Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Archetype-Based Semantic Integration and Standardization of Clinical Data

David Moner, Jose A. Maldonado, Diego Bosca, Jesualdo T. Fernández, Carlos Angulo, Pere Crespo, Pedro J. Vivancos, Montserrat Robles

Abstract—**One of the basic needs for any healthcare professional is to be able to access to clinical information of patients in an understandable and normalized way. The life-long clinical information of any person supported by electronic means configures his/her Electronic Health Record (EHR). This information is usually distributed among several independent and heterogeneous systems that may be syntactically or semantically incompatible. The Dual Model architecture has appeared as a new proposal for maintaining a homogeneous representation of the EHR with a clear separation between information and knowledge. Information is represented by a Reference Model which describes common data structures with minimal semantics. Knowledge is specified by archetypes, which are formal representations of clinical concepts built upon a particular Reference Model. This kind of architecture is originally thought for implantation of new clinical information systems, but archetypes can be also used for integrating data of existing and not normalized systems, adding at the same time a semantic meaning to the integrated data. In this paper we explain the possible use of a Dual Model approach for semantic integration and standardization of heterogeneous clinical data sources and present LinkEHR-Ed, a tool for developing archetypes as elements for integration purposes. LinkEHR-Ed has been designed to be easily used by the two main participants of the creation process of archetypes for clinical data integration: the Health domain expert and the Information Technologies domain expert.**

I. Introduction

EXCHANGE of healthcare information among health professionals or clinical information systems is nowadays a critical process for the healthcare sector. Due to the special sensitivity of medical data and its ethical and legal constraints, this exchange must be done in a meaningful way, avoiding all possibility of misunderstanding or misinterpretation. Two main problems arise when pursuing that objective.

On the one hand, in many hospitals there is not a unified information system available. Health data is distributed across several heterogeneous and autonomous systems whose interconnection and integration is difficult to achieve. This may be resolved by setting up a new and integrated information system for all the organization but it would represent a great economic cost, a traumatic upgrade of existing applications and a difficult adaptation of current users to the new system. Other solution is to use a federated information system strategy that links the heterogeneous data sources offering a unified view as a virtual centralized repository [1].

On the other hand, a clinical information system may lack of a comprehensive semantic definition of the information which it contains, up to the point of making impossible a semantic interoperability between different systems. Due to the complexity of the health domain this has not an easy solution. There are O(100.000) – O(1.000.000) terms and probably 200.000 – 300.000 concepts and they are never likely to be completely defined [2]. This implies that a traditional information model will never be completely adapted to the clinical requirements and their continuous evolution.

II. The Dual Model Architecture Approach

The Dual Model architecture [2] intends to solve these problems. It defines a clear separation between information and knowledge. The former is structured through a Reference Model that contains the basic entities for representing any information of the EHR. The latter is based on archetypes, which are formal definitions of clinical concepts, such as discharge report, glucose measurement or family history, in the form of structured and constrained combinations of the entities of a Reference Model. It provides a semantic meaning to a Reference Model structure. Examples of Dual Model architectures are CEN/TC251 EN13606 [3] and openEHR [4]. Although HL7 v3 [5] can not be considered a true Dual Model standard, it uses a similar approach.

A. Reference Model

A Reference Model is an Object Oriented model that is used to represent the generic and stable properties of health record information. It comprises a small set of classes that define the generic building blocks to construct EHRs. It specifies how health data should be aggregated to create more complex data structures and the context information that must accompany every piece of data in order to meet ethical and legal requirements. It does encode what it is

This work was supported in part by the Spanish Ministry of Education and Science under Grant TSI2004-06475-C02 and by the Spanish Ministry of Health under Grant PI050815.

D. Moner, J. A. Maldonado, D. Bosca, C. Angulo, P. Crespo and M. Robles are with the Biomedical Informatics Group, Technical University of Valencia, Valencia, Spain (e-mail: {damoca, jamaldo, diebosto, cangulo, pedcremo, mrobles}@upv.es).

J. T. Fernández and P. J. Vivancos are with the Departamento de Informática y Sistemas, Universidad de Murcia, Murcia, Spain (e-mail: jfernand@dif.um.es, pedroviv@um.es).

1-4244-0033-3/06/$20.00 ©2006 IEEE.

meant, not how it is intended to be presented. Typically, a Reference Model contains:

1) A set of primitive types.
2) A set of classes that define the building blocks of EHRs. Any annotation in an EHR must be an instance of one of these classes; we will call them entities. For instance EN13606 defines six different types of entities: folder, composition, section, entry, cluster and element.
3) A set of auxiliary classes that describe the context information to be attached to an EHR annotation including versioning information.
4) It may contain classes to describe demographic data and to communicate EHR fragments.

