



SHARE

TECHNOLOGY COMPONENTS ROADMAP II

Document ID:	SHARE-D3.4_v1.1
Date:	31/07/07
Authors:	I. Blanquer, V. Breton, V. Hernandez, Y. Legré, N. Jacq, M. Olive, T. Solomonides
Activity:	WP3: Infrastructure and security
Document status:	FINAL
Document link:	http://eu-share.org/deliverables.html
Confidentiality:	Public
Keywords:	Roadmap, healthgrid, technology, security

Abstract: This document presents the second and final version of the technical roadmap for the adoption of grids in the healthcare sector. The document builds upon the work performed in SHARE work packages WP3, WP4, WP5 and WP6. It integrates the feedback of user communities in the field of epidemiology and innovative medicine. Recommendations for concrete implementations are proposed.

**Document Log**

Issue	Date	Comment	Author
0.1	03/07/07	First version	V. Breton
1.0	25/07/07	Second version	V. Breton
1.1	31/07/07	Modifications on content and form. Delivered for internal reviewing	N. Jacq
1.2	25/09/07	Addition of roadmaps; addition of executive summary.	M. Olive
1.3	27/09/07	Addition of commentary on road maps.	T. Solomonides



CONTENT

1. INTRODUCTION..... 4

1.1. PURPOSE..... 4

1.2. APPLICATION AREA..... 4

1.3. REFERENCES..... 4

1.4. DOCUMENT EVOLUTION PROCEDURE..... 4

1.5. TERMINOLOGY..... 4

1.6. ACKNOWLEDGEMENTS 6

2. EXECUTIVE SUMMARY..... 7

3. HEALTHGRID VISION..... 8

3.1. DEFINITIONS 8

3.1.1. Grid..... 8

3.1.2. Computing Grid..... 8

3.1.3. Data Grid..... 8

3.1.4. Knowledge Grid..... 9

3.1.5. Healthgrid..... 9

3.2. THE DIFFERENT LAYERS OF A HEALTHGRID..... 9

4. RATIONALE 11

5. INTERNATIONAL CONTEXT..... 14

5.1. BIOMEDICAL ROADMAPS..... 14

5.1.1. *The Innovative Medicines Initiative (IMI) Strategic Research Agenda*..... 14

5.1.2. *STEP: a roadmap to the Virtual Physiological Human*..... 14

5.2. TECHNICAL ROADMAPS..... 14

5.2.1. *Reports from Next Generation Grid expert group*..... 14

5.2.2. *GridCoord*..... 15

6. USERS REQUIREMENTS 16

6.1. REQUIREMENTS FROM EPIDEMIOLOGY..... 16

6.1.1. *Introduction*..... 16

6.1.2. *Research Requirements*..... 16

6.2. REQUIREMENTS FROM INNOVATIVE MEDICINE..... 17

6.2.1. *Introduction*..... 17

6.2.2. *Research Requirements*..... 18

6.3. REQUIREMENTS FROM VIRTUAL PHYSIOLOGICAL HUMAN..... 18

6.3.1. *Introduction*..... 18

6.3.2. *Research Requirements*..... 19

7. ROADMAPS 21

7.1. ANALYSIS OF USER COMMUNITIES REQUIREMENTS..... 21

7.1.1. *Research challenges for computing grids*..... 21

7.1.2. *Research challenges for data grids*..... 23

7.1.3. *Research challenges for knowledge grids*..... 25

7.1.4. *Deployment milestones*..... 27

7.2. PROPOSED ROADMAPS..... 28

8. CONCRETE IMPLEMENTATIONS..... 31

1. INTRODUCTION

1.1. PURPOSE

The purpose of the document is to describe a roadmap towards the adoption of grid technology in Health and Life Sciences.

This document is written in preparation of the final roadmap that will include technology, but also ethical, legal, social and economic issues. .

1.2. APPLICATION AREA

The document is intended for internal and external use. It will be circulated within the SHARE project and the user communities for further discussion and amendments.

1.3. REFERENCES

[1] Seeding the Europhysiome: A Roadmap for the Virtual Physiological Human, STEP Consortium, <http://www.europhysiome.org> and http://www.biomedtown.org/biomed_town/STEP/Reception/step_presentations/RoadMap/vph_roadmap_v2b.pdf

[2] The Virtual Physiological Human: a true grand challenge for large scale grid infrastructures, Marco Viceconti, Peter Coveney, Gordon Clapworthy (STEP Consortium) Vincent Breton, Yannick Legre (SHARE Consortium), <http://eu-share.org/deliverables.html>

[3] SHARE epidemiology roadmap, deliverable D5.2a, <http://eu-share.org/deliverables.html>

[4] The Innovative Medicines Initiative (IMI) Strategic Research Agenda. Creating Biomedical R&D Leadership for Europe to Benefit Patients and Society, 15 September 2006 available at <http://www.imi-europe.org/DocStorage/PublicSiteAdmin/Publications/Innovative%20Medicines%20Initiative%20SRA%20Version%202.0.pdf>

[5] SHARE Innovative Medicine Roadmap, deliverable D5.2b, <http://eu-share.org/deliverables.html>

[6] BIRN project : <http://www.nbirn.net/>

[7] DEISA project, <http://www.deisa.org>

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 authors.

1.5. TERMINOLOGY

Abbreviation List

CA	Consortium Agreement
CC	Cost Claims
EAC	External Advisory Committee
MB	Management Board



PD	Project Director
PEC	Project Executive Committee
PM	Person Month
PM xx	Project Month xx
PO	Project Office
PR	Periodic Reports
QA	Quality Assurance
QR	Quarterly Report
VPH	Virtual Physiological Human
WP3	SHARE Technology and Security Activity

Definitions

The following definitions are useful to understand the document content.

- **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 virtualising resources
- **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.
- **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.
- **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 like 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
- **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 which allow the output of one component to be used as the input of another, or control links which state some conditions on the execution of a component.

1.6. ACKNOWLEDGEMENTS

We would like to acknowledge numerous contributions to this document.

2. EXECUTIVE SUMMARY

The HealthGrid vision defines a multi-layer service oriented architecture in order to enable the creation of computing, data and knowledge grids for use by healthcare professionals for research. In order to achieve this vision, a number of technical and organisational bottlenecks and research challenges have been identified, leading to the creation of a roadmap for future research in Europe. Rather than the linear progression from computing to knowledge proposed previously, computing and data grids are now seen as the ‘pillars’ on which knowledge grids will be built; as a result, a number of research tasks are expected to occur in parallel for computing and data grids.

In this second version of the technical roadmap, the views and user requirements of other research communities have been incorporated, including not only the roadmaps developed in response to epidemiology and innovative medicine communities (see deliverables D5.2a and D5.2b) but also from the Virtual Physiological Human (VPH) community. Roadmaps from other expert groups and European projects such as Next Generation Grid (NGG) and GridCoord have also helped inform this second roadmap.

After examining the research requirements of each community in detail, with a particular emphasis on user requirements, research challenges have been identified for computing, data and knowledge grids, grouped by common themes. User-driven milestones for deployment have also been defined. The milestones and research challenge themes are shown on a timeline according to their complexity, as well as on diagrams to show overlaps and concurrency. In addition to these research milestones, a number of concrete implementations are also recommended.

This roadmap will, in addition to the map of ethical, legal and socio-economic (ELSE) issues identified (see D4.3), feed into the final integrated roadmap (D6.2), the final deliverable of the project.

3. HEALTHGRID VISION

In this chapter, we would like to present the HealthGrid vision that this roadmap wishes to reach. For this purpose, we introduce first a number of definitions to help the reader.

3.1. DEFINITIONS

3.1.1. Grid

In this document, we use the definition of the grid used by the CoreGRID project: “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 virtualising resources.”

