INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 80 (2011) 205–226

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Applying the archetype approach to the database of a biobank information management system

Melanie Bettina Späth, Jane Grimson*

Centre for Health Informatics, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland

ARTICLE INFO

Article history:
Received 26 July 2010
Received in revised form
1 November 2010
Accepted 2 November 2010

Keywords:
Biological Specimen Banks
Biobanks
Electronic Health Record
*open*EHR archetypes and templates
Biobank information management system

ABSTRACT

Purpose: The purpose of this study is to investigate the feasibility of applying the *open*EHR archetype approach to modelling the data in the database of an existing proprietary biobank information management system. A biobank information management system stores the clinical/phenotypic data of the sample donor and sample related information. The clinical/phenotypic data is potentially sourced from the donor's electronic health record (EHR). The study evaluates the reuse of *open*EHR archetypes that have been developed for the creation of an interoperable EHR in the context of biobanking, and proposes a new set of archetypes specifically for biobanks. The ultimate goal of the research is the development of an interoperable electronic biomedical research record (eBMRR) to support biomedical knowledge discovery.

Methods: The database of the prostate cancer biobank of the Irish Prostate Cancer Research Consortium (PCRC), which supports the identification of novel biomarkers for prostate cancer, was taken as the basis for the modelling effort. First the database schema of the biobank was analyzed and reorganized into archetype-friendly concepts. Then, archetype repositories were searched for matching archetypes. Some existing archetypes were reused without change, some were modified or specialized, and new archetypes were developed where needed. The fields of the biobank database schema were then mapped to the elements in the archetypes. Finally, the archetypes were arranged into templates specifically to meet the requirements of the PCRC biobank.

Results: A set of 47 archetypes was found to cover all the concepts used in the biobank. Of these, 29 (62%) were reused without change, 6 were modified and/or extended, 1 was specialized, and 11 were newly defined. These archetypes were arranged into 8 templates specifically required for this biobank. A number of issues were encountered in this research. Some arose from the immaturity of the archetype approach, such as immature modelling support tools, difficulties in defining high-quality archetypes and the problem of overlapping archetypes. In addition, the identification of suitable existing archetypes was time-consuming and many semantic conflicts were encountered during the process of mapping the PCRC BIMS database to existing archetypes. These include differences in the granularity of documentation, in metadata-level versus data-level modelling, in terminologies and vocabularies used, and in the amount of structure imposed on the information to be recorded. Furthermore, the current way of modelling the sample entity was found to be cumbersome in the sample-centric activity of biobanking.

* *Corresponding author*. Tel.: +353 1 896 3466.
E-mail address: spaethm@tcd.ie (M.B. Späth).

doi:10.1016/j.ijmedinf.2010.11.002

The archetype approach is a promising approach to create a shareable eBMRR based on the study participant/donor for biobanks. Many archetypes originally developed for the EHR domain can be reused to model the clinical/phenotypic and sample information in the biobank context, which validates the genericity of these archetypes and their potential for reuse in the context of biomedical research. However, finding suitable archetypes in the repositories and establishing an exact mapping between the fields in the PCRC BIMS database and the elements of existing archetypes that have been designed for clinical practice can be challenging and time-consuming and involves resolving many common system integration conflicts. These may be attributable to differences in the requirements for information documentation between clinical practice and biobanking. This research also recognized the need for better support tools, modelling guidelines and best practice rules and reconfirmed the need for better domain knowledge governance. Furthermore, the authors propose that the establishment of an independent sample record with the sample as record subject should be investigated. The research presented in this paper is limited by the fact that the new archetypes developed during this research are based on a single biobank instance. These new archetypes may not be complete, representing only those subsets of items required by this particular database. Nevertheless, this exercise exposes some of the gaps that exist in the archetype modelling landscape and highlights the concepts that need to be modelled with archetypes to enable the development of an eBMRR.

1. Introduction

The primary purpose of electronic health records (EHRs) is to provide a documented record of care to support present and future healthcare of a subject of care [1]. However, one of the major advantages of EHRs is that the data that is being recorded as part of healthcare delivery presents a valuable source of information that can be reused for a number of secondary purposes, including for scientific research and candidate selection for clinical trials [1,2], whose results feedback to improve healthcare.

To enable biomedical knowledge discovery, such as the investigation of the fundamental mechanisms of complex diseases, researchers need to combine clinical patient data, such as medical history and lifestyle data found in the patient's health records with the results from molecular experiments so that correlations can be drawn about the influence of genes and/or the environment on disease pathology [3–7].

Biobanks, or bio-repositories, play a central role in combining these two streams of information [8]. Biobanks collect, store and distribute biological specimens, such as blood, urine and tissue and associated patient data, such as clinical history and lifestyle information. The roots of biobanking can be found in clinical pathology, but biobanking today is a young industry, which is evolving into a separate research area with many specialized components and dedicated personnel [9,10]. Biobanks differ in size, ranging from small disease-specific collections of biospecimens to large population-based biobanks. They also differ according to their purpose; for example, those that mainly support clinical healthcare, such as pathology archives for medical diagnosis, or those that have been set up primarily for research purposes. And finally, different biobanks collect different types of biological material [9,11]. Recent advances in biotechnology, such as the emergence of high-throughput technologies, have increased the demand for high-quality, well-annotated human biospecimens in biomedical research [3,12–16]. As a consequence, biobanking activity is increasing and new biobanks are being created all over the world, often focusing on specific diseases, resulting in a large number of small sample collections [11].

The data in biobanks is managed by a Biobank Information Management System (BIMS). The BIMS stores the clinical background information of the patient/donor, such as the disease, treatment and patient outcome and information about the samples, e.g. the composition of the sample, and sample handling and administrative information [9]. Clinical information that is pertinent to the research being carried out is generally manually extracted and imported into the BIMS from the patient's health record and/or from questionnaires and/or interviews with the patient/donor.

Similar to early EHR implementations, current BIMS solutions tend to be bespoke disease- and study-specific proprietary implementations of varying sophistication (e.g. [6,7]), reflecting the heterogeneity of biobanks. Thus, major resources and effort are being invested in setting up and populating a new BIMS every time a new study is initiated or a new biobank is established. Furthermore, there is an increased need for biobanks to collaborate and share samples and information, especially in the case of studies concerning rare diseases to ensure a sufficiently large population cohort to be statistically significant [3,4,6,11,17–24]. Indeed, the lack of a sufficient number of biospecimens restricts the amount of translational research that can be done [25], such that the pace of scientific advance cannot be matched with its exploitation in medical research [9].

However, as with many existing EHR implementations, due to the heterogeneity of the systems and underlying databases in biobanks, information cannot easily be shared between collaborating biobanks [11], thus restricting the scope and scale of research that can be carried out [3,26].

In addition, there is often no provision to feed back research results into the BIMS, so that they are kept scattered over individual researchers' computers, often lost when the researcher leaves the institute, and thus unavailable for further analysis. However, more and more researchers believe that there is value in storing this research data, together with the clinical data, for future research and analyses. In fact, some researchers believe that we are now on the verge of being able to combine the results of molecular analyses retrieved with high-throughput genotyping technologies with life-style and demographic data on a large scale [7], requiring the storage of today's research results for tomorrow.

It was thought that a possible solution to this problem could be derived from current EHR solutions, due to the overlapping domain and the similarity of problems they need to solve and requirements they need to fulfil. Both need to store and communicate clinical information. In the domain of EHRs, researchers are currently tackling the problem of storing and communicating patient information in a safe and reliable way to support healthcare delivery through the development of EHR standards. Several major standards development organizations, such as CEN, ISO and HL7, have taken on this challenge resulting in the development of a number of different approaches. One of these approaches, promising to enable the development of future-proof, semantically interoperable EHRs, is the archetype approach, based on a novel two-level modelling methodology pioneered by *open*EHR[1] [27–30]. The CEN standard, EN 13606, which has later also been approved as an ISO standard, is based on this approach.

The first level constitutes the information level, modelling the semantics of information, such as generic data types and structures in a small and stable Reference Model [28,31]. It specifies the global and stable characteristics of the components of the health record, how they can be aggregated/composed and what kind of information needs to be provided to meet requirements of legal, ethical and provenance nature. These components are defined in the Reference Model as a set of classes that form the generic information building blocks of the EHR [32]. The Reference Model is hierarchically organized with the EHR EXTRACT acting as root of the aggregation hierarchy and acts as the top-level container of the EHR of a single subject of care. The EHR EXTRACT contains COMPOSITIONs that can optionally be organized in a FOLDER hierarchy. A COMPOSITION is the unit of committal and represents a single encounter or record documentation session. FOLDERs are used to organize COMPOSITIONs in the EHR, e.g. by episode of care. Next, COMPOSITIONs contain ENTRYs that may optionally be organized in a SECTION hierarchy. The actual clinical information, such as clinical statements, is recorded in the ENTRYs. SECTIONs represent clinical headings and are used to organize ENTRYs within a COMPOSITION. The *open*EHR specification defines several ENTRY subtypes, such as ACTION, OBSERVATION, and EVALUATION. Finally, ENTRYs contain ELEMENTs, which can optionally be organized within a CLUSTER hierarchy as a means to organize multi-part data and to represent columns of a table [31,33]. Elements are thus the leaf nodes of the EHR EXTRACT hierarchy and contain the actual Data Values, while the other levels in the EHR EXTRACT hierarchy provide the context of the individual clinical data points, i.e. they position the data into its clinical context.

The second level is the knowledge level and constitutes formal definitions of clinical content in the form of archetypes and templates [28,29]. Archetypes are used to model domain knowledge, which may be volatile and evolving, e.g. clinical concepts such as "laboratory report" or "blood pressure measurement". Archetypes are defined as "constraint-based domain models" or "constraint-based concept definitions" that configure and constrain valid combinations of the information building blocks defined in the Reference Model to formally represent distinct and complete domain-level concepts with regard to their naming, attribute information, optionality and multiplicity [27,30,32]. Since an archetype modelling a specific domain concept aims to be complete, i.e. it aims to cover all possible variations of the concept, it should be reusable, thus reducing development effort [28]. Archetypes can also be composed of lower level archetypes to form other archetypes using a facility called archetype slots [27,28,30,34]. Furthermore, archetypes can be specialized from general concepts into more specific concepts [27].

In practice archetypes are ultimately deployed through templates that combine several archetypes to represent screen forms, documents, printed reports or messages, such as "discharge summary" [29,30] and are usually developed locally to customize archetypes to the local context [35]. For example, default values can be set at a template level and optional archetype items can be deleted or mandated depending on local usage requirements. Data entry and validation usually happens on the template level. These templates "soften the rough edges" inherent in using a small information reference model by adapting archetypes to the situation at hand.

Only the first level, i.e. the Reference Model, is concretely implemented in software objects and database schemas, whereas archetypes and templates are used at runtime by the system, where their primary purpose is to validate data during data capture, so that data that is finally committed to the data store conforms to the archetype definitions [29,30]. It is these characteristics that the creators of the two-level modelling approach claim make the EHR flexible to changes in the domain and thus future-proof [27–29].

Archetype development is an ongoing activity within *open*EHR and *open*EHR also provide an online archetype repository of developed archetypes.[2] The UK National Health Service (NHS) is also active in archetype development and has created its own online archetype repository [36,37].

In this paper, the authors propose the development of a generic electronic biomedical research record (eBMRR), using

[1] The *open*EHR Foundation is an international not-for-profit foundation working towards realizing the interoperable, life-long electronic health record. The foundation's activities include the development of open specifications, open-source software and knowledge resources, and contributing to international standards development and clinical implementation projects. More information about the *open*EHR Foundation and its activities can be found at http://www.openehr.org.

[2] Clinical Knowledge Manager, *open*EHR archetype repository, http://www.openehr.org/knowledge/.

the *open*EHR archetype methodology, to support biomedical knowledge discovery, in the same way as the *open*EHR methodology is used to support the construction of a generic electronic health record to support patient care. This eBMRR can integrate both clinical and research information and can be customized to a specific biobank or study. Such a research record would, in the short term, reduce development effort whenever a new biobank or study is set up and alleviate the problem of interoperability between collaborating biobanks. At the same time, research information could be stored in a way that it can be made available and used for future research.

The information in the proprietary database of the prostate cancer biobank of the Irish Prostate Cancer Research Consortium (PCRC) used in this research is modelled with *open*EHR archetypes and templates. The PCRC is a multi-disciplinary, trans-institutional collaboration involving several Dublin-based research institutions and hospitals. The main research aims of the PCRC are the discovery and validation of novel prostate cancer biomarkers for disease diagnosis, prognosis and treatment success and to understand the molecular basis of the disease and its recurrence, as well as the development of new therapies.[3] Central to the PCRC is the longitudinal collection of biospecimens of prostate cancer patients. The PCRC collects samples of prostate tissue, urine and blood from several hospitals. These samples are further processed, e.g. tissue is preserved and/or embedded, urine is treated with a protease inhibitor and blood is separated into plasma and serum, following standard operating procedures, before being stored in the biobank's freezers. The bio-resource currently holds over 350 tissue samples with matched serum and plasma samples as well as over 90 urine samples and includes comprehensive clinical information about the patients and their disease progression. For their biobank information management system, the PCRC uses the product Distiller developed by the Slidepath software company,[4] which provides a PHP point-and-click interface that enables non-technical users to create their own database table definitions without having to know any database query languages or the underlying relational database implementation.

