Computing and Informatics, Vol. 31, 2012, 3–15

EXCHANGING DATA FOR BREAST CANCER DIAGNOSIS ON HETEROGENEOUS GRID PLATFORMS

Damià Segrelles, Ignacio Blanquer, José Salavert
Vicente Hernández

Universitat Politècnica de València (I3M)
València, Spain
e-mail: {dquilis, iblanque, vhernand}@dsic.upv.es, josator@fiv.upv.es

José Miguel Franco, Guillermo Díaz, Raúl Ramos

CETA-CIEMAT, Centro Extremeño de Tecnologías Avanzadas
Trujillo, Spain
e-mail: {josemiguel.franco, guillermo.diaz, raul.ramos}@ciemat.es

Rosana Medina, Luis Martí

Universitary Hospital Doctor Peset
Valencia, Spain
e-mail: rosana.medina@hotmail.com, Luis.Marti@uv.es

Miguel Ángel Guevara, Naymi González

Faculty of Engineering, University of Porto
Porto, Portugal
e-mail: {mguevaral, nlgezposada}@inegi.up.pt

Joana Loureiro, Isabel Ramos

Faculty of Medicine, University of Porto
Porto, Portugal
e-mail: joanaploureiro@gmail.com, radiologia.hsj@mail.telepac.pt

Abstract. This article describes the process of defining and implementing new components to exchange data between two real GRID-based platforms for breast cancer diagnosis. This highly collaborative work in development phase pretends to allow communication between middleware, namely TRENCADIS and DRI, in different virtual organizations. On the one hand, TRENCADIS is a Service-Oriented Architecture in which the usage of resources is represented with Grid services based on the Open Grid Service Architecture specification (OGSA).On the other hand, DRI is a software platform aimed at reducing the cost of hosting digital repositories of arbitrary nature on Grid infrastructures. TRENCADIS has been deployed in the Dr. Peset Hospital (Valencia, Spain) and DRI has been deployed in the São João Hospital (Porto, Portugal). The final objective of this work in progress is to share medical images and its associated metadata among geographically distributed research institutions, while maintaining confidentiality and privacy of data.

Keywords: TRENCADIS, DRI, GRID, compatibility

1 INTRODUCTION

Nowadays, radiologists use computer technologies to improve the image diagnosis process. Image diagnosis involves medical images and reports with the description of the findings in the image. There are two main research lines in this field: the storage and processing of images; and the integration of diagnostic reports through semantic interoperability.

With respect to the processing and storage of digital images, approaches concentrate on the use of Peer to Peer (P2P) and Grid technologies to allow the federation of distributed data storages and computing resources.

Talking about data sources, PACS (Picture Archiving and Communication Systems) and RIS (Radiology Information Systems) organise data according to patient-centric information, as they are oriented to healthcare delivery. Also, its security systems do not support multi-domain structures.

Regarding Data Grid technologies, SRB (Storage Resource Broker [1]) and gLite [2] provide general-purpose tools for storing large amounts of data in distributed environments. On the one hand, several projects like BIRN (Biomedical Informatics Research Network) [3], a North American initiative that aims at building a virtual community of shared resources in the field of brain degenerative diseases, are based on SRB. On the other hand, the NeuroLOG project [4], which is a French middleware aimed at sharing and processing brain disease images, uses MDM [5] (Medical Data Manager) to manage information. MDM uses components from gLite (AMGA, LFC, Hydra...). Moreover, the NeuGRID project [6] is a European initiative whose goal is to develop public electronic infrastructures needed by the European neuroscience community. It is composed of several services that follow the SOA (Service Oriented Architecture) philosophy and it also uses gLite components (gLiteUI, AMGA, CE, SE, BDII, DPM and WN). Other projects have developed their own components based on Grid technology. CaBiG [7] is creating a network that will connect the entire cancer community. This project has developed a Grid Layer, namely CaGRID [8].

