MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

Design of a Decentralized Reusable Research Database Architecture to Support Data Acquisition in Large Research Projects

Jimison Iavindrasana [a], Adrien Depeursinge [a], Patrick Ruch [a], Stéphane Spahni [a], Antoine Geissbuhler [a], Henning Müller [a]

[a] *Geneva University Hospitals, Geneva, Switzerland*

Abstract

The diagnostic and therapeutic processes, as well as the development of new treatments, are hindered by the fragmentation of information which underlies them. In a multi-institutional research study database, the clinical information system (CIS) contains the primary data input. An important part of the money of large scale clinical studies is often paid for data creation and maintenance. The objective of this work is to design a decentralized, scalable, reusable database architecture with lower maintenance costs for managing and integrating distributed heterogeneous data required as basis for a large-scale research project. Technical and legal aspects are taken into account based on various use case scenarios. The architecture contains 4 layers: data storage and access are decentralized at their production source, a connector as a proxy between the CIS and the external world, an information mediator as a data access point and the client side. The proposed design will be implemented inside six clinical centers participating in the @neurIST project as part of a larger system on data integration and reuse for aneurism treatment.

Keywords:

Integrated information management systems, hospital information system, multi-institutional research databases.

Introduction

The diagnostic and therapeutic processes, as well as the development of new treatments are hindered by the fragmentation of information which underlies them. Many developments have emerged in the genomics, proteomics and medical imaging domains during the last 10 years. Linking distributed, multi-format, and multi-scaled data from genetics, proteomics, the individual, and epidemiological data for diagnosis and treatment is a new challenge in the biomedical informatics domain [1]. Many research projects are working on the integration of genomics knowledge into clinical practice. The INFOBIOMED[1] project is studying the relationship between bioinformatics and medical informatics in provision of individualization of healthcare. Integrating multi-scale data also means a need to integrate expertise of various participants: clinicians, biologists, statisticians, etc.

To develop new diagnostic or therapeutic processes, prospective clinical research can be conducted across multiple institutions. The electronic patient record can be used as initial data input and the data can be used for more than the treatment of a single patient. A multi-institutional data collection permits to have a statistically significant number of cases in a shorter time period. On the other hand, sharing clinical data for external researchers creates many problems: 1) patient data privacy and confidentiality need to be ensured; 2) there is no standard secure communication protocol and format for data representation as HL7[2] and DICOM[3] are mainly used for communication inside a hospital and do not cover all aspects required, particularly concerning terminologies to be used.

In multi-institutional clinical research, most of the data are collected inside the clinical information system (CIS) or are exported to create databases for one particular task. Specialized data collections such as genetic sequencing are rarely incorporated into routine medical practice. These data need to be entered from the outside of the clinical centers. Moreover, connections with existing public knowledge bases such as SWISSPROT[4] are rarely done. Designing a research database infrastructure for managing heterogeneous data formats, heterogeneous distributed data sources, access by multiple experts to the entire data located in various geographical regions is a big challenge.

Two categories of architectures exist to conduct multi-institutional clinical research: centralized [2] and distributed [3, 4]. The main drawback of a centralized architecture is the maintenance and data security [5]. In a distributed architecture, data can be stored in a local system and all participating institutions use the same data model, query language and the same database management software (distributed database system - DDBS) [6]. In the DDBS architecture, the maintenance complexity of the centralized architecture is reported on the level of each clinical center. In the real world it is also difficult to have the same database management system installed on the

[1] http://www.infobiomed.org/

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

[3] http://medical.nema.org

[4] http://www.expasy.org/sprot/

same operating system across various participating institutions, leading to a heterogeneous database system (HDBS) [6]. Mediators have been proposed to allow access to heterogeneous data sources [3].

For security reasons, a CIS is generally a closed system. However, more and more clinical research is conducted inside hospitals and an increasing number of research networks are set up to share clinical data among institutions. Generally, in DDBS or HDBS, data are stored in a secured zone outside the CIS for secondary use. Nevertheless, it is risky to leave clinical data outside of the CIS because it is difficult to achieve a complete anonymity of clinical data. Genetic sequences, for example, are unique for each individual. Clinical data are collected over time and a visit date can be linked to other databases and can permit to identify a patient. Moreover, in provision of a patient-centered healthcare, there is a need to go back to the patient to inform him about new discoveries that might arise in a research project and as a new treatment of a disease.

