INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 76S (2007) S417–S424

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

Non-invasive lightweight integration engine for building EHR from autonomous distributed systems

Carlos Angulo[*], *Pere Crespo*[a], *José A. Maldonado*[a], *David Moner*[a], *Daniel Pérez*[b], *Irene Abad*[b], *Jesús Mandingorra*[b], *Montserrat Robles*[a]

[a] *BET Group, Technical University of Valencia, Spain*
[b] *Consorcio Hospital General Universitario de Valencia (CHGUV), Valencia, Spain*

ARTICLE INFO

Article history:
Received 8 January 2007
Received in revised form 2 May 2007
Accepted 8 May 2007

Keywords:
Information management
Medical record systems
Hospital information system
Clinical data integration
Data exchange

ABSTRACT

In this paper we describe Pangea-LE, a message-oriented lightweight data integration engine that allows homogeneous and concurrent access to clinical information from disperse and heterogeneous data sources. The engine extracts the information and passes it to the requesting client applications in a flexible XML format. The XML response message can be formatted on demand by appropriate Extensible Stylesheet Language (XSL) transformations in order to meet the needs of client applications. We also present a real deployment in a hospital where Pangea-LE collects and generates an XML view of all the available patient clinical information. The information is presented to healthcare professionals in an Electronic Health Record (EHR) viewer Web application with patient search and EHR browsing capabilities. Implantation in a real setting has been a success due to the non-invasive nature of Pangea-LE which respects the existing information systems.

1. Introduction

Healthcare is a very data-intensive sector, producing and consuming a great amount of information. In healthcare organizations, especially hospitals, the big amount of data gets increasingly obscure due to their decentralized organization which has allowed different departments to meet specific or local requirements. This has led to fragmented and heterogeneous data resources, so called islands of information, which contain health data about patients, making the access and aggregation of data across systems very difficult. This situation has created a large gap between the potential and actual value of information contents of Electronic Health Record (EHR) systems.

A classic solution to the problem of information distribution is the acquisition or development of large and centralized information system, but past investments in existing IT infrastructure are not leveraged and departments lose their freedom to select the software that matches their requirements best. Furthermore, the best-of-breed approach can be very suitable for large organizations, letting departments meet their own business needs more easily and allowing bigger flexibility within decentralized institutions such as hospitals. The challenge is to find how these systems can efficient and meaningfully exchange health information in such a way that health professionals can access to the relevant information at the point of care, while the data can be held in specialized departmental systems or small enterprise systems [1].

Data integration is the problem of combining data residing at different sources and providing the user with a unified view of these data [2]. This unified view may be created from a set of existing data repositories in order to facilitate information

* *Corresponding author at*: Grupo BET, Instituto ITACA, Edificio 8G, Universidad Politécnica de Valencia, Valencia 46022, Spain. Tel.: +34 963877007.
E-mail address: cangulo@fis.upv.es (C. Angulo).
1386-5056/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2007.05.002

access through a single information access point, or when a certain information need arises it may be constructed by combining different but complementing repositories in order to meet the need. Data integration can be considered as the basis for other types of integrations such a functional or presentation [3].

There is a huge body of literature regarding data integration not only in the area of computer science but also in biomedical informatics [4–7]. There exist different approaches to data integration; broadly speaking they can be divided into three: data warehouses, database federations, and peer-to-peer data management systems.

A data warehouse is a central consolidated physical data repository. Data from local data sources are extracted, transformed and loaded in it. The main backward of this approach is that data are not available in the global repository until extracted from data sources, which may caused serious update difficulties and can create problems in those queries that need to be performed on the latest update. Also, it may be difficult to create a global schema that encompasses the high variability and complexity of the data sources.

