# SHARE

## INTEGRATED ROADMAP II

| | |
|---|---|
| Document ID: | **SHARE-D6.2** |
| Date: | **2008/08/05** |
| Authors: | **All Partners** |
| Activity: | **WP6: Integrated Road Map** |
| Document status: | **V5.0** |
| Document link: | **http://eu-share.org/about-share/deliverables-and-documents.html** |
| Confidentiality: | **PUBLIC** |
| Keywords: | **Healthgrid; Biomedical Informatics; Grid Computing – Research Challenges; Innovation and Technology Management; Ethical, Legal, Social and Economic Issues (ELSE); Epidemiology; Innovative Medicine; Genomic & Individualised Medicine; Security; Regulatory Compliance; Collaboration** |

**Abstract:** *The HealthGrid White Paper was published at the third annual HealthGrid conference in Oxford in 2005. Starting from the conclusions of the White Paper, the EU funded SHARE project (http://www.eu-share.org) has aimed at identifying the most important steps and significant milestones towards wide deployment and adoption of healthgrids in Europe. The project has defined a strategy to address the issues identified in the action plan for European e-Health (COM(2004).356) and has devised a roadmap for the major technological and ethical and legal developments and social and economic investments needed for successful take up of healthgrids in the next 10 years.*

*A "beta" version of the road map underwent full review by a panel of twentyone prominent European experts at a workshop in December 2007. An executive policy summary of the final draft road map was also presented for policymakers at the February 2008 meeting of the i2010 e-Health subgroup.*

*This final edition has sought to reconcile likely conflicts between technological developments and regulatory frameworks by bringing together the project's technical road map and a conceptual map of ethical and legal issues and of social and economic prospects. A key tool in this process was a collection of case studies of healthgrid applications. The document concludes with recommendations to address the various challenges identified.*

## CONTENT

# 1. THIS DOCUMENT

## 1.1. PURPOSE OF THIS DOCUMENT

The purpose of the document is to describe a roadmap towards the adoption of grid technology in biomedical research, healthcare and more generally in the life sciences.

This document is the final roadmap including technology, ethical, legal, social and economic issues prepared for review by the European Commission in April 2008.

## 1.2. RELEASE

This document will be made public after EC review.

## 1.3. DOCUMENT EVOLUTION

This is the final substantive deliverable from the SHARE project: an integrated road map to establish healthgrids in Europe as the infrastructure of choice for biomedical research in the first place and healthcare in the longer term.

Versions of this document have been formally presented at an Expert Workshop (see Appendix A.1 for membership and A.2 for the principal criticisms) in December 2007 and to policymakers at the last i2010 eHealth subgroup meeting in Brussels in February 2008.

We are grateful to the EC reviewers for insightful comments on the Review Edition of the Road Map and for the invitation to publish this as well as a compact version of the road map for policy makers. The latter will be termed D6.2a and will be available on the project web site.

## 1.4. DOCUMENT LOG

Major stages in the evolution of this document:

| Issue | Date | Comment | Author |
|---|---|---|---|
| 0.0 | 01/10/07 | Contributions from previous documents | I. Andoulsi; I. Blanquer; V. Breton ; A. Dobrev; C. Van Doosselaere; V. Hernandez; J. Herveg; N. Jacq; Y. Legré; M. Olive ; H. Rahmouni; T. Solomonides; K. Stroetmann; V. Stroetmann; P. Wilson |
| 1.0 | 15/11/07 | Expert Workshop draft | M. Olive, H. Rahmouni, T. Solomonides |
| 1.0 | 20/11/07 | Expert Workshop release | I. Andoulsi; I. Blanquer; V. Breton ; A. Dobrev; V. Hernandez; J. Herveg; N. Jacq; Y. Legré; M. Olive ; H. Rahmouni; T. Solomonides; K. Stroetmann; V. Stroetmann; P. Wilson |
| 1.1 | 01/02/08 | Draft for i2010 | M. Olive, H. Rahmouni, T. Solomonides |
| 1.1 | 15/02/08 | i2010 finalized | Y. Legré |
| 1.2 | 29/02/08 | Ready for final editing | M. Olive |
| 2.0 | 31/03/08 | Final draft | T. Solomonides, M. Olive, H. Rahmouni |
| 2.5 | 12/04/08 | Review final draft | Y. Legré |
| 3.0 | 15/04/08 | Release copy | T. Solomonides |
| 4.0 | 05/08/08 | Public edition | T. Solomonides, M. Olive, V. Breton, I. Andoulsi |
| 4.0 | 12/08/08 | Review | I.Blanquer, V.Hernandez, J.Herveg, Y. Legré |
| 5.0 | 15/08/08 | Public Edition (FINAL) | T. Solomonides, M. Olive, H. Rahmouni |

# 2. EXECUTIVE SUMMARY

## 2.1. SUMMARY

Grid computing ('the grid'[1]) is an exciting new technology promising to revolutionise many services already offered by the internet. This new paradigm offers rapid computation, large scale data storage and flexible collaboration by harnessing together the power of a large number of commodity computers or clusters of other basic machines. The grid was devised for use in scientific fields, such as particle physics and bioinformatics, in which large volumes of data, or very rapid processing, or both, are necessary. Unsurprisingly, the grid has also been used in a number of ambitious medical and healthcare applications. While these initial exemplars have been mainly restricted to the research domain, there is a great deal of interest in real world applications. However, there is some tension between the spirit of the grid paradigm and the requirements of medical or healthcare applications. The grid maximises its flexibility and minimises its overheads by requesting computations to be carried out at the most appropriate node in the network; it stores or replicates data at the most convenient storage node according to performance criteria. On the other hand, a hospital or other healthcare organisation is required to maintain control of its confidential patient data and to remain accountable for its use at all times. The very basis of grid computing therefore appears to threaten certain inviolable principles, from the confidentiality of medical data through the accountability of healthcare professionals to the precise attribution of 'duty of care'.

Despite these hurdles, pioneer projects have not been discouraged from exploring and demonstrating the potential impact and relevance of grids to such outstanding healthcare issues as early diagnosis of breast cancer or improved radiotherapy treatment planning. Grids are expected to bring a significant added value in the development of individualised medicine which requires the exploitation of biological and medical data, but this is still a research field. A 'healthgrid' is an innovative use of this emerging information technology to support broad access to rapid, cost-effective and high quality healthcare. Many areas of biomedical research and healthcare provision are expected to benefit from healthgrid technology, including medical imaging and image processing; modelling the human body for understanding, for surgery and therapy planning; pharmaceutical research and development, including specialisation of drugs to individuals; epidemiological studies; and genomic research and personalised treatment development.

Grid technology has been identified as one of the key technologies to enable and support the 'European Research Area'. The impact of this new paradigm is expected to reach far beyond eScience, to eBusiness, eGovernment and eHealth. However, a major challenge is to take the technology out of the laboratory to the citizen. The concept of healthgrids[2], i.e. grids for healthcare and biomedical research, was first developed in Europe in 2002 and has been carried forward through the *HealthGrid* initiative [1]. This European collaboration has edited a white paper setting out for senior decision makers the concept, benefits and opportunities offered by applying newly emerging grid technologies in a number of different applications in healthcare [2]. Starting from the conclusions of the White Paper, the EU funded SHARE project [3] aimed at identifying the important milestones towards wide deployment and adoption of healthgrids in Europe. The project has devised a strategy to address the issues identified in the action plan for a European e-Health [4] and has devised a roadmap for the major technological developments, legal and ethical barriers, and socio-economic investments needed for successful uptake of healthgrids in the next ten years.

[1] As a reference to a uniquely identifiable entity, 'the grid', unlike 'the Internet', is a misnomer; however, it is a convenient means of referring to the concept and the associated technologies.

[2] The term 'eHealth' was already in use with a broader meaning, hence the further neologism.

The roadmap proposed by the SHARE project expresses certain measurable goals and objectives for the *HealthGrid* community, provides an analysis of the technical gaps to be bridged in order to achieve a number of staged technical objectives, explores the ethical, legal, social and economic (ELSE) conditions of such developments, analysing the extent to which technology and its environment will need to be reconciled, and articulates a strategy for the concurrent achievement of these goals and objectives subject to realistic contextual conditions.

This roadmap has been developed from three major inputs:

1. an analysis of user requirements in a carefully triangulated set of domains through current projects and scripted use-cases;
2. a technical road map which sets out the key objectives for a viable 'knowledge healthgrid' to be achieved in a span of 10-15 years; and
3. a conceptual map of ELSE conditions, constraints and requirements which must be addressed before a knowledge healthgrid can be deployed in a real healthcare setting.

The conceptual map of ethical, legal, social and economic issues considered the regulatory challenges that any real healthgrid must meet:

- Legal challenges concerning rights to privacy and confidentiality, 'right to know' and duty of care.
- Ethical challenges concerning primary and secondary use of data whether individual or aggregated.
- Legal and ethical challenges concerning provenance and quality of information.
- Legal, ethical and economic challenges to the use of healthcare data in commercial and public research, including questions of ownership of data.
- Legal and ethical challenges in the communication of genetic information and the resultant 'lateral leakage' of information.
- Legal and ethical challenges to the communication of medical data across borders.
- Social and legal challenges concerning the formal professional competencies of different healthcare actors.
- Business challenges to the adoption of new information, communication and knowledge technologies in healthcare environments.
- Legal, ethical and socio-economic challenges of 'exceptional cases', such as assisted reproduction, organ donation and transplantation.

The proposed road map brings all these concerns together into one strategic plan.

## 2.2. WHAT ARE HEALTHGRIDS?

The White Paper [2] defines the concept of a healthgrid as follows:

> *Healthgrids are grid infrastructures comprising applications, services or middleware components that deal with the specific problems arising in the processing of biomedical data. Resources in healthgrids are databases, computing power, medical expertise and even medical devices.*

A *healthgrid* is an environment in which data of medical interest can be stored and made easily available to different actors in the healthcare system, physicians, allied professions, healthcare centres, administrators and, of course, patients and citizens in general. Such an environment has to offer all appropriate guarantees in terms of data protection, respect for ethics and observance of regulations; it has to support the notion of 'duty of care' and may have to deal with 'freedom of information' issues. Working across member states, it may have to support negotiation and policy bridging.

Early grid projects, while encompassing potential applications to the life sciences, did not address the specificities of an e-infrastructure for health, such as the deployment of grid nodes in clinical centres and in healthcare administrations, the connection of individual physicians to the grid and the strict regulations ruling access to personal data. However, a community of researchers did emerge with an awareness of these issues and an interest in tackling them.

## 2.3. THE *HEALTHGRID* INITIATIVE

Pioneering projects in the application of grid technologies to the health area have been completed over the past few years, and the technology to address high level requirements in a grid environment has been under development and making good progress. Because these projects had a finite lifetime and the healthgrid paradigm required a sustained effort over a much longer period, and besides because there was an obvious need for these projects to cross-fertilise, the '*HealthGrid* initiative', represented by the *HealthGrid* association (http://www.healthgrid.org), was initiated to provide the necessary long-term continuity. Its goal is to encourage and support collaboration between autonomous projects in such a way as to ensure that requirements really are met and that the wheel, so to speak, is not re-invented repeatedly at the expense of other necessary work.

The *HealthGrid* community identified a number of objectives [5]:

- Identification of potential business models for medical grid applications.
- Feedback to the broader grid development community on the requirements of the pilot applications deployed by the European projects.
- Development of a systematic picture of the broad and specific requirements of physicians and other health workers when interacting with grid applications.
- Dialogue with clinicians and those involved in biomedical research and grid development to determine potential pilots.
- Interaction with clinicians and researchers to gain feedback from the pilots.
- Interaction with all relevant parties concerning legal and ethical issues identified by the pilots.
- Dissemination to the wider biomedical community on the outcome of the pilots.
- Interaction and exchange of results with similar groups worldwide.
- The formulation and specification of potential new applications in conjunction with the end user communities.

Apart from research, where the value of grid computing is well established, a healthgrid may be deployed to support the full range of healthcare activities, from screening through diagnosis, treatment planning to epidemiology and public health. For example, anticipating that population trends, air pollution and global warming may lead, through extremes of heat, to increased risks for the elderly, a grid-based intelligent monitoring service may be deployed to track conditions and medical episodes in hot summers.

The results of several major studies of the interface between bioinformatics and medical informatics have been published with a remarkable promise of synergy between the two disciplines, leading to what had already begun to be referred to as 'personalised medicine' [6, 7, 8]. From the point of view of *HealthGrid*, this made clear the need to unify the field and to put its various elements in perspective: how would they – improved evidence bases, imaging, genetic information, pharmacology, epidemiology – fit together? What was their relative importance in the unfolding programme of work?

Given the source of the concept of grid in the physical sciences, many requirements arising out of the biomedical and healthcare fields were not a central concern to the grid development community. Indeed, even today, when these requirements have been fed through to the middleware services community, they are not always or necessarily a priority for the developers. Thus *HealthGrid* has been actively involved in the definition of requirements relevant to the development and deployment of grids for health and was among the first to identify the need for a specialist middleware layer, between the generic grid infrastructure and middleware and the biomedical or health applications.

Among data related requirements, the need for suitable access to biological and medical image data arose in several early projects, but for the most part these are present in other fields of application also. Looking to security requirements, most of these are special to the medical field: anonymous or private login to public and private databases; guaranteed privacy, including anonymization, pseudonymization and encryption as necessary; legal requirements, especially in relation to data protection, and dynamic negotiation of security and trust policies while applications remain live. Most administrative requirements are common to medicine and eScience, although the flexibility of ‘virtual grids’, i.e. the ability to define sub-grids with restrictions on data storage and data access and also on computing power, is more obviously required in healthcare. Medical applications also require access to small data subsets, like image slices and model geometry. At the (batch) job level, medical applications need an understanding of job failure and means to retrieve the situation.

## 2.4. HEALTHGRID WHITE PAPER

The European community working on applications of grid computing to health and biomedicine joined forces in *HealthGrid* and defined a vision of grids as the infrastructure of choice for biomedical research in the first place and healthcare delivery in the longer term. Nevertheless, adoption of grids for healthcare is still in its infancy. There are many reasons for this: an obvious first reason is that grid technology is still immature and is neither robust nor secure enough to offer the quality of service required for routine clinical use. Another important reason is that all grid infrastructure projects are deployed on national research and education networks which are both separate from the networks used by healthcare structures and very much less secure than they would need to be. Another potential obstacle is the legal framework in the EC member states which has to evolve to allow the transfer of medical data on a European healthgrid. It should also be borne in mind that grids, despite their virtual nature, still require human beings to make choices. Accordingly the economic and benefit case for the use of grids must be made and finally the real work environments and habits of people must be able to accommodate grid based working.

In all these areas, grid technology has the potential either to reduce significantly the cost or time to produce results and evidence, or even to provide resources that are able to deliver services that cannot be economically delivered using conventionally networked information systems. Moreover, the emergence of this technology opens up new possibilities for interdisciplinary research at the crossroads of medical informatics, bioinformatics and system biology to impact healthcare.

Along with many infrastructure projects, a growing number of grid applications are under development, with several completed and deployed in life sciences and medical research. Within the European Union and its member states, many applications have benefited and still benefit from substantial funding from the European Commission and some individual state funding bodies. Among the present projects, those relevant to health can be roughly classified into three categories:

- Infrastructure projects that aim to offer a stable distributed environment for scientific production. These infrastructures offer a generic multidisciplinary environment where biomedical applications can be deployed.
- Technology projects aim at developing new grid-enabled services and environments relevant to the needs of life sciences and healthcare.

- End-user projects that focus on specific life science or healthcare issues and integrate grid technologies wherever they appear relevant.

An extremely fruitful link should also be underlined here. At about the time the first European grid projects were being funded, a number of others ([6, 7, 8]) reported their results, demonstrating the benefits from convergence in our view of levels of biological – even biosocial – organisation (molecule, cell, tissue, organ, person, population, with pathogens also in the hierarchy), knowledge disciplines (from biochemistry through genetics, biology and pathology to epidemiology) and familiar informatic paradigms (bioinformatics, imaging, medical informatics, public health informatics). It was in the spirit of this work and work that had already begun to impact on policy, that *HealthGrid* commissioned the White Paper in 2004 and this was duly delivered in 2005.

## 2.5. THE SHARE PROJECT: FROM WHITE PAPER TO ROAD MAP

In the White Paper, the *HealthGrid* community expressed its commitment to engage with and support modern trends in medical practice, especially 'evidence-based medicine' as an integrative principle to be applied across the dimensions of individual through to public health, diagnosis through treatment to prevention, from molecules through cells, tissues and organs to individuals and populations. In order to do this, it had to address the question how to collect, organise, and distribute the 'evidence'; this might be 'gold standard' evidence, i.e. peer reviewed knowledge from published research, or it might be more tentative, yet to be confirmed knowledge from practice, and, in addition, would entail knowledge of the individual patient as a whole person. The community also had to address the issues of law, regulation and ethics, and issues about crossing legal and cultural boundaries, finding ways to express these in terms that translate to technology – security, trust, encryption, pseudonymization. Then it had to consider how the services of the healthgrid middleware would satisfy these requirements; and, if it was to succeed in the real world, how to make the business case for healthgrid to hard-pressed health services across Europe while they are struggling with their own modernisation programmes.

The vision of health that informs the thinking of the White Paper and the work of HealthGrid since its publication has been defined in the 'Action Plan for a European e-Health Area' [4] as follows:

> *"… the application of information and communication technologies across the whole range of functions that affect the health sector. e-Health tools or 'solutions' include products, systems and services that go beyond simply Internet-based applications. They include tools for both health authorities and professionals as well as personalised health systems for patients and citizens. Examples include health information networks, electronic health records, telemedicine services, personal wearable and portable communicable systems, health portals, and many other information and communication technology-based tools assisting prevention, diagnosis, treatment, health monitoring, and lifestyle management."*

In the light of the White Paper and its impact, the EC funded a 'specific support action' project, SHARE, to explore exactly what it would mean to realise the vision of the White Paper, investigate the issues that arise and define a roadmap for research and technology which would lead to wide deployment and adoption of healthgrids in the next ten years.

## 2.6. SHARE: AIMS AND OBJECTIVES

Based on the assumption that healthgrid should be the infrastructure of choice for biomedical and eHealth applications within the next ten years, the two objectives of the project have been:

- a roadmap for research and technology to allow a wide deployment and adoption of healthgrids both in the shorter term (3-5 years) and in the longer term (up to 10 years); and

- a complementary and integrated roadmap for e-Health research and technology development (RTD) policy relating to grid deployment, as a basis for improving coordination amongst funding bodies, health policy makers and leaders of grid initiatives, avoiding legislative barriers and other foreseeable obstacles.

SHARE has to define what has to be done, when – and in what sequence – by whom, and how? Thus the project must address the questions:

- *What research and development needs to be conducted now?* and
- *What are the right initiatives in eHealth RTD policy relating to grid deployment?*

— with all that implies in terms of coordination of strategy, programme funding and support for innovation. It turns out that action required in several domains: technical research and development; standards and security for real world deployment; squaring up to ethical and legal issues; and action to win community acceptance and economic investment. The conclusions of the project were presented for discussion at an Expert Workshop (see Appendices) and further insights gained there have been incorporated.

## 2.7. SUMMARY OF RECOMMENDATIONS[3]

### 2.7.1. Technical recommendations

#### 2.7.1.1. Communities

On the interaction of technical researchers and developers on one hand and users in the biomedical research and healthcare domains, we recommend

- promotion of cross community interaction, in order to build meaningful dialogue between grid developers and health researchers;
- joint development of prototype applications and test cases;
- broader availability of tools and infrastructures to those willing to experiment.

#### 2.7.1.2. Pioneer projects

Pioneers have already used existing grid infrastructures for scientific production in the fields of epidemiology, medical imaging and drug discovery. We recommend that:

- more attention be paid to such initiatives so that they may influence the evolution of the technology to make it better fit the communities' needs;
- two projects within the framework of the Europhysiome initiative be identified that could directly benefit from the computing and data management resources of the EGEE and DEISA infrastructures; these should be deployed in parallel on the two infrastructures in order to investigate interoperability issues and identify bottlenecks.

#### 2.7.1.3. e-Science approach

In terms of bringing biomedical applications closer to full exploitation of grids, we recommend:

- linking certain advanced health domains to an e-science infrastructure;
- adaptation of epidemiology data sources to grid models and grid-enabled gateways to epidemiological data, using medical informatics-related connectors, such as HL7, DICOM, ENV13606, or similar.

[3] References to particular projects, standards and standard-setting bodies are provided in the chapter on recommendations at the end of this document.

### 2.7.1.4. Bringing grids closer to the biomedical/healthcare communities

In the same spirit, in order to bring grids closer to the biomedical research and healthcare communities, we recommend:

- the release of open-source components for medical data interfacing;
- building a core reference database of validated experimental and clinical research data extracted from the literature of innovative medicine and to explore whether a grid infrastructure could support this activity;
- creation of disease-specific European imaging networks to aid in the establishment of standards, validation of imaging biomarkers, development of regional centres of excellence in innovative medicine and exploration of grid infrastructures to support such activity.

### 2.7.1.5. Security and standards

We recognise concerns about security and standards (at least) which predate the use of grids. Security is already receiving a good deal of attention, although not necessarily in the context of healthgrid specifically. In the field of standards, we consider that the *HealthGrid* initiative provides the right framework to coordinate the development of different standards in collaboration with the OGF and the various medical informatics standardisation bodies. We recommend:

- active pursuit of standards for the sharing of medical images and electronic health records on the grid within the already existing medical informatics standardisation bodies;
- active pursuit of ontology matching and development in the context of healthgrids.

### 2.7.2. Ethical Recommendations

- Guidelines should be established at EU level on methods for the appropriate balancing of key ethical duties of respect for autonomy, beneficence and justice in the development and use of health grids. To this end the European Group on Ethics in Science and New Technologies (EGE) should adopt a position on healthgrids.
- In order to facilitate the cross-border component of healthgrids, an EU-wide Healthgrids Ethics Committee system should be established which could monitor and harmonise local ethics committees' responses to healthgrid usage. To this end, EGE may consider the development of such a system.
- An EU wide education programme on ethics in healthgrids should be established, based on a network of ethics committees and academic researchers across the EU who could provide baseline materials to be adapted and adopted at local and national level.

### 2.7.3. Legal Recommendations

### 2.7.3.1. Liability

- A stepwise approach should be taken to develop the liability framework, distributing legal responsibility appropriately across healthgrid users and service providers.
- The European Commission should adopt policy tools encouraging the use of the RAPEX system for the testing of healthgrids.
- The European Commission should consider supporting the adoption of EU level guidelines that would identify the various parties involved in delivering healthgrid services and annex services and establish the various liabilities that each party must accept. Such guidelines should be widely disseminated in order to develop users' confidence in the use of healthgrids in general. In particular it should be investigated whether specific guidelines on those specific services could be drafted under the provisions for a Code of Conduct established in Directive 2000/31 on eCommerce.

### *2.7.3.2. Medical Devices*

- Special guidelines should be issued in order to clarify the application of medical devices legislation to specific tools used in healthgrids.

**Data Protection**

- Efforts should be made to harmonise these approaches across the EU so that meaningful cross-border work can support the health of all EU citizens.
- Efforts should be made to harmonise national standards on the technical and organisational provision for data security.
- Legal guidelines should be established in order to clarify the way in which professionals can make further use of personal data related to health in the interests of public health. Such guidelines should allow for secondary uses even where such uses could not have been foreseen at the time of data collection.
- The European Commission should co-ordinate the adoption of specific rules for the processing of health information to allow for proper balancing of patients' and public health interests, without relying on the concept of consent.
- A Directive or Code of Conduct on Privacy and Health Information Infrastructure should be developed within the context Directive 95/46/EC and could take the form of either a dedicated Directive or could be an EU-level Code of Conduct to be approved by the European Working Party on Data Protection set up under article 29 of the Directive. Any such Directive or Code would be complementary to Directive 95/46/EC on Data Protection and Directive 2002/58/EC on Privacy and Electronic Communications.

### *2.7.3.3. Intellectual Property*

- The EU Commission should develop guidelines for the use of open licensing and open standards, which could address the potential conflict between the intellectual property rights of developers and the needs of the grid technology.
- The EU commission should provide guidelines that would determine, in case of collaborative research, what research results every actor is entitled to exploit according to his contribution to the research

### *2.7.4. Social and Economic Recommendations*

We believe that technology transfer between EC projects should receive more prominent and active encouragement. In particular, we recommend:

- the Commission to implement collaboration measures in the funding mechanism for research projects in the fields of grid computing and healthgrid, healthcare, biomedical and health informatics;
- that the links between organisations across the EU and beyond that are established through specific technology-driven or otherwise motivated experimental and pilot projects be exploited and further developed;
- targeted capacity building so that projects may access grid resources on demand, without previous agreement or request; European grid infrastructures should be freely accessible to European projects;
- porting of one or two biomedical grid applications, already successfully deployed on grid infrastructures, to e-science environments using OGSA-compliant grid toolkits.

