



SHARE

TECHNOLOGY BASELINE REPORT

Document ID:	SHARE-D3.2_revised_FINAL
Date:	31/05/07
Authors:	I. Blanquer, V. Breton, V. Hernández, Y. Legré, M. Olive, T. Solomonides
Activity:	WP3: Infrastructure and security
Document status:	PUBLIC
Document link:	http://eu-share.org/delivrables.html
Confidentiality:	Public
Keywords:	Technology, grid, applications, Healthgrid

Abstract: This document presents an analysis of the status of the technologies relevant to health grids as well as a status of the deployment of biomedical applications on grid infrastructures.

Document Log

Issue	Date	Comment	Author
0	05/9/06	First version	V. Breton
0.3	01/10/06	First Review	I. Blanquer, V. Hernández
0.4	02/10/06	Small add-ons and typos	Y. Legré
1.0	09/10/06	Small reorganization – further edition	V. Breton
1.1	18/10/06	Additions and corrections	M. Olive
2.0	20/10/06	Review and additions	T. Solomonides
Final	25/10/06	Finalisation	Y. Legré
Revised_v0	18/4/07	Corrections required by reviewers	V. Breton – I. Blanquer
Revised_v0.1	30/04/07	Additions to sections 4.4 (main projects) and 2.4 (roadmaps)	M. Olive
Revised_1.1	21/05/07	Version ready for internal review	V. Breton
Revised_1.2	24/05/07	Review by WP4	C. Van Doosselaere
Revised_1.3	25/05/07	Review by the technical coordinator	N. Jacq
Revised_FINAL	31/05/07	Finalisation	Y. Legré

Document Change Record

Issue	Item	Reason for Change
Revised v0	Chapter 2	Addition of 2 sections Section 2.2: e-science environments Section 2.4: technology roadmaps
Revised_v0	Chapter 4	Reorganization of the chapter Additional projects described: Chemomentum, OpenMolGRID
Revised_v0.1	Chapter 4	Additional projects described: CLEF, eDiaMoND, GEMSS, GIMI, IBHIS, Integrative Biology, SIMDAT and WISDOM. Additions to Health-e-Child.
Revised 1.1	Chapter 1	Addition of grid definition in the terminology



TECHNOLOGY BASELINE REPORT

Doc. Identifier:
SHARE-D3.2_revised_FINAL

Date: I. Blanquer, V. Breton,
V. Hernández, Y. Legré, M.
Olive, T. Solomonides



CONTENT

1. INTRODUCTION..... 6

1.1. PURPOSE 6

1.2. APPLICATION AREA 6

1.3. REFERENCES 6

1.4. DOCUMENT EVOLUTION PROCEDURE 8

1.5. TERMINOLOGY..... 8

1.6. ACKNOWLEDGEMENTS 11

2. EXECUTIVE SUMMARY..... 12

3. STATE OF THE ART OF GRID TECHNOLOGIES..... 13

3.1. STATUS OF WEB SERVICES 13

3.1.1. WSDL 14

3.1.2. UDDI..... 15

3.1.3. Web Service Specifications..... 15

3.1.4. Web Services Resource Framework 16

3.2. E-SCIENCE ENVIRONMENTS 16

3.2.1. *myGrid*..... 17

3.2.2. VL-e..... 17

3.3. ISSUES SPECIFIC TO HEALTH GRIDS 17

3.3.1. Expression of User Requirements 17

3.3.2. Interface with Medical Informatics Existing Standards (IHE, DICOM, HL7, ENV 13734, CEN/TC251 EN13606)..... 18

3.3.3. Security..... 18

3.3.4. Medical ontologies 20

3.4. GRID TECHNOLOGY ROADMAPS..... 21

3.4.1. Reports from Next Generation Grid expert group..... 21

3.4.2. The Challengers project 21

4. STATE OF THE ART OF GRID INFRASTRUCTURES..... 23

4.1. EGEE 23

4.1.1. Overview..... 23

4.1.2. Services 23

4.1.3. Access to Resources..... 24

4.2. DEISA..... 25

4.2.1. Overview..... 25

4.2.2. Services 25

4.2.3. Access to Resources..... 26

4.3. NORDUGRID 27

4.3.1. Overview..... 27

4.3.2. Services 27

4.3.3. Access to Resources..... 27

4.4. OSG 27

4.4.1. Overview..... 27

4.4.2. Services 27

4.4.3. Access to Resources..... 27

4.5. TERAGRID 28

4.5.1. Overview..... 28

4.5.2. Services 28

4.5.3. Access to Resources..... 28

4.6. BIRN..... 28

4.6.1. Overview..... 28



4.6.2. Services	28
4.6.3. Access to resources	29
4.7. ISSUES SPECIFIC TO HEALTH GRIDS	29
4.7.1. Installation of Grid Nodes in Healthcare and Medical Research Centres.....	29
4.7.2. Security.....	29
4.7.3. Technological Requirements (Network and Data Storages)	29
5. STATE OF THE ART OF THE DEPLOYMENT OF BIOMEDICAL APPLICATIONS ON GRIDS .	31
5.1. INTRODUCTION	31
5.2. ADOPTION OF GRIDS FOR BIOMEDICAL SCIENCES	31
5.2.1. Life science	31
5.2.2. Medical research.....	31
5.2.3. Drug discovery	32
5.3. ADOPTION OF GRIDS FOR HEALTHCARE	32
5.4. MAIN PROJECTS	33
5.4.1. European projects funded within FP5.....	33
5.4.2. European projects funded within FP6: ICT for Health Unit.....	34
5.4.3. European projects funded within FP6: Research and Infrastructure Unit	36
5.4.4. European projects funded within FP6: Grid Technology Unit	37
5.4.5. European projects funded within FP6: DG-Research.....	39
5.4.6. Projects funded by National Grid Initiatives or national funding agencies	39
5.4.7. International collaborations.....	41
6. DISCUSSION	43
6.1. TECHNICAL BOTTLENECKS	43
6.1.1. Development of Grid Data Management Services	43
6.1.2. Development of Grid Nodes in Health Care Centres	43
6.1.3. Development of services compliant with medical informatics standard specifications based on the web services technology.....	43
6.1.4. Issues with Grid infrastructures under heavy use	44
6.1.5. Recording and ensuring consent	44
6.1.6. Anonymisation and pseudonymisation	44
6.2. ORGANIZATIONAL BOTTLENECKS	45
6.2.1. Organization of the healthcare and medical research community.....	45
6.2.2. Development of best practices.....	45
6.2.3. Technology Transfer between EC Projects	45
6.2.4. Worldwide Open Standards in Medical Informatics	46
6.3. COMMUNICATION BOTTLENECKS.....	46
7. CONCLUSION	47



1. INTRODUCTION

1.1. PURPOSE

The purpose of the document is to present an analysis of the status of the technologies relevant to health grids as well as a status of the deployment of biomedical applications on grid infrastructures.

This document is written in preparation for a roadmap towards the adoption of grid technology in Health and Life Sciences.

1.2. APPLICATION AREA

The document is intended for internal and external use. It will be used as a dissemination tool for the Share project.

1.3. REFERENCES

- [1] Web Services Description Language. Available at <http://www.w3.org/TR/wsdl>
- [2] Simple Object Access Protocol. Available at <http://www.w3.org/TR/wsdl>
- [3] Web Services Interoperability Basic Profile Version 1.0:
<http://www.ws-i.org/Profiles/BasicProfile-1.0-2004-04-16.html>
- [4] Document/Literal Wrapped Style:
<http://www-128.ibm.com/developerworks/webservices/library/ws-whichwsdl/>
- [5] WS-ResourceSpecification, http://docs.oasis-open.org/wsrf/wsrf-ws_resource-1.2-spec-os.pdf
- [6] WS-ResourceProperties Specification, http://docs.oasis-open.org/wsrf/wsrf-ws_resource_properties-1.2-spec-os.pdf
- [7] WS-Resource Lifetime Specification, http://docs.oasis-open.org/wsrf/wsrf-ws_resource_lifetime-1.2-spec-os.pdf
- [8] WS-ServiceGroup Specification, http://docs.oasis-open.org/wsrf/wsrf-ws_service_group-1.2-spec-os.pdf
- [9] WS-BaseFaults Specification, http://docs.oasis-open.org/wsrf/wsrf-ws_base_faults-1.2-spec-os.pdf
- [10] Evolving Web Services Standards for Managing System Resources, <http://www-128.ibm.com/developerworks/webservices/library/specification/ws-roadmap/>
- [11] WSRF Primer, Banks T, <http://docs.oasis-open.org/wsrf/wsrf-primer-1.2-primer-cd-02.pdf>
- [12] W3C Semantic Web Activity, Health Care and Life Sciences Interest Group:
<http://www.w3.org/2001/sw/hcls/>
- [13] Open Biomedical Ontologies: <http://obo.sourceforge.net/>
- [14] Swiss BioGrid, <http://www.swissbiogrid.org>
- [15] OSG: <http://www.opensciencegrid.org>
- [16] VDT: <http://vdt.cs.wisc.edu/>
- [17] SIMDAT, <http://www.scai.fraunhofer.de/simdat.html>
- [18] MyGrid, <http://www.mygrid.org.uk/>
- [19] MammoGrid, <http://mammogrid.vitamib.com/>
- [20] GEMSS, <http://www.gemss.de/>



- [21] Jean Salzemann, Nicolas Jacq, Gaël Le Mahec, Vincent Breton, “Replication and Update of Molecular Biology Databases in a Grid Environment”, Proceedings of the Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics 2006 workshop, Calgari, July 2006
- [22] GPS@, C. Blanchet et al, Proceedings of HealthGrid Conference, IOS Press, Vol 112, 2005 - <http://gpsa.ibcp.fr/>
- [23] Embrace, <http://www.embracegrid.info>
- [24] Vincent Breton, Vinod Kasam and Nicolas Jacq, “High Throughput Grid Enabled Virtual Screening: Successes and Challenges”, Proceedings of the Distributed Applications, Web Services, Tools and GRID Infrastructures for Bioinformatics 2006 Workshop, Calgari, July 2006
- [25] OGSA-DAI middleware, available at <http://www.ogsadai.org.uk/>
- [26] S. Dasmahapatra, D. Dupplaw, B. Hu, H. Lewis, P. Lewis, M. Poissonnier, N. Shadbolt
Ontology-Based Decision Support for Multidisciplinary Management of Breast Cancer. International Workshop of Digital Mammography, 18-21 June 2004, Chapel Hill, North Carolina, USA.
- [27] OpenGALEN, <http://www.opengalen.org>
- [28] NOESIS, <http://www.noesis-eu.org>
- [29] NCI Thésaurus, <http://ncimeta.nci.nih.gov/>
- [30] BIRN project : <http://www.nbirn.net/>
- [31] SRB : http://www.sdsc.edu/srb/index.php/Main_Page
- [32] Embrace: <http://www.embracegrid.info>
- [33] UMLS: http://www.nlm.nih.gov/research/umls/umlsdoc_intro.html
- [34] <http://www.clinicaltrials.gov/>
- [35] Database of requirements: <https://savannah.cern.ch/support/?group=egeeptf>
- [36] Biopattern: <http://www.biopattern.org/index.asp>
- [37] A. P. Chaudhuri, K. M. Senthil Kumar and V. Singh, “An approach to Web Services non-functional requirements using WSDL annotations”,
<http://www.theserverside.com/tt/articles/article.tss?l=WebServicesWSDL>
- [38] M. Turner, F. Zhu, I. Kotsiopoulos, M. Russell, D. Budgen et al, “Using Web Service Technologies to create an Information Broker: An Experience Report”, Proceedings of the 2004 International Conference on Software Engineering
- [39] K. H. Bennett, N. E. Gold, P. J. Layzell, F. Zhu, O. P. Brereton et al, “A Broker Architecture for Integrating Data Using a Web Services Environment”, Proceedings of the 2003 International Conference on Service Oriented Computing
- [40] M. Turner, D. Budgen and P. Brereton, “Turning Software into a Service”, IEEE Computer Volume 36 Issue 10, October 2003
- [41] Web Service Semantics - WSDL-S, W3C Member Submission, 7 November 2005,
<http://www.w3.org/Submission/WSDL-S/>
- [42] F. Jacq, F. Harris, V. Breton, J. Montagnat, R. Barbera et al, “EGEE application migration progress report (DNA4.3.2)”, 2005, <http://www.cern.ch/edms/document/641261>
- [43] D. Britton, P. Clarke, J. Coles, D. Colling, A. Doyle et al, “A Grid for Particle Physics – from testbed to production”, Proceedings of the 2004 UK e-Science All Hands Meeting
- [44] D. Russell, P. Dew and K. Djemame, “Access Control for Dynamic Virtual Organisations”, Proceedings of the 2004 UK e-Science All Hands Meeting
- [45] D. Kalra, P. Singleton, D. Ingram, J. Milan, J. MacKay et al, “Security and confidentiality approach for the Clinical E-Science Framework (CLEF)”, Proceedings of the 2005 UK e-Science All Hands Meeting
- [46] A. Simpson, D. Power, M. Slaymaker and E. Politou, “Towards fine-grained access control for health grids”, Proceedings of the 2005 Ottawa Workshop on New Challenges in Access Control
- [47] M. Brady, D. Gavaghan, S. Harris, M. Jirotko, A. Knox et al, “eDiaMoND: The Blueprint



- Document”, 2005, <http://www.ediamond.ox.ac.uk/publications/blueprint-Final.pdf>
- [48] M. Jirotko, R. Procter, M. Hartswood, R. Slack, A. Simpson et al, “Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study”, Computer Supported Cooperative Work (2005) Volume 14, Springer
- [49] Health-e-Child Proposal – Part B, March 2005
- [50] J. Fingberg et al, “GEMSS Deliverable D6.3b – Edited Final Report”, 2005, <http://www.ccril-nece.de/gemss/Deliverables/D6.3b.pdf>
- [51] A. Rector, A. Taweel, J. Rogers, D. Ingram, D. Kalra et al, “Joining up Health and BioInformatics: e-Science meets e-Health”, Proceedings of the 2004 UK e-Science All Hands Meeting
- [52] R. Taira, A. Bui and H. Kangarloo, “Identification of Patient Name References within Medical Documents Using Semantic Selectional Restrictions”, Proceedings of the American Medical Informatics Association Annual Symposium, 2002
- [53] Future for European Grids: GRIDs and Service Oriented Knowledge Utilities, third report from Next Generation Grids Expert Group, January 2006
- [54] Conclusions from Challengers Workshop on “Interaction of European and International Grid communities” in Pisa, October 2006, http://www.challengers-org.eu/images/Documents/challengers_pisa_workshop_public_report_v0_3.pdf
- [55] Stevens, R.D., et al., myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 191(1) 302-304 (2003)
- [56] Rauwerda, H., et al, The Promise of a virtual lab, *Drug Discov Today*. 11(56)22836 (2006)

1.4. DOCUMENT EVOLUTION PROCEDURE

This document will be updated incrementally via WP3 Activity as new information becomes available. Comments should be sent to the Vincent Breton (breton@clermont.in2p3.fr)

1.5. TERMINOLOGY

Glossary

WP1	SHARE project Management Activity
PO	Project Officer
PD	Project Director
MB	Management Board
PEC	Project Executive Committee
QR	Quarterly Report
EAC	External Advisory Committee
CC	Cost Claims
PR	Periodic Reports
CA	Consortium Agreement
PM	Person Month



PMx	Project Month x
-----	-----------------

Definitions

The following definitions are useful to understand the document content.