With respect to the second research branch (semantic interoperability for diagnostic reports integration), imaging repositories have strongly evolved, providing tools for sharing and processing studies of images and its associated metadata, and allowing data-mining techniques that improve diagnosis and therapy. In this sense, evidence-based medicine is a case-based methodology that relies on high quality and organized medical knowledge. Data annotation is often performed by centrally storing metadata (in the case of MDM on centralized AMGA databases). Some projects, like BIRN, define their own lexicon in order to manage complex data representations. There are also collaborative data collection projects that define diagnostic report templates, Health-e-Child [9] focuses on creating a common database of pediatric disorders with the support of many medical centers.

Other approaches also consider structured reports to enable context-based retrieval (such as Mammogrid [10] or NeuroBase [11]). In those approaches, metadata required for context-based searching is also located on central repositories that store the annotation of the medical images. In both research branches, developments based on the DICOM standard [12] (Digital Imaging and Communications in Medicine) have demonstrated to be a convenient and widespread solution among the medical community. In fact, the majority of hospitals and vendors choose the DICOM standard to store and exchange digital images (CT, MRI, X-Ray, . . .).

Our work aims at developing an Iberian collaborative network on breast cancer diagnosis by sharing medical images and their associated reports metadata among geographically distributed research institutions from Spain and Portugal, while maintaining the confidentiality and privacy of data.

To achieve this goal, we collaborate to allow communication between two different middlewares in different virtual organizations, namely TRENCADIS [13] and DRI [14, 15]. This results in the definition and implementation of a platform that allows medical data to be shared between geographically distant infrastructures. Both middlewares use gLite-compliant [2] storage resources as backend (such as EGI [16] or NGI [17]).

Two real deployments are presented, as well as the specification of new middleware compatibility components for DICOM data sharing. The first deployment is based on TRENCADIS and has been performed in the Dr. Peset Hospital (Valencia, Spain). The second deployment is based on DRI and has been performed in the São João Hospital (Porto, Portugal). Both deployments aim at the management of breast cancer mammography reports.

Section 2 shows the concrete objectives of this article. Section 3 summarizes the DRI technology and its deployment in the São João Hospital. Section 4 describes the TRENCADIS technology and its deployment in the Dr. Peset Hospital. Section 5 specifies the steps to follow in order to define the components needed for data exchange between the platforms, detailing the steps already completed. Finally, expected benefits and conclusions are presented.

2 OBJECTIVES

In this paper, the main objective is to set a procedure to develop data exchange components for TRENCADIS and DRI. These components must allow information exchange among both technologies, while improving the diagnosis of the breast cancer process.

For the attainment of this general objective, the following targets have been defined:

- To study two real infrastructures, TRENCADIS and DRI, in order to allow connectivity between them. These infrastructures have been already deployed in Dr. Peset Hospital and São João Hospital, respectively, leaving the door opened for the incorporation of new medical institutions.
- To identify and specify the TRENCADIS and DRI data fields needed to translate the information in a bidirectional way. This information is referred to as diagnosis of breast cancer mammography explorations.
- To design an interface in both middleware accommodated to the identified data fields. The new components will use this interface to allow compatibility.
- The remaining steps will be completed in future works.

3 DRI TECHNOLOGY

The Digital Repository Infrastructure (DRI) is a software platform developed by CETA-CIEMAT aimed at reducing the cost of hosting digital repositories of arbitrary nature on Grid infrastructures, providing both users and repository providers with a set of graphical and conceptual tools that easily define repositories and manage content.

The digital repository presented here is composed of a set of units of digitalized content annotated with metadata [18] described through an entity-relationship model. With DRI, a repository provider describes his data model in an XML file and has immediate access to a set of standard graphical user interfaces for browsing and managing repository content stored on a Grid infrastructure. On top of that, he could also develop custom tools to provide specific functionality for his repository (for content viewing, data analysis, etc.). This way, a repository of mammography studies is composed of digital content (mammograms) and metadata (patient info, diagnoses, etc.).

A deep description of the DRI architecture and services can be found in [14, 15].

3.1 Deployment of Infrastructure

The current deployment of infrastructure involves three centres: The Faculty of Medicine and the INEGI, from Porto, and the CETA-CIEMAT, located in Trujillo, Spain.

