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

Carlos Angulo[1], Pere Crespo[1], José A. Maldonado[1], David Moner[1], Montserrat Robles[1], Jesús Mandingorra[2], Daniel Pérez[2], Irene Abad[2]

[1] BET Group, Technical University of Valencia, Spain.

[2] Consorcio Hospital General Universitario de Valencia (CHGUV). Valencia, Spain.

Correspondence address:
Carlos Angulo Fernández
Grupo BET- Instituto ITACA
Edificio 8G, Universidad Politécnica de Valencia
Valencia 46022
Spain

Keywords: information management, medical record systems, hospital information system, clinical data integration, data exchange

To be considered for publication in the MIE 2006 special issue of the IJMI

[1] cangulo@fis.upv.es

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 XSL (Extensible Stylesheet Language) 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 EHR (Electronic Health Record) 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 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, whilst 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 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][5][6][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 EAI (Enterprise Application Integration). 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 DBMS (Database Management System) 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 XLM [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 connected to Pangea-LE do not required to be updated after any local data source change. These features make the Pangea-LE system a valuable solution to achieve fast, easy and secure data sharing and integration.

2. Materials and Methods

Pangea-LE acts like a mediator [8] between existing health data repositories and health professionals of an organization. It allows the definition and management of a global, structured XML view of all the clinical records stored for a patient. This view is presented to health professionals by means of an EHR Web application. Information views are created on demand and are role dependant so that enterprise-wide access roles can be defined for different professional profiles with specific access rights levels corresponding to different EHR views (clinicians, nurse, management, administrative, etc.)

We must stress that the purpose of this concrete use case is only to display EHR views. Information is presented in a human readable way and the clinician must provide his/her particular interpretation. Nevertheless, it is worthwhile noticing that it is also possible to keep the original meaning of shared EHR extracts for machine interpretation purposes, guided by some EHR standards. This use case has been tested only for experimental purposes according to the European standard CEN EN13606 [18] [19] for the communication of electronic health records.

Pangea-LE can be classified as a generic middleware that integrates clinical information. It is important to notice that there are only a few basic requirements to be met before Pangea-LE could be deployed into a healthcare organization:

- Unique patient identification throughout all the organization local information subsystems. In the event of there being none, the organization must provide a method in order to solve patient identification conflicts.
- Organization-wide user authentication and role assignment. This task is commonly executed by means of a directory service such a LDAP (Lightweight Directory Access Protocol) [20].

Pangea-LE offers read-only views only. Users access the system from standard desktop PCs with a Web browser. Security is implemented by the application server SSL channel via https protocol.

Pangea-LE architecture has four basic components (see Figure 1):

Wrappers: In the context of data integration a wrapper [21] is a program which understands the particular data organization of a data source. Its main purpose is to encapsulate data sources and offer a common interface for the extraction of data to the integration engine. In Pangea-LE heterogeneous data source access is achieved by using JDBC (*JAVA* Database Connectivity) drivers and through a set of configuration files, each of them describing a data source. The selection of JDBC is due to the predominance of relational databases in the health care sector. At the time of writing this paper there were 221 different JDBC drivers available [22] covering the vast majority of relational DBMS. For nonrelational data sources and repositories without a data manager such as file systems, ftp, messaging systems or even web services, corresponding JDBC wrappers have to be built in order to access them in a homogeneous way. For this purpose, we see data sources as a set of entities and attributes, as they are specialized in some concrete data organization, such as tables and columns in the case of relational data sources.

