SECURE FEDERATED DATA RETRIEVAL IN CLINICAL TRIALS

Anthony Stell, Richard Sinnott, Oluwafemi Ajayi
National e-Science Centre
University of Glasgow
United Kingdom
ajstell@dcs.gla.ac.uk
ros@dcs.gla.ac.uk
ajayio@dcs.gla.ac.uk

ABSTRACT
The clinical domain is one in which a plethora of data exists in repositories distributed across the globe, crossing institutional, regional and national boundaries. To be able to harness this data and move it across these boundaries has the potential to provide great scientific and medical insight, to the benefit of many protagonists in the field of clinical medicine. In this paper, we outline the challenges inherent in drawing together such data sets using Grid technology, focusing specifically on the issues surrounding security and data access.

A framework is outlined that makes use of Grid technology to achieve this "federation" of clinical data. It is described in the context of Virtual Organisations for Trials and Epidemiological Studies (VOTES), a project funded by the UK Medical Research Council (MRC) and involving the National e-Science Centre (NeSC) at the University of Glasgow. In this framework, several solutions are proposed to address the security issues specific to the clinical domain including fine-grained "anonymisation" services where identifying data in medical records are seamlessly de-identified based on the user privilege, leaving only statistically relevant data for viewing by un-privileged users.

KEY WORD
Database and Information Systems, Data Federation, Anonymisation

1 Introduction

Clinical data exists in many and varied formats across many domain boundaries: institutional, regional and national. Initiatives such as e-Health [1] exist across the developed world and are attempting to bring this data and knowledge together in a secure yet flexible manner. The problems that exist in trying to achieve this are suited to the application of Grid Computing. Though Grid technology has traditionally been applied to harness large-scale, heterogeneous compute resources or data storage, in the clinical domain, the focus is predominantly upon access to and usage of clinical information where transferral of data and knowledge across domains is of primary concern.

One of the central paradigms of Grid Computing is to support the establishment and subsequent management of Virtual Organisations (VOs). A VO is typically represented as a collaboration of partners, users and resources, delimited by common policies, in a dynamic and transient manner. Security is one of the key components used to establish the terms and agreements (rules) used to subsequently manage the VO. And a feature of a VO is the limited degree of trust between the parties. Though they are required to work together to achieve a common goal, there will almost certainly be data or specific areas of their respective domains, which are considered sensitive and are not necessary for the other partners to be aware of, i.e. site autonomy is an underlying principle of Grids. To establish and manage effective and useful VOs, numerous issues have to be addressed such as how are security information set and exchanged so that the partners can collaborate effectively without compromising their respective security policies?

VOs in the clinical trials domain are characterised by a much greater degree of emphasis on security, data access and data ownership. We term these Clinical Virtual Organisations (CVOs) since they place requirements not typical to other High Performance Computing-oriented VOs common to the wider Grid community. Rather than developing bespoke CVOs for each individual clinical trial, it is our intention within the VOTES project to develop a framework supporting a multitude of CVOs. Each of these CVOs will be derived from the framework and adapted depending on the needs of the trial or study being conducted.

The VOTES project [2] is a collaborative effort between e-Science, clinical and ethical research centres across the UK including the universities of Oxford, Glasgow, Imperial, Nottingham and Leicester. The primary focus of VOTES is to build an infrastructure to support a multitude of clinical virtual organisations. Common phases of many clinical trials and epidemiological studies, and the primary focus for core components that will exist in the

VOTES Grid framework, will cover three areas: patient recruitment, data collection and study administration.

In this paper, the security challenges raised by realising the VOTES plans are discussed and explored. The outline of a Grid framework implementation that attempts to address these issues is described, in terms of architecture and technology. Finally, different solutions to the problem of "anonymisation" are proposed, with a discussion of their inherent merits and drawbacks.

2 Security challenges

An important feature - probably *the* most important feature - of setting up a Grid framework is the security involved in protecting the resources and users of that system. Without this, sites will not trust one another nor users trust the system, and in turn the users will simply not use the infrastructure.

