Integration of Distributed Healthcare Records: Publishing Legacy Data as XML Documents Compliant with CEN/TC251 ENV13606

J.A. Maldonado, M. Robles, P. Crespo
Bioengineering, Electronics and Telemedicine Group
E.U.I., Technical University of Valencia, Valencia 46022, Spain.
jamaldo@upvnet.upv.es

Abstract

To support co-operative work among health professionals and institutions it is necessary to share healthcare information about patients in a meaningful way. But, nowadays, in most hospitals health data are distributed across several information systems whose interconnection is difficult to achieve, this leads to the so-called islands of information. This paper briefly describes the architecture, design and implementation of the PANGEA system. PANGEA allows healthcare professionals to access patient information stored in heterogeneous autonomous information systems through a set of formal aggregates of health data based on the European pre-standard of Healthcare Record Architecture ENV13606 from CEN/TC251. ENV13606 is also used as canonical model for the representation of healthcare information, therefore the overall system can also be considered as a system for publishing legacy relational data as XML-Electronic Healthcare Records compliant with ENV13606.

1. Introduction

Integration is still central for health information systems. Most modern hospitals have computerized records. However, these systems are usually proprietary and often only serve one specific department within the hospital. Hospitals may have dozens of individual systems that do not interoperate with each other. This leads to fragmented and heterogeneous data resources and services, which contain health data about patients, and adds to the creation of the so-called *islands of information*. The best-of-breed approach can be very suitable for large organizations letting departments meet their custom business needs more easily and allowing a greater flexibility within decentralized organizations such as hospitals. Possibly, each departmental system does a good job in meeting the department's needs of information. The challenge is finding how these systems can efficiently and meaningfully exchange health information about patients with one another.

Nowadays, the healthcare sector is undergoing a change. One new context in which a team of healthcare professionals from different disciplines and institutions is responsible for patient health is replacing the traditional single doctor-patient relationship. This new context requires a high level of interoperability and data sharing among professionals and institutions involved in the healthcare of a patient. This crucially depends on the ability to exchange information about patients while preserving its original meaning. In the absence of this information, tests may be repeated or prior findings ignored, and in emergencies lifesaving information may be unavailable. Briefly, what is required is that everyone

involved in the delivery of healthcare to a patient should be able to access all the relevant patient's healthcare information.

2. Electronic Healthcare Record Architecture

Basically, an Electronic Healthcare Record Architecture (EHCRA) is an information model or framework for the construction of electronic healthcare records. It models the generic features necessary in any electronic healthcare record for it to be communicable, complete, a useful and effective ethic-legal record of care, and may retain integrity across systems, countries, and time. It neither does it prescribe or dictate what should be stored in a healthcare record, nor how any electronic healthcare record system has to be implemented [1].

Much work has been done in the field of EHCRA. In Europe, Work group I of CEN/TC251 (European Committee of Normalization, Technical Committee 251), has developed a European pre-standard known as ENV 13606 [1]. ENV13606 is divided into four parts.

- ENV13606-1-Extended Architecture. It defines a conceptual information model that is capable of structuring any medical data in a uniform way, presenting the multitude of different facts while preserving meaning and context of the data.
- ENV13606-2-Domain termlist. It provides tables of suitable names for categorizing Record Components, thus making it more possible to carry out electronic processing of the record.
- ENV13606-3-Distribution rules. It defines measures for security and access to the EHCR intended by the author.
- ENV13606-4-Messages for the exchange of information. It defines a number of messages that support the exchange of data between systems (source and destination). This part also offers a set of DTDs for its message definitions.

The key concept upon which ENV13606 has been defined is interoperability. An EHCR is interoperable, if:

- The communication of it or parts of it is performed in such a way that the transferred information can be rendered human-readable by the receiving system. Furthermore, if the exchange of electronic healthcare record information is to support individual patient care, clinicians using the receiving information are able to read and understand it.
- The transferred information can be incorporated into a record held by the receiving system in a way that enables it to be processed and retrieved as an integral part of that record.

