ISO EN13606 standard for Electronic Health Record (EHR) communication [1–3] define, among other things, an *Archetype Model* [4], and specify how to construct archetypes to express constraints on clinical information in an EHR. The Archetype Definition Language (ADL) is a formal language that can be used to create archetypes that in turn express constraints on the EHR data. Archetypes are information modelling artifacts (see Section 2.1) that are used in advanced e-Health modelling methodologies, and the number of archetypes is growing steadily to cover many clinical specialties. In order to accommodate the growing number of archetypes, WEB-based archetype authoring and management platforms have been developed to maintain the repositories of archetypes [5]. These repositories also include the capability of organising and categorising archetypes under certain classifications.

SNOMED-CT [6], in contrast to the contents of an archetype repository, is intended to facilitate semantic interoperability by providing a comprehensive set of commonly understood clinical concepts. If SNOMED-CT is to deliver on this objective, it will increasingly cover the space of clinical findings, diagnosis, anatomy, drugs, and physical objects that relate to health-care activity. SNOMED-CT, with its large scope and user base, is a powerful external reference in the medical domain with a capability to enable communicating parties to unambiguously convey clinical concepts. The completeness of its clinical concept modelling in each medical domain provides health professionals with a sufficient number of concepts to express clinical statements in many clinical scenarios [7,8]. Thus, it can be considered an appropriate metric to use when measuring the coverage of clinical concepts of an archetype repository. However, no previous work reports the clinical content coverage of archetypes with respect to the terminology in clinical terminology systems, and the distribution of clinical concepts. Nor is there a detailed description in the literature of a method to obtain this coverage information. Such information would help to identify clinical specialities where archetypes provide insufficient support, and therefore more archetype development work is required.

* Corresponding author. Fax: +353 1 402 7980.
E-mail addresses: sheng.yu@dit.ie (S. Yu), damon.berry@dit.ie (D. Berry), jesus.bisbal@upf.edu (J. Bisbal).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.12.001

Many archetype terms are intended to link to universally understood medical references, such as SNOMED-CT, so that when sharing archetypes, the meaning can be conveyed to the communicating parties. This use of external coding systems has the potential to provide a variety of means to classify and categorise archetype repositories. However, not all archetype terms are linked or linkable with external terminological references. Nevertheless, the presence of two large and related information resources, an archetype repository and a terminological system, presents an opportunity to investigate how clinical information is modelled in contemporary repositories.

The primary focus of this paper is to identify the coverage of SNOMED-CT concepts in an archetype repository. By mapping terms that are defined internally in Archetypes to standard SNOMED-CT concepts, it is possible to obtain an approximate overview of the equivalent terminological content of the archetype repository. This work therefore harmonises two modelling approaches in health information systems. On the one hand, an archetype model is representative of clinical meta-data modelling methods that expresses the commonly agreed clinical contents of EHRs. On the other hand, SNOMED-CT is representative of clinical terminology, and models clinical knowledge using an ontological approach. Both modelling approaches are designed to work with EHR systems for different purposes, but gaps in coverage, overlaps and similarities are likely between the two approaches. As discussed in [9], increasing integration of the two approaches will facilitate semantic interoperability in e-Health.

The results reported here will allow both communities (archetype modellers and terminologists) to address interoperability issues that arise due to embedded codes. In addition, the outcomes of this research will allow the identification of which categories in SNOMED-CT are more thoroughly covered by modelling artifacts like archetypes, and which ones still require more attention. These results should help in focusing the significant modelling efforts undertaken by the community.

This paper is organised as follows. Sections 2 and 3 describe the main concepts involved in this work and provide an overview of the state of the art. Section 4 describes the calculation of coverage of an archetype repository over SNOMED-CT top level categories. Section 5 summarises the results of the investigation, while Section 6 discusses these results and analyses certain outstanding coverage results. Finally, Section 8 summarises the paper, and Section 7 outlines directions for future work.

2. Background

This section briefly introduces some concepts related to archetypes as well as SNOMED-CT, used in the remainder of the paper.

2.1. Two-level modelling

The concept of *Archetypes*, used in the previous section, is part of an advanced information modelling methodology referred to as *Two-Level Modelling* [10,2,3]. Instead of trying to capture all required information in a single large data model [1], this methodology advocates a separation between *information* and *knowledge*. The former, it is claimed, can be generically represented with a rather simple data model that is stable over time, while the latter is modelled as instances of the former. Information (data model) is the first level in this approach, and knowledge is the second level.

The term *Archetype* has become widely used to refer to the knowledge represented in the second level of this approach (e.g. blood pressure). As the nature of the information in an EHR is updated, archetypes are maintained accordingly. However, as these only describe instances of the underlying data model, commonly termed a *Reference Model* (composed by abstract building blocks like Element, Item, Entry, Section, and Composition), this update will not require changes to the clinical application. This approach, then, is expected to shield applications from large numbers of evolutionary changes that would be needed if using the traditional single model approach [1]. Also, Archetypes are the basis upon which information exchange and interoperability can be achieved.

2.2. Term binding in ADL

Archetypes may contain tens if not hundreds of freely designed data nodes that express clinical meanings. As part of the archetype specification, the syntax of the Archetype Definition Language (ADL) [4] includes a mechanism to allow annotation of clinical concepts in archetypes by defining local terms. These local terms are specified as 'AT codes', where the 'AT' stands for 'Archetype Term'. A dedicated section is provided in each archetype to expand the explicit meaning of these terms and occasionally a 'presentation name' for display on a screen. Archetype terms can also be optionally linked to terms in external terminologies such as SNOMED-CT, known as *term binding* in the ADL syntax. These bindings to local 'AT' coded terms in archetypes can be used to retrieve a commonly understood medical definition. The following example shows a snippet of such an archetype definition (in ADL syntax) where the locally defined code 'at0021' is linked to the SNOMED-CT code 162465004 for severity, while the actual values (mild, moderate, etc.) are local and do not map to SNOMED-CT.

```
...
ELEMENT[at0021] occurrences matches {0..1} matches
  {–Severity
  value matches {
    1|[local::at0044], – trivial
    2|[local::at0023], – mild
    5|[local::at0024], – moderate
    8|[local::at0025], – severe
    9|[local::at0045] – very severe
        }
...
term_bindings = <
    ["SNOMED-CT"] = <
        items = <
        ["at0021"] = <[SNOMED-CT::162465004]>
...
```

2.3. Structure of SNOMED-CT

To support the coverage assessment in this paper, the hierarchical structure of SNOMED-CT is studied and adopted as the base of the coverage calculation. Fig. 1 depicts the hierarchical structure of SNOMED-CT, which consists of a concept called 'SNOMED-CT' as the root node of all concepts, and its 19 first-level categories in the concept model range from *body structure* to *physical object*. Table 1 lists the information for all 19 first-level categories of SNOMED-CT. The numbers indicate the size, i.e. the total number of concepts under each first-level category, and these categories are listed in descending order.