2. Methods

The schema of the Irish Prostate Cancer Research Consortium BIMS database was first analyzed with regard to the structure, concepts and content. It was important to determine the exact meaning and context of each field, so that an exact correspondence between an archetype item and the database field could be established. The research nurses who populate the database and who have been involved in the design of the database schema had only just started in documenting this information, so that much of this information was retrieved through analysis of available meta-information in form of table names and fields, field contents, consultation of available literature on practices and artefacts of the medical domain in general and prostate cancer in particular. Interviews with the research nurses clarified the meaning of the outstanding data items and fed back into the nurses' documentation. It was necessary to return to the process of information refinement throughout the modelling process to make sure that the mappings between the items in the archetypes and the fields in the BIMS database schema were correct.

Due to the proprietary nature of the PCRC BIMS it was not possible to obtain the underlying relational database schema. However, since the tables that are created by the proprietary software in the underlying relational database do not correspond directly to the table definitions that the users created by using the PHP interface, a conceptual database schema that reflects the table structures as created and seen by members of the PCRC in the PHP interface was used. This conceptual database schema contains 18 interlinked, conceptually hierarchical tables, with 146 data fields in total, the organization of which can be seen in seen in Fig. 1a–d. It turned out that each table (apart from the root "Patient" table) contains an identity field that was not being used, thus reducing the number of information-bearing fields to 129. With regard to the content, the database contains:

- clinically relevant patient background information, such as past and present medical history,
- information detailing current medication, family history and current symptoms,
- information about patient treatment options and outcomes,
- information about specimen collection such as prostate tissue, urine and blood,
- specimen processing information,
- results of histo-pathological tissue analysis,
- and the results of biochemical tests that have been carried out on the specimen as part of the diagnosis and which are documented in the patient's health record.

It should be noted that the results of more sophisticated analyses for research purposes, such as omic experiments are not currently stored in the PCRC BIMS database, but rather on the individual researchers' computers and thus the current modelling exercise only concerns clinical and sample information. Of the 18 tables, 14 contain information that is mainly related to the sample, whereas the remaining 4 tables contain mainly clinical information. The database contained content for over 500 patients. The data types of the content were text, Boolean, numbers, e.g. float or integer, or date/time. Some of the text fields contained either unstructured text (e.g. for medical history and current symptoms), or the result of choosing a text item from a list (e.g. Caucasian, Asian etc. for "Ethnicity"). Many fields were left empty and there was no instance where data was entered in all available fields for one patient.

To consider how this conceptual database may be modelled with archetypes one has to bear in mind that each particular archetype models only one distinct concept [35]. An analysis of the conceptual BIMS database schema showed that often the tables in the BIMS database contain a number of distinct concepts. For example, the Biopsy Table contains information about the conditions surrounding sample procurement, a description of the sample, the results of pathological analysis, an evaluation of those results, clinical cancer staging information, and the findings of a physical examination of the

[3] Prostate Cancer Research Consortium, https://pcrc.tchpc.tcd.ie/.

[4] Slidepath, http://www.slidepath.com/.

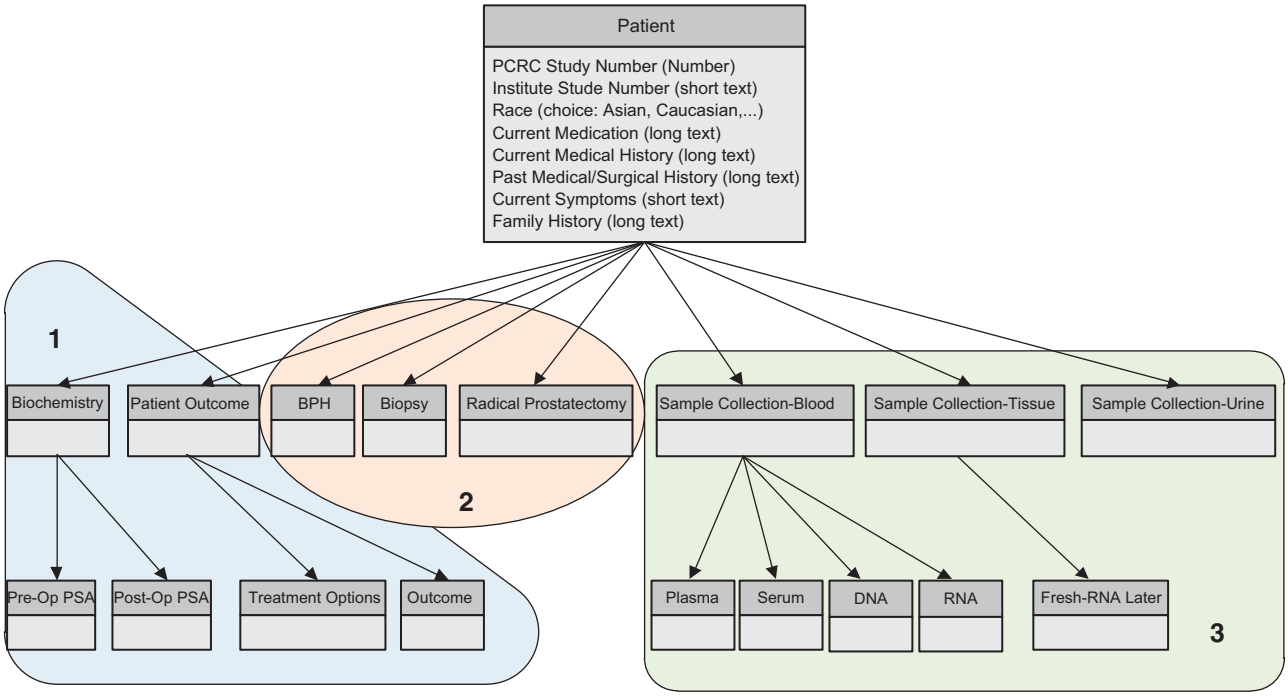


Fig. 1 – (a) Overview of the conceptual schema of the PCRC BIMS database (b–d) show parts 1–3 in detail. (b) Detail of the conceptual schema of the PCRC BIMS database—Part 1. (c) Detail of the conceptual schema of the PCRC BIMS database—Part 2. Pink ovals highlight concepts that reappear in different tables and blue rounded rectangles highlight the different concepts contained in the Biopsy table. (d) Detail of the conceptual schema of the PCRC BIMS database—Part 3. Green rectangles highlight the spread of one general concept over several tables.

patient, as highlighted in Fig. 1c with blue rectangles. On the other hand, the same concept was found to appear in several tables. For example, the fields related to the Gleason Staging system[5] appear in three different tables, as highlighted in Fig. 1c with pink ovals. Similarly, there are three tables concerned with sample collection: Sample Collection—Blood, Sample Collection—Tissue and Sample Collection—Urine, as marked in Fig. 1d with green rectangles, two of which model virtually the same information related to sample collection and sample processing such as "Date Collected", "Time of Voiding" or "Time of Collection", "Time of Spinning" or "Time of Processing" and "Time to process", with the third table modelling a subset of the former two. Other tables also contained very similar types of information, for example, the tables describing blood sample derivatives (Plasma, Serum, DNA and RNA) each contain sample location information and the number of tubes that contain the sample. These repeated information structures are due to the poor (non-normalized) design of the PCRC BIMS database that may have resulted from the ad hoc nature with which non-technical users can create table definitions using the PHP point-and-click interface. Additionally, information related to the same concept may be spread over several tables. For example, the results of the clinical histo-pathological analysis of prostate tissue are recorded in the BPH, Biopsy and Radical Prostatectomy tables. Thus, to facilitate searching for and identifying suitable candidate archetypes from public repositories and the creation of new archetypes, the tables and fields in the BIMS database schema were first regrouped under non-overlapping, distinct concepts. It helps here to have an idea of how archetypes are generally designed, such as attempting to make each cover a distinct concept in a generic way so that the archetype can be reused in different contexts. As part of this regrouping step the concepts that are repeated in several tables could be identified, such that the number of items to be modelled could be reduced, such as the "Location—Freezer Number" and "Location—Box Number" combination that appeared in five tables and the four fields involving the histo-pathological cancer staging tool Gleason Score, which appeared in three different tables. A generic archetype modelling Gleason Score for example can be reused in different contexts, e.g. to describe the Gleason Score of Biopsy Cores, or to describe the Gleason Score of resected tissue, as is required in this biobank.

This reorganization resulted in a structure of 23 high-level concepts that could be roughly divided into 13 clinical patient-centric (left side of mindmap) and 10 sample-centric (right side of mindmap) concepts (see Fig. 2a and b) with several specifications. The authors acknowledge that there is some overlap between the two sides, e.g. some of the concepts included on the sample side are clinically relevant to the patient, such as "Laboratory Test Results" and "TNM Cancer Staging". However,

[5] The Gleason Staging system is a used to grade the severity of prostate cancer. The pathologist assigns Gleason grades to the tumour patterns found in microscopic analysis of the prostate cancer tissue. The Gleason score is the sum of the individual grades.

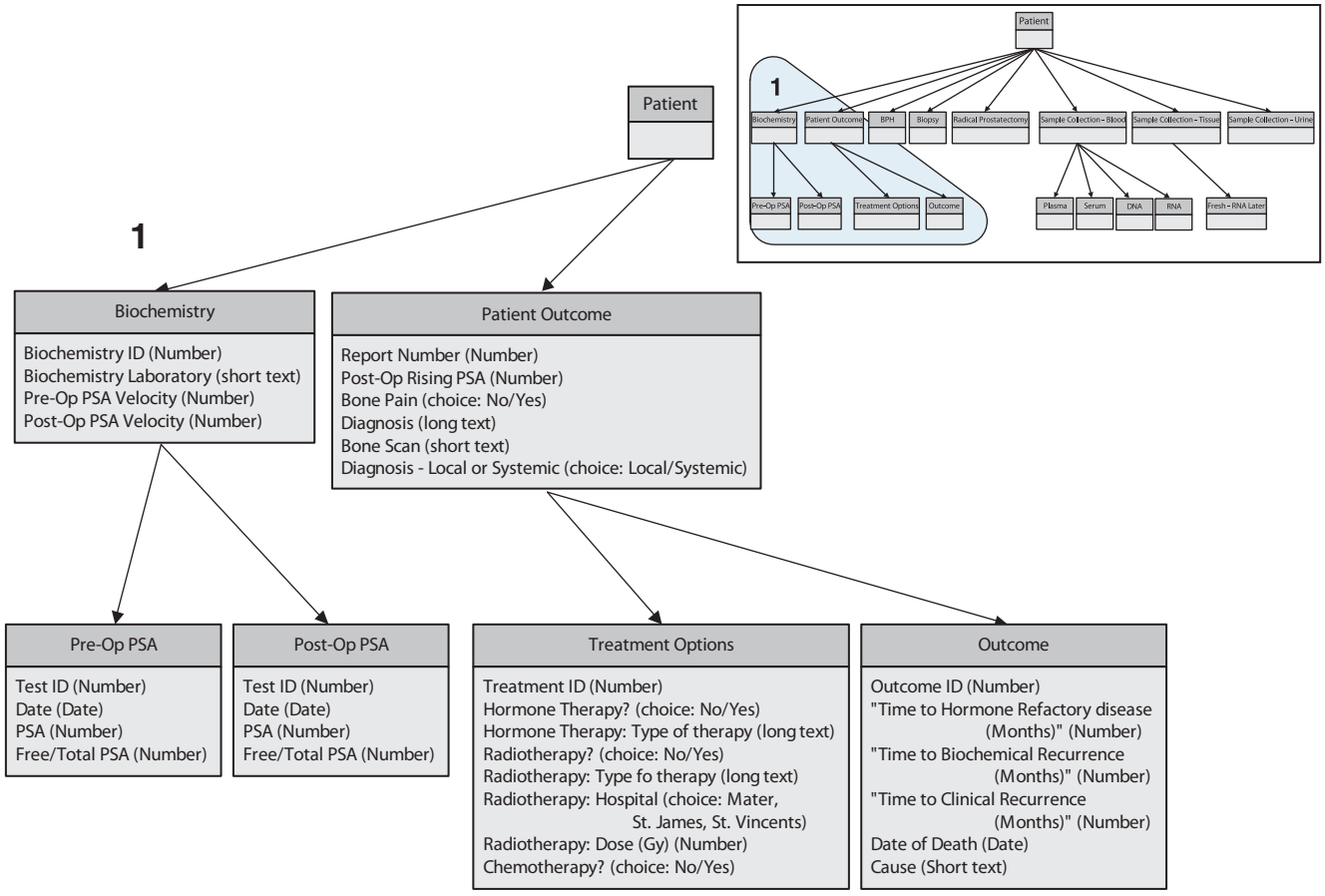


Fig. 1 – (*Continued*)

the justification for this classification is based on the observation that biobanking is a specimen centric activity where the subject of an examination or observation made, or of actions, carried out, is usually a sample, not a patient.

