Grant agreement no.211714

neuGrid

A GRID-BASED e-INFRASTRUCTURE FOR DATA ARCHIVING/ COMMUNICATION AND COMPUTATIONALLY INTENSIVE APPLICATIONS IN THE MEDICAL SCIENCES

Combination of Collaborative Project and Coordination and Support Action

Objective INFRA-2007-1.2.2 - Deployment of e-Infrastructures for scientific communities

Deliverable reference number and title: **D6.1 Distributed Medical Services Provision (Anonymization Service)**

Due date of deliverable: **Month 12**

Actual submission date: **31st January 2009**

Start date of project: **February 1st 2008** Duration: **36 months**

Organisation name of lead contractor for this deliverable: **University of the West of England, Bristol UK**

Revision: Version **1**

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

Intended Recipients

The WP6 workpackage entitled **"Distributed Medical Services Provision"** aims to design a group of *generic* services that can be used in a number of related medical applications. These will then be implemented in order to fulfil the neuGrid specific project requirements. The services will be built according to the design philosophy presented in the WP6 deliverable. This will help to enhance and promote their re-usability in other related applications.

This deliverable document presents a design philosophy that the generic services will follow, maps user requirements against suitable services and briefly presents a list of the services. An initial implementation of the services and their detailed API descriptions will be delivered in the year 2 deliverable.

The WP leaders, technical users and neuGrid developers are the intended recipients of this document. To a lesser extent, since indirectly concerned (through the natural abstraction of Workflow/ Pipeline authoring environments such as the ones proposed in WP6), neuro-scientists and prospective users (e.g. Pharmaceutical companies) as well as internal and external reviewers of the project activities, are anticipated as potential readers of this document.

11. The Anonymization Service

11.1 Introduction and objectives

The neuGrid platform is intended to handle a large quantity of sensitive data coming from heterogeneous sources. It is extremely important therefore, to ensure that medical information is not made available without appropriate ethical clearance. Such legal and ethical requirements mean that an efficient and common level of anonymization must be used throughout the project. Anonymization should therefore be considered at the following two levels:

a. Pseudonymization: This is defined by Wikipedia as "a procedure by which all person-related data within a data record is replaced by one artificial identifier (like a hash value) that maps one-to-one to the person. The artificial pseudonym always allows tracking back of data to its origins which is the difference with anonymized data, where all person-related data that could allow backtracking has been purged." [86]

1. Face scrambling: This is the process by which algorithms or manual processes are applied so that the face is removed from an MRI image, thus preventing the possibility of a subject being recognised.

In the context of neuGrid, anonymization is the means by which it is ensured that data cannot be traced back to the originating subject. This is a challenging task and is informed by the work of WP2 which is actively considering the level and type of anonymization that should be applied. Given the legal complexities that surround this subject, the design of the anonymization service has necessarily progressed at a somewhat slower pace than some of the other services. This is to be welcomed because it is essential that care be taken in complying with national and international policy. It is clear that for neuGrid to become a successful research infrastructure, such issues will need to be addressed thoughtfully and with care. It is with this purpose in mind that an initial outline is presented of what an anonymization service may contain. This is seen as an important step in stimulating discussion and driving a well founded design that will be refined and covered in more detail in subsequent project deliverables.

The purpose of the anonymization service is to facilitate the pseudonymization of the data that is stored within the neuGrid infrastructure in order to make it available to users, so that they can use it in their analyses whilst preserving the anonymity of the patients. The pseudo-anonymization process will include the checking of files to ensure that all the markers which can provide information to identify the patient are removed before the image can be made available. Most of the time it will be done by removing the image file's text headers containing metadata about the patient such as name, date of birth or any other information that could identify them. The face scrambling process will try to make sure that faces on images cannot be recognized. The anonymization service interacts mainly with the PACS Abstraction, the Grid Abstraction and the Face Scrambling services.

11.2 The Anonymization Service User Requirements

For a complete list of user requirements, the reader should refer to deliverable D9.1 due for delivery end of March 2009; a brief extract of the latest information that has been gathered

includes:

1.1.5 Provide software to enable the anonymization of data sets. The ability to easily anonymize the principal image fields defined by neuGrid ethical committee (if they are not already treated in some previous steps) ensuring that no identifiable patient information crosses the network (Images Scrambling and anonymization.)