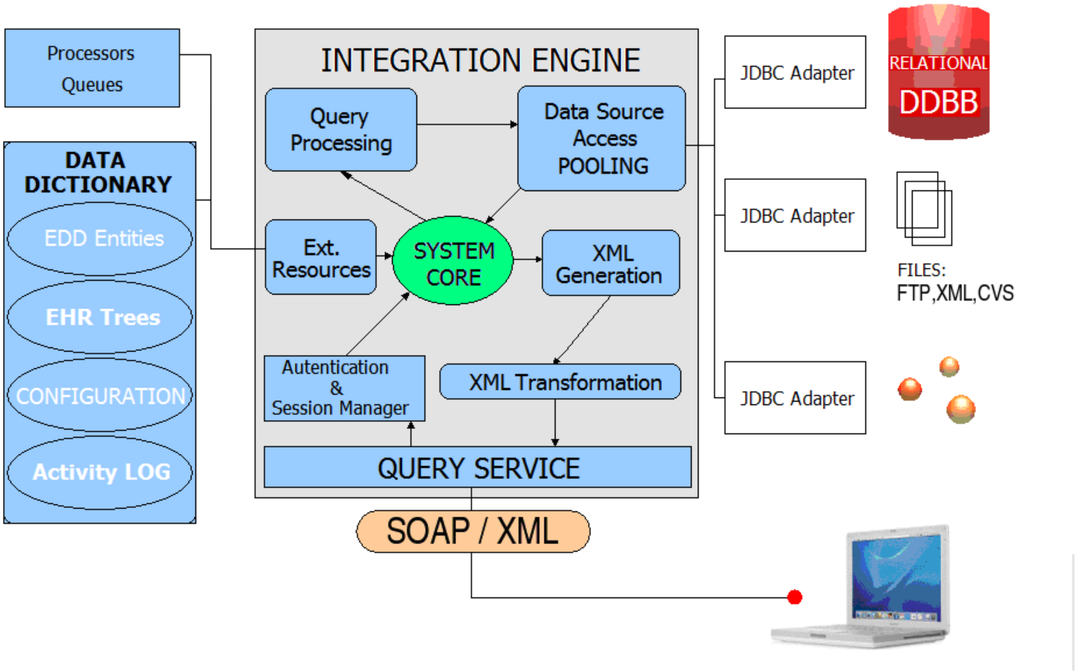
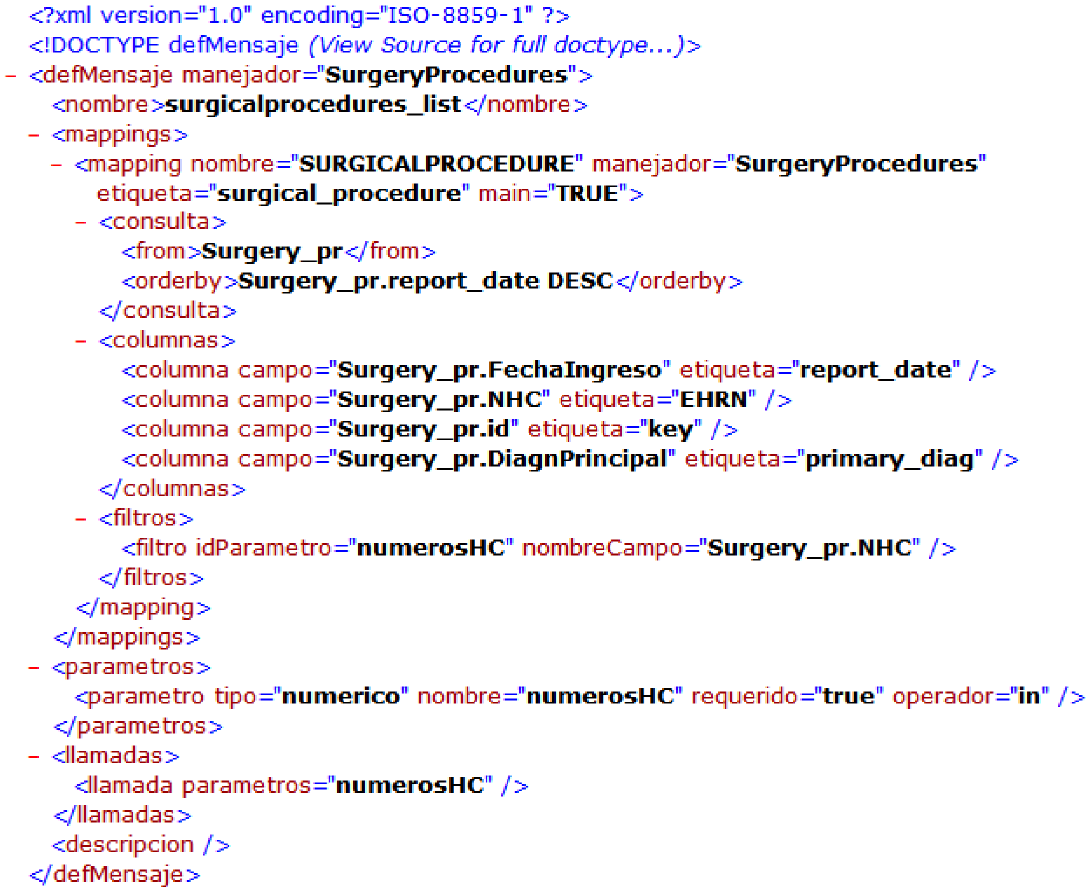


