From HealthGrid to SHARE: A Selective Review of Projects

Mark OLIVE, Hanene RAHMOUNI and Tony SOLOMONIDES[1]
CCCS / CEMS Faculty / UWE, Bristol / Coldharbour Lane / Bristol BS16 1QY / UK

Abstract. The SHARE project (Supporting and structuring Healthgrid Activities and Research in Europe) is an EC funded specific support action to define a roadmap for future healthgrid research, highlighting opportunities, obstacles and potential bottlenecks. The aim is identify technical, legal and ethical issues that would affect the wide deployment of healthgrids in the medical research community. The initial technical roadmap has proposed a series of technology, standards and deployment milestones, starting with the testing and development of a reference implementation of grid services, and ending with the deployment of a knowledge grid for medical research.

In this paper we review a number of projects from pre-healthgrid to the second generation of healthgrid projects and retrospectively consider their achievements and issues in the light of the first SHARE technical road map.

Keywords. healthgrid, e-health, grid applications

Introduction

In order to address the key technical issues that would affect the wide deployment of healthgrids for medical research, two technology milestones have been defined by SHARE for the development of grid middleware and services, two more milestones as examples of required grid standards for bioinformatics and medical informatics, and three deployment milestones increasing in complexity and scope.

In order to identify issues that would affect the development and deployment of healthgrids, we have examined a number of key national (UK) and European projects. This included medical applications designed to run on early generic grid environments (or 'pre-healthgrid'), the first true generation of healthgrid projects tackling generic medical problems that presented novel challenges, and the current second generation.

This paper will include technology and development issues identified in open publications from the following projects: GEMSS (Grid-Enabled Medical Simulation Services), CLEF (the Clinical e-Science Framework), ECIT (the ESHRE Classification of Infertility Taskforce), eDiaMoND (Digital Mammography National Database), IBHIS (Integration Broker for Heterogeneous Information Sources), GIMI (Generic Infrastructure for Medical Informatics), WISDOM (Wide In Silico Docking On Malaria), Health-e-Child and Integrative Biology.

This work has been carried out as part of the SHARE project. Work on the technical road map, and in particular the phasing into technology and deployment

[1] Corresponding Author: Tony Solomonides, CCCS, CEMS Faculty, University of the West of England, Coldharbour Lane, Bristol BS16 1QY, United Kingdom; E-mail: Tony.Solomonides@uwe.ac.uk.

milestones, has been led by our colleagues at CNRS-IN2P3. We are especially grateful to Vincent Breton and Yannick Legré at CNRS-IN2P3, Vicente Hernandez and Ignacio Blanquer at UPV, Petra Wilson formerly at EHMA, and Isabelle Andoulsi and Jean Herveg at FUNDP.

1. Milestones

A total of seven technical milestones have been defined for the first roadmap. Three deployment milestones were defined for the creation of demonstration production environments for a computational grid (MD1), a data grid (MD2) and a knowledge grid (MD3) for medical research. Two milestones dealing with specific standardisation efforts for medical imaging (MS1) and electronic health records (MS2) for grids were also defined. There is some overlap between the standardisation milestones and the second deployment milestone of a data grid production environment to represent concurrency and a cyclical relationship.

Additionally, two ongoing technology research efforts were identified and have been added as technical milestones prior to the creation of the first integrated roadmap, which will also incorporate ethical, legal and socio-economic considerations. These were the testing of grid middleware (based on web services, and with sufficient job and data management) with medical applications for robustness, scalability and compliance with EC country laws regarding reliability, recovery and backup (MT1), and the production of a robust, user-friendly and open source distribution of healthgrid services with appropriate user support (MT2).

In what follows, we consider the projects we have studied in depth with reference to these seven milestones.

2. Review of Projects Against Milestones

2.1. Milestone MD1

We begin by considering the first SHARE road map, whose first deployment milestone is the development of a demonstration computing grid production environment for the medical research community. The importance of a robust, well maintained infrastructure was stressed by eDiaMoND, with appropriate redundancy to address issues of availability. This is a necessity for any sustainable grid infrastructure, and so would naturally be a requirement of the first deployment milestone.

The necessity (and cost) of providing a suitable network infrastructure will certainly be an issue for the deployment of data grids in medical research centres. Due to the volume and size of digital mammograms, which must satisfy both legal and practical requirements, eDiaMoND required gigabit Ethernet, expensive monitors and high speed, high capacity storage. There is also the issue of medical data not typically residing in a convenient central location, but distributed among firewall protected hospital databases. Another concern was how hospital executive managers could be convinced to allow external systems to penetrate their firewalls.

