Access to Clinical Information Systems for Research in Life Science – Security and Privacy Considerations –

Jimison IAVINDRASANA[a,1], Luigi LO IACONO[b], Henning MÜLLER[a], Ivan PERIZ[c], Paul SUMMERS[d] and Jessica WRIGHT[e]
[a] *Hôpitaux Universitaires de Genève, Switzerland*
[b] *NEC Laboratories Europe, Germany*
[c] *Hospital Clínic i Provincial de Barcelona, Spain*
[d] *University of Oxford, England*
[e] *Durham University, England*

Abstract. To accommodate the complexity of contemporary and future research in life science, prospective clinical research must be conducted across multiple institutions, as multi-institutional data collection allows statistically significant numbers of cases to be collected in a shorter period of time, incorporates a more diverse study populations and reduces the bias induced by any individual researcher. Common architectural patterns and Data Grid technologies to federate and integrate distinct data sources have drawbacks especially in terms of setup and maintenance costs. An enhancement to this state of the art in health-related Data Grids has been proposed which introduces a decentralized database architecture allowing the access to the CIS or parts thereof. In this paper, security and privacy considerations unique to such an architecture are raised and discussed, including possible solutions to achieve the security level of common and widely adopted architectural approaches.

Keywords. CIS, EHR, E-health, Grid, HDBS, Privacy, Security

1. Introduction

The many developments in the domains of genomics, proteomics and medical imaging during the last ten years are difficult if not impossible for an individual clinician to track and consolidate, even within a single speciality. As a result of the fragmentation of the information on these developments, diagnostic and therapeutic practice, as well as the development of new treatments can be hindered. To develop novel diagnostic or therapeutic processes, prospective clinical research must be conducted across multiple institutions as multi-institutional data collection allows statistically significant numbers of cases to be collected in shorter times, incorporates a more diverse study population, and reduces the bias induced by any individual researcher. Linking distributed, multi-

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

format, and multi-scaled data from genetics, proteomics, the clinical domain (lab reports, images, clinical history), and epidemiological data for diagnosis and treatment is a new challenge in the biomedical informatics domain [1]. The @neurIST[2] project is working on the integration of information across several molecular, cellular, tissue, organ, person and population levels. This data is maintained in distinct repositories and is accessible through innovative IT infrastructure based on Grid and SOA (Service-oriented Architecture) related technologies applied and developed for this purpose. Inter-relating computational tools will make use of this federated data for diagnosis, analysis and decision support for the preventive treatment of cerebral aneurysms.

Designing a research database infrastructure for managing heterogeneous data formats, at distributed sites, while allowing for access to the entirety of data by multiple researchers, despite it and they being located in various geographical regions is one major challenge to be resolved towards prospective research in life science. Data integration is a growing issue in this context and may be defined as the problem of combining data originating from different sources, and providing the user with a unified view of these data [5]. Two main categories of database architectures exist for integrating data and conducting multi-institutional clinical research: centralized and distributed [2]. In the centralized architecture, data resides in the same physical place; and in the distributed architecture, data are stored at various independent sites and connected via communication networks. The main drawback of a centralized architecture relates to maintenance. The data are collected at different sources and have to be transformed into a consistent and coherent representation prior their integration in the repository. However, it is associated with a higher performance than the distributed architecture. The main advantage of the distributed architecture is the independence of each participating institution to keep control of the data they generate and store. A central problem in either context is that in the real world it is unlikely to have the same database management system using a common data schema installed on the same operating system across independent institutions, leading to a heterogeneous database system (HDBS) [3]. To federate such decentralized data sources, additional infrastructure components are required – which are typical of those developed in the Grid context – including e.g. mediators to enable access to heterogeneous data sources [4]. In multi-institutional prospective clinical research, a decentralized HDBS architecture is most flexible and the major architectural pattern in use.