3.1.2. Computing Grid

A **computing grid** is a fully distributed, dynamically reconfigurable, scalable and autonomous infrastructure providing location independent, pervasive, reliable, secure and efficient access to a coordinated set of **computing resources**.

The services offered by the computing grid are geared toward offering to its users the highest possible quality of service in terms of response time, crunching factor and interactivity. Data are temporarily moved and replicated on the grid in relation to the jobs submitted.

A very good example of a computing grid is DEISA. 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 jointly build and operate 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.

There have also been successful examples of computing grid applications in the health domain, particularly in the initial steps of drug discovery with projects such as WISDOM and OpenMolGRID being good examples of these.

3.1.3. Data Grid

A **data grid** is a fully distributed, dynamically reconfigurable, scalable and autonomous infrastructure providing location independent, pervasive, reliable, secure and efficient access to a coordinated set of **data repositories**.

The services offered by the data grid are geared toward offering to its users the highest possible quality of service in terms of secured data storage and access to distributed data. Computing resources are used only at a local level for the analysis of data previously queried on the infrastructure.

A very good example of a data grid is BIRN. Launched in 2001 with the support of the National Institutes of Health, the Biomedical Informatics Research Network (BIRN) [6] 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 hospital.

BIRN infrastructure is based on a client-server middleware developed at San Diego Supercomputing Centre which was designed for managing file collections in a heterogeneous, distributed environment.

3.1.4. Knowledge Grid

A **knowledge 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 knowledge** and virtualising resources

Compared to a data or a computing grid, the knowledge grid added value is to offer services which encapsulate knowledge built upon distributed data. Such services require regular query, analysis and integration of new data in order to update the knowledge. As a consequence, a necessary condition for the deployment of a knowledge grid is the availability of services for distributed data management and integration. The non local nature of the grid requires also the existence of agreed standards and vocabularies to describe information among the grid customers and data providers.

There is no knowledge grid presently deployed in the world.

3.1.5. Healthgrid

A healthgrid is an environment created through the sharing of resources, in which heterogeneous and dispersed health data as well as applications can be accessed by all users as a tailored information providing system according to their authorisation and without loss of information. The environment should allow multiple usages depending on the user community and provide means to share information and access resources on demand for any groups involved in medical and life sciences research.

The ultimate aim is the realisation of a healthgrid infrastructure supporting a distributed/federated, service oriented, and ontology driven architecture, providing a collaboration medium, facilitating effective computation and being capable of generating, organising and managing knowledge. Support for collaboration in particular is a significant dimension of healthgrids, where for example situations such as referral to a colleague for a second opinion on an x-ray image will require supporting the communication between both parties.

3.2. THE DIFFERENT LAYERS OF A HEALTHGRID

The HealthGrid vision is to create, through the innovative use of the grid technologies, an environment where information at all levels of the biological hierarchy (molecule, cell, tissue, individual, population) can be associated for the benefit of medical research and life sciences and to provide individualized healthcare.

To represent the Service Oriented Architecture of a healthgrid, it is convenient to introduce three layers of services as shown on figure 1:

- at the bottom, core services are provided by the infrastructure for job and data management. They include all the security mechanisms for accessing the grid. Generic grid services provided by the middleware are not specific to a healthgrid.
- healthgrid services are specific to the needs of the users in life and medical sciences. This for example could include services for pseudo/anonymisation or specific services for the storage of medical images.
- on top of these services are built the applications using these services. These applications integrate domain-specific and generic tools. In some cases, domain-specific services such as mammogram standardisation could be more appropriate at this level rather than at the more generic healthgrid service level.

Figure 1 illustrates also how the different kinds of grids defined in the previous section provide different families of services. Computing grids and data grids can be seen as the pillars on which

knowledge grids are built; knowledge grids will require the achievement of both computing and data grids.

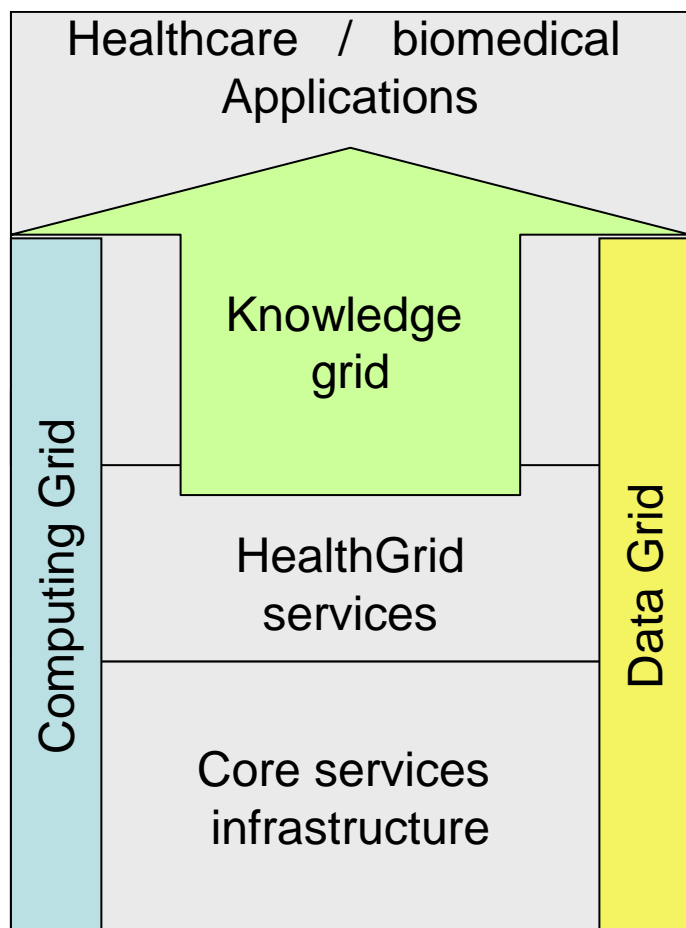


Figure 1: layers of services in a healthgrid

4. RATIONALE

This document proposes a roadmap to achieve a wide adoption of healthgrids in the life sciences and medical research communities. It is the result of the different steps highlighted below.

The content of the document builds upon 4 years of work which started in June 2003 when 40 experts in the fields of life sciences and biomedical research contributed to a collective reflection on the relevance and potential impact of grids in the health sector. The outcome of this reflection became the HealthGrid White Paper [REF] published in 2004.

Starting from the conclusions of the HealthGrid White Paper, the Technology Baseline Report D3.2 [REF] presents an analysis of the status of the technologies relevant to healthgrids as well as a status of the deployment of biomedical applications on grid infrastructures:

- Technologies (Web services, e-science environments), user requirements (standards, security...) and other grid technology roadmaps (SOKU, Challengers projects) are analyzed;
- Grid infrastructures (EGEE, DEISA, NorduGrid, OSG, TeraGrid...) are reviewed from a biomedical user perspective (services, access to resources);
- Successfully deployment of biomedical applications and healthgrid projects on grid infrastructures are analyzed (OpenMolGrid, GEMSS, SIMDAT, BIRN, Embrace, WISDOM...).

From this 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.

Starting from this analysis, The Technology Component Roadmap I document proposes a 10-year roadmap to achieve the goal to offer to healthcare professionals an environment created through the sharing of resources, in which heterogeneous and dispersed health data as well as applications can be accessed by all users as a tailored information providing system according to their authorisation and

without loss of information. The document identifies milestones and presents short term objectives on the road to this healthgrid.

We have identified 5 milestones on the road:

- MD1, called “Computing grid”, corresponding to the successful permanent deployment of computing grid nodes inside European medical research centres which we will call “grid”
- MD2, called “Data grid”, corresponding to the successful permanent deployment of data grid nodes inside European medical research centres
- MD3, called “Research K-Grid”, corresponding to the successful permanent deployment of knowledge grid nodes inside European medical research centres
- MS1, called “Grid DICOM”, corresponding to the production of a standard for the exchange of medical images on the grid
- MS2, called “Grid EHR”, corresponding to the production of a standard for the exchange of Electronic Healthcare Records on the grid

Achieving these different milestones require the availability of a grid operating system providing all the needed functionalities as well as an easy-to-install distribution of this middleware.