On the other hand, database federations leave data at the sources and provide querying access to the set of data sources through a virtual view (schema). Users pose queries on the virtual view and a query engine is in charge of decomposing and translating the query into an equivalent set of local subqueries that are executed against the local data sources and whose results are then combined. The main disadvantages of this approach are that data cleansing becomes difficult since it must be done on the fly and performance may be degraded because it depends on the query capacity of data sources. Nevertheless, it alleviates the temporal problem of data warehousing since it always facilitates fresh data. The federation may maintain a common data model and relies on schema mapping for integration of the data source [7], usually supported by wrappers [8], which are specialized programs that interface with the sources and hide their technical details. In order to overcome the difficulties that can arise in dealing with the heterogeneity of data sources (relational, XML, object oriented, etc.) some systems use a global conceptualization model, called mediated schema, for the data from all integrated databases [4] which describes the domain entities and their relationship. Data sources are mapped to the mediated schema by defining the entities they contain. The mediator/wrapper architecture is one of the most commonly used to achieve data federation. Wiederhold [8] defines a mediator as "a software module that exploits encoded knowledge about certain set or subset of data to create information for a higher layer of applications". A mediator can be considered as a read-only virtual database which is introduced between the data sources and the application using them. It is capable of answering queries about the underlying data, for this purpose it uses data sources (suitable interfaced by a wrapper), and/or other mediator to answer the queries.

Peer-to-peer management systems (PDMS) [9] are an evolutionary step in data integration systems. In a PDMS every data source (peer) needs to only provide a semantic mapping to either one or a small set of other data sources. More complex relationships emerge when different semantic paths, used by the system to answer queries, are traversed. They allow the creation of multiple local, specialized mediated schemas tailored to specific users and then mappings are used to glue together semantically related peers. Therefore, none of the peers must take the responsibility of both creating a maintaining a mediated schema and mapping it to data sources. Therefore, they offer a truly distributed architecture to exchange data.

One of the difficulties in setting up a data integration system is the definition of the semantic mappings between the sources and the federated virtual schemas. It requires both database expertise to express them in a formal language and domain knowledge to understand the meaning of the schemas being mapped [10]. Mappings between two schemas can be specified either procedurally, i.e. by programs that physically import objects stored in the underlying databases into corresponding objects of the global environment or declaratively, i.e. by defining a set of correspondences between entities from the local and global schemas. Procedural mappings allow dealing with a wide range of data source formats and transformations, but they are more difficult to maintain because the mapping is encoded in actual programs. On the other hand, declarative mappings are easier to maintain and can accommodate query optimization but they are less powerful in specifying complex data transformation.

In the health domain, there exist several examples of health data integration systems. It is prevalent the utilization of a set of standard messages, mainly HL7 [11], which are used to exchange data among information system [12,13]. A far smaller number of integration efforts have used the federation approach; some of them are ARCHIMED [14], IBHIS [1], Synapses [15] or Synex [16].

At industrial level data integration is known as Enterprise Information Integration (EII). The main objective of this industry is to provide tools to access and query data held by heterogeneous sources without having to first load all the data into a central warehouse [10]. A related, and more mature, sector is Enterprise Application Integration (EAI). EAI aims to enable the communication among computer applications in order to support workflows. EAI is based on a diversity of technologies such as message brokers and adapters. Currently major Database Management System (DBMS) vendors and IT companies offer complex and generally quite expensive data integration products for the health care sector, mainly message brokers for HL7.

In this paper we present an overview of the Pangea-LE data federation system, emphasizing its flexibility and quick deployment. Briefly, Pangea-LE is a data integration system that provides a virtual, integrated and global XML [17] view over distributed health data sources. Although a materialized solution is generally more efficient computationally, we prefer virtual approach as it does not involve data replication, which may cause data update and synchronization problems, a significant limitation in an EHR application where up-to-date information is needed.

In Pangea-LE, the virtual global view is easily customizable for different user groups whose needs may change over time. In an evolving environment as healthcare, the local databases may change often. The databases are designed and maintained to meet local needs and changes are almost made independently of the integrated global view. Applications con-