Another major issue in multi-institutional research in life science is sharing clinical data with external researchers, which creates many problems especially with respect to patient data privacy and confidentiality. For these reasons, a Clinical Information System (CIS) is generally a closed system. As a result, in multi-institutional clinical research, even when data is collected inside the individual CIS the relevant records are generally exported to a network area accessible from the Internet outside the CIS and the hospital network (also known as demilitarized zone, DMZ) to create a dedicated database for one particular secondary use which becomes part of the HDBS. Conversely, specialized data collections such as genetic sequencing are rarely incorporated into routine medical practice, in part because these data would often need to enter the hospital domain from outside institutions. Similarly, connections with existing public knowledge bases such as SWISSPROT[3] are rare.

[2] http://www.aneurist.org/
[3] http://www.expasy.org/sprot/

Often, half of the money invested in large scale clinical studies is used for data creation and maintenance, especially when adopting architectures which promote the use of dedicated replicas of parts of the CIS relevant to the research. In [8] a design of an enhanced HDBS has been proposed, introducing a decentralized and reusable research database architecture to overcome this management overhead. The architecture concerns an "open" CIS: amenable to change and direct data access, in the context of a multi-institutional research project (see Section 2). In this paper, security and privacy considerations unique to this architecture are raised and discussed including possible solutions to achieve the security level of common approaches (see Section 3).

2. Access to CIS

The research database architecture introduced in [8] has three major components (see Figure 1):

- the Public Data Service located in the hospital's DMZ,
- the Private Data Service located in the hospital's intranet, and
- the CIS located in the hospital's intranet as well.

The public data service is the access point for the data stored in the CIS. It is located in the hospital's DMZ and thus is accessible from public and open networks such as the Internet. It authenticates the user or the application querying the database. All transactions are monitored and logged to audit trails. The public data service cannot communicate directly with the CIS. It instead queues all authenticated and authorized incoming queries in a repository.

The private data service is responsible for fetching the incoming queries from the repository located in the DMZ, transforming them to reflect the internal data structure and representation (mediation and de-normalization), and sending the queries 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 the normalization and 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.

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.

The clinical centre's CIS serves as the local database and is only accessible by the private data service. The CIS stores primary data acquired inside the institution and patient-related derived data produced by researchers. The primary data structure and representation are not necessarily identical to those agreed in a project due to the re-use of existing patient data in the CIS and also the local data representation and language. The mediation service and the normalization and de-normalization service implemented in the private data service handle this issue. 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, and subject to both the clinical centre's policy and the patient's decision on whether they agree to the return of research results relevant to their health.

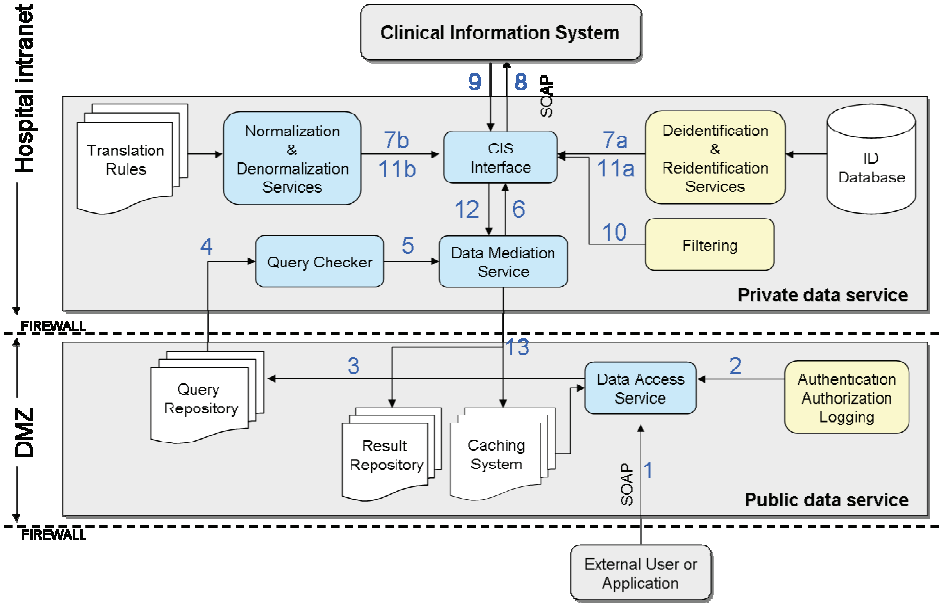

Figure 1: Clinical centre's database architecture and information flow to access the CIS for research in life science