The necessity of providing a user friendly interface for non technical users was recognised by a number of the projects examined. In particular, Integrative Biology, GEMSS and WISDOM saw hiding grid complexity and grid mechanisms as important.

However, WISDOM noted that considerable knowledge of grid mechanisms was required in order to quickly resolve unforeseen errors and failures, which would be a concern for projects where jobs are being submitted by users with limited computing skills, as may be the case in medical research centres or hospitals. Additionally, human supervision was considered mandatory, which could suggest that scalability would be an issue.

2.2. Milestone MT1

MT1 is the testing of grid middleware(s) with medical applications for scalability and robustness. A good example of a high throughput computing grid application is WISDOM, which demonstrated the capabilities of the EGEE (Enabling Grids for E-sciencE) infrastructure. However, a number of errors and efficiency issues were identified by both the project and EGEE reports. Data management has been identified as a bottleneck for biomedical data challenges on grid infrastructures that only support one Replica Location Service (RLS), such as the biomedical VO (Virtual Organisation) framework on the EGEE grid at the time of the WISDOM data challenge.

There were significant problems with automatic job resubmission both within WISDOM and the EGEE middleware at the time. WISDOM described this as a 'sink-hole' effect, which led to a significant number of aborted jobs and excessive job execution times. Checking, cancellation and resubmission of jobs had to be performed manually as a result, and automatic job resubmission was still absent from the second (avian flu) WISDOM data challenge.

WISDOM also highlighted a number of issues with grid Information Systems (IS) and Resource Brokers (RB) that will need to be addressed before the reference implementation can be finalised. The IS holds pertinent data to optimise performance and the RB makes decisions on the basis of this information. Ideally, an RB would use the IS data to select the best computing element for any job that arises. But if the IS fails or does not have the required information, the submission of jobs will be inefficient or the user will have to expend significant effort manually correcting the failure.

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

2.3. Milestone MT2

MT2 is the production of a reference distribution of healthgrid services, using standard web services technology. An important issue for this milestone will be the security mechanisms employed to protect medical and commercial data in a grid environment.

The earliest project examined, GEMSS, identified internal staff as the main security threat for grid-based healthcare. Through the use of use case scenarios, the eDiaMoND project went on to identify five different classes of user that represented a potential threat:

- Insiders making innocent mistakes, causing the accidental disclosure of confidential information
- Insiders who abuse access privileges

- Insiders who knowingly access information through spite or for profit
- An unauthorised physical intruder gains access to information
- Vengeful employees and outsiders

There is a need for each individual entity within a virtual organisation to be able to describe the access control policies associated with that site's data. The utilisation of the OASIS eXtensible Access Control Markup Language (XACML) is a partial solution to this problem, although the verbosity and complexity of XACML descriptions makes tool support essential. Another requirement is that it should be possible to give temporary access to data on a healthgrid. It is important that people given temporary access are not able to make unauthorised copies of restricted data. To address this, eDiaMoND proposed that data should have a lifetime, and would be deleted after use. If this were the case, it would then be following the fifth principle of the Data Protection Act. If temporary access to data is to be allowed, then either the policies must contain time-based information, or a secondary process would have to change the policies at the appropriate time. GIMI later proposed the use of Digital Rights Management (DRM) technologies to address this issue.

An aim of many grid projects, including Health-e-Child, is enabling 'single sign-on' for users, using a single identity credential that is valid for many infrastructures, communities and organisations. This is simply an issue of practicality; it would take extra time and effort for the user, and if they had to remember multiple passwords for different grid resources they could be tempted to write them on paper, opening the door to a security breach. However, eDiaMoND identified a number of issues with credentials that will need to be addressed. It is important that credentials are portable, so that researchers are able to move around their own centre or to access data when visiting remote sites without carrying a computer with them. Another problem associated with credentials is propagation – external services need to be trusted to pass users' credentials to other services. The typical grid mechanism for this task is the use of a proxy certificate, however the user has to rely on the intermediate services to use the proxy certificate as intended. Finally, it must be possible to revoke a set of credentials. This is especially important if doctors are carrying their credentials around with them. Current systems often rely on the service actively requesting lists of revoked certificates, leaving a window of opportunity for misuse. Use of the Online Certificate Status Protocol (OCSP) may reduce the window to a minimum but this has the drawback of requiring frequent communication.