We have identified major technical risks which can prevent the vision to happen:

- the absence of concentration of the critical mass of expertise to develop the grid middleware and its distribution
- the absence of agreed standards to share medical images and Electronic Health Records on the grid
- the non-adoption of the healthgrid infrastructures by the research community so that they are not used by medical applications

To these technical risks, we have added a fourth major risk which is the absence of evolution of the legal texts to allow sharing of medical data which can prevent the deployment of healthgrids.

Finally, we have proposed in the next 2 to 3 years to develop R&D activities along three lines

- Develop healthgrid infrastructures
- Deploy biomedical grid applications on the existing infrastructures
- Deploy biomedical grid applications using OGSA compliant grid toolkits and e-science environments

In parallel to these R&D activities, we recommend to pursue actively the definition of standards for the sharing of medical images and electronic health records on the grid in the already existing medical informatics standardization bodies. We consider that the HealthGrid initiative provides the right framework to coordinate the development of the different standards in collaboration with the OGF and the different medical informatics standardization bodies.

The Technology Component Roadmap I document was critically evaluated by user communities in the field of epidemiology and innovative medicine. Their analyses are delivered in the Epidemiology Roadmap document: D5.2a, and the Innovative Medicine Roadmap: D5.2b. These documents present an analysis of current technologies in grids and the ethical, legal, social and economic (ELSE) framework of grid technologies for their adoption in epidemiology and in innovative medicine. They propose milestones to be achieved for the uptake of healthgrids in these application domains. The feedback of user communities in the field of epidemiology and innovative medicine is so integrated in the present document: D3.4, chapter 6.

It is well understood by the research community promoting the Virtual Physiological Human (VPH) research program that grid technology is required to pursue effectively this ambitious goal. The concept of Virtual Physiological Human (VPH) indicates a methodological and technological



framework that once established will enable the investigation of the human body as a single complex system. In an attempt to bridge the gap between the VPH community and the grid community, a group of experts from the STEP (Strategy for The EuroPhysiome) and SHARE consortia exchanged views on the relevance of grids for VPH [2]. Requirements and feedback from the VPH community are added in this document (chapter 6) to enrich the final SHARE roadmap. It was not foreseen initially in the SHARE project.

5. INTERNATIONAL CONTEXT

In this chapter, we describe how this technical roadmap fits into other initiatives to propose biomedical or technical roadmaps.

5.1. BIOMEDICAL ROADMAPS

The Innovative Medicines Initiative and the EuroPhysiome proposed a roadmap for innovative medicines and for VPH respectively.

5.1.1. The Innovative Medicines Initiative (IMI) Strategic Research Agenda

At the request of the European Commission, the European Federation of Pharmaceutical Industries and Associations (EFPIA) has identified the main barriers to innovation in Life Sciences research in Europe with the objective of establishing a European technology platform for innovative medicines. A document was produced by all relevant stakeholders in July 2005 describing the Strategic Research Agenda for the Innovative Medicines Initiative [REF].

This document has been used as the basis for the analysis in the deliverable D5.2b of the present bottlenecks in the biomedical R&D process as well as the recommendations on how to address these bottlenecks. The document D5.2b compares the proposed roadmap for healthgrid adoption to the technology platform of the strategic research agenda of the Innovative Medicine Initiative. As a result of this comparison, it is recommended that the HealthGrid roadmap explicitly addresses the development of enhanced knowledge representation models and data exchange standards for complex systems, as well as the definition of standards for the federation of data bases.

5.1.2. STEP: a roadmap to the Virtual Physiological Human

STEP: a Strategy for The EuroPhysiome is a coordination action partially supported by the European Commission. The document, compiled through a consensus process that has involved more than 300 stakeholders from research, industry and clinical practice, aims to provide a roadmap for the development of the VPH. STEP represents a collective European response to the individual actions by creating an integrated framework - EuroPhysiome - which, while remaining true to the overall physiome concept, can accelerate the progress of the European teams by avoiding redundancy, enhancing compatibility, identifying deliverables and time scales, etc. The coordination action finished in march 2007.

It is well understood by the research community promoting the VPH research program that grid technology is required to pursue effectively this ambitious goal. In an attempt to bridge the gap between the VPH community and the grid community, a group of experts from the STEP and SHARE consortia exchanged views on the relevance of grids for VPH [REF].

5.2. TECHNICAL ROADMAPS

5.2.1. Reports from Next Generation Grid expert group

Starting in 2004, the European Commission has convened a group of high-level experts, named the Next Generation Grid (NGG) expert group, to develop a European vision for grid research. Driven by the need and opportunity of bringing grid capabilities to business and citizens, the NGG vision underpins the evolution of grid from a tool to solve compute- and data-intensive problems towards a general-purpose infrastructure enabling complex business processes and workflows across virtual organisations spanning multiple administrative domains. The expert group, which has been convened



three times since its inception in 2004 is comprised of leading figures from the grid research academic community and from industry.

According to these reports [REF], 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.

5.2.2. GridCoord

GRidCoord is an ERA pilot initiative on coordinated grid research in Europe. The first objective of the GridCoord project is to strengthen co-operation amongst the funding authorities in order to better coordinate the planning of future activities in the field of grid research, an ERA objective. A second objective is to enhance the already ongoing collaboration among the research actors and users. A third objective is to develop, based on the above, national and EU Programme visions and roadmaps enabling Europe to play a leadership role in grid technologies and applications. In September 2006, the GridCoord consortium published a summary report about the Grid R&D Vision and Roadmap. The vision is of a world-leading grid utility of resources, tools and applications that is a key enabler for European Research, Industry and Commerce.

6. USERS REQUIREMENTS

The requirements were collected from three user communities: epidemiology, innovative medicines and Virtual Physiological Human communities. They are described in the deliverables D5.2a and D5.2b and in [1]. They are summarized in this chapter.

6.1. REQUIREMENTS FROM EPIDEMIOLOGY

6.1.1. Introduction

As documented in SHARE deliverable D5.2a [3], epidemiology and more generally ICT-driven research that uses health data focuses on two areas:

- Patient-customised research: personalised therapy, advanced diagnosis, bio-simulation and genomic analysis are the main issues.
- Population-level research: epidemiological studies, surveillance networks and therapy assessments are the main study areas.

Both scenarios share in general the problem of access to distributed, critically sensitive and heterogeneous data, resulting in overall costly computing processes. Patient-centric analyses normally deal with smaller amounts of data and require a pre-existing knowledge of models of healthy and diseased organs or tissues. Population-level analyses normally deal with the integration of larger, poorer-quality data. Semantics are especially relevant in those approaches.

Users ought to be able to take for granted:

- that the security mechanisms are sufficient to protect their data. Other than being sensitive to security issues, they should not need to know anything in detail about encryption, secure transfer, delegation or other technical issues.
- that the system will meet the concerns of the ethical and legal committees of their research institutions.
- that the results of their research will be private and available to third parties only if desirable. They will want to be able to define groups and permissions at a global scale for their research community.
- that the services are reliable, efficient and permanent. They may not understand, or want a detailed explanation of why a service is down, or why a job is taking so long. They are expecting a quality of service similar to any other utility.
- that they do not have to change significantly their current procedures, protocols or workflow. They should be able to use the same tools as usual, but with an enhanced productivity.
- that the data is somehow automatically organised and gathered, thus available for further exploitation. They will be aware of problems such as lack of coding, heterogeneity or data distribution/delivery but will not need to provide solutions.

6.1.2. Research Requirements

Requirements in a broad sense can be summarised as follows:

- effective semantic annotation of data. Data is poorly coded and interoperability of coding is not trivial. Extracting knowledge from medical data, however, is a main objective.