This innovative and promising research database architecture possesses specific security and privacy requirements which need to be analyzed carefully to understand their impact and influence to the usability of such approach.

3. Security and Privacy Considerations

The CIS maintains and manages personal medical records in a digital format, containing in the first instance information relating to the current and historical health, medical conditions and medical tests of its subjects. It is primarily used in the treatment context in which the patient's identity data is protected by medical secrecy. But the CIS can also serve as a basis for other purposes denoted as secondary uses, such as clinical or epidemiological research projects, health care research, assessments of treatment quality or health economy.

Characteristically, for secondary use, the patient data leaves the control- and protection-sphere of the medical secrecy. As patient data contains some very sensitive personal data, in order to make it available outside of the treatment context, various legal and ethical aspects have to be considered. First of all, it is ideal that the patient gives informed consent to use his data for a secondary purpose. Second, the Personal Identifiable Information (PII) of the patient has to be removed from the Electronic

Health Record (EHR) before use in secondary contexts is allowed. Accompanying this, the data released should not extend significantly beyond that needed to address the research question to hand.

3.1. Pseudonymisation

Depersonalization must be performed prior to secondary use. Yet although the identity of the patient is not always important in secondary contexts, it is not always desirable to simply anonymise the EHR. There are many secondary use scenarios for which ability to form the correct association between a single patient and his EHR from distinct sources or distinct points in time is essential. Examples are the provision of follow-up data at a later point in time, the withdrawal of samples or data after a specific patient's request or the quality control of the data such as checking for double-entries. This usually prevents anonymisation and demands pseudonymisation schemes instead. Furthermore, in some research and ethical frameworks it may be necessary to maintain the possibility to re-contact patients in the event of results relevant to their health being obtained. Here, a reversible pseudonymisation system is required. In general, depending on the kind of research network and its requirements, distinct procedures for anonymisation or pseudonymisation are appropriate.

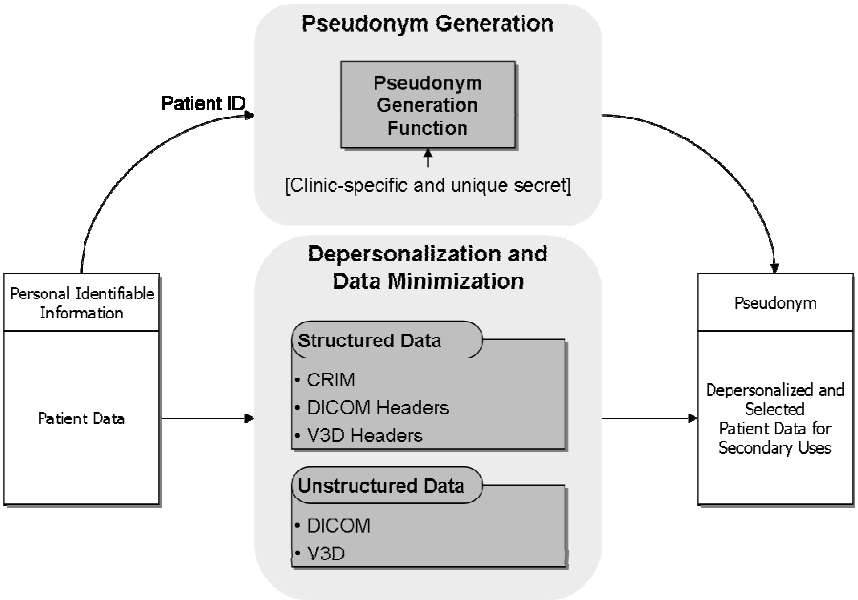

Figure 2: Conceptual view of pseudonymisation

The pseudonymisation process can be split into two steps (see Figure 2). In the first step, the data has to be depersonalized, meaning all contained PII has to be removed (see section 3.1.1). In the second step, a pseudonym is generated and attached to the depersonalized data (see section 3.1.2). In the following subsections the unique characteristics of the proposed research database architecture in respect to depersonalization and pseudonym generation are discussed.