The annotations in Fig. 2a correspond to the annotations in Fig. 1c–d and highlight how these fields have been reorganized. For example, the different concepts contained in the Biopsy table have been split up into seven different concepts, as highlighted with blue rounded rectangles in Fig. 2a and b. This reorganized schema was then used as basis for the archetype modelling process.

The next step was to model the concepts in the database with archetypes. The strategy was to try and reuse as many existing archetypes where suitable and to develop new archetypes where none were available. This decision is based on the principle that a single archetype should represent a single clinical concept and archetypes need to be designed in a generic and reusable way so that they can be used in different contexts to enable interoperability and thus the sharing of EHRs and EHR EXTRACTs [38]. This also means that the proliferation of archetypes modelling the same concept should be avoided. Therefore, it was decided to seek to reuse as many existing archetypes as possible rather than creating an entirely new set of non-generic archetypes based on the particulars of this prostate cancer biobank, which would have only shifted interoperability issues to the level of archetypes rather than solving them. This also avoids carrying out the unnecessary work of re-developing a parallel set of high-fidelity research archetypes de novo as well as risking the development of overlapping archetypes that could jeopardize interoperability [39,40]. Reuse of existing archetypes would also support the direct reuse of archetyped EHR data.

Two archetype repositories, the aforementioned *open*EHR archetype repository and the NHS archetype repository were searched for candidate archetypes. The *open*EHR online repository was searched using the available string search function or through manual inspection of the archetypes, using the provided mindmap representation or looking at the ADL files directly. The latest NHS repository was downloaded and searched both manually by visual inspection of the ADL files and by searching for the occurrence of suitable words that may identify the desired archetype.

When potential matches were found, these archetypes were analyzed in detail to see if they modelled the information presented in the regrouped PCRC BIMS fields or if they were incomplete or if there was a conflict in how the information was modelled in the PCRC BIMS database and the archetype. In principle, each archetype should model a complete clinical knowledge concept and cover the maximum data set possible, and thus there should be no incomplete archetypes. However, since the archetype approach is still quite new and good quality archetypes are the result of

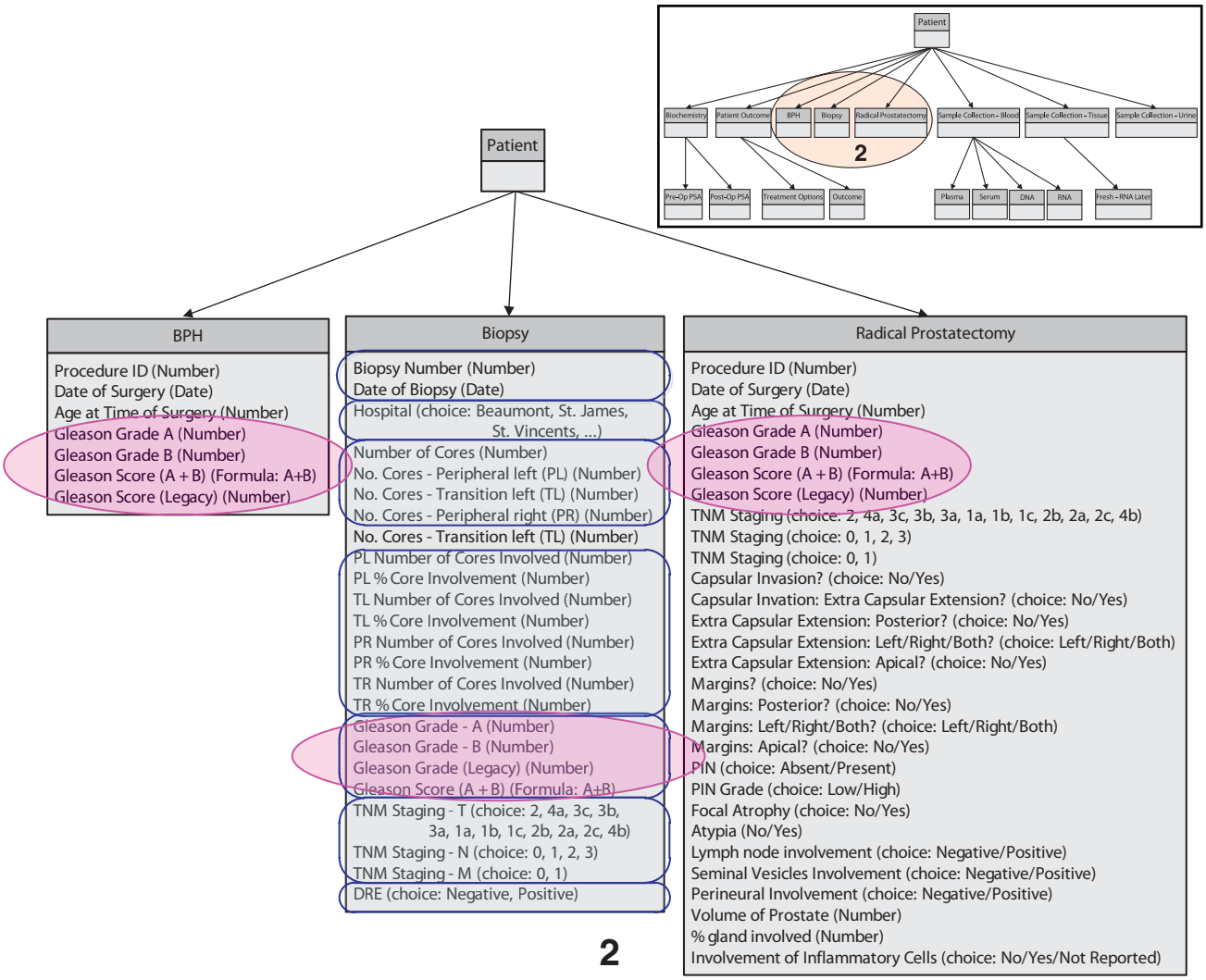


Fig. 1 – (*Continued*)

an ongoing consensus-based approach, the number of stable archetypes is small. For example, the NHS archetype repository contains only 1726 archetypes (as of January 15, 2010), of which 966 are of lifecycle_state = ⟨"0"⟩, 502 are of lifecycle_state = ⟨"Initial"⟩ and the remaining 222 archetypes are of lifecycle_state = ⟨"AuthorDraft"⟩, 32 archetypes are of lifecycle_state = ⟨"OrganisationDraft"⟩ and 4 archetypes are of lifecycle_state = ⟨"NotSet"⟩. At the time of writing, most archetypes in the *open*EHR archetype repository are still in a draft state (222 out of 244 active archetypes in total), with 15 in team review state and 7 published (as of January 15, 2010).

When an archetype was found to cover all fields of a certain BIMS concept, it was reused as is. When an archetype was found to cover only some of the fields in the regrouped BIMS concept, the repositories were searched again for further archetypes that covered those. The regrouping may have followed a different philosophy from that used to design these archetypes—it is not always clear where one concept ends and another one starts. However, if no other archetypes could be found to model the information in these fields, the archetype was deemed to be incomplete and was extended to contain the missing items, as they belonged to the same concept. If an archetype was found that covered a more general concept than one needed in the BIMS, the archetype was specialized to the more specific BIMS concept. Finally, when no archetype could be found for certain fields of the BIMS database, then a new archetype was designed according to the *open*EHR reference model described in the introduction. *open*EHR also provide a simple decision algorithm to help decide which archetype class should be used for a new archetype.[6]

The flow chart in Fig. 3 visualizes this approach. Initially the Linköping [41] archetype editor[7] and later the Ocean Informatics [42] archetype editor[8] were used to review, modify, specialize and create new archetypes.

[6] Decision algorithm: http://www.openehr.org/wiki/download/attachments/786529/Decision+algorithm.gif?version=3&modificationDate=1195620505000.

[7] http://www.imt.liu.se/mi/ehr/tools/.

[8] http://www.oceaninformatics.com/Solutions/ocean-products/Clinical-Modelling/ocean-archetype-editor.html.

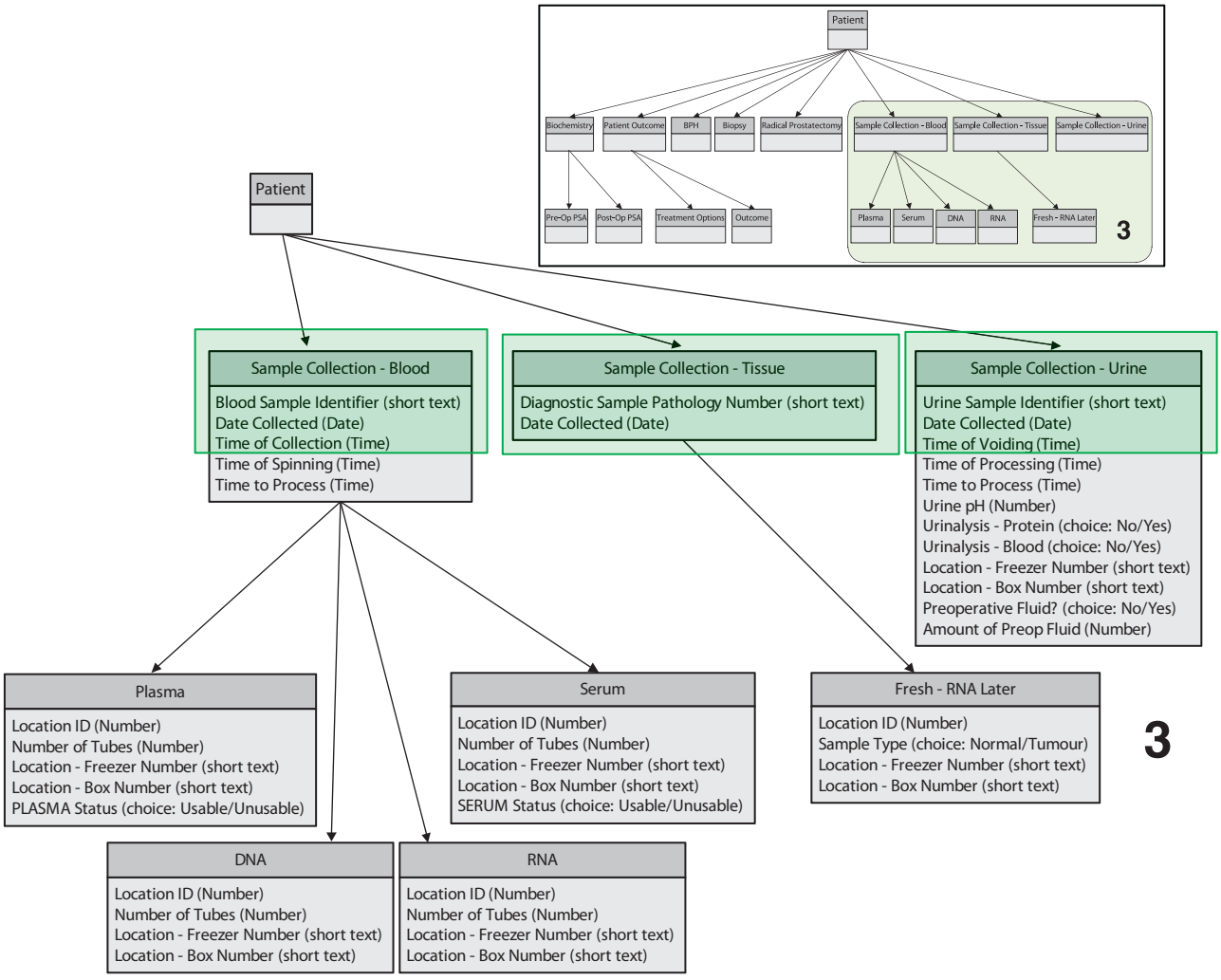


Fig. 1 – (*Continued*).

Once the set of archetypes to cover all the fields of the BIMS database was defined, a mapping was created between the fields in the PCRC database and the elements in this set of archetypes.

Finally, the set of archetypes was used to define templates to localize the archetypes to the specific requirements of the PCRC database. The Ocean Informatics Template Designer[9] (license required) was used for template design. Buck et al. [43] proposed essentially the same approach, the five-step odma approach, for modelling data for the development of *open*EHR-archetype based EHRs.

