Grid Infrastructures Supporting Paediatric Endocrinology across Europe

Anthony Stell, Richard Sinnott, Oluwafemi Ajayi
National e-Science Centre
University of Glasgow, UK
a.stell@nesc.gla.ac.uk

Abstract

Paediatric endocrinology is a highly specialised area of clinical medicine with many experts with specific knowledge distributed over a wide geographical area. The European Society for Paediatric Endocrinology (ESPE) is an example of such a body of experts that require regular collaboration and sharing of data and knowledge. This paper describes work, developed as a corollary to the VOTES project [1] and implementing similar architectures, to provide a data grid that allows information to be efficiently distributed between collaborating partners. This infrastructure also allows a variety of analyses to be run over the data-sets including addressing issues such as crossing domain boundaries and negotiating data access policies between administrations that only trust each other to a limited degree.

1. Introduction

An emergent result of the ever-shrinking, Internet-enabled world is the need for linking data sets distributed across technological and geographical boundaries. Nowhere is this need more apparent than in the various specialised areas of clinical medicine, where distributed experts require infrastructures for sharing data sets and collaborating with one another. Of paramount importance to the development of these infrastructures are ethical concerns and, as far as possible, ensuring their satisfactory realisation through associated security infrastructures.

The requirement for a technology infrastructure to underpin such data linkage and sharing is rapidly evolving from a desirable product to a necessity for future clinical research. Grid Computing addresses this requirement through the development of "data grids" – distributed groups of entities (commonly referred to as "Virtual Organisations") that facilitate data sharing, whilst enforcing rigid yet flexible security policies that allow parties to collaborate whilst maintaining a limited level of trust between each other.

Paediatric endocrinology has been identified as a specialised clinical area with distributed expertise that would greatly benefit from just such a technological infrastructure. The European Society for Paediatric Endocrinology (ESPE) [2] is composed of some of the leading experts in this particular field. The sites involved have isolated individual computer systems in place that serve the needs of local clinicians, specialists and general practitioners, and are located at variously distributed centres throughout Europe – Rotterdam in the Netherlands, Pisa in Italy, Luebeck and Kiel in Germany, and Glasgow and Cambridge in the UK. This distribution of centres, which provide a multitude of local services yet have this specialised data in common, exemplifies a particular issue that Grid technology deals with directly: the notion of limited trust between collaborating parties.

In this sense, the partners that are collaborating in a virtual organisation are only partners for the duration of that collaboration. The assumption is that the VO has a transient lifetime and that current partners could be potential competitors for other projects. As such, it is necessary to only give access to the data-sets that are required for the current collaboration. The notion of context must also be supported so that information is controlled relative to how it could be used in other projects. These are non-trivial research questions that provide the driving force for study into Grid technologies.

This paper describes the work that would be involved in implementing such a system, the

nature of the data and the schemas concerned, the detailed security aspects that must be considered, and the implementation work that has been conducted so far.

2. Distributed Paediatric Data

The paediatric data involved in the ESPE project focuses on the aspects of congenital anomalies occurring in children, the analysis of which essentially defines the gender of the child. The data focuses on the primary, secondary and tertiary causes of disorders of sexual development (DSD), with a DSD register that encapsulates the symptoms and diagnoses for the various patients. The individual data sets and computer systems that exist across Europe have been developed largely in response to the needs of local clinicians. As such, the schemas and data classifications that have evolved within each of the different sites, though describing broadly the same data, have many features idiosyncratic to their own region.

This classification issue is one of the central challenges when attempting to federate data and presents that initial building block upon which all other features must be founded (e.g. how can one enforce a security policy upon an object when the exact concept of how to classify and describe that object and its context is ill-defined?)

There are two main paradigms that could potentially solve the question of matching heterogeneous data descriptions: global schemas and ontologies.

The idea behind a global schema is that an over-arching and static definition exists that matches every possible data description to a common schema. This has the benefit of providing a clear and un-ambiguous classification of data objects within this field. However it is also a highly static solution – requiring regular maintenance of what could potentially be a large and unwieldy document. Scope for dynamically adding new resources to this setup is limited by the need to continually update the global schema model itself.

The construction of an ontology is to define an interface that allows differing schema sets to provide a common *lingua franca* and structuring of concepts and terms used. The benefit behind this is that the interface can be defined as per the needs of the schemas being added or removed, without impacting on other interfaces. However, it does also require an overhead of work in matching one schema to another each time another is added.

