



SHARE

EPIDEMIOLOGY ROADMAP

Document ID:	SHARE-D5.2a_v1.0
Date:	05/07/07
Authors:	I. Blanquer, V. Hernandez, Nicolas Jacq
Activity:	WP5: Applications
Document status:	FINAL
Document link:	http://eu-share.org/about-share/deliverables-and-documents.html
Confidentiality:	Public
Keywords:	Epidemiology, Roadmap HealthGrid

Abstract: This document presents an analysis of current technologies in grids and the Ethical, Legal, Social and Economic (ELSE) framework of grid technologies for their adoption in health problems and particularly on epidemiology. The document makes a critical analysis of the technologies and the ELSE framework and proposes milestones to be achieved for the uptake of grids in this application domain.

Document Log

Issue	Date	Comment	Author
0.1	13/4/07	First version	I. Blanquer, V. Hernández
0.2	30/4/07	Main Sections Completed	I. Blanquer, V. Hernández
0.3	16/5/07	First Release for Review	I. Blanquer, V. Hernández
0.4	25/5/07	Comments from Nicolas Jacq, Vincent Breton, Isabelle Andoulsi	I. Blanquer, V. Hernández, N. Jacq, V. Breton, I. Andoulsi
0.5	21/6/07	Comments from Celine Van Doosselaere	I. Blanquer, V. Hernández, C. Van Doosselaere
1.0	05/07/07	Final version	N. Jacq

0.1	General Structure and General Sections	Preparing first release
0.2	Main Sections Completed	
0.3	First internal release	
0.4	Comments from Reviewers	N. Jacq, V. Breton, I. Andoulsi
0.5	Comments from Reviewers	C. Van Doosselaere
1.0	Final version	N. Jacq



CONTENT

1. INTRODUCTION.....	4
1.1. PURPOSE.....	4
1.2. APPLICATION AREA	4
1.3. REFERENCES	4
1.4. TERMINOLOGY.....	4
1.5. SOURCES.....	5
2. EXECUTIVE SUMMARY.....	7
3. THE HEALTGRID VISION.....	8
4. BACKGROUND	11
5. USE CASE	13
6. ASSESMENT OF GRID TECHNOLOGY IN HEALTH	15
6.1. AUTOMATIC DATA GATHERING.....	15
6.2. ENHANCEMENT OF QUALITY OF DATA	16
6.3. SUFFICIENT SECURITY MANAGEMENT.....	16
6.4. EFFICIENT PERFORMANCE OF PROCESSING SERVICES	17
6.5. SEAMLESS INTEGRATION OF PROCESSING SERVICES	18
7. ASSESMENT OF ELSE ISSUES IN HEALTH.....	20
7.1. EFFICIENT DEPLOYMENT IN HEALTH ENVIRONMENTS	20
7.2. SUFFICIENT SECURITY AND ETHICAL MANAGEMENT	20
7.3. RELIABILITY AND QUALITY OF SERVICE IN INFRASTRUCTURES AND LONG-TERM EXPLOITATION	22
8. ROADMAP	23
8.1. RESEARCH TOPICS	23
8.2. DEPLOYMENT ACTIONS	23
8.3. TIME PLANNING	24
9. CONCLUSION	26

1. INTRODUCTION

1.1. PURPOSE

The purpose of this document is to analyse the requirements, needs and aims of the epidemiology use case defined in the deliverable D5.1a with respect to the technological questions and the ELSE issues identified in D3.3 and D4.2, respectively. Those documents presented the bottlenecks and challenges for Technology and Security (D3.3), and for Ethical, Legal, Social and Economic Issues in the adoption of grids in health. This document will revise the issues raised in those documents and analyse the extent to which they match the requirements expressed in the use case scenario. From this analysis, the evolved versions of the technology and security roadmap (D3.4) and the ELSE roadmap (D4.2) will be elaborated. This document will also be used for the integrated roadmap (D6.2).

This document has a corresponding version for Innovative Medicine identified under the code D5.2.b. This document is self-contained including the relevant information from D3.4 and D4.2.

This document is intended for both users (epidemiology researchers, public health authorities or industry) and developers (technologists and legal experts). It tries to outline the milestones that must be covered for the adoption of grids. The level of technical detail is presented in two steps, trying to attract the interest of policy makers, but providing the technical details that are necessary for the implementation.

All deliverables can be downloaded from the SHARE website at www.eu-share.org.

1.2. APPLICATION AREA

The document is intended for internal and external use. It will be used as a dissemination tool for the SHARE project, and for consultation in the process of elaborating the second versions of the deliverable and the integrated roadmap.

1.3. REFERENCES

- [1] The digital patient, special issue of ERCIM news, number 69, April 2007.
- [2] Pre-Standards for Vital Signs and Electronic Patient Record of the CEN TC251.
- [3] H. F. Gagliardi, F. A.B. da Silva, and D. Alves, Automata Network Simulator Applied to the Epidemiology of Urban Dengue Fever, ICCS'06.
- [4] TRENCADIS – Secure Architecture to Share and Manage DICOM Objects in a Ontological Framework Based on OGSA, IOS Pres, From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences, Studies in Health Technology and Informatics. ISBN 978-1-58603-738-3, ISSN: 0926-9630, Vol. 126, pp. 115-124, 2007.
- [5] V. Caviccholi Neto, H. F. Gagliardi, A. Rezende, E. S. Gonçalves, E. Gallo, F. B. Silva, I. Torres, D. Alves, Data Access Service in a Computational Grid Platform Applied to the Monitoring and Control of Epidemics on Georeferenced Dynamic Maps, e-Science'06.