- **Component:** A component is a reusable building block, that performs some well-defined function within a process, such as a workflow. A key feature of components is that they are atomic i.e. they cannot be split into smaller units.
- **Computing Element (CE).** A computing element is a service that provides a grid interface to CPU power. Often, by language extension, the CE is also referred to as the front-end or head node to a computing cluster. The actual grid jobs are not executed on the CE itself but on Worker Nodes (see below).
- **Data:** Data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.
- **Data model:** A data model is a model that describes in an abstract way how data is represented in an information system. A data model can be a part of ontology, which is a description of how data is represented in an entire domain.
- **Grid:** A fully distributed, dynamically reconfigurable, scalable and autonomous infrastructure to provide location independent, pervasive, reliable, secure and efficient access to a coordinated set of services encapsulating and virtualizing resources
- **Grid client library:** The grid client library is the main piece of software that is used by the biologist or bio-informatician in order to interact with the grid. Therefore, it can be anything that is considered to be the grid user interface (from software components to command-line tools, APIs, etc.).
- **Grid Workload Management Service (WMS):** The WMS can be regarded as a workflow management system that defines user jobs, their workflows and their successful execution on a grid environment. In more detail, suitable Computing Elements (see above) are selected to run end user applications. Internally, the WMS itself consists of several subcomponents that keep track of the job status. Different implementations of existing WMSs are architecturally different in the sense that they either provide a “stand-alone” service on dedicated grid nodes, or they are tightly coupled with the grid client library. For our discussion the exact architecture is irrelevant since the overall goal of different implementations is very similar.
- **Information Services (IS):** This service is used to discover and locate services that are available in the grid system. This service is relevant for the operator to ensure availability and for the user through the WMS (see above) to select the most appropriate resources.
- **Local Resource Management System (LRMS):** A grid WMS can be regarded as a global scheduler that identifies CEs to execute jobs. However, the CE itself represents only the *interface* to a local scheduler or a local batch system that in turn needs to select one of the available local Worker Nodes. This local scheduler is referred to as the Local Resource Management System. Possible implementations are LSF, PBS, Sun N1 Grid Engine, Condor, Torque, etc.
- **Metadata:** Metadata may be regarded as a subset of data, and are data about data. Metadata summarize data content, context, structure, inter-relationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.
- **Ontology:** Ontology is the systematic description of a given phenomenon, which often includes a controlled vocabulary and relationships, captures nuances in meaning and enables

knowledge sharing and reuse. Typically, ontology defines data entities, data attributes, relations and possible functions and operations.

- **Replica Location Service:** This service, composed of hierarchical Replica Catalogs, is used to locate data stored in Storage Elements. The concept of replicated data is very important in grids for improving performance and availability.
- **Serialization:** This is the process of transformation of an object in memory to a flat format (file, network message etc). The name comes from the transformation of parallel data into a serial stream.
- **De-serialization:** This is the inverse process of serialization. It is the process of transforming a serial (set of sequential data) into a parallel set of data. It means transforming a message or network sequence into an object in memory.
- **SOAP:** This is a protocol for exchanging XML messages over a network. It defines a certain structure of the XML messages (the SOAP envelope), and a framework that defines how these messages should be processed by software.
- **Storage Element (SE):** A storage element is a service that provides a grid interface to storage resource whatever their type (disks, tapes ...). All data that is accessible in the grid is stored in dedicated Storage Elements that offer grid transfer and data access protocols. Once files are stored there, they are typically registered in a replica catalogue to retrieve their locations.
- **Web Service:** A web service is a software system designed to allow inter-computer interaction over a network to perform a task. It has an interface described by a computer-readable definition language called WSDL (Web Services Description Language). Other computers interact with a web service, in a manner prescribed by the interface, using messages. These messages are enclosed in a SOAP envelope and are often conveyed by HTTP. Software applications can use web services to exchange data over a network.
- **XML:** It is an annotation technology used to describe structured data within a document using mark-ups and tags, similar to HTML. The main difference between the two is that the elements in XML can be given a definition depending on their usage which may be semantic rather than presentational. XML is a text format and can be read easily either by humans or machines.
- **XML Schema:** It is a definition of the structure of an XML document. A schema contains a set of rules that dictate how an XML document must look in order to be an instance of this schema. The relationship between a schema and an XML document implementing it can be compared with a class definition and an instance in object-oriented programming
- **Worker Node (WN):** A worker node consists of one or many processors that actually execute grid applications. Often, a WN has some local storage where intermediate results are stored but cleaned up after job execution. In order to store data permanently, storage elements (see above) are used.
- **Workflow:** This is a set of components and relations between them, used to define a complex process from simple building blocks. Relations may be in the form of data links that allow the output of one component to be used as the input of another, or control links that state some conditions on the execution of a component.
- **Web Services Interoperability (WS-I):** Web services interoperability basic profile proposes a set of rules to achieve interoperability of web services between different platforms.
- **Web Service Description Language (WSDL):** It is a language for describing the interface of a web service. It describes the structure of the data sent to and received from the service, the different operations supported on the service, how to communicate with the service, and finally where the service is located.
- **Web Services Resource Framework (WSRF):** It is a set of specifications to allow access to stateful resources through web services. It means persistent data values that can also evolve



through web service interactions. The goal of WSRF is to define conventions for managing state so that applications discover, inspect and interact with stateful resources in standard and interoperable ways.

1.6. ACKNOWLEDGEMENTS

We would like to acknowledge numerous contributions to this document. The status of web technology is inspired from the Embrace deliverable D3.1 [32] edited by Jean Salzemann and Jan-Christian Byrne. Some of the bottlenecks described in chapter 5 were identified during a BELIEF workshop at CERN. We would like also to acknowledge the work provided by the Health e-Child project and the UWE, Bristol Team.



2. EXECUTIVE SUMMARY

This document presents a state of the art of grid technology for healthcare. It reviews the main grid infrastructures currently deployed around the world as well as the status of the biomedical applications deployed around the world.

A grid is a fully distributed, dynamically reconfigurable, scalable and autonomous infrastructure to provide location independent, pervasive, reliable, secure and efficient access to a coordinated set of services encapsulating and virtualizing resources (computing power, storage, instruments, data, etc.) in order to generate knowledge. A health grid is the use of a grid infrastructure and specialized services to integrate data about the patient with emerging biomedical knowledge and protocols, leading it is hoped, to a more precisely targeted and individualized kind of medicine.

The technology of web services is the foreseeable best candidate to enable the HealthGrid vision. But most of the existing grid infrastructures presently deployed in the world do not offer a web service interface to their services.

From the state of the art, a set of bottlenecks are identified:

- Secure and robust data management on the grid using standard web service technology
- Deployment of grid nodes in healthcare and medical research centres
- Development of services compliant with medical informatics standard specifications based on the web services technology
- Development of knowledge management services using ontologies

These 4 challenges are addressed to different communities:

- The first challenge is clearly in the hands of the grid technology developers
- The second challenge lies in the hands of the e-infrastructure designers
- The third challenge lies in the hands of the medical informatics community deploying applications on the grid
- The fourth challenge lies in the hands of the user communities

Besides these technical challenges, we have identified organizational bottlenecks:

- organization of the medical research and healthcare communities
- development of best practices
- technology transfer between EC projects
- worldwide open standards in medical informatics

The importance of communicating about grid technology in the world of healthcare and medical research is also highlighted.

3. STATE OF THE ART OF GRID TECHNOLOGIES

A 'grid' – not *the* grid – is now understood to mean an Internet-like infrastructure that extends the concept of the Internet in several significant ways:

- like the Internet, a grid would provide access to information services but in addition would provide pooled storage, processing power and collaboration in so-called 'virtual organizations' (VOs);
- use of a grid will be reciprocal – while a user subscribes and takes advantage of services provided by a grid, the user's resources are pooled and are available to all grid subscribers;
- the process is transparent – the grid allocates resources and provides an interface to services which give the appearance that the user is accessing just one powerful machine.

In a commercial context, grid computing has been assimilated to 'utility computing' – services provided on tap, rather like electricity supply or a telephone service. In a scientific context, grid – or 'cyber infrastructure' in the United States – has been allied to the concept of 'e-science': science done in an environment in which scientists can come together in 'lawful' but loose associations (VOs), define workflows (processing observational data, simulations, etc), derive benefit from the results and move on.

Health grids are sometimes seen as the e-health equivalent to e-science: the use of a grid infrastructure and specialized services to integrate data about the patient with emerging biomedical knowledge and protocols, leading it is hoped, to a more precisely targeted and individualized kind of medicine. It is foreseen that health grids will be deployed on a European scale in the coming years. This deployment will be progressive. As a consequence, the technology on which to build them must be carefully chosen so it fits current trends and foreseeable developments in the IT world.

Major IT companies have agreed to develop web services as the technology to enable the deployment of services on the Internet. It has been also adopted by the Open Grid Forum, which is the acknowledged body to propose and develop standards for grid technology.

Moreover, web service technology provides the bridge between the grid world and the Semantic Web, which is about common formats for interchange of data and about language for recording how the data relates to real world objects.

It must be understood however that most of the existing grid infrastructures presently deployed in the world do not offer a web service interface to their services. In this chapter, we are going to concentrate our state of the art of grid technologies on web services because it is the relevant technology for the future. In chapter 3, we will discuss the status of existing grid infrastructures where we will provide information on the technologies they use and present the services they offer.

3.1. STATUS OF WEB SERVICES

The initial idea behind web services was to enable the World Wide Web increasingly to support real applications and a means for communication among them. The web services specifications recommended by the W3C propose a set of standards and protocols allowing interaction between distant machines over a network. These interactions are made possible through the use of standardized interfaces which describe basically what are the available operations in a service, what are the messages exchanged (requests and responses), and where the service is physically located on the network and through which support. This interface, which is just a conceptual representation of an application written in a given programming language, is written in WSDL (Web Service Description Language) [1]. The typical file extension is ".wsdl".

The glue between the services, or between a server exposing a web service and a client (any piece of software that will communicate with the web services), which enable them to communicate are these request and response messages (figure 1). They can be described in a standardized way on the network and be exchanged with a standard protocol over basic http or SMTP or any common Internet protocol. All the messages and description languages are based on XML.

The main language used to make web services communicate with each other is SOAP (Simple Object Application Protocol). SOAP has the advantage of being implemented in several languages and toolkits [2].

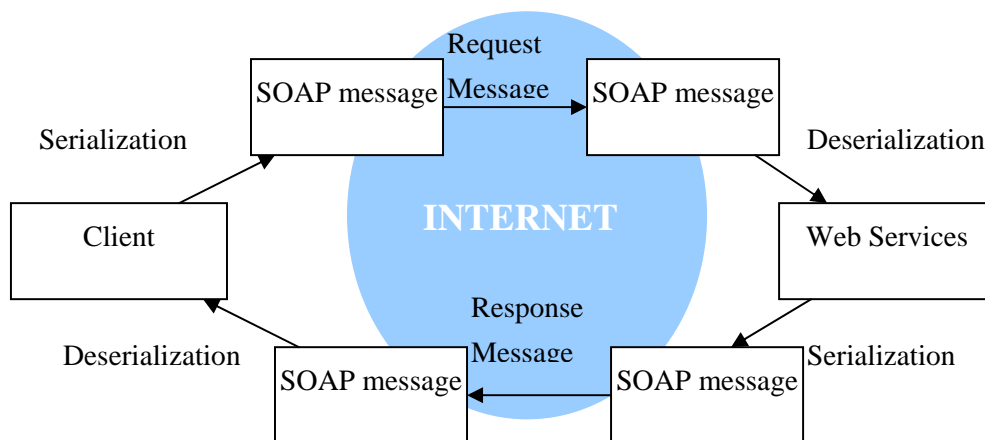


Figure 1: Messages Exchange Mechanism

3.1.1. WSDL

Web Service Description Language (WSDL) is the de facto standard used in web services to describe the service interface. It includes descriptions of the operations available in the service, the data formats used by the operations, and how and where the service can be accessed. WSDL files can be auto-generated but at present tool immaturity may lead to a need for manual editing of the WSDL file.