Several global initiatives are underway to attempt to standardise the data description languages currently in use in the clinical domain. SNOMED-CT [3] and HL7 [4] represent two of these standards. However, much work is still required before they are to be adopted on a widespread scale throughout the clinical community. It is the case that currently a range of standards exist for describing clinical data sets from imaging standards, disease classifications, drug classifications, right through to operational procedure classifications [13-15]. It is the case that data Grid infrastructures are required to overcome these differences and provide seamless access to clinical resources.

No matter how it is achieved, the motivation of having a classification to marry these various descriptions together is obvious – not only can data access across the distributed sites be effectively policed, but a variety of functions can be performed over the enhanced data-set that will bring immense value to the boundaries of knowledge within this field – e.g. statistical calculations, the ability to map visual graphs and correlations, and the ability to highlight similarities and discrepancies in the trends from differing regions.

Given the nature of the data being federated, the sensitivity must be treated as extremely high and the security and safe-guarding of this data, and the interests of the patients involved, must be paramount in all discussions of system design and implementation.

3. Security

As with all clinical data, security – particularly the security of the patient's interests – is paramount. The main challenge in enforcing security policies across distributed, heterogeneous environments, is where to strike the balance between having a rigid enough policy to effectively prevent system abuse, and being flexible enough to allow the benefits of data federation to be realised (e.g. the most secure system would be a locked box disconnected from any network – but this would be of little use for data sharing or collaboration).

3.1 "AAA"

As with all software solutions, a system can only be considered secure if adequate processes are in place that address the fundamental tenets of security. In Grid computing these are traditionally labelled as the "AAA" concepts:

- Authentication – establishing and verifying the identity of an individual.
- Authorization – once the identity has been verified, establishing what access controls that identity is subject to.
- Accountability – in the event of system abuse or security breach, being able to correctly tie the actions of a user to their identity.

With these three processes in place, it is possible to identify, limit access and have recourse to any users that attempt to abuse the system or any data stored within.

Authentication

The authentication process is covered in many systems by the use of username/password combinations, the identity of which is stored on a database within the system. However, in terms of applying this to a virtual organisation, there is a need for a more sophisticated distributed model. As such, the paradigm for single sign-on (SSO) – which allows a local user to have a remote identity – is now implemented for many of the projects undertaken at NeSC-Glasgow.

Shibboleth [5] is a technology that provides the ability to transfer attributes between entities signed up to a common federation, and is rapidly gaining widespread acceptance throughout the academic community. This allows users to authenticate to a local institution – who are more likely to be able to verify the veracity of that identity – but have access (subject to agreed policies) to resources at other institutions signed up to that same federation.

This distributed model of authentication allows a policy of flexible access to distributed resources, whilst also allowing tight locally-enforced powers of revocation to be established. The implications of being able to use a similar model in the context of the ESPE project are immediately obvious – distributed, yet limited, access to remote resources at a level controlled by the local resource-owner.

Authorization

One of the major factors in developing Grid solutions is the need for a solution that will scale to the potentially large number of users that will use the system. In specialised and focused cases such as this, the numbers are potentially lower than in generic solutions, which are often the main focus of Grid research.

However, the issue of scalability must still be addressed. In terms of security, the first step to make the lists of users manageable across a wide domain, is done by implementing the concept of role-based access control (RBAC). This immediately makes the number of entities to identify and restrict more manageable – the role must be secured rather than the individual user, and administration is de-coupled to the linkages between user and role, and those between role and resources.

In terms of data federation, the role-based approach allows pre-set views to be implemented that restrict the data sets available. This appears to be the most efficient way of securing access to individual data fields within databases in terms of grid services. The authorization process must be implemented on each role – delimiting what data fields can and cannot be viewed or updated. Within the VOTES project, the concept of an access matrix has been introduced to provide a framework that matches role against data sets, e.g. in a database. With this access matrix and knowledge of the data model and roles required, if a given parameter in the matrix is set, say to "1", this can be used to ensure that only users with that role can access that data item, as shown in Figure 1.

	R_1	R_2	R_3	R_4
U_1		h_1	h_2	h_3	h_4
U_2	U_1	0	0	1	0
U_3	U_2	0	0	0	1
U_4	U_3	1	1	1	1
	U_4	0	1	0	0

Figure 1: Conceptual representation of the access matrix framework: role versus parameter, and user versus role.

We note that it is possible for the access matrix to protect data at different levels of abstraction. Thus the parameter might be the database

itself, a table in the database or column/row of a table in that database. The "0" value indicates that this data item is not available to the user with that role.

