INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 78 (2009) 559–570

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

LinkEHR-Ed: A multi-reference model archetype editor based on formal semantics

José A. Maldonado[a,*], *David Moner*[a], *Diego Boscá*[a], *Jesualdo T. Fernández-Breis*[b], *Carlos Angulo*[a], *Montserrat Robles*[a]

[a] *Grupo de Informática Biomédica (IBIME), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia, Valencia, Spain*
[b] *Departamento de Informática y sistemas, Universidad de Murcia, Murcia, Spain*

ARTICLE INFO

Article history:
Received 30 May 2008
Received in revised form
9 March 2009
Accepted 15 March 2009

Keywords:
Electronic health records
Biomedical informatics
Standardization
Archetypes
Semantic description
Semantic interoperability

ABSTRACT

Purpose: To develop a powerful archetype editing framework capable of handling multiple reference models and oriented towards the semantic description and standardization of legacy data.
Methods: The main prerequisite for implementing tools providing enhanced support for archetypes is the clear specification of archetype semantics. We propose a formalization of the definition section of archetypes based on types over tree-structured data. It covers the specialization of archetypes, the relationship between reference models and archetypes and conformance of data instances to archetypes.
Results: LinkEHR-Ed, a visual archetype editor based on the former formalization with advanced processing capabilities that supports multiple reference models, the editing and semantic validation of archetypes, the specification of mappings to data sources, and the automatic generation of data transformation scripts, is developed.
Conclusions: LinkEHR-Ed is a useful tool for building, processing and validating archetypes based on any reference model.

1. Introduction

Health care is a sector where the need to share information is the norm rather than the exception. However, the information about the health of one person is usually scattered among all the health facilities where he or she has ever been attended. This leads to distributed- and heterogeneous data resources and makes the exchange of data across systems very difficult. This situation has created a large gap between the potential and actual value of the information contents of health data repositories.

Due to the special sensitivity of health data and the wide range of ethical and legal constraints, the exchange of data must be done in a meaningful way, avoiding the possibility of misunderstanding or misinterpretation. The faithful communication and multiple usability of electronic health records (EHR) crucially depend on the standardization of its syntax, structure and semantics, i.e. on the standardization of the EHR architecture and vocabulary used to communicate data [1].

Currently there are several international organizations working on the definition of an EHR architecture [2,3]. Health Level 7 (HL7) [4] maintains the XML-based clinical document

* *Corresponding author at*: Grupo de Informática Biomédica (IBIME), Instituto ITACA, Universidad Politécnica de Valencia, Camino de Vera s/n, Edificio 8G, Acceso B, Nivel 3, 46022 Valencia, Spain. Fax: +34 963877279.
E-mail address: jamaldo@upv.es (J.A. Maldonado).
1386-5056/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2009.03.006

architecture (CDA) [5] that specifies the structure and semantics of clinical documents for exchange. The Technical Committee 251 (health informatics) of the European Committee for Standardization (CEN/TC251) has completed a European Standard for the communication of the EHR called EN13606 [6] whose part 1 (reference model) [7] became an ISO standard in February 2008. The openEHR consortium [8] maintains architecture designed to support the constructions of distributed, patient-centered, life-long, shared care health records. Finally, ISO provides a set of clinical and technical requirements for an EHR architecture that support using, sharing and exchanging EHRs in the technical specification TS 18308:2004 [9].

OpenEHR and EN13606 utilize the dual model architecture to describe the EHR structure. The dual model or two-level methodology is a novel approach for the development of EHR systems. It aims to overcome the disadvantages of traditional "one-model" methodologies in which domain concepts are hard-coded directly into database models and software, when applied to complex and changing domains such as medicine. As stated by Rector [10], medicine is not only big but also open ended in breadth, depth and complexity. As a consequence health information systems that follow these methodologies need frequent, complex and expensive updates to accommodate new requirements which if not made, cause the systems to suffer "creeping obsolescence" and as result becomes obsolete. The dual model methodology distinguishes a reference model and archetypes [11]. In a broad sense, a reference model is an abstract representation of the entities and relationships of a domain which is designed to provide a basis for the development of more concrete models and implementations. The generality of the reference model is complemented by the particularity of archetypes [12,13]. Archetypes are formal definitions of clinical concepts in the form of structured and constrained combinations of the entities of the reference model.

The work reported in this paper is part of a bigger project called LinkEHR whose main objective is the utilization of the dual model methodology for semantic integration and normalization of health data. In LinkEHR we use archetypes to describe the semantics of legacy health data in a manner independent of the particular data organization in the underlying data repositories. This will enable users (mainly health professionals) to view and query data repositories at the level of its relevant semantic concepts. In this paper we describe the archetype editing framework of LinkEHR that includes a formalization of the definition section of archetypes and an archetype editor, called LinkEHR-Ed, which supports multiple reference models. The remainder of this paper is organized as follows. Section 2 provides an overview of the dual model approach. Section 3 deals with the formalization of the definition section of archetypes. Section 4 describes LinkEHR-Ed, a visual multi-reference model archetype editor based on the proposed formalization. Finally, Section 5 discusses and concludes the paper.

2. Background

The most remarkable feature of the dual model approach is the complete separation of information models, such as software models or database schemas, from domain models such as blood pressure measurement, discharge report, prescription or microbiology result. The information model is represented by a stable and small object oriented reference model that models the generic and stable properties of health record information. Only this model is hard-coded in database schemas or software. The possible numerous and volatile domain concepts, represented as archetypes, are modeled separately by domain specialists and their definitions are maintained in a shared repository [14]. Since the software is only bound to the reference model it has no direct dependency on domain concepts. Therefore, when new concepts are introduced or existing ones are altered the system does not need to be updated.

Archetypes are formal definitions of a distinct domain-level concept in the form of structured and constrained combinations of reference model classes. Their principal purpose is to provide a powerful, reusable and interoperable way of managing the creation, description, validation and query of EHRs. From a data point of view, archetypes are a means for providing semantics to data instances that conform to some reference model by assuring that data obey a particular structure and satisfy a set of constraints. The semantic description of domain concepts is achieved by linking the data structures and content to knowledge resources such as terminologies and ontologies [15].

Only those classes of the reference model that define logical building blocks of EHRs can be used to construct archetypes, we call them business concepts. For instance ISO/CEN EN13606 defines six business concepts, namely: Folder, Composition, Section, Entry, Cluster and Element. What is important here is that for each domain concept, a definition can be developed in terms of constraints on structure, types, values, and behaviors of business concepts. Archetypes may be specialized: an archetype is considered a specialization of another archetype if it specifies that archetype as its parent, and only makes changes to its definition such that its constraints are narrower than those of the parent.

Archetype Definition Language (ADL) [12] is a formal language developed by openEHR for expressing textually archetypes that has also been adopted by EN13606. The most important section of an archetype is its definition tree, where clinical concepts are defined by constraining a particular business class in several different ways. For a thorough explanation of ADL we refer the reader to [12,13]. We also refer the reader to ref. [16], an archetype repository with advanced search capabilities.

3. Methods

Although ADL was designed to be a formal language, the current ADL specification is not precise enough regarding some important issues such as archetype specialization and semantics. The main purpose of archetypes is to provide a mechanism to describe, validate, store and query health data. Since the data instances defined by archetypes are also instances of the underlying reference model, understanding the relationship between reference models and archetypes becomes crucial in order to manipulate archetyped data. We