Each category represents an abstract clinical classification, each of which are sub-classified in turn by second-level categories. Second-level categories are again sub-classified into child concepts, and this structure continues further down the concept hierarchy until very specific concepts are reached. The SNOMED-CT concept model allows multiple-inheritance, which means that concept 'a'

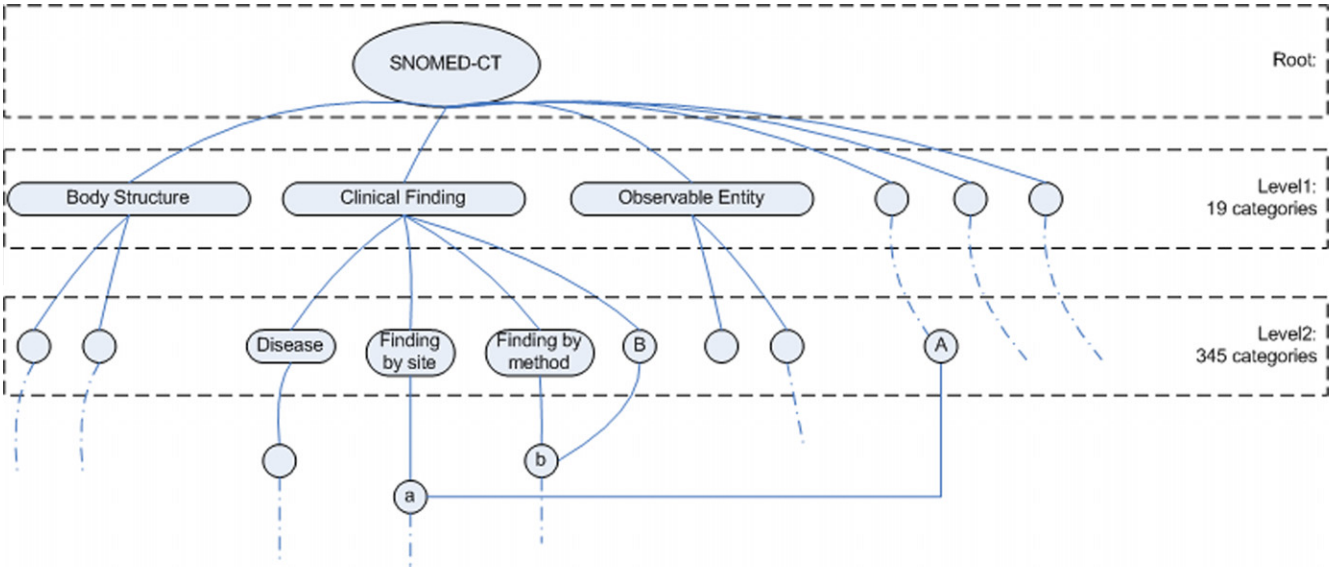

Fig. 1. A hierarchical representation of SNOMED-CT, showing *SNOMED-CT* as the root node of all concepts, that is categorised into 19 first-level, concepts and 343 second-level concepts.

Table 1
First-level categories in SNOMED-CT, release January 2008.

First-level category name	Size
Clinical finding (finding)	109,311
Special concept (special concept)	67,342
Procedure (procedure)	53,854
Body structure (body structure)	31,837
Organism (organism)	27,952
Substance (substance)	23,456
Pharmaceutical/biologic product (product)	19,084
Qualifier value (qualifier value)	8904
Event (event)	8447
Observable entity (observable entity)	7834
Social context (social concept)	5252
Situation with explicit context (situation)	4912
Physical object (physical object)	4515
Environment or geographical location (environment/location)	1741
Linkage concept (linkage concept)	1136
Staging and scales (staging scale)	1113
Specimen (specimen)	1055
Record artifact (record artifact)	202
Physical force (physical force)	172
Total concepts	378,111

can both belong to category ‘Finding by site’ and category ‘A’, as depicted in Fig. 1.

2.4. Archetype repository

An archetype repository manages the life-cycle of archetypes. The repository primarily maintains the inheritance and versioning of archetypes. All committed archetypes follow the naming convention that is defined in the specifications for the Archetype Object Model. The naming scheme allows generic archetypes to be extended by more specific archetypes. To extend an archetype, the resulting archetype must inherit information in its parent archetype and add more specific clinical content. Fig. 2, as an example, shows a *Pain symptom* archetype that could be an extension of the generic *Symptom* archetype with more specific clinical details. *Versioning* is a mechanism that allows different versions of archetypes to co-exist for the purpose of tracing their editing history as the development goes on. The latest version is assumed to be most mature in quality and content. As Fig. 2 demonstrates, both inheritance and versioning produce duplicate counts for the same archetype term that will be removed in the archetype term extraction process, described in Section 4.

3. Related work

Due to the textual nature of clinical information, association of medical text and codes to a standard terminology usually involves complex mapping processes. Much existing work involves the mapping of arbitrary clinical text with standard terminologies. This can be considered a generalised case of concept mapping between archetype terms and SNOMED-CT, whether it is done manually or automatically. Related work can be classified into the following three main categories, with slightly different aims.

3.1. General purpose mapping tools

RELMA is a helper application released by the Regenstrief Institute, Inc. for mapping local tests (and other observation codes) to the Logical Observation Identifier Names and Codes (LOINC) database on a one-at-a-time basis [11]. This manual process is aided by an additional component called the *Intelligent Mapper*, which processes the mapping automatically and has been algorithmically improved [12].

MetaMap is a highly-configurable program that has been developed by the National Library of Medicine (NLM) to map arbitrary free text in the biomedical field to the UMLS thesaurus [13]. During its evolution, various new techniques have been adapted, notably Natural Language Processing (NLP) technology such as word sense disambiguation (WSD) [14]. By design, MetaMap is tightly coupled with the UMLS [15]. The program itself bundles the pre-computed database of all UMLS concepts. Its algorithm is intended to work with UMLS concepts. Other terminologies that have not been incorporated into the UMLS (such as locally defined ones) are not well supported by the tool.

The focus of these tools is to provide a generic means to map unstructured clinical data, often free text, to specific terminologies such as LOINC and the UMLS thesaurus.

