How clinical information system can support life science research

Jimison Iavindrasana[a], Adrien Depeursinge[a], Rodolphe Meyer[a], Stéphane Spahni [a], Antoine Geissbuhler[a] and Henning Müller[a,b]
[a] *Service of Medical Informatics, University and Hospitals of Geneva (HUG), Switzerland*
[b] *Business Information Systems, University of Applied Sciences Sierre, Switzerland*

Abstract.
The management of the cerebral aneurysm can be improved by setting up an IT infrastructure capable of integrating all available knowledge related to the disease. This paper presents the infrastructure implemented in the University Hospital of Geneva to support the @neurIST project. The infrastructure permits to share patient data stored in the clinical information system for secondary use and integrate derived results. The requirements, the basic functionalities and the security aspect regarding the patient privacy and the clinical information system security are presented. The main advantages of the implementation are the re-usability, scalability and the lower maintenance cost while maintaining the patient privacy and the clinical information system security.

Keywords: clinical information system, cerebral aneurysm, secondary use of clinical data

1. Introduction

The wider availability of modern imaging solutions permits to diagnose more and more unruptured cerebral aneurysms. Their rupture has fatal and disabling consequences with high societal costs but the incidence of rupture is low. Over the past decades, impressive developments were achieved in various domains like the determination of the associated risk factors, imaging, rupture prediction based on hemodynamic simulations and treatment of this complex disease. The actual knowledge related to cerebral aneurysm spans length scales from molecular, through cellular, to tissue, organ and patient representations. However, the current management of the disease is based on very limited information such as the size and the localization of the aneurysm. Information technology can bring valuable improvement on the research and management of cerebral aneurysms by integrating all exploitable information and computational services. The @neurIST[1] project targets to provide a new IT infrastructure to manage, integrate and interrogate the data related to the disease. The @neurIST is an Information Society Technologies (IST) Integrated Project funded within the European Commission's (EC) Sixth Framework Programme.

Five pilot clinical centers are participating in the @neurIST project and are acting as patient recruiter and clinical data provider. A multi-institutional data collection permits to have a statistically significant number of cases in a shorter time period, to incorporate a

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

more diverse study population, and to reduce the bias induced by any individual researcher. The electronic patient record is used as initial data input and the data can be used for more than the treatment of a single patient.

In a multi-institutional prospective research, the electronic medical record of participating patients can be anonymized and exported on a centralized repository such as in [1] or stored in each site in an anonymized form such as in [2,3]. In the second case, the data are stored in a repository disconnected from the hospital information system. In the first model, the basic research results are also stored on the central repository. In the second solution, research results are stored where they were produced; a mediation program links all data sources according to the end-user query.

The second architectural solution provide more flexibility with respect to data management as each data producer is responsible of the maintenance of the distributed data source. However, the evaluation of the effectiveness of a new therapy or a disease management tool is done inside each clinical centre. To reduce the data management cost inside the clinical centre and to provide real time data to researchers, the clinical information system (CIS) can be used as data repository for both clinical management of the patient and for research. In this way, the electronic medical record of each participating patient is not duplicated in a specific data repository and all research results related to a patient are directly stored in the clinical information system. This later brings other issues concerning patient data privacy and confidentiality as the patient data leaves the control and protection-sphere of the medical secrecy and the security of the hospital information system itself.

A first architectural design was proposed in [4] to deal with access to data stored in a CIS for secondary use. In this paper, we present the requirements and the basic functionalities of the first implementation of architectural solution chosen by the University Hospital of Geneva with respect to data management for the @neurIST project (section 2 and 3). The security of the patient data and the hospital information system is also discussed (section 4).

2. Method

The implementation described in this paper was designed according to legal and technical constraints and based on use case scenarios proposed by the end users.

3. Architecture requirements

The architecture described in this paper provides two functional cycles: 1) share clinical data produced in the clinical centre to their potential users i.e. the researchers and/or external computation services; 2) store back research results related to participating patients. Four main requirements guided the design of the architecture: data categorization, patient identification, data integration and data communication and security.

In the @neurIST project, two categories of data are produced: clinical and research data. The CIS is only used to store the first category and research data related to one patient. However, it is possible that the later are voluminous (such as simulation files) and that only the metadata permitting to reproduce the result will be stored back in the

CIS. Moreover, other voluminous research data are produced according to experimental process and are not reproducible without the necessary materials. These results are stored where they were produced but can be linked with the patient data using a strong identification mechanism.

As a patient has the right to access his electronic health record, the derived data produced by researchers should be identified in other way than the clinical patient identifier. Furthermore, the personal identifiable information of the patient should be removed or reduced at the minimum required before secondary use. This implies that all data leaving the clinical context should be depersonalized and a mechanism to map the two identities should exist (one-to-one relationship).

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 [5]. In a data integration problem, all data sources are hidden to the end user and viewed as a single data source; the data representation is also unified so that the provenance of the data cannot be identified. As the primary data are collected inside the clinical centre, they are coded with the local terminological system. The normalization of the data i.e. their conversion into an agreed and unified format is done on-the-fly when the data leave the hospital boundary. The de-normalization process is also done in the same process.