3.1.1. Depersonalization and Data Minimization

The depersonalization step within the pseudonymisation process described above forms a natural point at which to reduce the information being released to the minimum data required to carry out the study in keeping with the data minimization and relevant principles of the European Data Protection Directive 95/46/EC. Functionally, it is useful to remember patients' privacy before considering what data items to depersonalize and so to reduce the data to the absolute minimum required for the targeted research prior to depersonalization. The less data that is exposed, the lower the risk of unauthorized intentional or unintentional re-identification. Taking genetics as an example: in studies, in which the whole DNA sequence is determined, but isn't needed to fulfill a given query the research, the response to the query must be reduced to the minimum subsequences required. Linked to this is the important principle of ensuring that the data provided is relevant to the research project. For example, if data on religion or a rare genetic illness is not useful for a medical research project, it should not be provided.

The depersonalization of patient data can be categorized according to the data type. In the case of structured data, the depersonalization is performed by removing PII from the data set or replacing it with something unlinkable such as e.g. a pseudonym. This approach needs also to be applied to structured data components within unstructured data like the header information contained in medical images. Unstructured data needs to be addressed as well but the nature of the depersonalization may be less straightforward. For example, images may need to be manipulated to prevent accidental recognition or the effective isolation (e.g. via genetics) of the patient as an individual. In the case of images, this is commonly achieved by cropping the image to the region of interest or introducing noise to hide identifying characteristics such as the face which are not subject of the research.

Whereas depersonalization of structured data does not need to be specifically adapted for the proposed research database architecture, some performance-related issues arise in respect to supporting its on-the-fly properties. The depersonalization of unstructured data on the other hand does require specific architectural consideration, due to the computational nature of this task. In order to provide depersonalized medical images in a timely manner, a caching mechanism may be needed to maintain copies of depersonalized images to serve subsequent requests efficiently.

3.1.2. Pseudonym Generation

Depending on the type of research network and its requirements the range of adequate pseudonymisation approaches might range from simple one-pass one-way pseudonymisation schemes to two-pass reversible pseudonymisation schemes. An overview on pseudonymisation schemes can be obtained from [9, 10].

In generating pseudonyms for the proposed research database architecture, specific characteristics have been taken into account. Pseudonymisation schemes which rely on a centralized pseudonymisation service such as dual-pass pseudonymisation approaches possess issues in terms of performance, due to the additional communication required. This is also true for the re-identification of patients, which is very unique to the proposed research database architecture as will be discussed in the subsequent paragraphs.

3.2. Re-identification

When clinically relevant information, which might have a direct impact on the treatment or future health of a patient, arises during the course of a study, ethical principles favor re-contacting and informing all relevant patients about the findings (as far as the return of results is subject to the clinical centre's policy since this can have significant cost implications with minor clinical benefit as well as the patient's consent since the additional knowledge may for instance affect a person's ability to obtain insurance or in the case of genetic results warrant unaware family members being informed of the patient's condition). In this setting, re-identification of a patient from their pseudonym may be required.