- effective integration of distributed and heterogeneous data. Integrating distributed resources requires exchange protocols, secure mechanisms, patient de- and re-identification, and automatic data analysis services.
- availability of efficient infrastructures and usage policies. Applications will require resources and reliable infrastructure to work on under a clear Quality of Service (QoS) promise.
- user-friendliness of applications and services. The tools should be available through protocols and interfaces similar to those used in the users' normal research. Not only must the applications be as compliant as possible with current systems and interfaces, but so must the technologies.
- ensuring that the research is done in a secure and legally-compliant framework. Legal and ethical constraints are misunderstood or ignored in some, perhaps most health research.
- reliability, scalability and pervasiveness. All the previous services must be robust and trustful and should be scaled without reducing performance

6.2. REQUIREMENTS FROM INNOVATIVE MEDICINE

6.2.1. Introduction

At the request of the European Commission, the European Federation of Pharmaceutical Industries and Associations (EFPIA) has identified the main barriers to innovation in Life Sciences research in Europe with the objective of establishing a European technology platform for innovative medicines. A document was produced by all relevant stakeholders describing the Strategic Research Agenda for the Innovative Medicines Initiative [4]. This document has been used in deliverable D5.2b [5] as the basis for the analysis of the research challenges in the biomedical R&D process as well as the recommendations on how to address these challenges.

Among other issues, the discovery and development of new drugs is very costly and attrition rates are high. Initiatives to reduce the rate of attrition during later phases of development are clearly desirable and if successfully implemented will reduce costs.

EFPIA's Research Directors Group has identified pre-competitive barriers to innovation. The objective for the future would be to identify as soon as possible in the pre-clinical phase:

- Reasons for lack of efficacy, despite promising pre-clinical data.
- The potential for adverse drug reactions and pre-clinical toxicity.

The identified key bottlenecks in the R&D process are the following:

- predictive pharmacology at the discovery research stage;
- predictive toxicology at the preclinical development stage;
- identification of biomarkers at the translational medicine stage;
- patient recruitment and validation of biomarkers at the clinical development stage;
- risk assessment with regulatory authorities at the pharmacovigilance stage.

In these areas, scientific and technological advances would be of direct benefit to the pharmaceutical industry by improving efficacy of tests and containing costs.

The knowledge management area is identified as key to leveraging the potential of new technologies such as genomics and proteomics and to analyse the huge quantity and diversity of information in an integrated way.

The report identifies two levels of knowledge management that need to be addressed:

- The capture, analysis and interpretation of knowledge generated regarding the physiology and pathophysiology related to disease stage or toxicological targets. Here the aim is to improve the understanding of the underlying process including the impact of pharmacogenomics in order to predict successfully the validity of a drug target and risk management for patient populations
- The capture, analysis and interpretation of knowledge generated for one potential drug candidate from discovery, non-clinical and clinical development all the way to lifecycle management. The aim here is to integrate all available knowledge at any given stage of the development process in order to make the best predictions possible for the chances of success of this molecule in the next stage. The know-how for an integrated model-based drug development tool is available in Europe but one of the major bottlenecks is the lack of availability of databases across R&D that might facilitate data integration.

6.2.2. Research Requirements

The levels of knowledge management identified in the previous section translate into scientific requirements:

- Capacity to search, query, extract, integrate and share data in a scientifically and semantically consistent manner across heterogeneous sources (public and proprietary) ranging from chemical structures and “omics” to clinical trial data,
- Capacity to integrate and share scientific tools (e.g., modelling, simulation) as modules in a generic framework and apply them to relevant dynamic data sets,
- Expressive data representation and exchange standards,
- Dynamic and customizable configuration of applications,
- Encapsulation of validated physiological models, when applicable,
- Flexible, secure (covering all aspects of data protection encountered in a biomedical context), and scalable IT infrastructure.

These requirements are not specific to grids but healthgrids can become relevant infrastructures for biopharmaceutical research and development provided the technology matures to support a distributed/federated, service oriented, and ontology driven architecture which provides a collaboration medium, facilitates effective computation and is capable of generating, organising and managing knowledge.

6.3. REQUIREMENTS FROM VIRTUAL PHYSIOLOGICAL HUMAN

6.3.1. Introduction

The concept of *Virtual Physiological Human* (VPH) indicates a methodological and technological framework that once established will enable the investigation of the human body as a single complex system. At the current state of consensus [1], such framework should fulfil three main attributes:

- Descriptive: a framework within which observations made in the laboratories, in the hospitals, and in the field all over the world can be collected, catalogued, organised, shared and combined in any possible way
- Integrative: a framework that allows experts to collaboratively analyse this observations and develop systemic hypotheses that involve the knowledge of multiple scientific disciplines

- Predictive: a framework that makes possible to interconnect predictive models defined at different scale, with different methods, and with different levels of detail, into systemic networks that provide concretisation to those systemic hypotheses, and make possible to verify their validity by comparison with other clinical or laboratory observations

It is well understood by the research community promoting the VPH research program that grid technology is required to pursue effectively this ambitious goal. In an attempt to bridge the gap between the VPH community and the grid community, a group of experts from the STEP and SHARE consortia exchanged views on the relevance of grids for VPH. We present here the main conclusions and recommendations coming out of this reflection [2].

6.3.2. Research Requirements

The vast scope and integrative approach of the VPH project can only successfully be addressed using the resource sharing mechanisms provided by a grid infrastructure. However, analysis of the present situation shows there are barriers to such deployment. We propose to overcome this situation by deploying on the existing infrastructures some grid services that could be of extreme usefulness for the VPH community; this should attract VPH researchers to the large scale infrastructures, and should help grid developers to become more aware of the special needs of this emerging scientific community. The collaboration between the VPH and grid communities should enlarge the computing and storage resources as well as the services made available to the VPH community and foster the identification of new scientific research areas to which a grid environment can be appropriate.

6.3.2.1. Requirements specific to grid computing

It is essential to take an approach that integrates all resources beyond the desktop into a cohesive infrastructure, accessible by all VPH researchers as necessary. This means allowing researchers access to resources in a uniform manner, from their local departmental cluster to the biggest HPC machines available on a national or EU basis, and including everything in between. Taking this approach will mean that researchers who currently have no wish to access resources beyond their local cluster have the least painful migratory path, if and when they decide they need to access more powerful resources provided by a grid.

The multiscale nature of the VPH project demands that access to such resources be provided in as seamless a way as possible, and where appropriate, mechanisms be developed to allow the automatic migrating of simulations between different scales (and by implication, between resources appropriate to run the simulation at a particular scale).

6.3.2.2. Requirements specific to grid data and knowledge management

In many grid contexts the data are transient in nature; they are produced by the simulation runs, but after being analysed, can be stored off line or even trashed. Persistent data collections must be provided to the VPH community. Large Scale Infrastructures should make available storage services designed to ease the upload and download of large binary objects, and their replication computationally near by the execution nodes.

The management of storage and execution resources should be designed to have inherent security and knowledge management features. Security is vital because of the sensitive nature of the clinical or genetic data VPH sometime involves. We need specific research projects aimed to develop security models that while mostly transparent to the users, ensure that the data are managed according to the laws and to the ethical principles. The accumulative nature of VPH imposes that everything is organised under solid knowledge management models, which make possible to keep organised and usable even very large information spaces.

6.3.2.3. Requirements relevant to grid technology adoption and application deployment

The VPH community (which is heterogeneous collection of academic communities, linked only by the interest for an integrative approach to biomedical research) largely ignores the large scale infrastructures, avoiding deploying large scale collections, and excluding the use of massive computational resources as an opportunity to solve some of its modelling problems. On the other hand, with a few notable exceptions, the High Performance Computing (HPC) infrastructures and the other grid stakeholders are so far failing to provide the services needed to handle the VPH community computing needs. What is required is the encouragement of cross community interaction, in order to build meaningful dialogue between grid developers and VPH researchers. Providing higher-level tools that allow VPH researchers to interact with the resources that they need to achieve their scientific objectives in a uniform manner, abstracting where necessary the underlying difficulties of dealing with grid middleware, will help engage with researchers who previously found that the grid was of no relevance to them.

