

Proposing a roadmap for HealthGrids

Vincent Breton¹, Ignacio Blanquer², Vicente Hernandez², Yannick Legré¹ and Tony Solomonidés³

¹*LPC, CNRS-IN2P3, Campus des Cézeaux, 63177 Aubière Cedex, France*

²*Universidad Politecnica de Valencia,*

³*University of West England, Bristol, Coldharbour Lane, Bristol BS16 1QY, United Kingdom*

Abstract. With the regular progress of technology and infrastructures, a growing number of grid applications are developed and deployed for life science and medical research. At the last HealthGrid conference in April 2005 in Oxford, many groups described successful usage of grids for compute intensive calculations. Very large scale deployment of a biomedical application in the area of drug discovery has been achieved on EGEE during 2005. On the other hand, beside a few pioneers, very few data grids have been deployed so far and knowledge grids are still at a conceptual level. This situation is expected to evolve quickly as many projects are focussed on developing data management services and knowledge management tools relevant to biomedical sciences. At this stage, it is important to identify the potential bottlenecks and to define a roadmap for the wide adoption of grids for healthcare. This article presents an analysis of the present adoption of grids for biomedical sciences and healthcare in Europe: it identifies bottlenecks and proposes actions that will be further assessed within the framework of the SHARE European project dedicated to the definition of a roadmap for HealthGrids.

1. Introduction

The emergence of grid technology opens new perspectives to enable interdisciplinary research at the cross roads of medical informatics, bioinformatics and system biology impacting healthcare.

A HealthGrid is an environment where data of medical interest can be stored, processed and made easily available to the different actors of healthcare, physicians, healthcare centres and administrations, and of course citizens. If such an infrastructure offers all guarantees in terms of security, respect for ethics and observance of regulations, it allows the association of post-genomic information and medical data and opens up the possibility of individualized healthcare [1].

This enabling integration tool for medical applications provides also the infrastructure for navigation space. Access to many different sources of medical data, usually geographically distributed, and the availability of computer-based tools that can extract the knowledge from these data are key requirements for providing an equal healthcare provision of high quality.

Born from discussions between grid application developers and medical informaticians, the concept of HealthGrid is now 3 years old. The yearly HealthGrid conferences are an opportunity to evaluate the growing usage of grids for life science and medical research. They allow also identifying the obstacles to a wider adoption. In

chapter 2, we illustrate the concept of HealthGrid on a very simple example where we highlight key issues related to the deployment of grids for healthcare. In chapter 3, we propose an analysis of the present adoption of grids by biomedical sciences. Recent accomplishments are also critically reviewed. Based on this analysis, we will propose some actions to address the present bottlenecks. In chapter 5, we will describe the SHARE project which aims at proposing a roadmap for HealthGrids. While the SHARE project will address all dimensions of a roadmap including legal, social and ethical issues, this paper will restrict itself to technical issues.

2. Concept of HealthGrid: illustration by an example

One of eHealth important goals is to allow the transfer of information between hospitals in Europe. A very simple example is a practitioner in Hospital 1 needing to transfer a patient Electronic Health Record (EHR) to Hospital 2 (figure 1). In this very simple use case, we consider that there is no legal issue for sake of simplicity.

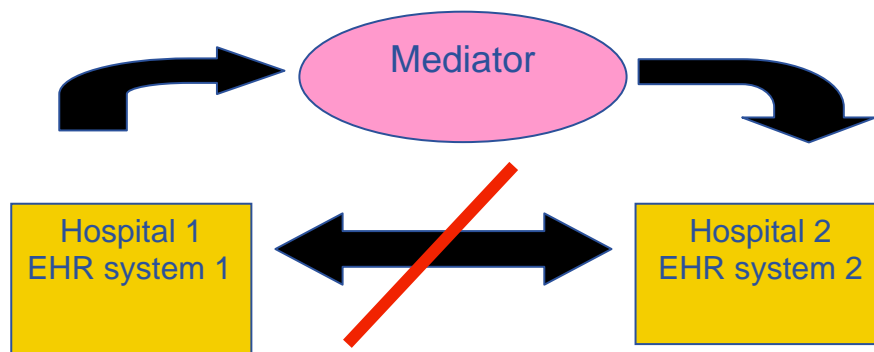


Figure 1.