Many authorization technology solutions are under development within the Grid community [6, 7] but most of these do not meet all the requirements of adequately securing **data** in a Grid environment. As such, the access matrix concept has been implemented in this project using a combination of simple database relationships (as shown in figure 2) and JDBC communication streams between portal, matrix and grid services.

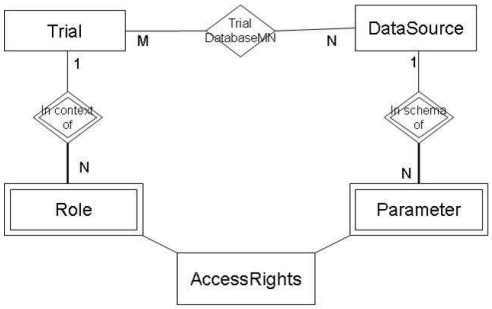

Figure 2: Entity-Relationship diagram representing the access matrix database. Note this is re-used in the VOTES project, so "Trial" here is represented by "Project".

The roles distributed so far include the differences between consultants and investigators as the most privileged roles, and nurses as the less privileged role. The low-privilege roles do not have access to the upload portal and only have limited access to the query portal. The roles and their impact upon data access policies across distributed data sets themselves will be determined on a VO by VO basis – subject of course to ethical agreements.

Accountability

A major requirement of the collaborative sharing of the ESPE data is an audit trail of who has accessed, read and edited the data. This is ordinarily an area that is left till the later stages of most Grid projects. However the demand from the production context has put it forward as of being of great importance.

The need for an audit trail of course points to the need for accountability in the event of a security breach. However the main driving motivation on the part of the ESPE project is to know who has originally uploaded the DSD records, so to check upon consistency and validity of records. This underlines to a certain degree, the notion of limited trust between parties mentioned in the section 1.

It also provides the opportunity to enable more mainstream security in terms of flagging suspicious behaviour, or retrospectively trailing malefactors in the event of a security breach. The audit trail itself will be available in logs within the grid server, accessible by the system administrator. However, the need for clinician accountability requires information – such as identifying the clinical user, the upload time and the read time – to be coded into the individual patient records returned.

Legal considerations

Because of the semi-autonomous nature of the sites involved in this project, and hence the between all the nodes within the VO, a static context is required which will establish a base-level of trust between the parties. Due to the security considerations above, it is necessary for users with administrative rights to grant rights to users over remote resources. To implement this in any system requires a high level of trust between the acting parties, which in turn will require a legal form of recourse should either party feel that their grant of administrative rights has been abused. Therefore, a static element to this dynamic system is unavoidable (and indeed necessary). The question then becomes what interactions between the node super-users are delineated in the static agreement and what can be assumed at run-time.

4. Implementation

Figure 3 shows the current architecture used that links the ESPE database to a web portal through the use of grid services.

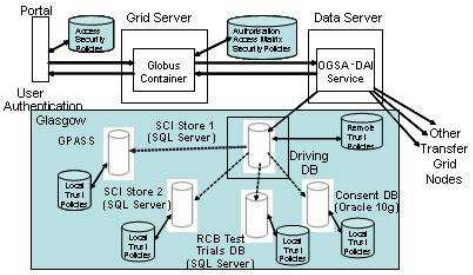

Figure 3: Architectural diagram of data grid constructed for the ESPE project.

The basic operation is as follows:

- The user logs into the portal at a particular node - this can be directly or via Shibboleth (not shown above).
- The portal server checks the local resource files to discover the available grid servers, data servers and driving databases.
- A request containing the user's role and a specific trial is sent to the grid server.
- The grid server consults the local access matrix and returns the parameters for the resources that the user can query for that trial. These are presented to the user as a list of check-boxes, with the option to specify conditions if desired.
- The user makes their selection of parameters and submits them. These are constructed into a single query distributed across the various resources.
- The query is sent from the grid server to the data server, where it is wrapped as an OGSA-DAI service request and is passed to the driving database.
- The driving database executes the distributed query over the resources under its guard and joins the various distributed results into one single result.
- This result is sent back to the data server and then to the grid server, where it is transformed into readable HTML and finally presented to the user through the portal again.

The technology used to implement this architecture is as follows:

- GridSphere [8] to implement the portal that provides user-friendly access to the system.
- Globus Toolkit v4 [9] to implement the grid service that links the UI layer to the data layer, whilst providing an intermediary layer, where the security principles can be implemented.
- OGSA-DAI [10] is used to filter the grid service SQL query into an XML format more easily used for rendering in the portal.
- The data store used in the Glasgow site is PostgreSQL [11]. However, this will differ across the various sites involved, depending on their local installations.