To foster grid adoption in the VPH community, it is highly recommended to identify a few VPH CPU intensive applications which would benefit immediately of the existing grid infrastructures like EGEE or DEISA.. The deployment of these applications will allow identifying the missing services on the existing infrastructures and will rise up the grid awareness in the VPH community.

6.3.2.4. Other requirements

The VPH roadmap [1] identifies a number of IT developments needed to address the scientific challenges of the EuroPhysiome initiative. Although not specific to grids, these developments should be integrated and/or transparently accessible on a healthgrid:

- databases or repositories of existing models
- frameworks for model communication
- knowledge management software / database
- visualisation tools

7. ROADMAPS

7.1. ANALYSIS OF USER COMMUNITIES REQUIREMENTS

The analysis of the user community requirements documented in the previous section show very clear patterns:

- Knowledge management is what researchers need. Computing and data storage resources are not sufficient although it is expected they can be accessed in a transparent and ubiquitous way.
- although the existing grid infrastructures do not provide all the services needed by the user communities, they already allow to perform a number of tasks of scientific relevance. As a consequence, deployment of scientific applications should be started as soon as possible in order to foster grid adoption and to clearly identify the existing gaps
- the technology complexity must be hidden from the users. Grids are perceived as potential infrastructures in so much as their use does not require adaptation or acquisition of skills
- the communities expressed the needs for developing the technology for distributed data management while the usage of grids for distributed computing is perceived as available but still very complex.

In the rest of this section, we have attempted to translate the requirements of the three communities: Europhysiome or VPH, Epidemiology or EPI, and Innovative Medecine or IMI, into a number of research challenges and deployment milestones:

- the research challenges are technical issues which need to be addressed in order for grids to offer services needed by the communities
- the deployment milestones are applications that should be deployed on grids in order to demonstrate their relevance, to identify existing limitations and to quantify the progresses made

The research challenges have been classified according to their relevance to computing, data and knowledge grids. We have also identified a number of them which are not specific to grids but which are needed for the deployment of knowledge grids like for instance the definition of agreed standards and ontologies in the research communities.

7.1.1. Research challenges for computing grids

Table 1 lists the Research Challenges identified from the requirements expressed by the research communities for Computing Grids (RCCG). They focus mainly on user friendliness, interoperability, quality of service and on demand access:

- user friendliness (RCCG4) is needed in order for the communities to use the grids without having to learn complex procedures. To make the grid user friendly, its operating system must be fault tolerant (RCCG6). The complexity should be hidden to the point the use of grids become transparent (RCCG5, RCCG2).
- The need to access resources on clusters and supercomputers raises the need for interoperability between grid infrastructures (RCCG2). The transfer of jobs between infrastructures should also be made transparent to the user to ease his work (RCCG3).

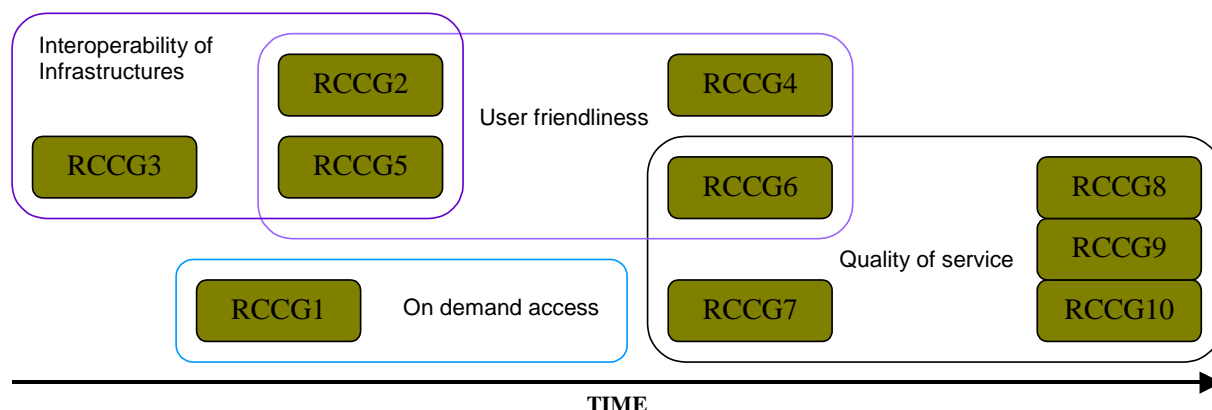
- The quality of service is particularly critical for biomedical applications in relation to healthcare (RCCG8). This includes the need for a scalable job scheduling system (RCCG9), the availability of a robust middleware easy to install in health environments (RCCG7) as well as resources with low latencies and high performance (RCCG10).
- On demand access to the resources (RCCG1) raises both technical and political as well as financial issues as to who pays for operating the infrastructures.

Research challenge name	Community expressing the requirement	Description of the requirement
RCCG1	VPH	Capacity to access grid resources on demand, without previous agreement or request. European grid infrastructures should be freely accessible to European projects
RCCG2	VPH	Capacity to submit jobs to cluster and supercomputer grids in a transparent way. Easy transfer of tasks between grid infrastructures
RCCG3	VPH	Automatic migration of simulations between different scales
RCCG4	VPH	User friendly access. Lower barrier to adoption.
RCCG5	VPH	Transparent access. The users should ignore they are using one grid or the other
RCCG6	EPI	Need for real fault-tolerant scheduling systems
RCCG7	EPI	Need for a grid middleware that can be installed in health environments seamlessly and without requiring exhaustive maintenance and administration.
RCCG8	EPI	Need for services in the infrastructures to define exploitation models and guarantee a Quality of Service. Need to consolidate the booking of resources in advance and to guarantee a pre-negotiated Quality of Service.
RCCG9	EPI	Need for scalable job scheduling system
RCCG10	EPI	Integration of resources with low latencies and high performance.

Table 1 Research challenges for a computing grid

The four key words we will keep to characterize the research challenges for computing grids are user friendliness, interoperability of infrastructures, quality of service and on demand access.

The timeline for these milestones, grouped by key words, can be seen in the diagram below:



Although largely arranged by level of complexity, certain milestones are prerequisites for others. For example, for true transparent access to multiple grids and infrastructures (RCCG5) and similarly the transfer of tasks between infrastructures (RCCG2), it will first be necessary to enable on demand access to grid infrastructures without prior agreement (RCCG1).

It should be noted that work towards achieving these milestones is expected to be done in parallel. Although this ideal timeline reflects dependencies (to some extent), it is also the case that the demand for various developments arises from different quarters with plans and development programmes working towards their achievement at different stages of progress. In any case, we observe that computing grids are at a more advanced stage in their development in general, so that work in progress here may fairly be expected to support and facilitate progress in data grids and, as they emerge, knowledge grids.

7.1.2. Research challenges for data grids

Table 2 presents the Research Challenges for a Data Grid (RCDG). Some of these challenges are common to computing grids like the need for quality of service (RCDG3), including the availability of a robust middleware easy to install in health environments (RCDG1).

Some challenges are related to basic data management services which are still to be developed such as scalable data cataloguing and data transfer (RCDG4) as well as upload and download of large binary objects (RCDG5). Further developments include services to provide security in the management of the medical data (RCDG2) related to the adoption of standards (RCDG7).

The need for distributed data models (RCDG2, RCDG6) is also expressed.