2.1 Authentication, Authorization and Accounting

The concepts involved in security in Grid are traditionally broken down into three areas, commonly known as the "Three A's": Authentication, Authorization and Accounting.

Authentication refers to the process of verifying that users are who they say they are. This is achieved using the concept of a PKI (Public Key Infrastructure), where public and private keys are used to digitally sign and verify signatures in exchanged security tokens (certificates).

Authorization refers to the process whereupon a user that has been satisfactorily authenticated is allowed to perform different actions on a resource depending on their identity and their associated privileges. There are many Grid application solutions that claim to provide a scalable, effective authorization solution, such as PERMIS [3], CAS [4], VOMS [5] and Akenti [6]. But so far, a clear area leader has not been established in the Grid community. In the VOTES project an idiosyncratic Access-Control Matrix has been devised to provide authorization on the services involved, used in preference mainly due to ease of implementation, described in section 4.

Accounting refers to the process of auditing, whereupon once a user has been properly authenticated and authorized, their actions on the resource are logged and monitored so that if any improper actions are conducted, they can be held accountable for their actions. This is also known as "non-repudiation".

Whilst these areas must be addressed in any production implementation of a Grid framework, they are still only the basic building blocks of securing such systems. As mentioned previously, within life sciences generally, and in the clinical domain specifically, there are other additional security concerns that need to be addressed before a system can be described as "secure".

2.2 Statistical Inference

One of these additional security concerns is the idea of statistical inference. This is where a sufficiently low number of records have been returned as a result of a query, and which contain enough readily available data for un-privileged users to infer the identity of the patient that the record refers to. The problem stems from the fact that this data on its own is not sufficient to identify an individual but when combined with data from other domains, also considered to be "safe", identification can be made.

There are currently methods to address this problem in the clinical trial domain. For instance, one solution taken by many clinical centres is to set a threshold number, whereby if the number of records returned is less than this number, the records will be with-held. This is often a poor approximation however, since it is often the case that *sufficiently* anonymised records will be omitted even though it may not be possible to identify the associated patients.

The concept of building an aggregation environment to address the issues [7,22] surrounding inference through statistical disclosure is one that is being investigated in NeSC, Glasgow.

2.3 Anonymisation

One of the major additional security concerns in the clinical domain, and a corollary to statistical inference, is the ability to identify patients. The information and data being processed in clinical trials and studies is of such a highly sensitive nature that any ability for un-privileged users to be able to tie a particular condition or treatment to a specific individual would be a major breach of privacy.

To address this concern, the idea of anonymisation is applied. This is a process that is already applied in many closed domains where such data is held (e.g. the NHS-Scotland clinical databases). Anonymisation is achieved by uniquely identifying fields that are to be encrypted according to a specific key under the control of that domain's administrators.

To apply this paradigm across the domain of a virtual organisation there must be strict controls as to what data should be available for research purposes (i.e. unencrypted). The questions that arise are what data should be available but encrypted? And what data should be known to exist but not necessarily be available to any user beyond the originating domain?

An example of this last type would be the unique reference number that identifies all patients across the NHS in Scotland – the Community Health Index (CHI). This parameter uniquely specifies an individual patient so is mandatory to make sure that records are not double-counted. However, because of its uniqueness, the ability to know what that CHI number actually was, would bring up the privacy issues alluded to earlier in the section.

Defining the boundary between *using* this unique index to collate statistical data, and *knowing its value* to be able to identify, treat and liase with specific patients is a complicated issue, not easily solved with current technology. Some possible solutions are proposed for addressing the problem of anonymisation in section 5 of this paper.

3 Securely Managing Data

In order to provide effective security within a virtual organisation, it is necessary to have a common security concept across the domain over which security policies will be applied. This can only be achieved across heterogeneous domains by either having a standard schema that all the domains within the VO subscribe to, or to have multiple schema mappings that allow inter-domain communication to occur in a meaningful context. [23]

3.1 Global standards

In terms of standards, there are numerous developments for the description of data sets used in the clinical trial domain. However, this can be an involved process depending on standards groups developing and acting on strategies put together through major initiatives such as Health-Level 7 (HL7) [8], SNOMED-CT [9] and OpenEHR (Open Electronic Health Records)[10].