Figure 5 (at the end of the paper) shows a screen-shot of the interface that is provided to the user when using the system. Two portals are used – one for querying the records and one for uploading data to the centralised repository of data.

The communication protocols used are SOAP (Simple Object Access Protocol) between entities and JDBC (Java DataBase Connectivity) for accessing the data store. The access matrix is implemented by using a set of relational tables within a PostgreSQL database. This delimits the roles, parameters and access rights associated with each role. The database is queried from an SQL statement, which returns the parameters relevant to that role, presented to the user in a user-friendly graphical interface.

In terms of the use of Shibboleth to facilitate the authentication mechanism, the portal has been made more user-friendly by inserting the roles for the different trials into attribute certificates maintained in an attribute authority at the University of Glasgow. When a user logs in to the University of Glasgow Identity Provider these attributes are returned to the portal along with a signed SAML assertion that the user has authenticated satisfactorily (or not). The attributes themselves are then used to restrict the data sets associated with that role across the federated data sets comprising the ESPE data Grid. The ability to have roles per VO overcomes the largely static model of security upon which Shibboleth federations are based. In the UK a core set of eduPerson attributes have been agreed for UK–wide Shibboleth federations. Extending this to support more dynamic VO-specific roles and attributes is essential and has been supported through projects such as NeSC DyVOSE [16] with its delegation issuing service.

Figure 4: the Shibboleth attributes relevant to the user "guest2". Note how the user has different roles for different trials, with varying levels of privilege.

4.1 Production Security

Because this is a system that will enter a production context in a relatively short space of time, the system must also be considered secure in terms of not just the unsolved

research questions, but also in terms of real protection against malefactors.

The main issues raised in this context are as follows:

- Validating all data that is entered into the portal. These points of entry provide the opportunity to enter escape codes or byte-code that could surrender control of the portal server to a malicious third-party (e.g. the modus operandi of the buffer overflow attack).
- Encrypting all communications between components of the architecture. This includes SOAP calls between portal and grid servers and JDBC connections from grid server to database.
- A list of grid servers exist that must be encrypted and protected. This would be an effective point of attack to redirect the operation of the system to another URL.
- The access matrix must be encrypted and protected. This is potentially the most important component in the system – if roles and authorization privileges can be modified by a malicious third-party, the consequences would be high indeed.

This list is by no means exhaustive but does highlight the security precautions that must be applied to all systems of this nature.

The distributed system also has failed over components to allow continuous operation in the event of the loss of a single component. The portal, grid servers and databases are co-located in several machines, with a connectivity test being performed upon portal load-up. An issue that still remains to be implemented is the ability to replicate the contents of the database at the end of a session, or during a session to allow consistently synchronised real-time, multi-user access.

5. Conclusion

This paper has outlined a system, shortly to enter a working context that effectively brings the paradigms of Grid technology to allow secure yet flexible data-sharing and collaboration between partners in a highly specialised yet geographically distributed field.

Given the sensitivity of the data involved, the primary emphasis of the system is on security. The authorization processes involved use a home-grown system that meets the idiosyncratic requirements of this field, as opposed to the authorization solutions currently on offer in the Grid community, none of which entirely meet the needs of this project. Similarly, Shibboleth has been adopted as a single sign-on solution, which, as an attribute transfer mechanism, provides an effective method of securely navigating throughout virtual organisations and federations of partner sites in just such a context as this.

The architecture and technology used here is believed to be to be a well-supported, robust framework that is ideal for use in a clinical environment. Other projects conducted at NeSC-Glasgow that involve using Grid technology in a similar context, are the VOTES project, studying patient recruitment, data collection and study management within clinical trials and studies, and the BrainIT project [12], which grid-enables access and linkage between neurological MRI scan data in the Southern General Hospital, Glasgow. The ultimate aim for all these projects is to create a single, unified Grid infrastructure that brings these many different strands of medical expertise together in a single, coherent and re-usable framework.

6. Acknowledgements

The authors would like to acknowledge the Department of Child Health at the Royal Hospital for Sick Children, Yorkhill, Glasgow, and in particular the contribution of Dr Syed Faisal Ahmed.