The data formats used by a web service are expressed using XML Schema definitions. These definitions can be combined in a single file, which can be easily imported into a WSDL file. This also simplifies the task of updating the data formats at a later stage.

WSDL is rich, resulting in many possible ways to describe interfaces and still be interoperable. The WS-Interoperability Basic Profile [3] provides a mechanism for restricting the WSDL language, and interfaces that are compliant with the profile are therefore more interoperable. A further restriction of WSDL design is the Document/Literal Wrapped style [4]. This is not a specification, but is to be considered as a best practice in WSDL design.

WSDL only describes a service in functional terms [38] – its data types, methods, parameters, message format, etc. Despite recent efforts to extend WSDL to include non-functional requirements as non-standard annotations [37], the WSDL standard still lacks flexible, semantic, non-functional descriptions required for a dynamic service-oriented environment, such as descriptions of a service's security requirements and quality of service [39]. Without semantic and non-functional descriptions, there can be confusion in the meaning of service and parameter names, and certain security

considerations – a key concern when dealing with medical data – could be neglected. Ontology-based description languages, such as DAML-S (now OWL-S) will provide a much more complete service description, but these are still in development and have limited tool and/or registry support [38]. OWL-S also assumes that ontologies are always represented using OWL, which is not always the case [41]. WSDL-S has been developed to support semantic descriptions [41], and WSDL 2.0 also promises to include non-functional requirements, but these are still in development and are not widely supported.

3.1.2. UDDI

Universal Description, Discovery and Integration (UDDI) is another de facto standard for web services, and provides a registry for service discovery [38]. Web services registries describe the available services and provide search facilities for finding suitable services. However, the current search functions in UDDI provide only limited support for automatic service selection decisions and cannot facilitate matching at the service capability level. A key limitation of UDDI is that it does not provide semantic searching; it is essentially limited to keyword and category-based searching [38, 41]. It also does not capture relationships between entities, and is not able to infer relationships using semantic information. To facilitate semantic searching, UDDI's capabilities can be extended using DAML-S/OWL-S [40], and ways to address these limitations are also being examined for upcoming versions of UDDI.

3.1.3. Web Service Specifications

The various web service specifications can be divided into first-generation and second-generation specifications. The first generation of specifications includes those mentioned above. These specifications are widely adopted and are fairly stable (WSDL will soon come in a new version).

The second generation of web service specifications are often called WS-*, because their names usually starts with WS-, like WS-Addressing and WS-Security. This set of specifications provides functionality for state, workflow composition, security, policies, attachments and more. The WS-* specifications take advantage of various “utility services” to perform the tasks they are designed for. Another feature of WS-* specifications is that they can require that a client also needs to have a web service available.

Specifications are currently becoming more standardized and stable, but in some areas there are still rapid developments. The main advantages of web services are:

- They offer great interoperability (mainly because of standardised specifications).
- They enable the communication of processes and transfers of data independently of the programming language of the underlying applications. Therefore, by extension, almost any piece of software can be exposed as a web service.
- They can be considered as firewall-friendly, because they are based on standard Internet protocols.

The main weaknesses of web services are:

- They are not adapted for transferring large volumes of data.
- Their performance can be worse with respect to other RPC based communication methods due to the overhead of sending XML messages and multiple encapsulations.
- The time taken for dynamic searching and composition – searching for, choosing, and binding services to satisfy user requirements – can be a factor [39].
- They are stateless, i.e. lack a persistent state.
- Many earlier web services stacks in use are unclear on which technologies should be used at which level, and even which technologies are compatible with each other [40].

3.1.4. Web Services Resource Framework

Web Services Resource Framework (WSRF) is a set of five specifications [5, 6, 7, 8, 9] that define conventions for modelling and accessing stateful resources using web services. The five specifications are: WS-Resource, WS-ResourceProperties, WS-ResourceLifetime, WS-ServiceGroup and WS-BaseFaults. WS-Notification is often related but not part of WSRF. WSRF 1.2 was approved as the OASIS standard in April 2006 and WSRF is also part of the future WS roadmap [10] backed up by HP, IBM, Intel and Microsoft.

Traditionally, web services have either been stateless or, if state handling, been implemented in a case-specific way. A common solution is to add state handling specific parameters for the operation calls. WSRF separates operations and state by introducing the concept of WS-Resource, which maintains the state information during several operation calls.

Simple services could be implemented as stateless services, thus avoiding the complexity of state handling altogether. However, with many services state becomes an issue that has to be dealt with. These include for example services with long-lasting operations and large input or output data. In addition to complexity, state handling can also make a web service substantially more flexible to use and monitor.

In March 2006 the major industry leaders within the field of web services agreed together with the Globus Alliance that the WSRF specifications should be merged together with the WS-Transfer set of specifications. This will be a process that is expected to be completed within 18-24 months.

The WSRF specification is already used in the grid world and has several widely used implementations. Industry partners recognise this and will work to simplify the process of merging with WS-Transfer. It should also be noted that the final WS-Transfer specification will be semantically very similar to WSRF and will operate with the same concept of resources, so the difference will mostly be in syntax.

The main advantages of WSRF are the following ones:

- Standard and interoperable way for implementing state in web services
- WSRF separates state information from the operations.

The main drawbacks of WSRF are:

- WSRF is still a fairly new specification
- Tool support for WSRF is not very good yet.
- WSRF will be merged with WS-Transfer

The WSRF Primer [11] is a good source for further information on WSRF technology. The WSRF specification version 1.2 as standardized by OASIS is compliant with WS-I Basic Profile 1.1 which proposes a set of rules to achieve interoperability of Web services between different platforms. However, the existing tools to produce WSRF services are not yet WS-I compliant.

3.2. E-SCIENCE ENVIRONMENTS

Europe has witnessed in the last few years the emergence of e-science environments designed as science portals for distributed analysis. It offers scientists the possibility to carry out their experiments in a familiar environment while using the most recent developments in Grid-computing (e.g., OGSA, database technology, data management and storage systems, workflow software and visualization

techniques). These problem-solving environments are designed in a way to easily adapt or include new application domains.

The most advanced e-science environments in Europe are ^{my}Grid [55] in UK and VL-e [56] in the Netherlands.

3.2.1. ^{my}Grid

The aim of ^{my}Grid is to provide the services for managing the complexities of experiments and analyses in the life sciences. Thus, ^{my}Grid does not perform the same tasks as, for instance, Globus or OGSA; it adds a layer of support services above such standards to meet the needs of life scientists in creating, running and managing experiments and analyses.

The ^{my}Grid middleware framework employs a service-orientated architecture. The various ^{my}Grid services can be used as a whole or in various combinations, depending on the needs of the application. This service model is uniform throughout the ^{my}Grid architecture, such that the networked biological resources act as services, as do all of the ^{my}Grid middleware components. These services need to be offered within a framework that accommodates their distribution and the variety of data formats within ^{my}Grid. The service model in ^{my}Grid is currently based on Web Services with a later migration path to the Open Grid Services Architecture (OGSA). Thus, ^{my}Grid has many of the most important attributes of the Grid, namely creating virtual organizations of distributed people, tools, data and other resources, but is currently awaiting developments in OGSA to exploit the potential of distributed computation.

3.2.2. VL-e

Integration of methodologies and infrastructure needed for e-science experimentation form the basis for the concept of a Virtual Laboratory (VL). a VL can be broadly defined as

“an electronic workspace for distance collaboration and experimentation in research or other creative activity, to generate and deliver results using distributed information and communication technologies.”

Using this terminology, the essential components of the total e-science technology chain are:

- e-science (in application) development areas
- a Virtual Laboratory development area
- a Large Scale Distributed computing development area, consisting of high performance networking and grid parts

The VL-e program in Netherlands (<http://www.vl-e.nl>) [56] aims at addressing these three layers of activity. A first program line has the objective of creating several research prototypes of advanced e-Science application-specific Problem Solving Environments (PSEs) in many areas including life and medical sciences. A second program develops the fundamental knowledge for the Virtual Laboratory while the third program focuses on the development of knowledge for large-scale distributed computing systems.

3.3. ISSUES SPECIFIC TO HEALTH GRIDS

3.3.1. Expression of User Requirements

Clinical requirements have been documented extensively by many health grid projects, supported both by the EC and national bodies, such as the e-Science core programme in the UK. These have led to the identification of many middleware service requirements to be addressed by infrastructure projects. Like in GEMSS, MammoGrid, CLEF, eDiamond, Health-e-Child, ... a large set of requirements relevant to bio- and medical informatics applications of grids have been collected within the framework of DataGrid and EGEE projects. These requirements are available for public use [35].

3.3.2. Interface with Medical Informatics Existing Standards (IHE, DICOM, HL7, ENV 13734, CEN/TC251 EN13606)

Medical Informatics has evolved significantly in the last years. The use of digital systems for clinical practice has led to the definition of standards in many areas that are being adopted at different speeds. Medical Imaging for example, is an exemplary case, in which the adoption of DICOM for the acquisition, connection and storage of medical images has been accepted worldwide. Other fields, such as medical records, are progressively considering HL7 for the exchange of data. Competence of standards, such as CEN/TC251 EN13606, more focused on the storage and structuring of clinical records prevents a wider uptake of the technology. Other fields suffer from similar problems, such as vital signs, in which, although there is a European Norm (ENV 13734) proposed to be an ISO norm, the uptake of the norm has been reduced to a few SMEs and modest attempts by large companies.

Moreover, initiatives such as IHE (Integrating the Healthcare Enterprise), promote the coordinated use of DICOM and HL7 by publishing best practices and guidelines.

3.3.3. Security

The field of (Grid) security is a rather broad one and can be divided into the following main domains:

- *Authentication*: In order to allow only registered users to use a certain service, authentication is the mechanism to detect who is the actual user, i.e. the main question of user identification is tackled. In the grid community, grid certificates based on the X.509 standards are used for both users and services (i.e. hosts) due to the mutual authentication process imposed by the Grid Security Infrastructure (GSI).
- *Authorisation*: Once a user is authenticated, i.e. once it is known who the user is, the next step is to determine what access rights this user has, whether it is a person, a machine or a service. Therefore, this answers the question, “What can the user do?” Authorisation is normally controlled at the level of the VO, although resource access is ultimately always a resource provider decision.
- *Confidentiality and Data Encryption*: Authentication is typically done only once per connection between a client and a server. In contrast, authorisation is often done for each request that requires access to resources. Once this is done, data can be exchanged between the client and the server. This exchange can either be done in clear text or using a data encryption standard. By default, grid file transfer protocols such as GridFTP transfer data in clear text.
- *Integrity and Non-Repudiation*: Medical data is also critical in questions of liability. Medical records are kept for several years after the death of a patient and the authenticity of the records is very important. Security mechanisms must ensure that the information is not changed or amended, or if it has been, to maintain an audit trail of the change, its author and context.

There are many mechanisms in use to address these security issues. The main technologies used in grid environments are:

- *Asymmetric Encryption*. Usage of different keys for encryption and decryption. Used in many secure Internet protocols, such as SSL or Https.
- *Digital X509 Certificates and Certificate Authorities*. Duly digitally signed credentials that identify user name and organisation in a chain of trust.
- *Public Key Infrastructure (PKI)*. Security infrastructure using Asymmetric Encryption and Digital Certificates to hold a private and a public key for encryption and decryption of data.
- *Globus Security Infrastructure (GSI)*. Security infrastructure that uses PKI and local authorisation methods to implement the authentication and authorisation system.
- *Authorisation Mechanisms (VOMS, PERMIS, ...)*. The balance between local resource

autonomy and global management of the authorisation is complex and has been addressed through many different approaches. From totally distributed to centralised approaches, authorisation systems aim at providing additional information with the certificates to let resource providers organise better the access rights of different users.

In earlier grid projects, it was common for grid users to be mapped to local security policies, for example the UNIX file system permissions associated with the local user that the grid user was mapped to. However, the problem with this approach was that it did not provide sufficiently fine-grained access control, and there was no way to support users with multiple authorisation profiles [43].

This issue of ‘identity management’ is a recurring issue with grid computing as a whole, and has been addressed in various ways [44]:

- Globus’ original *GSI* mapped user identities onto local users, and therefore suffers from the above problem of lack of fine-grained access control.
- Globus therefore produced a *Community Authorisation Service (CAS)* which provides highly detailed user permissions for grid resources via X.509 certificate extensions. However, these are so detailed that micro-managing such permissions across a large grid and certainly across multiple organisations could prove exceptionally difficult.
- The *Virtual Organisation Management System (VOMS)* is similar to CAS but extends the certificate with role and group attributes. The resource authenticating this certificate then needs to know the access policies for the roles and groups in order to make authorisation decisions.
- *Shibboleth* is an access control system predating grid research. Shibboleth federates access permissions across multiple organisations. Authorisation is exchanged via services rather than certificates, which means that requests can be trusted without having to verify the identity of the person who made it, enabling ‘pseudonymous access’ – the permissions of a user can be determined by a third party, but not the user’s identity. However, in the healthcare domain logging and accountability are strict requirements, so this mechanism would not be appropriate.
- *Akenti* is a ‘policy engine’, providing an authorisation decision when it receives a user request. Resources are protected by a Policy Enforcement Point (PEP), which connects to a Policy Decision Point (PDP) that uses the access policy certificate for the particular resource. The PDP locates and verifies all relevant certificates, then evaluates them and returns the access decision. Akenti uses LDAP (Lightweight Directory Access Protocol) for policy certificate management.
- The *PERMIS* system provides an application gateway that holds access control policies according to defined roles. Service requests pass through the gateway, which verifies authorisation for the target service before passing on the request. The gateway can contact multiple LDAP directories to include various policies from multiple organisations when determining access rights.