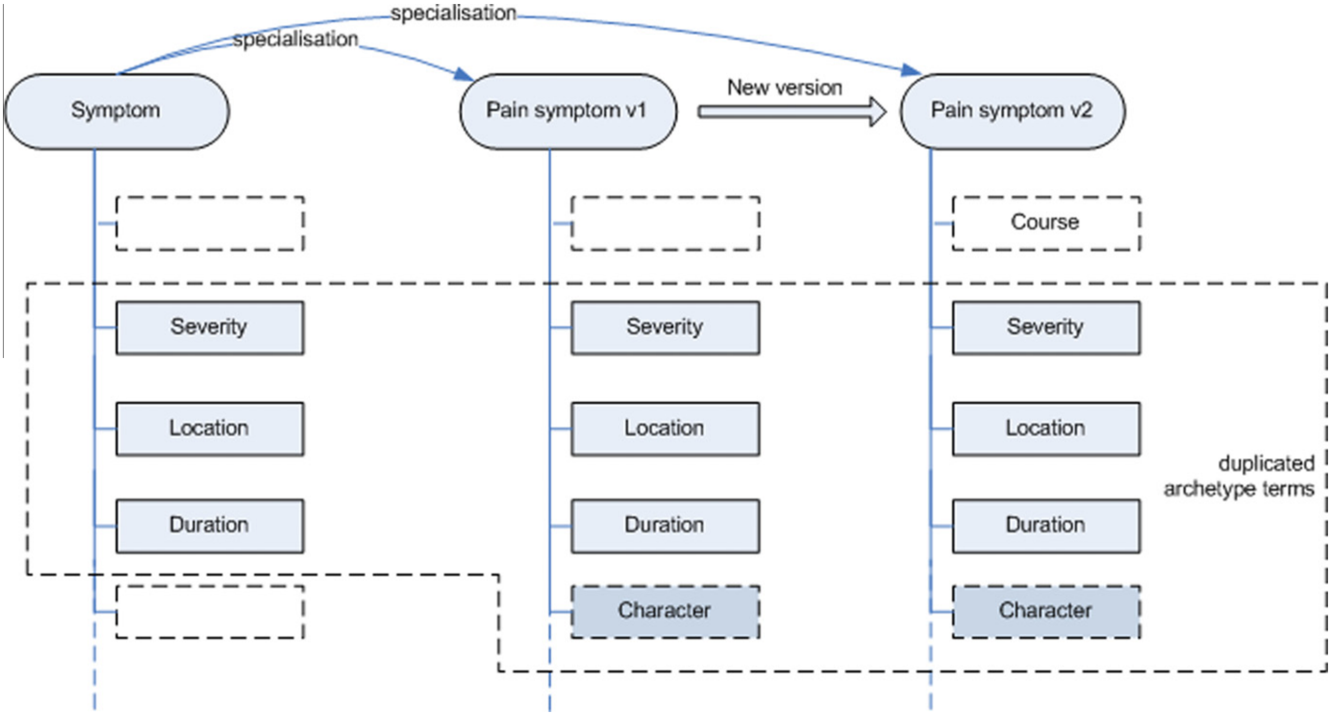

Fig. 2. archetype inheritance and versioning.

3.2. Automatic terminological concepts recognition from free text

Automatic terminological concept recognition from free text is a popular subject of research activity. Ruch et al. [16] focused on processing fragments of medical text to identify SNOMED-CT concepts by combining two search techniques: regular expression matching and a search engine called easyIR. The main source for concept recognition in Ruch's work is the free text that is produced in medical institutions, such as clinical notes or a discharge letters.

Friedman et al. [17] adapted an existing NLP system called *MedLEE* to automate the process of encoding clinical documents with the UMLS thesaurus.

3.3. Archetype and terminology association by using existing thesaurus

Lezcano et al. [18] used a lexical tool from the UMLS thesaurus, which stores pre-mapped arbitrary text and UMLS concepts in an index to associate archetypes with UMLS concepts. The resulting clusters of UMLS concepts were used to build a bipartite graph for the semantic classification of archetypes. Applications of such graph models include the enhancement of archetype browsing and comparisons.

Qamar et al. [19] built a system named MoST to semantically map archetype data to SNOMED-CT codes. The aim of their work was to utilise various technologies such as filtering of the mapping results to aid the mapping process. A prototype of the system was integrated into an archetype editor to enable SNOMED-CT code binding.

The major difference between the related work presented above and this paper is that this research aims to establish the distribution of SNOMED-CT clinical concepts in an archetype repository. The mapping method, outlined in Section 4, does not rely on a pre-mapped thesaurus (see Section 3.1) or (free) medical text processing software. Because it is an algorithmic method, other clinical terminologies could also be used, instead of SNOMED-CT. The above techniques are successful approaches for recommendation of terms, but they do not suit the purpose of this study. The method used here is geared towards the specific purpose of this work, and its performance has been quantitatively assessed [20].

4. Method

This Section describes how to apply the *Shadow* creation technique, which maps arbitrary archetype nodes to SNOMED-CT codes in order to obtain an overview of a repository. *Shadows* contain information on all archetype nodes as well as links to the corresponding SNOMED-CT concepts. Therefore, a quantitative study can be performed to reveal the coverage of clinical content according to the SNOMED-CT space.

The method involves the following steps:

1. Analysing the archetype repository.
2. Extracting text from archetype nodes.
3. Constructing the *Shadow* of each archetype onto SNOMED-CT, thus mapping archetype nodes onto SNOMED-CT codes.
4. Reporting the categories of the resulting codes in SNOMED-CT.
5. Computing the coverage per SNOMED-CT category.

These steps are detailed in the following subsections.

4.1. NHS CfH archetype repository

The National Health Service (NHS) Connecting for Health (CfH) project led to the generation of a repository containing a large number of archetypes, which is as far as the authors are aware, the largest archetype repository available. It covers a wide range of clinical areas and so it is well suited the experiments that are reported in this paper. These archetypes are available in the public domain,[1] and have undergone extensive internal review by expert clinicians prior to being approved for NHS usage as exemplar clinical models [21]. Table 2 lists the number of archetypes per *reference*

[1] http://www.openehr.org/svn/knowledge/archetypes/dev-uk-nhs/.

Table 2
Break-down of types of archetypes in the NHS-CfH repository.

Reference model concept	Number of archetypes based on this concept
Cluster	300
Composition	18
Element	5
Entry	557
Section	96
Structure	99
Total number of archetypes	1075

Table 3
Result of extraction/mapping for the NHS-CfH archetype repository.

Terms type	# Terms found	Percentage
Unique 'AT' codes (terms) extracted from repository	8362	–
Mapped concepts	7925	94.7
Mapped unique concepts	4982	59.6

model type (as referred to in Section 2.1) in this repository.

The NHS CfH repository was initially created with a focus on Accident and Emergency, and maternity care, but it evolved to centralise all modelling efforts that follow the two-level modelling approach. As such, it is not targeted to any specific clinical specialty. On the contrary, the two-level modelling methodology (see Section 2.1) precisely advocates a generic and application independent approach that is equally applicable to any specialty. Accordingly, this archetype repository was populated with additional archetypes as needed by applications that were built using this approach, without any centralised decision to model a given specialty.

The intended users of this repository are domain experts who engage in concept modelling. Due to the current most common use of archetypes, the final users of the applications that embed these artifacts are generally clinicians, recording and exchanging information for medical care delivery.

4.2. Extraction of archetype terms

The pre-processing of archetypes involves taking the latest version of an archetype and resolving inheritance (see Section 2.4) to reduce the redundant nodes that repeat their parent archetypes' base nodes. The resulting output comprises of a list of unique archetype terms that represent the clinical information in the repository as a whole. As shown in Table 3, the extraction process produced 8362 archetype terms.

