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].