Akenti and PERMIS both use LDAP, which has been criticised for difficult deployment across organisations and managing changes. SAML (Security Assertion Markup Language) and XACML (eXtensible Access Control Markup Language) should both be better alternatives in this respect, and Akenti and PERMIS were due to replace LDAP with a subset of these standards.

All of the above are designed to support large, static communities of users, accessing static resources. There could therefore be a concern about scalability in situations with large numbers of groups continually forming and disbanding. It is also unclear how well the above systems can support dynamic roles and policies.

- Argue the current storage systems. Problem of privacy.
- Other points: provenance, metadata, integrity and long term storage.

3.3.4. Medical ontologies

Standardization of interfaces with web services can drastically increase the interoperability between biomedical resources. However, interface standardization is only half the effort in making truly interoperable resources. By operating on standardized *data formats*, the bioinformatics resources can more easily be integrated into complete bioinformatics experiments by eliminating the restructuring of data between each service.

The construction of standardized data formats can be improved by defining a domain ontology that covers the concepts used within a given domain. An ontology is the systematic description of a given phenomenon: it often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse.

In this context the domain to be covered by an ontology is the set of services available on a health grid. From an agreed ontology it is possible to define a common data model that describes the format of the data used by all the services.

Another benefit of ontology is that it can also help to provide useful high level functionalities based on machine reasoning. The field of machine reasoning would be further enhanced if the functionalities of the services themselves were described in an ontology, and not only the data they operate on.

Examples of such functionalities are automatic service discovery, invocation and composition. The Semantic Web Activity at the World Wide Web Consortium is dedicated to these topics. In particular the Health Care and Life Sciences Interest Group (HCLSIG) [12] is relevant for health grids.

The value of an ontology is determined by the general support it has, because the interoperability between communication partners increases when there is a common understanding of the terms used for communication. Hence, the more support an ontology receives, the broader the possibilities for a user of a resource that also supports that ontology. Within life science and healthcare sector, the highest degree of general support is arguably the Open Biomedical Ontologies (OBO) [13].

Semantic Web technologies are just emerging in the field of medical research and healthcare. Open issues include how to integrate biomedical data using ontologies, how to combine different initiatives and how to employ advanced, semantic reasoning techniques for analysing medical data.

The majority of the biomedical applications currently using ontologies mostly deal with decision support, namely assisting health professionals in disease diagnosis, staging or therapy planning via preliminary detection services. Breast cancer diagnosis and treatment is one of the most advanced domains with the development of a Breast Cancer Imaging Ontology [26]. Anatomy is another field where ontological approaches have been explored. The interest is focussed on the representation of anatomical terminology and classification of surgical procedures, extraction of heart anatomical features, etc. The GALEN model [27] aims at developing advanced terminology systems for clinical information systems. The ontology for the GALEN model is designed to be re-usable and application independent. It is intended to serve not only for the classification of surgical procedures but also for a wide variety of other applications - electronic healthcare records (EHCRs), clinical user interfaces, decision support systems, knowledge access systems, and natural language processing.

Ontology approaches are also under development in the cardiological domain. For instance, NOESIS [28] aims at developing a platform for wide scale integration and visual representation of medical intelligence for research and cure of cardiac and cardiovascular diseases.

In most cases, biomedical ontologies function as terminology vocabularies, containing the domain knowledge required to build the classes, rules and relationships according to which the several concepts interact with each other. The Unified Medical Language System (UMLS) [33] facilitates the development of computer systems that behave as if they "understand" the language of biomedicine and

health. Developers use UMLS to build or enhance systems that create, process, retrieve, and integrate biomedical and health data and information.

A very good example is the NCI Thesaurus which is a public domain description logic-based terminology produced by the American National Cancer Institute. This thesaurus implements rich semantic interrelationships between the nodes of its taxonomies. The semantic relationships in the thesaurus are intended to facilitate translational research and to support NCI bioinformatics infrastructure [29].

3.4. GRID TECHNOLOGY ROADMAPS

In this last section, we are going to review a number of initiatives that produced or are in the process of producing roadmaps for the evolution of grid technology in the coming years.

3.4.1. Reports from Next Generation Grid expert group

According to these reports [53], the perspectives in the new generation of grids should focus on the development of utilities, as a directly and immediately useable service with a defined quality of service, by the convergence of Service Oriented Grid Architecture and Semantic Interoperability models, in what is called a Service-Oriented Knowledge Utility (SOKU). These utilities will be built on existing industry practices and emerging technologies to ensure their adoption and exploitation. The concept of knowledge rises from the intelligence that the components have to re-arrange themselves to achieve the objectives.

In this view, computer-based services are described semantically to let autonomous agents to self-monitor the system and take decisions to improve the performance. This ends up with a multi-dimensional mesh of components that breaks the strict hierarchy of layered traditional service and grid architectures. This vision can be implemented with standard technologies, as the ones previously commented such as OWL, RDF, WSRF, XML, etc. but requires new blocks to be developed.

This new vision is not yet implemented in the infrastructures, and even it requires an important development in the research of concepts such as the lifecycle, the trust and security in VOs, the adaptability and scalability, the pervasiveness or the semantic technologies. However, several of these research topics are aligned with the needs of Health grids, and it is therefore relevant to address them.

3.4.2. The Challengers project

Challengers is a Specific Support Action with the following objectives:

- Investigate enabling technologies and market trends against the mainstream focusing also on holistic view and complementarity or converging disciplines;
- Consolidate and describe the vision of the research community for the Grids technology for the coming decade;
- Increase awareness of the next decade vision for Grids among researchers of different but complementary or converging disciplines;
- Introduce and recommend a Research Agenda and a roadmap of key technology challenges, with prioritized topics, which will be considered as the advisory tools for paving the way towards the realization of the next decade Grid vision;
- Assess the business, economic and societal impact contributed by tomorrow's Grid technology, in conjunction and convergence with other key ICT;
- Address the needs of critical infrastructures public safety and security applications and life improvement.



TECHNOLOGY BASELINE REPORT

Doc. Identifier:
SHARE-D3.2_revised_FINAL

Date: I. Blanquer, V. Breton,
V. Hernández, Y. Legré, M.
Olive, T. Solomonides

The Challengers consortium has started its consultation work with a group of experts from the research and business community with well established experience and deep knowledge in the area of Grid technologies. A first report discussing the vision for a global grid for business in 2020 has been recently published [54]. The document identifies barriers and proposes a number of solutions to overcome them.

The project has also solicited the opinions of a significant number of academics and technologists, and made their ‘position papers’ publicly available on the Challengers web site. These position papers include key requirements and barriers to the successful commercial deployment of grid technologies as perceived by their authors, and some also propose solutions and possible timescales. The (lack of) user friendliness of grid user interfaces, the complexity of administering a grid infrastructure, and the need for further promotion and clarification of grid technologies are all recurring issues in these papers. Additionally, legal issues such as IPR and security concerns are also mentioned. The main concerns however seem to be a lack of a clear roadmap for the development of a production quality grid infrastructure (using appropriate standards), and the lack of a persuasive business model.

4. STATE OF THE ART OF GRID INFRASTRUCTURES

The goal of this chapter is to review some of the most relevant infrastructures from the perspective of health grids. Health grid applications can take advantage of the existing infrastructures provided interfaces can be developed between their services and the grid services made available on these infrastructures. We briefly review in this chapter the main grid infrastructures in Europe (partially USA) in terms of overall architecture, technology used, services offered and means to access resources.

4.1. EGEE

4.1.1. Overview

EGEE (Enabling Grids for E-scienceE) is a production grid project, funded by the European Commission that aims to build a grid infrastructure for e-Science. The project is a follow-up of EU DataGrid project (<http://www.edg.org>). The project also developed its own middleware, **gLite**, which offers services to build a grid. This means that EGEE is a heavy grid infrastructure built up from dedicated resources around the world in institutes, computing centres, laboratories etc. The resources range from simple desktop computers to clusters so that EGEE is now the biggest grid infrastructure in the world with more than 30000 CPUs and more than 10 Petabytes of storage.

The grid is organised hierarchically, with resource centres that are under the responsibility of Regional Operation Centres (one per federation) which themselves are coordinated by the Operations Management Centre (OMC). The goal of this hierarchy is to offer an efficient, responsive and scalable grid service to the users.

4.1.2. Services

The middleware is based on several services and protocols. Most of these protocols and services have been developed especially for gLite. The security mechanisms used by gLite are based on GSI. All users are identified by certificates, and grid map files, created in conjunction with the information registered on the VOMS servers (VO Membership Service), provide the authorization part. VOMS define the virtual organisation structure; it is quite similar to POSIX groups and users. There can be subgroups in the VO and users in these subgroups. Fine-grained access to lists is available on the resources.

gLite is a middleware specifically developed for (Scientific) Linux. Thus it is currently rather time consuming to deploy it under a different operating system but porting efforts to other operating systems are under way. The grid system interoperates with the underlying batch queues. We present a list of the main services and elements of the EGEE middleware:

- **Resource Broker:** The resource broker is the main service for research selection and job submission. It handles jobs from their submissions by users to the retrieval of the results. It is also a scheduler, doing match making to send the job to the right resources and monitoring the job as it executes on the worker node. It allows for running applications on remote computing resources, which typically deploy a batch system such as LSF, PBS, Condor, etc. End users only need to know the Job Description Language (JDL) and are not concerned with details of the underlying batch systems used in the grid.



- **Computing Element:** These are the services that handle job execution and provide information on job characteristics and status. They actually host the batch server, all the worker nodes behind being the batch clients.
- **Storage Element:** Data management is a very broad field and gLite only focuses on file based data management, i.e. relational databases are not supported by gLite. The SRM protocol is supported for (permanently) storing large amounts of data in grid storage elements. It is a service that allows virtualizing many types of storage; single or array disks, tapes servers, etc. Secure and reliable file transfers are mainly performed with GridFTP.
- **Single Catalogue or LFC (LCG File Catalogue):** This is the central service that registers files, replicas and logical names. It has a unique catalogue to manage replicas and files. In the grid each file is identified by a GUID, which is a unique identifier, that can be linked to a unique logical file name. One GUID can have several replicas, stored all over the grid. Each replica is identified by a unique physical name. The catalogue supports VOMS and GSI for authentication and authorization.
- **Secured Key Store (Hydra):** This is a special catalogue used to store encryption keys (for example, when a file is stored encrypted on the grid, the keys can be stored in the Hydra catalogue). The keys are split into at least three key stores, and one key can be reconstructed from the n-1 others: if one server is compromised, the files cannot be decrypted anyway.
- **Information System:** The information system keeps track of both user related as well as grid application specific metadata to discover resources and to get more detailed information about the resources.
- **Grid Security and Accounting:** Authentication, authorization and auditing (AAA) is supported by gLite using the Grid Security Infrastructure (GSI).
- **User Interface:** This is the entry point of the grid for the users. It is a set of command-line tools, GUIs, APIs to allow access to the main services of the Grid: Workload Management System and Data Management. The operations that can be performed from the user interface are the following: get information on the available resources, submit a job, cancel a job, get the status of the submitted jobs, retrieve job results, get information about the job, its submission and its execution, store files on the storage elements and replicate them, copy or delete files from the grid.

The gLite middleware uses essentially custom communication protocols, but some web services interfaces are available for the main services, that are to be contacted by users: Resource Broker, File Catalogue and Grid Information System. For interoperability issues, these interfaces are compliant with the WS-* specifications but not with WSRF.

4.1.3. Access to Resources

The grid users are organised in virtual organizations. When a user is member of a virtual organization, typically they can access all of their VO resources. To be able to join a virtual organization, and access the grid, users must own a personal certificate, issued by a recognized certification authority. There is a consortium, EUGridPMA, which gathers all the recognized authorities. Users should ask their certification authority for a certificate, then register into a virtual organization.

Once the user is registered and owns a certificate, they can then access a set of commands and APIs to interact with the grid through a user interface.

4.2. DEISA

4.2.1. Overview

DEISA (Distributed European Infrastructure for Supercomputing Applications) is a consortium of leading national supercomputing centres in Europe that are coordinating their actions in order to build and operate jointly a distributed Terascale supercomputing facility.

Scientists across Europe can use the bundled supercomputing power and the related global data management infrastructure in a coherent and comfortable way. A special focus is set on grand challenge applications from scientific key areas like material sciences, climate research, astrophysics, life sciences, fusion oriented energy research.

The integration of national research resources in the DEISA supercomputing grid operates at two levels:

- An inner level, dealing with the deep integration and strongly coupled operation of similar, homogeneous platforms, as well as global data management;
- An outer level, dealing with a looser federation of heterogeneous supercomputing resources.

4.2.2. Services

Each of the national supercomputing centres belonging to the DEISA consortium is primarily focused on scientific research, even if some of them have industrial users too. All partners offer both basic and extended services to their users (help desk, local documentation, training classes, technical and scientific workshops, and support from computer specialists to port, parallelize and optimize codes, particularly on machines with a large number of processors).