The proposed research database architecture already requires the deployment of a reversible pseudonymisation scheme in scenarios, where research data obtained through knowledge discovery processes are to be stored back in the CIS (including for example estimates of treatment and non-treatment risk based on information available at a specific point in the diagnostic process). Note, that the re-identification here is required to retrieve a form of patient ID which can be utilized to manage the patient's data in the CIS. This differs from other types of research database architectures in which the re-identification of patients is only performed for re-contact purposes. This process is usually very authoritarian, controlled by an ethical committee which decides whether the information should be released for re-contact or not. This includes the re-identification of the patient from the pseudonym. Consequently, clear and possibly distinct access control mechanisms are needed for the re-identification function itself and for the admission of derived research data back into the CIS. For the latter case it needs to be ensured, that this kind of data is managed only by authorized research personnel.

A typical strategy for the control of the re-contact of a patient is that a specific action on the part of the ethical committee is required in order to retrieve the patient's identity from the pseudonym by the adoption of adequate cryptographic primitives and a suitable key distribution. When applying for example public key encryption to generate the pseudonym, the public key in possession by the pseudonym generation component is used to generate the pseudonym and the private key in possession by the ethical committee is used to decrypt the pseudonym and gather the patient's identity from the pseudonym.

To allow for such a strict re-identification process in the proposed research database architecture, specific measures needs to be implemented since the re-identification is performed instantly – as for the pseudonymisation – and is hence not under full (technical) control of the ethical committee. Thus, in order to establish a security solution which achieves a comparable level of security for the re-identification of patients as in conventional research database architectures a strong authorization system and strictly defined access policies need to be in place. The process of re-contacting patients to feedback research results therefore is as follows: after the necessity and methods of re-contacting a patient have been favorably reviewed by an ethical committee (often internal to the research project), the Study manager is authorized to instigate the re-identification function to retrieve the patient ID and then the access rights to the research data within the CIS are reassigned to permit access by the health professionals in charge of treating the respective patient, who should follow appropriate methods for the return of results, which could include counseling.

It is obvious, that the described means include more potential for intentional misuse – in particular by insiders – then the separation of duties depicted by the pseudonym management example based on public key encryption given above. Thus, additional safeguards need to be considered to uncover abuses with certainty, to trace it back to the initiating source and to ensure non-repudiation. Before discussing these issues in section 3.4, access control aspects are examined in the following.

3.3. Access Control

Even though the patient data available for secondary uses does not contain PII and should not allow the linkage to a particular patient without disproportionate effort, the usage of this data is bound to the specific purposes of the associated study and must therefore be protected against unauthorized access. Hence, suitable access control policies need to be implemented and enforced.

From a conceptual viewpoint, the access control system for multi-institutional research in life science should follow the common patterns and principles for distributed cross-domain or cross-border information systems (including aspects such as the trust model), in which heterogeneous environments of multiple distinct institutions need to be interconnected. A widespread approach is to apply hybrid security models to such systems to respect the various security domains and their different security policy implementations. Generally, clinical centers have their own security, access rights management and privacy protection policy according to the role of the user [6], and have the know-how concerning data access and communication using standards such as HL7, DICOM, IHE, or business components such as Web services [7]. A hybrid security model underlies the combination of a local model and a distributed model. In other words, within a security domain, all the security is concentrated and placed under the responsibility of this domain, whereas between different security domains, the chosen approach consists in designating, in each domain, an security entity (which is known as Security Token Service, STS), who will be in charge of issuing and verifying short-term security tokens with the entities of the other security domains.

Figure 3 shows such a distributed architecture consistent with our application domain. Here the local security model relies on national e-health infrastructures such as Health Professional Cards (HPC) and local authorization systems to obtain a security token from the local STS which then can be used to access the distributed services residing in a distinct security domain or even in a distinct country.

The following discussion on access control focus on the distributed part only, since this is the unique aspect directly related to the proposed research database architecture. That is also why the indirection from the public data service to the CIS is not treated further, because this step relies on local access control mechanisms. The only addition to the local access control system to be mentioned is the requirement of access policies in the CIS for the derived research data. As indicated in section 2, derived research data is only to be accessible by research personnel. It is not viewable or accessible by unauthorized users such as clinicians or patients until it has undergone some form of clinical validation, ethical consideration and approval process.