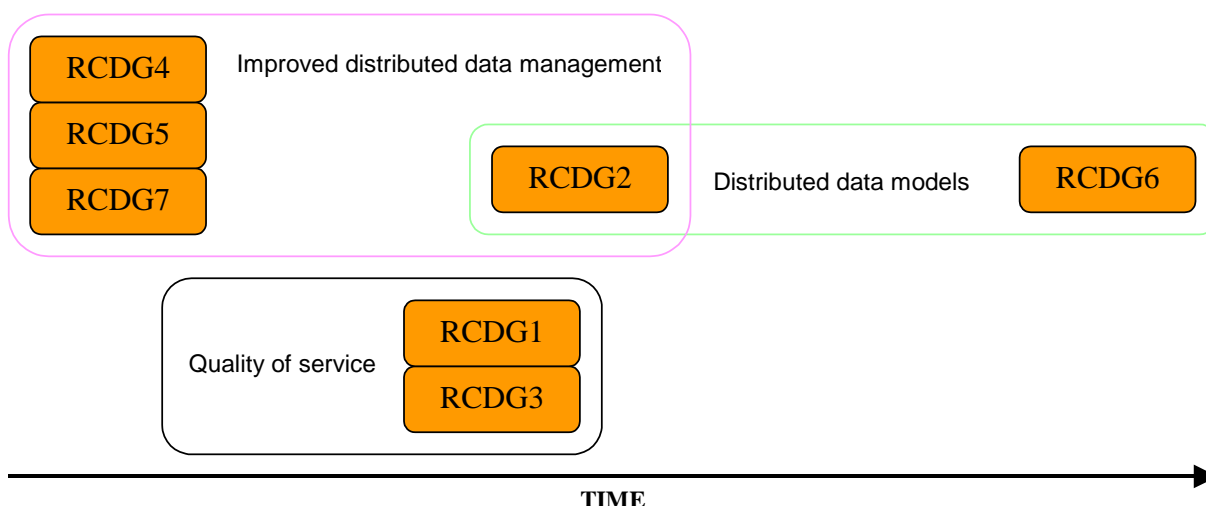
The key words we will keep to characterize the research challenges for data grids are improved distributed data management, quality of service and distributed data models.

Research challenge name	Community expressing the requirement	Description of the requirement
RCDG1	EPI	Need for a grid middleware that can be installed in health environments seamlessly and without requiring exhaustive maintenance and administration.
RCDG2	EPI - VPH	Need for data architectures and tools that implement private data dissociation, pseudo-anonymisation and encryption, and that are able to fulfil the legal requirements in the matter of data management.

RCDG3	EPI	Need for services in the infrastructures to define exploitation models and guarantee a Quality of Service. Need to consolidate the booking of resources in advance and to guarantee a pre-negotiated Quality of Service.
RCDG4	EPI	Need for scalable data cataloguing and data transfer.
RCDG5	VPH	Need for storage services designed to ease the upload and download of large binary objects
RCDG6	VPH	Need for distributed data models and repositories adapted to the multiscale nature of the data needed and generated by the Europhysiome community
RCDG7	IMI	Develop enhanced standards for data protection in a web services environment

Table 2 Research challenges for a data grid

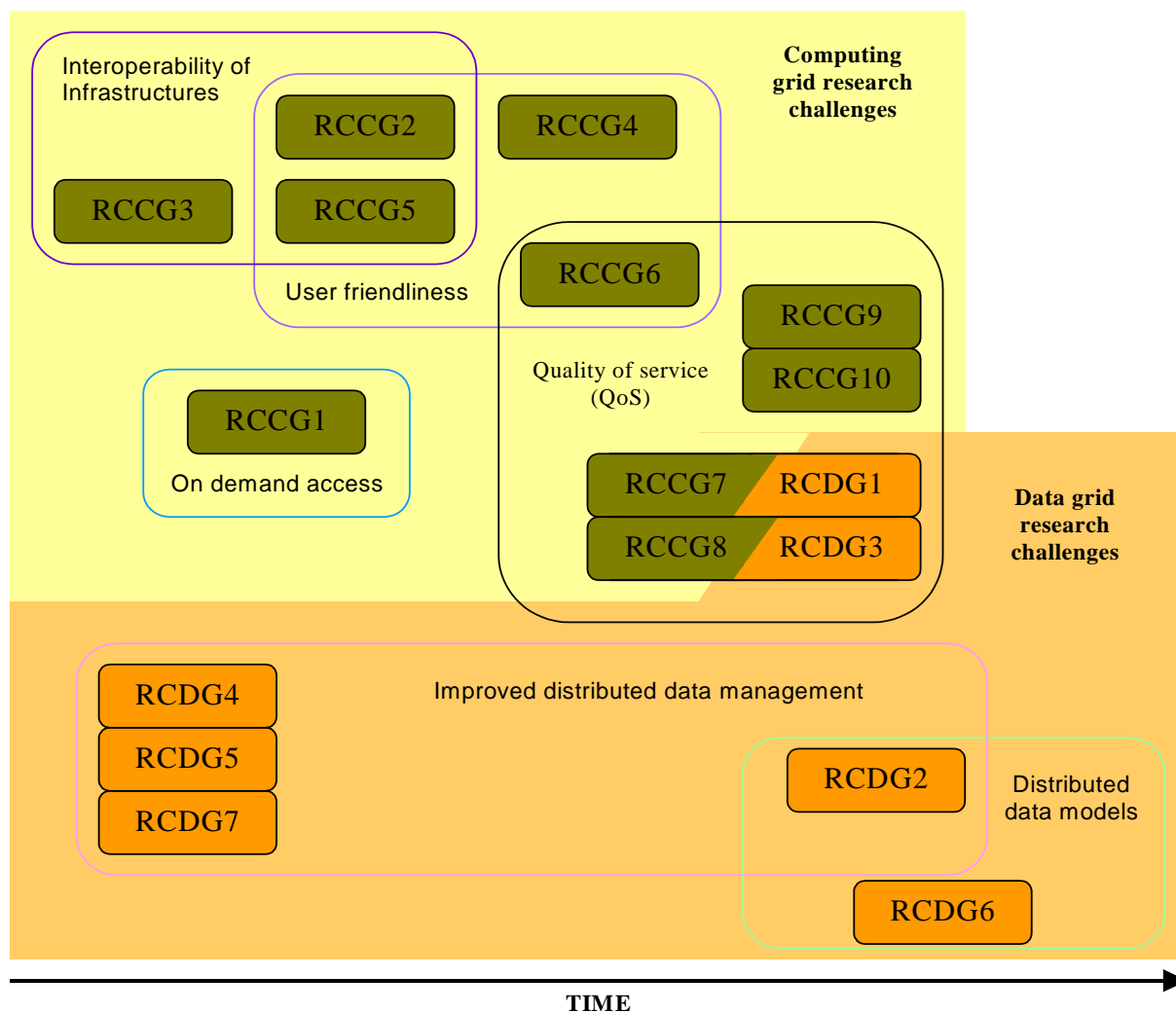
The timeline for these milestones, grouped by key words, can be seen in the diagram below:



It should be noted that quality of service (QoS), a key word for computing grids, is also an important area for data grids. The milestones RCDG7 / RCDG1 and RCDG8 / RCDG3 respectively are very similar, although there will be differences in the specific requirements for QoS between computing and data grids.

Naturally, there is a significant emphasis on the handling of data. Most questions will have already occurred in some guise or other in the field of distributed databases, but they reappear here with force in view of the autonomy of nodes within virtual organizations and especially the critical control that (non-virtual) organizations in the healthcare and biomedical domains must exercise over their data.

As noted above, developments in computing grids will support some of the work still necessary in the development of data grids. The following diagram illustrates the overlap between computing grid and data grid milestones:



As noted above, developments in computing grids are anticipated to support the evolution of data grids, although there is no simple correspondence between the different concerns and drivers in the two paradigms. Indeed, it is important to observe that the principal concern of data grids, the management of its transparently distributed data, may be addressed in parallel with the majority of issues in computing grids. Indeed, this is happening in several quarters in some cases independently of computing grid research and elsewhere in relation to it. Health-related projects dealing with imaging in particular, such as the EPSRC-funded Integrative Biology project and the EC-funded Health-e-Child project, have features common to both computing and data grids. These require significant data management facilities for distributed, possibly heterogeneous image data, associated annotations and metadata, but also require computational resources for biomedical modelling and simulations.