1.1.14 The system should allow new anonymization methods to be applied as privacy standards evolve.

1.1.17 Something similar to the linux/unix "string" command should be executed on at least one image in each series, to check for "hidden" patient information.

1.1.18 If face scrambling is required, a surface rendering tool should be available and used to show the effect of the face scrambling.

1.4.5.5 The anonymisation process shouldn't be visible to the final user. This step could be done within the neuGrid consortium and should not be accessible (except for special privileged personnel) by the end of neuGrid.

1.4.5.7 Provide provenance information related to modifications made to a data set, provenance information may include modification made for quality control, ethical compliance, anonymization, any format conversions that were necessary and related information.

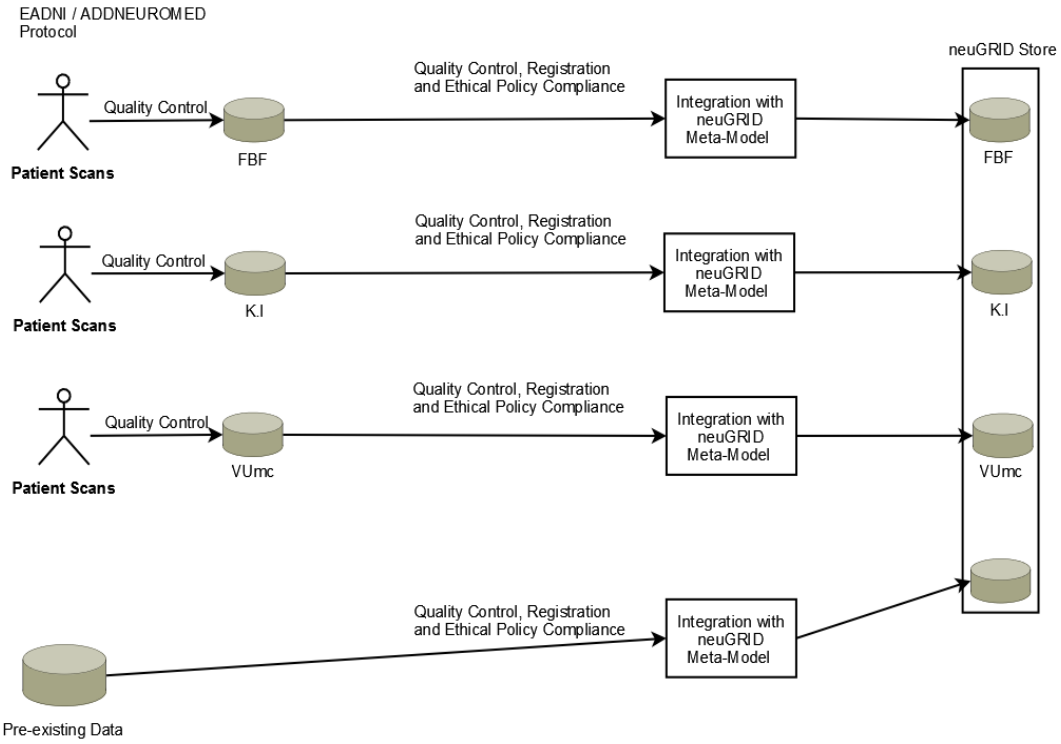

Figure 63: Data Entry into neuGrid

Data will primarily enter the neuGrid system in two ways (see figure 63). Firstly existing clinical

information from projects such as the E-ADNI Pilot and ADDNEUROMED will be uploaded into the neuGrid store. In this case a set of tools will need to be made available that will allow it to be integrated with the neuGrid standard for anonymization. The core labs (currently Karolinska, Brescia and Amsterdam) will play a vital role in this upload and verification process. Secondly as new images and data are acquired by individual clinical centres they will be passed to the core labs for processing before they are uploaded into the neuGrid store. The requirements that this places on the system are essentially the same as before in that the anonymization service will need to provide a set of tools which can be used to carry out and verify that appropriate anonymization has been carried out. WP2 has been tasked with considering the ethical policy for neuGrid and the **D2.3** document sets out some of the latest information regarding this. An excerpt of how the anonymization should be handled is now presented:

In order to anonymize clinical data and images, the following process should be put in place at the collecting centres level and at the core labs level:

11.2.1 Centre level

a. First anonymization of data subjects. It removes the identifiers listed below.
b. First coding
c. Transmission of clinical data and images to one of the core labs through CD, since the use of CD is more safe than web transmission.