DEISA is based on the UNICORE middleware. UNICORE has a vertically integrated three-tier architecture. It provides client and server components. The server consists of the Gateway, Network Job Supervisor (NJS) including an Incarnation Database (IDB), a UNICORE User Database (UUDB), and the Target System Interface (TSI). All components (except the TSI) are written in Java which makes possible to install UNICORE on almost every operating system with a Java Virtual Machine implementation available.

- **User Interface:** UNICORE provides a graphical client to allow job submission, job monitoring, setting up of the security environment etc. Before submission, the jobs are prepared: jobs requirements are defined and a match making is performed to find the suitable targets to run the job. It provides the user with the extensible application support, resource management of the target system and integrated security mechanism. With every submitted request is signed with the personal X.509 v3 user's certificate, the other server components can guarantee authentication and authorization in the UNICORE grid relaying on the PKI functionalities. The client is also able to perform data management and transfer through an intuitive GUI. This is the entry point of the grid.
- **Gateway:** This is the main contact point for all the UNICORE connections: it receives data from the clients and passes them to the UNICORE servers on the sites through an authentication and authorization procedure using certificates. It accepts SSL connections from clients and one or more NJSs but only if the incoming certificate is signed by a trusted Certification Authority (CA). Moreover when clients send abstract job objects (AJOs), it verifies they have been signed with trusted and valid certificates. If the verification is successful the AJO is redirected to the corresponding NJS, otherwise it is rejected.
- **Network Job Supervisor (NJS):** The Network Job Supervisor (NJS) deals as a UNICORE scheduler and is responsible for the virtualization of the underlying resources. It

receives/sends AJOs from/to the Gateway and sends a concrete batch job translation to the target system component called Target System Interface (TSI). This component dispatches jobs to a dedicated target machine or cluster (Virtual site, Vsite), and handles dependencies and data transfers for complex workflows. It transfers the results of executed jobs from the target machine and forwards them via the gateway to the UNICORE client. The abstract definition of the Job is translated to a concrete job in the NJS with the help of the Incarnation Database (IDB). The IDB contains all target system specific information regarding computing resources typology and availability of applications. Therefore each NJS has its own IDB that describes its specific target system. Finally the NJS implements the UNICORE security model for user authorization. All public user certificates are stored in the UNICORE User Database (UUDB) and they are mapped with an account existing on the target system. Every time the NJS receives an AJO, it checks if the signing certificate is present in the UUDB and it is forwarded to the target system as the associated account. **Target System Interface (TSI):** It accepts job components from the NJS and passes them to the local resource management system for execution. It is the interface between the NJS on one hand, and the batch system of the site on the other hand. It is composed by a set of Perl libraries that implements the specific target system commands for job submission, status query, file handling etc. There are already available several TSI implementations for different systems, e.g. LoadLeveler, LSF, PBS-Pro, Linux, and CCS.

4.2.3. Access to Resources

At present there are two ways to access the DEISA resources: the UNICORE middleware and the batch queue managers:

- UNICORE (Uniform Interface to Computing Resources) allows a user to define system independent jobs and to submit them to any DEISA resource. UNICORE has been developed by a European grid project and aims to provide a computational grid for science and engineering by combining large computational resources and making them available through the Internet.
- The batch queue manager allows the users to submit batch jobs using, for example, the LoadLeveler *lsubmit* command on AIX systems. To do this, users must first log in interactively to their DEISA site.

In the future, other access ways to DEISA will be provided by the project itself. In addition, several projects, mainly at European level, are working on this topic:

- UNIGRIDS (EC) has developed Unicore/GS. It can provide a WS-RF (Web Service Resource Framework) interface to UNICORE and can interoperate with other WS-RF compliant software components. The testing phase for the deployment of this middleware to DEISA will begin in the next year.
- LIBi (Bioinformatic Italian Laboratory) aims to provide, as an interoperability portal, a single point of access to the two major European Grid Infrastructures DEISA and EGEE. It plans to use the EnginFrame technology that, with the Genius plug-in can interact with EGEE and with the LL plug-in can interact with the LoadLeveler scheduler on DEISA.
- A-WARE (EC) will develop a new technology (stable, supported, commercially exploitable, high quality) able to give easy access to all the grid resources currently available.



4.3. NORDUGRID

4.3.1. Overview

NorduGrid is a grid research and development collaboration aiming at development, maintenance and support of the free grid middleware, known as the Advance Resource Connector (ARC). The 'NorduGrid grid' or the ARC-grid is formed by the individual grid projects that use ARC as their middleware. However, these individual grid projects may have very little to do with each other. Examples of grid projects using ARC include SweGrid, M-grid and NDGF.

4.3.2. Services

NorduGrid ARC provides all the common grid services including computing and storage elements and information services. Security is based on GSI and jobs are described using xRSL.

NorduGrid ARC does not offer web service interfaces. However, there are plans to implement a WSRF based job submission interface.

4.3.3. Access to Resources

Before using resources on a NorduGrid ARC based grid, the user needs to acquire a certificate and join a Virtual Organisation that is allowed by one of the grid projects. For example Finnish scientists can join the M-grid VO and then use the resources of the M-grid.

The SwissBioGrid project [14] uses ARC in its production grid infrastructure. Some of the tools developed within the EMBRACE grid project on profile-Hidden Markov Models have also been ported to ARC.

4.4. OSG

4.4.1. Overview

The Open Science Grid (OSG) [15] can be considered an American (USA) sister project to EGEE. OSG provides a production infrastructure to several scientific communities such as High Energy Physics, Earth Sciences, Life Sciences etc. The software infrastructure is mainly based on the Virtual Data Toolkit (VDT) [16], which also includes packages such as GT4 etc. The services offered by OSG are rather similar to the ones offered by EGEE (partly overlapping, partly complementary) and cover computing and storage services.

There is no specific support nor tool for the bioinformatics community as there are no bioinformatics groups involved in the project.

4.4.2. Services

Many of OSG's services provide web services (and/or WSRF) interfaces for several programming languages.

4.4.3. Access to Resources

This issue is not relevant to this report as the OSG grid is only deployed in the USA. However, resources can be partly accessed via EGEE.

4.5. TERAGRID

4.5.1. Overview

TeraGrid (www.teragrid.org), the US supercomputing ‘cyberinfrastructure’, is a collaborative infrastructure consisting of diverse set of resource providers; DEISA can be considered its European sister project. The TeraGrid system is an integrated and coordinated set of scientific resource that provide advanced capabilities to the end user that are driven by scientific requirements and delivered through a variety of software, middleware, policy and support functions. (‘Cyber infrastructure’ is increasingly used in the USA to mean an e-Science grid.)

4.5.2. Services

Several TeraGrid services could be accessed by WSRF interfaces using the Globus Toolkit 4 components deployed onto the grid resources.

4.5.3. Access to Resources

This issue is not relevant to this report as the TeraGrid grid is only deployed in the USA.

However, TeraGrid and DEISA have been linked by a common, scalable, wide-area global file system spanning two continents. The bridging of the USA and European communities was showcased during the Supercomputing Conference SC05 at Seattle.

4.6. BIRN

4.6.1. Overview

Launched in 2001 with the support of the [National Institutes of Health's](#) National Center for Research, the Biomedical Informatics Research Network (BIRN) [30] is prototyping a collaborative environment for biomedical research and clinical information management. The growing BIRN consortium currently involves 30 research sites from 21 universities and hospitals that participate in one or more of three test bed projects: Morphometry BIRN, Function BIRN, and Mouse BIRN. These projects are centred around structural and/or functional brain imaging of human neurological disorders and associated animal models of disorders including Alzheimer's disease, depression, schizophrenia, multiple sclerosis, attention deficit disorder, brain cancer, and Parkinson's disease.

BIRN is an end-user driven project based on a robust middleware and it addresses all dimensions from capacity building to service development. It is important to have projects on the model of BIRN where user communities can build grid infrastructures.

4.6.2. Services

BIRN infrastructure is based on the Storage Resource Broker (SRB), a client-server middleware developed at San Diego Supercomputing Centre that was designed for managing file collections in a heterogeneous, distributed environment [31]. SRB presents the user with a single file hierarchy for data distributed across multiple storage systems. It has features to support the management, collaboration, controlled sharing, publication, replication, transfer, and preservation of distributed data. The SRB system is middleware in the sense that it is built on top of other major software packages (file systems, archives, real-time data sources, relational database management systems, etc). The SRB has callable library functions that can be utilised by higher level software. However, it is



more complete than many middleware software systems as it implements a comprehensive distributed data management environment, including end-user client applications ranging from Web browsers to Java class libraries to Perl and Python load libraries.

4.6.3. Access to resources

BIRN is an open infrastructure. Already several research centres in Spain and UK have joined the infrastructure. Access to the data stored on the different BIRN sites is restricted to the BIRN participating institutes.

4.7. ISSUES SPECIFIC TO HEALTH GRIDS

4.7.1. Installation of Grid Nodes in Healthcare and Medical Research Centres

The installation of grid nodes in healthcare and medical research centres is required for the use of grids by the medical research community. However, there is no equivalent in Europe to the BIRN project [30], which federates a large number of American hospitals. The main obstacles to the deployment of grid nodes are the present complexity of the procedures and the fact that most of the healthcare centres don't have the resources in house to deploy the existing technologies.

4.7.2. Security

The security offered by the existing infrastructures does not yet allow the manipulation of medical data. Important progress is being made in terms of fine-grained Access Control and data encryption. Some prototype services are under development but they are not yet fully deployed.

A specific security feature implemented by hospitals is the restrictions in the access to Internet. Installation of grid elements behind the hospital firewall is incompatible with the present security model where outbound connection is not allowed through this firewall.

Ideally, the grid node would be located outside the firewall with only anonymised or pseudonymised data being stored on the grid. However, given the legal and ethical implications of storing any personal data on the grid, even if it is fully anonymised, further investigation will be required to determine if this is possible with current national and European policies and legislation. For example, even with the most stringent pseudonymisation and de-identification techniques there is still some risk of unauthorised re-identification by a person with sufficient knowledge from other sources [45]. WP4 deliverables D4.1 and D4.2 also address the issues of anonymisation and pseudonymisation, and in particular note the fact that pseudonymisation does not 'exist' legally.

Revocation of credentials and how to provide temporary access to data is still an open issue, and an important one for health grids [46,47]. There are a number of situations where users would temporarily require access to data that they would not normally have access to, such as a visiting expert to a breast cancer unit being shown an unusual case [48]. Certificate authorization servers have been developed in both 'Pull' mode (VOLDAP, GridSite LDAP, and VOMS-httpd), in which sites periodically pull a list of valid members from a central service, and 'Push' mode (VOMS attribute certificates), in which users obtain a short-lived attribute certificate that they present to sites to prove their membership [43]. However, both of these would leave a window where revoked or expired credentials could be used to gain unauthorised access. Several health grid projects have suggested that the data itself should have a 'lifetime' – users with temporary access should not be able to access the data (or a copy of the data) once their credentials have expired. This would be in agreement with the fifth principle of the UK Data Protection Act, 1998 and using a form of Digital Rights Management (DRM) has been proposed to provide this restriction [46].

4.7.3. Technological Requirements (Network and Data Storages)



TECHNOLOGY BASELINE REPORT

Doc. Identifier:
SHARE-D3.2_revised_FINAL

Date: I. Blanquer, V. Breton,
V. Hernández, Y. Legré, M.
Olive, T. Solomonides

There are many cases where medium to long-term grid storage of data will not be possible in order to protect ownership and custodianship, and to ensure the ethical control of information [38]. Medical data can be managed by autonomous, distributed organisations, and is constantly being changed and updated. Therefore the integration of data is an important issue for health grids. Additionally, biomedical data is proving more heterogeneous in nature than data from other research domains, with projects having to address semantic, syntactic, conceptual and temporal heterogeneity. New mechanisms for dealing with the integration of heterogeneous data, as well as dealing with missing, incomplete or uncertain data, are being developed [49]. It should also be noted that medical centres may not have enough bandwidth or the appropriate (technical) means for collecting and storing data in digital format.

WP4 deliverables D4.1 and D4.2 also address the issues of data storage which is legally considered as processing.

5. STATE OF THE ART OF THE DEPLOYMENT OF BIOMEDICAL APPLICATIONS ON GRIDS

5.1. INTRODUCTION

Grids benefit from a large funding from the European Commission and the member states. Among the present projects, the ones relevant to health can be roughly classified in three categories:

- Infrastructure projects aiming at offering a stable distributed environment for scientific production. Examples of such infrastructures are EGEE and DEISA in Europe. These infrastructures offer a generic multidisciplinary environment where biomedical applications can be deployed.
- Technology projects aiming at developing new grid-enabled services and environments relevant to the needs of life science and healthcare. Examples of such projects are SIMDAT [17] and MyGrid [18]
- End user projects focusing on specific life science or healthcare issues and integrate grid technology wherever they feel relevant. Examples of such projects are MammoGrid [19] and GEMSS [20].

5.2. ADOPTION OF GRIDS FOR BIOMEDICAL SCIENCES

Biomedical sciences have been identified very early as potential adopters of the grid technology. The wealth of data produced by life sciences in the last 10 years and its complexity requires more and more resources and services for their storage and analysis. Medical research is also evolving quickly with the generalized use of images and the growing integration of molecular biology in the perspective of individualized medicine.