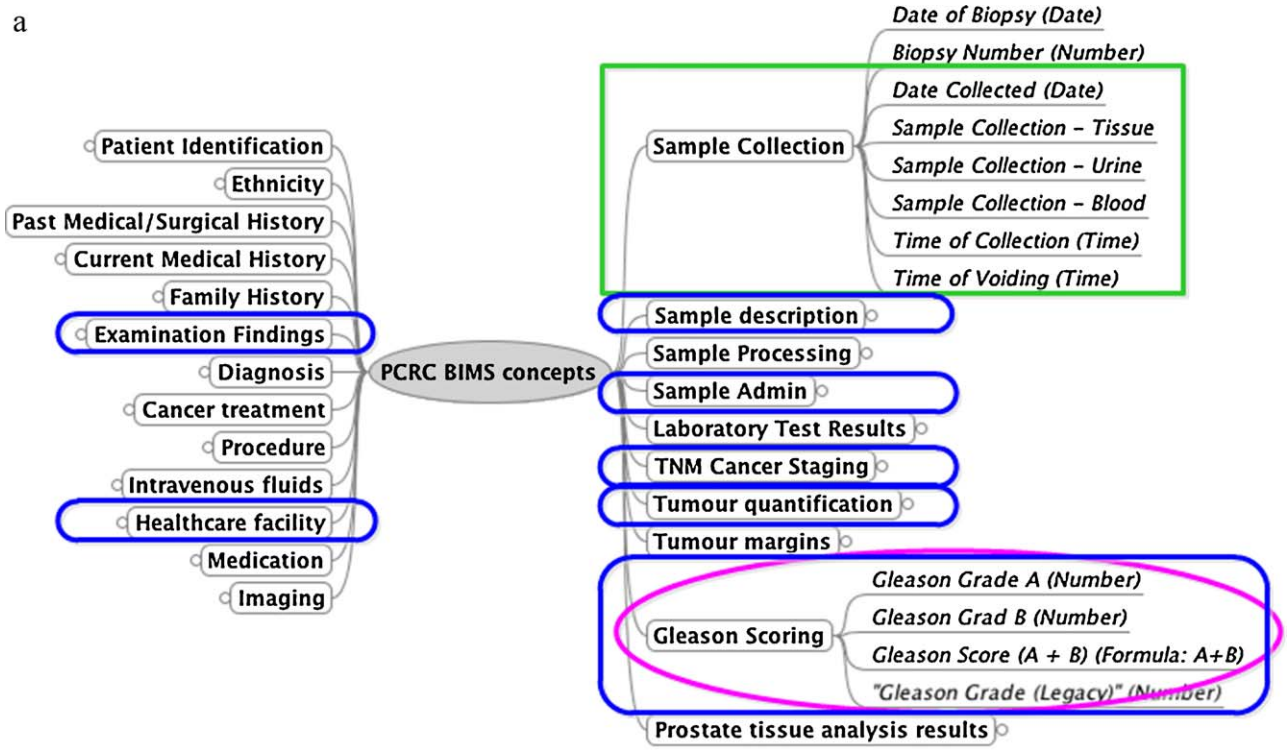
3. Results

A set of 43 archetypes that fully cover all the fields in the biobank database was identified (see Table 1). Many archetypes from the public repositories could be reused without change in the context of this biobank, making up the majority of the archetypes needed to cover the fields in this biobank database (27 out of 43, 63%). However, not all fields of the PCRC BIMS database were covered by existing archetypes, so that five new ENTRY (ACTION, ACTIVITY, INSTRUCTION, ADMIN_ENTRY, OBSERVATION, EVALUATION) archetypes needed to be developed, three for cancer therapy, one for PSA velocity, and an OBSERVATION archetype to hold the CLUSTER.specimen.v1 archetype, listed below:

- ACTION.chemotherapy.v1,
- ACTION.hormonetherapy.v1
- ACTION.radiotherapy.v1
- OBSERVATION.psa_velocity.v1
- OBSERVATION.specimen_collection.v1

Furthermore, four new organisational archetypes were defined to organise the data in the PCRC BIMS database in a sensible way. These archetypes (three COMPOSITION and one SECTION archetype) act as containers that point through the slot mechanism to other SECTION and/or directly to ENTRY type archetypes, thus providing context for the medical concepts modelled in these:

[9] http://www.oceaninformatics.com/Solutions/ocean-products/Clinical-Modelling/ocean-template-designer.html.



b

Biopsy	Concepts
Biopsy Number (Number) Date of Biopsy (Date)	Sample Collection
Hospital (choice: Beaumont, St. James, St. Vincents, ...)	Healthcare facility
Number of Cores (Number) No. Cores - Peripheral left (PL) (Number) No. Cores - Transition left (TL) (Number) No. Cores - Peripheral right (PR) (Number) No. Cores - Transition left (TL) (Number)	Sample description
PL Number of Cores Involved (Number) PL % Core Involvement (Number) TL Number of Cores Involved (Number) TL % Core Involvement (Number) PR Number of Cores Involved (Number) PR % Core Involvement (Number) TR Number of Cores Involved (Number) TR % Core Involvement (Number)	Tumour quantification
Gleason Grade - A (Number) Gleason Grade - B (Number) Gleason Grade (Legacy) (Number) Gleason Score (A + B) (Formula: A+B)	Gleason Scoring
TNM Staging - T (choice: 2, 4a, 3c, 3b, 3a, 1a, 1b, 1c, 2b, 2a, 2c, 4b) TNM Staging - N (choice: 0, 1, 2, 3) TNM Staging - M (choice: 0, 1)	TNM Cancer Staging - clinical
DRE (choice: Negative, Positive)	Examination findings

Fig. 2 – (a) BIMS fields regrouped into non-overlapping concepts (in bold). The left side shows clinical patient-centric concepts, and the right side shows sample-centric concepts. Some concepts on the right side are expanded to show the leaf items (*italic*) contained within them. These leaf items correspond to the PCRC BIMS database entities and attributes. The highlighting corresponds to the highlighting in Fig. 1c and d. (b) Detail of how the attributes of the Biopsy table (left) were divided into seven different concepts (right).

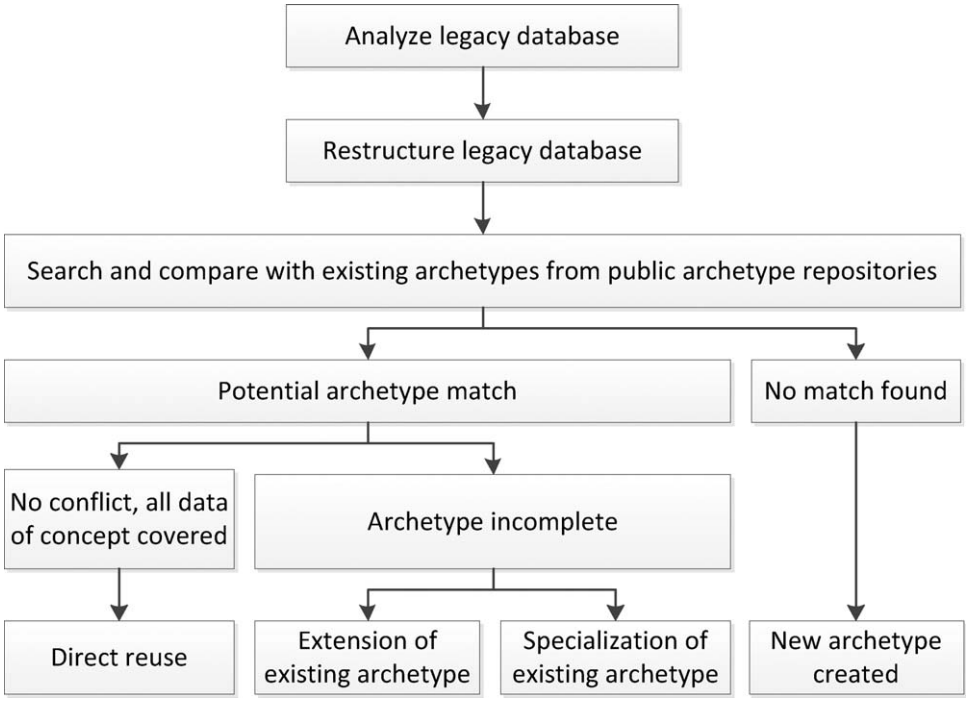


Fig. 3 – Flow-chart of modelling the concepts in a BIMS database with archetypes.

- COMPOSITION.family_history.v1
- COMPOSITION.patient_background.v1
- COMPOSITION.specimen_list.v1
- SECTION.history_medical_surgical.v1

Of a total of the nine newly developed archetypes, most could be categorized as belonging to the medical domain in general, whereas only the following two were thought to be specific for the biobank domain:

- COMPOSITION.specimen_list.v1
- OBSERVATION.specimen_collection.v1

COMPOSITION.specimen_list.v1 can be used to store descriptions of all specimens that are currently being stored in the biobank for a particular patient. OBSERVATION.specimen_collection.v1 acts as a holder for the existing CLUSTER.specimen.v1 archetype. Thus, a list of specimens for a particular patient can be stored by using the CLUSTER.specimen.v1 through the slot mechanism in OBSERVATION.specimen.v1 which itself is used through the slot mechanism in the COMPOSITION.specimen_list.v1 archetype (see Fig. 4). The intention here is to model a repository for available samples in the biobank. A discussion about the difficulties involved when the specimen entity is modelled with a CLUSTER archetype, as it currently is, is provided in the conclusion.

Finally, five archetypes were extended, one archetype was modified and extended and one archetype was specialized from a more generic archetype. Details about extensions and modifications to existing *open*EHR archetypes are listed in Table 2 and the templates created for the BIMS are listed in Table 3.

A mapping between the fields in the PCRC BIMS database to the elements in the set of archetypes as arranged in the templates was achieved in most cases, but sometimes a direct mapping was not possible, as discussed below. Details of the mappings can be found in Supplementary data.

4. Discussion

While it was possible to fully cover all the fields in the biobank database with archetypes, the process of modelling the information in the PCRC BIMS database with archetypes posed several challenges and various issues were encountered, which are discussed below. Some of these issues are due to the immaturity of the archetype approach, such as:

- Immaturity of modelling support tools.
- Lack of modelling guidelines or best practice rules.
- Overlapping archetypes.

Others, such as the difficulties encountered during the identification of suitable archetypes and establishing of a mapping between the PCRC BIMS database and archetype elements can be attributed to the fact that this research attempts to reuse archetypes, that have been designed for the clinical context, in a different (biomedical) context, although some of those may have been aggravated by the specific characteristics of the PCRC BIMS database.

Furthermore, the new archetypes developed during this research cannot be regarded as complete. This is particularly true for the three clinical cancer therapy archetypes developed above, for which no archetypes existed and for which the most basic archetypes were created, using only the informa-

Table 1 – Summary of archetypes used to cover the PCRC BIMS concepts.

Archetype	Source
EXTENDED:	
openEHR-EHR-CLUSTER.anatomical_location-precise.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.specimen.v1.adl	*open*EHR
openEHR-EHR-EVALUATION.problem-diagnosis.v1.adl	*open*EHR
openEHR-EHR-ITEM_TREE.procedure.v1.adl	*open*EHR
openEHR-EHR-OBSERVATION.urinalysis.v1.adl	*open*EHR
MODIFIED and EXTENDED:	
openEHR-EHR-CLUSTER.microscopy_prostate_carcinoma.v1.adl	*open*EHR
NEW:	
openEHR-EHR-ACTION.chemotherapy.v1.adl	
openEHR-EHR-ACTION.hormonetherapy.v1.adl	
openEHR-EHR-ACTION.radiotherapy.v1.adl	
openEHR-EHR-COMPOSITION.family_history.v1,adl	
openEHR-EHR-COMPOSITION.patient_background.v1.adl	
openEHR-EHR-COMPOSITION.specimen_list.v1.adl	
openEHR-EHR-OBSERVATION.psa_velocity.v1.adl	
openEHR-EHR-OBSERVATION.specimen_collection.v1.adl	
openEHR-EHR-SECTION.history_medical_surgical.v1.adl	
NO CHANGE:	
openEHR-DEMOGRAPHIC-CLUSTER.person_identifier_iso.v1	*open*EHR
openEHR-DEMOGRAPHIC-PERSON.person.v1	*open*EHR
openEHR_EHR-ACTION.intravenous_fluid_administration.v1.adl	*open*EHR
openEHR-EHR-ACTION.procedure.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.ethnic_background.v3.adl	NHS
openEHR-EHR-CLUSTER.exam-generic.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.lymph_node_metastases.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.organisation.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.palpation.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.physical_properties.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.specimen_preparation.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.symptom-pain.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.symptom.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.tnm_staging-prostate.v1.adl	*open*EHR
openEHR-EHR-CLUSTER.tumour_resection_margins.v1.adl	*open*EHR
openEHR-EHR-COMPOSITION.encounter.v1.adl	*open*EHR
openEHR-EHR-COMPOSITION.history_medical_surgical.v1.adl	NHS
openEHR-EHR-COMPOSITION.report.v4.adl	NHS
openEHR-EHR-EVALUATION.problem.v1.adl	*open*EHR
openEHR-EHR-EVALUATION.social_history.v8.adl	NHS
openEHR-EHR-ITEM_TREE.intravenous_fluids.v1.adl	*open*EHR
openEHR-EHR-OBSERVATION.exam.v1.adl	*open*EHR
openEHR-EHR-OBSERVATION.history_of_medications.v1.adl	NHS
openEHR-EHR-OBSERVATION.imaging.v1.adl	*open*EHR
openEHR-EHR-OBSERVATION.lab_test-histopathology.v1.adl	*open*EHR
openEHR-EHR-OBSERVATION.story.v1.adl	*open*EHR
openEHR-EHR-SECTION.adhoc.v2.adl	NHS
SPECIALIZED:	
openEHR-EHR-OBSERVATION.lab_test-psa.v1.adl	*open*EHR

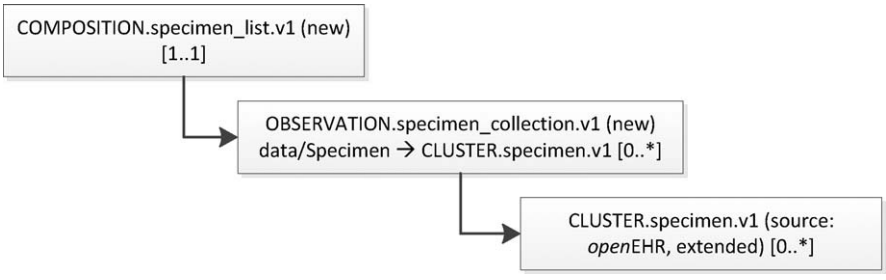


Fig. 4 – Containment hierarchy for modelling a specimen repository.