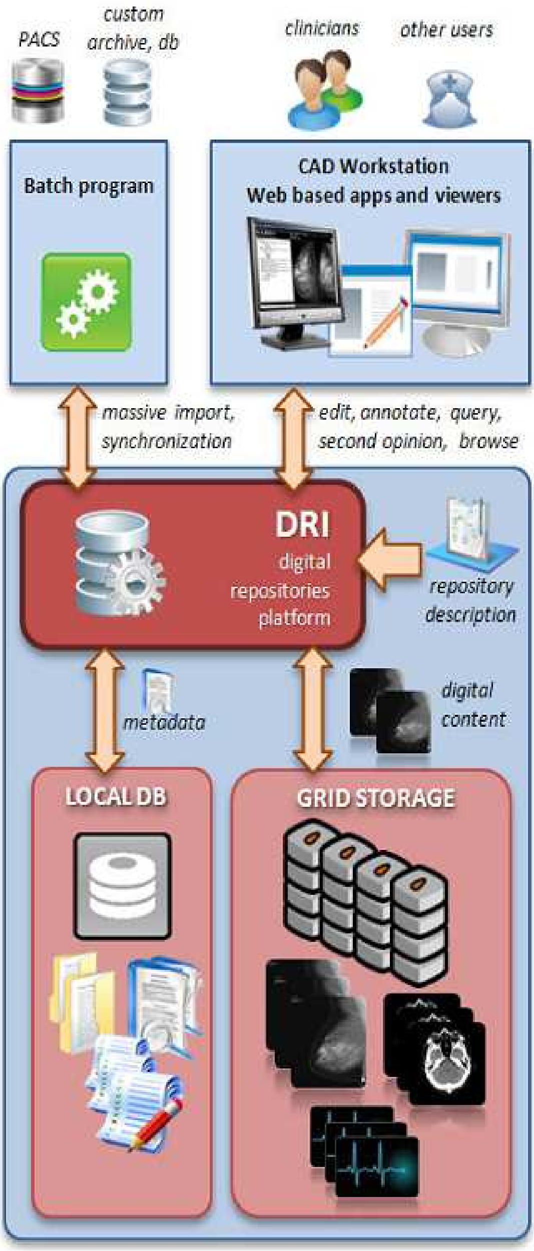

Fig. 1. General diagram of DRI architecture

There is a full deployment of the DRI platform located in the São João Hospital, in Porto, where the Faculty of Medicine is located. The platform is configured in a standalone mode, storing the repository metadata in a local MySQL database and the digital content in the local file system of the deployment machine.

The Mammography Image Workstation for Analysis and Diagnosis (MIWAD) is a rich DRI client doctors use as a frontend to interact with the repository platform. As a rich client, MIWAD implements image processing operations that are used for supervised mammography segmentation and classification.

Both INEGI and CETA-CIEMAT host other instances of the DRI platform used as replicas of the repository data. INEGI DRI configuration stores metadata in a MySQL database and digital content in a FTP server. CETA-CIEMAT DRI instance also stores metadata in a MySQL database, but digital content is stored in a gLite storage element.

3.2 Deployment Objectives

The aim of the previously mentioned deployment is validating DRI usage to support two scenarios:

1. managing large collections of federated mammograms studies and
2. building CAD systems that help doctors in mammograms analysis and diagnosis.

A sample deployment scenario and a description of the CAD system were presented in [19].

4 TRENCADIS TECHNOLOGY

TRENCADIS technology defines a horizontal architecture that organizes virtual repositories of DICOM objects. TRENCADIS is a Service-Oriented Architecture (SOA) in which the usage of resources is represented as Grid services based on the Open Grid Service Architecture specification (OGSA).

TRENCADIS is structured into several layers that provide the developers with different abstraction levels. Figure 1 shows a diagram of the TRENCADIS infrastructure. A description of the TRENCADIS architecture and its services can be found in [13].

4.1 Deployment of Infrastructure

The deployment of infrastructure involves two centres; these are the Polytechnic University of Valencia (UPV) and the Dr. Peset Universitary Hospital. The TRENCADIS Grid services have been distributed as follows:

VOMS Server: The VOMS service manages user memberships, groups and roles in the VO.