4.3. Shadows creation process

In our previous work [22,20], a vector-model [23] search tool called Lucene[2] was utilised to achieve automatic mapping of archetype terms onto SNOMED-CT concepts, without using an external thesaurus (a common limitation in popular tools, see Section 3.1, like for example MetaMap). This work also introduced the idea of a terminological *Shadow*. A *Shadow* is an artifact defined in [22] that contains the corresponding SNOMED-CT concepts that a single archetype or set of archetypes may be associated with. It is a result of mapping archetype nodes to their equivalent terminological references, for example SNOMED-CT concepts, to represent the semantics of an archetype in the language of clinical concepts.

One way of visualising the *Shadowing* operation is to 'project' archetype terms onto SNOMED-CT to see which category they belong to, as illustrated in Fig. 3. SNOMED-CT, as a well established and widely accepted terminology, is designed in particular to encode information that appears in a clinical record [1]. Although not complete, the clinical content coverage and expressiveness of SNOMED-CT has been assessed in many studies and it has been shown to be capable of coding up to 80% of information in many clinical scenarios [7,24,25]. Therefore, it is suitable for assessing the clinical content in archetypes.

The algorithm used for the mapping process [20] involves the following steps:

1. Given an archetype, extract the text of all its 'AT' codes from archetypes in the repository, and for each text string, search for associated SNOMED-CT concepts.
2. The search will normally return a list of candidate SNOMED-CT codes. Rank these codes in a list, by placing those that are considered best answers at the top of the list.
3. Select the top answer from the list for each archetype term in an archetype to produce its terminological *Shadow*.

This algorithm was evaluated in [20], which compared the manually defined bindings in some archetypes (recall from Section 2.2 that bindings are optional in archetype definitions), with those extracted automatically by the *Shadows*, on average over the entire NHS CfH repository. The rate for finding the correct category for each archetype term was up to 80% of cases, which is considered acceptable, taking into account the diversity of archetypes.

4.4. Shadowed SNOMED-CT categories

This step takes advantage of the terminological *Shadow* approach, described in Section 4.3. For the NHS CfH repository (see Table 3), of the 8362 archetype terms that were extracted, 7925 mapped directly to SNOMED-CT concepts, thus only 473 (i.e. below 5%) archetype terms were not matched with SNOMED-CT concepts. There are 4982 uniquely mapped SNOMED-CT concepts when duplicates are removed.

The process seeks the first and second level category of every mapped SNOMED-CT concept in its hierarchy. By tracing this hierarchy, the second-level category from the root of the result is stored as the category of the mapped archetype term. For example, if the text string 'Crutches' is mapped to the SNOMED-CT concept 'Crutches (physical object)', its second-level antecedent is *Device(physical object)*.

The reason for choosing second-level categories as the measure of coverage calculations is that due to the granularity of SNOMED-CT, as shown in Table 1, many of the 19 first-level categories are very large and too abstract to be useful in categorising archetypes. The 345 second-level categories that are indicated in Fig. 1, in contrast, while still rather general, are in the authors' view sufficiently specific to represent meaningful classifications of clinical topics corresponding to archetype information.

The mapping process may encounter a so-called *inactive concept*, that is nevertheless suggested as a mapping. This is due to the life cycle of SNOMED-CT concept modelling: certain concepts are marked as *inactive* but are not deleted. The algorithm will then try to replace the inactive concept with the most relevant concept, i.e. second concept on the list returned that is not inactive. Although the mapped SNOMED-CT concepts are the approximation of the original archetype terms, since it is an automated task compared to human operation, their second-level categories can still reveal the covered clinical content of an archetype repository.

[2] http://lucene.apache.org.

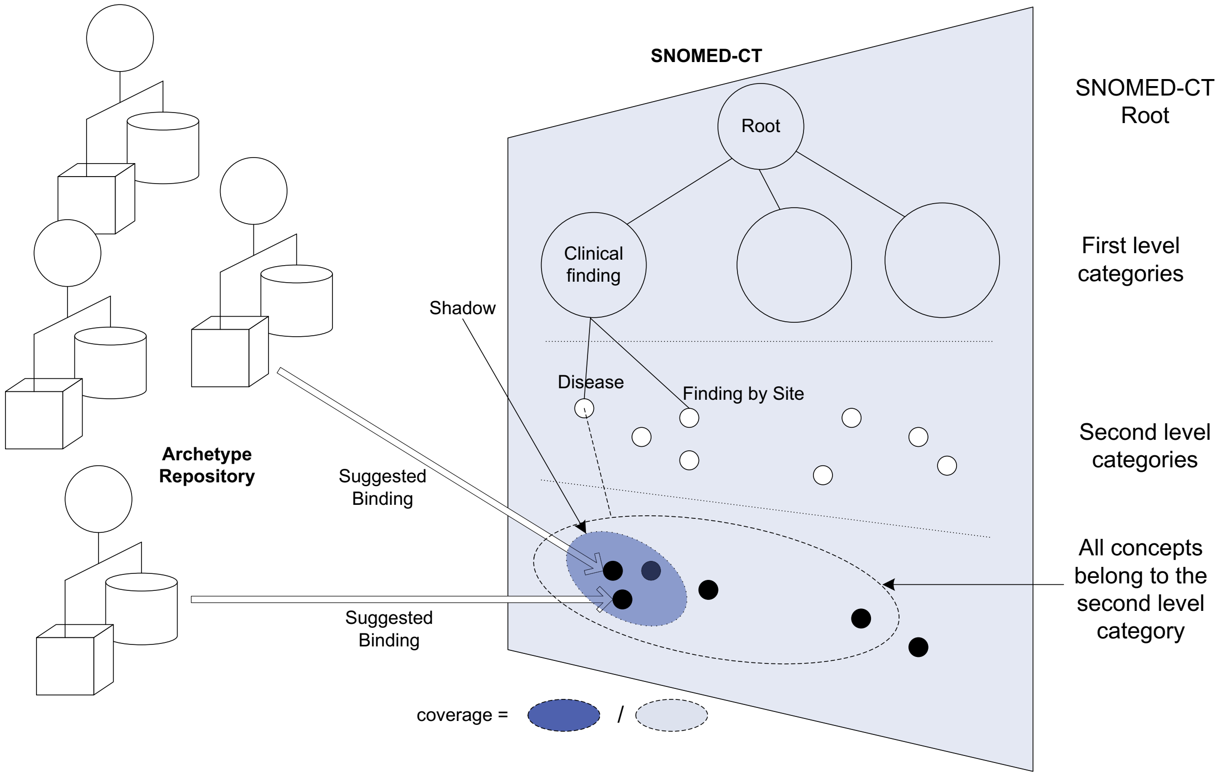

Fig. 3. *Shadow* of archetypes on SNOMED-CT showing the relationships between local terms in archetypes and their counterpart terms in a SNOMED-CT hierarchy.

4.5. Definition of SNOMED-CT coverage

From the data represented by the *Shadows*, the coverage of a category in an archetype repository is defined as the ratio of mapped unique SNOMED-CT concepts found in a category, to the size of that category, as defined by Eq. (1), where:

$$coverage = \frac{mapped\ SNOMED\text{-}CT\ concepts\ in\ T}{Size\ of\ T} \tag{1}$$

- *mapped SNOMED-CT concepts in T* is the total number of uniquely mapped SNOMED-CT concepts belonging to category *T*, and
- *Size of T* is the total number of concepts category *T* contains.