Within the @neurIST project, GRID and Service-Oriented Architecture (SOA) where chosen as the underlying technologies of the whole IT infrastructure [6]. The elements of the architecture are composed of a set of web services and compatible with GRID technology. A detailed description of the security aspect is provided in section 4.

4. Basic functionalities of the implementation

The architecture implemented to support the @neurIST project has three layers (see figure 1): the Public Data Service located in the hospital's demilitarized zone (DMZ), the Private Data Service located in the hospital's intranet, and the CIS located in the hospital's intranet as the data source. The mechanism to query data stored in the CIS is presented in the following paragraphs.

The public data service is the entry point to the data stored in the CIS. It is located in the hospital's DMZ and thus accessible from open networks such as the Internet. It authenticates the user or the application querying the database and all transactions are monitored and logged. A hospital information system is a closed one and the public data service cannot communicate directly with the CIS. It instead queues all authenticated and authorized incoming queries in a repository. This component is based on the GRID middleware OGSA-DAI[2].

The private data service is responsible for recovering all queued incoming queries located in the DMZ. Received queries are transformed to reflect the internal data structure and representation (mediation and de-normalization) and sent to the CIS. A novel feature of the current architecture is the ability of the private data service to mediate the reception of data generated outside the hospital for incorporation into the CIS - a process which mandates the ability to re-identify the patient. Thus a query may be either a request for data or a notification of data being available for integration. The private data service has two important resources for these steps: the translation rules for

[2] http://www.ogsadai.org.uk

the normalization & de-normalization service and the ID database which also contains the patient's consent preferences. This latter policy is required to control the access to the patient's data: only the @neurIST data related to patients registered in the ID database can be queried.

The patient data returned by the CIS (query results) are depersonalized and pseudonymized on-the-fly by the private data service. The query results are also filtered by this component. When the results are normalized – i.e. transformed into an agreed representation – they are written down in a result repository located in the DMZ and can be retrieved by the client.

A further access policy consideration is that derived data are not viewable or accessible by unauthorized users (which may, for unverified results, include the clinicians and/or patients) until reasons and methods for release are favorably reviewed by internal project ethics committee. These access are subject to both the clinical centre's policy and the patient's decision on whether they agree or not to the return of research health relevant results.

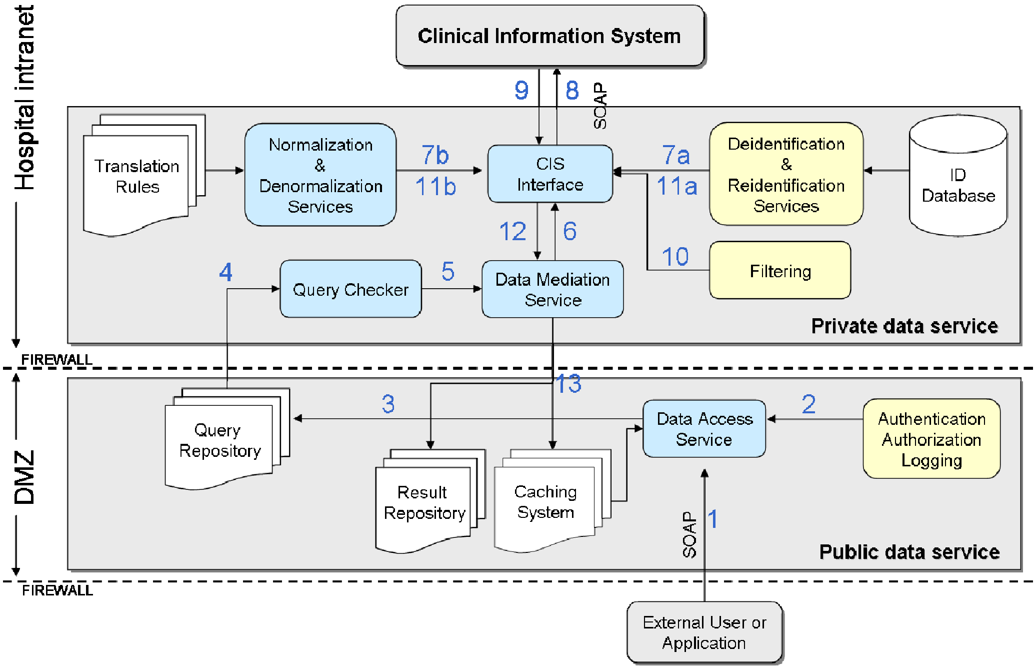

Figure 1: Access to @neurIST data stored in the clinical information system

5. Security and privacy

The CIS maintains and manages personal medical records in a digital format, containing the first instance information relating to the current and historical health, the medical conditions and the medical tests of its subjects. Various measures are taken to protect not only the privacy and confidentiality of patients participating in the project but also for the security of the whole hospital infrastructure.