Role Based Access Control (RBAC) is a common approach to authorisation in healthgrid projects. RBAC assigns permissions based on the 'role' of the user, e.g. 'medical researcher', 'district nurse', 'local doctor', etc. However, many healthgrid projects including IBHIS have found that RBAC, which might involve simply mapping a grid user to a local user account, is too inflexible for the health domain and does not provide fine-grained rules for access control, including how to support users with multiple authorisation profiles. The second IBHIS prototype used the Tees Confidentiality Model for finer grained rules, which was also used for mapping between security domains. The project also mentioned the need for managing changes in security policies and requirements, with user roles and authorisation levels changing dynamically to reflect organisational restructuring. eDiaMoND also mentions RBAC and proposed that the co-existence of local and global policies should be investigated, to support a flexible 'situated role-based access', where skills are delegated on a needs basis.

This leads to the issue of how to provide flexible, fine-grained access control, a recurring issue in grid based research. As mentioned by GIMI, this becomes problematic when securing services that are provided by a third party vendor that have their own access control mechanisms. These may not provide sufficiently flexible or fine-grained access control, and coordinating access control policies for many resources could be extremely difficult. Therefore, GIMI proposed that all access control for resources at a node should be determined by a single set of coordinated policies, and all access to existing vendor-provided web services will have to comply with these policies.

Another major task for this milestone will be the development of a stable and mature web services framework. IBHIS publications made several comments about a key component of web services, the Web Service Description Language (WSDL), used to describe the interface of a service. WSDL only describes a service in functional terms – its data types, methods, parameters, message format, etc. The WSDL standard lacks flexible, semantic, non-functional descriptions required for a dynamic service-oriented environment, such as descriptions of a service's security requirements and quality of service. Without semantic and non-functional descriptions, there can be confusion in the meaning of service and parameter names, and certain security considerations – a key concern when dealing with medical data – could be neglected. Ontology-based description languages, such as the DARPA Agent Mark-up Language for Services (DAML-S, now OWL-S) will provide a much more complete service description, but these are not fully mature and have limited tool and/or registry support.

IBHIS also noted a number of issues with the current version of UDDI (Universal Description, Discovery and Integration), which provides a registry for service discovery. The search functions in UDDI have only limited support for making automatic service selection decisions, cannot facilitate matching at the service capability level, and a key limitation of UDDI is that it does not provide semantic searching; it is essentially limited to keyword-based searching. It also does not capture relationships between entities, and is not able to infer relationships using semantic information.

Finally, IBHIS noted that the existing web services stack framework requires clarification in order to determine which technologies are usable at which level, and which are compatible with each other. The updated framework proposed by IBHIS is composed of protocols that use or extend WSDL, have roots in the semantic web's resource description framework and DAML-S, and include ebXML specifications.

Further development is currently underway in these areas, with WSDL-S supporting semantic descriptions, and WSDL 2.0 promising to include non-functional requirements. To facilitate semantic searching, IBHIS notes that UDDI's capabilities can be extended using OWL-S, and ways to address the other limitations mentioned are being examined for upcoming versions of UDDI.

2.4. Milestone MD2

The next deployment milestone is the development of a demonstration data grid production environment for medical research, including distributed storage and the querying of medical data at a distance. The distributed nature of the grid will be of particular relevance for this milestone, including how to deal with remote data sources, issues concerning the grid-based storage of medical data, and dealing with dynamic data providers and sources.

In addition to communication between clients and services, Integrative Biology discussed integration within and interaction with infrastructures that have their own security models. The authorisation mechanism for healthgrids should complement rather than conflict with these. The healthgrid projects examined here have typically chosen to use the authorisation mechanism that best suits their purpose rather than standardising on any particular one, which if not addressed could become an obstacle for inter-grid authorisation, or at least for single sign-on.

ECIT identified two security concerns with distributed storage. Local privileges may allow a user inappropriate access to sensitive data either locally or remotely, and data replication management can result in sensitive data being replicated and stored without permission, possibly from one country to another. According to ECIT, this may mean data on the grid will need to be censored rather than just encrypted to mitigate against data 'leakage'.

eDiaMoND notes that most system designers concerned with security and trust requirements fail to take organisational trust relationships into account. As a result, protection mechanisms may obstruct the effective use of the system and the activities it is designed to support. Where physical artefacts such as records and images are involved, the activity taking place can be overseen by others in the same domain – the work provides a "natural, locally visible account of itself"[7]. People also have a biographical familiarity with those they work with. Replicating this situation for digital artefacts is an important consideration, as it affects procedures for accountability, visibility of actions taken, informal practices and discussions, effective team working, and people's trust in the reliability and credibility of data and decisions. In particular the project recommended facilitating informal communication and collaboration between remote users.