5. Results

This section details the coverage results of the first and second-level SNOMED-CT categories obtained in this study. As a broad summary, the complete set of mapped archetype terms in the repository cover 186 second-level SNOMED-CT categories, out of the 345 available second-level categories.

5.1. First-level overview

Fig. 4 gives an overview of the distribution of first-level categories in the archetype repository. Each section of the chart shows the quantity of mapped concepts among first-level categories. The first number following the name of the category is the number of mapped concepts with duplicates that belong to this category. The second number is the number (percentage) of mapped terms in the current category divided by the total number of successfully mapped concepts, which is 7925 according to Table 3.

From this figure it can be concluded that the majority of the mapped concepts are from *Clinical finding*, which implies that most of the information in archetypes cover this area. This makes sense since, as referred to in Section 4.3, most archetypes in the repository have been created in the context of clinical care. Also, *Clinical finding* is the largest category according to Table 1.

The second most popular category is *Qualifier value*, which makes up one fifth of the total number of mapped concepts. However, referring to Table 1, this category is not among the largest in SNOMED-CT. The smaller but popular categories such as *Qualifier value* will be discussed later. The third largest group of mapped concepts belong to *Procedure* and other prominent categories are *Observable entity* and *Body structure*. Again, this can be explained due to the nature and most common usage of archetypes: communicating information for clinical care.

5.2. Second-level coverage

From the perspective of second-level categories of SNOMED-CT, all 7925 concepts that were mapped during this study are broken down into 186 categories. While there are 345 second-level categories in SNOMED-CT, the ones that are not mapped from any archetype term represent areas that the archetype repository does not cover. These are shown in Table 4 for categories with size greater than 100 concepts.[3] Notably among these, the first-level category *Pharmaceutical/biologic product* is largely missed by this repository, particularly, for example, *14833006 Cardiovascular drug (product)* with a size of 926 terms is the largest SNOMED-CT level-two category that is not covered.

[3] The reason for presenting this subset is to focus on categories that have a relatively large number of concepts, and to investigate their lack of coverage in the repository. Including all categories would make this list unnecessarily long.

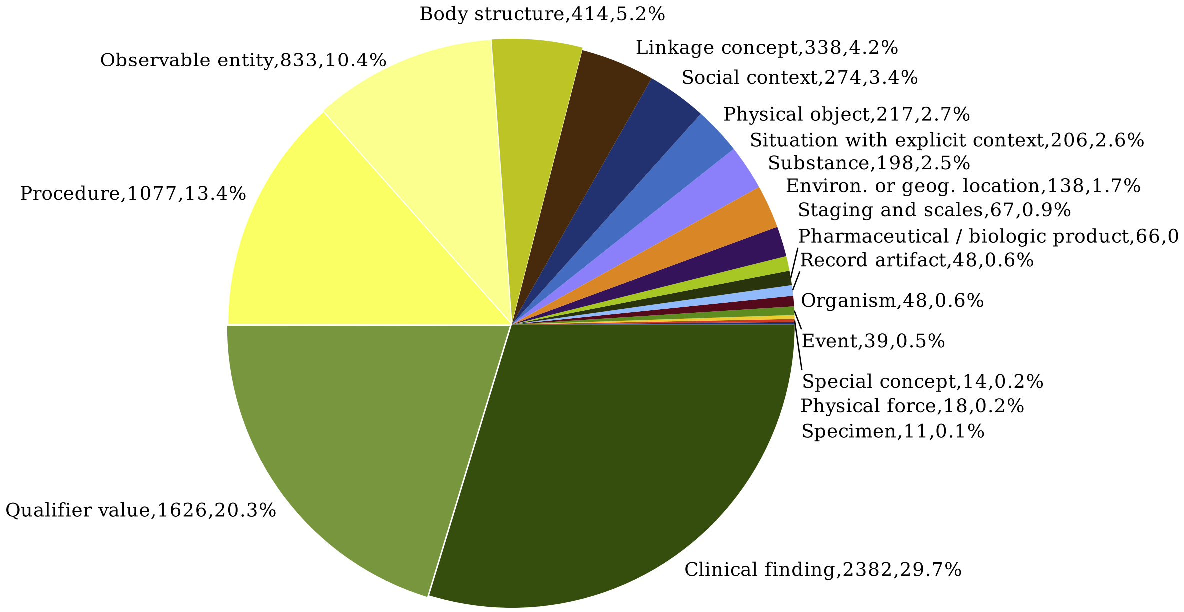

Fig. 4. Break-down of all mapped archetype terms by first level categories.

Table 4
Example SNOMED-CT categories not covered by the NHS CfH repository.

SNOMED-CT first and second level category name		Category size
a. Pharmaceutical/biologic product		
	Cardiovascular drug (product)	926
	Drugs used in the eye (product)	261
	Antidiabetic preparation (product)	249
	Vitamin preparation (product)	244
	Immunotherapeutic agent (product)	240
	Respiratory drugs (product)	229
	Chelating agents and antidotes (product)	163
	Drug groups primarily affecting the musculoskeletal system (product)	160
	Drugs used to treat addiction (product)	118
b. Event		
	Event of undetermined intent (event)	314
	Event related to biological agent (event)	155
c. Substance		
	Oil (substance)	149
d. Organism		
	Life-cycle form (organism)	135
e. Observable entity		
	Sample observable (observable entity)	126
f. Specimen		
	Tissue specimen (specimen)	120

Among the mapped categories, the coverage of some second-level categories is very low. Examples include *Geographical and or political region* and *Radiation therapy observable*, each of which has only one uniquely mapped concept respectively. They represent the minority of mapped second level categories of the archetype repository, in contrast to the most popular category, which is *Clinical history and observation findings*. Table 5 details these results, where:

- *Number of Uniquely Mapped Terms* is the number of uniquely mapped concepts in this category.
- *Category Size* is the total number of concepts under this category.
- *Coverage* is calculated using Eq. (1).

All second-level categories are grouped according to their membership of first-level categories and sorted in descending order of their coverage within each first-level category. Table 5 shows only those categories with a minimum size of 500 concepts.[4]

Based on these results, the coverage of second-level categories can be heuristically divided into three groups in relative terms: well covered, moderately covered and rarely covered. By observing

[4] The reason for presenting a subset only is analogous to that used for Table 4. However, the relative sizes of the different categories either covered or not covered advises the use of a different minimum number of concepts per category from that used in Table 4. This ensures that a sufficiently representative number of categories are shown, without making the list unnecessarily long. The complete results, including smaller-sized categories and their coverage, can be found in the Appendix.

Table 5
Coverage of SNOMED-CT categories with size greater than 500.