Figure 1 - Pangea-LE Architecture

EHR Extract Definitions (EED) which are used to specify the shared global view. Pangea-LE uses XML as canonical format for publishing and exchanging data, therefore the global view is a virtual XML view and legacy data should be transformed from the flat relational format to the hierarchical structure of XML. This is a typical data exchange problem (also known as data conversion, data translation or data transformation). Data exchange is the problem of taking data structure under a source schema and creating an instance of a target schema that reflect the source data as accurately as possible [23]. Data exchange requires at schema level an explicit representation of how the source and target schemas are related to each other; these explicit representations are called mappings. In Pangea-LE, mappings are defined in an EED.

An EED describes the structure of the XML instances of the clinical concepts to be shared in an integration project. It also contains the mapping to data sources, i.e. how the EED entities are related to data source entities. They combine the extraction part of SQL (FROM and WHERE clauses) with attributes (from data sources) to path mappings (from target schemas). Mappings are specified as a set of rules that associate an expression that may involve several columns from a data source to a leaf element of the target document. It is possible to define complex nested structures by using nested (correlated) SQL queries.



```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
  <!DOCTYPE defMensaje (View Source for full doctype...)>
- <defMensaje manejador="SurgeryProcedures">
    <nombre>surgicalprocedures_list</nombre>
  - <mappings>
    - <mapping nombre="SURGICALPROCEDURE" manejador="SurgeryProcedures"
        etiqueta="surgical_procedure" main="TRUE">
      - <consulta>
          <from>Surgery_pr</from>
          <orderby>Surgery_pr.report_date DESC</orderby>
        </consulta>
      - <columnas>
          <columna campo="Surgery_pr.FechaIngreso" etiqueta="report_date" />
          <columna campo="Surgery_pr.NHC" etiqueta="EHRN" />
          <columna campo="Surgery_pr.id" etiqueta="key" />
          <columna campo="Surgery_pr.DiagnPrincipal" etiqueta="primary_diag" />
        </columnas>
      - <filtros>
          <filtro idParametro="numerosHC" nombreCampo="Surgery_pr.NHC" />
        </filtros>
      </mapping>
    </mappings>
  - <parametros>
      <parametro tipo="numerico" nombre="numerosHC" requerido="true" operador="in" />
    </parametros>
  - <llamadas>
      <llamada parametros="numerosHC" />
    </llamadas>
    <descripcion />
  </defMensaje>
```

Figure 2 – EED Entity sample

It is also possible to use more than one data source in an EED, the only restriction being that every FROM clause must reference tables from a single data source, i.e. every nested query must use tables from a single database. An EED entity is specified in an XML file (see Figure 2) where some fixed descriptor elements define their behavior:

- Elements that specify a valid wrapper configuration for accessing the necessary information to populate the EED instances.
- Elements specifying data to be extracted from the data source (FROM and WHERE clauses).
- Input parameters accepted by query processor to execute filters on data sources.
- Valid calls (input parameter combinations which are allowed in order to build an EHR extract from an EED).
- Elements specifying data pre-processing which allows source atomic data to be combined or transformed before XML generation.
- Elements that describe the labeling and nesting format that constitutes the resulting XML document.

Each EED is a coherent information unit definition where the mapping between data source elements and global schema is established. Each EED can define as many mapping elements as needed to create the coherent and complete piece of information. Every data source involved in an integration project is a potential supporter for a set of clinical concepts, i.e. it may contain the data required to populate their instance, and therefore several EEDs can be defined for one single data repository. Each EED can only be shared as a whole. In other words, the minimum unit of information that can be shared through Pangea-LE is generated by an EED entity.