B. *Archetype Model*

Archetypes [6] are composed of three main sections: header, definition and ontology.

The *header* section basically contains metadata about the archetype, such as an identifier or authoring information.

In the *definition* section is where the clinical concept which the archetype represents is described in terms of Reference Model entities. This description is built by constraining the entities in different ways:

1) Constraints on the range of attributes of primitive types.
2) Constraints on the existence of attributes, i.e. whether a value is mandatory for the attribute in run time data.
3) Constraints on the cardinality of attributes, i.e. whether the attribute is multi-valuate or not.
4) Constraints on the occurrences of objects indicating how may times in runtime data an instance of a given class conforming to a particular constraint can occur.
5) Constraints on complex objects. They can be stated by constraining their attributes or by reusing previously defined archetypes or archetype fragments.

Finally, the *ontology* section is where the entities defined in the definition section are described and bound to terminologies.

Archetype specialization is the mechanism that allows re-using an archetype definition. This is achieved by providing further constraint on information already expressed by other archetype. There exists an underlying specialization hierarchy behind every archetype whose root is a Reference Model entity. Below it hangs up a parent-children succession of archetypes. The deeper the level of the hierarchy is, the more constrained or specialized the archetype is. Data created as an instance of a specialized archetype is also an instance of more general or parent archetypes and at the same time is compatible with the root Reference Model entity that we are archetyping.

III. Archetypes For Semantic Integration And Standardization Of Clinical Data

The general case of use of a Dual Model architecture is the creation of new clinical information systems, where a complete set of archetypes can be defined to satisfy the needs of the organization. All data would be stored in compatible or optimized repositories for a particular Reference Model. Thus, it is easy to recover the needed data in order to create an EHR extract following the structure and constraints of those previously defined archetypes.

It may occur that we want to adapt an existing system in production with its own data structure (also known as legacy systems) to a Dual Model architecture in order to provide a semantic layer that describes formally the stored information. This semantic layer will be also useful for publishing our legacy data in the form of EHR extracts compliant with some EHR architecture standard. It is feasible to achieve interoperability between standards [7]. Therefore, if we are able to standardize our EHRs it is possible to make public our data conforming to more than one standard.

We rely on archetypes to accomplish standardization and integration of clinical information. Mapping the structure of an archetype definition to the elements of the data sources from where related information can be extracted convert those archetypes into a mechanism of data integration, giving a semantic meaning to that linked data at the same time. This will allow, for example, using the semantics described by archetypes to do flexible query processing instead of taking a simple syntactical approach such as checking for the presence of a set of terms. In that case, if terms are not present in the EHR extract, it would not be retrieved even thought it may be relevant.

A. *Mapping Specification*

Since the health data to be made public resides in the underlying data sources, it is necessary to define some kind of mapping information that links entities described in the archetype (classes and attributes) to data elements in data repositories (e.g. elements and attributes in the case of XML documents, tables and attributes in the case of relational data sources). An integration archetype is considered to be a view that provides abstraction in interfacing between the data sources and the Reference Model used to communicate the EHR extracts. Since all three EHR standards (CEN/TC251 EN13606, openEHR and HL7 v3) consider that EHR extracts have an inherent hierarchical structure, we have chosen XML as canonical data model, i.e. EHR extracts and the content of data sources are viewed as XML documents. Furthermore, we have developed a type system based on [8] capable of representing all the constraints that can be used when defining an archetype, the specialization relationships between an archetype and the Reference Model, and between a pair of archetypes [9]. This type system allows the definition of a formal framework for checking the semantic validity of archetypes.

There exist two kind of mapping specifications: atomic attribute mappings and object mapping. Atomic attribute mappings define how to obtain a value for an atomic attribute of an archetype by using a set of values from the data sources. For this purpose transformation functions and conversion tables can be used. This kind of mapping is mandatory for each atomic attribute that must have a value in the data instances. For each constrained class there exist an object mapping which contains both the query to be used to retrieve all the data necessary for generating data instances and the set of attributes that identify univocally the instances of the class. The combination of both components allows the conversion from source data to XML documents compliant with the Reference Model. The query extracts the relevant information and for each different combination of values of the identification attributes a new instance of the class is generated.