To achieve this transfer, a first simple idea is to use a standard File Transfer Protocol. It will work only if the two hospitals EHR systems have the same data model. The EHR data model describes the content of each data field. If the data models are different, a mediator is needed to *interpret* the data coming out of the EHR system 1 and to *translate* it into the format used by the EHR system 2. This mediator is able to handle this translation provided the data models used by the 2 EHR systems are known. The mediator can not invent information so if the 2 EHR systems have different data fields, some data fields will not be filled or some data field may be unused and the data lost.

This use case illustrates very simply different needs for the transfer of information between healthcare centres in Europe:

- for Hospital 2 to request a patient record, it has to provide an identifier for this patient. This illustrates the need for a unique patient identifier allowing querying patient records while preserving their anonymity.
- For the mediator to be able to translate patient record stored in Hospital 1, the data models of both EHR systems 1 and 2 must be known. Even if the two EHR systems can be completely different, the mediator will reorganize information as needed. EHR data models must be made publicly available.

Even the precise definition of the data fields must be provided in order to allow the reliable translation. This requires a common vocabulary to define the data fields.

- EHR system 1 has most probably specific data fields which have no equivalent for EHR system 2. Therefore some data fields will not be filled for the patient record at Hospital 2. However, it is of utmost importance to have the most important data fields filled. This requires an agreed patient summary with an agreed vocabulary to describe it.

The HealthGrid is going to be the environment on which services and resources needed to enable the above picture are provided:

- when hospital 2 looks for a patient record, it does not know necessarily that hospital 1 is holding the patient record it is looking for. An information service is needed to provide the localization of the patient records in Europe. This critical service must be constantly updated and needs to be replicated in order to avoid being a single point of failure. The information service needs to have the relevant security features so that only authorized healthcare professionals are allowed to consult it.
- Another information service is needed to provide the data models for each healthcare centre storing medical patient record. This information service is consulted by the mediator before translating a patient record
- a network of mediators is needed to address all the requests for patient record transfers in Europe. These mediators must also be updated to follow the evolution of the EHR data models

This very simple example illustrates the role of a HealthGrid and the bottlenecks towards its deployment, including the interoperability of HER systems and the definition of a unique patient identifier and an agreed patient summary. These issues are presently addressed at a European level.

3. A perspective on the present adoption of grids

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

- infrastructure projects aim at offering a stable distributed environment for scientific production. Examples of such infrastructures are EGEE [2] and DEISA [3] in Europe. 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 science and healthcare. Examples of such projects are SIMDAT [4] and MyGrid [5]
- End user projects focus on specific life science or healthcare issues and integrate grid technology wherever they feel relevant. Examples of such projects are Mammogrid [6] and GEMSS [7].

3.1. Adoption of grids for biomedical sciences

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

3.1.1. Life science

Molecular biologists are facing a daunting challenge: the relevance of their research requires a constant access to the databases containing all the knowledge acquired up today. Comparative analysis is a mandatory step in most of the molecular biology data analysis workflows. This analysis has to be frequently repeated to keep up with the exponentially growing volume of data stored in the databases. Comparative analysis is often the first step of complex workflows needed to extract information from the data in genomics, transcriptomics and proteomics. At a basic level, grids can help distribute the databases in order to make them accessible to the biologists [11] and provide the computing resources required by data analysis. Bioinformatics portals like GPS@ [9] are presently under development on top of grid infrastructures.

The grid technology is also very promising to address biological data complexity. Indeed, the last years have witnessed the development of hundreds of databases providing specific representations of biological data. Interoperability of these databases is a key to the development of integrated approaches needed to start modelling living organisms. Projects such as Embrace [8] focus on addressing this interoperability issue using the grid technology.

Other projects such as MyGrid [5] have been developing tools and environments to ease the design of data analysis workflows for biologists. The next step is to achieve the integration and deployment of these high level interfaces on grid infrastructures so as to offer to the biologists the data and computing resources needed for their analysis.

3.1.2. Medical research

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

Attempts to use grids to confront patient medical and biological data are presently under exploration in several projects presented at this conference. The success of these approaches depends again on the capacity of the grid to provide the tools needed to manipulate these data.