Member States and the EC have a role to play in initiating and regulating the development of grids, as grids have a significant public good component. Thus, we recommend the following parallel investigations to be carried out:

- explore the best model for Member States and the EC to support the development and maintenance of healthgrid networks, with specific attention to public-private partnership and other service contracts dealing with allocation of risk.

- analyse alternative resource allocation options both from a societal and an organisational perspective.
- consider flexible government regulated budgets and reimbursement schemes to encourage cross-organisational collaboration, including beyond national borders, through healthgrids.
- analyse the organisational changes that need to be implemented across workplaces sharing healthgrids, with specific attention to new workflows and shared management systems.
- require decision makers to account for risks when assessing potential changes in resource portfolios, e.g. concerning the fit between participation in healthgrids and overall organisational objectives and strategy, especially in relation to ethical and legal risks.

The potential economic benefits of healthgrids and how they might fit with current organisational and business practices should be analysed, taking into account:

- pilot projects and prototype applications, which are an inherent part of the technology roadmap, need to be future oriented in the sense that the ultimate routine operation users have to be persuaded both of their value and their applicability. If necessary, technical specifications and/or functionalities must be adapted. The goal should always be to give users, especially clinicians, tools that they would consider using with patients in real healthcare situation.
- ex-ante analyses over time, based on initial pilot experience, have to focus on ensuring acceptance, technical and regulatory certainty, and sufficient private incentives in the steps to follow.
- ex-ante analysis should also estimate potential net benefits (i.e. expected benefits less expected costs over time), accounting for different risks and for optimism bias in estimations.

# 3. THE BENEFITS OF HEATHGRIDS

## 3.1. GRID COMPUTING

Grid computing has been given a variety of definitions, with some emphasising the infrastructure: *"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"* [9] while others emphasise the computation *"distributed computing performed transparently across multiple administrative domains"* [10]. These 'virtualised resources' and 'multiple administrative domains' must be coordinated, often by means of what are known as 'virtual organisations' (VOs), a term intended to convey a sense of agility and scalability.

Grid technology was invented to help scientists and engineers process large volumes of data. Given the impressive growth of the internet and the ability of the world-wide web to deliver information and services, it was natural to ask whether the high volumes of data generated by modern science, from particle physics to bioinformatics, could be captured, treated, stored and transmitted using web technologies. It was also natural to ask if some configuration could be devised which would enable the still high cost of high performance computing to be shared between users who had frequent but not constant need for it. In the UK, Sir John Taylor, then Director General of the Research Councils, invoked the need for orchestration of experimental and analytical tools on one hand and of increasing scientific collaboration on the other to motivate the concept of 'e-science', a term which he coined to describe "global collaboration in key areas of science and the next generation of infrastructure that will enable it".

There are several analogical views of the concept: the screen-saver that does useful work when the resources of the processor are not fully utilised is familiar. However, only the scientist who created the screen-saver can download jobs to it. A grid offers a similar service, but to all its members: one subscribes to a grid, so to speak, thereby assigning use of some of one's resources to it and using freely those that have been provided by others. Another view is the source of the technology's name: the 'national grid' for electricity provides electric power for any device, be it a radio, a lawn mower or an air-conditioning system, that is compatible with it. A grid in our terms provides computational, data management and collaboration services to the user of any device that is compatible with it. Compatibility here entails more than simple 'pluggability', it implies compliance with the technological standards and with the rules of the virtual organisation in question.

From an information technology perspective, medical applications are some of the most demanding multi-media applications due to the high volume of data and high processing demands. The transfer of this data presents very high security demands, and very often hard synchronization and latency demands. Grid technologies support many of the specific requirements for medical data, including workflow, load balancing, service quality, and security policies.

Regarding resource management, grids are able to provide security, interoperability and resource sharing. Filters can be executed locally to transform data into a homogeneous format. Grids preserve local administration and have few requirements on the user side while being robust at the same time. They provide access to distributed and replicated databases, multiple computing resources and can reallocate tasks.

Grids are an ideal platform for standard parallel computing applications and an effective job queue management system. But their benefits do not stop at this point; a grid can provide services whenever a coherent and secure access to distributed data is needed or when data processing needs to be done at each end. If a program has to run many times with variation in several parameters or for large coarse-grain parallel applications, grids are also of great value.

## 3.2. FOR THE HEALTHCARE PROFESSIONAL/BIOMEDICAL RESEARCHER

Healthcare systems both in developed and in developing countries face major economic and capacity challenges to maintain quality of care in the face of the growing demands of ageing populations and the increasingly sophisticated treatments available. Add to this the desire to improve access to new care methods, and the challenge of delivering care becomes significant. In an attempt to meet these demands, health systems have increasingly looked at deploying information technology to scale resources, to reduce queues, to avoid errors and to provide modern treatments into remote communities, for example.

From the individualized care point of view, in order for clinicians to make the best diagnosis and decide on treatment all the relevant health information of the patient needs to be available and transparently accessible to them regardless of the location where it is stored. Moreover, computer-aided tools are now essential for interpreting patient-specific data in order to determine the most suitable therapy from the diagnosis. [11]

To store and process medical images, genetic information and other patient data, a large amount of computing power is needed. Large computing resources are also needed for keeping statistics of patient records, for knowledge extraction using data mining, and for the simulation of organisms and diseases using complex biomedical models. Grid technology has undoubtedly much to offer medical professionals, as illustrated by the following examples. [12]

However, the modernisation process faces significant challenges:

- Connecting and understanding patient records across organisation structures and even national borders.
- Ensuring that information is secured and those accessing it are authorised and authenticated.
- Discovering trustworthy sources of information for comparison.
- Handling a huge volume of data, especially that involved in genetic medicine for instance.
- Applying traditional information networks and technology into healthcare.

The delivery of medical information and certain services through the internet is familiar. In healthgrid computing, we seek an extension of the concept to consider how to provide large scale services to the user on demand. Some examples will serve to illustrate:

i. Consider a radiologist who needs to manipulate an image: we want to provide a set of services, some of which may require heavy processing, making them available on her desktop ‘transparently’, as if they were programs simply running on her computer.

ii. Consider a public health service which monitors certain infectious diseases and has to trigger an alert in case of a suspected epidemic. The identification of unusual patterns would in many cases be the critical step to halting the problem.

iii. Consider a surgical simulation prior to maxillofacial surgery, to determine how the patient’s face may appear after one manoeuvre versus another, the presence of sufficient tissue to allow the operation or to demand transplantation, and even to involve the patient in the decision.

iv. Consider a ‘neglected disease’ like malaria. Malaria is neglected by the pharmaceutical industry because there is no prospect of profit in it. Relatively little progress has been made towards the eradication of this well understood disease, notwithstanding substantial investments of public funds in research projects. *In silico* lead generation may possibly be coupled with investment in plant by the poorer nations that suffer from it to lead to a locally sustainable solution.

v. Consider the possibility of linking genomic information to imaging in diseases like juvenile idiopathic arthritis. The genome will indicate susceptibility long before the disease is expressed, but equally, signs picked up from imaging may obviate the need for genetic screening, thus avoiding some of the most acute problems associated with it.

vi. Consider more abstractly the nature of evidence-based practice, the volume of scientific literature that provides the evidence base and the accumulation of evidence from practice that occurs as a matter of routine healthcare. How can these be integrated? How can they be used without violating any ethical restrictions on use of data, confidentiality, privacy, security? How can they be shared without violating any data protection laws?

However, there are problems even among these optimistic scenarios. Standards have not stabilised in the grid world, so data exchanges will present problems straight away. Codes and coding languages are also still not universally adopted, while the application provider will wish to protect investments in software licence rights.

## 3.3. FOR THE TECHNOLOGIST

The SHARE roadmap for the adoption of healthgrids constitutes a critical analysis of the status of grid and other supporting technologies for the advance on the integration and processing of large scale eHealth and biomedical data.

From the technologist's point of view, this document outlines the deficiencies, gaps and promising technological research lines that are necessary for achieving a reasonable degree of maturity in healthgrids. Therefore, it can be seen as a list of opportunities for collaborations, new working lines and technology transfer.

In this sense, four technological areas are considered:

- Computing challenges. Issues related to the reliability, quality of service, lightweight middleware and compatibility on health networks require specific actions which are outlined.
- Data grid challenges. Issues related to data federation, effective update of databases, scalability and privacy management are considered and analysed.
- Collaboration grid challenges. Issues related to workflow definition, threading processes, 'playing' with data, adjusting images, consulting colleagues, comparing and contrasting, have been analysed at greater length in eScience projects than in healthgrids.
- Knowledge challenges. The evolution to future knowledge services through the semantic integration of services, semantic data analysis and federation are medium and long term issues which should be started now through basic research.

These technological challenges are summarised in several milestones that are described in the final section of the roadmap.

## 3.4. SOCIO-ECONOMIC BENEFITS

Modern healthcare services are expected to be available around the clock, seven days a week, so that systems with pervasive access and near-absolute fault tolerance are indispensable. However, it is difficult for these applications to run non-stop with a high quality of service. Grids could help by providing a platform of collaboration, allowing the linking centres which co-operate to achieve better continuity and quality of service. Medical staff will then be able to share experience, knowledge and 'second opinion' with other internal and external staff. The distributed architecture of grids with the availability of high-bandwidth networks responds well to the requirements of healthcare provision. There are also optimistic stakeholders' views towards medical research, healthcare and computing capabilities combined to better satisfy the patient [11].

Healthgrids promise many benefits to mobile patients as well as citizens. It could help a travelling individual to receive the right treatment in an emergency situation, thanks to the ability of the grid to facilitate communication between the local hospital of the patient and the admitting hospital abroad in order to exchange necessary heath related information.

In addition, healthgrids enable the mobility of a patient within EU states and allow them to receive medical treatment in a country of their choice. This could help solve problems of long waiting lists in states with busy hospitals and lack of medical staff. In economic terms, the grid could provide an optimal solution for healthcare. It allows a better use of resources and maintenance of tasks, an improved global IT organisation, scalable costs, and a large and consolidated IT business within the healthcare organisation [12].

Heath tourism is a growing concept which can enrich the economy of countries where modern medical treatments (plastic surgery, dental surgery, reproductive medicine, laser surgery for vision correction, etc.) are evolving and having higher success rates than others. This domain can benefit from healthgrid technology as it facilitates the exchange of patient heath records and the communication between foreign hospitals and heath insurance companies to facilitate the referral and payment process.

Transferring medical images for the purpose of a second opinion to another hospital requires high bandwidth connections between hospitals. Healthgrid technology can provide automated workflows that could be considered a better alternative to manual workflows, such as agreement over the phone and fax transmission of data. These manual workflows are still used by clinicians at present, but are labour-intensive and can cause errors [11].

## 3.5. ON USER PULL AND TECHNOLOGY PUSH: STAKEHOLDERS AND PRIORITIES

The distinction between ‘demand pull’ vs. ‘supply push’ for innovation is well known in the fields of technology and technology management. It is an expression of a broader duality which manifests itself in our road map proposals in at least four different forms:

### *3.5.1. Business domain*

Here we find the traditional contrast between users’ perceived requirements and technologists’ imagined solutions to problems (and sometimes solutions to imagined problems!).

User pull may come from the medical staff (including scientific and medical research) when they develop or wish to use a new technology in order to enhance the quality of care or to reduce its cost or to improve their research. This pull originates from within the medical and scientific communities and must first prove persuasive to a significant body of opinion in its community to achieve recognition as a valid tool. If it proves efficient, such a tool may well be integrated in a commercial product. However, this way of producing new medical tools is likely to be very slow.

Another source of ‘pull’ may be the patients wishing to benefit from the latest technology. They may try to get their practitioner or researcher to use the new technology and in some circumstances they will also try to use it directly. Experience suggests that patient demand of this nature does not easily focus the attention of medical staff on new technologies, except possibly when accompanied by a campaign to raise funds for it. On the other hand patients may precipitate the use of a new technology if they can access it directly without an intermediary. Of course, this raises questions of a public health nature, notably about damage through patients’ misuse of medical devices.

Both public and private health system management boards may also intervene in the introduction of new technologies in the medical and scientific research sectors. This would be rooted in their responsibility to organise the best or most efficient public health system, taking into account both healthcare quality control and funding. Depending on political decisions, they will boost directly the use of new technologies or will create a legal and material framework that allows the development of these technologies in a safe way, but they will not be able to originate the new technologies.

Technology-based industries are another obvious actor in the introduction of new technologies. By their very raison d'être they have a massive impact on their creation and their dissemination, but somehow the case of the healthgrid technologies is extraordinary. Healthgrid technologies are promoted by the computing community, both those who favour grids as a platform and those focussing in particular on the medical sector. Indeed, we may witness the creation of new category of medical staff: they are computer scientists who take part in the delivery of healthcare or at least in medical research projects. Their technological push is quite coherent with their new position in the medical world.

The emergence of healthgrid technologies has benefited from a technological push by the European Commission and national funding bodies which have given it a formidable political and financial boost. Already in possession of the technology, industries have eagerly invested in the exploitation of this new technology in healthcare in response and then tried to interest practitioners and patients in the merits of these new technologies. If healthgrid technologies deliver on their promise, it will be a great advance for healthcare. Nevertheless, there is considerable risk in promoting a new technology with relatively little direct support from the most concerned communities: practitioners and patients.

### 3.5.2. Ethical-legal domain

Here we encounter both the requirement that technological solutions comply with regulatory frameworks and the pressure for regulatory frameworks to evolve as the technology makes new things possible. Does permission or prohibition of an activity lead or follow 'the climate of opinion'? In some cases prohibition is said to have led to a change in the social climate; the most prominent case may be smoking which was not affected either by research findings or by heavy taxation, but has come to be regarded as unacceptable following legislation banning it.

On the other hand, in the case of genetic and embryo research, technical possibility has led to changes in the law – often in favour, though sometimes against as if to give time to social attitudes to catch up with scientific successes. An implicit objective of this project is to highlight areas in ethical regulation, data protection legislation and research uses of healthcare information in which a change in formal arrangements would be beneficial but would not entail abandoning any major underlying principles.

### 3.5.3. Scientific domain

Here we find the development of so-called 'data-driven' science in contradistinction to 'hypothesis-driven' science. Although there are competing philosophies of science, the ones that scientists by and large subscribe to hinge on the predictive power of a theory ("hypothetico-deductive") and on its susceptibility to being proved wrong ("falsificationism"). In other words, they place a higher value on predictive/explanatory moves than on descriptive/phenomenological ones. However, the sheer quantity of data emerging from certain fields of science, arguably coupled with the intense competition in modern scientific careers, has led to an experimental approach to the analysis of empirical data, whereby one "theory" after another may be tried to find one that fits the measurements. The degree of sophistication of such approaches (consider the discovery of patterns through data mining or the extraction of rules through machine learning techniques) has shifted the ground in the judgement of what constitutes valid science.

The extent to which this may also happen in the biomedical field is uncertain, although there are examples from traditional medicine (e.g. reliance on personal tools, often case databases) which may suggest that it would be better to embrace the trend and seek a constructive alignment with tradition rather than to reject it outright.

### 3.5.4. Medical domain

Genomic medicine in particular may be considered a manifestation of ‘push’ in healthcare. It seeks to be proactive on the basis of informational content, the patient’s genome. Knowledge of the patient’s genome may lead to more accurate differential diagnosis (leukaemia being a particular case in point) and better targeted drugs – or, more accurately correlated dosages (even for some common medicines), as happens with the anticoagulant Warfarin. For children who, perhaps for family reasons, are susceptible to certain juvenile conditions (joint idiopathic arthritis, cardiac dysmorphology, etc) a genomic map may provide a lifeline through early aggressive treatment of diseases whose symptoms emerge slowly and subtly, going unnoticed until the pain reaches an unacceptable level.

Although the popular image of ‘personalised medicine’ – drugs designed for each individual – is hopelessly wasteful and unrealistic, pharmaceutical providers will be able to market alternative formulations with advice on which is best to prescribe for which particular form of a condition or for any particular patient depending on their genome. This evidently benefits the patient; indeed, thinking in terms of the ethical precepts of beneficence and non-malfeasance, this makes a major contribution to minimizing the risk of harm.

Another aspect of ‘push’ in the medical domain is in the ‘just in time’ delivery of knowledge which has already been deployed in advanced healthcare systems [13]. This can be taken much further in a healthgrid through the development of knowledge discovery from current practice data alongside the knowledge management processes that are already feasible. This would be a significantly enabling technology, allowing variance from integrated care protocols to be recorded automatically in context and re-emphasising the primacy of the professional’s decision-making at the point care.

# 4. USE CASES, USER REQUIREMENTS AND CHALLENGES

## 4.1. PARADIGMS AND EXEMPLARS

Grids are often differentiated into computational, data and collaboration grids. The ideal grid, envisaged as a servant of a new paradigm of scientific research called 'e-science', must provide transparent processing power, storage capacity and communication channels for scientists who may from time to time join the grid, do some work and then leave, so that the alliances they form in their scientific endeavours might be described as 'virtual organisations' or VOs for short. Different sciences have different needs, and the grid concept has become differentiated: particle physics generates enormous amounts of data which must be quickly stored, but not necessarily instantly processed; on the other hand, data in bioinformatics is not large by comparison – it is, of course, in plain terms, large – but requires intensive processing. In extending the application of grid computing to e-health, another feature becomes pre-eminently necessary: that of collaboration.

An important consequence of the fluidity of collaboration in grid computing has been in the choice of 'architecture' for grid systems. 'Architecture' is used loosely in computer systems to describe the manner in which hardware and software have been assembled together to achieve a desired goal. Favoured also in the commercial application of the web, the so-called 'Service-Oriented Architecture' has been widely adopted in grid applications. In effect, it means that needed services – software applications – once constructed, are provided with a description in an agreed language and made available to be 'discovered' by other services that need them. A 'service economy' is thus created in which both *ad hoc* and systematic collaborations can take place.

Compared with data from physics or astronomy, medical data is less voluminous, but requires much more careful handling. Among the services it therefore calls for are 'fine grained' access control – e.g. through authorisation and authentication of users – and privacy protection through anonymisation or pseudonymisation of individual data or 'outlier' detection and disguise in statistical data. There are, of course, many more specialist medical services, as some of our examples reveal.

Most of these examples are taken from projects which have been or are currently deployed in Europe or beyond. As a consequence, they are rather focussed on biomedical research rather than healthcare as a number of legal and ethical issues still prevent the deployment of grid services for clinical routine. An interesting recent application to clinical trials has been demonstrated by Richard Sinnott [67]. This illustrates ways in which grid computing principles may be adapted to comply with ELSE issues.

## 4.2. COMPUTING GRID: AN EXAMPLE FROM INNOVATIVE MEDICINE

Drug discovery is the long term, multi-stage and high cost process by which drugs are discovered and/or designed.. Drug candidates are intermediate products of the drug development process. Drug Development manages preclinical safety studies and clinical phases. Registration and Delivery are the last steps of the full process. Reducing the research time in the discovery stage and having enhanced information about the leads are key priorities for pharmaceutical companies worldwide. Collaborations with academic laboratories and small biotechnology or pharmaceutical companies are crucial, mainly in exploratory research, then in the lead discovery stage and progressively less during the drug development phases.

The drug discovery goal is to find new molecules that bind with specific macromolecules known to play a key role in a disease process, in a manner that changes their function, either to increase resistance to or to reduce the virulence of some pathogen.

Recent progress in genomics, transcriptomics, proteomics, high throughput screening, combinatorial chemistry, molecular biology and pharmacogenomics has radically changed the traditional physiology-based approach to drug discovery where the organism is seen as a black box[4]. The approach is now to understand how disease is controlled at the molecular and physiological level and to target specific entities based on this knowledge.

*In silico* drug discovery is one of the most promising strategies to speed up the Drug Discovery process. It is important to understand and control the *in silico* process; this is described below.

Figure 1 shows the different phases of the drug discovery process with their approximate duration, their success rate and the corresponding *in silico* contributions.

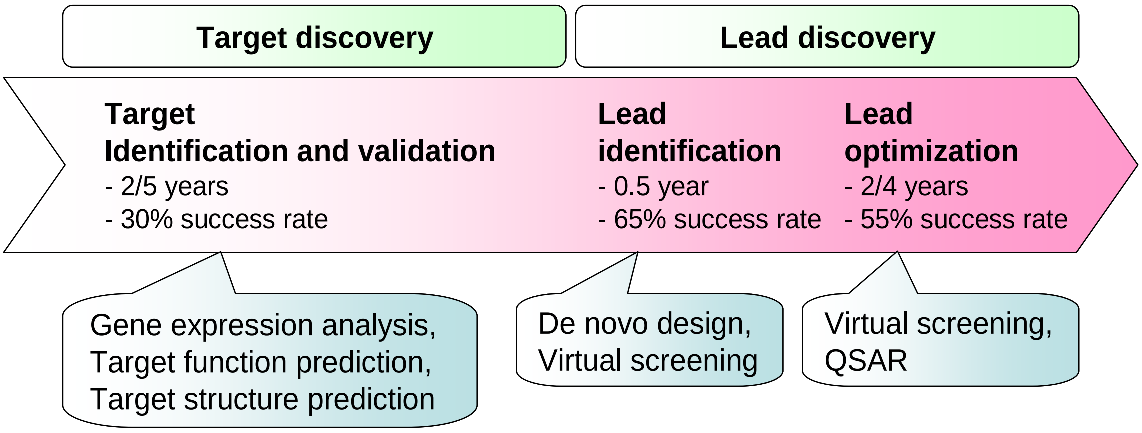


**Figure 1: Representation of the different phases of the drug discovery process with their duration, their success rate and the corresponding potential *in silico* contributions.**

A target is a cellular molecule which is believed to be associated with a desired change in the behaviour of a disease process and on which drugs usually act. Target identification and validation aims to isolate and select it. *In silico* drug discovery contributes to the target discovery by analysis of the gene expression data, target function prediction and target three-dimensional (3D) structure prediction.

A lead compound is a substance affecting the selected target. Two different *in silico* pipelines can be used for identifying it: the *de novo* design and virtual screening. *De novo* design builds iteratively a compound from the structure of a protein active site. Virtual screening selects *in silico* the best compound from a molecule database. These methods speed up the process and reduce costs avoiding time consuming and costly *in vitro* tests.

Lead optimisation addresses the development from the most promising lead compounds to a safe and effective drug. Instead of expensive and longer *in vitro* and *in vivo* tests, evaluation of the basic chemical properties can be achieved by virtual screening and using Quantitative Structure Activity Relationship (QSAR). QSAR can be used in a quantitative process correlating chemical structure with function in order to optimise pharmaceutical properties (Absorption, Distribution, Metabolism, Excretion and Toxicity) or efficacy against the target organism.

*In silico* drug discovery contributes to increasing biological system knowledge, to managing data in a collaboration space, to speeding up analysis and consequently improving the success rate compared with the traditional "wet" approach. The efficiency gains of such an integrated knowledge system

[4] This is pursued in greater detail in SHARE deliverable D5.2b The Innovative Medicine Case Study.

could result in 35% cost savings, or about US$300 million, and 15% time reduction, or two years of development time per drug.

Reducing the research time and cost in the discovery stage and enhancing information about the leads are key priorities for pharmaceutical companies worldwide. To achieve this goal, *in silico* drug discovery must meet the following requirements:

- **Data integration.** The *in silico* drug discovery process includes the management of a large variety and quantity of scientific data. For example: images, sequences, models, databases. Data integration is thus a challenge to increase knowledge discovery but also to ease the complex workflow. This implies data format standardisation, dataflow definition in a distributed system, infrastructure and software providers for data storage, services for data and meta-data registration, data manipulation and database updates, development and sharing of ontologies and knowledge representations. .
- **Workflow enactment**. The *in silico* drug discovery process also includes the management of a large variety and quantity of software. Software integration is another challenge to build efficient and complex workflows and to ease data management and data mining. Experts in different areas are absolutely necessary to maintain and update software and workflows, to propose new methods or pipelines, to use remote services, exploit outputs, and finally to propose compounds for assay. A software workflow will assist the scientist and the decision-maker in organising their work in a flexible manner, and in delivering the information and knowledge to the organisation.
- **Access to computing and data resources.** Deploying intensive computing is a challenge for *in silico* drug discovery. For instance, computing 1 million docking probabilities or modelling 1,000 compounds on one target protein requires on the order of a few TFlops for one day. Extensive computing resources are also needed to accurately describe protein structure models by computational methods based on all-atom physics-based force fields including implicit solution. Computing power is also required for bioinformatics resource centres where server access is saturated by the large number of short tasks requested by users.
- **Collaboration between public and private partners.** Joining the new Information Technologies with life science to enable *in silico* drug discovery requires strong remote collaboration between different public and private experts when addressing neglected and emerging infectious diseases. It also involves concrete sharing of resources: data and knowledge, software and workflow, and infrastructures such as computing, storage and networks. The collaboration space needs experts to maintain the resources. Having tools and data accessible to everyone in collaboration requires intuitive interfaces that need to evolve and be maintained. These interfaces reduce the development time of new methods. They also help the integration of data and software from *in silico* drug discovery but also from experimental processes. Of course, security is a key challenge for pharmaceutical industries but also for academic institutes in most cases. Effective protection of intellectual property and sensitive information requires, for instance, authentication of users from different institutions, mechanisms for management of user accounts and privileges and support for resource owners to implement and enforce access control policies.