Archetype designers are responsible of defining the atomic attribute mapping and the system tries to generate automatically [10] from them a set of candidate object mappings by taking into account the structure of the Reference Model entity, the constraints defined in the archetype and the integrity constraints of data sources. This approach alleviates the work of defining how to populate archetypes since it is easier for the designer to indicate which data elements of the data sources are relevant to a certain archetype attribute, rather than to specify the possible complex query required to extract and transform all the relevant information, which may involve many data structures possibly from several data sources.

A. Authoring

During the creation of integration archetypes two different actors should collaborate: the Health domain expert who knows the abstract clinical concept meaning and the constraints and elements of the Reference Model suitable to represent it, and the Information Technology expert who knows the structure of the non-standard data sources and who is able to define the correct mappings between the structure of the archetype and the data sources.

There exist tools to edit and create archetypes [11] but they do not support the integration archetype concept we have just presented. We have developed an editor to help carrying out this duty.

II. LinkEHR-Ed: A Tool For Developing Integration Archetypes

LinkEHR-Ed is a tool that brings together the different needs of Health domain experts and Information Technologies experts during the development of archetypes for clinical data integration, narrowing the gap that exists between the two knowledge domains. LinkEHR-Ed is a fundamental part of LinkEHR, an integrated system under development for the publication and intelligent access to existing health data of patients based on an EHR Dual Model architecture and formal models for the description of clinical meaning, both for health care and investigation purposes. LinkEHR is an extension of the Pangea system [12]. A conceptual model of archetypes for clinical data integration can be found in Figure 1.

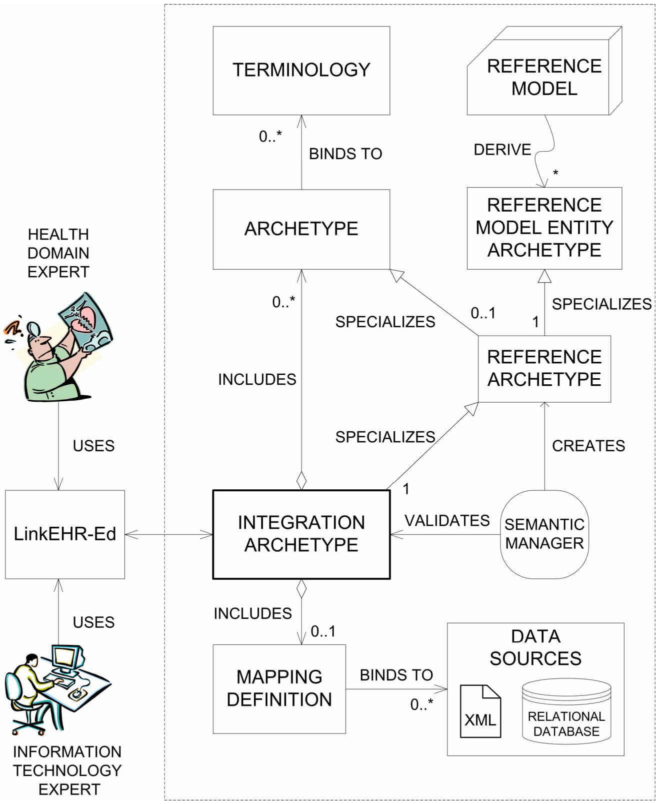

Fig. 1. Conceptual model of archetypes for clinical data integration.

A. Reference Model Independence

LinkEHR-Ed can work with several Reference Models. As explained before, we can consider the Reference Model entities as the most general or root archetypes of any specialization hierarchy. Then, constraining a Reference Model entity follows the same logic that is applied to any other archetype specialization, giving consistency to the full definition process. At the same time, it is assured that the created archetype is semantically valid with respect to the chosen Reference Model entity and will create valid data instances of it. An XML Schema defining a Reference Model can be imported into LinkEHR-Ed. This will generate archetype representations of non abstract Reference Model entities that will be immediately available as basis for a new archetype specialization.

B. Archetype Persistence

Archetype persistence is achieved using an *Archetype Definition Language* (ADL) representation [13]. ADL is a language expressly created for a textual representation of archetypes. The defined mappings can be represented in the ADL code adding a new *mapping section* to it.

C. Semantic Manager

Automatic validation of the developed archetype from a semantic point of view is done through a semantic management module. This module controls that the new constraint definitions are valid, i.e. narrower with respect to the parent archetype and the root Reference Model entity.

By definition, an archetype only contains those nodes and attributes of the Reference Model which have been constrained. The ones which are unconstrained are not explicitly present but they must be also validated and mapped to data sources if necessary. For doing that validation, the semantic manager generates a virtual *reference archetype*, which is created by merging the root Reference Model entity archetype with the parent of current archetype, if it exists. The reference archetype contains all the semantic constraints that the new developed archetype must fulfil.