Data integration is defined as the problem of combining data residing at different sources, and providing the user with a unified view of these data [7]. Generally, clinical centers have their own security, access rights management and privacy protection policy according to the role the user [8], and have the know-how concerning data access and communication using standards such as HL7, DICOM, IHE, or business components such as web services [9]. In multi-institutional prospective clinical research, a decentralized HDBS architecture is most flexible. It is also often the safest for the management of multimedia, multi-source, and distributed user data. In this architecture, the data are stored at their source i.e. the CIS of the institution where the data were created. Across institutions this can be in multiple databases, managed by various database management system installed on distinct operating system.

Often, half of the money of large scale clinical studies is paid for data creation and maintenance.

The objective of this paper is to propose a design of a distributed re-usable research database architecture, which permits managing and retrieving heterogeneous data to have real-time and up-to-date integrated data. Interoperability issues, research database maintenance and access are also discussed. The work is being carried out in the context of the European Union research project @neurIST and is part of a larger infrastructure within this project. The architecture proposed concerns "open" CIS: accessible for change and having direct data access, in the context of a multi-institutional research project.

Methods

The article describes the design and first implementations of a reusable research database architecture for a large-scale European Union funded clinical research project. Clinical data acquisition is planned in four countries at the first phase and then in a larger number once all system components are in place. To do so, many constraints were taken into account from these countries on the one hand side from a legal standpoint and also from a technical standpoint. The architecture is to be implemented between the clinical institutions that will produce the data and several research applications that are the potential users of the data.

A web-services HTTP/XML based approach was chosen for the communication. To model all possible scenarios a large number of use cases were defined among the participants of the research project. Based on these use cases, legal and technical aspects of the functionality were defined and an architecture planned that was proposed to all partners and that is currently under internal evaluation for possibly missing parts.

Results

Requirements

Critical tasks in a database project are: data collection and validation, data storage and communication. In a decentralized database architecture, most of the clinical data collection, data validation and storage are done inside the participating clinical centers.

Data categorization and storage

In a decentralized research database architecture, it is crucial to differentiate the treatment and the research cycle. The treatment cycle produces clinical data and the research cycle provides derived data.

Additional data such as genetic analyses and research results related to a specific patient are in our case collected outside of the clinical centers and have to be stored inside the participating clinical centers' information systems. Other global research data such as epidemiological study results or other public knowledge useful for the users of the database cannot be stored in any participating clinical center's information system.

Patient identification

In a patient-centered clinical context, a patient can have access to all data stored in the CIS related to his health. Derived data need to be separated from clinical data by the means of patient identifiers. An internal patient identifier is used to identify a patient in the treatment cycle and an external identifier is used to identify the patient for secondary use of its data. As other clinical data may be collected outside of the hospital, the internal identifier cannot be used to identify the patient during the upload. First, the internal identifier may be created from private information of the patient. Second, external clinical data need to be validated by a clinician before its insertion into the electronic health record. Other specific identifiers need to be set up to identify external clinical data. In prospective clinical research, a patient needs to be recruited and followed in a single center. At this stage, we do not consider mobile patients between centers. This problem could be resolved with a global identification solution but is discarded for now.

Data integration

One of the biggest issues in a multi-institutional research database is data integration. The decentralized research database architecture is implemented as a global-as-view [7] and is able to manage structural, naming, semantic and content differences of the data from different sources. Inside a participating clinical center, data are coded into the local terminological system. The normalization of the data into a common terminological system is done on-the-fly.

Communication and security

Interoperability is also a crucial point in the design of a research database. XML/HTTP is the de facto standard to achieve interoperability. The architecture will use XML for data encoding and HTTP as transport protocol.

Privacy of clinical data needs to be ensured by the architecture: all data leaving the CIS, for secondary use have to be anonymized and the process has to be done on-the-fly before the data leave the clinical center. However, the architecture needs to permit to re-identify the patient to store related research result in the CIS. The architecture also needs to allow identification of the end-user. All communication has to be encrypted.