The distributed part of the access control system of the proposed research database architecture is very similar to the common HDBS and does only include specifics in the filtering subsystem. Nevertheless, the following subsections describe the distributed access control system as a whole since it is an emerging approach to federate cross-

domain information systems which has not yet been widely adopted in research in life sciences.

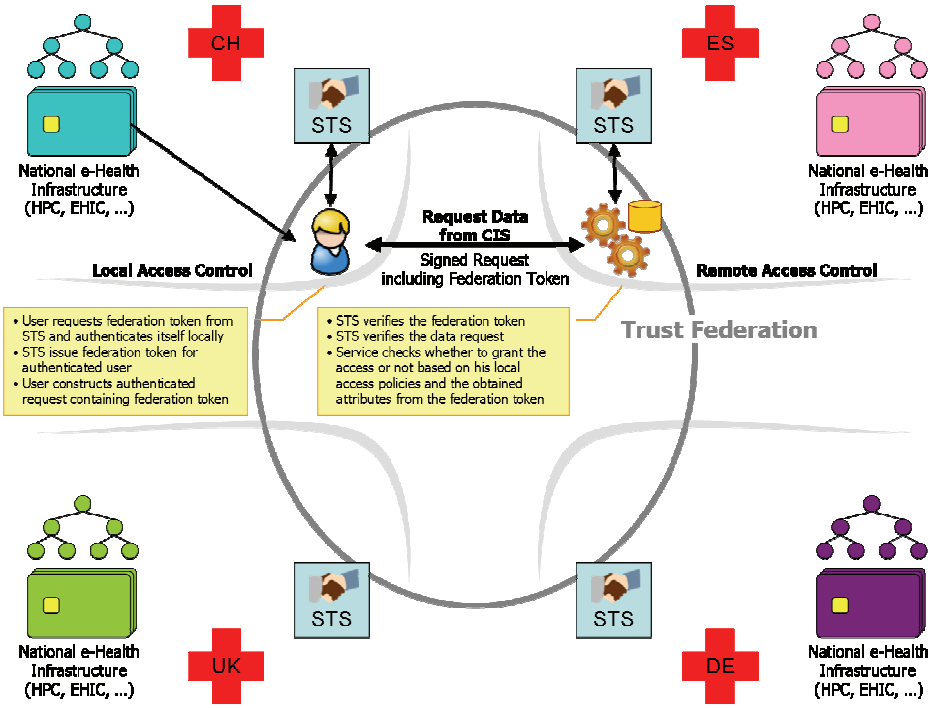

Figure 3: Access control architecture

3.3.1.Authentication

The authentication of external queries to the CIS is a multi-step process. The incoming request differs in form and content depending whether the data query is to be sent by the private data service directly from the CIS or via an instance of a data resource mediator.

In the first case, the data request contains a federation token signed by the requestor's STS and the data query itself. The whole data request is signed by the requestor with his personal signature device obtained by his national e-health infrastructure. The federation token certifies requestor-specific and project-wide attributes as well as the requestor's public key which corresponds to the private key stored securely in the personal signature device. The data request is then send to the CIS (see Figure 3). The public data service first verifies the federation token by involving its local STS. If the verification result is positive the public data service obtains authenticated attributes as well as the authenticated requestor's public key. This can then be used to authenticate the whole request.

In the second case, the described data request is enhanced with a delegation token signed by the requestor addressing the mediation service, enabling it to distribute mediated data requests to the attached data resources on behalf of the requestor.

For the sake of simplicity, but without the loss of generality, the following discussions are focused on the interaction between the data requestor and the CIS without any mediation involved.

3.3.2. Authorization

The authorization of access to the CIS is enforced locally at the clinical centre that exposes its CIS (denoted as remote access control in Figure 3, due to the viewpoint). Thus, the clinical centre maintains full control and can manage access privileges on its own. This can include for example black-listing certain users. The clinic can locally restrict the access privileges of individuals even over-writing global access rights within the research network that would have been sufficient to grant access. This is necessary in cases in which a certain user misbehaves, to prevent research results being accessed by clinical staff members, or where the local clinical center elects to selectively restrict access to individuals from a participating institution.