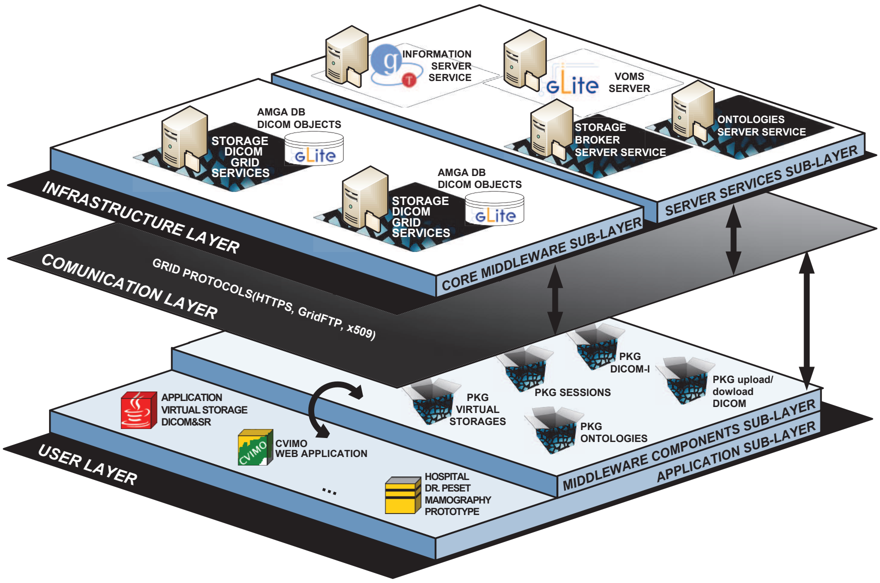
INFORMATION SERVER SERVICE
gLite
VOMS SERVER
ONTOLOGIES SERVER SERVICE
STORAGE BROKER SERVER SERVICE
SERVER SERVICES SUB-LAYER
AMGA DB DICOM OBJECTS
STORAGE DICOM GRID SERVICES
gLite
AMGA DB DICOM OBJECTS
STORAGE DICOM GRID SERVICES
gLite
CORE MIDDLEWARE SUB-LAYER
INFRASTRUCTURE LAYER
COMUNICATION LAYER
GRID PROTOCOLS(HTTPS, GridFTP, x509)
PKG DICOM-I
PKG SESSIONS
PKG VIRTUAL STORAGES
PKG upload/ dowload DICOM
PKG ONTOLOGIES
MIDDLEWARE COMPONENTS SUB-LAYER
APPLICATION SUB-LAYER
APPLICATION VIRTUAL STORAGE DICOM&SR
CVIMO
CVIMO WEB APPLICATION
...
HOSPITAL DR. PESET MAMOGRAPHY PROTOTYPE
USER LAYER

Fig. 2. General diagram of TRENCADIS architecture

Ontologies Server: This Grid service contains the federated report templates.

Storage Broker: This Grid service offers the infrastructure needed for indexing DICOM objects.

Index Information Service: This Grid service keeps information about the Grid services deployed in the infrastructure.

The services located in Core Middleware have been deployed in the Dr. Peset Universitary Hospital:

Storage DICOM: This Grid Service offers the infrastructure needed for sharing DICOM objects by using federated report templates.

AMGA Server: AMGA manages the metadata from files stored in the Grid (mainly DICOM images and DICOM-SR).

A more detailed description of all these Grid Services was presented in [13].

Also, a web application prototype has been developed and hosted in the UPV. The application is acceded from the Mammography Department of Dr. Peset Universitary Hospital and uses TRENCADIS middleware components to access the Grid services deployed in the infrastructure.