## 4.3. DATA GRIDS: EXAMPLES FROM EPIDEMIOLOGY AND PUBLIC HEALTH INFORMATICS

### *4.3.1. Epidemiology*

One relevant example of health applications for data grids is epidemiology. The epidemiology use case is defined as a system able to link the information from distributed and heterogeneous databases, identify patients and complete episodes, improving quality automatically without interrupting clinical practice. Complex epidemiology models are fed with this data and produce by simulation and other methods, in a reliable way, aggregated prospective results. The analysis of this section focuses on this use case.

This use case is representative of different applications and systems, such as:

- oncological information systems;
- infectious surveillance networks;
- pharma-epidemiology analysis of efficiency and cost; and
- study of propagation models for diseases.

The main users (from the highest-concept level to lowest one) are public health authorities, epidemiologists and pharmaceutical companies. The data is normally owned by clinical care (both public and private).

The steps that the use case must go through (from the point of view of the user) are:

- **Automatic data gathering** Data from different, geographically distributed sources (primary care, prescription, demographic information, hospital information, microbiological data, etc.) must be put together.
- **Data quality improvement** Data is poorly coded and must be corrected, completed and improved. Aberrations, incoherent, incomplete or inaccurate fields must be revised and corrected.
- **Processing of the data** From simple aggregation analysis to complex data-mining techniques, epidemiological data is used for prospective and retrospective analysis.
- **Presentation of results** The analysis must provide, by the end, well-known indicators. Cancer survival rates, epidemic secure intervals and other measures are typically obtained by well-known and widely used computer programs that are fed upwards with the results of the analysis.

The requirements of the use case are:

- **Automatic data gathering** The data should automatically be made available in a comprehensive way. The user should not have to extract the data, adjust formats or even trigger the data collection procedure. At this level, neither the technologies nor the architectures (centralised versus virtual storage, for example) are relevant.
- **Enhancement of quality of data** The availability of different complementary sources must be sufficient to achieve this task. Knowledge management tools should make the linking of different registers to assist on the correction of the mistakes. It could be necessary, and assumed by the users, that the expert must (or at least could) intervene to validate the process.
- **Sufficient security management** The user should be provided with simple and effective measures that will guarantee that the privacy of the data and results are not compromised.
- **Compliance with ethical guidelines** The use of the system must not occur in violation of legal regulations, whether in matters of data protection or other. It must be valid for the requirements of the respective ethics committees of the research centres. The system should advise on the compulsory documents, agreements and steps that should have been performed.

- **Efficient performance of processing services** The complex analysis (and even more basic analysis) must be performed in a reasonable time. Users will expect a utility-like behaviour of the service, so it must be guaranteed that the process ends in a maximum time window.

- **Seamless integration of processing services** It will be difficult or even impossible to support all the post-processing tools available in epidemiology. Users normally download the processed data from the previous analysis and feed applications that compute the indicators, graphs and charts that the epidemiologist are used to. This should still be possible or even made easier.

- **Reliability and long-term exploitation** The system should be reliable, not only at the user level, but also at the different steps (data gathering, data quality, etc.). Epidemiology systems are kept for long periods of time, so pervasiveness of the services in the long-term is required.

### *4.3.2. Surveillance network for avian flu*

A use case in the field of public health informatics is a surveillance network of data repositories offering services to the research communities working on avian flu and firing an alert in case of pandemic risk. Indeed, the ability of the international community to respond efficiently to the possible emergence of a human-to-human transmissible avian influenza virus depends on its capacity to quickly assess any evolution of the disease. Many countries have set up very efficient national networks for collecting data and monitoring outbreaks. However, there is presently no international surveillance network that allows the sharing of data collected at a national level.

The starting point would be to set up a data grid collecting public and private information on avian flu. Public data would be made available to all registered users while access to private data would be strictly controlled through grid authorization and authentication mechanisms. The repositories would share a common model allowing distributed queries. The two main concerns are the evolution of the virus and its capacity for human to human transmission:

- Once a new virus stream has been sequenced, its comparison to the previously identified streams allows the evolution of the virus to be measured. Services for molecular epidemiology are therefore greatly needed.

- Genome sequences can also be used to guide the search for vaccines and drugs. For instance, the 3D structure of the enzymes of the new virus streams can be derived from the structures available in the protein data bank database by homology modelling. Grid resources can be used for large scale virtual screening of these mutated targets.

- Disease epidemiology requires measuring the evolution of the disease in time and space using GIS (geographic information system) and environment data as well as understanding the disease transmission, reservoir, immunity and treatment. Based on these data, mathematical modelling and computer based simulations allow the testing of outbreak hypotheses.

The quality of the surveillance depends on several parameters:

- The design and deployment of a robust federation of databases on multiple sites worldwide.

- The reliability, relevance, and completeness of the data stored. Reliability depends on the mechanism to collect data. Relevance depends on the mechanism to update data while completeness is influenced by the capacity to collect data from multiple scientific disciplines and multiple countries.

- The relevance of the services offered. It is not sufficient to have the best, most up-to-date information on the disease. The data must be properly integrated and its exploitation must build upon expert skills in epidemiology.

- The user friendliness of the environment. Scientists are so busy they will not take time to contribute information and operate services if these services are not easy to use.

Several types of data should be integrated to achieve disease monitoring:

- medical / epidemiological data on human and animal cases
- geographical data for each outbreak: outbreak location, number of cases, geographic distribution of the casualties, environment, and population density
- molecular biology data: virus genome sequences, philogenetic trees, and proteomics

Text mining services would be used to extract information from this literature on demand.

The grid technology to build such a monitoring network has been developed in recent years. Development of standards for interoperability allows a joint deployment across grid infrastructures all around the world.

Such a world-wide surveillance network would involve many stakeholders:

- international organizations like WHO or FAO
- national public health institutes and centres for disease control
- research laboratories for infectious diseases

These stakeholders have different roles as data providers or customers of the alert services provided by the network.

## 4.4. COLLABORATION GRIDS: EXAMPLES FROM BREAST CANCER SCREENING, PAEDIATRICS AND VPH

### 4.4.1. MammoGrid

#### 4.4.1.1. Breast cancer and screening

Breast cancer as a medical condition, and mammograms as images, exhibit many dimensions of variability across a population. Likewise, the way diagnostic systems are used and maintained by clinicians varies between imaging centres and breast screening programmes, as does the appearance of the mammograms generated. A distributed database that reflects the spread of pathologies across a broad population is an invaluable tool for the epidemiologist; understanding the variation in image acquisition protocols is essential to a radiologist in a screening programme. Exploiting emerging grid technology, the aim of the MammoGrid project [14] was to develop a potentially EU-wide prototype database of mammograms to be used to investigate a set of important healthcare applications and to explore the potential of the grid to support effective collaboration between healthcare professionals. In particular, the project aimed to prove that grid infrastructures can be practically used for collaborative medical image analysis. This led to several technical issues, including the standardisation of mammograms, design of an appropriate clinical workstation and distribution of data, images and clinician queries across a grid-based database while respecting patient confidentiality and security protocols. MammoGrid effectively demonstrated the viability of the grid by harnessing its power to enable radiologists from geographically dispersed hospitals to share standardised mammograms, to compare diagnoses (with and without computer aided detection of masses and microcalcifications) and to perform sophisticated epidemiological studies across national boundaries.

The statistics of breast cancer diagnosis and survival appear to be a powerful argument in favour of a universal screening programme. However, a number of issues of efficacy and cost effectiveness limit the scope of most screening programmes. The predominant method in breast cancer screening is mammography (breast X-ray). In younger women glandular tissue is dense and largely X-ray opaque, so that in women under 50 signs of malignancy are far more difficult to discern in mammograms than they are in post-menopausal women. Consequently, most screening programmes, including the UK's, only apply to women over 50. Breast X-rays are rather challenging, both to take and to interpret. In mammography, the breast is compressed between two Lucite compression plates partly to immobilise it and partly to displace as much fat to the margins as possible. Carcinomas of different types attenuate X-rays of typical energies by about 5% more than functioning glandular tissue (parenchyma) and by up to twice as much as fatty tissue. Thus fat against parenchyma or tumour contrast very well, but it is very much harder to draw a clear distinction between signs of carcinoma and functioning tissue.

While clinically significant signs are subtle, many parameters also affect the appearance of an image. For mammograms, these include image acquisition parameters, such as degree of breast compression, tube voltage and beam intensity, and anatomical and physiological data, which show marked variation across the population, at different times in the menstrual cycle and throughout the course of a woman's life. The way diagnostic imaging systems are used and maintained by clinicians also varies between imaging centres and breast screening programmes. In order to study the epidemiology of breast cancer, it is necessary to understand this variability. This is also a prerequisite for the integration of computer-aided detection (CADe) tools and quality control in the process.

Radiographers ('radiologic technicians' in the US) adhere to certain codes of professional practice, but are responsible for maintaining the X-ray equipment and have freedom to determine machine settings in the course of their work. This makes comparison of images as taken rather difficult. Occasionally this may be a problem for images of the same patient at different times, but it is rather more serious if comparability of images is to be used for diagnostic purposes or in a radiological training programme.

#### *4.4.1.2. The project*

MammoGrid developed a collaborative grid-based image analysis platform in which statistically significant sets of mammograms can be shared between clinicians across Europe. The applications implemented can be thought of as addressing three main problems:

- image variability, due to differences in acquisition processes and to differences in the software packages (and underlying algorithms) used in their processing;
- population variability, which causes regional differences affecting the various criteria used for the screening and treatment of breast cancer; and
- support for radiologists, in the form of tele-collaboration, second opinion, training, quality control of images and a growing evidence-base.

In practical terms, the project proceeded as follows:

- first, it evaluated current grid technologies and analysed the requirements for grid-compliance in an EU-wide mammography database;
- then, it implemented a prototype MammoGrid database, using novel grid-compliant and federated-database technologies to provide access to distributed data;
- it then deployed versions of a standardisation system (SMF – the Standard Mammogram Form™) that enables comparison of mammograms in terms of tissue properties independently of scanner settings, and to explore its place in the context of medical image standards (e.g. DICOM [15]); and
- then, used the annotated information and the images in the database to benchmark the performance of the prototype system.

The European dimension was highlighted through provision of statistically significant numbers of exemplars even for rare conditions of cancer development and thus enabling more diverse epidemiological studies than hitherto had been possible. The project has thus paved the way for potential knowledge discovery in the diagnosis and understanding of breast cancer through a growing database of case histories.

Given the limited built-in security of early grid infrastructures, MammoGrid operated within a framework of informed consent, anonymised data, and certification, authorisation and authentication processes. In addition, the development of an efficient information infrastructure demanded data with integrity, quality and consistency and the project met these requirements by developing standard data formats and strict automated quality checks.

### 4.4.2. Health-e-Child

The Health-e-Child project [16] deals with Paediatrics, an area where not only the disease in question is changing, but the child as well as he or she grows. The project addresses specific diseases within certain medical domains, and through consultation with these communities a number of requirements for the project were identified.

#### 4.4.2.1. Cardiology

The role of a cardiologist as defined by the project is to decide on the best medical or surgical treatment for a patient, and determine a patient’s follow up schedule. They perform ultrasound examinations, and annotate heart images. Within the project, the focus has been on right ventricular overload and cardiomyopathy.

There can be many causes of right ventricular overload, such as atrial septal defects, anomalous pulmonary venous return, and tetralogy of Fallot. Many can have genetic causes, which can be considerably complex. Similarly, both dilated cardiomyopathy and hypertrophic cardiomyopathy have genetic links.

Cardiologists must provide second opinion, and the project argues that this could be aided by gathering data from similar cases, accompanied by the decisions taken in those cases. An evidence base of past cases would be built up, aiding cardiologists in their diagnoses and also serving as a training resource.

Regarding annotation, region of interest (RoI) measurements are often subjective, e.g. hand-drawn on images of the heart. A semi-automatic method of feature discovery could save the cardiologist time, and support junior clinicians.

Cardiologists must attempt to ascertain how severe a right-ventricular overload might be, and when surgery should be performed. Advanced right-ventricular size measurement tools have been developed which could be of considerable help, as could a prediction algorithm based on previous cases.

The primary requirements from cardiology are therefore related to *imaging* and *integrated disease modelling*.

#### 4.4.2.2. Rheumatology

Rheumatologists perform an assessment of the disease, decide the most appropriate treatment and determine the follow up schedule for patients. They view and annotate x-ray and MRI images of bones and joints. Within Health-e-Child, the focus has been on juvenile inflammatory arthritis (JIA).

JIA is not a single condition; all the various forms of chronic arthritis in children are grouped and classified according to clinical criteria. Informed by this classification of disease subtypes, a rheumatologist must determine what drugs are likely to slow or eradicate the disease. This classification is therefore of considerable importance, but is currently inadequate in that the identified subtypes are not sufficiently homogenous. As a result, predicting how different drugs might affect a particular patient is a difficult task. The project is investigating a new classification, using data from genomic, proteomic, imaging and clinical sources to construct the categories.

When treating JIA, predicting the evolution of the disease is a key concern for rheumatologists. Another is detecting and quantifying early damage in images of patients with the disease, something that is currently missing in clinical trials.

Requirements from rheumatology include the *construction of homogenous JIA subtypes* (which will require vertically integrated data), and models for *predictive disease outcome*, and for *progressive organ damage*.

### 4.4.2.3. Neurology

For neurology, the specific focus has been on gliomas, a form of brain tumour that originates in glial cells. There can be strong links between certain types of gliomas and type 1 neurofibromatosis, tuberous sclerosis and Li-Fraumeni syndrome. Treatment is typically a combination of chemotherapy, irradiation and surgery.

There is considerable variability in the age and gender distribution, physical location, growth potential, invasiveness, tendency for progression, and clinical course of gliomas. Underlying biological differences are in many cases responsible for this variability, and there is evidence that molecular classification could be used to determine these for individual patients. Molecular information could therefore be of significant benefit to neurologists, in addition to the histological and morphological data for the tumour that they currently use.

The project determined that a variety of models using integrated data would be useful for neurologists, including models for *surgical decision-making*, *post surgery treatment*, and models supporting *automatic tumour detection and change quantification*. The potential for *individualised brain models* has also been recognised, which will require deforming a generic brain atlas in order for it to match the geometry of a patient.

### 4.4.2.4. Other areas

In addition to these three key areas, other related domains are explored:

- Radiologists perform imaging examinations, including capturing images and providing expert opinion on the images created
- Geneticists analyse genetic data from tissues in addition to clinical information to provide expert opinion on genetics for individual cases. They make use of public gene data repositories.
- Biologists perform tests on biological samples to generate genetic and proteomic data for individual cases.

### 4.4.2.5. Project Requirements

As a result of the above requirements, the following *technical* requirements were identified by the Health-e-Child project.

#### *4.4.2.5.1. Vertical integration of clinical data*

This is central to the project's aim of gaining a comprehensive view of the patient's health, utilising genetic, clinical and epidemiological data. Practitioners within different medical disciplines require information at different levels and with different granularity. For example, a cardiologist might be interested in a model of the heart, whereas a bioinformatician requires data at a finer granularity, such as molecular or cellular data. However, integrating data from these disciplines, including genomic, imaging and proteomic data, could provide medical practitioners and researchers with a unified, coherent view of the patient's current and past health.

This integrated view of the patient should be suitable not only for direct use by clinicians, but also for use by (semi)automated disease modelling and decision support systems.

#### *4.4.2.5.2. Storage and sharing of biomedical information*

In order for biomedical data to be shared between several hospitals, for analysis and annotation for example, methods for managing the heterogeneity of data and for adequate pseudo/anonymisation of patient data will be required. Advanced searching and matching techniques must also be provided. Biomedical information should be accessible from geographically distributed sources, including information from the local hospital intranet (PACS, HIS, etc.) and databases such as gene databanks on the internet. Access to data must be secure, with appropriate levels of encryption, and access rights must be enforced. Additionally, data sources must have sufficient availability and responsiveness to assure an appropriate quality of service.

#### *4.4.2.5.3. Biomedical query processing*

Clinicians from different disciplines will require different views of the integrated record provided by the project. Some may require the facility to find similar cases, others to identify illnesses common to patients in specific populations, etc. Given the vertically integrated nature of the patient records, these queries will be considerably complex, but must be executed sufficiently fast for use by clinicians during consultations or treatment.

#### *4.4.2.5.4. Integrated disease modelling*

While various organ and disease models exist already, such as electro-mechanical heart models and statistical maps of brain changes for certain conditions, these do not currently use vertically integrated data. Disease modelling is required at all levels addressed by the project, from molecular modelling such as searching for gene defects, to in-silico physiological models of the whole body at a high level. These models should aid clinicians when attempting to determine if a patient is likely to develop a certain disease, whether the patient currently has the disease or not, and if so how it is likely to develop, and what the best treatment might be.

#### *4.4.2.5.5. Decision support systems*

Given the scope and volume of data involved, a decision support system based on the health-e-child vertically and horizontally integrated patient record would be complex but potentially a very useful tool for assisting with diagnoses and treatments. Computer-aided detection using the integrated disease models mentioned previously could be a significant aid for both cardiologists and neurologists.

#### *4.4.2.5.6. Tools to support image annotation*

Image annotation tools should support the creation and modification of annotations by cardiologists, rheumatologists and radiologists. The project suggests that annotations be stored as metadata.

#### *4.4.2.5.7. Queries to find records of similar cases*

Based on similarity criteria specified by the user, a mechanism should exist to find and display similar cases, with any accompanying annotations, metadata and records of clinical decisions taken.

#### *4.4.2.5.8. Macroscopic computational models for key organs and diseases*

Generic models should be constructed, which can then be adjusted according to patient-specific clinical data in order to produce individualised models.

### *4.4.3. Virtual Physiological Human*

The EuroPhysiome initiative has led to the concept of Virtual Physiological Human (VPH) [17], indicating a methodological and technological framework that once established will enable the investigation of the human body as a single complex system. VPH will provide a framework within which observations made in laboratories, hospitals, and in the field all over the world can be collected, catalogued, organised, shared and combined in a variety of ways. It should also allow experts to collaboratively analyse observations and develop systemic hypotheses that involve the knowledge of multiple scientific disciplines, and 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 verifiable systemic hypotheses.

The Virtual Physiological Human (VPH) community has defined a number of requirements for grids in order to achieve its vision [18]. The main aims are to allow:

- observations made in laboratories, hospitals and ‘in the field’ worldwide to be collected, categorised, shared and usefully combined,
- interested experts to analyse these observations and develop hypotheses that span multiple disciplines,
- the interconnection of predictive models defined at different scales, and
- the verification of the validity of such models using clinical and/or laboratory observations.

#### *4.4.3.1. Technical requirements*

When examining the requirements of VPH with respect to grid computing, the following technical requirements were identified.

#### *4.4.3.1.1. Access to resources*

Researchers require access to all available resources in a uniform way, from those provided by their own department to specialised HPC resources. Access to these should be as seamless as possible, with simulations at different scales being automatically migrated and appropriate resources being used as required.

#### *4.4.3.1.2. User friendly interfaces*

Current grid portals require the user to specify parameters such as memory to be allocated and execution time; this would not be appropriate for VPH users.

#### *4.4.3.1.3. Grid usage models*

The nature of VPH simulations means that timescales are an issue, and current models of HPC use would not be appropriate. Instead, models which permit a large number of grid nodes to be used for a relatively short time ('burst mode') with little or no waiting time should be established. Resource co-reservation will also be required, particularly where multiscale simulations that run over multiple sites are concerned.

#### *4.4.3.1.4. Shared storage for large data/model repositories*

The imaging datasets concerned can be several hundred megabytes in size, but after pre-processing to generate a predictive model, the modelling and simulation data can be as much as a hundred gigabytes in size. With potentially thousands of these, there is a clear need for multi-terabyte storage facilities, connected to distributed HPC resources via high speed networks. These should incorporate the required security and confidentiality measures.

#### *4.4.3.1.5. Methods to solve multiple predicted models, in a coupled way*

As the coupling of predictive models at different scales is central to VPH's description of human physiology, coupled methods to solve multiple models will be required. This is considerably complicated by the fact that the models concerned may be very different both in conceptual nature and mathematical nature. Even relatively simple VPH problems can be considerably complicated by variations between individual subjects and treatment procedures.

#### *4.4.3.1.6. Direct prediction from medical images*

Methods for transforming a medical imaging dataset into a subject specific predictive model that do not require the costly pre-processing phase are being developed, such as the Boltzmann Lattice in haemodynamics and voxel meshes for hard tissue simulations. However, these are enormously computationally intensive, requiring fifty or more teraflops of computational power to solve in less than a day.

## 4.5. KNOWLEDGE GRIDS: AN EXAMPLE FROM GENERAL HEALTHCARE

For any given domain, a distinction is often drawn between declarative knowledge ('know what') and procedural or operational knowledge ('know how'). In the domain of healthcare, both kinds occur. What is often referred to as 'the scientific basis' of medicine, that which must furnish the evidence in so-called 'evidence-based practice', is present in research publications to which different standards of credibility are attached. For example, research results based on a randomized, double-blind, controlled clinical trial are held to be the gold standard, provided they were also submitted to adequate peer review. Evidence based on one physician's own practice, although not negligible, would be considerably less reliable. On the other hand, 'best ways' of treating patients – in a particular context – are often described in integrated care pathways (ICPs). It is not unreasonable to claim that declarative knowledge in medicine tends to be disseminated through peer-reviewed publication and operational knowledge through such things as guidelines and care pathways. In a healthgrid environment, these are brought together for the better treatment of patients and at the same time to improve research; indeed, the interplay between healthcare and research, e.g. through appropriately controlled 'secondary use', would be an important element in a full healthgrid environment.

Chronic obstructive pulmonary disease (COPD) refers to an airway obstruction caused by chronic inflammation. It is usually progressive, not fully reversible, and often occurs as a result of smoking but other factors, such as air pollution, can also contribute to the development of COPD. In the UK almost 900,000 people have been diagnosed with the disease, and the true number of people suffering from the condition is estimated to be around 1.5 million.

According to NHS guidance, the management of the disease should be tailored to the individual, with adjustments being made based on responses to treatment. The guidance includes a large number of drugs, including some off-label drugs such as Beclometasone, Fluticasone and Budesonide.

In the UK, the National Institute for Health and Clinical Excellence (NICE) has issued national guidelines for the treatment of the disease, but these are frequently modified to account for local variations and priorities. As a result, the procedure for assessing and treating the disease will vary even within a single country, let alone between countries. The evidence on which this guidance is based comes from a variety of sources, such as national studies by NICE and systematic reviews with an international scope.

Two main concerns exist for general healthcare; supporting the travelling patient, such as migrating elderly populations, and enabling decision support systems that can account for local variations in best practice and clinical evidence.

### 4.5.1. Challenges and requirements

Evidence from national studies, such as the aforementioned NICE study, may not be available to a doctor from a different locale. In order to continue treating the patient concerned, the doctor (or decision support system) must be aware of the evidence and guideline/pathway that informed the plan of care for that patient, and any deviations from that plan that have occurred to date. This may not a trivial matter of simply retaining a link to the relevant material, as there may be language barriers, and local reasons why the guidance followed in one country would not be appropriate in another.

The guidance also mentions drugs that are not certified for the treatment of COPD (off-label) despite this evidence coming from high quality systematic reviews. Different drugs will be certified for the treatment of the condition in different countries, complicating the process of following a single guideline or pathway regardless of travel. In fact, the patient concerned may be travelling for the express purpose of receiving different or less costly treatment in another country.

Prior history of exacerbations and smoking are essential for properly treating the disease, and therefore the doctor concerned must be able to access, comprehend and update the patient’s record. This requires standardisation of electronic health records (EHRs) and electronic integrated care pathways (eICPs). When it comes to decision support, a standard interface format, such as the proposed HL7 vMR [19], will also be a necessity.