11.2.2 Core lab level

a. Check of the anonymization procedure performed in the collecting centre;
b. Implementation of the anonymization process, in case of incorrect/incomplete anonymization procedure:
c. Second coding.

List of the identifiers to be removed:
(A) Names;
(B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes;
(C) All elements of dates (except year) for dates directly related to an individual excluding: -birth date (month and year admitted);&- exams/visits date (day, month and year admitted); -date of death; (month and year admitted);
(D) Telephone numbers;
(E) Fax numbers;
(F) Electronic mail addresses;
(G) Social security numbers;
(H) Medical record numbers;
(I) Health plan beneficiary numbers;
(J) Account numbers;
(K) Certificate/license numbers;
(L) Vehicle identifiers and serial numbers, including license plate numbers;
(M) Device identifiers and serial numbers;
(N) Web Universal Resource Locators (URLs);
(O) Internet Protocol (IP) address numbers;
(P) Biometric identifiers, including finger and voice prints;
(Q) Full face photographic images and any comparable images; and
(R) Any other unique identifying number, characteristic, or code.

Pseudoanonymization following the HIPPA recommendations [87][88], ie. By removing the 18

identity related fields, is not sufficient to ensure that all the sensitive information has been removed. The fact is that the DICOM norm is an open and extensible format, defining a list of common fields but also allowing manufacturers to add whatever fields they choose. Such unreferenced and often unusual fields have the potential to contain some identity related data. This problem could pose a serious risk, and one solution might be to define a list of allowed fields to be kept instead of a list of fields to be removed. One drawback could be that some potentially non-standard interesting field could also be removed by this approach. It seems, however less problematic and more secure than keeping some identity related field. The de-facing of images will be handled by the Face Scrambling service of the WP5, therefore this service will not be covered in detail in this document.

11.3 Description and Justification of the proposed architecture

Due to the security and privacy constraints, the architectural possibilities are somewhat restricted. It is clear that the hospitals will not allow unanonymized data to leave their walls, and then they will certainly want to rely on their own anonymization process, but as the level of uniformity of these anonymizations cannot be assumed to be stable and reliable, the neuGrid architecture will need to handle this itself.

There are essentially two main possibilities:

- a standalone application or
- a Java applet running locally in the browser of the user.

The main advantage of the standalone application architecture is that it would be available to users even when an Internet connection is unavailable. An important drawback to this is that it will require the deployment of the application within the local computing environment as well as the management of the subsequent updates as they become necessary. If some sites are not fully up-to-date there is a danger that it could lead to differences occurring between the image anonymization levels at different core labs. In order to reduce the potential for such issues we could implement an automatic update mechanism for the stand alone applications. However, this would require an Internet connection for each machine on which the application is deployed in order to reduce the manual deployment cost which could not be undertaken by any of the hospitals.

For the applet-based architecture the main advantage is that it would not require a manual deployment and could be fairly easily updated. The code responsible for the anonymization process will be downloaded at run-time, which ensures that an up-to-date applet from a central repository is always used. In terms of requirement 1.1.14 an applet would be a good means of providing the degree of flexibility that is necessary for the anonymization service to continue to evolve as best practice progresses. The main disadvantage to an applet-based design is that there are some potential security issues that are associated with such techniques that may complicate the implementation of the service. This method of delivery is also tied to an Internet connection and might not therefore be suitable in all hospital working environments, although it is likely that this can be overcome with some additional efforts. Further analysis will be undertaken during the next phase of system development in neuGrid.

As both solutions seem interesting but have their own pros and cons, and regarding the new additions made by Java SE 6 in the Update 10, allowing users to drag an applet on the desktop [89] and making it independent of a browser, it sounds quite realistic to plan two steps. Firstly as

the project needs to have an infrastructure that is fully available through a simple browser, and as a non up-to-date anonymization risk is not acceptable, an applet is the best solution. Secondly with the benefits of the aforementioned possibility made available by Java SE 6 Update 10, the applet should be made completely draggable, in order to be used in a context where an Internet connection is not always available. Obviously, in order to be able to implement the draggable feature quickly and easily, the applet needs to be implemented with this in mind from the outset.