1.4. TERMINOLOGY

Glossary

Term	Definition
Data provenance	Reliability of a piece of data according to their origin and processing.
Data quality	Integrity, completeness and accuracy of a distributed set of data.
Effectiveness	Similar to Efficacy but under clinical practice conditions.
Efficacy	Validation of the suitability of a therapy to solve a medical problem based on a sample population under optimal conditions.
Efficiency	Analysis of the suitability of a therapy in the whole population and under the real conditions of medical practice.
EHR	Electronic Health Record.
Epidemiology	Scientific study of factors affecting the health and illness of populations.
Federation	Integration of different sources of information.
HealthGrid	Adoption of Grid Technologies in a Health IT environment.
ICD-9	International Coding of Diseases, Version 9.
ICD-O	International Coding of Diseases, oncology.
ICT	Information and Communication Technologies.
LOINC	Logical Observation Identifiers Names and Codes.
Medical Informatics	Usage of Computer Means for storing and processing medical information related with patients' health.
Morbidity	Damage on a patient's health.
Ontologies	A set of attributes and functions associated to a concept.
Prevalent cases	Cases that have been identified and diagnosed in the past.
SNOMED	Systematized Nomenclature of Medicine. A terminology for defining medical terms.
Terminologies	Subset of an ontology that defines a vocabulary.
UMLS	Unified Medical Language System, a set of Knowledge Sources for easing medical informatics developments.
Use-case	Example complete scenario that describe the problem, actors, data and procedure to tackle a specific problem.

1.5. SOURCES

For the preparation of this document, along with the information listed in the references and the previous deliverables, information from different projects, meetings and interviews with users has been used.

Along this information, main contributions come from:



- Events
 - HealthGrid 2006 SHARE Workshop and Advisory Panel.
 - International Symposium on Grid Computing 2006.
- Projects
 - Infrastructure projects DEISA, EGEE and EELA Project websites and associated documentation.
 - CVIMO (Valencian Cyberinfrastructure for Research and Epidemiologic Study of Medical Imaging in Cancer) <http://www.grycap.upv.es/cvimo>.
- Key Users
 - General Public Health Authorities of the Valencia Region
 - Centre for Research in Public Health.
 - Institute Cavanilles for the Study of Bio-diversity.
 - ABUCASIS (GAIA, SIP, SIC and SIA Components) Information System.
 - Project for the Epidemiological Study of the Treatment of Arthritis.
 - Institute of Molecular and Cellular Biology (IBMCP) of Valencia.
 - International Epidemiology Association.
 - European Public Health Association.

2. EXECUTIVE SUMMARY

Epidemiology constitutes one relevant use case for the adoption of grids for health. It combines challenges that have been traditionally addressed by grid technologies, such as managing large amounts of distributed and heterogeneous data, large scale computing requirements and the need for integration and collaboration tools. However, epidemiology presents important additional constraints that are not solved and makes the take-off of grid technologies difficult. The most important requirements are the problems on the semantic integration of data, the effective management of security and privacy, the lack of exploitation models for the use of infrastructures, the reduced Quality of Service of infrastructure and the seamless integration of the technology on the epidemiology environment.

Those requirements involve several technological, scientific and “political¹” actions which must be performed to achieve widely used effective grids for health.

However, not all the actions are equally urgent. The market studies on the adoption of IT suggest that effective pilots are needed to convince early users. Those pilots should evolve from the early prototypes that are being developed in emergent projects and must prove the benefits in a reliable and effective way. Thus, early users will need reliable platforms, a reasonable fulfilment of legal regulations, and clear exploitation models. Early users will accept new interfaces and even changes on their work schemes, but will react negatively against unpredictable behaviour or lack of services.

Once early users are convinced of the benefits of grids, larger collectives could be addressed. This will require reinforcing the Quality of Service models, the scalability and the accurate fulfilment of legal issues. They will require less modification on their workflow and tools used and will require a seamless integration of the grid on their procedures.

Finally, the deployment of grids for health for pure clinical practice will require new economic models and certification of technologies and components. This usage should be led by the industry.

This document details the requirements, milestones, communities and challenges producing the roadmap reflected in the above points. It provides a vision for the uptake of grids for health in the next 10 years.

¹ - “Political” decisions are those which relate to economics and legal regulations.

3. THE HEALTGRID VISION

Most healthcare systems in the developed world are facing multiple challenges in their attempt to maintain an acceptable level of care for their citizens. The principal challenges are often experienced and expressed in economic terms, such as issues of total cost, capacity and responsiveness, and allocation of limited resources.

Underlying these economic constraints is the moral challenge of priorities, as governments seek to balance the demands of:

- changing demographics, with an ageing population both surviving and remaining active longer;
- increasingly effective treatments for both acute and chronic conditions; and
- sophisticated – and sometimes unproven – novel treatments for conditions that not very long ago were considered untreatable.

In an attempt to meet these demands, health systems have increasingly looked to information technology to help, among other things, to optimize the distribution and use of resources, to reduce queues and waiting times, to record and thus avoid errors, and to provide modern treatments into remote communities.