The process of decision making relies on the authenticated attributes within the federation token, which can be e.g. roles or context-information like the targeted research network as well as the requestor's location which might e.g. be important in cases where the involved research institutes are located in countries with different privacy regulations.

3.3.3. Filtering

The critical process for the proposed research database architecture is the inspection of the query to the CIS by the Query Checker, to analyze whether the query itself poses security issues or not. Such filtering of incoming data queries is necessary, since even careless queries can have a serious effect on the operational status of the CIS which would of course influence also the patients' treatment. To prevent this and to introduce a countermeasure for this kind of incidents various approaches can be implemented within the Query Checker ranging from a set of pre-defined and static queries to full freedom of query specification. This includes the obvious trade-off between ease of implementation and the ease of management of and range of supported research queries.

To further strengthen the access control system, data access policies should not be enforced only on the incoming access requests, but also in relation to the resulting responses to those requests; thereby controlling outgoing data resources. Such filtering technologies, as complementary functionality of a comprehensive access control system, address requirements mainly in the context of privacy leakage protection. They can assist in preventing the unintentional disclosure of PII due to issues such as unidentified flaws in the on-the-fly depersonalization component. Another possible usage scenario is the enforcement of access policies to the result set of a particular query. Here, a policy for statistical queries might e.g. reject result sets which contain less than four entries to be exposed to the requesting user.

3.4. Logging and Monitoring

Another important component in the introduced research database architecture is the logging and monitoring of requests from the Internet to the public data service and onwards to the CIS. Since the CIS is queried through the introduced database

architecture directly, it is imperative that all queries to retrieve or store data are logged and monitored to identify and intervene potential misuse of the system.

4. Conclusion

Prospective clinical research conducted across multiple institutions is necessary in order to cope with the ever increasing complexity of medical data and to achieve the required accuracy and power in research results. Health-related Data Grids provide the technological basis to integrate distinct data sources and henceforth to enable multi-institutional research in life science. However, the current state of the art in Health Grids has drawbacks especially in terms of setup and maintenance costs. An enhancement has been proposed in [8], which introduces a decentralized database architecture allowing the direct access to the CIS or parts thereof, without the need for managing dedicated replicas of the CIS. The impact to the security and privacy requirements has been discussed. It has been shown, that the specific security and privacy requirements can be resolved adequately to achieve the security level of common approaches further underlining the suitability and importance of the proposed decentralized database architecture for prospective multi-institutional research in life science.

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.

The authors would like to thank Bernice Elger for many fruitful discussions as well as the anonymous referees for their constructive comments.

References

[1] Hunter P, Smith N, Fernandez J, and Tawhai M. *Integration from proteins to organs*: the IUPS Physiome Project. Mech. Ageing Develop. 2005:126(1):187-192.

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

[3] Sujansky W. *Heterogeneous Database Integration in Biomedicine*. J Biomed Inform 2001: 34(4): 285-298.

[4] 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: 317-320.

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

[6] 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: 251-256.

[7] 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 2001:10:735-739.

[8] Jimison Iavindrasana, Adrien Depeursinge, Patrick Ruch, Stéphane Spahni, Antoine Geissbuhler, Henning Müller, *Design of a Decentralized Reusable Research Database Architecture*, medinfo 2007, Brisbane, Australia, 2007.

[9] Pommerening K, and Reng M. *Secondary use of the EHR via pseudonymisation*. In: L. Bos, S. Laxminarayan, A. Marsh (Eds.), Medical Care Compunetics 1, IOS Press, Amsterdam, 2004: 441-446.
[10] Lo Iacono L. *Multi-centric Universal Pseudonymisation for Secondary Use of the EHR*, HealthGrid 2007, Geneva, Switzerland, 2007.