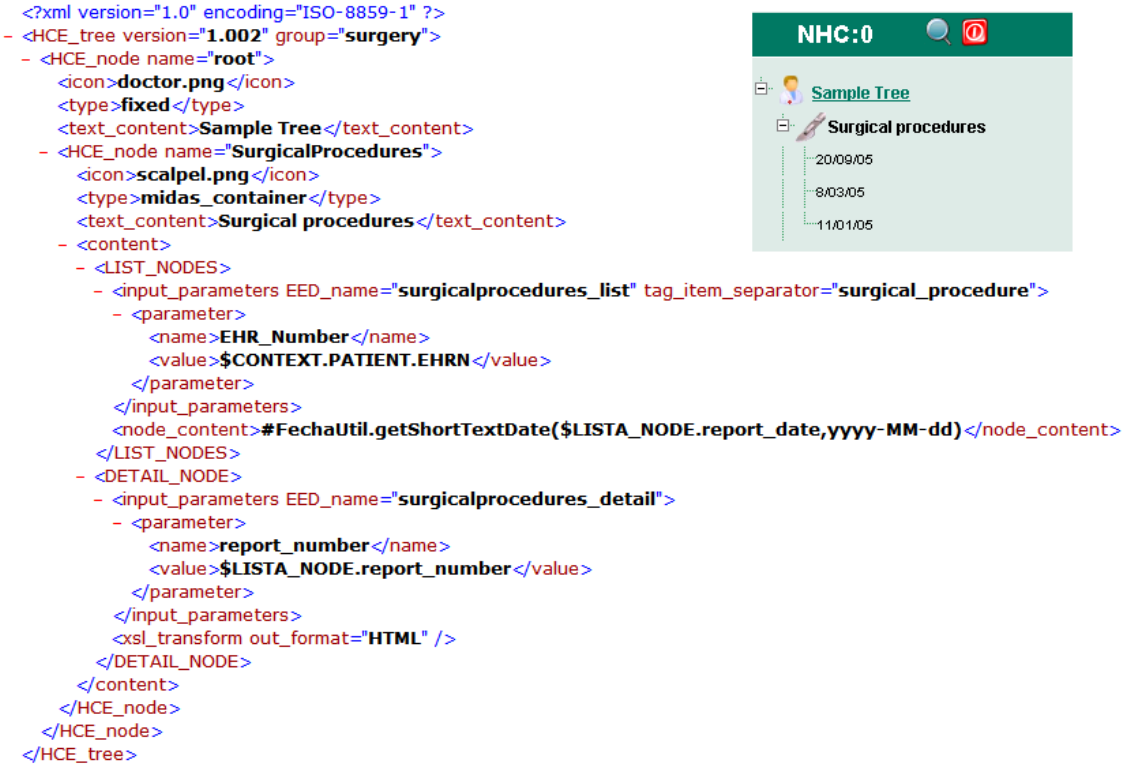
EED entities can be instantiated by a Pangea-LE exposed Web Service (WS) [24]. This WS accepts XML petition messages pointing to the desirable EED together with the required parameters (e.g. patient identification number). When the petition arrives to Pangea-LE and it is checked as a valid call, the Query Processing Module (see Figure 1) generates on the fly the SQL statements that are required to build the XML response. The extracted data is then structured and labeled according to the rules defined in the EED.

XSLT Transformations: Each EED can specify one or more XSL transformation files to be applied to the XML response message. This method allows Pangea-LE to adapt the output to different application message formats or devices (PDAs, Tablet PCs, etc.). Moreover, the same clinical entity can be formatted in different styles according to specific user access roles. Transformations can be applied on the server side or alternatively in the client application if it is allowed to perform this task.

EHR Browsing Trees: One of the most remarkable Pangea-LE features is the ability to organize the retrieved clinical information in a flexible tree structure. Once information has been extracted, different views can be configured: chronologically, data origin (emergencies, outpatient consultation, radiology…), etc. EHR trees are also defined in XML documents (see figure 3). They both guide the extraction process of the clinical data associated to one patient and specify how to structure, organize, aggregate and summarize data. At run time, each EHR tree constitutes a health record view for a particular role and it is mainly built by using EED entities. When a user wishes to access to the clinical data of a patient, the module that interprets EHR trees firstly retrieves only the minimum information needed to shape a summarized view of the patient's healthcare record. Subsequently, users can interact with each particular tree node to get a more detailed view of the particular clinical concept described by the corresponding EED entity. All these possible interactions are defined and controlled by EHR trees. Each structural component of the EHR tree data model corresponds to nodes in the visual tree control used in the EHR viewer application. Tree nodes are user interactive, therefore EHR views can be built interactively and on demand. A text and an icon can be assigned to nodes; furthermore they can be visible or invisible. Currently, three different types of nodes can be defined in an EHR tree: container, fixed and summary.

