Integrating Data Custodians in eHealth Grids – A Digest of Security and Privacy Aspects

Jochen Fingberg[1], Marit Hansen[2], Markus Hansen[2], Henry Krasemann[2], Luigi Lo Iacono[1], Thomas Probst[2], Jessica Wright[3]

[1]C&C Research Laboratories, NEC Europe Ltd.
Rathausallee 10, 53757 Sankt Augustin, Germany
{fingberg, lo_iacono}@ccrl-nece.de

[2]Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein
Holstenstraße 98, 24103 Kiel, Germany
{LD10, LD63, LD9, LD91}@datenschutzzentrum.de

[3]School of Law, University of Sheffield
Conduit Road, Sheffield, United Kingdom
jessica.wright@sheffield.ac.uk

Abstract: This work introduces Grid computing, shows its use in eHealth environments and elicits trends towards the integration of custodians in eHealth Grids. It considers security and privacy requirements for the use of Grid computing in eHealth scenarios and discusses the possible integration of different types of data custodians. Finally the paper concludes and gives an outlook on the development and deployment of eHealth Grids in the near future.[1]

1 Introduction

Grids can combine aspects of clustering (multiple physical entities operating as one logical entity) and virtualisation (multiple logical entities operating on one physical entity). In a Grid, multiple logical entities (Grid Nodes, GNs) that are not centrally administered interconnect, e.g., via Internet, and combine their resources to perform – among others – computational tasks [Fo02].

In life sciences, e.g., there is a compelling demand for the integration and exploitation of heterogeneous biomedical information for improved clinical practice, medical research, and personalised health-care. In this context Grid technologies are becoming a common infrastructure in order to federate different data sources to enable researchers as well as medical professionals to query and access distributed information in a unified and integrated way and to seamlessly provide computing resources [HGA04].

---

[1] This document is published under Creative Commons "Attribution-NonCommercial-NoDerivs 2.0" License (cf. http://creativecommons.org/licenses/by-nc-nd/2.0/de/). A longer version of this text with a much more comprehensive analysis is available online: cf. [FH+06].

The extremely distributed nature of Grids makes the control of personally identifiable information (PII) of a patient particularly difficult. Furthermore Grids incorporate an "amplifying" character meaning that the federated and integrated infrastructure also might enable unauthorised data collection and correlation which might enable mining pseudonymised patient data and turning them into PII by accumulating identifiable information. These specific conditions have to be considered when designing or deploying pseudonymisation mechanisms.

The medical practitioner sending information to the Grid has a duty of confidentiality towards his/her patients. The World Medical Association International Code of Medical Ethics states that a physician shall "preserve absolute confidentiality on all he knows about his patient even after the patient has died" [WMA49]. The patient him/herself has a right to informational privacy. This can be enumerated in many different ways, but is often conceived as a right for the individual to control "to what extent information about them is communicated to others" [We+70].

Considering these principles it is evident that the patient must give proper informed consent to the processing of information, and be able to control through choice the information that emanates from him/her. These principles can, however, be outweighed by other competing interests. These could include the patient's best interests and the public interest. It is often argued, for example, that medical research is in the public interest. However, the principle of personal autonomy is frequently highlighted as important in recent international statements on bioethics, for example the UNESCO Declaration on Bioethics and Human Rights states in Article 5 that "[t]he autonomy of persons to make decisions … is to be respected" [UN05]. Informed consent is therefore still seen as a *prima facie* rule when it comes to either medical treatment or research.

This means that, as far as possible, patients should be informed about and allowed control over the processing of their own information. Appropriate technical and security measures must be put in place to ensure safeguards to help protect privacy, as patient consent does not obviate the requirements that data be kept securely.

Current practice is that the patient's PII is provided for medical research after signing an informed consent form. This is bound to well-defined purposes and mostly to a rather short-term time-frame in accordance with the length of the research project. As long-term research collaborations are becoming increasingly important in some areas of medical research, the handling of informed consent forms might not be as feasible. The patient may not be able to take in all the information necessary to give proper consent to these future uses of his/her data, despite them being known at the time of collection. A larger discussion is needed on the impact of new Grid technologies and ICT-driven research on patient understanding. It is also questionable how it can be ensured that the patient data are used solely for the agreed purposes and that the patient does not lose the control over his data in the context of the Grid.