Standardization of Electronic Healthcare Record Architecture (EHCRA) is vital if the clinical information has to be transferred outside the organization or department where it was created but most healthcare data continue to be stored in relational databases system. It is not probable that this situation will change in the foreseeable future due to the high reliability and performance of relational databases. Consequently, some sort of semi-automatic mechanism is needed to publish legacy relational data in the form of XML documents compliant with ENV13606. PANGEA also addresses this issue.

3. Data Engineering

3.1. Archetypes

The use of ENV13606 in our system is based on the concept of archetype. An arquetype is a definition of an information structure used in a certain domain that is based on a reference model. In our case, this reference model is ENV13606. Two other similar experiences in the use of archetypes to model EHCR can be found in the literature [2][3]. Since the components of ENV13606 have been defined at a high level of abstraction they gives us a flexible model able to represent any entry in a healthcare record. Thus, they can be easily used to represent terms or concepts from the medical domain such as a discharge report, GP record, patient's demographics data, blood pressure, protein S level, etc. The archetypes constitute the core of our integration solution; their purpose is to make public the information stored in the underlying databases and, at the same time, to hide technical details, location and heterogeneity of the data repositories. They constitute a semantic layer over the underlying databases associating them with domain specific semantics. Specialist in the domain should define archetypes, for instance pathologists may define arquetypes to represent biochemical results but a college of general practitioners might define one for physical examinations.

3.2. Linking archetypes to data

Since the health data resides in the underlying databases, some kind of mapping information relating archetypes to database schemas should be defined. In database theory, views provide a user-defined subset of a large database. Thus, an archetype can also be considered as a view that provides sharing and abstraction in interfacing between the relational model and XML documents. We also want to use the views for instantiating extracts of EHCR that are compliant with ENV13606 from the data stored in the underlying database. To achieve this, and due to the heterogeneity between the relational model and XML model, views should provide apart from a set of queries, a mapping between the archetype and the underlying database schemas. Such a mapping is not a trivial task, because the two data models differ significantly. Relational data is flat, normalized into many relations, and its schema is often proprietary. By contrast, XML data is nested, unnormalized, and its schema is public [4], for example, the one proposed by ENV13606 for EHCR. Publishing XML data involves joining tables, selecting and projecting the data that needs to be exported, mapping the relational attributes names into XML elements and attribute and finally creating XML hierarchies.

In PANGEA the mapping is done by linking archetype attributes to table fields through a set of attribute mapping functions. The archetype designer defines a set of mapping functions between archetype attributes and table fields. From this set and the implicit structure defined by the archetypes the system is capable of:

- finding a candidate query that will allow the population of the EHCR extract defined by the archetype
- structuring and tagging the resulting XML document according to ENV13606 rules.

This approach alleviates the work of defining arquetypes since it is easier for the designer to indicate which field is relevant to a certain arquetype attribute, rather than to

specify the possible complex query required to extract all the relevant information for the archetype, which may involve many relations possibly from several databases.

When generating a candidate query for an archetype some issues should be considered:

- We are trying to generate extracts of electronic healthcare records; therefore not all the information held by the data repositories is relevant. We are only interested in data for which we can determine the patient that the data is about.
- We should keep the existing relationships among the data, typically expressed by foreign keys.
- We should not lose information.
- The query should facilitate the future generation of XML documents.

The problem is related to that of computing the natural outerjoin of many relations in a way that preserves all possible connections among facts and how to simplify the resulting query according to a set of conditions and filters imposed by both archetypes and users. The full disjunction of a set of relations R, denoted by FD(R), is defined as the maximum information without redundancy that can be obtained from the relations in R [5]. As demonstrated in [6] the full disjunction is unique for a set of relations and generally can be computed as a single stream of full outerjoins. It is possible to simplify the full disjunction [7][8] if we take into account the condition imposed by the users, properties of foreign keys and archetypes and that not all the information held by the underlying data repositories is relevant, we are only interested in data directly or indirectly related to a patient. Generally, the resulting query is a single stream of left and inner joins that can by computed by most database engines.