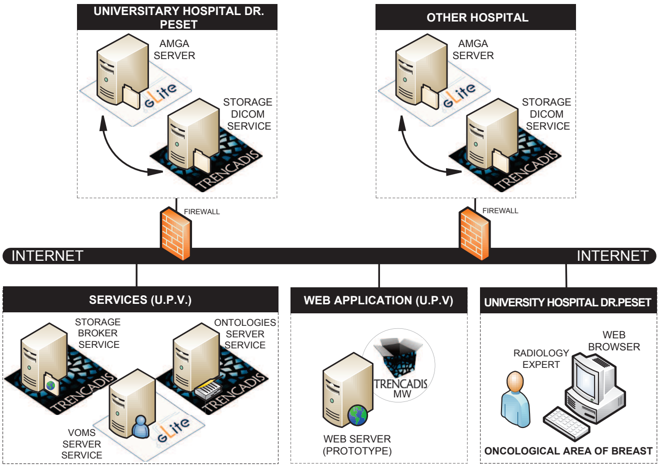

Fig. 3. Scheme of the TRENCADIS infrastructure deployment

4.2 Deployment Objectives

The two main objectives of this deployment are to federate structured report schemes among different centers and to provide a framework for sharing semantic annotations of breast cancer images. At the end, a comprehensive view of images and diagnostic reports is achieved along the whole infrastructure.

5 DATA EXCHANGE COMPONENTS FOR DRI AND TRENCADIS

As mentioned in previous sections, this article describes the current process in which involved institutions are defining components to exchange data between the deployments from Porto and Valencia, in order to improve breast cancer diagnosis. This is a highly collaborative work that involves the DRI and TRENCADIS development teams and the doctors from both hospitals.

The design of data exchange components between the deployed instances in São João Hospital and Dr. Peset Hospital must deal with two main issues:

- Both systems were not designed to interoperate from the beginning.
- Data confidentiality and privacy must be preserved between them.

In order to deal with the first issue of the data exchange between both systems because of their different architectures, we chose to define a common data exchange schema, XML based, so that the data exchange components will be developed as client modules of both infrastructures that will be able to import and export data from and to the defined data exchange schema. On the other hand, in order to preserve confidentiality and privacy of data to exchange, a set of rules was defined in collaboration with doctors from both hospitals to mark all sensitive data from patients so that data exchange components will ignore it when exporting data.

Once all data to exchange is converted to follow the data exchange schema, it will be properly encrypted and transmitted through a secure transmission channel. The connection between components will be point to point.

Therefore, several steps have been established to afford the development of the data exchange components. These steps are as follows:

1. Defining a data exchange schema. This schema should include all data fields from both systems to be exchanged, starting the mapping from similar meaning fields and including the rest of data as new fields.
2. Modifying both data models to support non-existing fields.
3. Designing and developing components to export system data to the data exchange format and to import data received in that format.
4. Determining data security and integrity between systems.
5. Defining data replication policies.

Next sections describe the progress of these steps.

5.1 Definition of the Data Exchange Schema

The first step to complete in the process of developing data exchange components between DRI and TRENCADIS deployments is the definition of a data exchange schema, as both systems were not designed to interoperate at the beginning. Both middlewares have different components and are based on different technologies. Moreover, even though gLite storage elements are used to store data, the organization of data is quite different.

The data exchange schema should allow data exchange between both systems without losing information that would be relevant to the diagnosis process. After analysing the data models of each system, it was decided to design a data exchange schema based on XML. The election of XML as data exchange format was because it is an open standard, highly extensible and widely used in web services interoperability. The flexibility offered by XML allows any future adaptation of the data schema while minimizing changes impact.

Doctors from both hospitals have played an important role in the definition of the data exchange schema, because without their expertise it would be a really hard work to define the mapping rules between the fields of each system data model and the fields of the data exchange schema. These rules will be used by the exchange components in the data importing and exporting process. Also, the exclusion rules that will be applied over sensitive data from patients during the exporting process have been defined. A diagram of the XML based data exchange schema is shown in Figure 4 (just the top elements of the schema hierarchy are shown).