7.1.3. Research challenges for knowledge grids

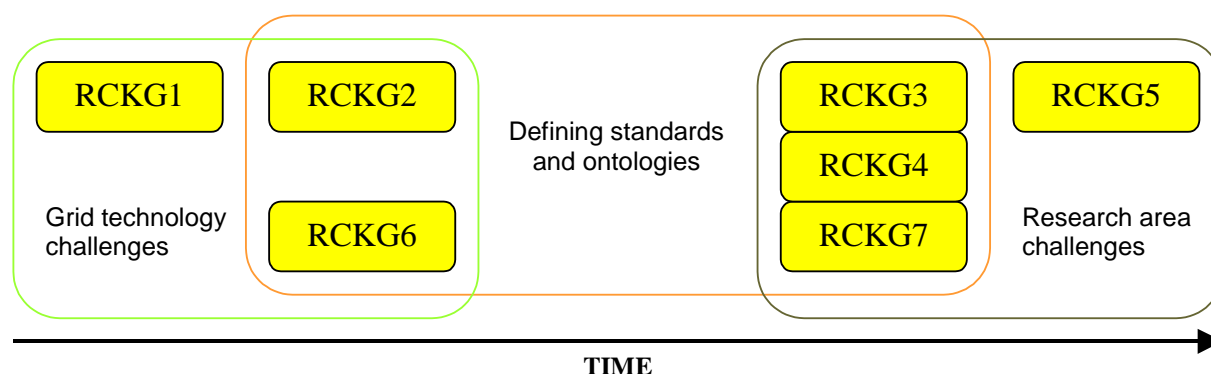
Table 3 provides a list of Research Challenges for Knowledge Grids (RCKG). These challenges refer repeatedly to data integration and knowledge management. Many of these challenges include the definition of standards and ontologies (RCKG2, RCKG3, RCKG4, RCKG6, RCKG7). Some challenges are directly related to the grid technology itself (RCKG1, RCKG2, RCKG6) while others are more relevant to the research area (RCKG3, RCKG4, RCKG5, RCKG7) and therefore not specific to the grid technology. In that case, it seems the healthgrid should benefit from the knowledge management services once they have been developed by the research community.

The key words we will keep to characterize the research challenges for a knowledge grid are data integration tools and standards as well as knowledge management tools and standards. In addition, we will use the concept of domain specific knowledge management tools and ontologies to characterize the developments which are not specific to grids but are needed to enable a knowledge grid.

Research challenge name	Community expressing the requirement	Description of the requirement
RCKG1	EPI	Need for knowledge-driven grid catalogues and integration based on the metadata.
RCKG2	IMI	Develop standards and models for exposing web services (semantics), scientific services, and the properties of data sources, data sets, scientific objects, and data elements
RCKG3	IMI	Develop enhanced knowledge representation models and data exchange standards for complex systems, presently largely inconsistent or incomplete, looking for synergies with other initiatives
RCKG4	IMI	Develop new, domain-specific ontologies, built on established theoretical foundations and taking into account current initiatives, existing standard data representation models, and reference ontologies
RCKG5	IMI	Develop advanced text mining tools for capturing implicit information about complex objects, relationships and processes, as described in patents and literature, beyond and above simple pair-wise relationships between entities
RCKG6	IMI	Design standards for and build an expert tool (ontology/schema/rules negotiator) for exposing the properties of local sources in a federated environment
RCKG7	IMI-VPH	Design standards for and build an expert tool (services/data negotiator) to guide users through the complexities of the data, data models, simulation and modelling tools, etc.

Table 3 Research challenges for a knowledge grid

The timeline for these milestones, grouped by areas, can be seen in the diagram below:



While work has been done towards many of these milestones, they remain significant challenges due to incomplete implementations and immature standards. Many of these challenges are as much in the Artificial Intelligence domain as in grid computing, with issues ranging from ‘knowledge-driven’ resource and service management to ontologies and meta-ontologies for medical knowledge.

7.1.4. Deployment milestones

Table 4 provides a list of deployment actions which were recommended by the research communities. These actions are perceived as milestones on the road to healthgrid adoption as their success will pave the way to the adoption of the technology.

Some actions are more geared towards computing grids (MD3) some are related to data grids (MD1, MD4, MD5) while others require from the beginning knowledge management (MD2, MD5)

These actions could be started on the existing grid infrastructures in view of the present state of the art of the grid technology. However, the quality of the services as well as their portfolio is expected to increase progressively with the evolution of the technology.

Deployment Milestone name	Community identifying the milestone	Description of the milestone
MD1	EPI	Need for successful pilots on epidemiology that will demonstrate the benefits of the technology.
MD2	EPI	Need for epidemiology data sources adapted to grid models and grid-enabled gateways to epidemiological data using medical informatics-related connectors, such as HL7, DICOM, ENV13606, etc.
MD3	VPH	Need for pilot applications in relation to VPH to foster adoption of grids in the community and to identify limitations of existing infrastructures
MD4	IMI	Build a core reference database of validated experimental and clinical research data extracted from the literature
MD5	IMI	Creation of disease-specific European Imaging Networks for establishment of standards, validation of imaging biomarkers and development of regional centres of excellence.

Table 4 Deployment milestones

7.2. PROPOSED ROADMAPS

In this section, we are going to present a technical roadmap for the adoption of the grid technology for healthcare. In the previous section, for three families of grids, computing, data and knowledge grids, we have identified a number of research challenges which have been characterized by a few key words.

Computing grids:

- user friendliness
- interoperability of infrastructures
- quality of service
- on demand access

Data grids:

- improved distributed data management
- quality of service
- distributed data models

Knowledge grids:

- data integration tools and standards
- knowledge management tools and standards
- domain specific knowledge management tools and ontologies

On the model of figure 1, figure 2 represents how research challenges address different layers of services from core infrastructure to applications. The following comments can be made from the picture:

- interoperability as well as improved distributed data management must be core functionalities of the infrastructure
- Quality of service is required from both core and healthgrid services for successful healthcare / biomedical applications
- Healthgrid services should be accessible on demand, in a user friendly way. Distributed data models need to be provided as well.
- Data Integration Tools and Standards are healthgrid services which stand at the interface between data and knowledge grids.
- Knowledge Management Tools and Standards require the availability of proper job and data management tools. They stand at the interface between generic healthgrid services and the application specific developments.
- Domain Specific Knowledge Management and Ontologies are under the responsibility of the research communities. Their interface to the knowledge grid is achieved using the Knowledge Management Tools and Standards.