There are often a wide range of legacy data sets and naming conventions which impact upon standardisation processes and their acceptance. The International Statistical Classification of Disease and Related Health Problems version 10 (ICD-10) [11] is used for the recording of diseases and health related problems and is supported by the World Health Organisation. In Scotland, ICD-10 is used within the NHS along with ICD version 9 and Read codes in the SMR data sets for example. ICD-10 was introduced in 1993, but the ICD classifications themselves have evolved since the 17th century [12].

These standards initiatives go some way to addressing the problems of idiosyncratic hierarchical classifications. But the results of these projects will only become apparent on a time-scale of years rather than in the shorter term. To provide meaningful communication of meta-data in the short term requires the manual mapping of data set schema to each other. The VOTES project attempts to do this by gathering data sets that are significantly representative of the data warehousing schemes that exist throughout Scotland. With a view to widening the scope of the project, it is hoped that such representative data schema could be other countries around the globe, if the work in VOTES proves to be successful and popular.

3.2 Joining distributed results

An issue inherently wrapped in the problems of data standardisation and classification is that of having unique references to records that can be matched across the schema that exist in different domains.

In Scotland, a parliamentary initiative is underway to give every member of the population in Scotland, an associated Community Health Index (CHI) number, which is unique across the entire Health Service in the country. With this standard reference, not only can databases have a common value upon which joins can be made across boundaries, but it also provides an implicit guarantee of correctly counting records for statistical purposes.

However, there are issues in the practicalities involved in rolling out this unique index. Some regions in Scotland have already rolled this initiative out, with nearly all patients having a CHI number assigned. Others have not, but the initiative, which hopes to have this number assigned to every citizen by the 6th June 2006, is unlikely to cover all eventualities. For instance, 8 million CHI numbers are currently assigned; the population of Scotland is roughly 5 million people. Because not everyone has a CHI number, this suggests that significant mis-counting has occurred already, in the designation of this number. Some patients may have two numbers, whilst others may be assigned to people that have died.

Again, this index potentially provides a single, unique point of reference for identifying patients so should be treated as highly sensitive – essentially it is the key to all sensitive patient data. However, by its nature, it is also the necessary component for joining data from across domains. This brings up the issue of whether administrators from other domains should be able to access the CHI or whether other more sophisticated methods of joining data sets should be found.

4 VOTES Implementation

As a first step to addressing the issues discussed in the previous two sections, the basic architecture of the Grid framework implemented in the VOTES project is presented here. The system supports federated queries in a user oriented, but secure, environment, as depicted in Figure 1. This infrastructure is hosted on a trial test bed at the National e-Science Centre (NeSC) at the University of Glasgow.

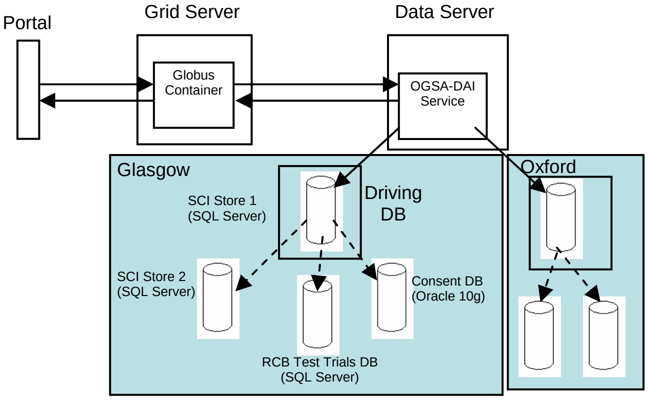


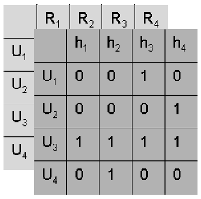
Figure 1: Software architecture schematic. The "Oxford" box indicates how other institutions will be added to the current design – the current implementation only incorporates the test databases running in Glasgow.