SNOMED-CT first and second level category name	#Uniq. mapped SNO. concept	Categ.	Coverage size
a. Body structure			
1. Body structure, altered from its original anatomical structure (morphologic abnormality)	55	4671	1.177
2. Physical anatomical entity (body structure)	229	27,005	0.847
b. Clinical finding			
1. General clinical state finding (finding)	32	642	4.984
2. Administrative statuses (finding)	114	2404	4.742
3. Clinical history and observation findings (finding)	699	16,319	4.283
4. Finding by method (finding)	157	4430	3.544
5. Neurological finding (finding)	61	1904	3.203
6. Finding by site (finding)	432	53,966	0.801
7. Disease (disorder)	174	25,287	0.688
8. Wound finding (finding)	13	2643	0.492
c. Environment or geographical location			
1. Environment (environment)	82	1133	7.237
2. Geographical and/or political region (geographic location)	1	607	0.165
d. Event			
1. Accidental event (event)	9	5106	0.176
2. Exposure to potentially harmful entity (event)	2	2046	0.0978
e. Linkage concept			
1. Attribute (attribute)	141	1127	12.511
f. Observable entity			
1. Function (observable entity)	84	1404	5.983
2. Clinical history/examination observable (observable entity)	205	3937	5.207
3. Feature of entity (observable entity)	22	773	2.846
g. Organism			
1. Trophic life form (organism)	1	503	0.199
2. Kingdom Animalia (organism)	24	13,256	0.181
3. Kingdom Plantae (organism)	3	1935	0.155
4. Pathogenic organism (organism)	1	678	0.147
5. Microorganism (organism)	9	11,413	0.0789
h. Pharmaceutical/biologic product			
1. Veterinary proprietary drug AND/OR biological (product)	18	2533	0.711
2. Replacement preparation (product)	3	576	0.521
3. Analgesic (product)	4	981	0.408
4. Hematologic drug (product)	2	510	0.392
5. Autonomic drug (product)	2	546	0.366
6. Biological agent (product)	2	829	0.241
7. Hormones, synthetic substitutes and antagonists (product)	2	1271	0.157
8. Skin agent (product)	1	698	0.143
9. Gastrointestinal drug (product)	1	778	0.129
10. CNS drug (product)	1	1252	0.0799
11. Diagnostic aid (product)	1	1282	0.078
12. Anti-infective agent (product)	1	1785	0.056
i. Physical object			
1. Device (physical object)	117	3758	3.113
j. Procedure			
1. Regimes and therapies (regime/therapy)	75	1169	6.416
2. Administrative procedure (procedure)	47	1313	3.58
3. Procedure with a procedure focus (procedure)	42	1276	3.292
4. Procedure by intent (procedure)	50	2185	2.288
5. Procedure with a clinical finding focus (procedure)	18	1080	1.667
6. Procedure by method (procedure)	220	22,223	0.99
7. Procedure by device (procedure)	38	4619	0.823
8. Laboratory procedure (procedure)	67	8677	0.772
9. Procedure by site (procedure)	61	10,093	0.604
k. Qualifier value			
1. Descriptor (qualifier value)	256	1579	16.213
2. Spatial and relational concepts (qualifier value)	59	1075	5.488
3. Intellectual concepts and systems (qualifier value)	29	660	4.394
4. Unit (qualifier value)	40	1137	3.518
5. Ranked categories (qualifier value)	30	964	3.112
l. Situation with explicit context			
1. Procedure with explicit context (situation)	50	1113	4.492

(*continued on next page*)

Table 5 (*continued*)

SNOMED-CT first and second level category name	#Uniq. mapped SNO. concept	Categ.	Coverage size
2. Past history of (situation)	20	672	2.976
3. [V]Factors influencing health status and contact with health services (situation)	33	1237	2.668
4. Finding with explicit context (situation)	39	1813	2.151
m. Social context			
1. Occupation (occupation)	82	4395	1.866
n. Staging and scales			
1. Assessment scales (assessment scale)	52	884	5.882
o. Substance			
1. Dietary substance (substance)	26	2000	1.3
2. Drug or medicament (substance)	10	1669	0.599
3. Substance categorised functionally (substance)	10	1742	0.574
4. Materials (substance)	8	1420	0.563
5. Allergen class (substance)	8	1708	0.468
6. Substance categorised structurally (substance)	47	11,871	0.396
7. Biological substance (substance)	6	2304	0.26

the spread of results, it is apparent that the coverage values for different categories, vary from 0.1% to 5.0%.

The boundary between the well and moderately covered categories could be placed in relative terms, at those categories whose coverage is around 1.0%. The boundary between the moderately and rarely covered is at 0.1%. These parameters have been fine tuned through heuristic experience and experimental evaluation during the course of this work. By using these values, it is easy to visually identify categories at different points on the coverage spectrum. The following are examples of identified categories with outstanding coverage. The indicator in the curly brackets points to the location of the item in Table 5:

1. Well covered categories
   (a) Clinical history and observation findings (finding) {b.3}
   (b) Administrative statuses (finding) {b.2}
   (c) Finding by method (finding) {b.4}
   (d) Clinical history/examination observable (observable entity) {f.2}
   (e) Descriptor (qualifier value) {k.1}
   (f) Unit (qualifier value) {k.4}
   (g) Attribute (attribute) {e.1}
   (h) Environment (environment) {c.1}
2. Rarely covered categories
   (a) Microorganism (organism) {g.5}
   (b) Kingdom Animalia (organism) {g.2}
   (c) Anti-infective agent (product) {h.12}
   (d) Diagnostic aid (product) {h.11}
   (e) CNS drug (product) {h.10}

Among highlighted well covered categories, category *b.3* has a relatively large size, of 16,319 concepts, with a coverage of 4.28%. The results also show that certain relatively small categories, such as *k.1*, *k.4* and *e.1*, tend to be well covered.

Notably in the SNOMED-CT model, concepts under category *e.1 Attribute* have special usage that allow users to compose refined new concepts. Complex clinical statements can be created by using attribute concepts to link existing concepts. This mechanism, known as *post-coordination* [6], relies on linkage concepts like *Attribute* to refine an existing concept and expand it by adding other concepts. Category k.1 *Descriptor (qualifier value)* is well covered and it also has a special role in the SNOMED-CT model. For all second-level categories under *Qualifier value*, they can be used as qualifiers in post-coordination to refine an existing concept. From these results, the archetype repository appears to also cover these categories relatively well.

Regarding rarely covered categories, it is clear that *h.12*, *h.11*, *h.10* are all product related. Category *g.5 Microorganism* is very poorly covered and this could be of interest to archetype modellers.

5.3. Frequency of term occurrence – StopWords

While extracting archetype terms from all the archetypes in the repository, manual observation shows that certain ambiguous terms exist such as *Location* and *Result*. Table 6 lists the most frequently used terms. These terms appear to play significant roles in general archetype semantics, but not in a medical sense. In the authors' view, this is analogous to "stop words" such as 'and', 'of' in natural language processing. This material can be beneficial to separating terms of this kind from less ambiguous terms. In our experiment, however, these terms are not excluded from the mapping process because it is important to understand how well they are covered by SNOMED-CT.