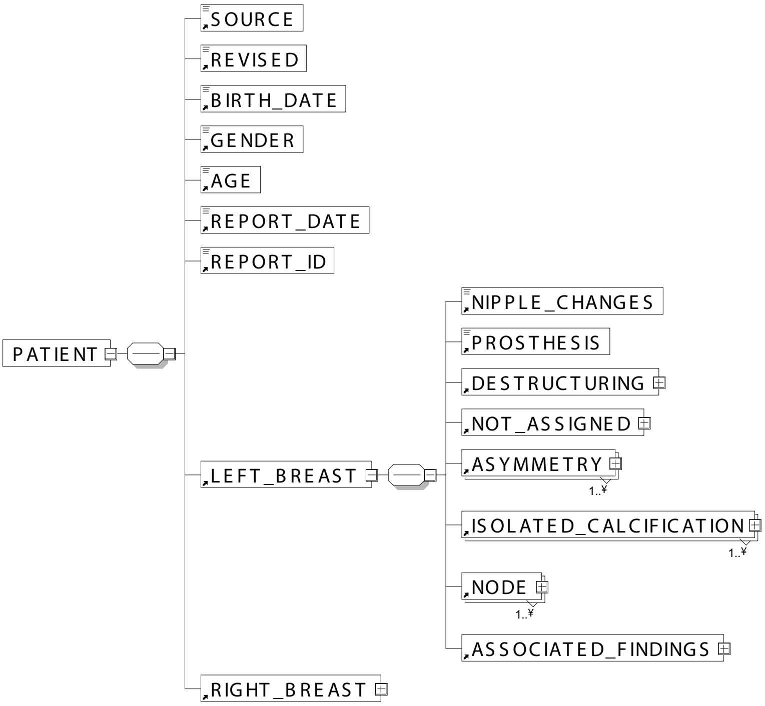

Fig. 4. Diagram of the data exchange schema

This process is being deeply influenced by the feedback of the doctors from both hospitals. Data models are continuously changing in order to fit expert needs and, therefore, the development speed is not as quick as expected.

5.2 Update of Both Systems Data Models

Data exchange between the two systems requires the modification of the data models in order to accommodate the unique fields of each system. Otherwise, there would be a loss of data relevant to the diagnosis. Currently, the changes made to data models have not modified the relations of previously existing models; they have simply added fields to existing entities, such as Study or Lesion.

At the moment, just the first two steps established to develop the data exchange components have been completed. Data models will continue changing to meet doctors needs, so future changes are expected.

5.3 Current Status of the Development of the Data Exchange Components

At this point, the development of data exchange components is in progress. Early versions of these components can import and export information in the format defined by the data exchange schema.

As mentioned before, the design determines that these components will be implemented as platform clients that interact with APIs provided by middleware, abstracting the underlying storage layer used. For example, in the DRI case, the same data component can be used to import and export data from the São João Hospital DRI deployment (local file system based), the INEGI replica (FTP based) and the CETA-CIEMAT replica (gLite storage element based).

According to data integrity and security between both systems, the use of an XML based data exchange schema allows data to be transmitted through any secure protocol such as SCP, SFTP, GSIFTP, SOAP + WS-Security, etc. Data replication policies are not defined yet.

6 CONCLUSIONS

Interoperability is a key feature in Grid middleware. However, the integration of different Grid systems with different internal representation models is not straightforward and requires an important effort. Despite of this effort, the integration of different environments is a target that presents large benefits. In this collaboration, it is envisaged that the annotated mammographic database deployed in Portugal, using the DRI technology, will be linked to the annotated mammographic database deployed at the Dr. Peset Valencian Hospital, using TRENCADIS technology.

In this sense, this work has allowed to define a data exchange schema. This schema includes all data fields from both systems (DRI and TRENCADIS) to be exchanged, modifying both data models to support non-existing fields.

This will end up with larger deployments, enriched databases and increased processing tools, without strongly compromising the autonomy of the centres in terms of structure and policies.

REFERENCES

[1] Moore, R. W.—Wan, M.—Rajasekar, A.: Storage Resource Broker; Generic Software Infrastructure for Managing Globally Distributed Data. Local to Global Data Interoperability – Challenges and Technologies, 2005. ISBN: 0-7803-9228-0.

[2] gLite. Lightweight Middleware for Grid Computing. Available on `http://glite.cern.ch/introduction`. Last visited on January 2011.

[3] Grethe, J. S.—Baru, C.—Gupta, A. et al.: Biomedical Informatics Research Network: Building a National Collaboratory to Hasten the Derivation of New Under-

standing and Treatment of Disease. Studies in Health Technology and Informatics, Vol. 112, 2005, pp. 100–109.