To overcome these problems, trusted third parties in the form of electronic data custodians in charge of taking care of confided data or managing security configuration might provide an efficient solution.

2 eHealth Grids and Custodians

2.1 Scenarios

Scenarios for eHealth Grids can be categorised according to their goals which include the improvement of clinical practice, medical research and personalised health-care. To enhance, for example, clinical practice and medical research, new technologies are used to incorporate imaging and simulations into diagnosis [HGA04]. Assume, for example, that including blood flow simulations based on scanned images in a patient's examination would aid the clinician to deduce the most appropriate and maybe cost-efficient treatment plan. Since these kinds of simulations are very complex and require a lot of computational resources, they are usually conducted by specialised service providers located outside of the attending hospital or physician. Thus, eHealth Grids serve – instead of an ordinary laboratory – the needs of the hospital or medical practitioners, and are commissioned by them.

- In the *clinical treatment scenario* the patient is referred to a specialist who retrieves the patient's data for further analysis by invoking – possibly located outside the specialist's domain – corresponding compute, analysis and simulation services. Furthermore the specialist's decision-support system will include other information sources so that he/she is finally able to give the diagnosis and then suggests treatment options.
- A path through a *medical research scenario* may include the federation of various biomedical data sources such as gene sequences in order to find correlations between the patient characteristics and the targeted research goal.
- Industrial *scenarios including drug discovery or health equipment design* may access biomedical databases and medical research resources in order to improve their existing products or even develop totally new ones.

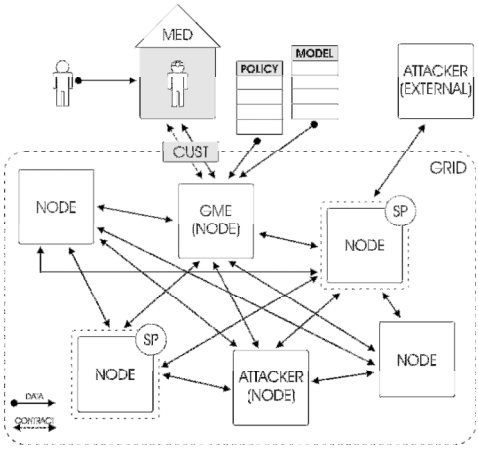

Figure 1: Main Roles in the Clinical Treatment Grid Scenario

Focusing on the clinical treatment scenario, the main roles, machines, and objects involved in health-care environments supported by eHealth Grids can be deduced (see Figure 1): The *patient* presents himself and thereby his PII to a *medical practitioner (MED)*, e.g., in a hospital. To analyse the data, the medical practitioner establishes a contract with a *Grid provider* who is operating a *Grid NODE (GN)* as a *Grid Management Entity (GME)*. The PII as well as a processing *model* (offered by the *model provider* who also may be the practitioner himself) and a *policy* get transmitted to the GME. The GME allocates resources within the Grid and transmits data and processing instructions to other GNs. The GNs might have a *security policy (SP)* of their own to prevent them from providing resources for unsolicited tasks. The GME should be aware of such restrictions and should not allocate resources.

The GNs are interconnected and – according to the processing instructions they received – communicate with each other to solve the computational problem and transmit the results back to the GME which combines them and forwards them to the medical practitioner. Furthermore, *attackers* (external or within the Grid) may want to intercept communication, gain access to the data, or manipulate data, processing and results to spy out or even sabotage certain research. A possible task of a *custodian (CUST)* is to manage identifiers or pseudonyms, see e.g., [PR+05]: As an intermediary, he/she pseudonymises medical data before they are transmitted from the medical practitioner to the GME and the GNs and de-pseudonymise their reported results before delivering them to the MED.

2.2 Relevant Privacy and Security Properties of eHealth Grids

According to our setting, the principal (i.e., the medical practitioner) is not in charge of processing patient data through a Grid model himself. So the principal commissions a Grid to perform the task by making a contract with the GME. This contract typically includes privacy and security requirements to be fulfilled by the GME itself and the GNs (for an example analysis of privacy and security issues in a medical Grid cf. [HC+04]). Involving a custodian will require appropriate contracts between principal, custodian and GME. Depending on the tasks of the custodian, some of privacy and security requirements might be burdened onto the custodian. An important factor is the degree of technical and/or organisational control of the GME over the GNs [cf. e.g., EGA05], or in other words the degree of autonomy of the GNs and their providers.