5.2.1. Life science

Molecular biologists are facing a daunting challenge. The relevance of their research requires constant access to current databases containing all the knowledge acquired up to the moment. Comparative analysis is a mandatory step in most of the molecular biology data analysis workflows. This analysis has to be frequently repeated to keep up with the exponentially growing volume of data stored in the databases. Comparative analysis is often the first step of complex workflows needed to extract information from the data in genomics, transcriptomics and proteomics. At a basic level, grids can help distribute the databases in order to make them accessible to biologists [21] and provide the computing resources required by data analysis. Bioinformatics portals like GPS@ [22] are presently under development on top of grid infrastructures.

Grid technology also promises to address the complexity of biological data. Indeed, the last years have witnessed the development of hundreds of databases providing specific representations of biological data. Interoperability of these databases is a key to the development of integrated approaches needed to start modelling living organisms. Projects such as Embrace [23] focus on addressing this interoperability issue using grid technology.

Other projects such as MyGrid [18] have been developing tools and environments to ease the design of data analysis workflows for biologists. The next step is to achieve the integration and deployment of these high level interfaces on grid infrastructures so as to offer biologists the data and computing resources needed for their analysis.

5.2.2. Medical research

Grid technology entry points into medical research have been most often related to the need to manipulate large cohorts of medical images. The volume of medical images produced in European hospitals is comparable to the volume of data expected from the CERN Large Hadron Collider, which is of the order of several Petabytes per year. Storing these images and running algorithms to extract their features require more and more resources. Attempts to distribute storage of medical image databases on the grid have been frustrated by the very limited data management services available on grid infrastructures in Europe. Encouraging prospects are opening with the addition of data management services on infrastructures like EGEE but adoption of grids in medical research depends heavily on the availability and extension of such services.

Moreover, the availability of medical data in digital format is increasing steadily. On one side, the collection of clinical data in “production” is a reality in several countries at hospital level, primary care or even at national level. This information is progressively encompassing complex data such as vital signs, microbiological analysis and genomics, along with images for diagnosis, staging and treatment, subsequent diagnosis, treatment and actual prescription. On the other side, medical databanks are being created for oncology and blood samples in many countries, requiring not only the management of the physical samples but also additional information in electronic format. The use of this information is unexploited due to the strong requirements in processing, but also due to standardisation and interoperability problems. Attempts to use grids to confront patient medical and biological data are presently under exploration in several projects. The success of these approaches depends again on the capacity of the grid to provide the tools needed to manipulate these data.

5.2.3. Drug discovery

In silico drug discovery is one of the most promising strategies to speed-up the drug development process. Virtual screening is about selecting *in silico* the best candidate drugs acting on a given target protein. Screening can be done *in vitro* but it is very expensive as there are now millions of chemicals that can be synthesized. If it could be done *in silico* in a reliable way, one could reduce the number of molecules requiring *in vitro* and then *in vivo* testing from a few millions to a few hundreds.

In silico drug discovery should foster collaboration between public and private laboratories. It should also have an important societal impact by lowering the barrier to the development of new drugs for rare and neglected diseases. New drugs are needed for neglected diseases like malaria where parasites keep developing resistance to the existing drugs or sleeping sickness for which no new drug has been produced for years. New drugs against tuberculosis are also needed as the treatment now takes several months and is therefore hard to manage in developing countries.

In silico drug discovery on grids is a growing field. Grids like EGEE are ideally suited for the first step where docking probabilities are computed for millions of ligands. The relevance of the grid has been clearly demonstrated during the summer 2005 by the WISDOM initiative on malaria [24] where 46 million ligands were docked for a total amount of 80 CPU years (1 Tflop during 6 weeks).

It is possible to anticipate within the foreseeable future *in silico* drug discovery on the grid forming part of a development or production pipeline for new drugs. Such a pipeline would allow rapid identification of promising compounds or modifications to suit particular conditions. The first stage, which will be explored notably within European projects like BioInfoGrid, EGEE and Embrace, is the deployment of a virtual screening platform that would take advantage of the European grid infrastructures for docking and of a supercomputer for molecular dynamics computations.

5.3. ADOPTION OF GRIDS FOR HEALTHCARE

Adoption of grids for healthcare is still in its infancy. There are many reasons for this situation. A first obvious reason is that grid technology is still immature and is neither robust nor secure enough to offer the quality of service required for clinical routine. Another important reason is that all grid

infrastructure projects are deployed on national research and education networks that are separate from the networks used by healthcare structures. There are a number of legal issues at EC level and at EU member state level surrounding the international transfer of patient data between states that must be addressed, either through further research to determine ways the legal requirements can be satisfied, or by introducing specific legislation at a European level to support health grids [50].

This has not stopped pioneer projects exploring and demonstrating the potential impact and relevance of grids to address such outstanding healthcare issues as the early diagnosis of breast cancer [19] or to improve radiotherapy treatment planning [20]. Grids are expected to bring a significant added value in the development of individual medicine that requires the exploitation of biological and medical data, but this is still a research field. Adoption of grids for healthcare will follow their adoption for life sciences and medical research provided also the legal and ethical framework of the member states allows their deployment.

5.4. MAIN PROJECTS

This section presents the main projects that are currently exploring the usage of grids for medical research and healthcare. The projects are listed according to their funding scheme.

5.4.1. European projects funded within FP5

5.4.1.1. OpenMolGRID

OpenMolGRID (<http://www.openmolgrid.org>) was a pioneering project with the objective to speed-up, automatise, and standardise the drug-design using Grid technology. The design of molecular compounds relies on the knowledge that the properties of compounds are determined by the properties of the molecular fragments and their interaction. Molecular engineering makes use of this fact by building candidates for chemical compounds with predetermined target properties from appropriate fragments according to established rules. All constructed candidate structures are validated by quantitative structure-property/activity relationship (QSPR/QSAR) models. This process involves a large number of calculations in a single design task, thus the Grid approach is vital.

OpenMolGRID used UNICORE to integrate the various applications needed, from databases to molecular engineering and prediction modules. Building on the basic Grid middleware, substantial functionality were added to the client, and in the abstraction layers were used to hide the system complexity from the user.

The OpenMolGRID system is available at [SourceForge](https://sourceforge.net/projects/unicore/) (<https://sourceforge.net/projects/unicore/>), including:

- base UNICORE Client and Server software
- OpenMolGRID_WorkflowSupport client package
- OpenMolGRID_CLI command line interface package

5.4.1.2. GEMSS

GEMSS (Grid-Enabled Medical Simulation Services) was funded by the FP5 Programme under the cross-program theme Grid Testbeds. The project aimed to demonstrate advanced simulation and image processing services using grid technologies for improved preoperative planning and near real-time surgical support for practitioners and researchers. Six medical grid service prototypes were deployed on the GEMSS grid infrastructure, including maxillo-facial surgery simulation, neurosurgery support, radio-surgery planning, inhaled drug-delivery simulation, cardiovascular simulation, and advanced image reconstruction.



The GEMSS Grid infrastructure is based on web services, but has yet to be extended to be compliant with OGSA. GEMSS services are defined via WSDL and securely accessed using SOAP messages. Further integration with NHS systems in the UK (e.g. to allow use of the NHS hospital authentication system) was planned.

GEMSS dealt with highly confidential and private information, such as images of a patient's head. The legal situation regarding this was examined in detail by the project. To operate within EU law, medical data must be anonymised where possible and not held for longer than is required to achieve the purpose of the grid processing. Legal analysis determined that full anonymisation was not possible for head image scans or other inherently personal data, so these always had to be treated as confidential personal patient data. Also, although service security is guaranteed by the GEMSS middleware, the limited anonymisation of DICOM medical image data is still a manual process. Therefore implementing best practice security and having well defined contracts between the data controller and processors was seen as essential. No associated image data could be transferred outside the user's clinical department.

GEMSS has a three step, client-driven job submission process. The initial business step deals with opening accounts and fixing payment details. The pricing model can also be chosen at this stage. The next step is Quality of Service (QoS) negotiation, where a job's required turnaround time is specified. When a contract is in place, the job itself can finally be submitted and executed. Soft-coded workflows have been examined for the negotiation step, using the Scuf language for workflow definition (using Taverna), and FreeFluo (from the myGrid project) for workflow orchestration.

The project also suggested a number of possible business models for a commercial grid, supporting pay-per-use, resource reservation, and a legally admissible audit trail for all business transactions.

5.4.2. European projects funded within FP6: ICT for Health Unit

5.4.2.1. @neurIST

@neurIST (European co-funded Integrated Project) targets the management, analysis and evaluation of the exponentially growing volume of data describing all aspects of human disease processes, including their understanding, diagnosis and management. The data spans all scales, from molecular, through cellular to tissue, organ and patient representations. @neurIST will integrate all of the available information about a particular disease by gathering together a unique set of experts from all the involved disciplines: genetics, biology, medicine, bioinformatics, medical informatics, computational medical imaging, computational physiology, computer science and computerized decision making. The goal is to develop a vertical and integrative approach to knowledge discovery, personalized risk assessment, patient guideline generation and treatment design. Although vertical integration across data structures and across physical scales is the primary theme of this project, there is horizontal integration at every level of abstraction, from access to information sources, to evidence processing, knowledge representation, structuring and fusion. The chosen clinical application of this project is that of cerebral aneurysms and subarachnoid haemorrhage. While having intrinsic interest due to the societal impact of combating this disease, this scenario has also a number of interesting challenges that make it attractive as proof-of-concept of the envisaged approach. The project expects that such a focused effort can credibly address the expected vertical integration and allows identifying clear exploitation paths both industrially (e.g., in our case, decision support systems and advanced design of medical devices) and in supporting further medical research and knowledge discovery (e.g. in linking the molecular level of a disease with the disease process itself).

5.4.2.2. ACGT

The completion of the Human Genome Project sparked the development of many new tools for today's biomedical researcher to use in finding the mechanism behind disease. While the goal is clear, the path to such discoveries has been fraught with roadblocks in terms of technical, scientific, and sociological challenges.

ACGT brings together internationally recognised leaders in their respective fields, with the aim to deliver to the cancer research community an integrated Clinico-Genomic ICT environment enabled by a powerful grid infrastructure. In achieving this objective ACGT has formulated a coherent, integrated work plan for the design, development, integration and validation of all technologically challenging areas of work. Namely: (a) **Grid**: delivery of a European biomedical grid infrastructure offering seamless mediation services for sharing data and data-processing methods and tools, and advanced security; (b) **Integration**: semantic, ontology based integration of clinical and genomic/proteomic data - taking into account standard clinical and genomic ontologies and metadata; (c) **Knowledge Discovery**: Delivery of data-mining grid services in order to support and improve complex knowledge discovery processes.

The technological platform will be validated in concrete setting of advanced **clinical trials** on **cancer**. Pilot trials have been selected based on the presence of clear research objectives, raising the need to integrate data at all levels of the human being.

ACGT promotes the principle of open source and open access, thus enabling the gradual creation of a European biomedical grid on cancer. Hence, the project plans to introduce additional clinical trials during its lifecycle. It is in line with EU priorities and the objectives of the IST program. It targets the fulfilment of urgent needs of the cancer research community, a key area of societal importance and with a view to strengthening the integration of the European Research Area.

5.4.2.3. Health-e-child

There is a compelling demand for the integration and exploitation of heterogeneous biomedical information for improved clinical practice, medical research, and personalised healthcare for the citizens of EU.

The Health-e-Child project aims at developing an integrated healthcare platform for European Paediatrics, providing seamless integration of traditional and emerging sources of biomedical information. The long-term goal of the project is to provide uninhibited access to universal biomedical knowledge repositories for personalised and preventive healthcare, large-scale information-based biomedical research and training, and informed policy making.

The general objectives of Health-e-Child are the following:

- To gain a comprehensive view of a child's health by vertically integrating biomedical data, information, and knowledge, that spans the entire spectrum from genetic to clinical to epidemiological;
- To develop a biomedical information platform, supported by sophisticated and robust search, optimisation, and matching techniques for heterogeneous information, empowered by the grid;
- To build enabling tools and services on top of the Health-e-Child platform, that will lead to innovative and better healthcare solutions in Europe:
 - Integrated disease models exploiting all available information levels;
 - Database-guided biomedical decision support systems provisioning novel clinical practices and personalised healthcare for children;

- Large-scale, cross-modality, and longitudinal information fusion and data mining for biomedical knowledge discovery.

The project focus will be on individualised disease prevention, screening, early diagnosis, therapy and follow-up of paediatric heart diseases, inflammatory diseases, and brain tumours. The project will build a grid-enabled European network of leading clinical centres that will share and annotate biomedical data, validate systems clinically, and diffuse clinical excellence across Europe by setting up new technologies, clinical workflows, and standards.

The project has identified the following gaps in available technology, required to support the kind of heterogeneous biomedical data involved and the proposed medical applications:

- Search, similarity matching, and resource or quality of service optimisation of queries and processing requests,
- Decision support systems based on all available information for a patient,
- Knowledge discovery tools for heterogeneous data sources on a grid,
- Disease models capable of capturing the evolution of proteins, cells, tissues, organs, etc. as a child grows.

The project intends to develop algorithms, models and tools to fulfil these requirements, building upon the work of other projects including BIRN, InfoGenMed and IBHIS.

The following issues related to integrating distributed heterogeneous data sources must also be addressed:

- data-related issues – distribution, acquisition, normalisation, aggregation, and curation
- access-related issues – transaction management and query processing
- network-related issues – location and fragmentation
- privacy and security

Existing integration solutions will be examined, and both informatics and medical standards (i.e. HL7, DICOM, etc.) will be followed where feasible. However, the integration of biomedical information will be complicated by the fact that there is no universally accepted biomedical data model.