Table 2 – Changes carried out on existing archetypes.

Archetype	Source	Description of change(s) carried out
EXTENDED: openEHR-EHR-CLUSTER.anatomical_location-precise.v1.adl	openEHR	1. Added the value "Apical" to the list of values in [Cluster]Relative location:[Text]Aspect
openEHR-EHR-CLUSTER.specimen.v1.adl	openEHR	1. Added URI-resource identifier (Parent specimen) under Identifiers 2. Added Cluster:Storage Container:[Text]Type of Container and.:[Quantitity] Amount of specimen in this container 3. Added URI-resource Identifier (Recorded observation on sample)
openEHR-EHR-EVALUATION.problem-diagnosis.v1.adl	openEHR	1. Added URI-resource identifier (Observation used in diagnosis) under [Data]:[Cluster]Diagnostic criteria
openEHR-EHR-ITEM_TREE.procedure.v1.adl	openEHR	1. Added [Duration]Age at time of surgery
openEHR-EHR-OBSERVATION.urinalysis.v1.adl	openEHR	1. Added URI-resource identifier (Specimen used in analysis) 2. Added the value "8/Positive" to [Data]:[Cluster]Proteins:[Ordinal]Protein results 3. Added the value "8/Positive" to [Data]:[Ordinal]Blood
MODIFIED & EXTENDED: openEHR-EHR-CLUSTER.microscopy_prostate_carcinoma.v1.adl	openEHR	1. Added [Cluster]Core Details:[Slot for Cluster.anatomical_location_precise]Location of cores in prostate, added [Cluster]Core Details:[Quantity]Number of positive cores at this location, added [Cluster]Core Details:[Proportion]Proportion of positive cores at this location 2. Modified Definition for Total Gleason Score under [Cluster]Gleason Score:[Count]Total Gleason Score to say "The sum of the individual Gleason scores." instead of "The sum of the primary, secondary and tertiary Gleason scores." 3. Added [Cluster]Capsular Invasion:[Text]Presence:Present/Absent 4. Added [Cluster]Extra Capsular Extension:[Text]Presence:Present/Absent and [Cluster]Extra Capsular Extension:[Slot for CLUSTER.anatomical_location_precise.v1]:Locations of extra capsular extension 5. Added [Cluster]Prostate Intraepithelial Neoplasia:[Text]Presence:Present/Absent and.[Text]Grade:High/Medium/Low 6. Added under [Cluster]Additional findings:[Choice]Additional findings:Perineural Involvement/Focal atrohpy/Atypia

tion required in the PCRC biobank. These archetypes should be regarded as "stubs" to be improved upon by clinical domain experts in cancer therapy. Furthermore, there already exists a SECTION.history_medical_surgical.v1 archetype developed by the NHS, but while it modelled the desired concept, it was not generic enough to be used in this case. Thus, a new archetype has been developed under the same name, as this is exactly the desired concept to be modelled, demonstrating the problem of allowing archetype development to be carried out in an uncoordinated fashion. As discussed below, while some anarchy in the development of archetypes may be necessary at the start, there seems to be a need for coordination in the archetype development process, followed by certification of high-quality archetypes. Also, the changes carried out on existing archetypes can be considered fairly minor except for the CLUSTER.microscopy_prostate_carcinoma.v1 archetype, which describes microscopic analysis of prostate cancer tissue. The PCRC BIMS database provided much detailed information on microscopic analysis of prostate cancer tissue and most changes were carried out on the aforementioned archetype (see Table 2).

All archetypes that were reused directly or modified and/or extended were in the draft or team review state, and as such not stable. Thus the archetypes reused at the time of the project may have changed in the meantime.

These issues will now be discussed in more detail below.

4.1. *Immaturity of modelling support tools*

It was found that the Windows based (.NET) Ocean Informatics Archetype Editor was more mature and reliable than the java-based Linköping archetype editor. An alternative editor is the LinkEHR archetype editor [44], which will be tested for use in the future and more archetype editors are becoming available,

Table 3 – Templates created for the PCRC BIMS.

TEMPLATES
Biochemistry.oet
Family History.oet
Past Medical and Surgical History.oet
Patient Background.oet
PCRC Patient Admission Information.oet
Specimen List.oet
Tissue Analysis.oet
Urinalysis.oet

such as that developed by the software firm Medical Objects[10]. However, there is minimal documentation available on how to use these modelling support tools. The immaturity of the current range of modelling tools was also reported by Buck et al. [43] and Garde et al. [45]. Good tool support is imperative to promote participation of non-technical domain experts in the design of high-quality archetypes.

4.2. Lack of modelling guidelines or best practice rules

There are many different ways of modelling information, and no guidance or rules are provided to help decide which ways are best. These issues concern common database modelling principles, for example how to model Boolean type information. While the *open*EHR specification provides for a Boolean data type (DV_BOOLEAN), which allows the values "true" and "false", Boolean type information could just as well be modelled using structured text with the two values "true" or "false" for choice. Similarly, there are different ways to model a certain piece of information. For example, the presence of inflammation in a diagnosis archetype could be modelled by a DV_TEXT item called findings and the value to be filled into the archetype instance could be "Inflammation present". Alternatively, the fact that inflammation is or is not present could be modelled explicitly with a DV_CODED_TEXT item called "Inflammation" with the values "Present" or "Not present" for choice at instantiation, or even with a DV_BOOLEAN item called "Inflammation present" which could then either be "true" or "false". There are currently no guidelines that help to decide which alternative should be used. The implications of the lack of modelling guidelines or rules will make it harder to maintain consistency and a high level of quality amongst archetypes, a weakness of the approach already identified by the NHS [46]. Modelling guidelines or modelling rules should be developed and made readily available. A process or organization should perhaps be put in place to provide quality assurance on developed archetypes. How this might work is discussed further in Section 4.3.

4.3. Overlapping archetypes

In relation to the previous issue, archetypes that model the same or overlapping concepts between the two sources were found, e.g. the CLUSTER archetypes openEHR-EHR-CLUSTER.body_site.v4 (NHS) versus openEHR-EHR-CLUSTER.anatomical_location-precise.v1 (*open*EHR), both of which are essentially modelling the same concept. Currently there is no governance process or authority that decides which of several possible archetypes describing basically the same concept should be used. In the absence of domain knowledge governance, the proliferation of incompatible, overlapping archetypes for related or the same concept will become an issue. In effect, using incompatible and overlapping archetypes only shifts the problem of semantic interoperability onto the archetype level, requiring the effort of establishing mappings between these archetypes. Thus, to avoid 'rank growth' of archetypes that could jeopardize the goal of achieving semantic interoperability promised by the *open*EHR archetype approach, there is a clear need for an internationally coordinated archetype development and maintenance process to provide quality assurance for archetypes [40,47]. Garde et al. [48] suggest that ideally existing organizations should take on the task and responsibility of managing and publishing archetypes, such as national health information organizations for national scope archetypes and international organizations, such as the *open*EHR Clinical Review Board, or the Cochrane Collaboration[11] for international level archetypes.

Furthermore, Buck et al. [43] present a model of distributed but cooperative archetype development and adoption to enable the international and inter-professional coordination of archetype development and maintenance. Their model is based on a process involving a hierarchical structure of professional committees, where every committee represents a particular division of the health domain and is responsible for a certain set of archetypes.

However, while such structured process may be desirable in the long run, some members of the archetype community feel that a degree of anarchy in archetype development may be a necessary phase in the beginning to test alternatives and to achieve high-quality archetypes. There are clearly risks associated with finalizing archetypes too early, such as necessitating further changes to finalized archetypes that are not due to new knowledge creation, but to already existing knowledge. This may have a negative impact on downstream software, such as advanced information processing features, that rely on previous archetype versions, since adapting advanced information processing features to new versions can be a very resource consuming process, e.g. requiring redesign, reimplementation and retesting of software. Thus, changes to archetypes should only need to be carried out when new knowledge warrants archetype modification. Input from researchers in clinical archetype design, especially regarding archetypes that overlap between clinical practice and research, e.g. pathology archetypes, may help in achieving a clinical archetype design that is also well suited for reuse in research, thus improving reusability of the knowledge in archetype-based EHRs in the research context. This avoids having to develop a parallel set of research archetypes on the same concepts. It can also be envisaged that research generated new data definitions, especially in the field of prognostic or diagnostic biomarkers may feed back into the clinical domain, possibly meaning that existing clinical archetypes may need to be augmented with these. Archetype governance approaches should take this into account.

4.4. Mapping PCRC BIMS database to archetypes

The process of identifying suitable candidate archetypes that contain elements that correspond to fields in the BIMS database and of establishing a mapping between these two presented considerable challenges. The process turned out

[10] http://download.medical-objects.com.au/Template-Setup.exe.

[11] http://www.cochrane.org/—International not-for-profit organisation preparing, maintaining and promoting the accessibility of systematic reviews of the effects of health care.

to be manual and time-consuming, involving much detailed work. Currently, the only way provided to find out if specific elements are modelled in archetypes is through string search online on the *open*EHR archetype repository[12] or on the downloaded NHS repository, or manually scanning through the archetype names in the hope that they will indicate possible containers of the desired element(s). Often only a small subset of elements covered by an archetype is needed, and it is not always clear from the name of the archetype whether or not it contains these elements. This means that if a search does not return any results, it is difficult to know if no archetype yet exists that contains a particular element or if the containing archetype has not been found. There are many different ways of structuring and organizing data and many alternative terms to describe the same concept or item, so it is easy to miss an archetype or an element in an archetype if it is described in an alternative way. As the number of archetypes increases, this problem will become more serious. While this issue has also been identified by Buck et al. during the process of modelling an electronic patient record to support an innovative individual care concept for premature infants using the *open*EHR approach [43], it is aggravated in this research due to the fact that archetypes are used in a different context than originally envisaged. Archetypes are generally defined with a clinical, as opposed to a research setting in mind. Thus, although the biobank database contains the same clinical data on donors and samples, it may be organized quite differently, which was possibly complicated by the poor (non-normalized) design of the PCRC BIMS database, leading to repeated information structures across the database. This issue was partially alleviated by first reorganizing the database fields into more archetype-friendly concepts, as described in Section 2. Although clinical archetypes model generic concepts and thus should also be reusable even in a different context, the way they are structured and organized may reflect the structure and organization of a typical EHR. Either way, this issue means that there is a danger that new archetypes are being developed for concepts that have already been modelled with archetypes, thus increasing the danger of creating overlapping and incompatible archetypes. This issue needs to be tackled by the archetype community, possibly by providing automated searching and mapping to save modelling effort as proposed by Buck et al. [43]. Some research groups are working on easing this process, e.g. Fernandez-Breis et al. [49] present a Semantic Web System, using OWL ontologies, for managing clinical archetypes in a repository, allowing archetypes to be semantically annotated and semantically searched, but these functionalities are not commonly available in the current archetype repositories.

Some of the issues encountered in the mapping process are listed below, followed by a detailed discussion of each of them:

- Granularity of documentation.
- Deduced information.
- Metadata-level versus data-level modeling.
- Terminologies and vocabularies.
- Unstructured text.
- Mandatory items.

4.4.1. *Granularity of documentation*

Differences in the granularity of documentation between the PCRC BIMS database and existing archetypes created problems in mapping database fields to archetype elements. For example the existence of PIN (prostatic intraepithelial neoplasia, a pathological condition) and its grade in resected prostate tissue is modelled in the PCRC BIMS table "Radical Prostatectomy" with two fields: one field labelled "PIN" that can carry the values "Yes" or "No", and the field "PIN grade" that can carry the values "Low" or "High". By contrast the archetype modelling the microscopic findings of resected prostate tissue only contains a field under "Additional findings" for which the value "High grade Prostatic intraepithelial Neoplasia" is given as a possible selection only. Furthermore, while in the original database, the non-existence of margins at the apical location can be explicitly expressed by choosing the value "No", this cannot be done explicitly within the archetype, only by assumption, i.e. if no margin is mentioned on the apical side, then it may be assumed that there is no margin on the apical side.

While the more detailed information may not be necessary in the clinical context, in this case it is required in the research context. The consequence of using archetypes, which do not cater for that level of detail, is that information cannot be recorded accurately, leading to loss of information in the archetype layer. To solve this problem, the archetypes in question have been extended accordingly. This issue also argues that the research community should be involved in the design of clinical archetypes, so that these may be more readily reused in the research context.