2.5. Milestones MS1 and MS2

The standards milestones that begin during MD2 and continue until the start of the final deployment milestone are required for the sharing of electronic health records (EHRs) and medical images (in this case, using DICOM) on the grid. These milestones are only two examples of the bioinformatics and medical informatics standards that may need to be modified or extended in order for them to be used on a healthgrid.

eDiaMoND identified a large number of standards, sometimes conflicting and often overlapping, that must be disambiguated and adhered to for a grid deployed in the UK. There are NHS-wide standards such as the NHS data dictionary and compatibility with NHSnet, standards endorsed by the National Programme for IT (NPfIT) including STEP standards, e-GIF standards, BS7799 for information security management, and HL7. For example, the DICOM (Digital Imaging and Communications in Medicine) standard for medical images is endorsed by e-GIF, but overlaps with HL7 v2.

For medical imaging, DICOM as used by eDiaMoND and GEMSS has already been accepted by UK and worldwide bodies as the accepted standard for the acquisition, connection and storage of images. SMF (Standard Mammogram Form) has also become something of a de-facto standard for the standardisation of mammograms with different procedures, film types and processing systems.

Although this paper focuses on the technical roadmap, it is important to note that a number of important legal and ethical issues, such as pseudonymisation and the recording of patient consent, both examined by CLEF, will also require standards.

2.6. Milestone MD3

The final deployment milestone is the development of a demonstration knowledge grid production environment for medical research. The synthesis of knowledge from data will require sophisticated data integration, data mining, and image processing applications, and may also involve the use of techniques from artificial intelligence to derive relationships between data from different sources and in different contexts.

According to Health-e-Child, the integration of biomedical information will be complicated due to the fact that there is no universally accepted biomedical data model. An integrated health record such as the one proposed by Health-e-Child will need to make use of data that is not just semantically and syntactically heterogeneous, but also conceptually and temporally heterogeneous, posing a major challenge for data integration. Managing the distribution, acquisition, normalisation, aggregation, and curation of distributed heterogeneous data are all issues requiring further attention. How to cope with missing, incomplete, conflicting and uncertain data is also an issue that will need to be addressed for this milestone.

Image processing is another important technology for knowledge grids, and eDiaMoND has demonstrated that there has been considerable success in this area already. The normalisation of images (using SMF from Mirada, for example) has been successfully used to correct differences in images due to contrast, tube voltage, equipment and other variations. This also allows images to be compared more easily, which can aid radiographers in their diagnosis.

Although primarily an activity for the medical research community, the implementation of medical ontologies will be hugely important for a medical knowledge grid. These will allow relationships between concepts and nuances in meaning to be captured, greatly enhancing the opportunities for communication, knowledge sharing and machine reasoning. Ontology mapping discovery, mapping between ontologies, and semantic integration of biomedical data have been identified as bottlenecks for generating integrated case data, and are currently being explored by several projects, including Health-e-Child.

3. Conclusion

Mapping the issues identified in a selection of existing healthgrid projects against the SHARE project's initial technical roadmap has shown that there are considerable challenges for each of the milestones described. However, important progress has already been made, with the current generation of projects promising to deliver early examples of data grids and even knowledge grids for particular medical research areas.

Ethical, legal and socio-economic issues were not examined in this paper, but will continue to be a concern. Further research and standardisation in this area will certainly be required to ensure these do not hinder the efforts of future projects.

Given that development is occurring in all areas within individual projects, closer collaboration and technology transfer between EC projects will be essential for rapid progression through the milestones described here. Also, raising the profile of grids and publicising examples of successful projects in the wider medical community will be vital for the future adoption of grids for medical research.