- Container nodes are the main EHR tree components since they define how to obtain the data to populate the EHR view. They are associated with an EED entity. In runtime, they will have a child for every EED instance generated.
- Fixed nodes are nodes that are not attached to an EED entity but they can have children nodes of any type. Their purpose is to customize the EHR tree structure.
- Summary nodes can collect EED instances from any number of container nodes and show them in the tree as child nodes. Summary nodes are very useful when different data sources are needed to be integrated as a single one, allowing a virtual union of sources. For instance, different types of discharge reports, held by different data sources, can coexist in the same organization. We can

define an invisible container node for each type of discharge report and collect their instances by using a summary node. Thus, the EHR tree view will have a single node that contains all the discharge reports.



```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
- <HCE_tree version="1.002" group="surgery">
  - <HCE_node name="root">
      <icon>doctor.png</icon>
      <type>fixed</type>
      <text_content>Sample Tree</text_content>
    - <HCE_node name="SurgicalProcedures">
        <icon>scalpel.png</icon>
        <type>midas_container</type>
        <text_content>Surgical procedures</text_content>
      - <content>
        - <LIST_NODES>
          - <input_parameters EED_name="surgicalprocedures_list" tag_item_separator="surgical_procedure">
            - <parameter>
                <name>EHR_Number</name>
                <value>$CONTEXT.PATIENT.EHRN</value>
              </parameter>
            </input_parameters>
            <node_content>#FechaUtil.getShortTextDate($LISTA_NODE.report_date,yyyy-MM-dd)</node_content>
          </LIST_NODES>
        - <DETAIL_NODE>
          - <input_parameters EED_name="surgicalprocedures_detail">
            - <parameter>
                <name>report_number</name>
                <value>$LISTA_NODE.report_number</value>
              </parameter>
            </input_parameters>
            <xsl_transform out_format="HTML" />
          </DETAIL_NODE>
        </content>
      </HCE_node>
    </HCE_node>
  </HCE_tree>
```

Figure 3 – EHR tree definition sample and result tree for EHR_Number = 0 instance

EHR tree definitions guide the EHR viewer appearance. Therefore, they may be used to define customized EHR views for different user groups or roles. Figure 3 shows a simple example. The main structural item in an HER tree definition is the *HCE_node* element. *HCE_node* specifies the node appearance in the resulting tree by defining its text content, type, icon and visibility. Content element defines the data to be showed. Two elements are nested inside of it. The *LIST_NODES* specifies the EED name (in the example surgicalprocedures_list) and the parameters that must be used to obtain the list of "user-friendly" instance identifiers, in the example the list of surgical procedure dates. List nodes are automatically executed when a user accesses to the EHR of a patient. This allows us to give the user a complete summary of the EHR. On the other hand, the *DETAIL_NODE* element stipulates the EED (in the example surgicalprocedures_detail) and the parameters that must be used to obtain a single full instance. Note that the parameter report_number is bound to the report number value which was previously extracted by the instantiation of the surgicalprocedures_list EED.

3. Results

Pangea-LE has been deployed in the Consorcio Hospital General Universitario de Valencia (CHGUV). CHGUV is a medium sized hospital serving a population of 350.000 inhabitants with 592 beds, 21 operating rooms and 470 physicians. The deployment process started at the end of 2004, since then the system has evolved from a prototype to a full operational system which currently serves the whole hospital.

In order to close the design-reality gap [25] a multidisciplinary work team has been created to make the integration project a success. The team is composed of representative specialists form clinical, management, documentation and computing departments. The team is in charge of the coordination of resources, planning and priorities selection. The deployment process itself has been progressive and strongly coordinated. Integration has been guided by the documentation and computing departments. The former is in charge of designing EHR trees and validation of data sources while the latter supplies the data source access parameters and schemata in order to build EED entities. Once a data source is selected to be integrated one or more EED must be created to extract data from that source. When the EED is ready, the appropriate transformations are implemented and tested to offer a suitable data presentation for human readable report. Finally, the relevant EHR trees are updated to include the new EDD.