6. Discussion

The results of the first-level category overview and second-level category coverage provide researchers who are interested in either in both, archetypes and SNOMED-CT, an opportunity to identify areas of interest in a number of ways. From an archetype developer's viewpoint, results such as those given in Table 5 offer a mechanism to determine the status of an archetype repository. The terminological *Shadow* that is exploited in this work can provide a snapshot of how much clinical content of SNOMED-CT has been covered within a particular repository. As the development

Table 6
Most frequently used archetype terms.

Archetype term	Frequency	Archetype term	Frequency
Tree	171	Normal statements	27
Any event	133	None	27
Event series	113	Findings	27
no text for this at-code	77	Name	25
Comment	53	Clinical description	25
Other	48	Normal statement	24
Description	43	Right	23
List	42	Procedure outcome	23
Comments	40	Procedure comments	22
Details	39	Left	22
History	32	Person name performing procedure	20
Normal	29		

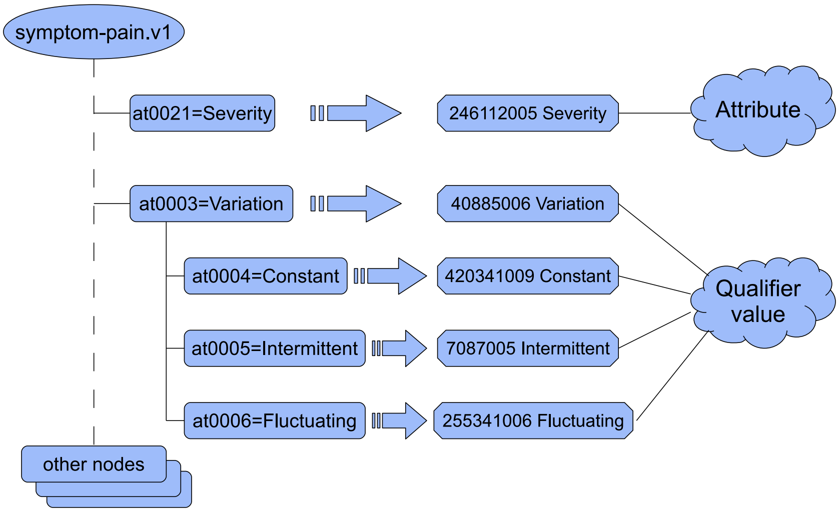

Fig. 5. Example of heavily used concepts.

of an archetype repository progresses, it can provide constant monitoring of medical topics that are being modelled to prevent redundant information or unbalanced development. This technique was only applied to archetypes in this case, but it could easily be extended to co-called *templates*[5] when they become available in sufficient numbers. The approach presented here is sufficiently flexible to be applied to any archetype repository. The threshold for determining when a category is well or rarely covered can be adjusted to particular use cases. For this study the upper limit of coverage for certain categories is as high as 7%. But there is as yet no way to determine a satisfactory threshold for a SNOMED-CT category. For example, a national project's archetype repository should have higher coverage than a clinic's repository. In any case, the idea of an acceptable level of coverage for a whole repository is also still undetermined.

For SNOMED-CT modellers, this approach provides feedback from application of the SNOMED-CT concepts in archetype modelling. It reveals the focused areas of medical concepts that are being modelled in a well-known EHR model. The popular categories reflect the overlapping areas where two communities have a common focus, while the under-covered categories may indicate differences in modelling priorities. Guidance on SNOMED-CT concept usage can be provided where coverage is not satisfactory. *Microorganism*, as an example, exhibits much lower coverage than other categories, therefore investigation may reveal why so few archetypes use concepts from this large category. Medical professionals in specific areas can also pay attention to their fields. For example, *Kingdom animalia* is a category with reasonably large size but is little covered in this repository. However, archetypes may be demanded by practitioners of veterinary medicine that require usage of many concepts from this category.

Finally, the results of this study are a resource for archetype creation planning. For instance, there are 11,413 types of *Microorganism* in the release of SNOMED-CT used here. When archetype modellers decide to create archetypes and templates for microorganisms, the results show that there is likely to be a requirement for either a query based binding mechanism or else a relatively large quantity of bindings even to achieve a coverage of 1% of the total number of terms in this category. A similar analysis can be performed on many of the first-level and second level categories with similar results. Archetype authors may need to incorporate specific queries, large hierarchical archetypes or even more sophisticated mechanisms to make these concepts available. Assuming that the archetype modellers' goal is to cover as much clinical content as possible to be used with EHRs, the results of the coverage of current archetypes can indicate the likely complexity of future development and implementation.

Based on the coverage results, the areas identified as well and rarely covered also lead to discussion and comparison of the two health modelling patterns. The Archetype Object Model is record-oriented, which means its model resembles the structure of document records. In contrast, SNOMED-CT is ontology-oriented, which means it represents a network of medical phenomena that is based on logic. However, similarities can be found between the two modelling approaches because the patterns of organising information show some correspondence at certain levels.

The fact that categories such as *Clinical finding* are well covered in the studied repository could be taken as evidence that the archetypes in the repository mainly focus on clinical finding information (e.g. for clinical care applications). The results also show that the proportion of archetype terms mapped to the *Qualifier value* category is significantly higher than other categories regardless of its small size. *Attribute*, has similarly been well covered despite its relatively small size. One possible reason for this coverage information is that many parts of archetypes use these concepts frequently, for qualifying answers such as "Mild", "Nil significant" or attributes like "Severity". Fig. 5 gives one example from the *openEHR-EHR-CLUSTER.symptom-pain.v1* archetype [26], to show how these concepts are linked to archetypes.

The limitation associated with this approach is that the automatic mapping process has limited precision when compared to the judgement of human experts. However, in the authors' opinion, the development and improvement of a mapping algorithm should be kept as a separate work, and their performance should preferably be evaluated against human judgement for this task.

7. Future work

Future work will investigate how to derive and use additional contextual information that can enhance the effectiveness of the *Shadow* approach for nodes in certain parts of an archetype. For instance, at the leaf level of an archetype, one is likely to find constraints on an element of a data type to store clinical data. In many cases this takes the form of a standard coded medical term or a condition. The most frequently occurring group of codes in the category

[5] openEHR specifications also include the idea of a template that allows archetypes to be combined and further constrained to suit particular scenario that may arise in healthcare. These templates also allow bindings to terminology systems.

Clinical history and observation findings are examples of SNOMED-CT codes that would comfortably map into this part of an archetype. The majority of SNOMED-CT codes are candidates for this type of mapping. On the other hand, the 'verbs' or 'link' concepts in the SNOMED-CT model that connect core concepts and their potential modifiers are called *linkage concepts*. The highly utilised *Attribute* category has the potential to relate closely to archetype organisational elements according to its coverage information. Another possible context information could include whether a searched term was a member of the *Person* category, which is associated with demographic model in the Archetype Model.

The authors see a potential for terminological *Shadows* of archetypes to be used to aid better integration between the EHR information model and clinical terminology concept models such those found in SNOMED-CT. Therefore in the future the authors will incorporate investigation of the relationship between the meta information of archetype terms and the mapped equivalent SNOMED-CT concepts. This meta information may include the data types of the particular EHR information model that the archetype is constraining, the intended applicability of SNOMED-CT concepts in its concept model. It is intended to improve the mapping of appropriate SNOMED-CT concepts according to the context of the archetype term.