Beyond these essentially resource-oriented uses, information technology is also seen as an essential ingredient in a change in medicine itself. In the course of the last two decades, the practice of medicine and healthcare provision in general have moved away from reliance on the doctor's personal knowledge and craft skill to requirement of a scientific basis in diagnosis and treatment, in what has come to be known as 'evidence-based practice'. The evidence a doctor or nurse must now take into account is:

- published medical knowledge;
- their knowledge of the patient; and
- practical knowledge of what is available by way of procedures, protocols, and so on, in their environment.

In this context also, governments have initiated programmes to create information-driven healthcare systems.

However, these modernization processes face a number of challenges:

- creating and populating, connecting and understanding patient records across organization boundaries and, in due course, across different national health systems;
- increasing the openness and accessibility of systems - e.g. providing patients with ownership of their healthcare record – while:
 - ensuring privacy, confidentiality and ethical compliance in the socio-legal plane;
 - maintaining data integrity, security and authenticity (e.g. provenance and semantics) in the technical plane;
- providing appropriate levels of authorization and authentication of users across all the services and the citizen;
- discovering, grading and certifying trustworthy sources of knowledge and case information to guide future action; and
- winning the trust and commitment of the medical professions at a time of immense change and economic pressure;

One immediate limitation is in the application of traditional information networks and technology in healthcare. Governments have naturally focussed on the technical issues that are reasonably well understood, even if solutions are not always easy to obtain: robustness of networks, scalability of systems, readiness to handle a very large volume of data. However, in many respects, these reproduce in the technology some of the problems of the traditional paper-driven systems: inflexibility, misdistribution of resources, failure to understand the needs of medical practitioners, failure to support effective collaboration, and an ultimately simplistic equation of quality with 'choice', while minimal provision is made for just-over-the-horizon future technologies such as genomic medicine and individualized prescribing.

In the face of these challenges, a computational innovation, 'grid' technology, or *the grid*, has become available to clinicians in the last few years, first as a research tool and then, in the not-too-distant future, as a serious healthcare infrastructure. The grid is not one technology but many, and the use of the singular is somewhat misleading, but it is convenient inasmuch as it echoes 'the internet' to which it is closely related. Just as the internet, or more precisely the World Wide Web, has provided a massive information platform whose exploitation is limited only by economics (and, in some cases, politics) grid technology promises to scale this up to the provision of unprecedented computational power, online storage and collaboration opportunities. The informatics grid approaches the provision of computational, information and communication services through resource sharing in a seamless and transparent manner, much as the electricity 'grid' provides power to any device plugged into it, irrespective of its purpose or design. Grid computing aims at the provision of a global ICT infrastructure that will enable a coordinated, flexible and secure sharing of diverse resources, including computers, applications, data, storage, networks, and scientific instruments across dynamic and geographically dispersed organizations and communities (sometimes known as 'virtual organizations' or 'VOs'). Grid technologies promise to change the way organizations tackle complex problems by offering unprecedented opportunities for resource sharing and collaboration. Just as the World Wide Web transformed the way we exchange information, the grid concept takes parallel and distributed computing to the next level, providing a unified, resilient, and transparent infrastructure, available on demand, in order to solve increasingly complex problems.

A 'healthgrid' is an innovative use of this emerging information technology to support broad access to rapid, cost-effective and high quality healthcare. In particular, the areas of healthcare provision and research that can be beneficially affected by healthgrid technology include:

- medical imaging and image processing;
- modelling the human body for therapy planning;
- pharmaceutical research and development;
- epidemiological studies; and
- genomic research and treatment development.

In all these areas, grid technology can either significantly reduce the cost or time to produce results and evidence, or even 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 new perspectives to enable interdisciplinary research at the crossroads of medical informatics, bioinformatics and system biology to impact healthcare.

With the regular progress of technology and infrastructures, 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 aimed at developing new grid-enabled services and environments relevant to the needs of life sciences and healthcare; and
- end-user projects that focus on specific life science or healthcare issues and integrate grid technologies wherever they appear relevant.

Born from discussions between grid application developers and medical informaticians, the concept of healthgrids is now over three years old. The annual HealthGrid conferences are an opportunity to evaluate the growing usage of grids for life science and medical research. Adoption of grids for healthcare is expected to follow their adoption in the life sciences and medical research, provided the legal and ethical framework of member states allows their deployment.

The next section of this document analyses this vision and presents a set of user requirements common to data-centric health applications. These requirements will be carefully analysed with respect to the technology and ELSE frameworks in the following sections.

4. BACKGROUND

This document focuses on the epidemiology use case, which has been selected due to its relevance in the health arena. However, it is worth considering the problems of Information and Communication Technologies (ICT) in health research before starting a more detailed description for epidemiology.

ICT-driven research that uses health data generally 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, requiring a pre-existing knowledge on models of sound and ill organs or tissues in order to perform their function. Population-level analyses normally deal with the integration of larger, poorer-quality data. Semantics are especially relevant on those approaches.

Users ought to be able to take for granted:

- that the security mechanisms are sufficient to protect their data. They do not need to know almost anything about encryption, secure transfer, delegation or other technical issues.
- that the system will not raise the concern of the ethical and legal committees of their research institutions.
- that the results of their research will be private and available to third parties only if desirable. They will want to be able to define groups and permissions at a global scale for their research community.
- that the services are reliable, efficient and permanent. They will not understand 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. 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 spreading but will not need to provide solutions.