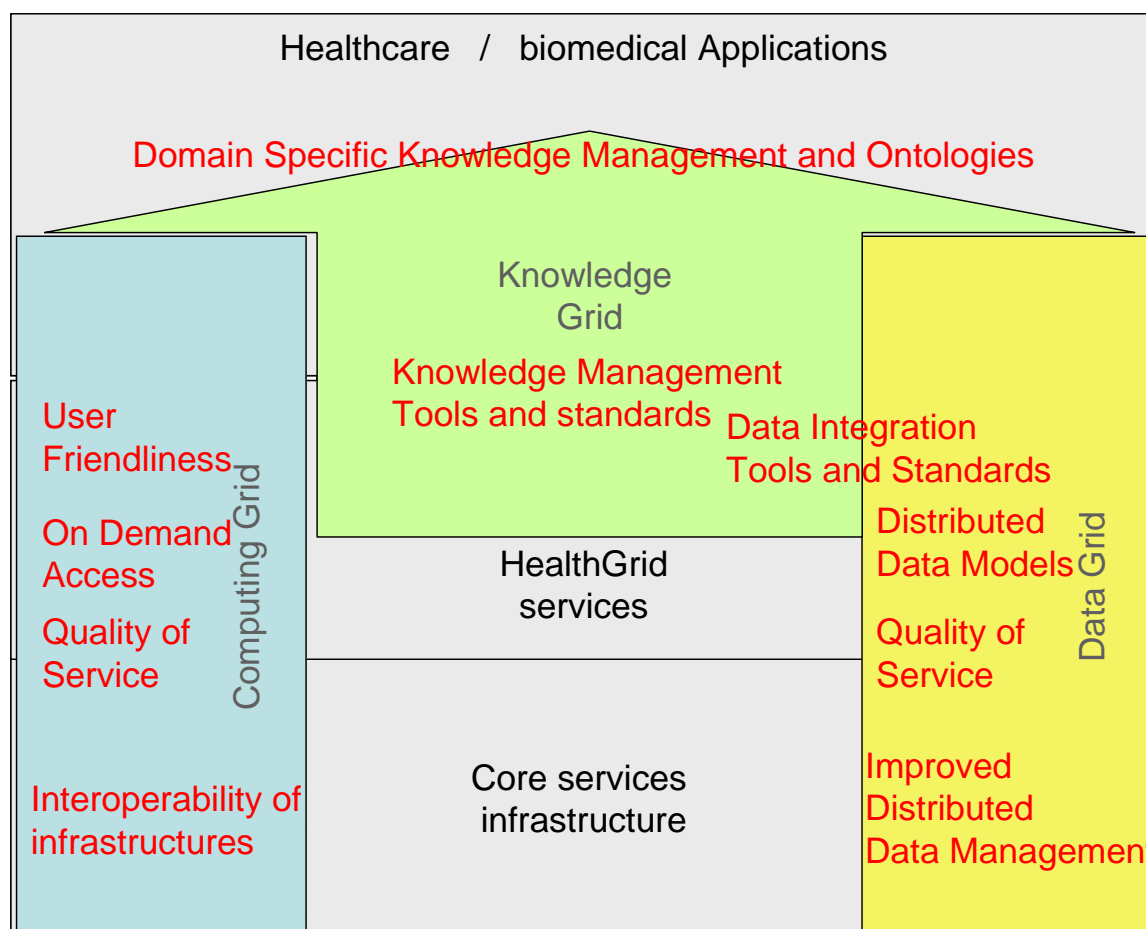


Figure 2: Representation of research challenges and healthgrid layers of services

On the basis of this analysis, we have represented on figure 3 the research challenges according to their complexity and an estimated time when they should be overcome. The figure inspired from SHARE deliverable D5.2a [3] also indicates the level of adoption by the research communities. As can be seen clearly from the picture, we identify two distinct roadmaps:

- research and development for computing grids should allow offering the quality of services needed for biomedical research and healthcare at a 5-year horizon
- data grids are expected to reach maturity at a 10-year horizon as the core technology is not yet mature.
- knowledge grids depend on the quality of services for distributed data integration and the capacity of the research communities to agree on standards and ontologies. As a consequence, their maturity is not expected before 15 years.

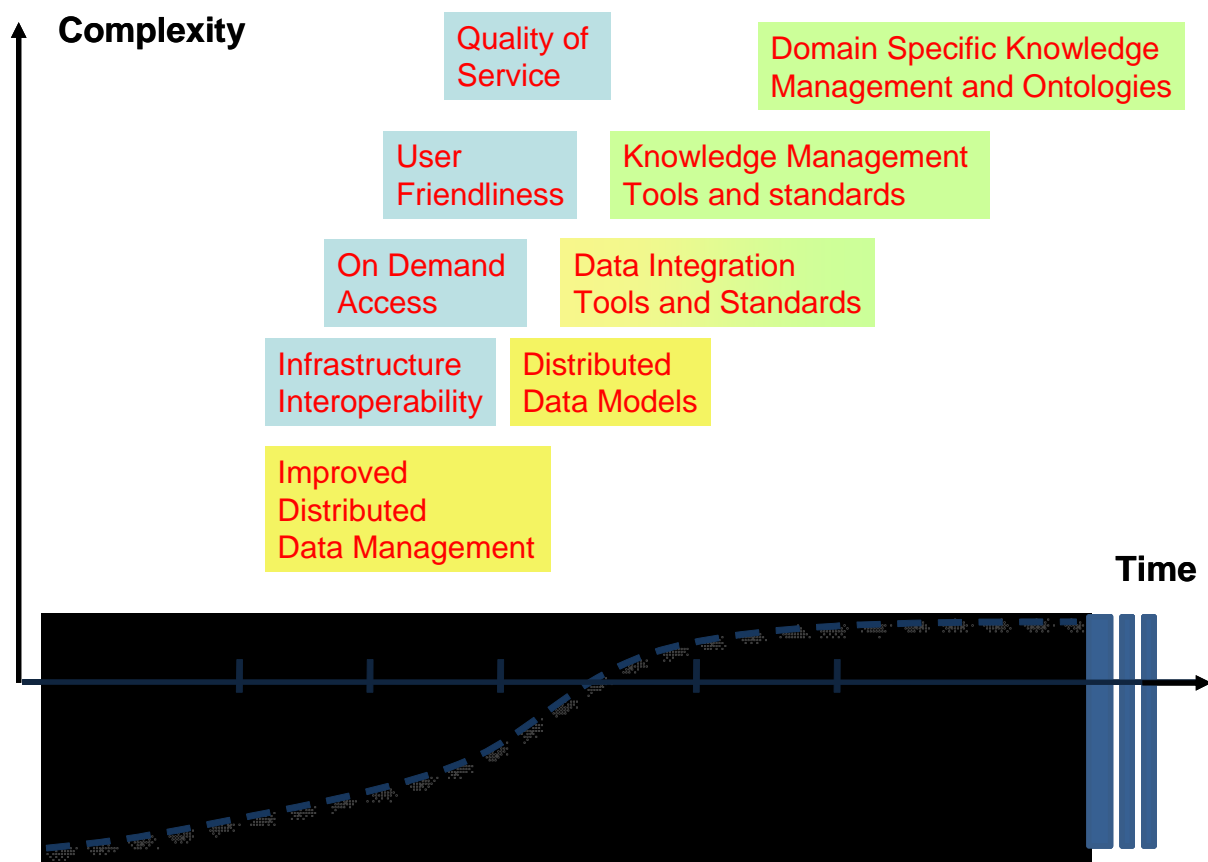


Figure 3: research challenges as a function of time and complexity

8. CONCRETE IMPLEMENTATIONS

It is extremely important that technical research and development be driven and constantly validated by user groups.

The three research communities involved in the definition of the roadmap expressed their interest and motivation for the deployment of prototypes and test cases on the existing grid infrastructures.

As a matter of fact, some projects are already using the DEISA and EGEE infrastructures for scientific production in the field of epidemiology, medical imaging and drug discovery. However, these initiatives come from pioneers and are not sufficient to achieve a wider adoption in the research communities. Moreover, they are some times perceived to fail to influence the evolution of the technology to make it better fit the community needs.

We recommend therefore the following concrete implementations:

- we recommend to identify one or two projects within the framework of the Europhysiome initiative that could directly benefit of the computing and data management resources provided by the EGEE and DEISA infrastructures. These projects would be deployed in parallel on the two infrastructures in order to investigate the interoperability issues and identify the bottlenecks.
- We recommend
- We recommend

The important message is that healthgrid applications can be deployed now. The infrastructures and the tools are available.

In the VPH roadmap, they list the following items:

- the infrastructure
- the data
- the models
- the validation
- long-term sustainability models
- the people

For HealthGrid, the following items could be stressed

- recommend concrete actions for each application area
- computing grids
- data grids
- knowledge grids