Once the files have been edited to remove the identity related fields that are defined in the D2.3 document, they need to be uploaded into the neuGrid data store. When the image files enter the neuGrid architecture the system needs to ensure that appropriate pseudonymization has been carried out on them, and the images should be made available for the ID and study protocol check. If the files are valid, they can be processed by the Face Scrambling service, then they should be manually quality controlled, and lastly made available for processing by the neuGrid users. The data should have appropriate access restrictions set and applied where necessary. The data anonymization process can be explained through the following steps:

- The first step of anonymization will take place inside the hospital, a web site with a Java applet could remove all the unwanted fields.
- Once files have been anonymized they could either be burnt on to a CD for a later submission to the neuGrid architecture or be directly put into the platform over an Internet connection.
- Once the files have been submitted into the platform, they are sent through a Web Interface to a Web Service that will ensure that no unwanted fields are present.
- Then the files are processed by the normal LORIS workflow, there is an ID and Study protocol check, then the Face stripping happens and after they have been Quality Controlled, they are made available for processing.

The Workflow drawn in the Figure 64 describes all the anonymization steps that are required:

- an imaging device produces images that could be associated or not with an ID
- the images are stored into the DICOM server of the hospital, usually a PACS system
- the images are pseudo-anonymized using an applet
- the images are either burnt on a CD for a letter upload to the neuGrid architecture or sent directly through the Web
- the images are upload into MDM which will do a second pseudonymization and make them available to the LORIS software
- the IDs and protocol checks take place
- the images are face stripped
- the Quality Control is done
- the images are made available for processing by neuGrid users

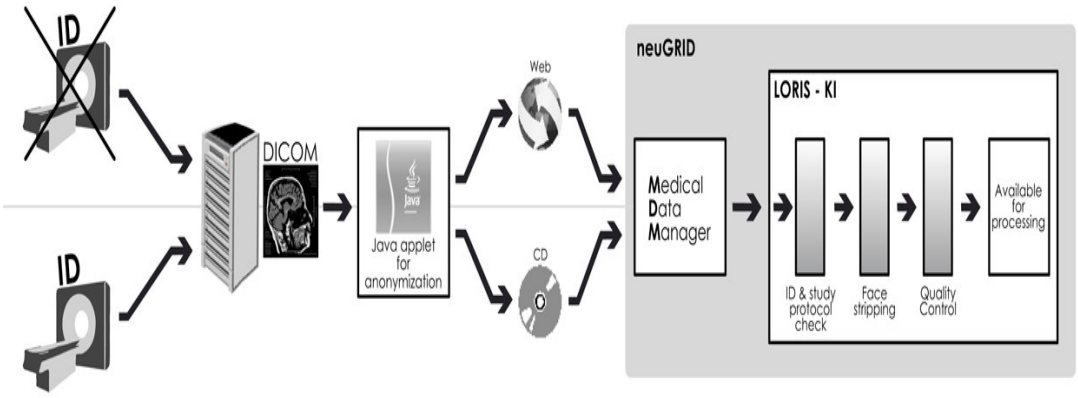

Figure 64: Workflow of the data acquisition process

The following components may be used within the Anonymization Service**:**

- Web Interface allowing to request the anonymization of images through the use of a Java applet
- Java applet responsible for the first anonymization
- Web Interface allowing to upload images into the neuGrid architecture through the call of a Web Service
- Web Service responsible for the second anonymization of the image and the storage of the files

The anonymization applet will allow users to select one or more local files and anonymized them. It will take a list of images as input and it will either output anonymized images into one chosen repository or send them to the neuGrid architecture. The Web Service responsible for the second anonymization of the images and their storage for further treatments will receive an array of anonymized images from authenticated and authorized users and upload them into the platform.

11.4 Technology Evaluation and Technical Choices

The technologies will be presented according to the order described into the Figure 64, ie. Through the steps that images will go during a typical image upload. The first step takes place into the hospital and is handled by the Java applet.

11.4.1 Pseudo-anonymization at the hospital

In order to leverage the work and reuse what has already been done, the applet will use some existing Java library to pseudo-anonymize the images. There are at least two libraries that are interesting for such a task: the LONI Java image I/O plugins and the dcm4che2 DICOM toolkit.