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-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.
- 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 to 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 most health research.
- reliability, scalability and pervasiveness. All the previous services must be robust and trustful and should be scaled without reducing performance.

Most of these requirements are relevant in the epidemiological use case and will be discussed in detail in the next sections.

5. USE CASE

The epidemiology use case is defined as a system able to link the information from distributed and heterogeneous databases, identify patients, complete episodes and improving automatically quality without interrupting clinical practice. With this data, complex epidemiology models are fed and simulated producing, in a reliable way, aggregated prospective results. The analysis of this section is concentrated in 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.
- conformity with guidelines on ethics. 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.

These requirements will be analysed with respect to the technology and security roadmap and with regard to the ELSE analysis in the following sections.

6. ASSESMENT OF GRID TECHNOLOGY IN HEALTH

Now that the use case and related requirements have been defined, we will review the status of the technology, as described in deliverable D3.3, with respect to these requirements and propose actions to be developed in further technologically research actions for those requirements identified in the use case and that are not yet matched technologically. In the next section, we will analyse the requirements with respect to other issues studied in deliverable D4.2 and also propose actions where needed.

Deliverable D3.3 identifies five milestones that focus on general and broad ranging issues that are affecting the take-off of grids in health. These five milestones (computing grid, data grid, research knowledge grids, grid DICOM and grid EHR) are long steps that must be achieved in general. The selection of a specific (but representative) use case, such as this epidemiology roadmap, gives the opportunity to analyse each one of the steps in detail. The milestones are divided into more detail, thus describing precise steps that affect the use case.

Computing grid issues (i.e., the first milestone) are considered in the analysis of the efficiency of the performance of processing and on the seamless integration. Data grid issues (i.e., the second milestone) are examined when we consider the automatic data gathering, enhancement of data quality and sufficient security management necessary for grid use in epidemiology. Finally, the seamless integration of epidemiological procedures and tools, the integration of data and the data quality enhancement are the basis for the research knowledge grids.

All these issues are analysed in detail in the following sections.

6.1. AUTOMATIC DATA GATHERING

Automatic data gathering implies that data sources can be automatically fetched, the data properly retrieved and the integration of the different data fragments performed. This relates with the technical bottlenecks identified in sections 5.1.1 (Development of Grid Data Management Services) and 5.1.2 (Development of Grid Nodes in Health Care Centres) of deliverable D3.2.

Regarding automatic fetching, this implies normally setting up agents and middleware on the local repositories. Setting up software will require direct access to databases, which is not always feasible, especially in the context of proprietary systems. This could require the manual execution of data queries, sometimes exporting the data to plain formats that could be read by the mentioned agents. Automatic fetching could require remote-control applications that could, under a secure environment, log in and fetch the data safely.

In any case, it will be necessary to set up a federated schema for accessing data. OGSA-DAI has been used in the particular case of epidemiology although there are, to date, no large-scale implementations that could have demonstrated the reliability of such systems. The adaptation to different formats, the delegation of authorization, the management of privacy and the efficiency and reliability at large scale has thus not been clearly demonstrated yet.

Once the data is available in the distributed environment, the integration of the data presents additional problems. Data will rarely present unique keys, and private information will be required, raising legal and ethical issues. Moreover, unique keys could easily reveal a patient's identity. However, technological solutions might provide the answer: there are, for example, pseudo-anonymisation techniques that could be used to create identifiers on trustful contexts, matching data from different sources. For example, two databases could merge the results from complementary records matching a global identifier using asymmetric cryptography. This would ensure that no private information is released, that only trustful entities are connected, and that records are correctly matched.

6.2. ENHANCEMENT OF QUALITY OF DATA

Provided that the data is properly gathered and the individuals and cases properly identified, the different pieces of information must be integrated. The different pieces could be inconsistent, so data curation, i.e. keeping the most reliable and complete piece of data, must be considered. This requires solving the following issues:

- reliability. Data provenance must be implemented. Data sources should be ranked according to factors such as nature of the source, level of certainty of the diagnosis, date, methods used, etc. This will require an advance on the characterization of provenance in distributed database integration. Dynamic adjustment of the reliability will be important in production environments.
- coherence. A detailed medical ontology of terms is needed. This ontology should be understandable by computers and should define the possible and impossible relations. It should be important to have weak and strong relations that could define terms that are normally compatible and rare associations. For examples, there should not be any relation between “cervix” and “male”, there should be a strong relation between “brain” and “temporal lobe”, and weak relation between “male” and “breast cancer”.
- coding. Data is not always coded, but it could be correctly and accurately described in natural text. Specialized text mining tools should be able to propose codes, which should not replace, but complement natural text. Standardised nomenclatures do exist, although they are maintained and updated to cover advances in treatment and diseases.

The advances on the semantic web and semantic grid could partially solve the problems of characterizing relations and ontologies (OWL, RDF). More effort is needed on dynamic characterization and provenance. However, the main issue lies on the deployment of these ontologies, rules and text mining tools.

6.3. SUFFICIENT SECURITY MANAGEMENT

Sufficient security management relates to the technical bottlenecks 5.1.5 (Recording and Ensuring Consent) and 5.1.6 (anonymisation and pseudo-anonymisation) of deliverable D3.2.