The most important departmental systems were the first to be integrated (admission, emergencies, outpatient consultations agendas, pathological anatomy and a wide range of discharge reports). Currently, the vast majority of patient health data is available electronically through Pangea-LE, including: alerts, emergencies, inpatient episodes, outpatient consultations, laboratory results, biopsy and cytology test results, magnetic resonance reports, mammographies, endoscopies and most of the discharge reports from specialty units. At present, more than twenty local systems have been integrated and 36 EED have been defined. Table I shows the distribution of requests per data source.

DATA SOURCE	%
LABORATORY	30,00%
PATHOLOGICAL ANATOMY	11,30%
AMBULATORY CONSULTS	11,30%
EMERGENCIES	10,60%
ADMISSION	9,40%
CLINICAL REPORTS	9,40%
MAGNETIC RESONANCE REPORTS	9,00%
ENDOSCOPY	2,10%
OTHER SOURCES	6,80%

Table I. Distribution of requests per data source.

Easy access, completeness and immediate information retrieval have made EHR viewer application rapidly extend organization-wide, which is, in fact, the major evidence of the project's success. In order to gain precise knowledge of the actual use of Pangea-LE we have analysed the system log to get statistical data about the use of the EHR viewer. At the time of performing the analysis, late 2006 (two year after deployment), the number of potential medical users is 970, 470 of them being medical practitioner and 500 nursing staff. The number of regular users is 750 (approximately 75% of potential users) and nearly 300 of them use the EHR viewer daily. The system serves on average approximately 13000 EHR per month and more than 80% of the activity is concentrated in the range from 9 a.m. to 5 p.m. A detailed analysis of distribution of EHR petitions per data source and day of the week shows an anomaly on Sundays because the emergency activity is much higher than on the other days. All this clearly reflects that Pangea-LE is mainly used for everyday care delivery and demonstrates that the EHR viewer has become an important tool in the doctor's office.

Two evaluations have been carried out over the last two years. For this purpose a questionnaire was designed and sent to a selected representative user group (100 health professionals from diverse hospital units) to assess their opinion, confirm the former results and detect any weakness or deficiency. The first one, which was handed out three months after installation, was especially focused on information priorities in order to determine which data sources should be integrated first. The second one was completed two years after initial deployment. The main topics and results of the second evaluation were:

- Information availability: Questions to detect information that is needed by clinicians but not available in the EHR viewer application. 53% of users feel that not all information needed is available. This mainly refers to information held by external systems (91%). On the other hand, 88.2% of users find the information that they are searching for easily.
- User satisfaction: Questions about system performance and faults, if detected. System speed was evaluated on a (0-9) scale (the higher the better), the mean being 6.48.
- Added value: Questions about benefits for patients due to the use of the EHR viewer were also presented on the (0-9) scale. Patient benefit received a score of 8.26 while the advantages compared to paper based records achieved a score of 7.05.

4. Discussion

We introduced Pangea-LE, a light data integration engine based on free software components. 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 [26]. Our solution is based on the definition of a set of data aggregates, each of them defining a clinical concept, and mapping them with the heterogeneous structures found in the autonomous information systems.

There are tools that address the data transformation between different data schemas. Although these tools can help us to define mappings, for example between a relational data source and a target XML Schema, they usually show a lack of flexibility. Thus, we can use them to design a particular extraction and transformation algorithm (an XSLT transformation, an XQuery or some particular source code) to embed it in our own software but it is not a generic code which can be used for data integration and querying. If the source or target schema changes, we would be forced to regenerate the transformation algorithm and modify our application by hand. Pangea-LE resolves these cases the edition of the affected EED (a single XML document).

Pangea-LE allows a non-invasive integration mechanism, completely respectful of the already existing autonomous subsystems of a health organization with a very specific purpose: to define secure, global and unified views, organized by flexible criteria, over all EHR dispersed among the manifold subsystems of a healthcare organization. Nevertheless, due to its design, it is capable of evolving towards more advanced functionalities such as clinical research support and as a helping tool in extract, transform and load (ETL) processes required in clinical or economical research. Pangea-LE may ease the always laborious load stage of data warehouses and data marts. The strongest feature of the system is its high flexibility and scalability: flexibility to specify multiple formats for the presentation, to structure health data according to normalized formats to communicate EHR (HL7-CDA or EN13606) etc, and to organize the clinical history view according to different criteria; scalability to deploy the engine for inter-institutional interoperation.