Programmers who develop neuroimaging software applications frequently encounter the need to read/write data from/to different file formats. The LONI Java Image I/O plugins [90] give Java programmers an easy access to the metadata and image data stored in many common file formats (AFNI, ANALYZE, DICOM, ECAT, GE, INTERFILE, MINC, and NIFTI) produced by scanners and neuroimaging software. Following the Java Image I/O API Specification [91], all metadata is represented as XML DOM trees and all images are returned using the same standard class. It is not exactly an anonymization tool, but it could be useful to convert

images between formats and leave only allowed fields in the metadata. dcm4che2 [92] is a high performance, open source implementation of the DICOM standard. It is developed in the Java programming language. Version 2.x of this toolkit is the next generation of the popular dcm4che-1.x [93] DICOM toolkit. The toolkit has undergone some architectural changes with improvement over the 1.x version in the areas of speed, memory usage, simplicity, and a more robust DICOM dictionary implementation. This open source toolkit has already been demonstrated in Health-e-Child project so there is already some good knowledge of its usage into the consortium. With these two libraries all the functionalities needed to process the images, either to convert them into the required format or to pseudo-anonymize them will be made available. After this first pseudonymization, the images are sent to the web service that is responsible for the second pseudonymization.

11.4.2 Pseudo anonymization in the neuGrid architecture

The Medical Data Manager (MDM) [94] is an interface between DICOM compliant storage and the gLite middleware. It aims at:

- providing access to medical data sources for computing purpose without interfering with clinical practice
- ensuring transparency so that accessing medical data does not require any specific user intervention
- ensuring a high data protection level to preserve patients privacy

This service exploits the DICOM standard for medical image transfers on the clinical side and the Storage Resource Management (SRM) on grids. It bridges these two standards by translating *on-the-fly* grid file read accesses into DICOM transactions. It benefits from the EGEE middleware capability in managing distributed files, thus enabling the federation of many DICOM servers geographically distributed and it provides a unified view of the archived data. It exploits state of the art encryption and fine grain ACL-based mechanisms to ensure both data protection and access control. MDM has been designed to work with gLite (and it uses a lot of EGEE software), and its anonymization and encryption capabilities are of real interest to neuGrid designers. MDM separates the metadata from the image, stores the metadata into AMGA and the encrypted image on the Grid. Access can be configured using fine-grained ACLs for image and metadata access. Users are authenticated using their Grid certificates. An overview of MDM is given in [94]. MDM seems to be the only DICOM anonymization and Grid aware software available and there is also existing knowledge of its usage and production configuration in the neuGrid project. Once the images have been made available to LORIS, and before being made available to the neuGrid uers, the following steps are carried out:

- ID and protocol checks
- Face scrambling
- Quality control

11.5 Conclusion

Privacy of patients is one of the key requirements that the neuGrid project must address in order to be positively accepted and widely used by medical communities which could benefit from the neuGIRD infrastructure. If users can utilise the architecture to widely share their patients' data in confidence, with care taken on the privacy of their data, this will greatly assist the adoption and the usage of such an architecture. By providing different steps of pseudonymization and face scrambling, taking place at different levels of the data acquisition process, a high level of privacy protection can be achieved.

As highlighted by the neuGrid protocol for Data Protection document (**D2.3**) all the possible ethical and legal problems related to the patient privacy have been resolved, both by following the HIPPA's recommendations during the writing of the appropriate anonymization

architecture and by requiring the informed consent of the subject. Finally, this architecture is as close as possible to the implementation of what the consortium agreed upon at the CERN meeting on December 2008 where all possible architecture models were discussed. This architecture will hopefully address the requirements of the project in the best possible manner.

11.6 References

[86] Pseudonymization http://en.wikipedia.org/wiki/Pseudonymization

[87] *Health Insurance Portability and Accountability Act (HIPAA)*, 1996 (Public Law 104-191)

[88] Health Information Privacy http://www.hhs.gov/ocr/privacy/index.html

[89] The New Draggable Applet Feature in the Java SE 6 Update 10 Plug-In http://java.sun.com/developer/technicalArticles/javase/6u10_applets/

[90] LONI Java Image I/O Plugins http://www.loni.ucla.edu/Software/IO_Plugins

[91] Java Image I/O Framework Specification http://jcp.org/en/jsr/detail?id=15

[92] dcm4che2 DICOM Toolkit http://www.dcm4che.org/confluence/display/d2/dcm4che2+DICOM+Toolkit

[93] dcm4che2 DICOM Toolkit 1.x http://www.dcm4che.org/confluence/display/d1/Home

[94] Johan Montagnat et al. A Secure Grid Medical Data Manager Interface to the gLit Middleware, *Journal of Grid Computing* (JGC), Kluwer, 2007