Design

The architecture we present in this paper has 4 layers: 1) the *data source layer* inside the CIS or at a node of the research network; 2) the *connector* on each participating clinical center which is mainly responsible for the horizontal data integration; 3) the *information mediator,* which is mainly responsible for vertical data integration and 4) the *client* (Figure 1). For security reasons and re-usability, operations are decentralized at the appropriate layer.

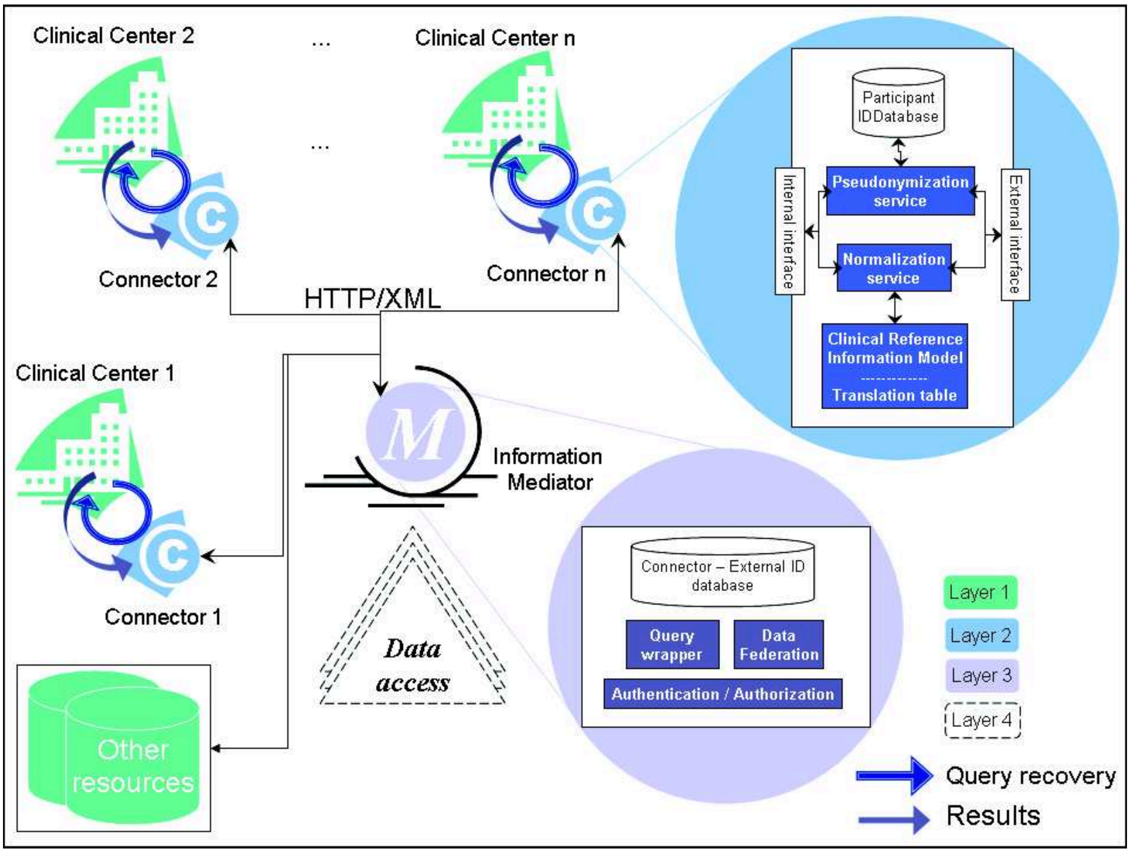

Figure 1 – A 4- layer architecture composed of: 1) the data source in the CIS or at a node of the research network; 2) the connector on each participating clinical center ensuring data normalization and pseudonymization; 3) the information mediator responsible for the query and user management and 4) the client

Data source

A large part of the data are stored inside the CIS and accessed using the existing data access components. Clinical data are stored in the electronic health record and derived data in a research database. The latter is integrated in the CIS because research results may contain private information such as gene sequences. There is no possibility to link clinical and derived data inside the clinical center: querying derived data using private information of the patient is prohibited. Even if derived data related to a patient are stored in the research database inside the CIS, it cannot be queried directly from the inside: it is only accessed through the information mediator. Other global or public derived data, not related to a specific patient are stored at their source.