# 5. WHERE WE ARE NOW: THE BASELINE

## 5.1. THE TECHNOLOGY

A ‘grid’ – not *the* grid – is now understood to mean an Internet-like infrastructure which 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 organisations’ (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.

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 [20] 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 [21] which is about common formats for interchange of data and about language for recording how the data relates to real world objects. Although many current grid infrastructures do not offer a web service interface to their services, we will concentrate our ‘state of the art’ on web services because it is the relevant technology for the future. We will then go on to discuss the status of existing grid infrastructures, the technologies they use and the services they offer.

### *5.1.1. 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 standardised 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) [22].

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. They can be described in a standardised 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 [23].

#### *5.1.1.1. 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. Data formats are expressed using XML Schema definitions which can be combined and 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. There are mechanisms, e.g. to restrict the WSDL language and interfaces that are compliant with the profile are therefore more interoperable. First generation WSDL only describes a service in functional terms and lacks the 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. Thus 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 [24]) may provide a much more complete service description. WSDL 2.0 has now been released and makes provision for semantic requirements.

#### *5.1.1.2. UDDI*

Universal Description, Discovery and Integration (UDDI) is another de facto standard for web services, and provides a registry for service discovery [25]. 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. 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, and ways to address these limitations are also being examined for upcoming versions of UDDI.

#### *5.1.1.3. Web Service Specifications*

The various web service specifications can be divided into first- and second-generation specifications. The first generation of specifications includes those mentioned above, which are widely adopted and fairly stable. The second generation of web service specifications are the so called WS-* because of the form of their names. 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 may require a client also to have a web service available. Specifications are currently becoming more standardised 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.

- 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.

#### 5.1.1.4. Web Services Resource Framework

Web Services Resource Framework (WSRF) [26] is a set of five specifications which define conventions for modelling and accessing stateful resources using web services. WSRF is part of the future WS roadmap backed by HP, IBM, Intel and Microsoft. 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 process is expected to be completed within two years.

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

### 5.1.2. Projects

We now consider a number of representative projects in grid computing, especially those that have already engaged with the biomedical and healthcare domains.

#### 5.1.2.1. EGEE

EGEE (Enabling Grids for E-sciencE) [27] 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**, that 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 68000 CPUs and more than 600 Petabytes of storage (data at August 2008).

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.

### 5.1.2.2. DEISA

DEISA (Distributed European Infrastructure for Supercomputing Applications) [28] 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.

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.

### 5.1.2.3. NorduGrid

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.

### 5.1.2.4. OSG

The Open Science Grid (OSG) [29] 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) [30], 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.

### 5.1.2.5. TeraGrid

TeraGrid [31], 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.)

With more than 750 teraflops of computing capability and more than 30 petabytes of total data storage, TeraGrid claims to be the world's "largest and most comprehensive distributed cyberinfrastructure for open scientific research" [31]. The project began in 2001 with an award from the US National Science Foundation (NSF), and in 2004 entered full production mode for academic research. This general purpose grid has been used for Image Guided Therapy [63] and is linked to Indiana University's Indiana Genomics Initiative. It provides gateways for biology and biomedical science, the National Biomedical Computation Resource (NBCR), and the Special PRiority and Urgent Computing Environment (SPRUCE). SPRUCE is developing and deploying technology to provide TeraGrid communities with fast, immediate access to resources to support large-scale models that can assist with urgent decisions impacting public health, safety, and security. This gateway provides massive resources on short notice, to applications that cannot simply run on a smaller set of resources for a longer period of time. Initial applications include LEAD and epidemiological pandemic modelling.

### *5.1.2.6. BIRN*

Launched in 2001 with the support of the National Institutes of Health's National Center for Research, the Biomedical Informatics Research Network (BIRN) [32] 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.

Within the BIRN project, biomedical researchers are standardising imaging protocols, and populating large distributed databases where they retain control of their own data. The BIRN portal provides a workflow and application integration environment, providing seamless access to the computational power required to perform large-scale analyses and to visualise and perform analysis on data stored anywhere on the BIRN virtual data grid. Complex, interactive workflows are supported, and provenance data is stored during data processing.

A major task of the BIRN coordinating centre (BIRN-CC) has been to develop a data integration system to enable researchers to make complex queries that include multiple data sources. The project also abides by the guidelines and regulations governing the sharing and storage of sensitive data, such as that of human subjects, through the use of encryption, auditing, and the security and integrity mechanisms present in its middleware.

### *5.1.2.7. Cancer Biomedical Informatics Grid (caBIG)*

Supported by the National Institutes of Health (NIH), the CaBIG infrastructure [64] provides many biomedical applications, including clinical trials, integrative cancer research, tissue banks and pathology tools, and clinical imaging of patients. An emphasis is placed on semantic and syntactic interoperability, with defined terminology, metadata, and information model standards. Although focussed on cancer research, the project aims to provide components that are applicable outside of this area.

The goals of caBIG are to connect scientists and practitioners through a shareable and interoperable infrastructure, to develop standard rules and a common language/vocabulary to aid the sharing of information, and to build or adapt tools for collecting, analysing, integrating, and disseminating information associated with cancer research and care. caBIG aims for its software and resources to be available to everyone in the cancer research community, with institutions maintaining local control over their own resources and data. Tools and infrastructure are being developed through an open, participatory process, making use of existing resources whenever possible.

The project also attempts to address the legal, regulatory, policy, proprietary, and contractual barriers to data exchange through it’s Data Sharing & Intellectual Capital (DSIC) workspace, preparing best practice guidelines and providing education services to caBIG participants.

The caBIG/caGrid infrastructure was used and extended for use by the cardiovascular research grid (CVRG) [65], which also makes use of BIRN.

### *5.1.2.8. National Biomedical Computation Resource (NBCR)*

Funded by the National Center for Research Resources (NCRR), the mission of the National Biomedical Computation Resource (NBCR) [66] is to conduct and enable multiscale biomedical research using a cyberinfrastructure. The development of this cyberinfrastructure is driven by multiscale modeling applications, which focus on scientific research ranging in scale from the molecular to organ level. Examples include the calculation of protein electrostatic potentials with APBS, protein-ligand docking studies with AutoDock, cardiac systems biology and physiology modelling with Continuity, and molecular visualizations using PMV. Projects include understanding the mechanism of action of HIV protease and integrase, neuromuscular junction research in myopathy, heart arrhythmia and failure, and emerging public health threats. NBCR also aids in the development of ontology and semantic mediation tools such as Pathsys and OntoQuest for data integration and interoperability, which may be coupled with application services provided for the biomedical community.

Large scale computation problems may be launched transparently on national scale infrastructures such as TeraGrid.

# 6. TECHNICAL, ETHICAL, LEGAL AND SOCIO-ECONOMIC ISSUES

## 6.1. TECHNICAL ISSUES

### *6.1.1. Standardisation issues*

Standards are absolutely necessary for the deployment of services which integrate data in bioinformatics and medical informatics, but are also vital for data coming from different medical disciplines and even data coming from different countries in Europe. These standards are needed for building data models, producing ontologies and for the development of knowledge management services. The adoption of standards for the exchange of biological and medical information is still limited to a few specific fields. Moreover, they need to be compatible with grid standards so as to allow their implementation on healthgrids.

### *6.1.2. Communication issues*

Lack of information about grids and grid technologies is frequently identified as one of the key reasons why there has been very little interest in them from the field of medical research. It is essential that all relevant actors to be kept well informed by the HealthGrid community of the potential benefits of the technology to them. Success stories demonstrating the impact of grids for medical research will be vital for convincing medical researchers of these benefits. As a result, there is a need for a suitable demonstration environment, offering very easy access to the grid for non experts and providing services that will help convince the medical research community. On this dissemination environment, dedicated efforts to promote the technology can then be developed.

### *6.1.3. Security issues*

Deployment of a data grid for medical research will only be possible when the middleware can provide all the necessary guarantees in terms of management of personal data. We perceive the specific technical requirements related to the handling of medical data on the grid to be as follows.

- Manipulation of personal data on the grid must obey strict regulations. These regulations vary between European member states.
- Services for the anonymization and pseudonymization of medical data must be provided.
- Medical data is the property of the patient. A mechanism must be set up to allow individuals to access their data on the grid.
- For healthcare purposes, the authentication of healthcare professionals on the grid cannot be handled by requesting all of them to get a grid certificate. A mechanism must be set up so that professional cards can be used to provide authentication on the grid.

## 6.2. ETHICAL, LEGAL AND SOCIO-ECONOMIC ISSUES

### *6.2.1. Ethical Issues*

Ethical issues in healthgrids may be summarised in three well known ethical principles: autonomy, beneficence and justice, in which the three each have individual value but in which the three must be taken as a whole offering a system of ethics in which the needs of the individual are balanced with the needs of society.

#### *6.2.1.1. Autonomy and Healthgrids*

Most common belief systems give a special place to the autonomy of the individual - the right of the individual to control his or her own person. The concept autonomy is intimately tied up with the legal duties of consent and confidentiality, both of which could prove difficult in the context of healthgrids. Thus, in healthgrids, one of the key ethical issues will be in the possible compromise of the patient's autonomy that will arise from sharing his or her data with people who are yet to be identified. It is worth noting that it has been argued, notably by the European Article 29 Data Protection Working Party [33], that consent may have only a very limited place as a justification of the sharing of health related data in the electronic age. The limitation is based on the argument that to exercise autonomy one must be able to make decisions unfettered by coercion. If it is accepted that sharing health data allows doctors to provide better care, then a patient who refuses permission to share such information will be *de facto* opting for a lower quality of healthcare, arguable therefore he or she is not able to withhold consent freely and is therefore not able to act autonomously. It is argued therefore that robust system of security of information and ethical practice should be adopted in which patients will be able to trust, notwithstanding that their information is shared, and providing for special opt-out possibilities when the nature of the information is especially sensitive.

#### 6.2.1.2. Beneficence and Non-Malfeasance

The ethical duty of Beneficence and Non-Malfeasance is the duty to do good – or in the words of the Hippocratic Oath at least to do no harm. This ethical duty is frequently used to justify the adoption of health technologies which allow doctors to better treat their patients. The argument with respect to healthgrids is, that in order to act ethically, a healthcare professional, would is obliged to use suitable grid applications if they are available. A healthcare professional refusing to use standard medical technology or refusing to prescribe antibiotics would be considered in breach of his or her duty of beneficence, thus, as the sophistication of grid aided diagnosis develops we will one day arrive at a time when a healthcare practitioner not linked to the appropriate grid networks will be in breach of his or her duty.

However, until we have reached a time when grid applications are stable, well ‘fed’ with data and fully integrated into the evidence base of good clinical practice such arguments will not apply. At present, in the more experimental stages of the healthgrid it will be important to ensure that the use of the applications does no harm, but perhaps most importantly to ensure that the patient is aware of any possible medical and social.

#### 6.2.1.3. Justice

The ethical principle of justice concerned with the duty to achieve a fair distribution of resources as well as the need to develop an overall just medical system in which the greatest health benefit of the greatest number is achieved is the principle of justice. In most legal systems this ethical principle is used to support social systems of distributive justice which provide for tools as taxation to distribute wealth on such a way that all may be afforded an acceptable minimum of social care. The developments of applications such as MammoGrid have established that the sharing of a very large number of mammogram images across a wide network that allows radiologists to test suspect images against a known and tested database of cases significantly contributes to the early detection of breast cancer. The healthgrid in this case not only acts to the benefit of the known patient whose suspect image is submitted to the tool, but to the overall health of the population.

### 6.2.2. Legal Issues

The legal issues presented here and further analysed in SHARE deliverables D4.1 and D4.2 were chosen for their relevance to healthgrid technologies. Other legal concerns such as competition issues are of relevance for grid technology, but in order for a full and complete analysis to be made, these were intentionally omitted. Regarding competition issues in relation to eHealth, please see the reports and analysis from the Legally eHealth project [68].

#### 6.2.2.1. Data Protection

The ethical principle of autonomy is legally underpinned by the duty of data protection. The EU has taken this principle very seriously, and as well as including privacy with the European Charter of

Fundamental Rights has developed robust EU level law in the Directive on Data Protection to promote privacy. While this current EU level legislation is adequate for the development of healthgrids, it is not ideal for promoting the use of healthgrids.

As noted above, autonomy is balanced by beneficence and justice, thus when healthgrids are used for treating patients or planning care, the balance of rights weighs in favour of data collection - that is, it is assumed that the patients' general interest in obtaining treatment or advancing medical care outweighs interests in privacy.

The current legislation is not, however, adequate to support most of the longer running research initiatives around which healthgrids are based[5]. As the current EU level legislation stands, Member States can enact specific legislation covering specific tools such as healthgrids in order to exempt scientists and medical practitioners using healthgrids from some of the more onerous duties of the Directive.

No Member State has addressed legislation to this particular issue and so healthgrids are burdened with onerous data protection requirements which could deter scientists from using adopting healthgrid technology and using its enhanced computational and data acquisition power.

#### 6.2.2.2. Liability

In line with the ethical duty of beneficence a legal system of liability has been developed in all legal systems in which the duty not to harm is shored up by systems of compensation to support those who may nonetheless be injured. At EU level legislation on Liability for Goods and Services is reasonably well developed, but does not in its present form lend itself well to the healthgrid domain. One of the reasons for this is, of course, that health services are organised at national or regional level and that the European Union has no legal competence to draw up legislation which states specifically how a health service should be organised. However, the EU does have a range of legislation designed to protect citizens from harm resulting from goods offered on the market [34]. Steps could be taken using guidelines, or even specific legislation, to address distributed computing services, such as healthgrids that would seem at present to be only marginally covered by the existing rules. Accordingly it is important that the existing European framework of general product safety be re-examined to consider its applicability to distributed networks such as healthgrids.

#### 6.2.2.3. Intellectual Property

The ethical principle of justice is concerned with ensuring a fair distribution of the needed and desirable, whilst also respecting individual interests (autonomy) and the duty to do no harm. In modern legal systems this principle is used to develop legislative tools which seek to balance individual work and cost with equitable access to goods. This gives rise to the concept of intellectual Property Rights which provides systems of sharing the fruits of intellectual endeavour (such as software code) with the interests of rewarding those whose individual labour was used to create the good.

In the EU this has resulted in a system where the owner of the copyrighted software running a healthgrid has the exclusive rights to reproduce his work, prepare derivative works, distribute copies to the public, perform the work publicly and display the work publicly. Under these circumstances any natural or legal person would have to pay to use computer programs while they constitute one of the most important compounds of healthgrids. Given that most Grid applications will depend on shared access to multiple-copyrighted programmes it is unlikely that such a model of copyright is useful in protecting the entirety of a healthgrid application.

[5] Analysis of this issue is offered in SHARE deliverables D4.1 and D4.2.

An open standards approach to software co-development could help the development and implementation of healthgrids. The open source licensing model actually uses copyright and contract principles to retain control of the work while enabling its use effectively for free and could thus encourage use and development.

### 6.2.3. Socio-Economic Issues

Legal fine tuning, whether through standardised contracts, special data sharing agreements or open source software development models, will be of little use in driving forward the development and implementation of healthgrids if the social and economic setting does not provide incentives, or if it presents other barriers to development or use. As these issues have not yet been examined in detail, it is necessary to analyse these aspects of healthgrid settings thoroughly in order to develop fully weighed up cost-benefit and cost-utility assessments of the use of healthgrids in healthcare delivery. In particular, the social and economic drivers and barriers (notably private incentives) must be examined and, where necessary, altered by different levels of policy intervention – from awareness-raising to direct financial support for specified initiatives.

From a *socio-economic perspective*, it becomes obvious that the uptake of healthgrid systems and solutions will also heavily depend on the extent to which they can help address problems and challenges of health systems [35]. Such impact is presumed, yet there is little evidence of its scope. Detailed analysis of existing applications, as well as ex-ante assessments of the benefits from the future use of healthgrids will be essential for mobilising the required will and enthusiasm among research funding entities, political organisations and society at large. Potential benefits include timesaving, particularly important in cases of potential pandemics, and access to better quality clinical and research data, leading to improvements in the quality of clinical outcomes.

Another inhibitor to a widespread adoption of healthgrid solutions that needs attention is the lack of (knowledge about) private incentives. A business case for the routine use of grid technologies in the health sector is essential for moving from project-based, exemplary utilisation to a widespread uptake of healthgrid based solutions. As has been acknowledged in the literature [36], private incentives play an important role in healthcare, and will also play a major role in this business case.

### 6.2.4. Organisational, social and cultural issues in the use of healthgrids

Both at the individual and the societal level, issues like universality of availability of full healthcare services to all citizens, equal access to healthcare, and equal high quality of services rendered are key issues [37]. Geographic factors relate mainly to equal access to quality care independent of location of living. ICT-based systems pose new problems like access to EHR by insurance companies or employers, and even police and prosecutors. Opinions and attitudes of patient and citizen associations and lobbying groups, often magnified by the media, can have strong impacts through public (policy) discussions of these topics on the implementation and diffusion of healthgrids.

The organisational level is always complex. Perspectives, confirmed by two most recent research studies, include:

- Changing care pathways that need new information, skills, knowledge and process in healthcare providers
- Changing roles of healthcare professionals, teams and healthcare organisations
- Transfer of roles between healthcare professionals, teams and healthcare organisations
- Increased collaborative working and exchange of information between providers
- New relationships between citizens and healthcare professionals and organisations
- New strategic partnerships for third party payers and healthcare providers.

Finally, cultural issues are a key factor in health services, including the great diversity of attitudes, behaviour and knowledge exchange among professional and non-professional staff involved in healthcare, and the impact this has on the quality, efficiency and processes of services. Education and training, professional standards and bodies, rules and regulations, attitudes and behaviour all have an influence here.

### 6.2.5. Technology transfer

There is evidence of recognition that the European Union has much to gain from encouraging greater cooperation between its funded research and development projects. There is a case to be made for specific support mechanisms, among which we have considered:

- A specific forum for collaboration issues possibly with a funding stream attached.
- Proportional leveraged funding for collaboration between projects (e.g. an additional 5-10% achievable only if projects enter into collaboration).

# 7. ONCE OVER LIGHTLY: A FIRST APPROACH TO A ROADMAP

## 7.1. A FIRST TECHNOLOGY ROADMAP: DEVELOPMENT AND DEPLOYMENT

### 7.1.1. Introduction

In order to better understand the development of the integrated roadmaps produced by SHARE, it is useful to revisit the initial draft of the technical roadmap, produced after the technology baseline had been established. Although it placed too much emphasis on the technological push from researchers rather than the pull from user communities, and was ultimately too simplistic, this roadmap provided a very clear view of the steps we believed would be required in order to reach the goal of the deployment of generalised healthgrids for medical research.

This initial technical roadmap consisted of a series of interlaced technical and deployment milestones. Figure 2 shows the four technology milestones that were defined for the implementation and development of grid services and required standards (purple), and the four interlaced deployment milestones (green) relate to the computing, data and knowledge grid paradigms. SHARE predicted that initial phase would be achievable in a fairly short amount of time, whereas the challenges of the second phase would require more time to address. We estimated the journey from a sustainable computing grid to a generalised knowledge grid would take from seven to fifteen years, although others have commented that in reality the timescale could be far longer.

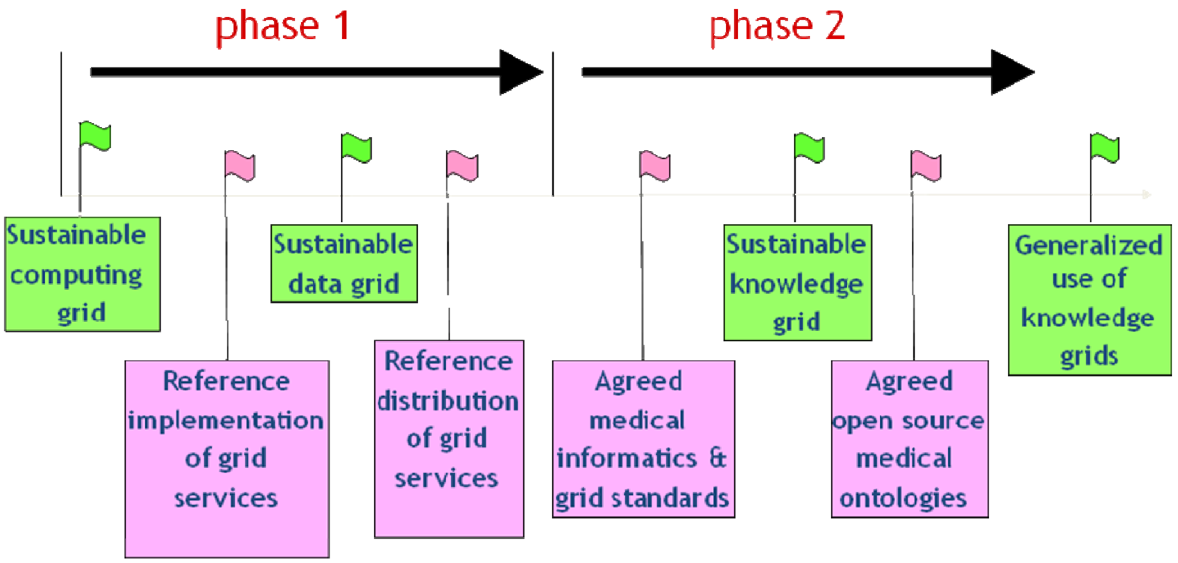

phase 1
phase 2
Sustainable computing grid
Reference implementation of grid services
Sustainable data grid
Reference distribution of grid services
Agreed medical informatics & grid standards
Sustainable knowledge grid
Agreed open source medical ontologies
Generalized use of knowledge grids

**Figure 2: the initial technology roadmap diagram**

### 7.1.2. Sustainable computing grid

The first step defined was to achieve a sustainable computing grid infrastructure for the medical research community, which we predicted should be an achievable goal in the near future given the success of computing grid applications on existing general purpose grid infrastructures. Challenges for the successful deployment of a computing grid within a hospital or clinic would include convincing management of the benefits of grid technology, ensuring the computer and network infrastructure is sufficient (enough bandwidth, fast enough storage, etc), and ensuring user interfaces, and the installation and administration of grid nodes are simple enough for non-grid experts.

### 7.1.3. Reference implementation of grid services

The first technology step was the development of a reference implementation of grid services, using standard web service technology and allowing computation and secured manipulation of distributed data. The important issues for this milestone were the use of web service standards (and the level of tool support), the maturity of the web services

For example, web service description languages and registries such as WSDL (Web Services Description Language) and UDDI (Universal Description, Discovery and Integration) are still lacking when it comes to semantic queries and descriptions, non-functional descriptions, and ontology-based searching.

This reference implementation would also need to address basic security issues such as secure data transfer, secure mechanisms for access, authentication, and authorisation, as well as sites for secure data storage, all sufficient for medical data. The potential storage of anonymised or pseudonymised data one grid nodes outside of a hospital’s firewall would need to be addressed.

### 7.1.4. Sustainable data grid

The next step was to develop a sustainable data grid for a specific, well defined medical research area. Healthgrid projects have created prototype data grids for medical research, but these are far from production quality. Limited data management services have hampered the storage of medical images for example, and high speed links between data providers and consumers will be a prerequisite. The geographic distribution of data inherent in data grid storage means that many legal and ethical issues will need to be resolved, such as the ownership of patient data, ethical control of information, the patient’s right to access or be informed about data that concerns them, and confidentiality issues.

### 7.1.5. Reference distribution of grid services

A reference distribution of grid services would then be produced for installation on grid nodes in medical research centres. This distribution should be sufficiently tested for scalability and robustness, and the underlying complexity must be hidden from grid users, with administration of grid nodes also made as simple as possible. Scalability, particularly regarding medical applications, is still a concern for grid middleware based on the Open Grid Services Architecture (OGSA) [38] such as GT4 [39] and GRIA [40]. Middleware such as gLite [41] and Unicore [42] on the other hand have been deployed on large scale infrastructures in Europe and have demonstrated their scalability and robustness, but are still awaiting migration to web services.

### 7.1.6. Agreed medical informatics and grid standards

The use of computer-based tools for clinical research has led to the definition of standards for the exchange of data in many areas. However, such standards are in many cases not universal, with different disciplines and countries adopting different standards. The exchange of data between bioinformatics and medical informatics is an area where standards are particularly limited. Medical imaging is an exemplary case, in which the adoption of DICOM [15] for the acquisition, connection and storage of medical images has been accepted worldwide. Medical records are another area where standardisation would have clear benefits, with HL7 [43] being the favoured standard for the exchange of data. However, previous standards such as CEN/TC251 EN13606 [44] focused more on the storage and structuring of clinical records and have prevented a wider uptake of HL7. A particularly important consideration for both of these standards is their compatibility with grid technologies, and how they could be implemented on a healthgrid. Both DICOM and HL7 developers are just starting to study the interface between their standards and web services technology.