Further research will also utilise *Shadows* to compare individual archetypes to identify similarity among heterogeneous archetype designs [27]. Archetypes could be compared by matching their *Shadows* and analysing the SNOMED-CT concepts in a normalised form [28]. The result of this process can provide a better automated means to analyse archetypes that are created in different development backgrounds that may contain overlapping information for the same clinical setting that is presented in different ways.

8. Conclusion

This paper applied a previously published method [22,20], referred to as the terminological *Shadow* approach, that automatically maps archetype terms to SNOMED-CT concepts to generate the overview of SNOMED-CT concept coverage in an archetype repository. This paper effectively identified a number of under-covered and well-covered SNOMED-CT categories in the chosen archetype repository. The authors believe, as a result of their experiences, that the *Shadow* approach can be used more generally to identify the coverage of SNOMED-CT concepts, and potentially, other terminology systems among clusters of archetypes that are being created.

This contribution is applicable to the management of large archetype repositories and it will guide archetype developers to pay attention to the relative size of equivalent terminological categories when assessing the effort required in creating terminological bindings for particular archetypes. For example, if a seemingly significant SNOMED-CT category, such as *Microorganism (organism)*, is not covered well but is used extensively in a particular clinical scenario, developers in that area must consider enhancing their archetype set in order to cover more concepts in this category. One possible, simple, solution is to embed queries to link one archetype term to multiple external references [29].

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.12.001.

References

[1] Bisbal J, Berry D. An analysis framework for electronic health record systems: Interoperation and collaboration in shared healthcare. Methods Inf Med 2011;50(2):180–9.

[2] Schloeffel P, Beale T, Hayworth G., Heard S., Leslie H. The relationship between CEN 13606, HL7, and openEHR. In: Proceedings of HIC 2006 bridging the digital divide: clinician, consumer and computer; 2006. p. 24–8.

[3] Eichelberg M, Aden T, Riesmeier J, Dogac A, Laleci G. A survey and analysis of electronic healthcare record standards. ACM Comput Surveys 2005;37(4):277–315.

[4] Beale T, Heard S. The openEHR archetype model: archetype object model; 2007. Last visited: October 2011. <http://www.openehr.org/releases/1.0.1/architecture/am/aom.pdf>.

[5] Garde S, Hovenga E, Granz J, Foozonkhah S, Heard S. Towards a repository for managing archetypes for electronic health records. In: Proceedgins of HIC 2006 and HINZ 2006, Health Informatics Society of Australia; 2006. p. 61–5.

[6] The International Health Terminology Standards Development Organisation. SNOMED Clinical Terms User Guide; 2008. Last visited: October 2011. <http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Technical_Docs/SNOMED_CT_User_Guide_20080731.pdf>.

[7] Sampalli T, Shepherd M, Duffy J, Fox R. An evaluation of SNOMED-CT in the domain of complex chronic conditions. Int J Integr Care 2010;10:e038.

[8] Wasserman H., Wang J. An applied evaluation of SNOMED-CT as a clinical vocabulary for the computerized diagnosis and problem list. In: Proceedings of the Medical Informatics Association (AMIA) Annual Symposium; 2003. p. 699–703.

[9] Sundvall E, Qamar R, Nyström M, Forss M, Petersson H, Karlsson D, et al. Integration of tools for binding archetypes to SNOMED-CT. BMC Med Inform Decis Mak 2008;8(Suppl. 1):1–10.

[10] Garde S, Hovenga E, Buck J, Knaup P. Expressing clinical data sets with openEHR archetypes: a solid basis for ubiquitous computing. Int J Med Inform 2007;76(Suppl. 3):S334–41.

[11] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49:624–33.

[12] Vreeman D, McDonald C. Automated mapping of local radiology terms to LOINC. In: Proceedings of the medical informatics association (AMIA) annual symposium; 2005. p. 769–73.

[13] Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the medical informatics association (AMIA) annual symposium; 2001. p. 17–21.

[14] Aronson A, Lang F. An overview of MetaMap: historical perspective and recent advances. J Amer Med Inform Assoc 2010;17(3):229–36.

[15] Bhatia N, Shah N, Rubin D, Chiang A, Musen M. Comparing concept recognizers for ontology-based indexing: MGREP vs. MetaMap. In: Proceedings of the medical informatics association (AMIA) annual symposium; 2009.

[16] Ruch P, Gobeill J, Lovis C, Geissbuhler A. Automatic medical encoding with SNOMED categories. BMC Med Inform Decis Mak 2008;8(Suppl. 1):S6.

[17] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Amer Med Inform Assoc 2004;11(5):392–402.

[18] Lezcano L, Sánchez-Alonso S, Sicilia MA. Associating clinical archetypes through UMLS metathesaurus term clusters. J Med Syst 2010:1–10.

[19] Qamar R, Rector A. MoST: A system to semantically map clinical model data to SNOMED-CT. In: Proceedings of semantic mining conference on SNOMED-CT; 2006. p. 38–43.

[20] Yu S, Berry D, Bisbal J. Performance analysis and assessment of a TF-IDF based archetype-SNOMED-CT binding algorithm. In: Proceedings 24th IEEE international symposium on computer-based medical systems; 2011. p. 1–6.

[21] Leslie H. openEHR – the world's record. PulseIT 2007:50–5.

[22] Yu S, Berry D, Bisbal J. An investigation of semantic links to archetypes in an external clinical terminology through the construction of terminological 'Shadows'. In: Proceedings of international association for development of the information society e-health; 2010. p. 9–17, ISBN: 9789728939168.

[23] Baeza-Yates RA, Ribeiro-Neto B. Modern information retrieval. Addison-Wesley; 1999.

[24] Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. J Amer Med Inform Assoc 1996;3(3):224–33.

[25] Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED-CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006;81(6):741–8.

[26] United Kingdom National Health Service (NHS) Connecting for Health program. openehr-ehr-cluster.symptom-pain.v1; 2008. Last visited: October 2011. <http://www.openehr.org/svn/knowledge/archetypes/dev-uk-nhs/adl/openehr/ehr/cluster/openEHR-EHR-CLUSTER.symptom-pain.v1.adl>.

[27] Bisbal J, Berry D. Archtype alignment: a two-level driven semantic matching approach to interoperability in the clinical domain. In: Proceedings of the international conference on health informatics, HEALTHINF 2009. INSTICC Press; 2009. p. 216–21.

[28] Andrews J, Patrick T, Richesson R, Brown H, Krischer J. Comparing heterogeneous SNOMED-CT coding of clinical research concepts by examining normalized expressions. J Biomed Inform 2008;41(6):1062–9.

[29] Beale T., Heard S. The openEHR Archetype Model: Archetype Definition Language ADL 1.4. 2007. Last visited: October 2011; URL http://www.openehr.org/releases/1.0.1/architecture/am/adl.pdf.