The main requirement for publishing relational data as XML documents is the need for a specification that dictates how to perform the conversion from relational data to XML documents. The specification should describe how to structure and tag data from one or more table as a hierarchical XML document. In PANGEA the attribute mapping functions define this specification. Therefore, they are not only used to generate a potential query but also they are expressive enough to define how to structure and tag the resulting XML document. PANGEA compiles a mapping specification for each archetype in design time, when the archetype is instanciated for a particular patient the specification is run and the XML document compliant with ENV13606 is generated.

4. System architecture and implementation

PANGEA is a middleware between the application and databases making the former independent of the data sources, helping them to communicate in a more meaningful and efficient way. It allows the publishing of health data distributed among several departmental information systems as EHCR extracts compliant with part IV of ENV13606. The functionality needed to accomplish this is for example transformation and sub-setting of databases using view definitions, methods to access and merge data from multiple databases and support for abstraction and generalization of underlying data [9], thus PANGEA can be considered as a mediator in the sense of [9].

The basic architecture is illustrated in Figure 1. The main components of the system are:

- The metadata server is the module that is in charge of managing the system's data dictionary. It manages an object-oriented database that basically contains the archetypes definition, the underlying databases schemas, the archetypes-databases schemas mappings, information about the location of patients' social-demographic data and general information about the underlying databases. Two visual tools have been developed to assist in the creation and management of metadata: the archetype editor and the schemata manager. The former facilitates the edition of archetypes, it allows the creation of new ones from scratch or the re-use of the existing ones, it validates their correctness, it allows their classification into groups in order to make the search easier, define the mappings with the component databases schemas and it controls the versioning (all prior versions are kept for legal reasons). The latter allows the retrieval and caching of the underlying database schemas, the enrichment of the schemas by defining new relationships between attributes (foreign keys), define where the social-demographic data about the patients, such as name, surname, address, SSN, date of birth etc., are located in order to allow the matching of patient identifiers and define which data from the underlying database is shared.
- The Electronic Healthcare Record Server (EHR server) is the core of the whole system. It is layered between the client applications and the data repositories. This server retrieves, by request, all the relevant patient information wherever it is located and presents back the information as a XML document compliant with ENV13606. Client applications ask for health information about a particular patient as one or more instances of any archetype defined in the data dictionary. The EHR server obtains the definition and mapping specifications of the requested archetype from the metadata server. Afterwards, it builds and populates, by interpreting the mapping specifications, the XML documents that contain the healthcare record extract. The EHR server offers a set of web services that can be used by client applications. Basically, a web service is an interface that describes a collection of operations that are network accessible through standardized XML messaging. The protocol designed to manage the XML messaging is called SOAP (Simple Object Access Protocol). This protocol defines a standard structure, encoding rules and associations to transport XML documents through other protocols such as HTTP, SMTP, etc. SOAP allows a high level of interoperation between heterogeneous applications. Therefore, it suits perfectly to accomplish the desirable requirement that between-organization communications needs to be achieved, ideally using the same technological solution as for intra-organizational communication.

5. Conclusions

Healthcare is fast becoming more distributed in nature, thus the ability to share health data about patients effectively, meaningfully and securely is the key issue in providing good and cost-effective healthcare. The above outlined system model and architecture define a clinical data access system that provides a single point of entry for providing an integrated virtual view of distributed patient's healthcare records across an institution. The EHCR architecture used is ENV 13606 from CEN/TC 251, however some extension has been developed to cope with the problem of data distribution among several pre-existing information systems. Our solution is based on defining a set of formalized aggregates of

data with specific semantics and associating them with the heterogeneous structures found in the autonomous information systems. The system is in the line of the virtual approach and read-only view systems, i.e. systems that support read-only views of data held by multiple databases [10].