### 7.1.7. Sustainable knowledge grid

The next deployment milestone is the successful transition from a data grid for a well defined research area to a knowledge grid. The synthesis of knowledge from data will require sophisticated data mining, modelling and image processing applications, and may also involve the use of techniques from Artificial Intelligence (AI) to derive relationships between data from different sources and in different contexts. The deployment of a knowledge grid will be a significant step as none currently exist.

### 7.1.8. Agreed open source medical ontologies

The last technology milestone will require open source medical ontologies to be agreed and implemented. 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.

### 7.1.9. Generalised use of knowledge grids

The final milestone is the generalisation of the knowledge grid produced in the previous deployment step, allowing it to be used outside of the defined medical research area. The development of medical ontologies required by the previous milestone will allow relationships between concepts and nuances in meaning to be captured, greatly enhancing the opportunities for communication, knowledge sharing and machine reasoning. However, the transition from a knowledge grid for a single research area to a generalised grid for medical research will not be a simple task.

## 7.2. QUESTIONS OF STANDARDS (MEDICAL PART)

Along with the migration of health applications to grid environments, there are many tools that do not benefit from the migration to grids. These tools, normally used in medical informatics environments for the access and processing of health data, should be however, somehow compatible with the grid. It will be important to develop gateways to standard formats of medical data exchange, such HL7, DICOM, IHE or CEN TC251 norms.

The most relevant interfaces needed are:

- ***hospital data*** Although medical databases have different storage formats, there exist de-facto standards and other standards under development that regulate the exchange of medical data. Hospital information, for example, is needed for epidemiological research. The availability of HL7 v2.5 and HL7 v3 (de facto standard), prENV 13606-4 (norm under development by the CEN TC251), CDA, RIM and OpenEHR gateways will ease the integration of the medical resources on the grid. Support to other vital signs exchange formats, such as ENV13734 or IEE11073 will also be important. Other important standards related to the continuity of care are CCR and CONTSYS EN 13940.
- ***medical imaging*** Screening for early treatment of cancer in breast, colon, lung or prostatic cancer is habitual in many areas. There are several attempts to develop DICOM-conformant interfaces to grid-storage systems, such as DICOM-SRM [45], MEDICUS [46] or TRENCADIS [47]. However, DICOM is a large and complex specification, and current approaches only cover parts of the standard. Moreover, DICOM components need to be certified for their use in production.
- ***statistical tools*** Medical research is a main aim for Healthgrids. Thus, along with the standardisation of the interfaces to data, it is important to provide interfaces to the statistical

tools most widely used in the medical community. Along these programs, the tools of the Centers for Disease Control of Atlanta and other related tools (EPIInfo, EPIMap, SIGEpi, EPIDAT) are widely spread. The support for software using "R" and "S" statistical languages will also improve the interoperability of the infrastructures.

The availability of those interfaces will ease the process of integrating grids for health infrastructures in the health environments without affecting severely the current processes, thus quickly providing enhanced performance.

## 7.3. MAPPING ELSE REQUIREMENTS INTO A TECHNOLOGY ROADMAP

### 7.3.1. ELSE Roadmap

Addressing the issues listed above is a challenge as the advice of medical and legal bodies is crucial. Thorough planning and structuring to the necessary actions and steps is needed to make the processes easier. The ELSE roadmap could be the answer, but it will not be enough unless it is harmonised with the technical roadmap milestones.

In this section we discus the different requirements we suggested to constitute the ELSE roadmap.

#### 7.3.1.1. Ethico-Legal Requirements

The primary concerns will be establishing systems in which the ethical principles of autonomy, beneficence and justice can be achieved through adequate legal and social tools. This will depend on legal structures to support data sharing while maintaining privacy as well as adequate tools for determining the responsibilities of the healthgrid actors, so that good may be achieved and harm may be compensated should it occur. In addition, systems will need to be developed that allow the fair and just distribution of the benefits of a healthgrid whilst still compensating those who build it.

Where patient identifiable data are used in a healthgrid, patient consent is crucial to the legitimacy of such medical data processing and transfer; therefore verifying that the patient has unambiguously expressed his/her consent should take place prior to any data manipulation. A technical way of supporting this could include adding a flag or metadata to the patient record to indicate whether he has any objection to the processing of his/her personal data.

Appropriate and user friendly ways of allowing patient access to data is also recommended. This will help patient not to feel totally dispossessed of data and information that concerns them and excluded from the data processing process. Thus more public trust will be added to research carried out within the healthgrid domain.

Robust legal solutions also need to be developed with respect to liability so that possible damage to the patient arising from the use of a healthgrid could be outlined along with some preventative measures. Logging and auditing must be addressed early to monitor whether enough testing was done to healthgrid services and products.

#### 7.3.1.2. Data Protection Requirements

Our concerns here are with patient privacy and how it could be best protected within the healthgrid environment. Patient identification issues should be discussed and good analysis and evaluation of the current de-identification software and tools should be produced. We suggest a start with medical images de-identification as it might contain recognisable parts of the patient body. This action should start at an early stage so the deployment milestones could benefit from any recommendations.

At this stage researchers need to make sure robust anonymisation, pseudonymisation and other identity protection techniques are developed and deployed in the grid infrastructure. The eDiamond project [48] suggests that a semantic understanding of the reasons why a person may be accessing particular pieces of data is crucial to the legitimacy of data processing in the healthgrid environment.

#### *7.3.1.3. Ethical Control Requirements*

The requirement for ethical oversights and monitoring should be dealt with at an early stage. As a first task, focus will be on the requirements and tools to facilitate oversight, with automation being explored. Then the effort will be oriented to satisfy the arrangements for automated ethical control for a data grid which will be more complex with long-term data storage.

#### *7.3.1.4. Policy Requirements*

These will cover data processing and transfer issues such as legitimacy, accessing the minimum data required, the ethical transfer of data, quality assurance, compliance with confidentiality rules and limiting the period of data storage.

This should be dealt with some time before each deployment milestone as requirements might differ when changing from a computational grid to a data grid or a knowledge grid.

### 7.4. QUESTIONS OF SECURITY

The security management has several issues to cover, including:

- ***authentication*** Although user authentication is a problem well solved in public key infrastructure (PKI) environments, in which most grid infrastructures sit, it will be important to analyze how these procedures are being implemented in health infrastructures. Normally, health users do have (or will very soon have) a means of digital identification used for accessing clinical records in their daily practice. Trusting different certifying authorities is feasible and should not present additional problems. However, authentication must go beyond the plain concept of grids and integrate with the rest of digital identities of the individuals. Structuring identity views and federating identity issuers should be addressed in the context of the identity 2.0 concept.

- ***single sign on*** Users must be able to provide their credentials only once and let processes act on their behalf. The use of proxies, proxy repositories and X509 standard attribute extensions are sufficient to deal with these requirements, provided that the authorisation model could manage the same model, as described in the next point. This must allow the verification of the authenticity off-line. Risk of theft of credentials is much more severe in healthgrids, since the compromise of privacy for one short period could be sufficiently attractive for fraudulent users and disruptive for the whole system.

- ***authorisation*** Management of the authorisation has not been effectively addressed yet in multiple-decoupled institutions. The use of attribute extensions in a central authorisation system (such as VOMS) reduces the flexibility of the management of the authorisation - which must be set-up at each site in a coordinated way- as well as the flexibility on the membership – which might not be scalable when the number of users increase and a need for quick reactions is needed. Trustee authorisation entities schemas, such as combining Shibboleth and PERMIS systems, could reduce the problems in deploying large-scale VO membership and delegating on trusted authorisation mechanisms but they have not yet demonstrated their viability when scaling up to thousands of users (as medical institutions have).

- ***delegation*** The delegation of credentials is a well-known problem that has been reasonably solved in many situations. Processes should be able to act on behalf of third parties who started them, with different levels of capabilities. Delegation could be full or limited and last for a defined period of time. However, the delegation of authorisation, of keys for accessing back-ends, and of roles has not been completely solved yet, and these issues have an important impact when accessing third-party applications whose security levels include additional features, such as login and password.

- ***privacy*** Privacy management is the hardest problem regarding security. Legal regulations impose, from a technological perspective, data dissociation, pseudo-anonymisation and encryption. Since any medical data is potentially personal –since further research could discover particularities unique to a patient– and considering that the processing starts with the storing of the data, state-of-the-art technical means must be applied to protect data from unauthorised access, both on the storages and on the network. Many schemas have been proposed (perroquet, MDM, TRENCADIS) aiming at data encryption and decryption on the fly, multiple key shares, reliable services, etc. Large-scale deployment of these techniques should be performed. Finally, the management of genetic data introduces more problems and difficulties due to the potential re-identification of data. It must be ensured that the issues outlined in Council of Europe Recommendation R (97) 5 on the Protection of Medical Data (Feb. 13, 1997), are taken into account.
- ***non repudiation*** This concept is especially important in the health context, in which users should not be able to deny the authorship of an action. In the case of epidemiology, in which the objective is not patient care, this concept is not so critical. It could however be applied to the data collection and the surveillance networks, in which the responsibility of the correct value of the sources has deeper impact.
- ***integrity*** Permitting access to data should not mean permitting its modification or deletion. Any loss or change to patient data could result in serious consequences leading in the worst cases to death. Measures should carefully be taken to insure malicious or un-intentional altering to the data is detected and forbidden. Attacks that threaten data integrity could target the data transmission channels or the data storage. These threats should be thought about while designing both message level security and data security.
- ***queries logging and auditing*** Recording and logging either queries requesting access to sensitive data or the results of such queries are not sufficient to detect forbidden access to sensitive data. There is a high requirement for the use of semantic and logical technologies to allow the auditing system to combine query logs, result logs and other backlogs of the database to generate audit trails identifying the user, recipient, purpose and time of query and the exact information disclosed by each query.
- ***policy compliance*** All the policies enforced at the different levels of the grid security infrastructure, need to be designed in a way to comply as much as possible with the regulation. Privacy policies have proven success in preventing malicious access to organisational resources in the business sector; however those policies are still inadequate to strongly protect a patient privacy. Before being enforced in a healthgrid domain, privacy policies need to go trough a whole process of refinement and enhancement to better cover or comply with regulatory obligations. The use of automated ways of auditing compliance at the different security infrastructure of the grid participants is also required.
- ***proxy certificate lifetime*** Most security issues of grid computing are related to the nature of Virtual Organisations (VOs) using a grid. The dynamic nature of VOs is advantageous but presents additional security risks. A large number of users can be rapidly added or removed from a grid thanks to grid mechanisms such as proxy certificates and identity chaining. The 12-hour lifetime of proxies is generally seen as too short to allow users to fully benefit from grid capabilities [60], i.e. being able to execute long experiments without manual supervision. Conversely, users who have had their access rights withdrawn may still be able to execute applications until the proxy expires. Security mechanisms are needed to ensure no protected data is extracted from grid resources when users are removed or blocked. With such mechanisms, the proxy certificate lifetime could be safely extended, and users might also be

able to specify the lifetime of the proxy certificate according to the needs and types of jobs to be executed.

- ***the virtual organisation as a policy domain overlay*** Virtual organisations can be long-lived or short-lived. When a VO is created, the users and resources involved will be governed by the rules of the classic organisations to which they belong. In most cases the controls of classic organisations will not be adequate for coordinating and organising the effective sharing of different resources constituting the VO. Conflicting rules might also exist due to the diversity of organisational interests and geographical considerations. Classic organisations may outsource some policy controls to promote resource re-use and sharing; the VO is then considered as a policy domain overlay [61]. However, the concept of outsourced controls might not be appropriate in a healthcare and medical research context given the highly sensitive nature of the data concerned. Also, in the case of short-lived VOs with a small number of users and resources, the participants might prefer to specify and negotiate the governing policies with the help of trusted grid mechanisms.

Grid computing presents a multitude of security issues that must be studied and addressed before the impact of these on the future of grid-based healthcare and biomedicine can be determined. Grid security is currently a very active research area, and many proposed solutions to these issues are emerging and merit further attention.

# 8. REVISITING THE PARADIGMS: CHALLENGES FROM USER REQUIREMENTS

The requirements were collected from three user communities: epidemiology, innovative medicine and the Virtual Physiological Human community. They are described in the deliverables D5.2a and D5.2b and are summarised in this chapter.

## 8.1. CHALLENGES FOR THE COMPUTATIONAL PARADIGM: INNOVATIVE MEDICINE

### *8.1.1. 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 [49]. This document has been used in deliverable D5.2b 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.

### 8.1.2. 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 customisable 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.

## 8.2. CHALLENGES FOR THE DATA PARADIGM: EPIDEMIOLOGY

### 8.2.1. Introduction

As documented in SHARE deliverable D5.2a, 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 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 system will meet the concerns of the ethical and legal committees of their research institutions.
- 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.

### 8.2.2. 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.

## 8.3. CHALLENGES FOR THE COLLABORATION PARADIGM: VPH

### 8.3.1. 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 [17], 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 [18].

### 8.3.2. 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.

#### 8.3.2.1. 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 or different platforms (and by implication, between resources appropriate to run the simulation at a particular scale).

#### 8.3.2.2. 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. Current technologies are insufficient to protect the privacy of the data outside the health network barriers, according to legal regulations and ethical principles request. Technologies are also not scalable when dealing with fine-grain authorisation, and delegation methods in current practice could be not sufficient for medical applications. Finally, 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.

#### 8.3.2.3. 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.

#### *8.3.2.4. Other requirements*

The VPH roadmap [17] 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

### 8.4. ELSE CHALLENGES

***Liability in a Healthgrid System***

Using grids blurs the liability issues in terms of medical practice. While the EU has a range of legislation designed to protect citizens from harm resulting from goods offered on the market, the construction of healthgrids makes it difficult to ascertain at which EU level legislation would apply to each part of the system.

This is particularly the case with the law on medical devices, which is unclear with respect to healthgrids. In September 2007, the European Parliament and the Council adopted Directive 2007/47/EC of 5 September 2007 amending Council Directive 90/385/EEC on the approximation of the law of the Member States relating to active implantable medical devices, Council Directive 93/42/EEC concerning medical devices and Directive 98/8/EC concerning the placing of biocidal products on the market (OJ, L 247/21, 21.09.2007). In particular, it is stated in Recital 6 of this directive that *"it is necessary to clarify that software in its own right, when specifically intended by the manufacturers to be used for one or more of the medical purposes set out in the definition of a medical device, is a medical device. Software for general purposes when used in healthcare setting is not a medical device."* An amended definition of what a medical device is is also to be found in the directive.

***Data Protection*** As regards data protection issues, we argued that in broad terms the current EU level legislation was adequate but not ideal for promoting healthgrids as it does not address any particular issue related to healthgrids' systems and services.

However, the European Working Party on Data Protection, established under article 29 of the Directive and composed of the national data protection authority of each Member State, has recently acknowledged that some special rules may need to be adopted for key eHealth applications. This will have an impact on the way healthgrids are designed and particularly on the way the data are collected and processed in these systems.

***Consent Management*** The Article 29 Working Party does not see consent as a valid basis for processing data in an EHR. It considers that, as the creation of medical records is a necessary and unavoidable consequence of the medical situation, a health professional may have to process personal data in an EHR, and thus withholding consent may be to the patient's detriment. The Working Party argues that consent might not be valid if it is given for general processing of the EHR and for sharing with unnamed healthcare professionals (HCP). It argues that valid consent is limited to the sharing of data with a specific HCP and for a specific purpose. It would seem therefore that even sharing a record with several HCPs in the course of the treatment of a disease or condition may not be covered by a general consent where those HCPs and the nature of their intervention is not known by the patient at the time consent is given.

***Intellectual Property*** The collaboration between private and public institutions will be particularly significant for scientific research in healthgrids, and may create numerous problems with respect to ownership of intellectual property. As pointed out in D4.2, there is a contradiction between the intellectual property rights and the needs of the grid technology, which would require that the access to databases and to software is free of rights. The challenge for EU and/or national legislators is therefore to find a way of balancing the two competing sets of rights

***Ethical Challenges*** Deliverable 5.2b noted that the potential benefits to society through the use of grid computing were significant. Taking as example the drug discovery field, benefits are not only in terms of the potential to alleviate suffering and illness, but also in the economic impact on reduction of illness and the drug development industry itself. It is therefore important that in the development of such technology, due consideration is given also to the ethical impact of failing to use grid technology. If drugs can be discovered more quickly and more efficiently using the technology is it not within the ethical duty of beneficence for governments to support such developments?

The ethical duty of justice, which is concerned primarily with the fair allocation of resources, could also call for the use of a technology that can lead to quicker and more efficient drug development, again taking the ethical argument out of the private domain and into the political arena of public funding and support.

***Trust and Acceptance*** Trust is a very important element in any interaction between the different members of a society. In the market context, trust is crucial for successful business to business collaborations. Similarly, in a healthgrid domain a good collaboration will not be achieved unless a trust relationship is built between the different users and stakeholders. Legal and ethical uncertainty could lead to the rejection of such technology.

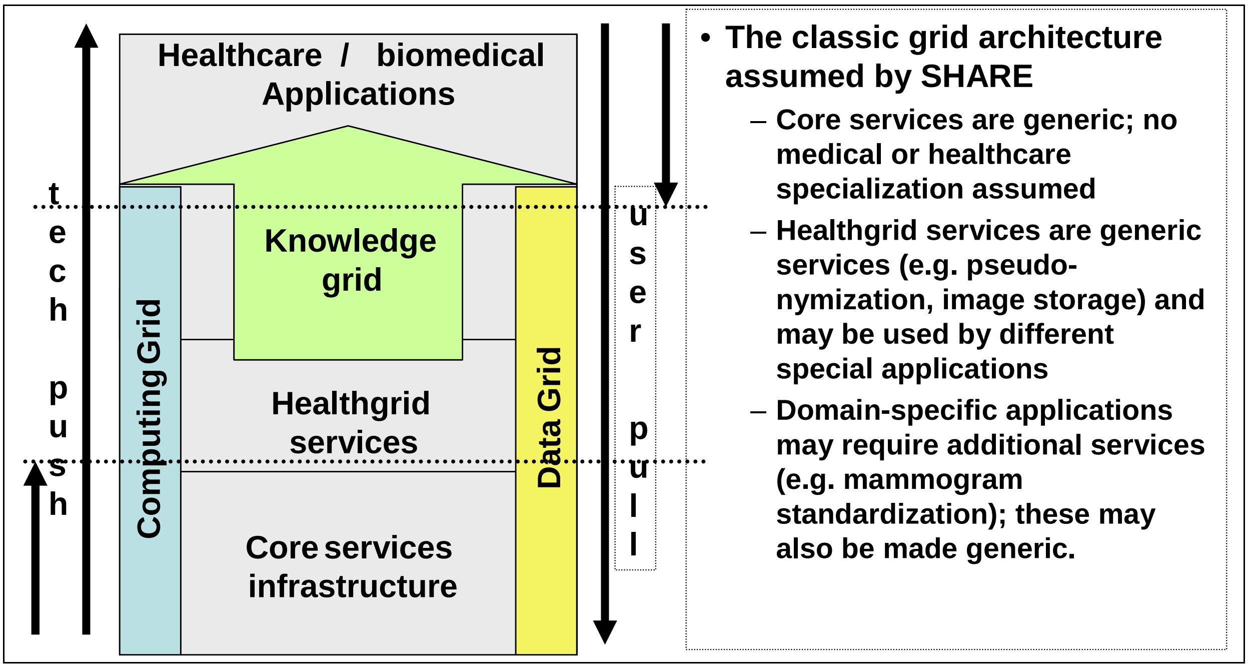
***Socio-Economic Sustainability*** D4.2 stressed that a key factor towards socio-economic sustainability is to ensure that healthgrids, as well as the services delivered over the grid infrastructure, respect the private interests of all stakeholders. More detailed steps towards that goal include appropriate economic and business analyses, accuracy and vigour of processes, user friendliness, and building of confidence. The move towards sustainability of healthgrids needs to be demand driven. Currently, the development of healthgrids is driven mainly by technology scientists rather than eventual users. As a consequence, the financial flows and other resource availability are based in the "wrong" field, when looked at from a long term perspective

***Sustainable Business Cases*** One of the main socio-economic themes requiring attention is the need for business cases for all stakeholders involved in developing and using healthgrids. This is critical because any stakeholder can de facto veto the whole process, and end-users can prove to be reluctant to fully endorse a new service that asks them to change their working processes. And the latter is essential for reaping benefits form healthgrids. No matter how advanced the technology solution is, if end-users do not see the benefits to them exceeding the costs and efforts, healthgrids will not have a future

## 8.5. CHALLENGES ON THE ROAD TO A KNOWLEDGE GRID

There is an underlying model implicit in the HealthGrid vision represented by the frame below. For some time, and throughout the SHARE project, the community has assumed a schematic architecture for healthgrid applications. This separates concerns into separate layers:

- A layer of infrastructure; at this level everything should be regarded as a resource.
- A middleware layer of generic grid services; there are not special to healthgrid.
- A layer of healthgrid services (e.g. pseudo/anonymisation or medical imaging).
- A layer of medical research and healthcare applications sitting on top of all these.
- At the same time, we take advantage of the image to reinforce the point that a knowledge grid assumes the achievement of computational and data grids, and implicitly of collaboration grids.



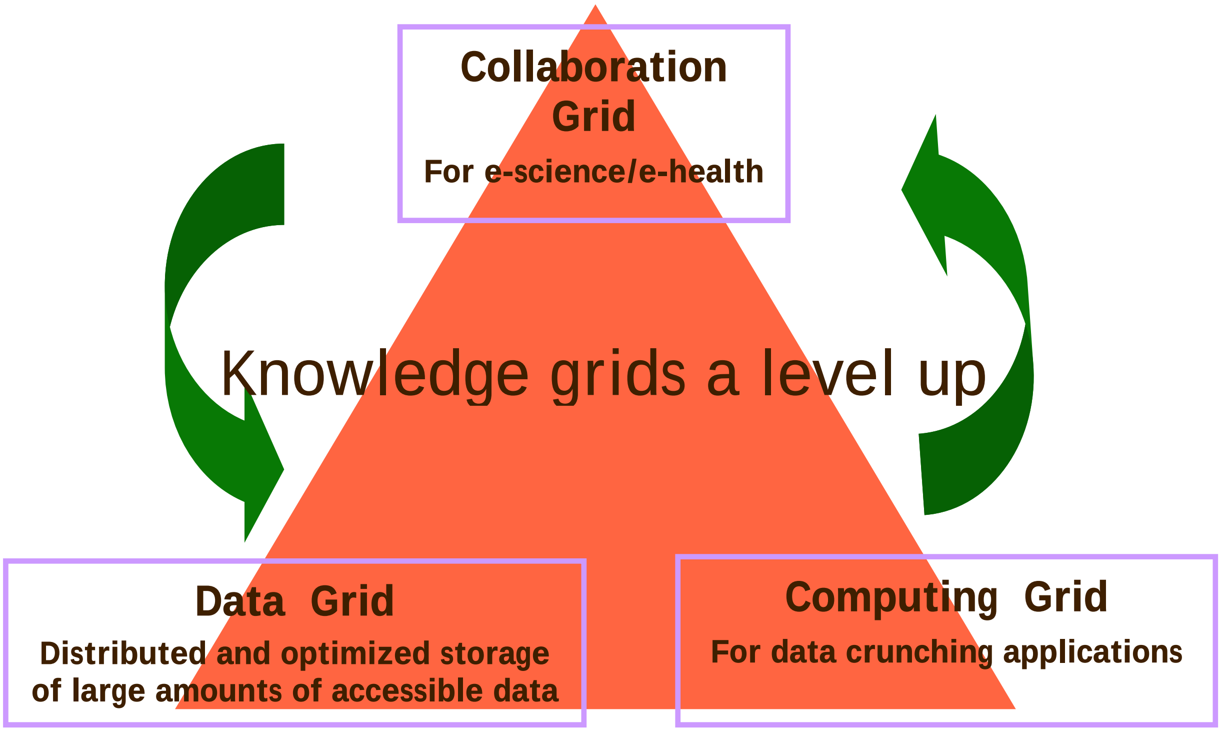
- The classic grid architecture assumed by SHARE
  - Core services are generic; no medical or healthcare specialization assumed
  - Healthgrid services are generic services (e.g. pseudo-nymization, image storage) and may be used by different special applications
  - Domain-specific applications may require additional services (e.g. mammogram standardization); these may also be made generic.