3.1.3. Drug Discovery

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

testing from a few millions to a few hundreds.

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

In silico drug discovery on grids is a growing field. Grids like EGEE are ideally suited for the first step where docking probabilities are computed for millions of ligands. Grid relevance has been clearly demonstrated during the summer 2005 by the WISDOM initiative on malaria [12] where 46 million ligands were docked for a total amount of 80 CPU years (1 TFlop during 6 weeks).

A foreseeable future is to enable a complete in silico drug discovery pipeline on the grid. Such pipeline would allow very quickly identifying promising compounds. The first stage, which will be explored notably within European projects like BioInfoGrid, EGEE and Embrace, is the deployment of a virtual screening platform that would take advantage of the European grid infrastructures for docking and of a supercomputer for Molecular Dynamics computations.

3.2. Adoption of grids for healthcare

Adoption of grids for healthcare is still in its infancy. There are many reasons to this situation. A first obvious reason is that grid technology is still immature and is neither robust nor secure enough to offer the quality of service required for clinical routine. Another important reason is that all grid infrastructure projects are deployed on National Research and Education Networks which are separate from the networks used by healthcare structures. Another major obstacle is the legal framework in the EC member states which has to be evolved to allow the transfer of medical data on a European HealthGrid.

This did not stop pioneer projects to explore and demonstrate the potential impact and relevance of grids to address such outstanding healthcare issues as the early diagnosis of breast cancer [6] or to improve radiotherapy treatment planning [7].

Grids are expected to bring a significant added value in the development of individual medicine which requires the exploitation of biological and medical data, but this is still a research field.

Adoption of grids for healthcare will follow their adoption for life sciences and medical research provided the legal and ethical framework of the member states allows their deployment.

4. Technical bottlenecks and proposed actions for a wider adoption of grids

The HealthGrid vision relies on the setting up of grid infrastructures for medical research and healthcare. The present bottlenecks towards this vision are the following:

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

The HealthGrid vision can not be achieved without a close collaboration of the projects developing grid middleware, deploying grid infrastructures and developing end-user oriented biomedical grid applications.

4.1. Technical bottlenecks

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

4.1.1. Lack of grid data management services

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

4.1.2. Lack of grid nodes in healthcare centres

Another bottleneck is related to the installation and maintenance of grid nodes in healthcare centres. Such deployment is still in its infancy because the configuration of a grid node is rather complex and requires significant manpower. Moreover, as stressed above, secure services for data management are still under development.

4.1.3. Lack of standards in medical informatics

Chapter 2 of this paper illustrated on a very simple example the role of a HealthGrid to exchange information between two hospitals in Europe. It also highlighted the need for a unique patient identifier allowing querying patient records while preserving their anonymity, for EHR data models publicly available and for an agreed patient summary with an agreed vocabulary to describe it. Work is under way at a European level to address these issues. For the HealthGrid vision to happen, standards must be agreed upon in the medical informatics community. This precludes the development of applications obeying to these standards, using the grid services and available from the grid nodes located in the healthcare centres.

4.2. Organizational bottlenecks

4.2.1. Insufficient technology transfer between EC projects

As a consequence of the technical bottlenecks previously identified, very few projects led by biomedical end users are deployed on the European grid infrastructures available today. This is due most notably to the limited data management services offered by the infrastructures, their still user-unfriendly interfaces and the lack of information and training on grids in the biomedical community. Interesting data management services are under development by some technology oriented projects but the mechanism by which they will be deployed on existing grid infrastructures is unclear.

4.2.2. Lack of coordinating bodies

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

4.3. First proposed actions

We recommend the creation of a dedicated infrastructure for medical research. From the beginning, the infrastructure should offer services such as database federations, distributed computing and data replication. Nodes of this infrastructure should be located in hospitals and healthcare centres. This infrastructure should host pilot medical research applications.

A model for such an infrastructure is the BIRN project [13] of the National Institutes of Health's National Center for Research Resources.

Launched in 2001 as an initiative, the BIRN 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 centered 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.

We also recommend to set-up a HealthGrid coordination body with a real power to make choices for standards and middleware deployment on this dedicated infrastructure.