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.
- 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 authorization model could manage the same model, as described in the next point.
- authorisation. Management of the authorization has not been effectively addressed yet in multiple-decoupled institutions. The use of attribute extensions in a central authorization system (such as VOMS) reduces the flexibility of the management of the authorization - 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 authorization entities schemas, such as combining Shibboleth and PERMIS systems, could reduce the problems in deploying large-scale VO

membership and delegating on trusted authorization 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 authorization, 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 unauthorized 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.

Finally, auditing, logging and monitoring are key issues to ensure compliance to regulations and proper response to adverse events. Current systems are mainly focused on the availability of the resources, rather than on the access policies. Intrusions, non-conformal uses or security leakages are detected through manual inspection. Since there exist effective mechanisms in other technological areas, the adaptation of such techniques to grids for health should be considered and studied.

6.4. EFFICIENT PERFORMANCE OF PROCESSING SERVICES

The performance of applications related to health (and particularly to epidemiology) would require solving two main issues: the integration of the maximum amount of resources and the Quality of Service. This relates with the technical bottlenecks outlined in sections 5.1.4 (Issues with Grid Infrastructures under Heavy Use) and 5.3 (Communication Bottlenecks) of deliverable D3.2.

Although the efficiency of fine-grain parallelism in grids has not been absolutely demonstrated, these currently unsolved technical issues are not critical in the epidemiology scenario.. MPI (Message Passing Interface) on the grid versions faces many problems regarding the configuration of sites, network latencies and scheduling. However, these unsolved technical issues are not critical in the epidemiology scenario. Applications in model simulation or statistical inference normally rely on a high-throughput model, rather than on a fine-grain coupled parallel mode. Provided that the infrastructures solve the problems of Quality of Service and efficiency, current infrastructures could provide enough resources for solving the problems faced in epidemiology.

On the contrary, most implementations of currently popular software for epidemiology are not “grid-enabled”. A few of them use parallel computing, although their computing model is fairly uncoupled. The migration of those applications into more decoupled, high-throughput models will enable to adapt them and to run them easily on the grid. The use of widely accepted interfaces, such as DRMAA for job spanning, will also improve the portability of such applications.

In the simulation of epidemiological models, it will be necessary to access data and to perform complex calculations. Since data will normally be stored at the local repositories and made available to the grid under demand, network, storage and computing resources must be guaranteed for any processing to start. Epidemiology is not bound by interactive processing needs, although best-effort processing will be needed. The patterns of access to the resources will generally be predictable and regular (weekly or daily updates of the information and consolidation, periodic analysis), although there could be at least two additional cases: large, schedulable and computing intensive analysis could be requested by researchers on specific studies; and shorter, unpredictable, computing intensive and time-limited analysis could be requested by epidemiologists when an epidemic burst appears in order to evaluate counter means or predict the expansion of a disease.

In the case of predictable or regular executions, there is a need to provide resource reservation and Quality of Service. Grid middleware and infrastructures should provide the means to reserve, block or checkpoint resources to attend those demands. Network reservations should be necessary to fetch and load temporarily data on the grid. In the case of unpredictable, quick need for resources when facing an emergency, the system should allow for pre-emptive scheduling and priority management. This will enable higher-priority actions to be performed in an effective time.

Main problems on the implementation of such policies lie on local resources, which are owned and administered by the resource providers. This should be implemented at the level of the resource managers and pushed up to the middleware level.

This model of high-throughput computing requires effective and fault-tolerant scheduling. An important fraction of time is still consumed by resource brokers and workload managers to discover the resources and schedule the jobs. Moreover, fault-tolerance, although implemented, is rather inefficient, and normally the developers end up creating their own solutions. Computing in epidemiology, as described in the document, does not require generally interactive nor short deadline jobs (although it could benefit in specific cases). Current scheduling models can be considered sufficiently effective. Thus, efforts must be concentrated on reliability and integrity as in transactional systems. It must be ensured that a job is completely finished, or, if it has intrinsic errors, no effect has been produced over the system. A user must be sure that, when a job fails, the error is due to the job itself, and not to the infrastructure.

6.5. SEAMLESS INTEGRATION OF PROCESSING SERVICES

Along with the migration of epidemiology applications to grid environments, there are many tools that do not benefit from the migration to grids. These tools, normally used for the final post-processing of the results or their presentation, as well as the tools for accessing the data repositories, 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 or CEN TC251 norms. This will ease the integration of the tools with infrastructures. This relates with the technical bottlenecks outlined in sections 5.1.2 (Development of Grid Nodes in Health Care Centres), 5.1.3 (Development of Services Compliant with medical Informatics Standard Specifications Based on the Web Services Technology), 5.2.1 (Organisation of the Healthcare and Medical Research Community), 5.2.2 (Development of Best Practices) and 5.2.4 (Worldwide Open Standards in Medical Informatics) of deliverable D3.2.

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 (de facto standard) and prENV 13606-4 (norm under development by the CEN TC251) gateways will ease the integration of the medical resources on the grid. Support to other vital signs exchange formats, such as ENV13734 will also be important.
- medical imaging. Although not directly related with the epidemiology use case proposed, medical imaging is also widely used in epidemiology. 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, MEDICUS or TRENCADIS. 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. Along with the standardization of the interfaces to data, it is important to provide interfaces to the statistical tools most widely used in the epidemiological 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. ASSESMENT OF ELSE ISSUES IN HEALTH

This section analyses the previous requirements with respect to the Ethical, Legal, Social and Economic (ELSE) Issues described in D4.2. Not all the requirements defined for the epidemiology use case are relevant to ELSE issues, so only those with impact will be discussed.