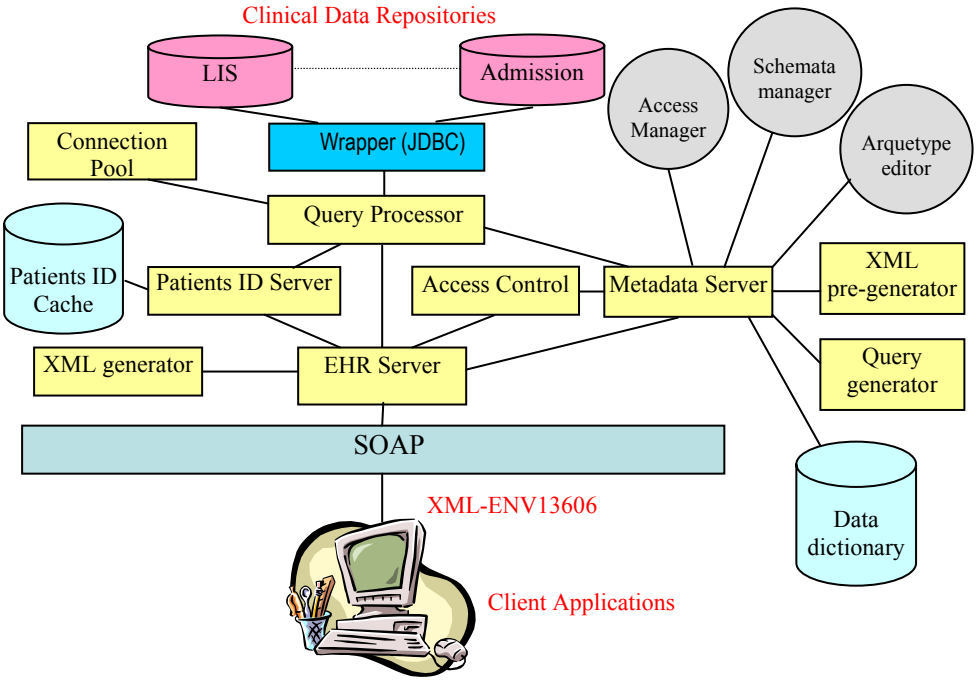

Figure 1. PANGEA architecture

6. References

[1] CEN/TC251 WG I.: Health Informatics-Electronic Healthcare Record Communication- Parts 1, 2, 3 y 4. Final Draft prENV13606, 1999.

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

[3] Grimson, J., William, W., Berry, D., Stephens, G., Felton, E., Kalra, D., Toussaint, P., and Weier, W.A. CORBA-based integration of distributed electronic healthcare records using the Synapses approach. Biomedicine, 1998; 2 (3): 124-138.

[4] Shammugasundaram, J., Shekita, E., Barr, R., Carey, M., Lindsay, B., Pirahesh, H., and Reinwald, B. Efficiently publishing relational data as XML documents. The VLDB Journal 2001; 10(2-3): 133-154.

[5] C.A. Galindo-Legaria. Outerjoins as disjunction. Proceedings of the 1994 ACM-SIGMOD International Conference on the Management of Data, Minneapolis, USA, pp. 248-258.

[6] Rajaraman and J.D. Ullman. Integrating information by outerjoins and full disjunction. PODS: 1996: 238-248.

[7] Galindo-Legaria, C.A., Rosenthal, A. Outerjoin Simplification and reordering for query optimization. ACM Transactions on Database Systems, 1997: 22(1): 43-73.

[8] Lee, B.S., Wiederhold, G. Outerjoins and filters for instantiating object from relational databases through views. Transactions of Knowledge and data engineering, 1994: 6(1): 108-1119

[9] Wiederhold, G. Mediators in the Architecture of Future Information Systems. IEEE Computer, 1992: 25(3): 38-49.

[10] Hull, R. Managing semantic heterogeneity in Databases: a theoretical perspective. ACM PODS, 1997: 51-61.