5. Proposing a roadmap for HealthGrid: the SHARE project

European leadership on grid deployment is recognized at a world level. This leadership is also internationally acknowledged in the area of HealthGrid. The concept of grids for health was born in Europe in 2002 and has been carried forward through the HealthGrid initiative. This European initiative has edited, in collaboration with CISCO, a short version of the 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.

Starting from the conclusions of the White Paper, the EU funded Share project aims at identifying the important milestones to achieve the wide deployment and adoption of HealthGrids in Europe. The project will devise a strategy to address the issues identified in the action plan for a European e-Health [10]. It will also set up a roadmap for technological developments needed for successful take up of HealthGrids in the next 10 years. The widest audience will be solicited for comments and validation during most of the preparation phases.

Grid infrastructures are designed at a world level and the consortium is therefore planning to involve at a later stage American and Asian participants in order for the resulting roadmap to have relevance beyond Europe.

The HealthGrid roadmap will cover the domain of RTD and uptake of Grid applications in healthcare comprehensively, including infrastructure, security, legal, financial, economic and other policy issues.

Each section of the roadmap will detail actions to be taken in terms of objectives and possible methods or approach as well as recommended milestones for completion, stakeholders responsible, appropriate methods of coordination etc.

As a first view, the sections of the roadmap will cover the following domains: networks, infrastructure deployment, Grid operating systems, services to end users, standards requirements, security measures, legislative development and economic issues.

The conceptual work during the start-up phase of the project will also specify in detail both the general scope and specific features of the roadmap. The roadmap will focus on identifying requirements for further research and technology development, but it will also sketch a realistic picture with respect to desirable applications/ICT implementations and indicate which technologies may have the potential to make a substantial contribution in this context. This will be supported through the presentation of good practice examples. To ensure that the RTD roadmap ultimately to be generated will actually yield positive results and desired impacts it will be based upon and, wherever possible, justified by empirical evidence from the research domain and a bottom-up assessment involving relevant stakeholders. In a sequential process, relevant research communities and communities of practice at EU, national and global levels will be joined up to enable an iterative refinement and extension of the initial road map.

The HealthGrid roadmap is to be developed in a three stage process based on two iterations (roadmaps I & II) and one synthesis, resulting in a full-scale validated and integrated roadmap. The technical roadmap component has to address the different levels relevant to such an infrastructure:

- The network must provide end-to-end high bandwidth connectivity between the Grid nodes. The services offered to the HealthGrid users will ultimately depend on the service level agreements between the network providers and the resources providers at each of the HealthGrid site.

- The Grid infrastructure is made of resources distributed geographically on the different Grid nodes. These resources share the Grid common operating system which is the hidden low-level part of the middleware, called sometimes “underware”. The services offered to the HealthGrid users depend on the functionalities offered by this operating system, the amount and nature of the resources made available to the Grid. At this “underware” level, most of the functionalities needed are common to all Grid infrastructures just like the DOS operating system used for PCs in hospitals is the same as for all other PCs. However, HealthGrid exceeds already e-science requirements at this level in areas such as security features for Access, Authentication and Authorization, performances and quality of service.
- The tools offered to the HealthGrid end users are made available through Grid interfaces. They are specific to medical research and healthcare. Their relevance, conviviality and performances are keys to the HealthGrid success. User friendliness of these services requires calling high level services taking care of knowledge management which themselves call lower level Grid services for access to distributed data and resources. Most of these high level middleware services, sometimes called upperware, are specific to HealthGrids.

In the definition of the roadmap, particular attention must be paid to security and standards in the choice of HealthGrid operating system and technology:

- Security is not a choice but a mandate for HealthGrids. Security is an issue at all technical levels: networks need to provide protocols for secure data transfer, Grid infrastructure needs to provide secure mechanisms for Access, Authentication and Authorization, sites for secure data storage. The Grid operating system needs to insure access control to individual files stored on the Grid. High level services need to properly manage legal issues related to the protection of medical data.
- Standards must be respected and promoted on the road to HealthGrids. Standards are needed for European wide compatibility and faster take-up. High level middleware services dealing with medical data need to conform to Grid standards but also medical informatics standards such as HL7 or DICOM.

RTD activities to address issues limiting the full exploitation of HealthGrid technologies across Europe will be structured into a first version of the technology roadmap to be discussed at the HealthGrid conference in Valencia and submitted to the European Commission in the fall of 2006. .