The source can be queried using traditional communication protocols such as HL7 or DICOM or other components used inside the clinical center. The data access component inside the CIS communicates with the connector using XML/HTTP. For security reasons, the CIS always initiates the communication with the connector and pools the results. The CIS checks for new queries transmitted by the connector at fixed time intervals.

The connector

The connector is installed in the DMZ of each participating clinical center and uses the clinical reference information model (CRIM) for query and data translation. The CRIM contains the semantics of all existing data attributes: name, type and definition. These attributes are mapped to the local data attributes in the translation table. The normalization service is a bidirectional conversion of data format in the local or the global terminological system and it is invoked during the transition of the data inside the connector. The mapping has to be strong enough to ensure data integrity.

The pseudonymization service removes all private information from the local data before their delivery outside of the DMZ. It uses an ID database containing the mapping between various patient identifiers in the provision of re-identifying the patient for update tasks. The pseudonymization service can be extended as a patient identifier generator (external and other identifiers) but does not store any private information concerning the patient other than the internal ID.

The connector has two web service interfaces: the internal interface communicates with the CIS and the external one communicates with the information mediator. The connector does not identify the end-user: this task is centrally done at the information mediator level and the connector accepts only queries from the information model.

The information mediator

The information mediator is the data access point. It manages queries to various data sources: CISs and other public or private resources stored outside clinical centers. The user authentication and authorization are done at the information mediator level. The mediator also contains a database with the mappings between the external patient ID and all connector identifiers having information about the patient. This limits the network usage when a query is performed for a specific patient. It uses the clinical reference information model to formulate queries and it can be extended using an ontology to enable semantic mediation.

Discussion

We detailed in the preceding sections the design of a reusable research database for managing heterogeneous data sources to have real-time and up-to-date integrated data and address interoperability issues. The architecture permits to address data structure, naming, semantic and content differences across various clinical centers. Storing an internal identifier in the DMZ can raise legal issues. It has to be mentioned in the study protocol and the patient consent that the identifier will be stored in the DMZ and that it is never published beyond.

Limits of the approach

Using production databases for secondary use is badly perceived by IT decision makers in the health field because of: 1) technical constraints that the CIS is not open to change; 2) the sensibility of clinical data. However, with the increasing number of clinical and research networks managing and duplicating clinical data for secondary use will become increasingly expensive. For this reason most participating clinical centers prefer to export anonymized clinical data into a database in the DMZ instead of doing on-the-fly data extraction from the CIS. The research database will be hosted in a secured zone as shown in figure 2.

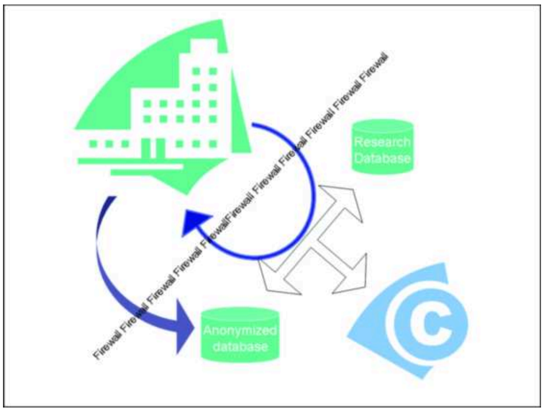
Firewall Firewall Firewall Firewall Firewall Firewall Firewall Firewall Firewall
Research Database
Anonymized database
C

Figure 2 – Second storage architecture

In this second storage solution, the pseudonymization service of the connector is activated to pseudonymize data before their export into the DMZ. When there is a need to go back to the patient's data in the CIS, a message is sent by the CIS to collect queries from the connector. This process can be automated by setting up a service that will check for the existence of new queries at a fixed time interval.

In the proposed design, the communication between the CIS and the connector is unidirectional and intermittent. It can affect the query timing especially for *select* queries. Opening a permanent communication between the 2 layers of the architecture can raise security risks.