In order to facilitate the semantic integration of biomedical information to generate viable integrated case data, Health-e-Child will make extensive use of ontologies, and will investigate a number of approaches to mapping/bridging between ontologies.

5.4.2.4. Biopattern

The BIOPATTERN Network of Excellence [36] vision is to develop a pan-European, coherent and intelligent analysis of a citizen's bioprofile; to make the analysis of this bioprofile remotely accessible to patients and clinicians; and to exploit bioprofile to combat major diseases such as cancer and brain diseases.

A biopattern is the basic information (pattern) that provides clues about underlying clinical evidence for diagnosis and treatment of diseases. Typically, it is derived from specific data types, e.g. genomics information and vital biosignals such as the EEG. A bioprofile is a personal 'fingerprint' that fuses together a person's current and past medical history, biopatterns and prognosis. It combines data, analysis and predications of possible susceptibility to diseases.

BIOPATTERN proposes to provide novel computational intelligent techniques for biopattern analysis and a pan-European integrated, intelligent analysis of an individual's bioprofile. Information from distributed databases will be made available, securely, over the Internet to provide on-line algorithms, libraries and processing facilities for such analysis

5.4.3. European projects funded within FP6: Research and Infrastructure Unit

5.4.3.1. BioInfoGrid

The BIOINFOGRID SSA will establish a common ground for collaboration between the European grid Infrastructure providers and the Bioinformatics research user community in various fields of bioinformatics applications (Biology, Computational Chemistry, Medicine and Biotechnology). This will be achieved through specific feasibility studies for each reference application in the grid domain that will make it possible to implement meta-laboratories in which experts of various disciplines can collaborate on the solution of highly complex problems.

The EGEE/EGEE II project will deliver the stable grid based analysis service in Europe. The project will include applications for distributed laboratory management systems for microarray technology for gene expression studies and bioinformatics technology for data mining, gene discovery, sequence similarity searching of DNA and protein in the grid. The programme covers the most contemporary uncharted fields of investigation in biological and medical research. Such areas of interest have a well-established priority on the grounds of their scientific, economic and social relevance.

An international conference for grid Bioinformatics applications will be organized at the end of the project by inviting a large Bioinformatics user community as main dissemination activity.

5.4.3.2. EGEE-II

The Biomedical community has rapidly established itself as a key user of grid technology throughout Europe. Many suitable applications exist in this sector, which covers a broad applications domain from genomics through to medical imaging and healthcare applications.

In EGEE-II, the biomedicine application group will continue to steer middleware evolution with its specific and complex requirements for which further development is needed. The application development process will be focused around three flagship areas that were identified in EGEE: drug discovery, portals and workflows for medical image database analysis, and interactive/short-deadline jobs. Scientific output in the field of biomedicine will be privileged and a specific effort will be allocated to ease the access to the developed, grid-enabled biomedicine tools to a wider community of users.

5.4.4. European projects funded within FP6: Grid Technology Unit

5.4.4.1. SIMDAT

SIMDAT addressed the problem of multidisciplinary product design, where complex problems are often co-related. In these situations, the need to access and analyse data generated within different departments or sites is a key issue for many industries. Distributed data access and a clear semantic definition of the data sources involved would be required, in order to enable the retrieval of relevant information regardless of where it is stored. This kind of integration through a 'data grid' would require not only the mapping of semantics between the data repositories involved, but would also require tools to analyse and mine this data, facilitating the successful federation of multiple industrial problem-solving environments for product and process design. Security was also a key concern for this project, particularly where information about confidential processes could be leaked due to a supplier having need-to-know access.

The primary objectives of SIMDAT were:

- Testing and enhancing data grid technology for product development and production process design
- Developing federated versions of problem-solving environments
- Exploiting data grids for distributed knowledge discovery
- Promoting de-facto standards for these enhanced grid technologies

- Raising awareness for the advantages of data grids in key industrial sectors

Four application sectors were selected to cover a wide range of issues in design, development and production of complex products and services. These were the *aerospace*, *automotive* and *pharmaceutical* industries, as well as *meteorology*. A complex problem was defined as a use-case for each sector.

SIMDAT's integrated grid infrastructure software was successfully deployed for four SIMDAT prototypes for these sectors, in order to demonstrate how grid technology can support the collaborative development of complex products. The integrated grid infrastructure software is based on GRIA, and was found to be well suited for SIMDAT's industrial application sectors. The security approach adopted by GRIA was also found to fulfil largely the operational security requirements of the industries examined. SIMDAT determined that GT4's Community Authorisation Service (CAS) and gLite's Virtual Organisation Management Service (VOMS) are not well suited for supporting relationship management where there is no centralised authority to operate VO services, e.g. when each party must be seen as equal. An infrastructure based on WS-Trust/WS-Federation was therefore suggested.

SIMDAT found that GRIA would need improvements in its robustness, usability and maintainability in order to support commercial grid deployment, with functionality enhancements to support the specific requirements of the application areas. It was also noted that compliance with current grid standards and OGSA/WSRF would not be sufficient to overcome the problem of how to enable the dynamic federation of information and computational resources between different grid middleware. The application requirements showed that SIMDAT users would need grid infrastructure software that was portable between a number of different operating systems and hardware platforms, but this would require significant work and was therefore only partially addressed by the project.

Feedback has shown that the version of GRIA used had some weaknesses that can be traced to the specific Quality of Service (QoS) model and management mechanisms. For example, users find it hard to predict the QoS requirements for each job and data store, and resource reservation is not supported by the current resource model.

5.4.4.2. Chemomentum

To make Grids more useful for knowledge-oriented applications such as decision support systems and risk assessment, more effort in the areas of semantics, metadata and knowledge management in distributed, heterogeneous environments are needed.

The Chemomentum project aims to fill these gaps by taking up and enhancing state-of-the-art Grid technologies and applying them to real-world challenges in computational chemistry and related application areas. It will help drive the transformation of computing paradigms in these areas towards collaborative research and Grid computing.

Chemomentum will:

- Provide an integrated Grid solution for workflow-centric, complex applications with a focus on data management and knowledge. Place the end users into the focus, enabling them to use powerful tools in a natural and transparent fashion;
- Provide Grid-enabled applications, data services and knowledge management solutions, offering integrated decision support services for risk assessment, toxicity prediction and drug design;
- Setup and operate a stable pilot installation, accessible for external users. Proactively gather and evaluate feedback from these users. Simplify administration and management of the Grid;
- Spread awareness of the Chemomentum aims, scientific and technical approach, results and success stories in relevant industries and communities. Ensure maximum exploitation of the

services and possible products developed in Chemomentum;

- Test-drive the developed services in the context of the European regulatory initiative, Registration and Evaluation of Chemicals (REACH), promoting the REACH initiative aimed at reducing animal testing, by developing in silico, Grid-based tools.

5.4.5. European projects funded within FP6: DG-Research

5.4.5.1. Embrace

The objective of this network [32] is to draw together a wide group of experts throughout Europe who are involved in the use of information technology in the biomolecular sciences. The EMBRACE network aims at optimising informatics and information exploitation by pure and applied biological scientists in both the academic and commercial sectors.

The network works to integrate the major databases and software tools in bioinformatics, using existing methods and emerging grid service technologies. The integration efforts are driven by an expanding set of test problems representing key issues for bioinformatics service providers and end-user biologists. As a result, groups throughout Europe will be able to use the EMBRACE service interfaces for their own local or proprietary data and tools.

5.4.6. Projects funded by National Grid Initiatives or national funding agencies

5.4.6.1. BIRN

BIRN has been described in the previous chapter as it is also providing an infrastructure.

5.4.6.2. CLEF

CLEF (the Clinical e-Science Framework) was a research project from the UK, funded by the Medical Research Council, to produce a scalable architecture for a distributed, grid-based computing environment for clinical research. The project defined 'policies for information governance', dealing with issues of confidentiality, access, authentication, consent and security, and worked with the Judge Institute, Cambridge on the 'ethico-legal requirements for research use of clinical information'.

In the CLEF model, anonymisation is a two-step process. Automated 'pseudonymisation' is carried out by the originating institution to remove identifiers, and then residual identifiers (including confidentiality threats to third parties such as relatives, carers, doctors, or organisations) are manually removed ('depersonalisation'). Researchers can only access anonymous pooled information, not full individual records. Clinicians can view summaries and make queries, but answers are monitored to prevent re-identification by cross-reference or data mining, referred to as 'statistical disclosure control'. However, even after CLEF's rigorous pseudo/anonymisation, records must still be considered 'sensitive personal data' according to the UK Data Protection Act 1998. There is always the possibility that unusual or unique data concerning an individual might make a patient's record identifiable to someone with sufficient knowledge from other sources. Given that no technical solution can be perfect, CLEF determined that confidence in organisational measures is the most critical criterion.

All records in CLEF's repository contain detailed metadata on the level of consent granted for their use by patients, and one of CLEF's aims is to seek agreed standards for metadata on consent.

5.4.6.3. IBHIS

IBHIS (Integration Broker for Heterogeneous Information Sources) was a UK project funded by the Engineering and Physical Sciences Research Council. The project was intended as a proof of concept demonstration for the software-as-a-service (SaaS) and data-as-a-service models from the 1990's, and a key goal of the project was to provide a service-based architecture for the integration of information

that could easily adapt to changes in organisational/data structures. A series of prototypes were created to explore the issues in distributed information management in healthcare. These prototypes in particular revealed much about the current state of the art of web services.

IBHIS had to integrate data from different organisations, each legally bound by data protection issues and potentially different information policies. Integration issues included the format, semantics, meaning, importance, quality, ownership (and custodianship), cost and ethical control of sources.

The initial IBHIS prototype used a web service-based federated database management system. A Federated Scheme Service (FSS) maintained the Federated schema and all mappings with the export schema. Global integration however was a complex, manual process that was error prone and would take considerable time. The subsequent prototype used a broker service, which dynamically located and bound data sources that were not compile-time fixed. IBHIS addressed changing data sources through 'dynamic binding' with semantically described, published data-providing services.

Authorisation in IBHIS is handled primarily through Role Based Access Control (RBAC). However, RBAC was found to be too inflexible for the health domain, so IBHIS supplements it with a more complex set of user-based access rules to form a user profile. The second prototype used the Tees Confidentiality Model for finer grained rules, and for mapping between security domains.

During the development of IBHIS, it was determined that web service description languages and registries are not yet mature enough, particularly Web Services Description Language (WSDL) and Universal Description, Discovery and Integration (UDDI). These are mentioned elsewhere in this report.

5.4.6.4. eDiaMoND

eDiaMoND (Digital Mammography National Database) was a UK project funded by the Engineering and Physical Sciences Research Council, the Department for Trade and Industry, and a grant from IBM. The project explored the use of grid technology to support a 'real world' environment, namely the National Health Service's Breast Screening Programme. It focused on the development of a prototype for a national digital mammography archive, to aid clinicians in the screening and assessment process, and as a training resource.

Normalisation of image data was required due to differences in image contrast (which can make identification of microcalcification clusters harder) as well as other variables including tube voltage and the age of equipment. The image standardisation process used was Standard Mammogram Form (SMF) from Mirada, where 'Interesting tissue' is represented as a surface. This software was also used by the similar EC funded project, MammoGrid.

The project addressed a number of key security, confidentiality and ethical issues when dealing with patient data. Requirements such as the ability to provide time-limited access to data, limiting the amount of data provided without accidental disclosure, and portable and revocable credentials were recognised, and the project explored the legal and ethical issues in the UK regarding patient consent. The exploration of a number of these issues, as well as the legal and ethical obligations of healthcare professionals, was continued by the following project, GIMI.

5.4.6.5. GIMI

GIMI (Generic Infrastructure for Medical Informatics) is a UK project funded by the Department for Trade and Industry, and builds upon earlier work undertaken by eDiaMoND. The project is developing software for the secure and ethical aggregation of data from distributed sources, supporting applications for medical problems including diabetes, asthma, and image analysis for cancer research/training.



The key technical challenges for GIMI are interoperability and security; the system must be able to interface with other systems deployed within the NHS, and do so in a secure and ethical manner. The GIMI middleware layer provides security features referred to as an 'ethical firewall'. Its controls are based on legal and ethical requirements defined by the Data Protection Act (1998) and the Caldicott Guardian, amongst other sources.

Issues previously identified by eDiaMoND, such as how to provide temporary access to data, how to manage withheld information, and how best to enable portable credentials are being further explored by this project.

5.4.6.6. Integrative Biology

Integrative Biology (IB) is another UK project funded by the Engineering and Physical Sciences Research Council. The aim of the project was to demonstrate the usefulness of grid technology to enable computer simulations of the function of whole organs based on molecular and cellular level models, in order to increase understanding of the causes of heart disease and cancer. Multi-scale models of the heart and cancer tumours would be developed, spanning the range from genes to whole organs, with simulations of these models being manipulated at a spatial and temporal resolution not previously feasible.

A key requirement was to hide the complexity of the underlying IT infrastructure, but still make it easy to link to and exploit tools and techniques for complex, collaborative in-silico experiments. Enabling co-scheduled access to combinations of distributed resources would also be a significant challenge for the project.

Computational steering, workflow, and visualisation techniques from earlier projects including RealityGrid, myGrid, gViz, Godiva and Geodise would be used to exploit efficiently distributed resources.

The San Diego Supercomputing Centre's Storage Resource Broker (SRB) forms the basis of IB's secure data repository, for access control and managing virtual file locations. Authorised users can access a file in 'SRB space' from any networked computer, and file owners retain complete control over who can see and/or alter their files.