4.4.2. *Deduced information*

Some information can be deduced from other information in the record and may not have been modelled explicitly in an archetype, thus preventing the establishment of a direct mapping. For example, the field "Time to process" in the PCRC BIMS urine and blood sample collection tables is not currently a field modelled in existing archetypes for sample collection and as such cannot be mapped. However, this information can be computed from the "Time of processing/spinning" minus "Time of voiding/collection", which are modelled in the archetypes. The question is whether there is a semantic need to create an extra field in the corresponding archetype to model this information or to provide a mechanism for calculating this value outside the realm of the archetype layer or to let the users deduce or calculate it themselves. The *open*EHR Reference Model does support calculations within the archetype through an "invariant" section that supports first order predicate logic statements in the archetype in ADL 1.4 [50] (which is replaced by a "rules" section in ADL 1.5 [51]). This section in an ADL archetype can introduce assertions that reference several elements in an archetype and can relate them through mathematical or logical formulae that can be evaluated to a Boolean result at runtime [50]. Unfortunately, currently neither archetype editor used in this research supports adding such rules to an archetype, so they would need to be added manually to the

[12] http://www.openehr.org/knowledge/.

ADL file and this feature is not currently used in existing archetypes.

4.4.3. *Metadata-level versus data-level modelling*

Problems were also encountered due to several cases of mismatches between metadata- or schema-level modelling versus data-level modelling between candidate archetypes and the original database. This concerns situations where information modelled in the schema of the PCRC BIMS database e.g. as a field or table name, is not modelled as meta-data in the archetype, e.g. as the name of an element, but as data value in an archetype element under a more generic element. This situation often leads to one-to-many relationships: to model one field from the database, several fields, possibly over several archetypes, may be needed.

One such example concerns the PCRC BIMS database field "Margins: Posterior" that can contain a choice of "Yes" or "No" as possible data values. The field name, i.e. the meta-data, combines two items of information, the name of the histo-pathological finding (margin visible) and the location of this histo-pathological finding in the prostate (posterior), with the possible data values either confirming or denying the existence of the margin at the posterior location in the prostate. To model this information with existing archetypes, one needs to use the openEHR-EHR-CLUSTER.anatomical_location-precise.v1 archetype with the value of the item "Relative location:Aspect" set to "Posterior", which is embedded in the openEHR-EHR-CLUSTER.tumour_resection_margins.v1 archetype in the "Margin location" slot, which itself is embedded in the openEHR-EHR-CLUSTER.microscopy_prostate_carcinoma.v1 archetype through the slot mechanism at the "Resection margin detail" slot under the "Surgical resection margin" cluster that is contained under the "Surgical resection margins" cluster in the openEHR-EHR-CLUSTER.microscopy_prostate_carcinoma.v1 archetype. It is easy to see that the task to correctly model a field from the original database in the correct context with archetypes can become very complex.

4.4.4. *Terminologies and vocabularies*

The terms used in the PCRC BIMS database differed in several cases from those used in the archetypes in the public repositories. Thus, it is of utmost importance to obtain a detailed meta-data description of the fields in the source database to be able to determine if two different terms essentially mean the same thing, e.g. "Involvement of Inflammatory Cells" (in the original database) and "Inflammation" (in the openEHR-EHR-CLUSTER.microscopy_prostate_carcinoma.v1 archetype) or to differentiate between things that have the same name but mean different things. While archetypes themselves are terminology neutral and can be developed independently of external terminologies, they can be linked to multiple external terminologies [30,45,52]. The issues encountered above highlight the importance of using the option of binding archetypes and data in archetypes to common medical vocabularies, such as SNOMED-CT or LOINC to uniquely define their meaning. Many of the terms in the PCRC BIMS database could be easily modelled with pre- or post-coordinated terms from medical terminologies (cf. [53–55]). Thus, it would have helped if both the archetypes and the original database had provided detailed term-bindings to a recognized medical terminology, such as SNOMED-CT, to ensure equivalence.

4.4.5. *Unstructured text*

The data in the PCRC BIMS database was often stored in a free-text, unstructured form, whereas the archetype approach inherently promotes a very structured way of documenting clinical data in the EHR, which can present a problem when legacy records are to be converted into an archetype-based record, a problem previously recognized by Bird et al. [38]. For example, the fields "Current Medication", "Current Medical History", "Past Medical/Surgical History", "Current Symptoms" and "Family History" all allow a free text entry, whereas corresponding archetypes follow a much more structured approach. Furthermore, sometimes the information provided as unstructured text in one field of the PCRC BIMS database may include such diverse data that it could be modelled with one of several possible archetypes. For example, five potential archetypes were found to potentially model the medical histories provided as unstructured text in the "Current Medical History" or "Past Medical/Surgical History" field in the "Patient" table, depending on the particular content:

- EVALUATION.problem-diagnosis.v1
- EVALUATION.problem.v1
- EVALUATION.problem-genetic.v1
- EVALUATION.injury.v1
- EVALUATION.problem-diagnosis-histological.v1

Two additional archetypes, ITEM_TREE.procedure.v1 embedded in ACTION.procedure.v1 could be used to model the past surgical history potentially provided in the "Past Medical/Surgical History" field in the "Patient" table, increasing the number of possible archetypes that model the information in this particular field to seven. Since a direct mapping between the content of the fields of an existing database to elements in an archetype is often not possible, this may represent a problem when, as part of converting to an archetype-based record or as part of exposing the legacy database through an archetype layer, the archetype-based database is to be populated with data from the original database or when the original database is to be queried through an archetype layer. It should be noted that these free text fields as they are recorded in the current PCRC BIMS database are in fact of limited use. The information in these text fields appears to be used mainly by the researchers to scan for any previous occurrences of cancer or cancer in the family, which could have been easily recorded in a structured way, which would greatly facilitate searching. While extracting this information manually may be feasible on projects involving small numbers of donors, it would not be feasible at this moment to extract it on a large basis, which in a future scenario of connected databases may become necessary.

In the future, advanced natural language processing techniques, information extraction techniques and semantic technology (e.g. [56–62]), may be used to automatically extract information from the unstructured text fields in an existing database and populate the correct archetype fields with it. However, advances in this area are still very much con-

fined to the research environment [62] and outside the scope of this research. It would also be unfeasible to manually go through the content in each PCRC BIMS database field and distribute the information chunks into the appropriate fields in the structured archetype.

Two workarounds, both less than ideal, have been considered. First, another archetype could be developed which basically models the fields "Current Medical History" and "Past Medical/Surgical History" as unstructured text fields, labelled "Current Medical History" and "Past Medical/Surgical History", respectively, so that a one-to-one conversion is possible. The approach leads to the development of an archetype that models the same information as the above archetypes, only in a much less structured format. However, this approach would lead to the development of overlapping archetypes, which would cover the same concept, thus shifting the interoperability problem to the archetype layer.

The second alternative is to decide on one archetype that seems best to cover the content provided, e.g. EVALUATION.problem-diagnosis.v1 and use the most suitable field, e.g. the "data/Diagnosis" field to carry all the information provided in the "Current Medical History" or "Past Medical/Surgical History" fields in the PCRC BIMS database. In this alternative, the one chosen in this research, one has to accept that the content provided in the PCRC BIMS database and filled into the "data/Diagnosis" field may contain more information than just the diagnosis itself or not even a diagnosis, e.g. the description of an injury that would be better modelled with the EVALUATION.injury.v1 archetype.

While this current research is concerned with modelling the data that is currently present in the PCRC BIMS biobank with archetypes, for future data entry, all potential archetypes should be provided, so that new data can be entered into the appropriate fields in the appropriate archetypes.

4.4.6. *Mandatory items*

Some items that are mandatory in the archetypes, such as information committer, are not provided or recorded in the PCRC BIMS database, as these may be regarded as irrelevant or unimportant in the research context, so these fields need to be labelled "unknown", when it is impossible to provide this information retrospectively. This problem was previously experienced also by Bird et al. [38].

4.4.7. *Summary of mapping issues*

Although the information to be recorded in clinical practice and biobanks overlaps, clinical archetypes reflect the information needs and recording practices in clinical practice, and thus do not always fit the recording needs and requirements in biobanks, which may be different. Thus, the mapping issues discussed above (Sections 4.4.1–4.4.6) could be attributed to reusing archetypes that have been developed for use in the clinical context in a different context, the research context, although some may have been aggravated by the particular characteristics of the PCRC BIMS database. In fact, many of challenges faced above are commonly encountered during the integration of distributed and heterogeneous information systems that have been designed for different use contexts. For example, Park and Ram [63] discuss several types of semantic conflicts that can occur and need to be mediated during system integration efforts. According to Park and Ram [63], semantic conflicts can either occur on the data level, such as data-value conflicts, data representation conflicts or data-unit conflicts etc., or on the schema level, such as naming conflicts, entity-identifier conflicts and schema-isomorphism conflicts etc. Using this classification, the issue described in Section 4.4.1, "Granularity of documentation", could be categorized as a data precision conflict, which is a data level conflict, issues described in Section 4.4.3, "Metadata-level versus data-level modelling", and Section 4.4.5, "Unstructured text", could be classified as schematic discrepancies, which are schema-level conflicts, and the issues described in Section 4.4.4 "Terminologies and vocabulary", are naming conflicts, which also occur on the schema level.

Today, most current EHRs and BIMS databases do not follow any recording standard but on the contrary use bespoke databases that have been designed specifically for the situation at hand and are only relevant to a particular hospital, study or biobank, often in proprietary format, reuse of the data in the database in other contexts is not possible. The *open*EHR archetype approach offers a possible solution to the sharing of information between biobanks or between EHRs, but since most BIMS databases and EHRs are bespoke implementations, the effort involved in achieving a mapping between the existing databases to existing archetypes, even when all the archetypes needed to cover the fields in the databases have already been developed and are stable, can be considerable, as discussed above. However, if archetypes were only to be used in a "green-field situation", i.e. in situations where new biobank databases or EHRs are being set up, then the wealth of information that is contained in existing databases could potentially be lost. In addition, the same conflicts may still arise in this situation, due to differences in information and documentation needs between healthcare facilities and biobanks, reconfirming the need for researchers to take part in the design of archetypes in the overlapping domains to make them reusable in the research context. Thus, if the biobank community chose to use the *open*EHR archetype approach to share information, then the use of a stepped approach towards enabling archetype support in the biobanks is recommended. For example, a set of basic archetypes could be established that each participating biobank should use and implement and which could be mapped to existing database schemas. Then more archetype support could be added gradually with time and as more stable archetypes become available. This means that for some time some of the information in an existing BIMS database will be available through the archetype layer while other information will only be available through the existing information system. If it is desired to switch over to an archetyped solution completely in the near future, i.e. in the absence of stable archetypes, as a bridging solution local archetypes, or integration archetypes may be used for a while until universally valid archetypes become available. Chen et al. [64] present a strategy of adding archetype support in legacy EHRs that could easily be adopted for adding archetype support in legacy databases in biobanks. Decisions also need to be taken on a case-by-case basis if the effort required to properly transform legacy information into archetyped information is worth the effort, such as deconstructing unstructured text to fill appropriate fields in the archetypes rather than dumping

the unstructured text into the "comment" field. On the positive side, the mapping effort only needs to be achieved once between each database and archetypes.

4.5. *Further observations*

Finally, the following observations have been made during this research. Much use of the openEHR-EHR-SECTION.adhoc.v2 archetype was made in the development of the templates for this biobank. This archetype is a "helper" archetype with no structure or constraints and is used in COMPOSITION archetypes through the slot mechanism by renaming it to the appropriate heading. While this ad hoc way of organizing information in COMPOSITIONs allows for the quick development of templates without having to develop a set of new SECTION archetypes, the use of this archetype should probably be reduced and replaced by well-defined SECTION archetypes when universal organizational structures of information in biobanks become clearer. Thus the number of new archetypes needed to model the information in the biobanks is expected to be much higher in the number of SECTION archetypes, if all necessary SECTION archetypes had been properly defined. However, since SECTION archetypes, which correspond to document headings, are generally used to organize the clinical information contained in ENTRY archetypes [35], the definition of universally valid high-quality SECTION archetypes was deemed much less important than the development of primary archetypes that contain the clinical concepts and agreement on the structure of SECTION archetypes is deemed much less critical for interoperability purposes.

Another important question deals with the issue of assumed or locally known context. It was found that a lot of context was not explicitly recorded in the PCRC BIMS database, but assumed or known locally by the database users. For example, since the biobank is only concerned with prostate cancer, all sample donors will be male, a fact that is not explicitly stated in the database, since it can be assumed from the context. Similarly, as discussed in Section 4.4.3, the table named "BPH" in the PCRC BIMS database describes the tissue analysis of prostate tissue that was collected through a procedure called TURP (TransUrethral Resection of the Prostate). The database users know from other database users that the tissue was collected through TURP, but this fact is also not recorded in the database. The authors only found out this information during enquiries with the nurses that enter data into the database. However, it is this implicit, assumed or locally known information that needs to be recorded alongside the explicitly provided information to provide the context so that the information in the database becomes meaningful independently of the local knowledge, e.g. if the eBMRR for the PCRC biobank is meant to be accessible outside the realm of this biobank. Similarly, to facilitate intelligent reasoning or data mining, this type of information needs to be stated explicitly. In the PCRC biobank, this context information has now been documented by the nurses and is supplied in form of an accompanying Standard Operating Procedure document to new database users. However, with appropriate archetypes, this information can be encoded alongside the explicitly modelled information. During this research, some attempt was made to explicitly model some of the implicit context in archetypes where feasible, e.g. to record TURP in the patient's surgical history and to record BPH (Benign Prostatic Hyperplasia) in the patient's diagnosis. However, this process is not complete and a much more complete context of the data in the PCRC BIMS database could be provided in archetypes, requiring more work to be carried out. In this process a balance needs to be found between the semantic gain of modelling this context information in archetypes and the effort needed to do so.