**Figure 3. The HealthGrid Stack**

Knowledge differs from information because it is generally understood to entail human agency. In the domain of computer science, it arises in a variety of forms: at the least demanding end, a ‘knowledge base’ is little more than structured information plus rules; at the most ambitious, it is claimed that systems exhibit intelligent behaviour based on stored knowledge. Information engineers would also recognise an intermediate form in the claims of ‘knowledge management’, where a restricted type of corporate knowledge is maintained through some form of collective memory system.

In the case of knowledge grids for a domain which itself has a body of knowledge, such as biomedicine or healthcare, we have to consider two different ways in which ‘knowledge’ may feature in the system. One is in the knowledge-based management of the grid, its resources and services. The second is management of the domain knowledge.

To begin with the latter first, the vague description of the more ambitious claims to the use of the term ‘knowledge’ in computer systems may be made a little more precise in our context. A distinction is often drawn between declarative knowledge ('know what') and procedural or operational knowledge ('know how'). In the domain of healthcare, both kinds occur. What is often referred to as 'the scientific basis' of medicine, that which must furnish the evidence in so-called 'evidence-based practice', is present in research publications to which different standards of credibility are attached. For example, research results based on a randomized, double-blind, controlled clinical trial are held to be the gold standard, provided they were also submitted to adequate peer review. Evidence based on one physician's own practice, although not negligible, would be considerably less reliable. On the other hand, 'best ways' of treating patients – in a particular context – are described in 'clinical care pathways'. It is not unreasonable to claim that declarative knowledge in medicine tends to be disseminated through peer-reviewed publication and operational knowledge through such things as guidelines and care pathways. In a healthgrid environment, these are brought together for the better treatment of patients and at the same time to improve research; indeed, the interplay between healthcare and research, e.g. through appropriately controlled 'secondary use', would be an important element in a full healthgrid environment.



Collaboration Grid

For e-science/e-health

Knowledge grids a level up

Data Grid

Distributed and optimized storage of large amounts of accessible data

Computing Grid

For data crunching applications

**Figure 4. Underpinning a Knowledge Healthgrid**

Among the 'knowledgeable' services a knowledge healthgrid would be expected to offer are – on the declarative level:

- understanding a scientific abstract, e.g. for the purpose of assessing its applicability to a particular case or to summarise;
- mine patients' healthcare records for specific or for unexpected patterns;

and – on the operational level:

- explore patients with a specific disease to assess the effectiveness of different care pathways;
- track 'variance' from pathways, where clinicians consciously depart from a pathway because they consider it inappropriate for the patient and/or context.

These services would be appropriately threaded together to provide the ‘knowledge server’ part of the highest level of healthgrid.

There is also scope, as we observed above, to use knowledge-based approaches in support of the grid itself. For example, we may consider how the grid optimises its resource allocation, how it effects its service description, recognition and subscription processes, how it negotiates interoperation with other grids, and so on.

Among the knowledge-based services needed to support such a healthgrid are:

- integrated data management by reasoning with metadata (provenance, paradata[6], computed and associated data, and metadata proper);
- semantic web services to identify the right service for an appropriately described task, or to thread in suitable workflow through a knowledge-based editor;
- semantic data sources and data objects to be matched to services, with or without the use of a workflow editor;
- design standards and tools for the provision of such functionality.

[6] Data about the data collection process, such as degree of (un)certainty.

# 9. RECOMMENDATIONS AND ROADMAPS

Objectives have been formulated in terms of milestones according to a number of key criteria:

- Is the proposed healthgrid essentially a computational grid, a data grid or a collaboration grid? Could it potentially develop into a knowledge grid?
- Is the necessary development to achieve any given stage likely to be delivered by generic grid research or is it particular to healthgrids?
- Is some prerequisite standard or other agreed framework necessary for the achievement of any particular milestone?

## 9.1. HEALTH RESEARCH CHALLENGES FROM 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 permit 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 technological complexity must be hidden from users. Grids are perceived as potential infrastructures in so much as their use does not require adaptation or acquisition of skills;
- The communities examined expressed the need for developing the technology for distributed data management, and while the usage of grids for distributed computing is perceived as available it is still very complex.

In the rest of this section, we have attempted to translate the requirements of three communities (epidemiology, innovative medicine and VPH) into a number of health research challenges and deployment milestones.

- The health research challenges are technical issues which need to be addressed in order for grids to offer services needed by the health communities.
- The health deployment milestones are health 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, such as the definition of agreed standards and ontologies in the research communities.

### *9.1.1. Health 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 (RCCG5) 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 (RCCG4, RCCG3).

- The need to access resources on clusters and supercomputers raises the need for interoperability between grid infrastructures (RCCG3). The transfer of jobs between infrastructures should also be made transparent to the user to ease his work (RCCG1).
- 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 (RCCG2) raises technical and political as well as financial issues as to who pays for operating the infrastructures.

| **Research challenge name** | **Description of the health research challenges** |
|---|---|
| RCCG1 | Automatic migration of simulations between different scales and platforms |
| RCCG2 | Capacity to access grid resources on demand, without previous agreement or request. European grid infrastructures should be freely accessible to European projects |
| RCCG3 | Capacity to submit jobs to cluster and supercomputer grids in a transparent way. Easy transfer of tasks between grid infrastructures |
| RCCG4 | Transparent access. The users should be able to ignore whether they are using one grid or the other |
| RCCG5 | User friendly access. Lower barrier to adoption. |
| RCCG6 | Real fault-tolerant scheduling systems |
| RCCG7 | Grid middleware that can be installed in health environments seamlessly and without requiring exhaustive maintenance and administration. |
| RCCG8 | 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 | Scalable job scheduling system |
| RCCG10 | Integration of resources with low latencies and high performance. |

**Table 1 Health Research challenges for a computing grid**

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

Dependencies for these challenges, grouped by key words, can be seen in figure 5.

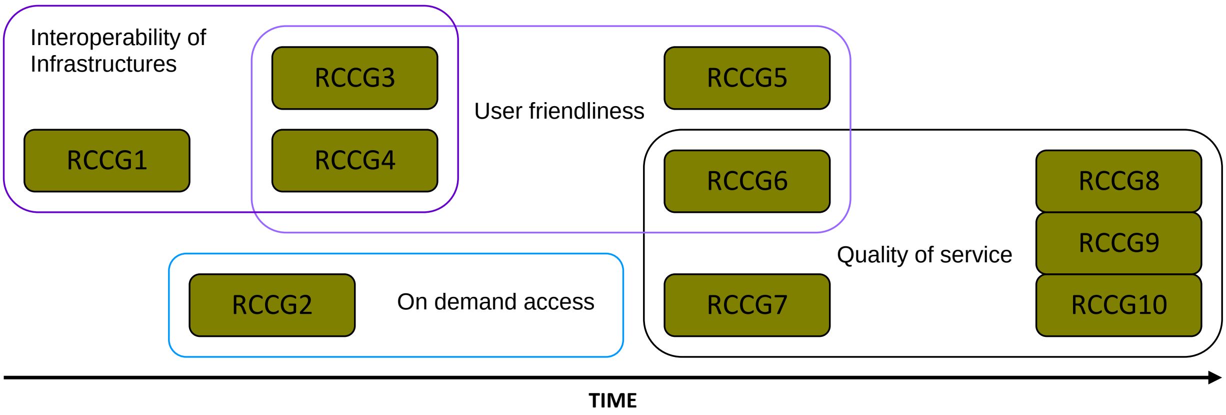


**Figure 5: Dependencies for computing grid research challenges**

Although largely arranged by level of complexity, certain milestones are prerequisites for others. For example, for true transparent access to multiple grids and infrastructures (RCCG4) and similarly the transfer of tasks between infrastructures (RCCG3), it will first be necessary to enable on demand access to grid infrastructures without prior agreement (RCCG3).

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.

### 9.1.2. Health research challenges for data grids

Table 2 presents the research challenges for a data grid (RCDG). Some of these challenges seem to be common to computing grids like the need for quality of service (RCDG5), including the availability of a robust middleware easy to install in health environments (RCDG4). But these challenges require different skills and content.

Some challenges are related to basic data management services which are still to be developed such as scalable data cataloguing and data transfer (RCDG1) as well as upload and download of large binary objects (RCDG2). Further developments include services to provide security in the management of the medical data (RCDG6) related to the adoption of standards (RCDG3).

The need for distributed data models (RCDG6, RCDG7) is also expressed.

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

| **Research challenge name** | **Description of the health research challenges** |
|---|---|
| RCDG1 | Scalable data cataloguing and data transfer. |
| RCDG2 | Storage services designed to ease the upload and download of large binary objects |
| RCDG3 | Develop enhanced standards for data protection in a web services environment |
| RCDG4 | Grid middleware that can be installed in health environments seamlessly and without requiring exhaustive maintenance and administration. |

| | |
|---|---|
| RCDG5 | 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. |
| RCDG6 | 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. |
| RCDG7 | Distributed data models and repositories adapted to the multiscale nature of the data needed and generated by the health community |

**Table 2 Health Research challenges for a data grid**

The dependency diagram for these milestones, grouped by key words, can be seen in figure 6.

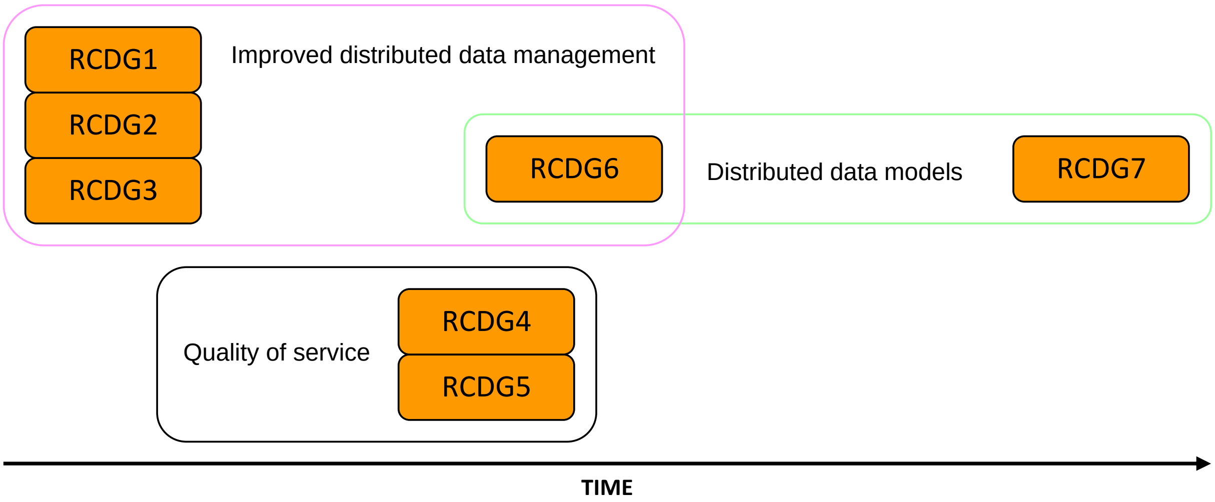


**Figure 6: Dependencies for data grid research challenges**

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 RCCG7 / RCDG4 and RCCG8 / RCDG5 respectively are 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 organisations and especially the critical control that (non-virtual) organisations 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. Figure 7 illustrates the overlap between computing grid and data grid milestones.

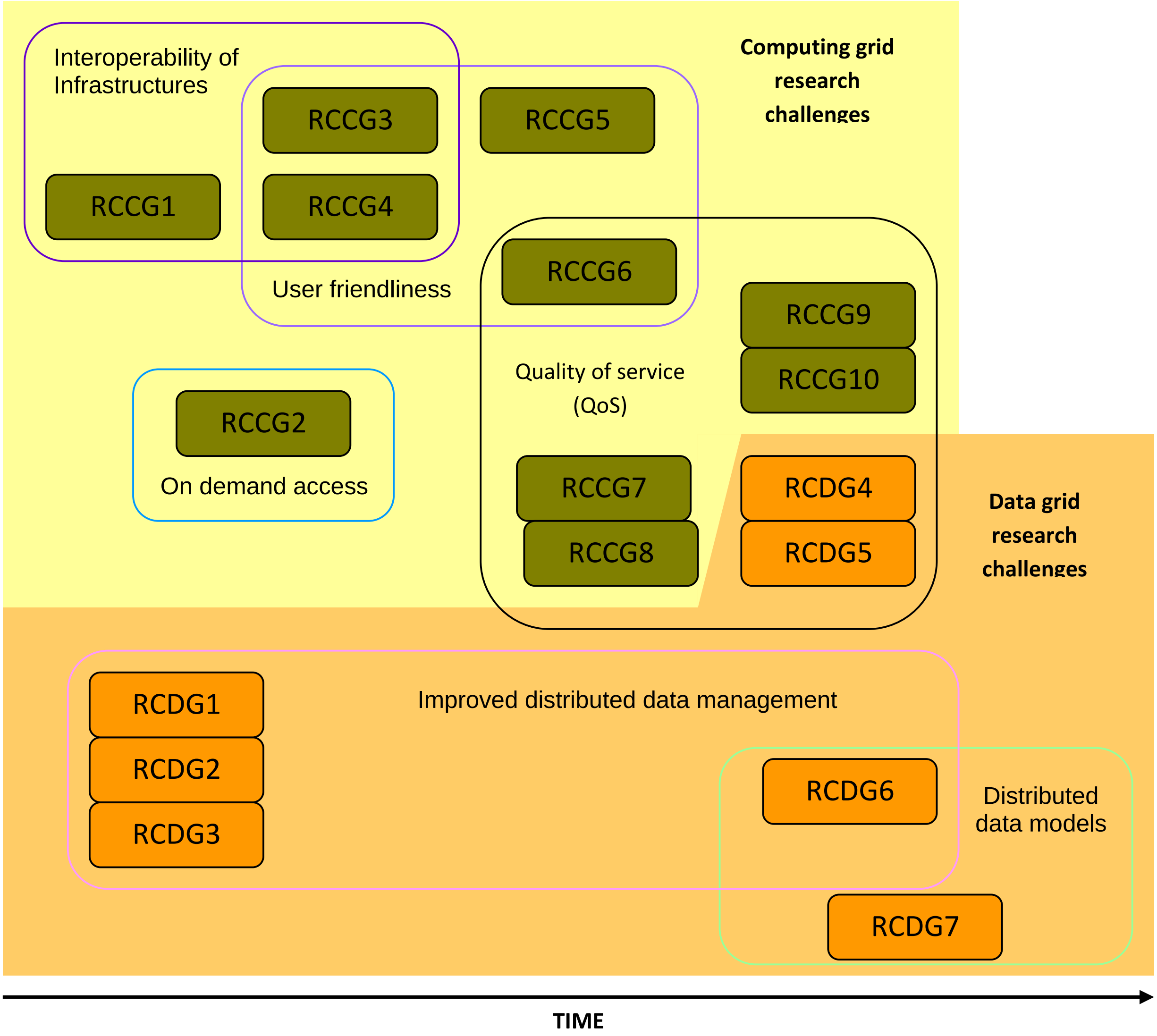


**Figure 7: Combined dependency diagram for computing and data grid research challenges**

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. 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 [50] and the EC-funded Health-e-Child project [16], 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.

### 9.1.3. Health research challenges for collaboration grids

Table 3 below catalogues the principal research challenges for collaboration grids, i.e. for grid services to support collaboration (RCLG). Biomedical research and healthcare are often highly cooperative, multidisciplinary activities, underpinned by informal as well as formal networks. While in some healthgrid projects collaboration has been built into the design from their very conception, in other cases the need for collaboration will arise in the same informal fashion as has arisen in the past. At the same time, many modern influences in medicine (e.g. evidence-based practice) have led to the definition of 'protocols' and 'care pathways' which may readily be recognised as workflows, thus providing a context for some collaborations. None the less, there is a good deal of scope for knowledge and technology transfer from heavily data-driven branches of e-science, where collaborative workflow engines have begun to be established.

| **Research challenge name** | **Description of the health research challenges** |
|---|---|
| RCLG1 | Migration of e-science workflow engines to biomedical research to encompass end-to-end processes, e.g. stages on the road from drug discovery to clinical trial. |
| RCLG2 | Natural mapping of healthcare/medical protocols to workflows for remote collaboration, education or quality control. |
| RCLG3 | Certification of medical workflows, complying with relevant legal and ethical obligations, to ensure they are reliable, validated and updated when required. |
| RCLG4 | Natural mapping of public health distributed decision support to facilitate coordinated action. |
| RCLG5 | Natural mapping of guidelines, protocols and integrated care pathways to validate practice against constantly updated evidence base. |
| RCLG6 | *Ad hoc* integration of heterogeneous sources of information where no prior coordination has been provided. Integration of different levels or modalities of medical data towards multidisciplinary diagnosis and treatment planning. The management of language issues. |
| RCLG7 | Workflow repositories to retain and maintain defined workflows and to enhance reuse, repurposing and recycling. Retain workflow histories and outcomes. |
| RCLG8 | Support for persistent collaborations, esp. in relation to rights management and participant privileges. |
| RCLG9 | Integration and management of workflows with implications in different domains, e.g. conflict between medical and ethical calls. |
| RCLG10 | A forum for the discussion of health/medical workflows, including provenance data, and a broader means of discussion and communication between collaborators. |

**Table 3 Health Research challenges for a collaboration grid**

Issues of collaboration arise in the context of diagnosis by different specialists, second opinion, treatment and surgery planning. Examples would include pipelining second reading or second opinion in breast screening; bringing in additional expertise if appropriate – e.g. staging a cancer; or quality control of the consistency of histopathology findings, by analysis of reports and checking them against guidelines. Monitoring in the public health domain may be concerned with MRSA-type epidemics or with avian flu or heat wave emergencies. All these call for a different kind of joint action, but all may benefit from decision support. More sophisticated epidemiology may be possible through analysis of associated data, as in the suggestion that avian flu passes more readily among genetically related individuals than among others despite close contact [51]. These suggested requirements would be satisfied through a combination of knowledge management and workflow management tools, linking the two where necessary. If the collaboration requires the sharing of data produced in different languages, particular care must be taken in the way these data are presented to healthcare professionals.

In another dimension of collaboration, the promise of modern biomedicine to relate genomic data to disease phenotypes, is being explored in such projects as Health-e-Child (HeC) and ACGT [51]. In HeC, there is a need to bring together information not only from different levels but also from different modalities, such as genome and imaging data. Thus collaboration here will also mean ability to coordinate different tools and modalities as well as integration of knowledge and data.

Finally, a further development is possible in the context of this discussion, to coordinate 'publication' of services and certification/licence issues.

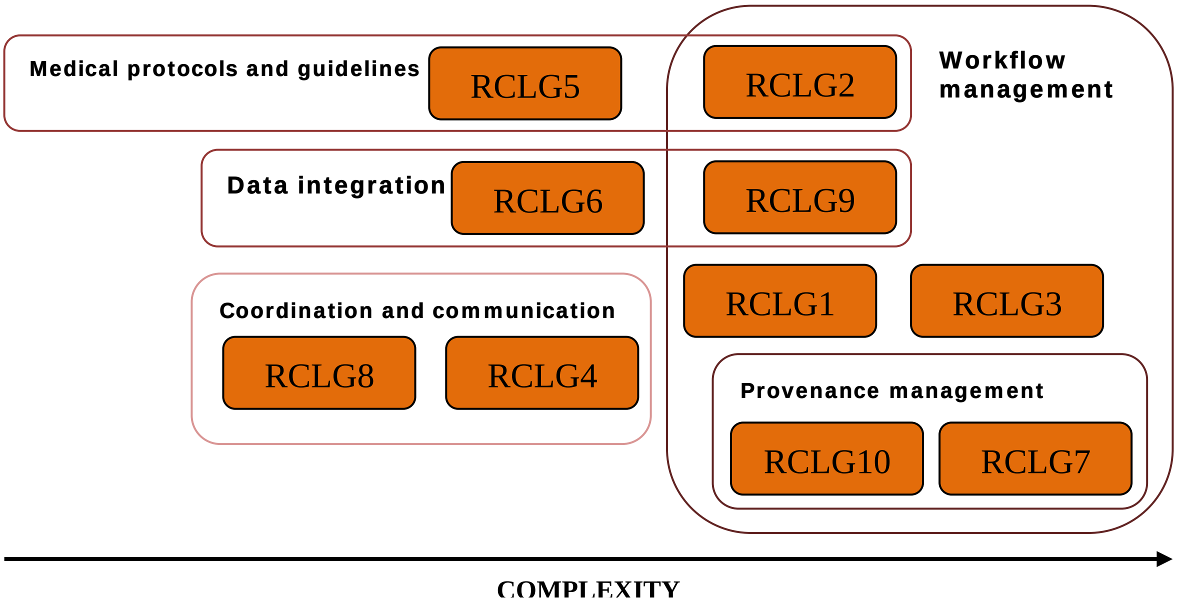


**Figure 8: Dependencies for collaboration grid research challenges**

### *9.1.4. Health research challenges for knowledge grids*

Table 4 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, RCKG5, RCKG6). Some challenges are directly related to the grid technology itself (RCKG1, RCKG2, RCKG3) while others are more relevant to the research area (RCKG4, RCKG5, RCKG6, 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 characterise 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 characterise the developments which are not specific to grids but are needed to enable a knowledge grid.

| **Research challenge name** | **Description of the health research challenges** |
|---|---|
| RCKG1 | Knowledge-driven grid catalogues and integration based on the metadata. |
| RCKG2 | 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 | Design standards for and build an expert tool (ontology/schema/rules negotiator) |

| | |
|---|---|
| | for exposing the properties of local sources in a federated environment |
| RCKG4 | Develop enhanced knowledge representation models and data exchange standards for complex systems, presently largely inconsistent or incomplete, looking for synergies with other initiatives |
| RCKG5 | Develop new, domain-specific ontologies, built on established theoretical foundations and taking into account current initiatives, existing standard data representation models, and reference ontologies |
| RCKG6 | 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. |
| RCKG7 | 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 |

**Table 4 Research challenges for a knowledge grid**

The dependency diagram for these milestones, grouped by areas, can be seen in figure 9.

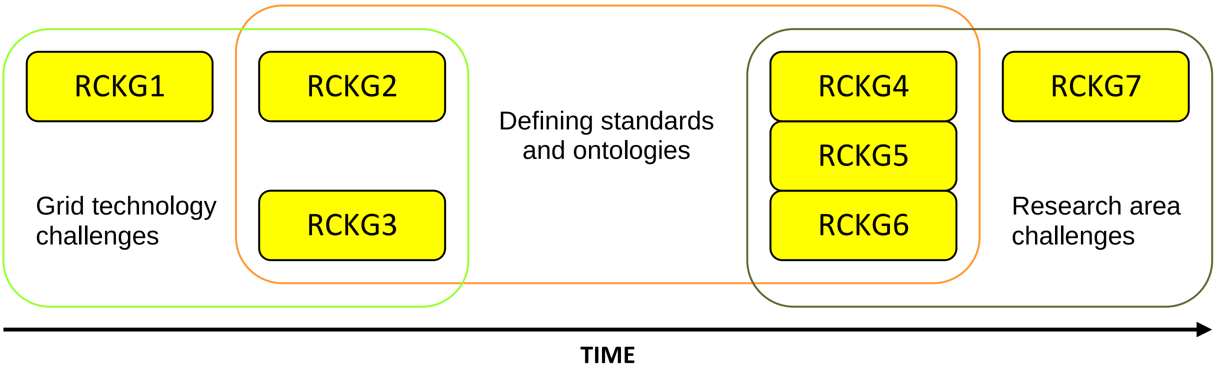


**Figure 9: Dependencies for knowledge grid research challenges**

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.

### *9.1.5. 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) 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** | **Description of the milestone** |
|---|---|

| | |
|---|---|
| MD1 | Need for successful pilot applications on epidemiology and VPH that will demonstrate the benefits of the technology to foster adoption of grids in the community and to identify limitations of existing infrastructures. |
| MD2 | 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 | Build a core reference database of validated experimental and clinical research data extracted from the literature |
| MD4 | Creation of disease-specific European Imaging Networks for establishment of standards, validation of imaging biomarkers and development of regional centres of excellence. |

**Table 5 Deployment milestones**

## 9.2. 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 characterised 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

Extending the model of figure 3, figure 10 below represents how research challenges address different layers of services from core infrastructure to applications. The following comments can me 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 and 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.