Even in the case of full digital control of the GME over the GNs there are differences to the scenario where the tasks are fulfilled directly in a laboratory associated to the hospital, as the GNs may be located in various places:

- The GME usually has no full physical control over the remote GN machines.
- The GNs may be located in multiple nations and therefore various legislative areas.
- The GNs require network access to exchange software and data which also opens ways for potential attacks.

The custodian can address some of these differences, e.g., by managing PII and keeping them away from the GME and the GNs. This would bypass problems of different

legislative areas. Also security issues such as software distribution and security configuration of GNs distrusting their GME might be handled by a custodian. However, other security aspects, e.g., the correctness and availability of computational results from GNs cannot be guaranteed by the custodian.

3 Approaches for Solutions

Especially in cases of PII to be distributed and processed, the GME must be able to safeguard that the data cannot be accessed by unauthorised third parties, e.g., operators of other GNs. As technical considerations will face worse problems than those the media industry tries to conquer, using Digital Rights Management or similar methods to protect PII will be required. This implies that the GME has to define a policy the GNs have to comply with and that the Grid environment, i.e. the set of protocols and tools that allow for interoperation of the nodes, has to have a policy enforcement mechanism. Also, the GNs should have a policy of "acceptable" tasks, e.g., "military research tasks will always be rejected while medical research tasks can be accepted if they do not have to do with birth control and if idle resources are available". Furthermore, the Grid environment should be aware of policies the nodes have to comply with, such as different legal implications in different countries. For example, PII may only be exported from the European Economic Area under certain circumstances.

Taking into account privacy and security principles, data processing in eHealth Grids has to be supported by a variety of measures. The current solution of getting the individuals' consent for processing their PII is questionable because this would require that everybody really understands the risks. Instead, data minimisation techniques should be applied which may rely on a third party as a custodian.

We understand "custodian" as an independent and trustworthy third party taking care of provided data (including software and configuration data), processing them in an agreed-upon manner, ensuring that provided data are used only for the agreed-upon purpose in the agreed-upon time period, are not forwarded to unauthorised parties, and are protected from external and internal attacks.

A primary task for a custodian could be to (reversibly) detach PII from the data for the duration of the processing. There are several possibilities, depending on the structure of the medical data and the computational task.

1. *Pseudonymisation*, i.e. exchanging names and other identifiers through the use of pseudonyms, and back, when transmitting data between practitioner and GME, to enable linkage between data relating to the same pseudonyms and to make re-identification possible e.g., for communicating the data processing results to the user. Pseudonymisation includes the administration of the relationship between identifying data and pseudonyms. If necessary, multiple pseudonyms can be used. This task might include not only the modification of meta data such as file names containing PII, but also modification of medical data that are considered as "originals", e.g., change of patient names in a X-ray picture. Depending on the data

structure, this can be a difficult task e.g., if the patient's name is included in the picture as a watermark.

2. *Segmenting* the computational tasks and processes and dispatching them to the GME or GNs in a way that the tasks reveal no PII, (e.g., dispatching an image in small parts to different GNs/GME ([HGA04], Chapter 8.11)). It depends on the computational model if segmentation into pieces can prevent identification.
3. *Pre- and post-processing* of computational tasks in a way that the remaining data cannot reveal PII. This is similar to the previous option, could also include "encryption" / "decryption" processes by manipulating the computational tasks and data in a way that they can be performed by the GNs or GME (e.g., a random change of scale), but neither input data nor results reveal PII.

Note that while reliably removing the link from PII to the related patient is relatively easy with most alphanumeric data, it is impossible with, e.g., biometric data such as a tomographic scan of a head, where – while the single "slice" does not necessarily allow one to recognise the person – the whole set of "slices" allows for computation of a 3D model making the person identifiable. Just removing the name from certain data does not make them anonymous, i.e., non-identifiable.

Another task for a custodian could be to offer a trustworthy archive for the huge data amounts which may occur in Grid computing, e.g., a central repository storing medical data from different hospitals as a third party (outsourcing). This requires multi-client capability of the registrar in order to keep files separate between different clients.