D. *LinkEHR-Ed Interface*

LinkEHR-Ed provides two different interfaces or perspectives, one for the Health domain experts and one for the Information Technologies professionals.

On the one side, Health domain experts will be in charge of generating the archetype structure and definition tree and so they must have some knowledge about the EHR Reference Model they are working with. But the main idea for this perspective is to hide the underlying complexity of the system and the Dual Model architecture logic involved in the designing of an archetype as we can not presuppose any computer management skills for this expert. A set of available constraints and applicable restrictions is provided during design time.

On the other side is the Information Technologies experts, who knows the structure of data sources of the organization and his role is to map the archetype definition tree nodes to them. A mapping definition interface fills nearly all this edition perspective. It is composed by a graphical representation of the archetype definition tree and a graphical representation of the diverse data sources available. Users can then add or modify mapping transformations between elements of both representations.

III. Future Improvements

Some features of LinkEHR-Ed can be improved in terms of usability and functionality.

Health domain expert perspective may be enhanced providing a higher abstraction of the archetype structure. This can be achieved defining an ontology layer for the archetype and Reference Model entities. If the ontology layer is abstract enough it could be also used for a semiautomatic interoperability between archetypes based on different Reference Models.

A standard definition for an archetype repository service is needed if we want to take full advantage of the Dual Model approach.

Access to demographics servers would be a useful tool for defining high quality archetypes since their definition trees can include references to demographic entities identifiers which perform a role in a health care activity.

Finally, instantiation of EHR extracts based on the edited archetype may be a good practice for validating them before publishing.

IV. Conclusion

Possibly, a Dual Model Architecture approach is nowadays one of the best options for building new clinical information system due to its capability for clinical information interchange. Archetypes give us a semantic layer for common understanding and mutual communication of clinical data structured as a formal clinical concept definition decided by Health domain experts, achieving at the same time an automatic semantic interoperability among clinical Information Systems. But archetypes are also a valid approach for upgrading already deployed systems in order to make them compatible with an EHR standard, considering the archetypes as clinical data integration components. The benefit of this approach is to maintain in-production systems and applications without any changes while providing a mean for extracting clinical information from those systems in the form of standardized EHR extracts, hiding technical details, location and heterogeneity of the data repositories. At the same time, they constitute a semantic layer over the underlying databases associating them with domain specific semantics. Thus, it is possible to combine in an easy manner the formal representation of knowledge of a Health domain expert, represented by an archetype, with the mapping information to data sources where clinical data is stored and use them together for semantic integration and standardization purposes.

References

[1] W. Sujansky, “Heterogeneous database integration in biomedicine,” *Journal of Biomedical Informatics*, vol. 34, no. 4, pp. 285-298, 2001.

[2] T. Beale (2001, Aug 21). “Archetypes, Constraint-based Domain Models for Future-proof Information Systems”. Available: http://www.deepthought.com.au/it/archetypes/archetypes.pdf

[3] http://www.centc251.org

[4] http://www.openehr.org

[5] http://www.hl7.org

[6] T. Beale and S. Heard (2003, Dec 20). “The openEHR Archetype System”. Available: http://www.openehr.org/

[7] V. Bicer, G. B. Laleci, A. Dogac, Y. Kabak, “Artemis message framework: Semantic interoperability of exchanged messages in the healthcare domain,” *Sigmod Record*, vol. 34, no. 3, pp. 71-76, 2005.

[8] G.M. Kuper and J. Simeon, “Subsumption for XML types,” in Proc. *8th International Conference on Database Theory (ICDT'01)*, Heidelberg, 2001, pp. 331-345.

[9] J. A. Maldonado, “Historia Clínica Electrónica Federada Basada en la Norma Europea CEN/TC251 EN13606,” PhD. Dissertation (in Spanish), Technical University of Valencia, 2005.

[10] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernández and R. Fagin, “Translating web data,” in Proc. *28th VLDB Conference*, Hong Kong, 2002, pp. 598-609.

[11] http://www.oceaninformatics.biz

[12] J. A. Maldonado, M. Robles and P. Crespo, “Integration of distributed healthcare records: publishing legacy data as XML documents compliant with CEN/TC251 ENV13606,” in Proc. *16th IEEE Symposium on Computer Based Medical Systems*, New York, 2003, pp. 213-218.

[13] OpenEHR. (2006, March). “The Archetype Definition Language Version 2 (ADL2)”. Available: http://www.openehr.org