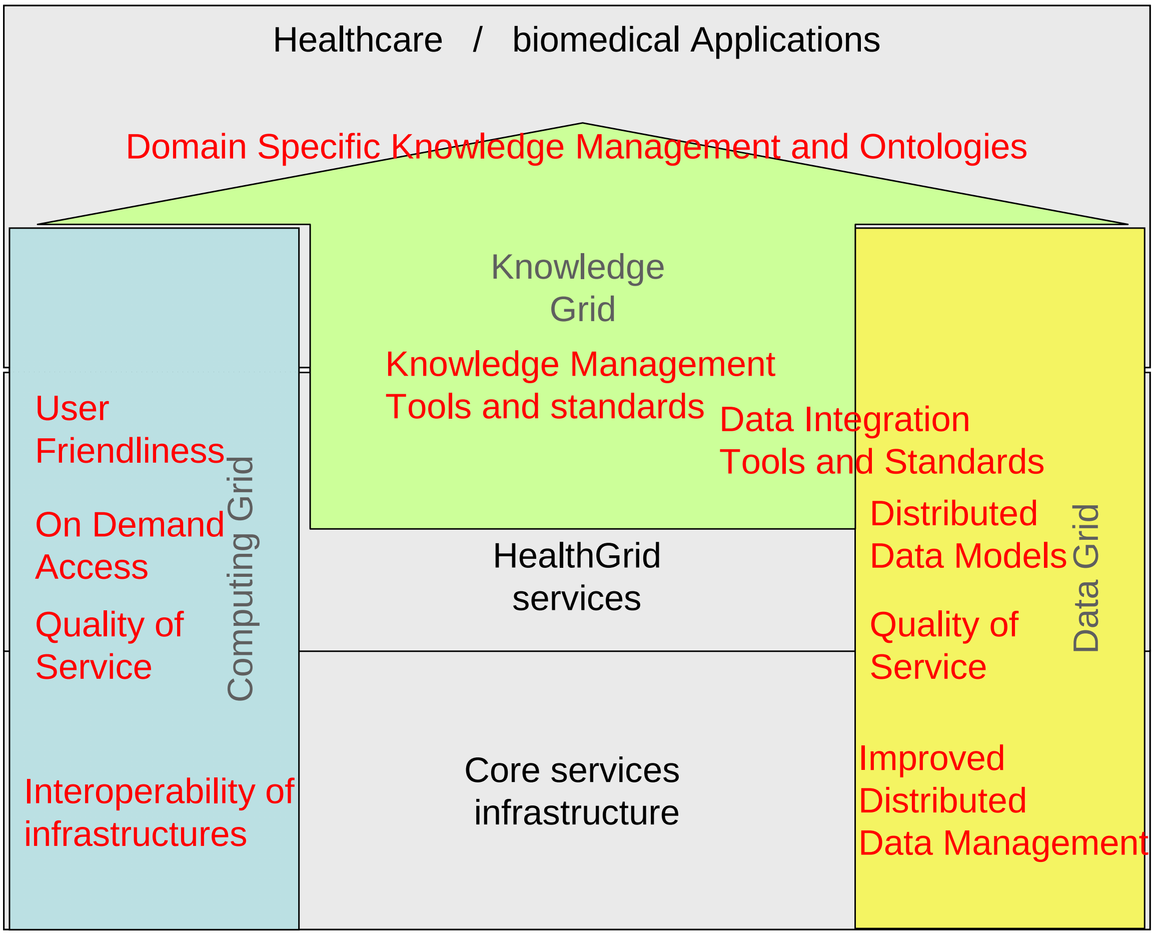

Healthcare / biomedical Applications
Domain Specific Knowledge Management and Ontologies
Knowledge Grid
Knowledge Management Tools and standards
Data Integration Tools and Standards
User Friendliness
On Demand Access
Quality of Service
Computing Grid
Interoperability of infrastructures
HealthGrid services
Core services infrastructure
Distributed Data Models
Quality of Service
Data Grid
Improved Distributed Data Management

**Figure 10: Representation of research challenges and healthgrid layers of services**

On the basis of this analysis, we have represented in figure 11 the research challenges according to their complexity and an estimated time when they should be overcome. The figure, inspired from the Innovative Medicine Case Study[7], also indicates the level of adoption by the research communities. As can be seen clearly from the picture, we identify several distinct roadmaps:

[7] SHARE deliverable D5.2a

- 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.
- collaboration grids are achievable with different levels of sophistication at different stages.
- 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.

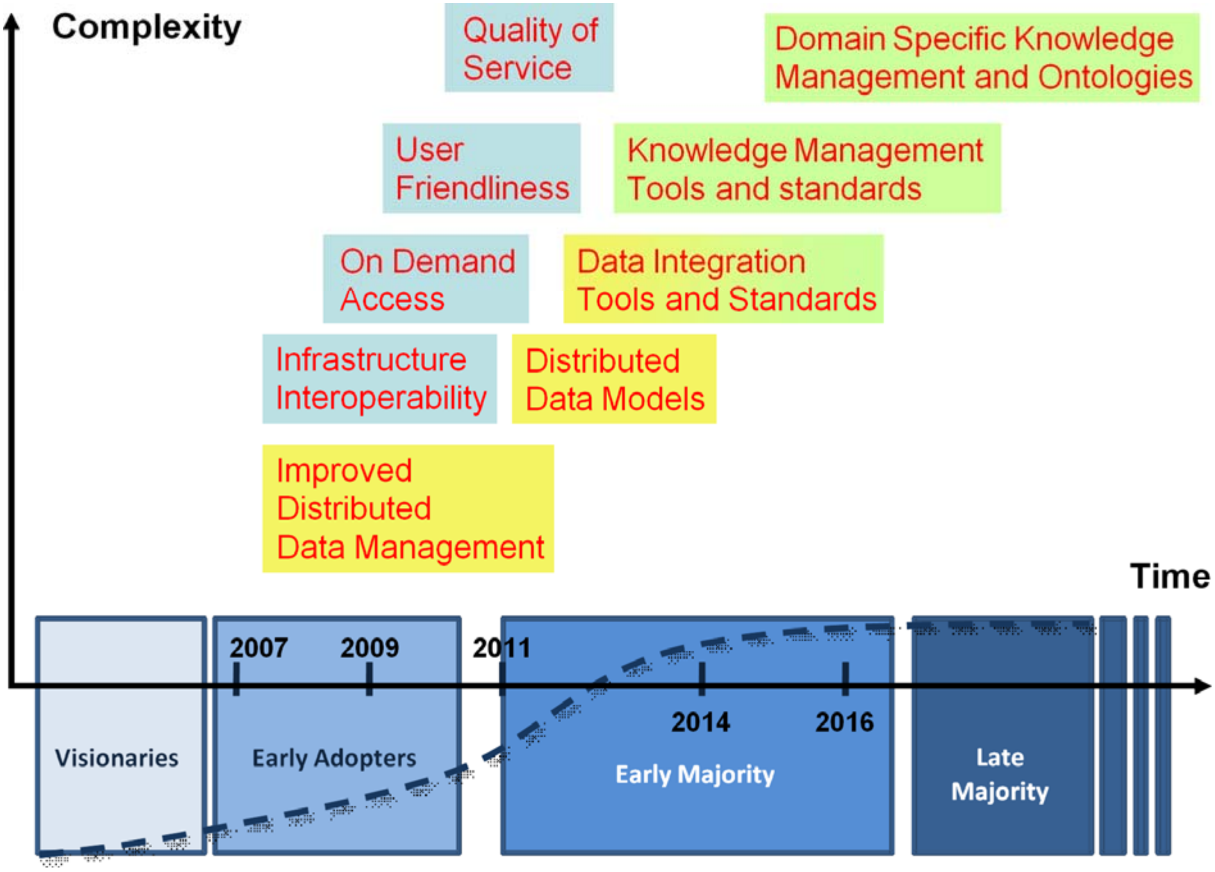

Complexity
Quality of Service
Domain Specific Knowledge Management and Ontologies
User Friendliness
Knowledge Management Tools and standards
On Demand Access
Data Integration Tools and Standards
Infrastructure Interoperability
Distributed Data Models
Improved Distributed Data Management
Time
2007
2009
2011
2014
2016
Visionaries
Early Adopters
Early Majority
Late Majority

**Figure 11: Research challenges as a function of time and complexity**

This model arises in the field of innovation studies, and distinguishes between: *visionaries*, pioneers both in research and applications; *early adopters*, who recognise the potential for rapid benefits and take up the technology quickly, often introducing further innovation; *early majority* who incur relatively little risk in adopting the technology; and *late majority*, those who take virtually no risk and finally adopt a technology because there is no other option.

The diagram depicts priorities: even for early adopters, infrastructure interoperability and distributed data management are already necessary; on demand access, usability and quality of service are at the first point of inflection, before rapid expansion, while with sophisticated AI tools in the later stages, a second inflection occurs and the technologies become routinely accepted.

## 9.3. MAPPING ELSE REQUIREMENTS:

### 9.3.1. Liability Issues

The current state of EU legislation does not cover liability issues that are specific to healthgrids. The following tasks could help minimise liability concerns for healthgrid usage.

- Analysis and prediction of risks and possible damage to patient health and privacy should begin from the outset
- Outlining examples of liability concerns specific to the use of healthgrids in order to encourage the introduction of new legislation and policies
- Good testing strategies for services and products, including testing for product safety
- Good quality assurance for services and products
- The use of standard techniques for detecting bugs and faults while dealing with infrastructure interoperability.

### 9.3.2. Trust and Acceptance

- Pilot projects and prototype applications to demonstrate the use of grid services in clinical and research workflows
- The use of trust based technologies to increase the trustworthiness of grid infrastructures
- Providing feedback and documentation to provide users with clear answers to any security concerns.

### 9.3.3. Cost and Benefit Estimation

- Ex-ante analyses over time, based on initial pilot experiences. These have to focus on ensuring acceptance, technical and regulatory certainty, and sufficient private incentives in the steps to follow

- Analysis to estimate potential net benefits (i.e. expected benefits less expected costs over time), accounting for different risks and for optimism bias in estimations. Such studies will facilitate access to initial funding, but can also be beneficial in the necessary dissemination work with the health sector.

### 9.3.4. Data Protection

- Ensuring the use of standard means of data security within the different data management systems of participating infrastructures (encryption, anonymisation, pseudonymisation, access control, etc.)
- Use of data quality assurance mechanisms
- Adopting formal ways to audit the regulatory compliance of operational level controls
- Automation of the collection of patient consent, and ways to allow patients to opt in, opt out and withdraw consent
- The use of evolving privacy enhancing technologies.

### 9.3.5. Sustainability Guarantees

- The development and deployment of data grids will benefit from more focused prospective assessments of the socio-economic impact in order to identify existing and potential barriers

- Convincing business cases ensuring sustainability.

- An organisational milestone can be defined here in the move from technology/science towards service provision. By that stage, a notable amount of legal and regulatory certainty has to be achieved, so that private incentives can be assessed and adjusted (including via government intervention) if necessary.

### 9.3.6. Education and New Skills Requirements

- Training and educational programs to increase users' confidence in the use of healthgrid products and services
- Adequate documentation and guidance must be available, and where grid infrastructures are distributed amongst geographically remote sites there must be sufficient communications methods such as video conferencing to ensure problems and concerns raised during deployment are dealt with quickly and efficiently
- Investing in technical staff within hospital and research centres to provide help with technical problems.

### 9.3.7. Intellectual Property

- There is a contradiction between intellectual property rights and the needs of grid technology, which will require that access to databases, knowledge and software is free of rights. Contract law and agreements could be an option to regulate the IP issues related to knowledge integration, ontologies and software reuse

### 9.3.8. Governance and Delegation

- The working document on the processing of personal data relating to health in electronic health records (EHR**)** recommends that in the case of health care systems that adopt a decentralised data storage model, it could be necessary to appoint one central body to be responsible for steering and monitoring the whole system and also for ensuring the operation of the system is compatible with data protection. It would also be useful if data subjects could address their data protection queries to a central body instead of having to search and identify the relevant controller among many. The architecture of a healthgrid system is similar but even more complex than a distributed system. The idea of one data controller might be preferable but more challenging. A discussion and analysis highlighting the main issues and benefits surrounding the idea of a unique data controller for data stored within a healthgrid domain needs to be produced. Linking this to the technological component of the roadmap, this could impact on the process of granting permission to access the data.

### 9.3.9. Policies and Codes of Conduct

- Discussions should take place between different healthgrid stakeholders to decide on the importance and benefit of applying for new legislation to address healthgrid related legal and ethical issues
- Once a decision is reached, a framework will need to be distributed to legal bodies showing why healthgrid services and products should be considered different from other marketed products. It could also present scenarios showing that the current legislation does not ideally cover these issues
- Evolving technologies for the automation and enforcement of policies at different infrastructure layers should be explored.

### 9.3.10. Dissemination and Publicity

- Dissemination and publicity programmes need to precede the deployment of knowledge grids. This includes workshops, conferences, and magazines to attract the user community and build awareness of healthgrid facilities

- Demonstrate the effectiveness of grid applications in providing healthcare and research services while preserving the users' autonomy.

### *9.3.11. Ethical Control and Auditing*

- In the UK, every health organisation is now required to have a privacy and data protection officer, the so-called "Caldicott guardian". The establishment of similar roles throughout Europe would be a major step towards harmonisation of ethical practice and compliance in the member states. This may be supplemented by the creation of a Europe-wide ethical body composed of these European Caldicott guardians, although there is a question about how this would relate to the article 29 working party. Operating at a healthgrid level, they would be able to judge matters in a European context. This would in a sense provide a useful bottom-up approach to confidentiality and privacy protection across healthgrids, as opposed to top-down European directives
- Before the deployment of Data grids should start, the requirements for ethical oversight and monitoring should be determined, and the automation of oversight facilities should be explored.

### *9.3.12. Benefiting from the Technology*

Despite the best efforts of policy makers to structure and control the use of personal data (in this case patient data), incidents of identity theft, identity base fraud, and the sale and misuse of data are still occurring, as highlighted by recent stories in the media. This may in part be due to a lack of enforcement of high-level legal obligations concerning personal data, and also as a result of the variability of privacy laws due to cultural and national considerations.

The recent push to enforce legal rules within enterprise infrastructures and business processes has initiated several EU and international projects, each aiming to help enterprises, organisations and governments to benefit from the use and exchange of personal data without compromising individual privacy. As a result, many privacy-enhancing technologies have been developed which have proven efficiency across many domains including e-health. The possibility of deploying similar technologies on a grid infrastructure should therefore be investigated.

#### *9.3.12.1. Privacy-Enhancing Technologies (PETs)*

The approach to privacy typically taken by more advanced systems involves enforcing privacy policies at the application level; the filtering of sensitive data before providing the query result to the user. But this could still reveal enough information for an intelligent person to identify individuals. For better protection of the data, access control policies have to be enforced at the data level. Traditional databases provide access control at the table level and use the view mechanism to restrict access to certain columns or rows of the table, but this is still inadequate. Hippocratic databases [53] provide a more advanced "limited disclosure" approach. They permit enforcement at a very fine level of granularity. Privacy policies could be enforced at the level of an individual cell in a relational table. Hippocratic databases also allow privacy policies to be stored and managed in the database as metadata.

Sticky policies [54] have emerged as one approach to enhance privacy preservation in distributed computer systems. The underlying notion behind *sticky policy enforcement* is that the policy applicable to a piece of data travels with it and is enforceable at each point it is used. Recent work done by the IBM Almaden Research Laboratory has improved this approach, making it adequate for the needs of a healthcare environment.

They have added new functionality to handle data disclosure to a party with well-defined constraints that allows data to be released to less privileged parties without requiring the originator's involvement.

A crucial task prior to the disclosure was identifying the applicable privacy policy constraints for the document(s) to be shared and sticking them together, forming a single entity for transfer. This avoids the potential pitfall of having to contact a (potentially) large number of third parties before making a decision to disclose a specific piece of information.

The PRIMA (PRIvacy Management Architecture [55]) System was developed by the IBM Almaden research lab in order to exploit *policy refinement* techniques to gradually and seamlessly embed privacy controls into clinical workflows based on the actual practices of the organisation in order to improve the coverage of the privacy policy. PRIMA attempts to improve policy coverage by gradually embedding new policy statements, which were discovered through the process of policy refinement, into the clinical system. Stakeholders define the privacy policies, which are embedded in *privacy controls* that are integrated into the clinical environment. One of these privacy controls is an auditing function that automatically generates entries for the system's audit logs. These logs are either periodically replicated or PRIMA-enabled by the construction of a consistent, consolidated view of them. In the simplest case, there is just one log. At regular intervals or at the request of stakeholders, the *policy refinement* component extracts input from the *audit management* component and the *privacy policy definition* component and outputs a list of definitions, if any exist, that should be included in the policy definitions.

Enterprise Privacy Authorisation Language (EPAL) [56] was designed to enable the translation of privacy policies into an XML based computer language. The resulting coded translation of human policy into information technology policy allows complex descriptions of the internal data handling practices needed for enforcing the privacy policy. The expressiveness of EPAL has been tested against a set of "real world" scenarios such as of the Ontario Freedom of Information and Protection of Privacy Act (FIPPA). This has demonstrated the effectiveness of EPAL in

- Linking access control to natural text policies
- Creating precise, fine grained description of the policy
- Enabling complex, context driven conditions on policy rules
- Creating portable, reusable policies
- Allowing for sector/legislation specific policy vocabularies
- Enabling policy negotiation

### 9.3.12.2. Trust-Based Technologies

To qualify as a trustworthy system [57], the healthgrid security infrastructure needs to adopt technologies that are able to fulfil the following needs of users and resource providers. Before sending the job request, the user needs to:

- Know whether the resource provider host in the resource provider domain (to be visited) is "trustworthy" in terms of faithfully executing the user code and completing the task
- Know whether the resource provider host(s) will have enough trust in the user to cooperate with them (i.e. a *code trust* question involving *trust symmetry* problem). In many cases such a trust relationship is often implicitly assumed
- Ensure that the resource provider host(s) will not tamper with the user code and/or computation result

Before running the job request on the resource provider node, there can be two *code trust* questions that the resource provider node should ask:

- Is the job requesting user trusted to produce benevolent and competent code that will not harm the grid?
- Has the user program been tampered with before it is allocated?

After completion of the job result:

- Both user and resource provider(s) need to update their relevant trust relationship knowledge
- The user needs to check the integrity of the completed job or result to update its *execution trust* with the resource provider(s). Resource providers need to update the *code trust* for the user.

PETs have shown they can be efficient when deployed across a variety of domains. However, each technology typically only deals with a stand-alone privacy or security issue. In order to optimise patient privacy protection, many PETs will need to be integrated into a single privacy framework.

The PRIME project[8], an EC FP6 project, aims to reconcile privacy and accountability of users' electronic interactions in Europe. The project addresses these goals by providing an architecture integrating several privacy enhancing technologies and emerging systems that include human-computer interfaces, ontologies, authorization and cryptology, anonymous communication, privacy-enhancing identity management architecture, and assurance methods. The PRIME project also recognised the need for solutions to be compatible with the existing legal framework in order for them to have real world relevance. Therefore legal requirements were considered from a very early stage, and the PRIME solution integrated legal rules to be more efficient and to form a "privacy-protecting framework that has a real impact on business practices".

[8] See https://www.prime-project.eu/

# 10. CONCRETE RECOMMENDATIONS

## 10.1. TECHNICAL RECOMMENDATIONS

It is important that technical research and development be conducted in close collaboration with user communities. At certain stages it must be driven and validated by user groups, although there is always scope for innovators to introduce unforeseen possibilities to users. The research communities involved in the definition of the roadmap expressed their interest and support for the deployment of prototypes and test cases on existing grid infrastructures. We recommend that these infrastructures and tools continue to be made available to applications requiring computing services and data management.

Indeed, some projects are already using the DEISA and EGEE infrastructures for scientific production in the fields of epidemiology, medical imaging and drug discovery. However, these initiatives come from pioneers and are not sufficient to achieve a wider adoption in these research communities. We recommend that:

- More attention be paid to such initiatives so that they may influence the evolution of the technology to make it better fit the needs of the community;
- Two projects within the framework of the EuroPhysiome initiative be identified that could directly benefit from the computing and data management resources of the EGEE and DEISA infrastructures; these should be deployed in parallel on the two infrastructures in order to investigate interoperability issues and identify bottlenecks.

In terms of encouraging biomedical applications to fully exploit grids, we recommend:

- Linking certain advanced health domains to an e-science infrastructure;
- The adaptation of epidemiology data sources to grid models and grid-enabled gateways to epidemiological data, using medical informatics-related connectors such as HL7, DICOM, ENV13606, or similar.

In the same spirit, in order to foster the uptake of grids in the biomedical research and healthcare communities, we recommend:

- The release of open-source components for medical data interfacing;
- Building a core reference database of validated experimental and clinical research data extracted from the literature in innovative medicine and to explore whether a grid infrastructure could support this activity;
- The creation of disease-specific European imaging networks towards the establishment of standards, validation of imaging biomarkers, development of regional centres of excellence in innovative medicine and exploration of grid infrastructures to support such activity.

We recognise that there are a number of concerns (for example: security and standards) in which problems exist irrespective of the use of grids. It is important to understand the nature of these problems and the extent to which the use of grids complicates them. Results could be concrete implementation recommendations (for example: security improvement) and a suggested list of health applications requiring security which may be able to be deployed on a grid. In the field of standards, we believe that the HealthGrid initiative provides the right framework to coordinate the development of the different standards in collaboration with the OGF and the various medical informatics standardisation bodies. We recommend:

- The active pursuit of standards for the sharing of medical images and electronic health records on the grid within the already existing medical informatics standardisation bodies;
- The active pursuit of ontology matching and development for healthgrids. A survey of existing

biomedical ontologies (disease, phenotype and genotype ontologies) would be extremely useful to understand the status of research in this field and its maturity. Attempts at gathering open source ontologies like the Open Biomedical Ontology (OBO) foundry are headed in the right direction.

We believe that technology transfer between EC projects should receive more prominent and active encouragement. In particular, we recommend:

- The commission implements collaboration measures in the funding mechanism for projects;
- Targeted capacity building so that projects may access grid resources on demand, without previous agreement or request; European grid infrastructures should be freely accessible to European projects;
- Porting of one or two biomedical grid applications, already successfully deployed on grid infrastructures, to e-science environments using OGSA-compliant grid toolkits.

Finally, to return to a frequent theme in this analysis, we recommend:

- The encouragement of cross community interaction, in order to build meaningful dialogue between grid developers and health researchers.
- The set up of training programmes for staff and students in the biomedical field. A European summer school on biomedical grids is already taking place yearly under the initiative of the BioinfoGRID project. We propose turning this summer school into a yearly event, and the organisation of tutorials as satellite sessions to the main conferences in the field.

## 10.2. LEGAL RECOMMENDATIONS

***Liability in a healthgrid system*** Using grids blurs the liability issues in terms of medical practice. A stepwise approach should therefore be taken to develop the liability framework, distributing legal responsibility appropriately across healthgrid users. Such an approach would help to favour the reliance on the system while providing legal certainty for all stakeholders, including patients. Moreover, the European Commission should consider supporting the adoption of EU level guidelines that would identify the various parties involved in delivering healthgrid services and annex services and establish the various liabilities that each party must accept. Such guidelines should be widely disseminated in order to develop users' confidence in the use of healthgrids in general. In particular it should be investigated whether specific guidelines on those specific services could be drafted under the provisions for a code of conduct established in directive 2000/31 on eCommerce [58].

***Product safety*** As mentioned in D4.2, in the framework of the European level legislation applicable to product safety, national authorities have been established to monitor product safety and to take appropriate measures to protect consumers. Under these circumstances, an information system has been put in place that imposes collaboration between distributors, producers and the national authorities but also between member states and the European Commission (RAPEX) [59].

At present, this system is not used at all for products used in the composition of grid systems. The European Commission should thus adopt policy tools encouraging the use of the RAPEX system for such products.

***Healthgrid as a medical device*** As outlined in the introduction to this document, the law on medical devices is very unclear with respect to healthgrids. While it may be argued that a healthgrid could fall within the ambit of the current medical devices directive [48] in that it is a software tool that impacts on a medical act, the whole construction of the directive is based upon physical goods (which might have a software component) that are placed on the market for purchase or lease. In this situation, many of the currently available monitoring devices are covered only by general product liability, but not by specific liability provision.

In this framework, special guidelines should be issued in order to clarify the application of medical devices legislation to specific tools used in healthgrids.

***Patient consent*** In February 2007, the European working party on data protection, established under article 29 of the directive issued a working paper looking at the applicability of data protection legislation to Electronic Health Record (EHR) systems [33]. In its report, the working party noted in particular the limitation of the use of consent to permit the processing of heath data. The working party notes that if processing health data in an EHR system is the primary way of processing health data in a given health system, then a patient's care may be compromised if he or she opts-out of such a system by not giving his or her consent to the creation of an EHR. Accordingly, consent should not be used as it cannot be said to be truly and freely given.