A further issue encountered in this research concerns the modelling of a sample repository with archetypes. Currently the structure of the eBMRR is based on the patient as the root subject of the record, the record subject. One of the issues that arose during the archetyping process was how to deal with the fact that biobanking is a sample-centric activity that treats the specimen as an independent entity, which can have its own observations, evaluations, instructions and actions associated with it. The question is how can current archetypes be used to model a biobank's sample repository? The *open*EHR repository provided the CLUSTER.specimen.v1 archetype to model the specimen entity. Since it is an archetype of class CLUSTER, this archetype is a "leaf" archetype, towards the bottom of the EHR EXTRACT hierarchy. Currently, actions carried out on the sample are modelled with the CLUSTER.specimen_preparation.v1 archetype that is used in the CLUSTER.specimen.v1 archetype through the slot mechanism. Since the CLUSTER.specimen.v1 archetype is a "leaf" archetype, it can only be used within some ENTRY archetype, such as some laboratory OBSERVATION archetype.

In this regard, the specimen entity is modelled as a part of the OBSERVATION or some other ENTRY archetype. It was found that in this way it would be cumbersome to model sample workflow, actions, and experimental procedures carried out on the sample entity in a biobank, which may store samples for future observations, but which may have no observation associated with them currently. As a workaround, the OBSERVATION.specimen_collection.v1 archetype was developed to act as a holder for the CLUSTER.specimen.v1 entity, without providing any other information. Furthermore, the CLUSTER.specimen.v1 was extended to contain a field that can be linked to observations recorded on this sample, e.g. before it was stored for future use. Similarly, the observational archetypes that did not contain a slot for the specimen used (e.g. OBSERVATION.urinalysis.v1) have also been extended with fields that can be used to link to the specific sample instance used in the observation, which is stored under the OBSERVATION.specimen_collection.v1 archetype. It was also found cumbersome and unclear on how to model the fact that child samples can be created by sampling from previous samples with the current archetype structure: For example, although blood is the original sample collected, it is not stored as blood in the biobank, but it is processed and spun down and stored as its constituents serum and plasma. Similarly, the sample may be thawed, a part taken out of it and the rest refrozen, effectively creating a modified sample instance and a derived sample instance from the previous one. Should information about these child sample instances be stored in two new instances of the CLUSTER.specimen.v1 archetype embedded in a new ENTRY archetype instance? Should the previous CLUSTER.specimen.v1 record be updated? Would such action

conflict with the idea that the parent ENTRY archetype concerned the parent sample instance? How can the sample parent–child hierarchy and actions and observations on the samples be modelled easily? It was found that the current way of modelling sample related information is not suitable in the research context and an alternative solution is proposed in Section 5.2.

Finally, as part of analysing the database, some flaws were detected in the PCRC BIMS database, such as using different terms for the same meaning or inconsistencies in handling the semantics of null values as well as the use of unstructured text fields where structured fields would have been more appropriate. The archetype approach is designed to help avoid such issues, e.g. by providing well-defined consensus-agreed archetypes.

5. Conclusions

5.1. *Overview*

In this research, the archetype approach was applied to model the information in the existing database of a prostate cancer biobank, so that the information in this biobank can be completely represented with archetypes. To the best of our knowledge this is the first attempt to apply the archetype approach to model the information in a biobank database and to reuse archetypes that have been developed for the creation of EHRs in the context of a biobank. In this regard it was very promising that the majority of the information in the PCRC BIMS database, including sample-related information, could be represented with existing archetypes which could be reused without change. This indicates that the underlying principle of reuse of archetypes is valid and that archetypes developed to represent clinical concepts in EHRs can generally be reused to represent clinical concepts in biobanks. Some archetypes needed to be extended and modified, which shows that archetypes are still not mature and a set of high-quality archetypes with which to model clinical data is yet to be achieved. Real-world implementation projects such as this can help improve clinical archetypes to make them reusable in different contexts, such as within the research context in the eBMRR, thus promoting the reuse of clinical information in research. Basing the proposed eBMRR on the same technology as is currently being proposed for EHRs provides a parallel solution for sharing information amongst collaborating biobanks, but it also allows to import relevant background or phenotypic information about the patient directly from an archetype-based patient's/donor's EHR and vice versa.

Several issues with the openEHR approach were encountered during this research, including immature modelling support tools, lack of modelling guidelines or best practice rules and the lack of domain knowledge governance. However, the majority of issues were encountered during the process of mapping the PCRC BIMS database to archetypes. The identification of suitable archetypes is still quite a time-consuming manual process. Adding archetype-support to an existing system is a challenging task, facing problems commonly encountered during integration of heterogeneous and distributed systems, such as differences in the granularity of documentation, mismatches between schema and data modelling, differences in terminologies and vocabularies used, differences in structure imposed on information to be recorded, and differences in mandatory information. The value of the archetype approach as candidate approach for information sharing between hospitals or biobanks could be much increased if the process of converting existing data sources into archetype-based format, once they become available, can be eased, e.g. through appropriate annotations using common medical terminologies in both the existing system and archetypes and the development of automated mapping tools. Possibly, research into solving semantic interoperability conflicts to facilitate the integration of distributed and heterogeneous information systems (e.g. [63]) could be leveraged for this. Furthermore, there is a need for the development of high-quality archetypes to cover more clinical and research domain knowledge.

On the positive side, the process of archetyping can expose flaws in legacy databases and thus provide a common basis for discussion about what kind of information should be recorded and how, independent of technical implementation details.

Finally, a major advantage offered by the use of EHRs is that they support the reuse of vast amounts of clinical information including outcomes data for different purposes, although the jury is still open on whether the reuse of clinical data for such purposes is feasible, due to the different requirements regarding completeness and granularity. However, the results of this research provide positive evidence to support this potential in the context of biomedical research and thus contribute to the debate regarding the reuse of clinical data for biomedical research.

5.2. *Future work*

Current archetype developments of the archetype community focus on the development of archetypes for the clinical domain and thus do not cover research information, such as omic data. However, the eBMRR proposed in this research intends to augment the clinical information with research information created on patient samples. Thus, future work towards the eBMRR will be concerned with the development of archetypes to specifically cover experimental protocols and data, especially those of high-throughput technologies, such as genomic, proteomic and other omic data, which can then be customized to the specific experimental protocols and research results obtained in a particular biobank. These are currently not modelled in the biobank database itself, but provided in separate documents.

In addition, the idea of creating a separate and independent sample-based record for each specimen entity will be further investigated. In this alternative solution, the sample, not the patient, acts as the subject of the record. Such a sample-based record would contain its own record hierarchy with ENTRYs, such as OBSERVATIONs, EVALUATIONs, INSTRUCTIONs that would describe observations, evaluations and instructions on the sample rather than on the patient/donor. Since further sub-samples can be created from existing specimens by sampling, and the resulting child specimens may inherit some characteristics from their parent sample, these sample-based records need to be linked to each other hierarchically, reflect-

ing their parent–child relationship, and linked to the regular patient-based record at the top of the hierarchy. This results in a hierarchical "record-within-record" structure. The *open*EHR reference model allows both for a different (i.e. non-patient) subject of care and for referring to other records using the DV_EHR_URI data type and the LINK class [29]. This would effectively turn the current way of dealing with the sample entity "on its head", so that it would no longer be treated simply as an attribute of an observation, but rather as the subject on which an action was carried out or an observation was made. It would also reflect the fact that biobanking is a sample centric activity where a sample is treated as an independent entity, which may be kept for future research, even after the donor's death (see Section 2). The authors are of the opinion that this is a much more natural way of recording information regarding a sample and that in this way sample processing workflows and results of experimental analysis could be modelled more easily. For the creation of archetypes for the sample-based record, other standardization work in the specification of specimen information and structure will be consulted. Possible candidates for input into the creation of such archetypes are the common specimen model defined by the IHE/HL7/DICOM collaboration [65] or the "specimen domain" [66] and "specimen Common Message Element Type" [67] specifications from the HL7 v3 standard. Harmonizing between the different specifications would also improve and promote interoperability between these.

Furthermore, the work achieved in this research will feed into the implementation of a prototype of the eBMRR, which will be customized to the PCRC biobank. There are generally two alternatives to the implementation. One possibility is to implement an "archetype layer" above the database schema, with mappings between the legacy database and the archetype items, and the legacy database can then be queried and populated through this archetype layer. This approach has the advantage that current software and database users using the database can keep using the original database, but the data can also be exposed through the archetype layer and thus populated and queried through it. The second possibility is to implement a new system with a new archetype-based underlying database that is populated and queried through the archetypes directly. In this case, the information from the original database needs to be migrated over into the new database and new data needs to be added through the archetype layer and would generally be more useful in a "green-field situation". For purposes of proof of concept, this research intends to follow the first alternative in the development of the eBMRR prototype. However, due to the complex proprietary nature of the PCRC BIMS database implementation, a test database based on the conceptual database schema provided will be used. In either case, the archetype/template layer will be used for data validation upon retrieval/import from the test PCRC BIMS database [29] where possible.

Finally, to investigate and improve the genericity of the prostate cancer domain specific and general biobank archetypes developed and reused in this research for the database of the Irish Prostate Cancer Research Consortium the authors intend to repeat this process of archetyping existing biobanks with another prostate cancer biobank as well as other biobanks from different domains.

Summary points

"What was already known on the topic"

- The *open*EHR archetype approach is considered a promising approach to generically model clinical information for the construction of semantically interoperable EHRs.
- Development of *open*EHR archetypes to model clinical content is an ongoing activity, with large organizations, such as the UK National Health Service taking part in this process.

"What this study added to our knowledge"

- The *open*EHR archetype approach can also be used to model the clinical information about sample donors and sample related information in biobanks.
- Many archetypes originally developed for the clinical context in an EHR can be reused in the context of research for the construction of an electronic biomedical research record (eBMRR), thus validating existing archetypes.
- The identification of suitable archetypes and the process of mapping the biobank database to these faces common integration challenges, such as differences in the granularity of documentation, in metadata-level versus data-level modelling, in terminologies and vocabularies used, and in the amount of structure imposed on the information to be recorded.
- An alternative, sample-centric way of modelling the sample entity and sample related information within the realm of the openEHR Reference Model is being proposed.
- This research recognized the need for better support tools, modelling guidelines and best practice rules and reconfirmed the need for better domain knowledge governance.

Authors' contributions

Melanie Spath performed the analysis, the identification of matching archetypes, the development of new archetypes, the mapping of the database fields to archetype elements in the templates and wrote and revised the manuscript. Jane Grimson participated in the design of the study and critically reviewed the final manuscript and its revision. An earlier version of this paper was presented (the proceedings were not published) at the Health Informatics Society of Ireland's Annual Conference 2009 and was awarded the best paper prize.

Competing interests

The authors declare that there are no competing interests, financially or otherwise.

Acknowledgements

The authors would like to thank the members of the Prostate Cancer Research Consortium for their help in clarifying the meaning of the database fields and entries. The authors would also like to thank Damon Berry for his helpful comments on an earlier version of the manuscript. This research was supported by an Irish Council for Science and Technology Embark Post-graduate Scholarship who have no financial or other interest in this research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijmedinf.2010.11.002.