7. References

[1] VOTES – Virtual Organisations for Trials and Epidemiological Studies, http://www.nesc.ac.uk/hub/projects/votes
[2] ESPE – European Society for Paediatric Endocrinology, http://www.eurospe.org/
[3] SNOMED-CT – http://www.snomed.org
[4] HL7 – http://www.hl7.org
[5] Shibboleth – http://shibboleth.internet2.edu
[6] PERMIS – http://sec.isi.salford.ac.uk/permis/
[7] VOMS - http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html
[8] GridSphere – http://www.gridsphere.org
[9] Globus v4 – http://www.globus.org
[10] OGSA-DAI – http://www.ogsadai.org.uk
[11] PostgreSQL – http://www.postgresql.org
[12] BrainIT, part of the GLASS project (GLASgow early adoption of Shibboleth)– http://www.nesc.ac.uk/hub/projects/glass
[13] DICOM (Digital Imaging and Communications in Medicine) – http://medical.nema.org
[14] OPCS (Office of Population, Censuses and Surveys) - http://www.connectingforhealth.nhs.uk/systemsandservices/data/clinicalcoding/classifications/opcs_4

[15] ICD 10 – International Statistical Classification of Disease and Related Health Problems (ICD-10), http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd_10
[16] DyVOSE – http://www.nesc.ac.uk/hub/projects/dyvose

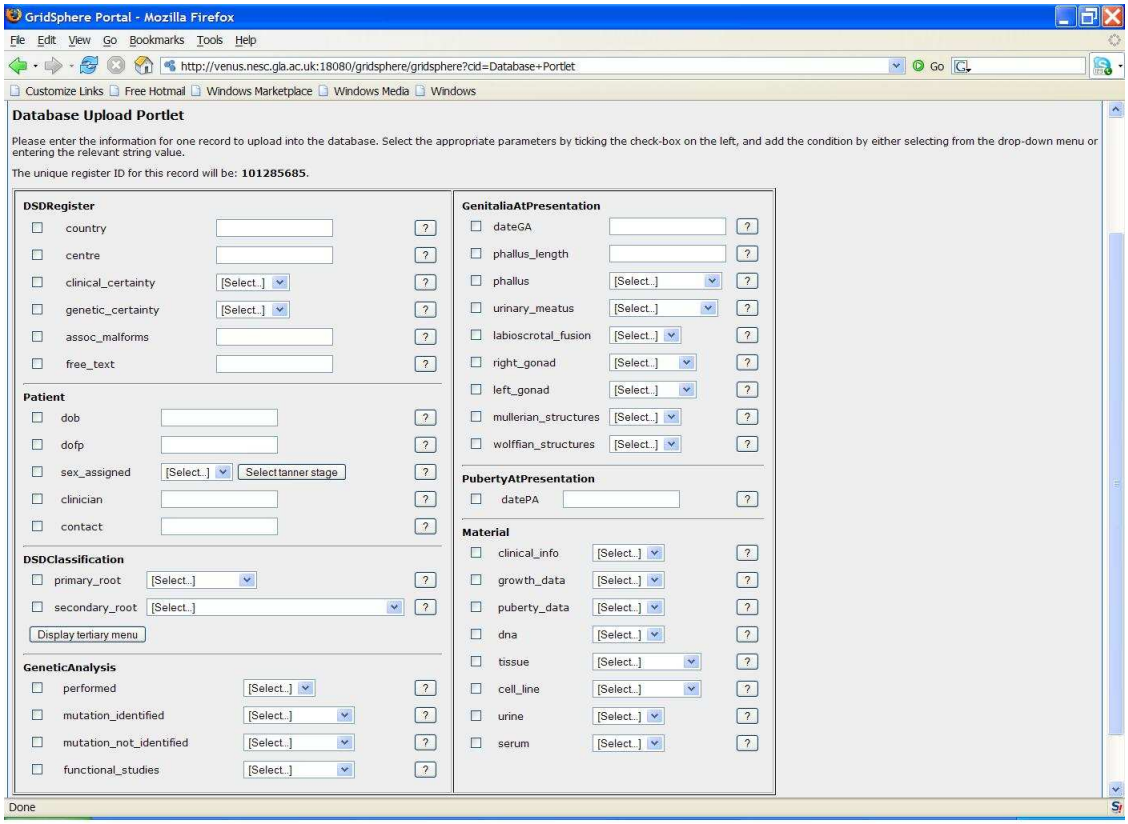

Figure 5: A screen-shot of the ESPE data upload portlet. The parameters show the details that are pertinent to the data set involved. A unique ID for each record uploaded is shown at the top. The query portlet replicates this information but allows a search on identifying number and the masculinisation score (calculated from the data collected in the GenitaliaAtPresentation section).