The remaining provisions setting aside the general prohibition on article 8 of the directive 95/46/CE [62] can also be said to pose some problems – notably the idea that a patient ought to know the full finality of the use of data before his or her data may reasonably be used. But, as noted by the data protection working party there are some problems in using consent as a valid basis for processing data in eHealth applications. Indeed, if the creation of, for example, electronic medical records is a necessary and unavoidable consequence of the medical situation, withholding consent may be to the patient's detriment.

***Specified and explicit purposes*** According to the data protection directive, data may only be collected for specified and explicit purposes. If healthgrids can be used for risk detection, disease monitoring and preventive care, legal guidelines should be established that clarify the circumstances in which professionals can make further use of personal data related to health in the interests of public health. Such guidelines should allow for secondary uses even where such uses could not have been foreseen at the time of data collection.

***Technical and organisational security measures*** Efforts should be made to harmonise national standards on the technical and organisational measures of data security. While the data protection directive calls for such standards to be adopted, little has been done at a regulatory level to harmonise guidelines across the EU.

***Intellectual property rights*** It might be desirable for the commission to develop guidelines for the use of open licensing and open standards, which could address the tension between the intellectual property rights of developers and the needs of the grid technology. Such an open standards software approach could then be a solution to help the development and implementation of healthgrids.

On the other hand, the use of healthgrids in the drug discovery sector raises the issue of the ownership of both methods used to discover the medicines and the results achieved. Indeed, all the grid nodes that contribute resources to compute the docking probabilities could claim some ownership of the results and the designers of the software used in the process would certainly be in position to claim ownership of the method. In this context, one may ask whether it is important to know, say, which grid node was the one to identify a particular candidate molecule.

In this context, it is of essential interest, notably in patents, to determine guidelines that would determine, in case of collaboration in the research, what every actor is entitled to according to his contribution to the system.

***Privacy policies and codes of conduct*** As suggested above, a directive or code of conduct on privacy and health information infrastructure should be developed within the context of directive 95/46/EC and could take the form of either a dedicated directive or could be an EU-level code of conduct to be approved by the European working party on data protection set up under article 29 of the directive. This could help to solve the problem of data processing legitimacy. In particular, it could provide possible bases of legitimacy other than the data subject's consent. It could also provide the following solutions:

- Appropriate safeguards to allow for the further processing of personal data (and especially of medical data) for substantial public interests (without requiring the data subject’s consent) like scientific research. An example of appropriate safeguard would be a first coding by the initial data controller and a second coding by a trusted third party gathering all the data from the data controllers before sending them to the researchers,
- Appropriate safeguards to allow keeping the data for longer periods for scientific use; terms under which identification numbers or other identifiers may be used; terms under which (coded) personal data may be transferred to third parties for scientific research.

### 10.3. SOCIO-ECONOMIC RECOMMENDATIONS

***Trust and acceptance*** Trust is a very important element in any interaction between the different members of a society. In the market context, trust is crucial for successful business to business collaborations. Similarly, in a healthgrid domain a good collaboration will not be achieved unless a trust relationship exists between the different users and stakeholders. Pilot projects and prototype applications, which are an inherent part of the technology roadmap, need to be future oriented in the sense that the ultimate routine operation users have to be persuaded both of their value and their applicability, i.e. their ability to fit into real clinical or research workflows. This has to be taken seriously from the very beginning, even in proof-of-technology demonstrators: the goal should always be to give users, especially clinicians, tools that they would consider using with patients in real healthcare situations. Trust and acceptance can be greatly enhanced by the establishment of appropriate ethics committee structures to advise on the observance of ethical principles.

***Estimation of costs and benefits*** Ex-ante analyses over time, based on initial pilot experience, have to focus on ensuring acceptance, technical and regulatory certainty, and sufficient private incentives in the steps to follow. An inherent part of such assessments should be to estimate potential net benefits (i.e. expected benefits less expected costs over time), accounting for different risks and for optimism bias in estimations. Such studies will facilitate access to initial funding, but can also be beneficial in the necessary dissemination work among the health sector.

***Sustainability guarantees*** Work towards achieving the next milestone in complexity – data grids – will benefit from more focused prospective assessments of socio-economic impact in order to a) identify already existing, as well as potential barriers, and b) build convincing business cases ensuring sustainability. The analysis of alternative resource allocation options from a societal perspective, but also on organisational level, becomes necessary.

An organisational milestone can be defined here in the move from technology science towards service provision. By that stage, a notable amount of legal and regulatory certainty has to be achieved, so that private incentives can be assessed and adjusted (including via government intervention) if necessary.

***Cross-organisational interoperability*** The effective deployment of knowledge grids will crucially depend on collaboration between institutions, meaning more than “simple” access to each others’ data and computing resources. This collaboration requires the utilisation of human resources and in some cases a significant strategic re-orientation and re-organisation of working processes and even management structures. As the health sector, including clinical research and public health, is (and should be) highly regulated, policy makers on regional, national, and EU level should review the existing regulatory framework against the requirements arising from the exploitation of knowledge grids. Particular attention should be given to flexibility of government regulated budgets and reimbursement schemes. The latter should encourage cross-organisational collaboration, including such beyond national borders, by means of using knowledge grids.

# A Appendices

## A.1 Expert Workshop

The following scientists, clinicians and experts in ethical, legal, social and economic issues participated at a workshop in December 2007 organised to provide a critical review of a draft final road map document:

| | |
|---|---|
| Prof Roberto Amendolia | Scientific Attaché, Italian Embassy in London, Italy |
| Dr John Brooke | University of Manchester, UK |
| Prof Iain Buchan | Northwest institute for Bio-Health Informatics, UK |
| Dr Joan Dzenowagis | World Health Organization, Switzerland |
| Dr Fabrizio Gagliardi | Microsoft Europe |
| Prof Martin Hofmann-Apitius | SCAI Fraunhofer, Germany |
| Dr Martin Huber | Siemens, Germany |
| Prof Julian Jenkins | Merck Serono International SA, Switzerland |
| Dr Tom Jones | Tanjent, UK |
| Dr David Lam | TATRC, USA |
| James Lawford-Davies | Clifford Chance, UK |
| Dr Keith McCormack | Medical Physics, University of Sheffield, UK |
| Dr Isabel Muñoz | Conselleria de Sanitat, Generalitat Valenciana, Spain |
| Prof Giacomo Pongiglione | Giannina Gaslini Institute, Italy |
| Prof Simon Rogerson | De Montfort University, UK |
| Dr Louis Schilders | Custodix, Belgium |
| Dr Simon Shiu | HP Labs, Bristol, UK |
| Dr Peter Singleton | Cambridge Health Informatics, UK |
| Dr David Wallom | Oxford e-Science Centre, University of Oxford, UK |
| Dr Pieter E. Zanstra | Kermanog, The Netherlands |
| Prof Pedro Zapater Hernández | Clinical University Hospital of Alicante, Spain |

The Expert Workshop draft road map received a number of comments, concerning the output of the project itself as well as suggestions for further work in healthgrid research and the wider health informatics domain.

### A.1.1 Additional Roadmap Use Cases

It was felt that the final report would benefit from the addition of a use case dealing with public health and another dealing with general healthcare, perhaps involving a scenario where a patient undergoing treatment moves between different regions. As a result, a use case for a surveillance network for avian flu was added (4.3.2), as was a general healthcare use case for Chronic Obstructive Pulmonary Disease (COPD, 4.5). The addition of the avian flu use case addressed the concerns that use cases should include scenarios involving multiple stakeholders, and that complex collaborations should be represented. The COPD use case included the travelling patient scenario, and also showed the potential for direct benefit to patients, addressing other concerns mentioned at the workshop.

### A.1.2 Observations

Some of the participants at the expert workshop felt that ‘road blocks’ should be emphasised in the use cases, rather than just what is possible in the current socio-legal landscape. However, many of these road blocks, such as issues concerning the collection and management of consent, are common concerns and therefore it was decided that it would be better to deal with these separately rather than in great detail in individual use cases. The nature of the project has meant that many of the ELSE issues have been presented as barriers or rules that must be complied with, but it has been suggested that the ethical dimension of risk mitigation deserves further attention in any analysis of ELSE issues by a subsequent project.

A standardisation issue concerning the language used to present data to medical professionals was mentioned, and research challenge RCLG6 was modified as a result.

One topic discussed both at the workshop and within the consortium was whether healthcare and research requirements should be addressed together in healthgrid development, or whether they should be more separate. Legal and ethical issues are largely the same when it comes to handling personal data, whether it is for research or clinical motivations, and this was the main reason why the project did not separate both requirements. There can also be a natural synergy between healthcare and research, where the observations and output of both (ideally) feed into each other to enhance the quality of care provided and the relevance of research objectives.

Another topic was whether the emphasis should be on current technologies, or whether a more abstract view of what is trying to be achieved should be adopted given that technologies are continually changing and moving on. Within SHARE, there was a deliberate choice not to give a more abstract view like the SOKU vision but to adopt a user perspective. Also, current technologies will be the only real choice for the new pilots proposed, given the relatively short timescale proposed by the roadmap.

It was noted that the training requirement has been underestimated in previous grid projects, and a point concerning this has been added to section 10.1. It was also felt that US projects were under represented in the report, and therefore the survey of healthgrid projects in section 5.1.2 was expanded to include additional projects from the US.

### A.1.3 Further work

The remaining comments from the SHARE expert workshop can be divided into recommendations for further work within the healthgrid community and in the wider health informatics domain.

Specific to healthgrids, the following research was suggested:

- A detailed analysis of business processes to highlight the socio-economic cost and benefits of healthgrids.
- Further research concerning the management of rights in a healthgrid context.
- Other communities should be analysed in addition to those included in the SHARE use cases to determine the additional requirements of other stakeholders.
- The potential for federating and linking between the use cases mentioned in the report should also be explored. This could be an additional function of HealthGrid conferences.
- It would be beneficial to the research community to expand the HealthGrid knowledge base to capture key ELSE issues and monitor emerging best practices.
- The withdrawal of authorisation and consent by patients and research subjects was seen as a potential stumbling block for any live healthgrid, and more research will be required to determine how this could best be handled. Related concerns that should be addressed by

research in this area include mechanisms for the correction of data, how incorrect data could be detected by data controllers, how patients and other stakeholders can have access to data, and models of data governance addressing questions of trust.

- The sharing not only of data, but also of *derived* data should be addressed by any standardisation efforts. This will involve agreement on particular tools and how they can be used.
- A survey of schemas and ontologies within individual medical domains will be required to determine how many are defined adequately across Europe. This recommendation was added to section 10.1 of the report.

Concerns for the wider health informatics community include:

- How to deal with vulnerable populations and equality of healthcare provision across Europe.
- Problems concerning the sharing of derived data, and agreement/certification of tools used to aid this.
- Exploring alternative methods for data collection, such as incorporating data collected during a workout at the gym into a patient's record.
- Access to data is a critical area for all stakeholders; dealing with offline, archived data is a recurring problem in health research, and the right for patients to access their own data is also frequently mentioned.
- Confidentiality and security concerns mentioned in this report, including problems with anonymisation and pseudonymisation, are not specific to healthgrids.
- The best ways to raise public interest in developments in health informatics, and what constitutes 'public good'. The wide promotion of successful pilot projects could be effective in this respect.

## A.2 References

[1] Information on the *HealthGrid* initiative available at http://www.healthgrid.org.

[2] V. Breton, K. Dean and T. Solomonides (eds), *The HealthGrid White Paper*, in *Proceedings of Third HealthGrid Conference*, **Studies in Health Technology and Informatics Vol 112** IOS Press (2005) pp 249–321; preliminary version edited by V. Breton, K. Dean and T. Solomonides, issued as a joint CISCO/*HealthGrid* paper available at http://www.healthgrid.org.

[3] Information on SHARE project available at http://www.eu-share.org.

[4] *Action plan for a European e-Health Area*, COM(2004) 356, European Commission, http://europa.eu.int/information_society/doc/qualif/health/COM_2004_0356_F_EN_ACTE.pdf.

[5] V. Breton, A. E. Solomonides & R.H. McClatchey. *A Perspective on the HealthGrid Initiative*. Second International Workshop on Biomedical Computations on the Grid, the **Fourth IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2004)**, Chicago, USA, April 2004

[6] A. Sousa Pereira, V. Maojo, F. Martin-Sanchez, A. Babic & S. Goes. *The INFOGENMED Project*. **ICBME 2002**, Singapore, 2002.

[7] F. Martin-Sanchez, V. Maojo & G. Lopez-Campos. *Integrating Genomics into Health Information Systems*. **Methods of Information in Medicine 41**: pp. 25-30, 2002.

[8] BIOINFOMED Study Project EC-IST 2001-35024 *Synergy between Medical Informatics and Bioinformatics: Facilitating Genomic Medicine for Future Healthcare*, 2003, available from: http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=308

[9] CoreGrid Network of Excellence. See http://www.coregrid.net/mambo/content/view/2/25/

[10] Peter V. Coveney *Scientific Grid Computing* **Phil. Trans. R. Soc. A** (2005) 363, 1707–1713; see also http://www.realitygrid.org/.

[11] W. Leister & A. Skomedal, *Grid Systems in Health Care*, **Norsk Regnesentral** (Norwegian Computing Centre) from http://www.nr.no/files/dart/mmmc/facts-medgrid.pdf.

[12] Ignacio Blanquer & Vicente Hernández. *The Grid as a Healthcare Provision Tool*. **Methods of Information in Medicine 44**: pp. 144-148, 2005.

[13] Thomas Davenport and John Glaser *Just-in-Time Delivery Comes to Knowledge Management* **Harvard Business Review** (2002); reprinted in HBR *Managing Health Care* HBS Press (2007)

[14] The MammoGrid project, see http://mammogrid.vitamib.com/

[15] The DICOM (Digital Imaging and Communications in Medicine) standard, see http://medical.nema.org/

[16] Health-e-Child project, see http://www.health-e-child.org/

[17] VPH framework definition from STEP, see http://www.biomedtown.org/biomed_town/STEP/Reception/step-definitions/VirtualPhysiologicalHuman

[18] 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/about-share/deliverables-and-documents.html

[19] Christine Huanga, Laura A. Noirotb, Kevin M. Hearda, *et al. Implementation of Virtual Medical Record Object Model for a Standards-Based Clinical Decision Support Rule Engine* **Proceedings of AMIA Symposium** (2006)

[20] Open Grid Forum (OGF), see http://www.ogf.org/

[21] W3C Semantic Web Activity, see http://www.w3.org/2001/sw/

[22] Web Services Description Language, see http://www.w3.org/TR/wsdl

[23] Simple Object Access Protocol, see http://www.w3.org/TR/wsdl

[24] OWL-S (Ontology Web Language with Semantic markup), see http://www.w3.org/Submission/OWL-S/

[25] Introduction to UDDI: Important Features and Functional Concepts (2004) http://uddi.org/pubs/uddi-tech-wp.pdf

[26] The Web Services Resource Framework Primer 1.2 (2006) http://docs.oasis-open.org/wsrf/wsrf-primer-1.2-primer-cd-02.pdf

[27] Enabling Grids for E-sciencE (EGEE), see http://www.eu-egee.org/

[28] Distributed European Infrastructure for Supercomputing Applications (DEISA), see http://www.deisa.org/

[29] OSG (Open Science Grid), see http://www.opensciencegrid.org

[30] VDT, see http://vdt.cs.wisc.edu/

[31] TeraGrid project, see www.teragrid.org

[32] BIRN (Biomedical Informatics Research Network) project, see http://www.nbirn.net/

[33] The Article 29 Data Protection Working Party, set up by Directive 95/46/EC, see http://ec.europa.eu/justice_home/fsj/privacy/workinggroup/index_en.htm, esp. *Working Document on the processing of personal data relating to health in electronic health records (EHR).*

[34] Treaty of the European Union Art. 152 provides that matters of health services organisation are subject to the rule of subsidiarity and limits the role of the EU to supporting and co-ordinating the activities of the Member States.

[35] A comprehensive treatment of the subject of health systems challenges in forthcoming in a report to the "Scenarios4Health - Scenarios for ICT-Enabled New Models of Health Care" project (IST-150644-2006-F1SC-DE), http://www.scenarios4health.eu/

[36] For example, see Ettner, S.L. and M. Schoenbaum. *The role of economic incentives in improving the quality of mental health care*, in Jones, A.M. (ed) *The Elgar Companion to Health Economics*, Edward Elgar Publishing, 2006

[37] "Council Conclusions on Common values and principles in European Union Health Systems", Document (2006/C 146/01), Official Journal of the European Union on 22 June 2006, pp. 1 – 5

[38] Open Grid Services Architecture (OGSA), see http://www.globus.org/ogsa/

[39] Globus Toolkit, version 4 (GT4), see http://www.globus.org/toolkit/

[40] Grid Resources for Industrial Applications (GRIA), see http://www.it-innovation.soton.ac.uk/projects/past-projects/gria/gria

[41] gLite, see http://glite.web.cern.ch/glite/

[42] UNICORE (Uniform Interface to Computing Resources), see http://www.unicore.eu/

[43] Health Level 7 (HL7), see http://www.hl7.org/ and http://www.hl7.org.uk/

[44] CEN Technical Committees, see http://www.cen.eu/cenorm/businessdomains/businessdomains/isss/committees/tcs.asp

[45] J. Montagnat, A. Frohner, D. Jouvenot, C. Pera, P. Kunszt, B. Koblitz, N. Santos, C. Loomis, R. Texier, D. Lingrand, P. Guio, R. Brito Da Rocha, A. S. de Almeida, Z. Farkas *A Secure Grid Medical Data Manager Interfaced to the gLite Middleware* Journal of Grid Computing (Kluwer) 6 (1), no. pages 45--59 (2008).

[46] S. Erberich et al. *Globus MEDICUS - Federation of DICOM Medical Imaging Devices into Healthcare Grids*, Proceedings of HealthGrid 2007, IOS Press.

[47] I. Blanquer, V. Hernandez, D. Segrelles & E. Torres, *TRENCADIS – Secure Architecture to Share and Manage DICOM Objects in an Ontological Framework Based on OGSA*, Proceedings of HealthGrid 2007, IOS Press.

[48] Michael Brady et al *eDiamond: a grid-enabled federated database of annotated mammograms*, in *Grid Computing: Making the Global Infrastructure a Reality* F Berman, G Fox and T Hey (eds), Wiley, 2003.

[49] 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

[50] Integrative Biology project, http://www.integrativebiology.ac.uk/

[51] Hua Wang et al Probable limited person-to-person transmission of highly pathogenic avian influenza A (H5N1) virus in China The Lancet (published online April 2008)

[52] ACGT project http://www.eu-acgt.org/

[53] Tyrone Grandison, Christopher Johnson, Gerald Kiernan. *Hippocratic Databases: Current Capabilities and Future Trends* **Handbook of Database Security: Applications and Trends**. Michael Gertz, Sushil Jajodia (eds), 2007.

[54] Tyrone Grandison, Ranjit Ganta, Uri Braun, James H. Kaufman. *Protecting Privacy while Sharing Medical Data Between Regional Healthcare Entities*, **Proceedings of MedInfo 2007**, IOS Press, 2007.

[55] Rafae Bhatti & Tyrone Grandison *Towards Improved Privacy Policy Coverage in Healthcare Using Policy Refinement* **LNCS Vol 4721** *Secure Data Management* (2007)

[56] Calvin Powers, Steve Adler, Bruce Wishart, *EPAL Translation of the The Freedom of Information and Protection of Privacy Act* (IBM) available from the Ontario Information and Privacy Commissioner's website: http://www.ipc.on.ca/images/Resources/up-EPAL_FI1.pdf

[57] Lin.C et al Enhancing grid security with trust management IEEE Conference on Services Computing, SCC2004, Proceedings, IEEE (2004)

[58] Directive 2000/31 of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, O.J. L178, of 17 July 2000, 1-16.

[59] See http://ec.europa.eu/consumers/dyna/rapex/rapex_archives_en.cfm Council Directive 93/42 of 14 June 1993 concerning medical devices, O.J. L169, of 12 July 1993, 1-43.

[60] Welch, V. et al. *Security for Grid services* in Proceedings of 12th IEEE International Symposium on **High Performance Distributed Computing** (2003).

[61] Wenbo Mao, Fei Yan, Chunrun Chen *Daonity: grid security with behaviour conformity from trusted computing* Proceedings of the first **ACM workshop on Scalable Trusted Computing** (2006)

[62] Directive 95/46/CE of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and free movement of such data. OJ L 281 of 23 November 1995, 31-50.

[63] A Majumdar, S Warfield et al (2004). *Grid Enabled high Performance Computing for Image Guided Therapy*. **5th International MRI Symposium**, Oct. 15-16, 2004, Boston, MA.

[64] The Cancer Biomedical Informatics Grid project (caBIG), see https://cabig.nci.nih.gov

[65] The Cardiovascular Research Grid (CVRG), see http://www.cvrgrid.org/

[66] The National Biomedical Computation Resource (NBCR), see http://www.nbcr.net

[67] R. Sinnott, O. Ajayi, A. Stell, A. Young *Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data* in **Global Healthgrid: e-Science Meets Biomedical Informatics - Proceedings of HealthGrid 2008** (Tony Solomonides et al, Eds) Studies in Health Technology and Informatics Volume 138 IOS Press 2008

[68] The Legally eHealth project, see http://ec.europa.eu/information_society/activities/health/docs/studies/legallyehealth-fp6book.pdf

## A.3 Terminology

### Abbreviation List

| | |
|---|---|
| DICOM | The Digital Imaging and Communications in Medicine Standard |
| EC | European Commission |
| EHR | Electronic Health Record |
| ELSE | Ethical, Legal, Social and Economic [—issues, actions, etc] |
| EPAL | Enterprise Privacy Authorization Language |
| EU | European Union |
| HL7 | The Health Level 7 Standard |
| HPC | High Performance Computing |
| IPR | Intellectual Property Rights |
| OGSA | Open Grid Services Architecture |
| QoS | Quality of Service |
| SHARE | Supporting and structuring Healthgrid Activities and Research in Europe |
| SOKU | Service Oriented Knowledge Utility |
| VPH | Virtual Physiological Human |
| W3C | World Wide Web Consortium |
| WPx | Work Package x |
| WP3 | SHARE Technology and Security Activity |
| WP4 | SHARE Health Policy, Legal, Social and Economics Activity |
| WP5 | SHARE Applications Activity |
| WP6 | SHARE Roadmap Synthesis and Validation Activity |

### Definitions

The following definitions are useful for understanding the document content.

- **Authentication**: Verifying and confirming the identity of a grid user.
- **Authorisation**: Restricting access to resources based on what a user has been granted access to.
- **Data**: 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 them.
- **Data controller:** The person or organisation responsible for the manner in which any personal data is processed.
- **Data mining**: Automatically searching large volumes of data for patterns or associations.
- **Data model:** 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.

- **Data processor:** Any person who processes data on behalf of a data controller.
- **Data subject:** An individual who is the subject of personal data.
- **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.
- **Informed consent:** A legal term referring to a situation where a person can be said to have given their consent based upon an appreciation and understanding of the facts and implications of an action.
- **Metadata**: May be regarded as a subset of data, and are data about data. Metadata summarise 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.
- **Middleware**: A software stack composed of security, resource management, data access, accounting, and other services required for applications, users, and resource providers to operate effectively in a grid environment.
- **Ontology**: 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.
- **Processing:** Obtaining, recording or holding the data, or carrying out any operation on the data, including organising, adapting or altering it. Retrieval, consultation or use of the data, disclosure of the data, and alignment, combination, blocking, erasure or destruction of the data are all legally classed as processing.
- **Roadmapping**: An extended look at the future of a chosen field of inquiry, leading to an outline or map of how and by what means to achieve certain goals.
- **SOAP:** A protocol for exchanging XML messages over a network. It defines the structure of the XML messages (the SOAP envelope), and a framework that defines how these messages should be processed by software.
- **The Article 29 Data Protection Working Party**: A working party established by article 29 of directive 95/46/EC. It is the independent EU advisory body on data protection and privacy. Its tasks are laid down in article 30 of directive 95/46/EC and in article 14 of directive 97/66/EC.
- **Virtual Organisation:** A group of grid users with similar interests and requirements working collaboratively and/or sharing resources regardless of geographical location.
- **Web Service:** A software system designed to allow inter-computer interaction over a network to perform a task. Other computers interact with a web service, in a manner prescribed by its interface, using messages which are enclosed in a SOAP envelope and are often conveyed by HTTP. Software applications can use web services to exchange data over a network.
- **Workflow:** 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.
- **XML**: 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**: 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.