The roadmap will identify key short-term (2-5 years) and medium-term (4-10 years) RTD needs to achieve deployment of e-health systems in a Grid environment. It will also analyse unsolved RTD issues arising in the context of realistic approaches to priority clinical and public health settings (reflecting on models of use, benefits expected, concrete application experience and lessons learned; relevance of open source model) and detail actions to be taken for networks, infrastructure deployment, Grid operating systems, services to end users, standards requirements and security measures

This first roadmap will recommend a number of case studies on specific aspects of technology issues requiring further investigation because they are identified as potential bottlenecks.

Its recommendations will be validated against several use case scenarios. As a result of this validation, new technological bottlenecks should be identified, requiring further RTD activities and a revision of the proposed technology roadmap.

The revised roadmap will implement a process to present, discuss, and validate the identified RTD needs and the resulting roadmap with the relevant RTD community. Actors of the Grid development will be asked to validate and prioritise areas of future work on the basis of highest expected short and medium term impact. Their endorsement is critical to the successful achievement of the proposed roadmap at the levels which are hidden to the user: networks, infrastructure deployment and Grid operating systems. Security as well must be implemented at all levels. The project technology partners will present and promote the revised roadmap in the different consortia where they are involved (EGEE, DEISA, UK e-science, Globus, national Grid initiatives...) to trigger RTD activities identified.

6. Conclusion

This paper aimed at giving an overall analysis of the present status of HealthGrids in Europe. Through the simple example of the transfer of a patient health record between two hospitals, we have demonstrated the importance of a unique patient identifier allowing querying patient records while preserving their anonymity, the need for EHR data models publicly available and for an agreed patient summary with an agreed vocabulary to describe it as well as for interoperability mechanisms.

We have also stressed the need for improved data management services on grid infrastructures. Indeed, the last HealthGrid conference has witnessed several success stories in the usage of grids for compute intensive tasks but data grids are still to come.

The analysis work started in this document will be further developed and enlarged to social, legal and ethical issues within the framework of the EU funded Share project in order to produce a roadmap for the adoption of HealthGrids in Europe.

Acknowledgments

Many of the ideas expressed in this document have been further refined in discussions with members of the HealthGrid consortium. We particularly acknowledge fruitful exchanges with Veli Stroetman and Sofie Nørager.

References

- [1] V. Breton, K. Dean and T. Solomonides, editors on behalf of the HealthGrid White Paper collaboration, "The HealthGrid White Paper", Proceedings of HealthGrid conference, IOS Press, Vol 112, 2005
- [2] Fabrizio Gagliardi, Bob Jones, François Grey, Marc-Elian Bégin, Matti Heikkurinen, "Building an infrastructure for scientific Grid computing: status and goals of the EGEE project". Philosophical Transactions: Mathematical, Physical and Engineering Sciences, Issue: Volume 363, Number 1833 / August 15, 2005, Pages: 1729 – 1742, DOI:10.1098/rsta.2005.1603
- [3] DEISA, <http://www.deisa.org>
- [4] SIMDAT, <http://www.scai.fraunhofer.de/simdat.html>
- [5] MyGrid, <http://www.mygrid.org.uk/>
- [6] Mammogrid, <http://mammogrid.vitamib.com/>

- [7] GEMSS, <http://www.gemss.de/>
- [8] Embrace, <http://www.embracegrid.info>
- [9] GPS@, C. Blanchet et al, proceedings of HealthGrid conference, IOS Press, Vol 112, 2005 - <http://gpsa.ibcp.fr/>
- [10] 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
- [11] J. Salzemann, V. Breton, N. Jacq and G. Le Mahec, "Replication and Update of molecular biology databases in a grid environment", submitted to FGCS, 2006
- [12] N. Jacq, J. Salzemann, Y. Legré, M. Reichstadt, F. Jacq, M. Zimmermann, A. Maas, M. Sridhar, K. Vinodkusam, H. Schwichtenberg, M. Hofmann and V. Breton, *In silico* docking on grid infrastructures: the case of WISDOM, submitted to FGCS, 2006. <http://wisdom.eu-egge.fr>
- [13] BIRN: <http://www.nbirn.net/>