Considering the conclusions of D4.2, there are several issues that have to be solved in the ELSE framework. In terms of economic issues, the main concern is the reimbursement and financial implications and the service-oriented business model. These issues are discussed below in the reliability and deployment of health environments. In response to the legal and ethical analysis made in D4.2, we analyze the impact of the legislation, looking in particular at the management of responsibilities and intellectual property (IP) rights. These issues are considered below in the analysis of the security and ethical management. Finally, in terms of social issues, recommendations are made mostly towards user-friendliness and training. Those issues are considered in the previous section on the seamless integration of the research and data tools (6.5).

The following subsections make a detailed analysis of these issues.

7.1. EFFICIENT DEPLOYMENT IN HEALTH ENVIRONMENTS

Deployment of grid infrastructures in health have two main barriers: the complexity of the middleware and system maintenance. The complexity of the middleware affects installation and configuration, but also maintenance. The use of non-standard protocols or port configurations makes the middleware unacceptable in many hospital environments, in which firewalls and independent networks are hardly configured. Moreover, the introduction of complex software stacks is normally viewed as a threat by system administrators, since maintenance and interactions in the medium term can cause security problems or malfunctions in systems that need to be permanently active. Finally, maintenance should not require a large amount of additional resources.

The previous issues lead to the following necessities:

- lightweight and compact middleware releases. The concept of virtual machines is interesting for isolating resources and for local access. The management of dependencies should enable re-using external packages while maintaining the quality and reliability. Environments for multi-component cross-compilation, such as ETICS, could also be explored.
- Integration of grid-like protocols in restricted environments, such as the hospital and primary care networks (main source for epidemiological data), implies solving problems of firewalls, NAT (Network Address Translation) and the like. Restrictions on the use of standard ports and protocols and the implementation of robust network communication environments able to deal with multiple private networks is another important issue that is preventing the deployment of grid environments. As an example, GridFTP is reliable and efficient in public networks between peers with public IPs, but the range of ports required and the problems in managing local IPs makes it difficult to use in those environments.

7.2. SUFFICIENT SECURITY AND ETHICAL MANAGEMENT

Although there are many developments in the context of security management on grids for health, it is important to reach a consensus and to develop de-facto standard solutions, such as in authentication. The previous section raised the technical problems in this matter, proposing some research lines. This section tries to go beyond these aspects and analyses the security means that would be needed to comply with the law.

Management of patient consent is a key issue for personal data. Since medical data in general could become private in the future, as new developments in research could discover particularities that could not guarantee that the identity of a data owner could not be revealed, it will be important to treat medical data in general as personal. This will require implementing ethical measures such as patient consent and patient feedback. However, it does not mean that it will be convenient to pseudoanonymise and dissociate the data (the contrary could be illegal).

- The management of patient consent should be improved. Data controllers should be advised on the requests that they have to make to the patients in order to ensure that their data is usable for research, storing and distribution over different countries. The management of consent is a key issue for being compliance to data protection on the long term, since anonymised data could become “owner-identifiable” data if new correlations and further research is performed.
- The management of the feedback to the patient should also be considered. Patients have the right of being informed or not, if a further analysis reveals a potential disease. This will require that identity of the patients should be able to be reconstructed (which will also be possible in pseudoanonymised or directly identifiable data). The use of efficient pseudoanonymised keys has been successfully studied in several projects (Aneurist, Belgium Authorities).

In general, there should be electronic means to manage the patient consent easily and to ensure that the data subject’s rights are considered. It should be possible (although only allowable for very restricted users) to reconstruct the identity of data subjects if a threat on their health status is detected and the consent has been given.

Finally, interpretation of the law can have various impacts on the use of grids for epidemiology, such as, for example:

- The data gathered at the beginning of the study could be considered as excessive for a concrete study, since it is obtained for further studies not linked to the original purpose for which the data was collected. The law clearly states that the information should be the minimum necessary, but searching undiscovered correlations in data mining (which could be a great benefit of grids in epidemiology), would require many data whose relation a priori is not known.
- By definition, data gathered for epidemiological purposes must be kept for long periods, and can have identifiable personal data within. This is essential for specific research in epidemiology (e.g. survival rates on cancer studies require following on a patient for several years). However, the law explicitly discourages researchers from doing so, stating instead that all data gathered must be well justified and its purposes clearly explained.
- The federation of data repositories produces virtual, distributed repositories. The data controller is a unique figure that is liable for any privacy leakage or data misuse. According to the law, it is not clear that a virtual repository could have multiple individual data controllers, although considering the case of virtual repositories (the data is not copied to a global, central system, but extracted as needed from the local repositories), this case should be considered. Currently, data controllers must ensure that the data being extracted from their local repositories to the virtual global repository is performed with the proper permissions and at the level of an authorized user. There should be somehow the figure of a federated data controller who would be liable for the misuse of the data legally extracted from the local sites.

The above issues cannot be answered with technological solutions, and amending or creating the law is not easy. However, in order for grids to be used and deployed to their full potential in health settings such as epidemiology, there is a need for clear, widely accepted jurisprudence that could give the researchers a reasonable degree of confidence and reduced liability.

7.3. RELIABILITY AND QUALITY OF SERVICE IN INFRASTRUCTURES AND LONG-TERM EXPLOITATION

Reliability of service requires, on one hand, the development of robust services, and on the other, a stable infrastructure (both middleware and resources). The reliability of software components has been addressed in the technical part of this document. The reliability of the infrastructure is part of the quality of service policies to be implemented in the exploitation models.