Contribution

The main advantage of the proposed architecture is its reusability, scalability and lower maintenance cost for research projects. Adding a new clinical center means installing a standard connector and providing a mapping between local and global data attributes. Defining a new research database project consists of defining the CRIM for this domain and adding clinical centers.

The emergence of GRID network technologies can empower the analysis and treatment of complex and manifold multimedia data: heavy computing tasks can be externalized onto a GRID-computing environment [10] and might allow new possibilities in the reuse of data. This architecture can be interfaced with grid computing infrastructures that allow access to computational power that many institutions could not afford on their own [11]. In this case, the connector plays more than the role of a "data proxy". It receives computing requests from the CIS, pseudonymizes patient data if needed, and send the queries to a GRID-computing service provider. When the computation is finished, the connector receives the results, re-identifies the patient and sends them to the CIS.

Future developments

In the first part of the project, a patient is recruited and followed in a single clinical center. However, a patient may move between places after being recruited while continuing to take part in the study. There is a need to develop a unique external identifier to link information about this patient across clinical centers.

Conclusion

The design of a decentralized, scalable and re-usable, research database architecture with lower maintenance costs is described in this paper. The architecture has layers to manage heterogeneous data collection, storage and retrieval in a secured way and ensures the patient's information privacy. The architecture is suited for extensible clinical information systems. The structure is part of a larger @neurIST architecture.

Acknowledgement

This work was generated in the framework of the @neurIST Integrated Project, which is co-financed by the European Commission through the contract no. IST-027703.

References

[1] Hunter P, Smith N, Fernandez J, and Tawhai M. Integration from proteins to organs: the IUPS Physiome Project. Mech Ageing Dev 2005:126(1):187-92.

[2] Kerkri EM, Quantin C, Grison T, Allaert FA, Tchounikine A, and Yetongnon K. A virtual intranet and data-warehousing for healthcare co-operation. Medinfo 2001:10:23-7.

[3] Astakhov V, Gupta A, Santini S, and Grethe JS. Data Integration in the Biomedical Informatics Research Network (BIRN). In: Ludäscher B, and Raschid L, eds. Second International Workshop, Data Integration in Life Sciences. San Diego. Proceedings. Lecture Notes in Computer Science 2005: 3615: 317-20.

[4] Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, Kher M, Manisundaram A, Shanbhag K, Covitz P. caGrid: design and implementation of the core architecture of the cancer biomedical informatics Grid. Bioinformatics 2006: 22(15): 1910-6.

[5] INFOBIOMED. State of the Art on Data Interoperability and Management. Available at http://www.infobiomed.org / paginas_en/D11_State_of_Art_Data.pdf.

[6] Sujansky W. Heterogeneous Database Integration in Biomedicine. J Biomed Inform 2001: 34(4): 285-98.

[7] Hernandez T, Kambhampati, S. Integration of biological sources: current systems and challenges ahead, ACM SIGMOD Record 2004: 33(3): 51-60.

[8] Lovis C, Spahni S, Cassoni-Schoellhammer N, Geissbuhler A. Comprehensive management of the access to a component-based healthcare information system. Stud Health Technol Inform 2006: 124: 251-6.

[9] Geissbuhler A, Lovis C, Lamb A, Spahni S. Experience with an XML/http-based federative approach to develop a hospital-wide clinical information system. Medinfo 2001:10:735-9.

[10] Müller H, Garcia A, Vallée JP, and Geissbuhler A. Grid-Computing at the University Hospitals of Geneva. Proceedings of the First HealthGrid Conference, Lyon, 2003, pp 264-76.

[11] Müller H, Geissbuhler A, Ruch P. Report on the CLEF Experiment: Combining image and multi-lingual search for medical image retrieval. In CLEF Proceedings: Springer Lecture Notes in Computer Science, 2005:3491:718-27.

Address for correspondence

Jimison IAVINDRASANA
Service d'Informatique Médicale
Hôpitaux Universitaires de Genève
Rue Micheli-du-Crest, 24, Genève, Switzerland
e-mail : jimison.iavindrasana@sim.hcuge.ch