A GridSphere [13] portal front-end communicates to a Globus Toolkit [14] (v4.0) grid service, which in turn provides access to an OGSA-DAI [15] data service. This runs queries from the "driving database" using standard Simple Object Access Protocol (SOAP) message-passing, but also in turn runs queries from the subsidiary databases available from the pool for which it is responsible, using direct Java Database Connectivity (JDBC) connections.

The technology used in this implementation places strong emphasis on the use of grid services – essentially web services with the additional notion of permanent state. Within the Grid community this paradigm has been largely seen as the most effective solution to implementing transient and dynamic virtual organisations.

An example of this is the Web Services Resource Framework (WS-RF) [16] as implemented in version 4.0 of the Globus Toolkit. Issues of access control are integrated within this framework by means of a Security Assertion Markup Language (SAML), which allows a standard exchange of security assertions and attributes. A popular implementation of this standard has been the OpenSAML project [17], which is now following the latest release of SAML, v1.1, and is currently developing an implementation of v2.0 [18].

The back-end authorization framework developed is an infrastructure based on an access matrix as shown in Figure 2. This is currently a short-term model, created to allow a prototype to be developed rapidly, there being implementation overheads with most other Grid authorization solutions.


R_1
R_2
R_3
R_4
U_1
U_2
U_3
U_4

	h_1	h_2	h_3	h_4
U_1	0	0	1	0
U_2	0	0	0	1
U_3	1	1	1	1
U_4	0	1	0	0

$U_1(R_1 \Delta h_3) = 1$, $U_2(R_1 \Delta h_2) = 0$, $U_3(R_3 \Delta h_1) = 1$,
$U_4(R_2 \Delta R_3 \Delta h_4) = 0$
where Δ is a combination function, 0, 1 are bit-wise privileges,
R_X, h_X are resources and U_x is a subject
Figure 2: Access Matrix Model

The authorisation mechanism implements an access matrix model [19] that specifies bit-wise privileges of users and their associations to data objects in the CVO. The access matrix is designed to enforce discretionary and role based access control policies and has been constructed to be scalable for ease of growth parallel to the growth of the infrastructure as a whole. Comparison of this approach with other solutions such as Role Based Access Control solutions such as PERMIS will be undertaken, where user views of data sets will be mapped to CVO roles.

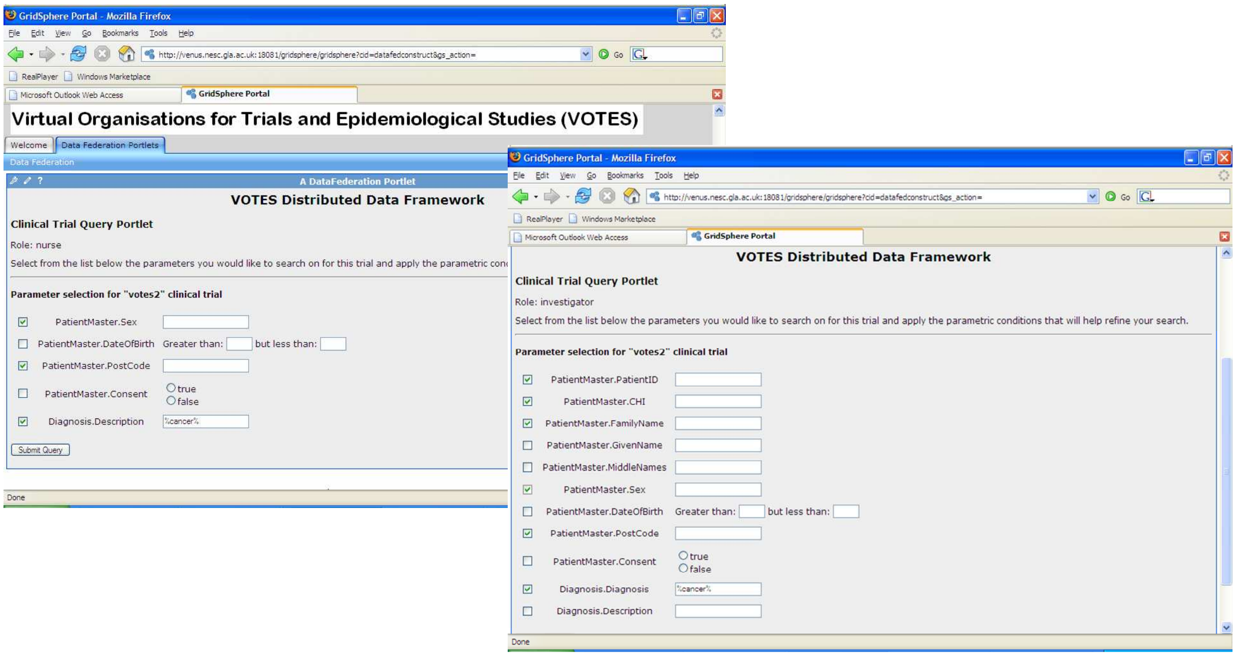
The portal operates as if to present a single unified resource to the end user. The user logs in to the portal, and is assigned a role according to their privileges within that domain. Upon selection of specific named clinical trial, they are presented with a set of parameters. What parameters these are depends on the meaning of their role within that context (i.e. the trial) allows them to see. In this way, role-based access-control is applied.