Currently, infrastructures are intended mainly for research and do not have a clear exploitation plan. The acceptance of applications is done through scientific committees, but the compromise of resources is not normally addressed.

When the infrastructures are to be used for production, they must ensure that a quality of service can be offered. This quality of service should be agreed as any other resource provider, defining the availability, the minimum, the nominal and the maximum parameters for performance and overhead. Response time to problems should also be defined.

Of course, these agreements will have a counterpart that should be defined by the structure providers in economic or other terms. The infrastructure should provide the technical means presented in the technical section to ensure that this could be offered.

The funding bodies (such as the EC or the National Agencies) should consider this exploitation model and provide the means in their calls both to support the infrastructure and to support the researchers in subcontracting the infrastructure (as it has been considered in the subcontracting of computing costs, but considering the different exploitation models).

It is clear that a factor that reduces the acceptability of those systems is the uncertainty in the long-term of the infrastructure. Many users are reluctant to migrate to grids since they do not see clearly if the infrastructures will be available in the future.

Infrastructure providers should develop this concept and work on the set up of long term agreements. Steps done in the EGI are very relevant issues and should be extended and consolidated.

Finally, it should be noted that the use of grids for health and healthcare provision will require certification. This will be a very complex step, since certification should involve resources, middleware and applications. Considering the shared nature of grids, this could not be done currently for health infrastructures. The global certification might need individual certifications of Internet connections (and providers), infrastructure resources, middleware and applications. There are companies offering hosting and “housing” of services and resources, which follow a similar approach.

8. ROADMAP

This section outlines the issues presented in the previous two sections and proposes a timeline for them. The timeline has been defined considering the models for the adoption of IT and analyzing the requirements for each group of users.

8.1. RESEARCH TOPICS

This subsection describes the milestones that involve issues that require further research to satisfy the requirements of the users.

MR1. Pilots on epidemiology.

- Need for successful pilots on epidemiology that will demonstrate the benefits of the technology.

MR2. Pervasiveness and reliability.

- Need for real fault-tolerant scheduling systems and pervasive services. Improvement of the existing services and adoption of standards to support transactional jobs.

MR3. Lightweight, compact and health-networks-compatible middleware.

- Need for a grid middleware that can be installed in health environments seamlessly and without requiring exhaustive maintenance and administration.

MR4. Secure data models compliant to regulations.

- Need of data architectures that implement private data dissociation, pseudo-anonymisation and encryption, and that are able to fulfil the legal requirements in the matter of data management.

MR5. Epidemiology data sources adapted to grid models.

- Need for grid-enabled gateways to epidemiological data using medical informatics-related connectors, such as HL7, DICOM, ENV13606, etc.

MR6. Quality of Service (priorities and resource reservation).

- Need for services in the infrastructures to define exploitation models and guarantee a Quality of Service. Consolidate the booking of resources in advance and to guarantee a pre-negotiated Quality of Service.

MR7. Epidemiology applications adapted to grid models.

- Need for grid-enabled applications used in epidemiology.

MR8. Scalability of resources.

- Need for scalable job scheduling, data cataloguing and data transfer.
- Integration of resources with low latencies and high performance.

MR9. Semantic Data Integration.

- Need for knowledge-driven catalogues and integration based on the metadata.

8.2. DEPLOYMENT ACTIONS

This subsection describes the milestones that involve issues in which the technology is mature enough but deployment actions are needed.

MD1. Health infrastructures and service agreements.

- Need for clear exploitation and service agreements between infrastructure providers and health users.

MD2. Patient consent management.

- Need for the development of patient-consent models to be valid for data to move across country borders and for data to be stored for long periods of time.

MD3. Support for health protocols.

- Need for the integration of the grids for health intrinsic features on the health protocols.

MD4. Accepted jurisprudence on the legal issues of health grids.

- Need for clear jurisprudence on the management and processing of distributed data in grids for health environments to create trust.

MD5. Certification and introduction of grids for health into medical practice in addition to Research

- Need for certification models able to deal with the multiple resources and components of grids for health.

8.3. TIME PLANNING

Considering the well-known curve for IT adoption, it is important to convince users progressively, segment by segment. The requirements for each segment vary, and give the timeline for the requirements to be solved.

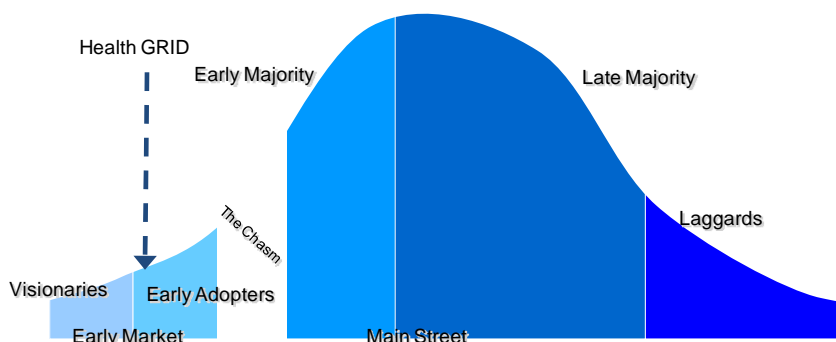


figure 1: Structure of the IT market. Horizontal axis refers to time and the size of the areas determines the size of the user community.