All these tasks (pseudonymisation, segmentation, pre-processing, storage etc.) may be executed by the hospital itself or on contractual bases by a data processor on behalf of the hospital. But there may be several advantages of using a custodian as defined in the beginning of this section:

- As a security and privacy expert, a custodian might have a deeper knowledge of the specific legal and security requirements and do a better service.
- A custodian can help to overcome internal conflicts in a hospital, including conflicts of interests of different departments (e.g., a demand of the finance department to get the actual address of a patient from research files).
- A custodian serving multiple hospitals can simplify and thereby cheapen the transmission of pseudonymised (or anonymised) data for research purposes in-between research facilities, as the data formats of the pseudonymisation/anonymisation are compatible. This includes also the uses of multiple GME infrastructures operating on the same data, e.g., for benchmarking computational complexity and accuracy of different algorithms.

From the legal perspective the custodian must not have own interests in the patients' data, but has to demonstrate its independency and reliability. Of course appropriate contracts which regulate the obligations of all parties involved must be set up before data processing starts. The custodian will also have to adhere to data protection law, including the support of the patient's right to access.

4 Conclusions and Outlook

Grid computing is becoming steadily more important for intensive computing tasks as well as distributed and federated data access. Its capabilities are very relevant for the eHealth sector, since modern medical treatment and research is demanding for high-capacity computing, e.g., for image processing.

Security and privacy requirements have to be considered when designing the workflow dealing with patient data. Preventing identifiability of patient data is not trivial. This is especially true for Grids, since due to their highly distributed nature, the control of the use of patient data becomes very complex if not impossible. Custodians can help to implement concepts for increased control and trustworthiness by keeping track of the patient's data and pseudonymising them in different ways, partially tailored according to the model to be used, and thereby separating the Grid context (calculation or data access) from the clinical context (treatment or research).

Many parts of security functionality are meanwhile addressed by Grid designers, but still aspects of multilateral security and legal requirements for privacy are not fully solved. Privacy policies which formulate requirements depending on location and national legislation of the Grid Nodes should be supported. Privacy Commissioners should be integrated early in Grid projects to give feedback in design phases.

References

[EGA05] Enterprise Grid Alliance Security Working Group: Enterprise Grid Security Requirements, Version 1.0, 2005. Available online at: http://www.gridalliance.org/en/workgroups/GridSecurity.asp.

[FH+06] Fingberg, J.; Hansen, M.; Hansen, M.; Krasemann, H.; Lo Iacono, L.; Probst, T.; Wright, J.: Integrating Data Custodians in eHealth Grids – Security and Privacy Aspects. 2006. Available online at: http://www.ccrl-nece.de/publications/paper/public/LR-06-262.pdf.

[Fo02] Foster, I.: What is the Grid? A Three Point Checklist. In GRIDToday, July 20, 2002. Available online at: http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf.

[HC+04] Herveg, J. A. M.; Crazzolara, F.; Middleton, S. E.; Marvin, D.; Poullet, Y.: GEMSS: Privacy and Security for a Medical Grid. In Proceedings of HealthGrid 2004, Clermont-Ferrand, France, 2004. Available online at: http://www.ccrl-nece.de/gemss/Reports/Herveg-healthgrid2004.pdf.

[HGA04] HealthGrid Association; Cisco Systems: Healthgrid White Paper; 2004. Available online at: http://whitepaper.healthgrid.org/.

[PR+05] Pommerening, M.; Reng, M.; Debold, P.; Semler, S.: Pseudonymization in medical research – the generic data protection concept of the TMF. In GMS Medizinische Informatik, Biometrie und Epidemiologie 2005; 1(3): Doc17. Available online at: http://www.egms.de/en/journals/mibe/2005-1/mibe000017.shtml.

[UN05] UNESCO: Declaration on Bioethics and Human Rights, 2005. Available online at: http://portal.unesco.org/shs/en/file_download.php/46133e1f4691e4c6e57566763d474a4dBioethicsDeclaration_EN.pdf.

[We+70] Westin, A.: Privacy and Freedom, London: Bodley Head, 1970, p. 7.

[WMA49] World Medical Association: International Code of Medical Ethics, 1949. Available online at: http://www.wma.net/e/policy/c8.htm.