Experience in a real setting has been shown. Pangea-LE has been in routine use for the last two years in the CHGUV, becoming the most important computer system for the medical professional who has to access patients' electronic health records. The deployment has demonstrated its flexibility and scalability. The system has allowed fast EHR integration throughout the whole organization. A key factor for the project's success has been the creation of a multidisciplinary working team involving medical, administrative and IT professionals in charge of coordinating the system deployment. The participation of clinical professionals has allowed rapid detection of requirements and failures, which has brought about the possibility of adapting the system to meet the needs of the organization.

Pangea-LE is not only helping in the delivery of healthcare but also in administrative tasks such as diagnosis codification using ICD-9. The admission department performs this task by processing discharge reports and any other relevant information. The EHR viewer has emerged as a key application because it eases the codification process by allowing codifiers to have immediate access to all the relevant information. Moreover, the demand on the hospital's archive of paper-based health records has been reduced significantly.

Interesting results were deduced from the questionnaire: in spite of the great amount of information available a large number of health professionals expressed the need to access information from data sources external to the hospital (mainly speciality outpatient centres). Users have found it easy to use the EHR viewer, which has fostered its rapid deployment throughout the hospital. Also, patient benefits were explicitly expressed, such as time-saving or the decrease of redundant diagnostic procedures. The most significant use of the EHR viewer occurs during patient encounters when doctors and nursing staff can have quick and complete access to information about past encounters and examinations.

The facility of use and a very intuitive graphic interface has made the EHR viewer a key application in the Hospital. The integration project has produced a very positive symbiosis between all participating parts. Even so, we have found some limitations in Pangea-LE:

- There is no terminology service available to map information to a shared knowledge source.
- Asynchronous data retrieval service is not supported. This would be a useful feature for clinicians who could be alerted, for instance, when a test result becomes available.
- System configuration files (EED, EHR Tree definitions, loggers, etc.) must be defined manually.

Prospects:

- Pangea-LE infrastructure can be a useful tool in the extract, transform and load (ETL) process of data warehousing.
- A visual management module is being developed in order to ease EED and EHR tree definition.
- A specific audit log viewer is also being developed so that a trace can be performed for a specific user or a specific patient.

5. Conclusions

In this paper we have presented Pangea-LE, a data integration system. Pangea-LE allows a non-invasive integration mechanism, completely respectful of the already existing autonomous subsystems of a health organization with a very specific purpose: to define secure, global and unified views, organized by flexible criteria, over all EHRs dispersed among the manifold subsystems of a healthcare institution. Nevertheless, due to its design, it is able to evolve towards the advanced functions described in the discussion section of this paper. But without a doubt the strongest feature of the system is its high flexibility and scalability: flexibility to specify multiple formats for the presentation of information, to structure the health data according to normalized formats to communicate EHRs (HL7 or EN13606) etc., and to organize the clinical history view according to different criteria; scalability to deploy the engine for inter-institutional interoperation.

Pangea-LE is highly flexible, fully scalable, with almost no preliminary requirements, and allows fast EHR deployment throughout an organization. Our experience in CHGUV has shown that the full specification of an integration project can be performed in just a few weeks. In fact, the main part of it was completed in a matter of a few days.

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Technology (MEC – TSI2004-06475-102-01) and the Spanish Ministry of Health (PI052245).

References

[1] M. Rigby, D. Budgen, M. Turner, I. Kotsiopoulos, P. Brereton, J. Keane, K. Bennett, M. Russell, P. Layzell, F. Zhu, A data-gathering broker as a future orientated approach to supporting EPR users. International Journal of Medical Informatics 76(2-3) (2007) 137-144.

[2] M. Lenzerini, Data Integration: A Theoretical Perspective, Proceedings of the 21st ACM SIGACT-SIGMOND-SIGART Symposium on Principles of Database Systems (2002) 233-246.

[3] R. Lenz, M. Beyer, K.A. Kuhn, Semantic integration in healthcare networks. International Journal of Medical Informatics 76(2-3) (2007) 201-207.

[4] R. Alonso-Calvo, V. Maojo, H. Billhardt, F. Martín-Sánchez, M. García-Remsal, D. Pérez-Rey, An agent- and ontology-based system for integrating public gene, protein, and disease databases, Journal of Biomedical Informatics, 40(1) (2006) 17-29.