5.1. Pseudonymization

Motivated by the need to store back derived data into the CIS (i.e. to re-identify the patient), the pseudonymization procedure was preferred instead of anonymization of the patient data for the secondary use. The pseudonymization procedure is split into two steps as shown in the figure 2. In the first step, all information permitting direct identification of the patient in structured and unstructured data such as images are removed or reduced to the minimum required to carry out the research. This step is called depersonalization and data minimization. In the second step, a pseudonym is generated from the patient ID and attached to the depersonalized data and is reversible to re-identify the patient. The re-identification of the patient from the pseudonym is necessary for query management but also needed by the ethical committee to re-contact the patient when important finding is obtained from research. It is important to notice that the pseudonym generated is specific for a participant recruited and followed at a single clinical center.

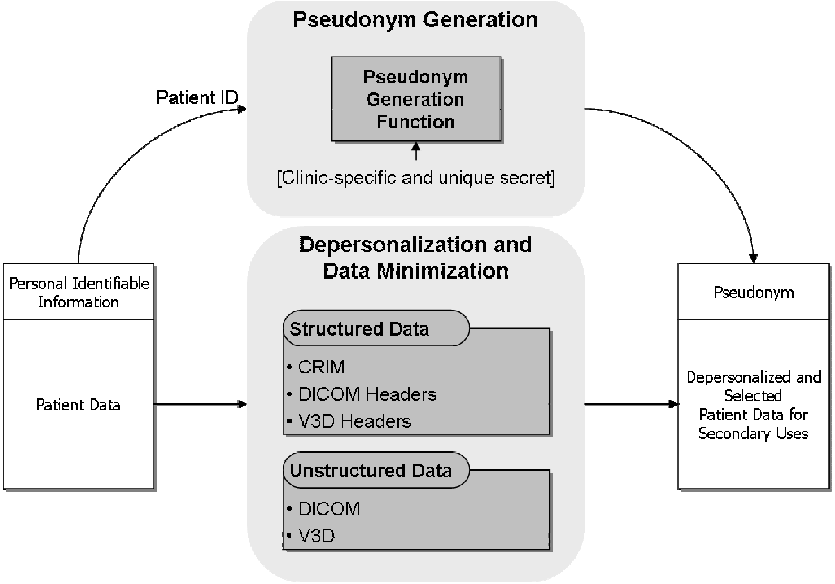
Pseudonym Generation
Pseudonym Generation Function
[Clinic-specific and unique secret]
Patient ID
Personal Identifiable Information
Patient Data
Depersonalization and Data Minimization
Structured Data
• CRIM
• DICOM Headers
• V3D Headers
Unstructured Data
• DICOM
• V3D
Pseudonym
Depersonalized and Selected Patient Data for Secondary Uses

Figure 2: Pseudonymization system of outgoing data

5.2. Access control

Even if the data accessible through the infrastructure does not permit direct identification of the patient, it should be protected against unauthorized access. Generally, clinical centers have their own security, access right management and privacy protection policy according to the role of the user [7]. However, these security models are designed for local use and should be combined with distributed security model to authenticate and authorize external users. Within a security domain, all the security is concentrated and placed under the responsibility of this domain. Between different security domains, the chosen approach consists in designating, in each domain, an security entity (Security Token Service), who will be in charge of issuing and verifying short-term security tokens with the entities of the other security domains.

Accessing an operational CIS may also affect its operational status if the query itself poses security issue. Incoming queries are analyzed to prevent such incidents and only valid queries are executed on the CIS. The access control is also strengthened with a control of outgoing results and can be considered as a complement of the privacy protection. Such filtering technologies can assist in preventing the unintentional disclosure of personal identifying information due to issues such as unidentified flaws in the depersonalization component. A policy for statistical queries might for example reject result sets which contain less than four entries to be exposed to the requesting user.

5.3. Logging and monitoring

Another important component in the architecture is the logging and monitoring of external requests to the public data service and onwards to the CIS. Even if the CIS has a logging and monitoring mechanism, it is imperative that all queries to retrieve or store data are logged and monitored to identify and intervene potential misuse of the system. This process is managed at the public data service level.

6. Discussion and conclusion

In this paper, we proposed the first implementation of the infrastructure for a better management of the cerebral aneurysm. The infrastructure permits to share clinical data acquired inside the hospital and integrate derived data produced by researchers. The main advantage of the implementation is its re-usability, scalability and lower maintenance cost. Most of the implementation are generic and can be adapted for other disease management. The security and privacy of the patient and the CIS security has been discussed. The implementation of such infrastructure requires an open access to the CIS from technical and political point of view. The next step is the integration of infrastructure in the whole @neurIST IT infrastructure.

7. Acknowledgements

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.

8. References

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

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

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

[4] J Iavindrasana, A Depeursinge, P Ruch, S Spahni, A Geissbuhler, H Müller. Design of a decentralized reusable research database architecture to support data acquisition in large research projects. Medinfo. 2007;12:325-9.

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

[6] A Arbona, S Benkner, G Engelbrecht, J Fingberg, M Hofmann, K Kumpf, G Lonsdale, A Woehrer. A service-oriented grid infrastructure for biomedical data and compute services. IEEE Trans Nanobioscience. 2007;6(2):136-41.

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