[4] Montagnat, J.—Gaignard, A.—Lingrand, D.—Rojas Balderrama, J.—Collet, P.—Lahire, P.: NeuroLOG: A Community-Driven Middleware Design. Studies in Health Technology and Informatics, Vol. 138, 2008, pp. 49–58.

[5] Montagnat, J.—Frohner, A.—Jouvenot, D. et al.: A Secure Grid Medical Data Manager Interfaced to the gLite Middleware. Journal of Grid Computing, Vol. 6, 2008, No. 1, pp. 45–59.

[6] Neugrid deliverable: D6.1b Distributed Medical Services Provision (Design Strategy). University of the West of England, Bristol (UK), Technical Report 2009.

[7] CaBIG – Cancer Biomedical Informatics Grid. `https://cabig.nci.nih.gov`. Last visited on January 2011.

[8] Infrastructure Overview Page. `https://cabig.nci.nih.gov/workspaces/Architecture/caGrid`. Last visited on January 2011.

[9] Freund, J.—Comaniciu, D.—Ioannidis, Y.—Liu, P.—McClatchey, R.—Morley-Fletcher, E.—Pennec, X.—Pongiglione, G.—Zhou, X.: Health-e-Child: An Integrated Biomedical Platform for Grid-Based Paediatrics. HealthGrid 2006.

[10] Amendolia, S. R.—Estrella, F.—Hassan, W.—Hauer, T.—Manset, D.—Clatchey, R.—Rogulin, D.—Solomonides, T.: MammoGrid: A Service Oriented Architecture based Medical Grid Application. Lecture Notes in Computer Science, ISBN 978-3-540-39090-9, Volume 3251, 2004, pp. 939–942.

[11] Barillot, C.—Valabregue, R.—Matsumot, J.-P.—Aubry, F.—Benali, H.—Cointepas, Y. et al.: NeuroBase: Management of Distributed and Heterogeneous Information Sources in Neuroimaging. DiDaMIC-2004, a satellite workshop of Miccai-2004.

[12] Digital Imaging and Communications in Medicine (DICOM), Part 10: Media Storage and File Format for Media Interchange. National Electrical Manufacturers Association, 1300 N. 17th Street, Rosslyn, Virginia 22209 USA.

[13] Blanquer, I.—Hernández, V.—Meseguer, J. V.—Segrelles, D.: Content-Based Organisation of Virtual Repositories of DICOM Objects. Future Generations Computer Systems, FGCS Journal, ISSN-0167-739X.DOI:10.1016/j.future.2008.12.004. 2009.

[14] Calanducci, A.—Martín, J. M.—Ramos, R.—Rubio, M.—Tcaci, D.: gLibrary/DRI: A Grid-Based Platform to Host Multiple Repositories for Digital Content. Proceedings of the Third Conference of the EELA Project, 3–5 December 2007, Catania, Italy.

[15] Barbera, R. et al.: The gLibrary/DRI Platform: New Features Added to the User Interface and the Business Layer. First EELA-2 Conference, Bogota (Colombia), Proceedings of the First EELA-2 Conference, ISBN: 978-84-7834-600-4, CIEMAT 2009.

[16] European Grid Infrastructure (EGI). Towards a Sustainable Grid Infrastructure. `http://www.egi.eu`. Last visited March 2011.

[17] National Grid Initiative home page. `www.es-ngi.eu`. Last visited March 2011.

[18] Arms, W. Y.: Key Concepts in the Architecture of the Digital Library. DLib Magazine, July 2005.

[19] Pollán, R. R. et al.: A Grid Application for Mammography CAD. BIOMED 2010, Innsbruck, Austria. (Proceedings); Biomedical Engineering, Vol. 1 and 2, ISBN: I: 978-0-88986-825-0/II: 978-0-88986-827-4.

Damià Segrelles is a Contracted Professor in the Department of Information Systems and Computation (DSIC), Universidad Politècnica de Valencia (UPVLC), and researcher in the Institute for Molecular Imaging Instrumentation (I3M). He has been involved in Grid Technologies and medical image processing since 8 years ago and participated in 7 national and European research projects.