Early market

- Visionaries. They will develop experimental case studies and can accept poor security and rudimentary user interfaces. They will be involved if they foresee that an advance on the knowledge could be obtained, assuming the risk of not having outstanding results, but learning from the experience.
- Early adopters. They will develop complete pilots that are functional and will integrate them into production. They are convinced of the benefits but require that security, ethical concerns and legal requirements be considered. They will also count on a level of reliability of the service and a reasonable user interface. They will not be concerned about scalability and could accept changes on the procedures.

Main street

- Early majority. They represent a community of users that understand the benefits and are proactive. They will require that changes to be procedures be small, and that scalability and reliability be high. These are the most difficult public to address.
- Late adopters. Most of the users will adopt the new technologies if they are forced to do so. They will be convinced by the early majority and will be more restrictive on the requirements.
- Laggards. They will never adopt the technology and are therefore not addressed in this roadmap.

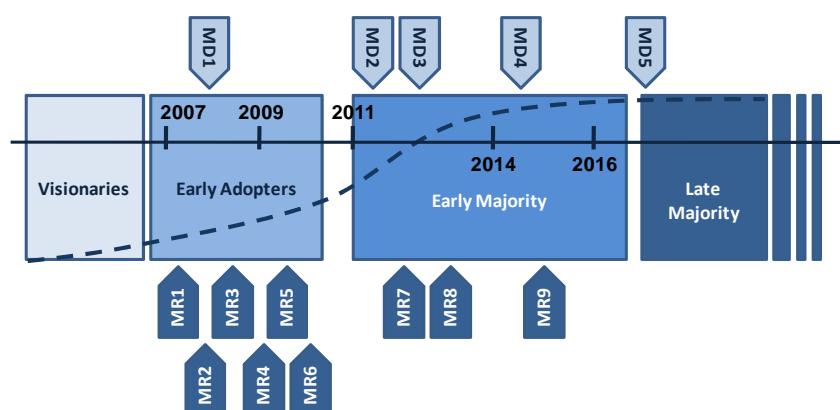


figure 2: Milestones and timeline for the roadmap of grids for health.

Figure 2 shows in a time-scale the position of the milestones. In the short term, it will be necessary to solve the main problems to deploy functional, usable and efficient pilot applications. This will require the availability of pilots, to deploy them on reliable infrastructures, partially on the health domain and partially on the e-Science domain, with a clear exploitation model, to solve the problems on data security and to link existing data repositories efficiently on the system.

These pilots will be the success stories needed to prepare the jump to the “main street” where the largest share of the users is located. Thus, scalability, improved usability and better semantic integration are needed. It would then be necessary for legal regulations to reach maturity and for enough references to be developed to ensure that the use of grids for health is consistent.

The jump to the “late majority” is quite complex and long. Even for disruptive technologies, such as the Web, this late majority has required more than ten years to start to be targeted. It is therefore out of the scope of this roadmap.

9. CONCLUSION

The current status of grids for health is in a decisive stage. The application of grid to health imposes several constraints in terms of robustness, security, Quality of Service, compliance to regulations and scalability that are also shared by other application areas, leveraging the interest in the development of such issues.

Epidemiology is a fairly representative case study, in that it involves large data management, data integration and intensive computing. It constitutes a very challenging target that can produce a clear revolution on the need and usage of e-Infrastructures.

This document agrees with the roadmaps presented in D3.3 and D4.2, and provides more details with respect to the use case of epidemiology. It describes the specific needs for persistent infrastructures for grid epidemiology and the associated middleware. While in terms of the technology, computational grids are a first step, the nature of epidemiological research will require mainly the use of data grids, which are thus the key issues to be considered. Finally, the concept of knowledge grids (i.e., the third major milestone) will come up with the generalisation of the federation of data and processing. In order to reach this “research knowledge grid”, along with the integration of general medical informatics protocols, such as image and EHR-related, integration of epidemiological tools and procedures should be considered.

The analysis done in this document with respect to the natural evolution of IT-based developments(feasibility studies, pilots, early deployments, production) reveals that effective grids for health pilots are still a very important issue to address. Meanwhile, early deployment could be targeted provided that several technological and legal issues are covered. The deployments on restricted environments are feasible, although do not exploit sufficiently the advantages of grids for health at large.

The first step is to target small but significant communities. Policy makers should foster, on one hand, the development of exploitation and Quality of Service models in order to make use of current e-Infrastructures, as well as standardising and deploying the already effective security and privacy guard technologies developed in several research lines, which have not been intensively tested to reach the desired level of maturity. On the other hand, pilots and early deployments still need to be developed in order to reach a wider audience and to bring more data and users to the epidemiological grids for health.

Long term research should take into account the problems of larger communities, focusing on solving scalability, integration of data and widely used tools at a large scale, as well as working on the jurisprudence that could give the stability in the ethical and legal framework. This will solve the issues outlined in the conclusions of D4.2 concerning the management of responsibility, the impact of legislation and the reimbursement through service business models.

Grids for health need standardisation and development, but mainly, grids for health need coordination with other medical informatics developments. From the synergy between the maturity and penetration of medical informatics in health, and the increased performance of grids, new health challenges in the study of treatments, the quick reaction against epidemic occurrences, or the long-term surveillance of the population’s health could be managed effectively and efficiently, for a significant added value.