[5] K.A. Karasavvas, R. Baldock, A. Burger, Bioinformatics integration and agent technology, Journal of Biomedical Informatics 36(3) (2004) 205-219.

[6] W. Sujansky, Heterogeneous database integration in biomedicine, Journal of Biomedical Informatics 34(4) (2001) 285-298.

[7] B. Louie B, P. Mork, F. Martin-Sánchez, A. Halevy, P. Aarczy-Hornoch, Data integration and genomic medicine. Journal of Biomedical Informatics 40(1) (2007) 5-16.

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

[9] S. Gribble, A. Halevy, Z. Ives, M. Rodrig, D. Suciu, What Can Databases Do for Peer-to-Peer?. Proceedings of the Fourth International Workshop on the Web and Databases, WebDB (2001) 31-36.

[10] A. Halevy, A. Rajaraman, J. Ordille, Data Integration: The teenage years. Proceedings of the 32nd International Conference on Very Large Data Bases (2006) 6-16.

[11] Health Level Seven (HL7), http://hl7.org (last visited December, 2006)

[12] K.U. Heitmann, R. Schwiger, J. Dudeck, Discharge and referral data exchange using global standards-the SCIPHOX project in Germany. International Journal of Medical Informatics 70(2-3) (2003) 195-203.

[13] Y. Sooyoung, K. Boyoung, P. Heekyong, C. Jinwook, C. Jonghoon, Design and implementation of HL7 based real-time clinical data integration system. Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS '03 (2003) 222-230.

[14] G. Thurler, F. Borst, C. Bréant, D. Campi, J. Jenc, B. Lehner-Godinho, P. Maricot, J.R. Scherrer, ARCHIMED: A network of integrated information systems. Methods of Information in Medicine 39(1) (2000) 36-43.

[15] W. Grimson, D. Berry, J. Grimson, G. Stephens, E. Felton, P. Given, R. O'Moore, Federated healthcare record server-the Synapses paradigm. International Journal of Medical Informatics 52(1-3) (1998) 3-27.

[16] Y. Xu, D. Sauquet, P. Degoulet, M.C. Jaulent, Component-based mediation services for the integration of medical applications. Artificial Intelligence in Medicine 27(3) (2003) 283-304.

[17] World Wide Web Consortium, XML, http://www.w3.org/XML (last visited December, 2006)

[18] J.A. Maldonado, M. Robles, P. Crespo, Integration of distributed healthcare records: publishing legacy data as XML documents compliant with CEN/TC251 ENV13606, Proceedings of the 16th

IEEE Symposium on Computer Based Medical Systems, IEEE Computer Society Press (2003) 213-218.

[19] European Committee for Standardization, Health informatics-Electronic health record communication- Part 1: Reference model, Draft European Standard for CEN Enquiry prEN13606-1, 2006.

[20] LDAP RFC, http://www.ietf.org/rfc/rfc2251.txt (last visited December, 2006)

[21] M.T. Roth, P.M. Schwarz, Don't scrap it, wrap it! A wrapper architecture for legacy data sources, Proceedings of 23rd International Conference on Very Large Database Systems (1997) 266-275.

[22] JDBC Data Access API, http://developers.sun.com/product/jdbc/drivers (last visited December, 2006)

[23] R. Fagin, P.G. Kolaitis, R.J. Miller, L. Popa, Data exchange: semantics and query answering, Proceedings of the 9th International Conference on Database Theory (2003) 207-224.

[24] World Wide Web Consortium, Web Services: http://www.w3.org/2002/ws (last visited December 2006)

[25] R. Heeks, Health information systems: Failure, success and improvisation, International Journal of Medical Informatics 75 (2006) 125-137.

[26] R. Hull, Managing semantic heterogeneity in Databases: a theoretical perspective, Proceedings of the ACM PODS (1997). 51-61.

Table I. Distribution of requests per data source.

DATA SOURCE	%
LABORATORY	30,00%
PATHOLOGICAL ANATOMY	11,30%
AMBULATORY CONSULTS	11,30%
EMERGENCIES	10,60%
ADMISSION	9,40%
CLINICAL REPORTS	9,40%
MAGNETIC RESONANCE REPORTS	9,00%
ENDOSCOPY	2,10%
OTHER SOURCES	6,80%