References

[1] T Solomonides et al (Eds). From Grid to HealthGrid, IOS Press, 2005

[2] S Cox & D Walker (Eds). Proceedings of the UK e-Science All Hands Meeting 2005, EPSRC, 2005

[3] D Kalra. CLEF – De-identifying the EHR: building a resource for research, AHM 2003

[4] A Rector, A Taweel, J Rogers, D Ingram et al. Joining up Health and BioInformatics: e-Science meets e-Health, AHM 2004

[5] P Jeffreys. UK, Oxford. Life Sciences and Healthcare, Forum Engleberg 2004

[6] J Declerck. MammoGrid & eDiaMoND: grid-enabled federated databases of annotated mammograms, Mirada Solutions Ltd (Siemens)

[7] M Jirotka, R Procter, M Hartswood et al. Collaboration and Trust in Healthcare Innovation: The eDiaMoND Case Study, Computer Supported Cooperative Work (2005) 14, Springer

[8] M Brady, D Gavaghan, S Harris, M Jirotka et al. eDiaMoND: The Blueprint Document, 2005

[9] J Fenner, R Mehrem, V Ganesan, P Melas, & L Walton. Practical Experience of Grid-enabled Health Computing with the GEMSS Grid, AHM 2004

[10] S Benkner, G Berti, G Engelbrecht, J Fingberg, G Kohring, S E Middleton & R Schmidt. GEMSS: Grid-infrastructure for Medical Service Provision, Healthgrid 2004

[11] GEMSS Deliverable D1.3b – Final Evaluation & Validation, GEMSS Consortium, 2005

[12] J Fingberg et al. GEMSS Deliverable D6.3b – Edited Final Report, GEMSS Consortium, 2005

[13] D M Jones, J Fenner, G Berti, F Kruggel, R A Mehrem, W Backfrieder, R Moore & A Geltmeier. The GEMSS Grid: An Evolving HPC Environment for Medical Applications, Healthgrid 2004

[14] S Benkner, G Engelbrecht, W Backfrieder, G Berti, J Fingberg, G Kohring, J G Schmidt et al. Numerical Simulation for eHealth: Grid-enabled Medical Simulation Services, ParCo 2003

[15] A Simpson, D Power, M Slaymaker & E Politou. GIMI: Generic Infrastructure for Medical Informatics, Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)

[16] A Simpson, D Power, M Slaymaker & E Politou. Towards fine-grained access control for health grids, Proceedings of the Ottawa Workshop on New Challenges in Access Control, 2005

[17] Health-e-Child Proposal – Part B, March 2005

[18] J Freund, D Comaniciu, Y Ioannis, P Liu, R McClatchey, E Morley-Fletcher, X Pennec, G Pongiglione et al. Health-e-Child: An Integrated Biomedical Platform for Grid-Based Pediatrics, Healthgrid 2006

[19] Enabling Grids for E-sciencE II (EGEE-II) Proposal – Part B, August 2005. Available at: http://delphi.pd.infn.it/~mmazzuca/proposal-25-8/EGEE-II-PartB-25-08-2005.pdf

[20] IBHIS 2nd Annual Review (2003-2004). February 2004

[21] I Kotsiopoulos, J Keane, M Turner et al. IBHIS: Integration Broker for Heterogeneous Information Sources, 27th Annual International Computer Software and Applications Conference, Dallas, 2003

[22] M Turner, F Zhu, I Kotsiopoulos, M Russell, D Budgen et al. Using Web Services Technologies to create an Information Broker, 26th International Conference on Software Engineering, Scotland, 2004

[23] K Bennett, N Gold, P Layzell et al. A Broker Architecture for Integrating Data using a Web Services Environment, First International Conference on Service-Oriented Computing, Italy, 2003

[24] M Turner, D Budgen & P Brereton. Turning software into a service, IEEE Computer, Vol. 36, 2003

[25] F Zhu, M Turner, I Kotsiopoulos, K Bennett, M Russell, D Budgen et al. Dynamic Data Integration Using Web Services, 2nd International Conference on Web Services, USA, 2004

[26] D Gavaghan, S Lloyd, D Boyd, P Jeffreys, A Simpson et al. Integrative Biology – exploiting e-Science to combat fatal diseases, AHM 2004

[27] D Mac Randal, D Gavaghan, D Boyd, S Lloyd, A Simpson & L Sastry. Integrative Biology – Exploiting e-Science to Combat Fatal Diseases, ERCIM News No. 60, January 2005

[28] IBVRE Infrastructure – Initial Design Report, October 2005

[29] F Jacq, F Harris, V Breton, J Montagnat, R Barbera et al. First revision of EGEE application migration progress report (DNA4.3.2), EGEE Collaboration, September 2005

[30] V Breton, N Jacq & M Hofmann. Grid added value to address malaria, Proceedings of the 6th IEEE International Symposium on Cluster Computing and the Grid (CCGrid), 2006

[31] H C Lee, J Salzemann, N Jacq, H Y Chen et al. Grid-enabled High-throughput in silico Screening against influenza A Neuraminidase, IEEE Transactions on Nanobioscience, Vol. 5, N°4, 2006