IB will be integrated within and will interact with different environments, each having its own security model. For example, there is a need to interact closely with the NGS infrastructure, where the Storage Resource Broker (SRB) data resides, which already has an extensive security model/infrastructure. Security enforcement will be context sensitive; it will attempt to determine not only who is accessing a source but also why.

IB uses a metadata catalogue based on CCLRC's Scientific Metadata Format (with extensions from myGrid) for contextual data, annotations and provenance metadata. Researchers manage their work using the metadata system, which builds the links needed to connect together all the relevant information about an experiment - newly created files are automatically associated with the appropriate logical metadata record. Provenance metadata is automatically captured about a simulation run and stored along with the results for later use.

5.4.7. International collaborations

5.4.7.1. WISDOM

WISDOM (World-wide In Silico Docking On Malaria) constitutes a first step toward an *in-silico* drug discovery pipeline on a grid infrastructure. Essentially, protein-ligand docking involves computing the binding energy of a protein target to a library of potential drugs using a scoring algorithm. The goal is to identify which molecules can dock on the protein's active sites in order to inhibit its action, and so



disrupt the pathogen. Critically, docking is currently the only application for distributed computing that has generated an interest in grid technology within the pharmaceutical industry. Docking is only the first step towards a full drug discovery pipeline; future challenges towards this goal will include molecular dynamics, lead optimisation and toxicity.

Three goals have driven the project. The biological goal was to propose new inhibitors for a family of proteins produced by *Plasmodium falciparum*. The biomedical informatics goal is the successful deployment of in-silico virtual docking on a grid infrastructure. Finally, the grid goal was the deployment of a CPU consuming application generating large data flows to test the grid operation and services.

This computationally intensive application was deployed successfully on the EGEE infrastructure during summer 2005; 46 million docking scores were performed in 6 weeks, equivalent to 80 years of computation time on a single PC. An improved version of the WISDOM production system using the later EGEE middleware gLite would be used for the next large-scale deployment in 2006.

A number of errors and efficiency issues with the EGEE middleware were identified during the initial WISDOM data challenge. The project noted that considerable knowledge of grid mechanisms was required in order to resolve quickly unforeseen errors and failures, which would be a concern for projects where jobs are being submitted by users with limited computing skills, as may be the case in medical research centres or hospitals. Additionally, human supervision was considered mandatory.

Data management has been identified as a bottleneck for biomedical data challenges on grid infrastructures that only support one Replica Location Service (RLS), such as the biomedical VO (Virtual Organisation) framework on the EGEE grid. There were significant problems with automatic job resubmission, described as a 'sink-hole' effect, which led to a significant number of aborted jobs and excessive job execution times. Checking, cancellation and resubmission of jobs had to be performed manually as a result. The project also highlighted a number of other issues with grid Information Systems (IS) and Resource Brokers (RB).

A number of improvements to the EGEE middleware were proposed, including better configuration/policy discovery by the grid IS, suggestions for improving the reliability of RBs, and better handling of errors/failures. These should be addressed by future versions of the gLite middleware.

6. DISCUSSION

In the previous chapter, we have provided an overview of the state of the art in terms of technology, infrastructure and application deployment. In this chapter, we are going to discuss a set of bottlenecks which we have identified on the road to wide grid adoption in the medical research and healthcare sectors.

6.1. TECHNICAL BOTTLENECKS

The HealthGrid vision (detailed on whitepaper.healthgrid.org and www.healthgrid.org) relies on the setting up of grid infrastructures for medical research and healthcare. The present bottlenecks towards this vision are the following:

- the availability of grid services, most notably for data and knowledge management;
- the deployment of these services on infrastructures involving healthcare centres such as hospitals, medical research laboratories, public health administrations;
- the definition and adoption of international standards and interoperability mechanisms for medical information stored on the health grid.

6.1.1. Development of Grid Data Management Services

Two worlds coexist today, the information world extensively using web services and the grid infrastructure world, which is slowly migrating to web services. Existing infrastructures in Europe are not yet based on this agreed standard because it takes years to develop a robust middleware and the migration to web services is a recent evolution in grid standards.

Adoption of grids for medical research and clinical routine depends on the capacity of grids to manipulate data in a secure and efficient way. Medical data are complex, highly sensitive and presented in multiple formats. Data management services offered by grid infrastructures must be very significantly improved in order to allow such manipulations. Importance of a large coordinated effort must be stressed to achieve this goal.

Integration of databases into the grid is a particularly critical issue. Initiatives such as SRB or OGSA-DAI [25] are particularly relevant but more efforts are needed in this direction, since the efficiency on large production schemas is not clearly addressed. On the other side, interfaces from grid storage elements (such as SRM) must pay attention to Medical Data storages, such as medical images, clinical data or vital signs.

6.1.2. Development of Grid Nodes in Health Care Centres

Another bottleneck is related to the installation and maintenance of grid nodes in healthcare centres. Such deployment is still in its infancy because the configuration of a grid node is rather complex and requires significant manpower. Moreover, as stressed above, secure services for data management are still under development, which could make firewall rules to be relaxed to enable the connections to grids. It is important to define new architectures and designs that will minimise data from being sent out of the hospital borders. Finally, there is a need of human resources to deal with these new services.

6.1.3. Development of services compliant with medical informatics standard specifications based on the web services technology

Chapter 2 of this paper illustrated on a very simple example the role of a health grid to exchange information between two hospitals in Europe. It also highlighted the need for a unique patient identifier allowing querying patient records while preserving their anonymity, for EHR data models publicly available and for an agreed patient summary with an agreed vocabulary to describe it. Work

is under way at a European level to address these issues. For the HealthGrid vision to be realised, standards must be agreed upon in the medical informatics community. This includes the development of applications obeying these standards, using grid services and available from grid nodes located in healthcare centres.

The development of standards is not a trivial task. The complexity of medical vocabularies and concept relations, and pressures of routine healthcare, make it rather difficult to adopt common procedures, protocols, coding schemas and templates. However, many initiatives have focussed on interoperability (IHE, DICOM, HL7, EN 13606, OpenEHR) and on coding schemes and nomenclatures (ICD9, ICDO, SNOMED, etc.). The adoption of these standards is progressive and does not cover the whole clinical scenario.

Finally, it is very important to outline that current standards are not “grid-aware”. DICOM and HL7 are based on client-server schema, EN13606 on a semantic structure for examples. It is important that grid-compliant protocols (such as WSRF) become part of the agenda of the normalisation teams.

6.1.4. Issues with Grid infrastructures under heavy use

The WISDOM data challenges are probably the best examples of health grid projects making heavy use of grid resources, and these have discovered a number of bottlenecks and efficiency issues [42].

Data management is a bottleneck for biomedical data challenges on grid infrastructures that only support one Replica Location Service (RLS), such as the biomedical VO framework on the EGEE grid at the time of the WISDOM data challenge. There have also been significant problems with automatic job resubmission. This has been described as a ‘sink-hole’ effect, and has led to a significant number of aborted jobs and excessive job execution times. Checking, cancellation and resubmission of jobs had to be performed manually as a result, and automatic job resubmission was still absent from the second (avian flu) WISDOM data challenge.

Specific knowledge of grid mechanisms was required in order to solve various unforeseen errors and failures, which would be a concern for projects where jobs are being submitted by users with limited computing skills, as may be the case in health grid projects. Often, error messages were unclear and did not provide enough information to diagnose the problem quickly. Additionally, continual human supervision was required, raising scalability concerns.

Reports from WISDOM have suggested a number of improvements to the EGEE middleware, including better configuration/policy discovery by the grid information system, suggestions for improving the reliability of Resource Brokers, and better handling of errors/failures. These should be incorporated into future versions of gLite.

6.1.5. Recording and ensuring consent

Although consent is largely an ethical and legal concern, how to record consent and ensure data is only being used for the purpose agreed to by the patient is something that must be addressed. Ideally, each record or piece of data would be accompanied by metadata specifying the level of consent granted, and although this has been proposed there is currently no agreement on a standard [51].

6.1.6. Anonymisation and pseudonymisation

Anonymisation is a very important aspect for health grids to protect patient privacy and confidentiality. Manual anonymisation or pseudonymisation can be a very time consuming task, particularly where patient narratives are involved, and therefore a number of methods have been employed to automate this process.

Various techniques are used, such as ‘named entity extraction’ and related techniques, such as those from AMBIT, and patient identifier algorithms using Semantic Selectional Restrictions [51, 52].

However, these are not foolproof and can incorrectly flag innocent data, such as proper names for drugs that could not be found in drug databases and medical conditions containing a recognised human name. They may also fail to identify risks when complicated grammar is used [52]. Further development of these techniques will therefore be required.

6.2. ORGANIZATIONAL BOTTLENECKS

6.2.1. Organization of the healthcare and medical research community

There is a gap between some of the needs of the medical research and the services offered by an infrastructure like EGEE. For instance, the resources that could be used in the hospital are not dedicated. Indeed, most of the computers in the hospitals are not used most of the time. Having the computers usable for research would be great but there is not much knowledge about grid technology among the people making decisions and taking decisions on the allocation of resources in the hospitals. Raising awareness requires disseminating information on grids outside the scientific world: achieving this dissemination requires having good success stories. There is also a need for easy to install demonstrators that people can test and play with.

Another important issue is the legal ownership of the patient record, and the patient's rights with regard to the data held about them. The legal ownership of the patient record is different in different countries, in some cases the patient record is said to belong to the patient, in others it belongs to the patient's hospital, clinic or trust. A patient may have the right to access and have corrections made to data that concerns them, and may also be entitled to feedback in the event that something of immediate clinical relevance is discovered. User groups within virtual organizations must be organized in an appropriate way to meet the legal requirements for manipulating patient data.

Finally, none of the resources inside the hospital are accessible from the outside world. As a consequence, a grid node in a hospital would have to be located outside the hospital firewall. Only anonymised or pseudonymised data could be transferred from inside the hospital to the grid node. Moreover, the connections to Internet of clinical centres are often intentionally low-bandwidth and limited. Security threats in medical data are very important and risks are avoided in such a manner that could prevent the set up of grid nodes or the transferring of data and results. Technical solutions must prove that security is not compromised and reduce the bandwidth limitations.

6.2.2. Development of best practices

Best practices must be defined on how to use both computing and data resources. Most of the computing resources in healthcare centres are available for clinical routine, but a model should be developed how to use part of the available power for medical research using the grid technology.

A recurrent issue is also how to make data available for research projects. Healthcare professionals are reluctant to share data. However, data and knowledge produced by European projects should be made available openly to the next generation of projects. Molecular biology databases are examples of successful knowledge repositories. Within FP7, funding will be allocated to knowledge repositories. It is necessary to set up infrastructures to collect, manage and keep available data produced by research projects specific to biomedical research. For instance, a Global Trial Bank of clinical trials [34] storing the results even negative results would help store and maintain the global knowledge.

6.2.3. Technology Transfer between EC Projects

As a consequence of the technical bottlenecks previously identified, very few projects led by biomedical end users are deployed on the European grid infrastructures available today. This is due most notably to the limited data management services offered by the infrastructures, their still user-unfriendly interfaces and the lack of information and training on grids in the biomedical community.



Interesting data management services are under development by some technology oriented projects but the mechanism by which they will be deployed on existing grid infrastructures is unclear. The present funding scheme of the European Community does not allow today to fund transversal projects like BIRN where all levels from hardware to end user applications are supported.

6.2.4. Worldwide Open Standards in Medical Informatics

We have demonstrated in chapter 2 how a European infrastructure such as a health grid depends on the definition of open standards. These open standards are needed to achieve interoperability of healthcare systems and records. The development of these standards requires coordination. The lack of agreed open standards in medical informatics will be an obstacle to any large-scale infrastructure deployment. The absence of a reference body or structure in charge of defining the connection between medical informatics standards and grid standards is a clear bottleneck to the development of grid technologies in healthcare.

6.3. COMMUNICATION BOTTLENECKS

There is an evident lack of information on the grid technology in the biomedical community. It is very important to target decision makers and budget holders to achieve the adoption of the grid technology. But to convince these people, it is essential to have success stories. Building success stories requires developing specific scenarios demonstrating the grid added value and to develop prototypes to be tested in “real” situations. But the relevance of the prototypes to address real medical issues depends on the involvement of medical experts and it is important to find incentives to involve these experts. The technology must offer services that are needed by the medical research community.

Communication on grids has been mostly focussed on the particle physics and computer sciences academic communities. It is important to start communicating on grids in events attended by opinion and decision makers attend in order to educate the field about the services that are coming.

7. CONCLUSION

This document has presented a state of the art of grid technology for healthcare. We have described the technology of web services that is the foreseeable best candidate to enable the HealthGrid vision. We have reviewed the main grid infrastructures currently deployed around the world as well as the status of the biomedical applications deployed around the world.

We have identified a set of bottlenecks that are going to be further documented in the next SHARE deliverable proposing a first technical roadmap for health grids:

- Secure and robust data management on the grid using standard web service technology
- Deployment of grid nodes in healthcare and medical research centres
- Development of services compliant with medical informatics standard specifications based on the web services technology
- Development of knowledge management services using ontologies

These 4 challenges are addressed to different communities:

- The first challenge is clearly in the hands of the grid technology developers
- The second challenge lies in the hands of the e-infrastructure designers
- The third challenge lies in the hands of the medical informatics community deploying applications on the grid
- The fourth challenge lies in the hands of the user communities

Besides these technical challenges, we have identified organizational bottlenecks:

- organisation of the medical research and healthcare communities
- development of best practices
- technology transfer between EC projects
- worldwide open standards in medical informatics

We have also highlighted the importance of communicating about grid technology in the world of healthcare and medical research.