REFERENCES

[1] ISO (ISO TC 215 Health Informatics), ISO/TS 18308:2004 Requirements for an Electronic Health Record Reference Architecture, 2004. Available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=33397, last accessed April 2010.
[2] J. Grimson, Delivering the electronic healthcare record for the 21st century, Int. J. Med. Inform. 64 (2001) 111–127.
[3] P.H. Riegman, M.M. Morente, F. Betsou, P.d. Blasio, P. Geary, the Marble Arch International Working Group on Biobanking for Biomedical Research, Biobanking for better healthcare, Mol. Oncol. 2 (2008) 213–222.
[4] S.R. Jayasinghe, A. Mishra, A. Van Daal, E. Kwan, Genetics and cardiovascular disease: design and development of a DNA biobank, Exp. Clin. Cardiol. 14 (3) (2010) 33–37.
[5] P. Founti, F. Topouzis, L. van Koolwijk, C.E. Traverso, N. Pfeiffer, A.C. Viswanathan, Biobanks and the importance of detailed phenotyping: a case study—the European Glaucoma Society GlaucoGENE project, Br. J. Ophthalmol. 93 (5) (2009) 577–581.
[6] G. Voidonikolas, M.-C. Gingras, S. Hodges, A.L. McGuire, C. Chen, R.A. Gibbs, F.C. Brunicardi, W.E. Fisher, Developing a tissue resource to characterize the genome of pancreatic cancer, World J. Surg. 33 (2009) 723–731.
[7] G. Olund, P. Lindqvist, J.-E. Litton, BIMS An information management system for biobanking in the 21st century, IBM Syst. J. 46 (1) (2007) 171–182.
[8] Z. Zimmerman, M. Swenson, B. Reeve, Biobanks: accelerating molecular medicine—challenges facing the global biobanking community, IDC Special Study 4296 (2004) 1–36.
[9] P.H. Watson, et al., Evolutionary concepts in biobanking—the BC BioLibrary, J. Trans. Med. 7 (2009) 95.
[10] M.M. Morente, P.L. Fernández, E.de Alava, Biobanking: old activity or young discipline? Semin. Diagn. Pathol. 25 (4) (2008) 317–322.
[11] I. Hirtzlin, et al., An empirical survey on biobanking of human genetic material and data in six EU countries, Eur. J. Hum. Genet. 11 (6) (2003) 475–488.
[12] D. Troyer, Biorepository standards and protocols for collecting, processing, and storing human tissues, Methods Mol. Biol. 441 (2008) 193–220.
[13] S.K. Mohanty, et al., The development and deployment of Common Data Elements for tissue banks for translational research in cancer—an emerging standard based approach for the Mesothelioma Virtual Tissue Bank, BMC Cancer 8 (2008) 91.
[14] W. Amin, et al., National Mesothelioma Virtual Bank: a standard based biospecimen and clinical data resource to enhance translational research, BMC Cancer 8 (236) (2008) e1–10.
[15] R. Ravid, Biobanks for biomarkers in neurological disorders: the Da Vinci bridge for optimal clinico-pathological connection, J. Neurol. Sci. 283 (1–2) (2009) 119–126.
[16] G. Davey Smith, S. Ebrahim, S. Lewis, A.L. Hansell, L.J. Palmer, P.R. Burton, Genetic epidemiology and public health: hope, hype, and future prospects, Lancet 366 (9495) (2005) 1484–1498.
[17] M.M. Morente, E. de Alava, P.L. Fernandez, Tumour banking: the Spanish design, Pathobiology 74 (4) (2007) 245–250.
[18] M. Yuille, et al., Biobanking for Europe, Brief. Bioinform. 9 (1) (2008) 14–24.
[19] P.H.J. Riegman, A.L. Bosch, O.T. Consortium, OECI TuBaFrost tumor biobanking, Tumori 94 (2) (2008) 160–163.
[20] B. Clément, G. Chêne, F. Degos, A national collection of liver tumours: lessons learnt from 6 years of biobanking in France, Cancer Lett. 286 (1) (2009) 140–144.
[21] R.F. Ozols, et al., Clinical cancer advances 2006: major research advances in cancer treatment, prevention, and screening—a report from the American Society of Clinical Oncology, J. Clin. Oncol. 25 (1) (2007) 146–162.
[22] M. Yuille, K. Dixon, A. Platt, S. Pullum, D. Lewis, A. Hall, W. Ollier, The UK DNA banking network: a "fair access" biobank, Cell Tissue Bank (2009).
[23] M. Asslaber, K. Zatloukal, Biobanks: transnational, European and global networks, Brief Funct. Genomic Proteomic 6 (3) (2007) 193–201.
[24] C.E. Teunissen, et al., A consensus protocol for the standardization of cerebrospinal fluid collection and biobanking, Neurology 73 (22) (2009) 1914–1922.
[25] M. Ponzoni, I. Kwee, L. Mazzucchelli, A.J.M. Ferreri, E. Zucca, C. Doglioni, F. Cavalli, F. Bertoni, A virtual tissue bank for primary central nervous system lymphomas in immunocompetent individuals, Pathobiology 74 (4) (2007) 264–269.
[26] J. Muilu, L. Peltonen, J.-E. Litton, The federated database—a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe, Eur. J. Hum. Genet. 15 (2007) 718–723.
[27] Beale, T., Archetypes constraint-based domain models for future-proof information systems, 2000, available at http://www.openehr.org/publications/archetypes/archetypes_beale_web_2000.pdf, last accessed April 2010.
[28] T. Beale, in: K. Backlawski, H. Kilov (Eds.), Archetypes: Constraint-based Domain Models for Future-proof Informatin Systems in Eleventh OOPSLA 2002 Workshop on Behavioural Semantics: Serving the Customer, Northeastern Univsersity, Boston/Seattle/Washington, USA, 2002, pp. 16–32.
[29] T. Beale, S. Heard, openEHR Architecture: Architecture Overview, 2007, available at http://www.openehr.org/svn/specification/TRUNK/publishing/architecture/overview.pdf, last accessed April 2010.
[30] T. Beale, S. Heard, Archetype Definitions and Principles, 2007, available at http://www.openehr.org/svn/specification/TRUNK/publishing/architecture/am/archetype_principles.pdf, last accessed April 2010.
[31] CEN/TC251 (European Committee for Standardization), EN13606-1:2006 Health informatics — Electronic health record communication — Part 1: Reference model, 2006.
[32] D. Kalra, Electronic health record standards, in: IMIA Yearbook of Medical Informatics, 2006, pp. 136–144.

[33] K. Smith, D. Kalra, Electronic health records in complementary and alternative medicine, Int. J. Med. Inform. 77 (9) (2008) 576–588.
[34] T. Beale, The openEHR Archetype Model: Archetype Object Model. (2.0.2), 2008, pp. 1–54.
[35] S. Heard, T. Beale, G. Freriks, A.R. Mori, O. Pishev, Templates and Archetypes: how do we know what we are talking about?, 2003, available at http://www.openehr.org/publications/archetypes/templates_and_archetypes_heard_et_al.pdf, last accessed April 2010.
[36] NHS, NHS Connecting For Health release report on openEHR Clinical Modelling Pilot, 2006. Available at http://www.ehr.chime.ucl.ac.uk/download/attachments/3833859/NHSCFH_13606-Pilot-Final-Rpt_v1-0.pdf, last accessed April 2010.
[37] Leslie, H., International developments in openEHR archetypes and templates, Health Information Management Journal, 2008. Available at http://www.thefreelibrary.com/International+developments+in+openEHR+archetypes+and+templates.-a0175874267, last accessed April 2010.
[38] L. Bird, A. Goodchild, Z.Z. Tun, Experiences with a Two-Level Modelling Approach to Electronic Health Records, J.Res. Pract. Inform. Technol. 35 (2) (2003) 121–138.
[39] S. Garde, E.J.S. Hovenga, J. Gränz, S. Foozonkhah, S. Heard, Managing archetypes for sustainable and semantically interoperable electronic health records, Electronic J. Health Inform. 2 (2) (2007) pe9.
[40] S. Garde, P. Knaup, E.J.S. Hovenga, S. Heard, Towards semantic interoperability for electronic health records: domain knowledge governance for openEHR archetypes, Methods Inform. Med. 46 (3) (2007) 332–343.
[41] P. Hurlen, K. Skifjeld, E.P. Andersen, The basic principles of the synapses federated healthcare record server, Int. J. Med. Inform. 52 (1998) 123–132.
[42] P. Lichtenstein, U. De Faire, B. Floderus, M. Svartengren, P. Svedberg, N.L. Pedersen, The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies, J. Intern. Med. 252 (3) (2002) 184–205.
[43] J. Buck, S. Garde, C.D. Kohl, P. Knaup-Gregori, Towards a comprehensive electronic patient record to support an innovative individual care concept for premature infants using the openEHR approach, Int. J. Med. Inform. **78** (2009) 521–531.
[44] D. Moner, J.A. Maldonado, D. Bosca, J.T. Fernández, C. Angulo, P. Crespo, P.J. Vivancos, M. Robles, Archetype-based semantic integration and standardization of clinical data, in: Conference proceedings of the 28th IEEE Engineering in Medicine and Biology Society Annual International Conference, New York City, USA, August 30–September 3, 2006, pp. 5141–5144.
[45] S. Garde, E.J.S. Hovenga, J. Buck, P. Knaup, Expressing clinical data sets with openEHR archetypes: A solid basis for ubiquitous computing, Int. J. Med. Inform. 76S (2007) S334–S341.
[46] NHS (NHS Connecting for Health), Investigating implementing CEN 13606 with HL7 V3 and SNOMED CT–Final Report, 2006, available at http://detailedclinicalmodels.org/documents/NHS_CFH_13606InvestigationRpt_v1-0.pdf, last accessed April 2010.
[47] C.D. Kohl, S. Garde, P. Knaup, Facilitating the openEHR approach—organizational structures for defining high–quality archetypes, in: EHealth Beyond the Horizon: Get It There: Proceedings of MIE2008 the XXIst International Congress of the European Federation for Medical Informatics, In: Studies in Health Technology and Informatics, vol. 136, 2008, pp. 437–442.
[48] S. Garde, E. Hovenga, J. Gränz, S. Foozonkhah, S. Heard, Managing archetypes for sustainable and semantically interoperable electronic health records, Electronic J. Health Inform. 2 (2) (2007) e3–e12.
[49] J.T. Fernandez-Breis, M. Menarguez-Tortosa, C. Martinez-Costa, E. Fernandez-Breis, J. Herrero-Sempere, D. Moner, J. Sanchez, V.-G. Rafael, M. Robles, A Semantic web-based system for managing clinical archetypes, in: Conference Proceedings: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2008, pp. 1482–1485.
[50] T. Beale, S. Heard, The openEHR Archetype Model: Archetype Definition Language, ADL 1.4, 2008, available at http://www.openehr.org/releases/1.0.1/architecture/am/adl.pdf, last accessed April 2010.
[51] T. Beale, S. Heard, The openEHR Archetype Model: Archetype Definition Language, ADL 1.5, 2010, available at http://www.openehr.org/svn/specification/TRUNK/publishing/architecture/am/adl1.5.pdf, last accessed September 2010.
[52] E. Sundvall, R. Qamar, M. Nyström, M. Forss, H. Petersson, D. Karlsson, H. Åhlfeldt, A. Rector, Integration of tools for binding archetypes to SNOMED CT, BMC Medical Informatics and Decision Making 8 (Suppl. 1, S7) (2008) 1–10.
[53] S.T. Rosenbloom, R.A. Miller, K.B. Johnson, P.L. Elkin, S.H. Brown, Interface terminologies: facilitating direct entry of clinical data into electronic health record systems, J. Am. Med. Inform. Assoc. 13 (3) (2006) 277–288.
[54] R. Cornet, Definitions and qualifiers in SNOMED CT, Methods Inform. Med. 48 (2) (2009) 178–183.
[55] R.H. Dolin, K.A. Spackman, D. Markwell, Selective retrieval of pre- and post-coordinated SNOMED concepts, Proc. AMIA Symp. (2002) 210–214.
[56] D. Dalan, Clinical data mining and research in the allergy office, Curr. Opin. Allergy Clin. Immunol. 10 (3) (2010) 171–177.
[57] S. Hyun, S.B. Johnson, S. Bakken, Exploring the ability of natural language processing to extract data from nursing narratives, Comput. Inform. Nurs. 27 (4) (2009) 215–223, quiz 224–225.
[58] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, C.G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (5) (2010) 507–513.
[59] K.J. Kristianson, H. Ljunggren, L.L. Gustafsson, Data extraction from a semi-structured electronic medical record system for outpatients: a model to facilitate the access and use of data for quality control and research, Health Inform. J. 15 (4) (2009) 305–319.
[60] S.B. Johnson, et al., An electronic health record based on structured narrative, J. Am. Med. Inform. Assoc. 15 (1) (2008) 54–64.
[61] I. Spasic, F. Sarafraz, J.A. Keane, G. Nenadic, Medication information extraction with linguistic pattern matching and semantic rules, J. Am. Med. Inform. Assoc. 17 (5) (2010) 532–535.
[62] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb. Med. Inform. (2008) 128–144.
[63] J. Park, S. Ram, Information systems interoperability: what lies beneath? ACM Trans. Inform. Syst. 22 (4) (2010) 1595–1632.
[64] R. Chen, G.O. Klein, E. Sundvall, D. Karlsson, H. Ahlfeldt, Archetype-based conversion of EHR content models: pilot experience with a regional EHR system, BMC Med. Inform. Decision Making 9 (33) (2009) e1–13.

[65] C. Daniel, M. García Rojo, K. Bourquard, D. Henin, T. Schrader, V. Della Mea, J. Gilbertson, B.A. Beckwith, Standards to support information systems integration in anatomic pathology, Arch. Pathol. Lab. Med. 133 (11) (2009) 1841–1849.
[66] Specimen Domain, HL7 Version 3 Standard: Specimen, Release 1, available at http://www.hl7.org/v3ballot2010MAY/html/domains/uvsp/uvsp.htm, last accessed September 2010.
[67] CMETS defined by Domain: Specimen Domain, available at http://www.hl7.org/v3ballot2010MAY/html/domains/uvsp/UVSP_do_cmets.htm#, last accessed September 2010.