The user then selects the parameters and conditions that they wish to apply and submit the query. What is returned on the final screen is the user's role, the trial selected, the databases used for the query, the SQL query constructed from their parameter selection and a table of the data returned from this query, as shown in figure 5 (at the end of the paper).

5. Anonymisation Solutions

The implementation outlined above is a first step towards addressing the security issues in the clinical domain. However, it is not clear whether it necessarily provides the most effective solution to the problem of anonymisation.

Other solutions for data anonymisation are also possible, given the central requirements of the problem. This is that an anonymised data set is that any fields in a patient record that could positively identify that patient should be inaccessible to any handler of that record that does not have the necessary privilege. However, as has been discussed, in order to gather data that can be meaningfully matched across domain boundaries and is not mis-counted, the technology must use unique reference handles that by their nature, necessarily identify the patient in that record.


GridSphere Portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://venus.nesc.gla.ac.uk:18081/gridsphere/gridsphere?cid=datafedconstruct&gs_action=
Go
RealPlayer
Windows Marketplace
Microsoft Outlook Web Access
GridSphere Portal
Welcome
Data Federation Portlets
Data Federation
A DataFederation Portlet
Submit Query
Done
GridSphere Portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://venus.nesc.gla.ac.uk:18081/gridsphere/gridsphere?cid=datafedconstruct&gs_action=
Go
RealPlayer
Windows Marketplace
Microsoft Outlook Web Access
GridSphere Portal
Done

Virtual Organisations for Trials and Epidemiological Studies (VOTES)

VOTES Distributed Data Framework

Clinical Trial Query Portlet

Role: nurse

Select from the list below the parameters you would like to search on for this trial and apply the parametric con

Parameter selection for "votes2" clinical trial

☑ PatientMaster.Sex
☐ PatientMaster.DateOfBirth Greater than: but less than:
☑ PatientMaster.PostCode
☐ PatientMaster.Consent ○ true ○ false
☑ Diagnosis.Description %cancer%

VOTES Distributed Data Framework

Clinical Trial Query Portlet

Role: investigator

Select from the list below the parameters you would like to search on for this trial and apply the parametric conditions that will help refine your search.

Parameter selection for "votes2" clinical trial

☑ PatientMaster.PatientID
☑ PatientMaster.CHI
☑ PatientMaster.FamilyName
☐ PatientMaster.GivenName
☐ PatientMaster.MiddleNames
☑ PatientMaster.Sex
☐ PatientMaster.DateOfBirth Greater than: but less than:
☑ PatientMaster.PostCode
☐ PatientMaster.Consent ○ true ○ false
☑ Diagnosis.Diagnosis %cancer%
☐ Diagnosis.Description

Figure 3: Parameter selection screen for the same clinical trial. In the screen on the left, the number is restricted to only a few non-sensitive parameters for the "nurse" role, whilst the screen on the right, for the "investigator" role, shows a much richer set of parameters, including readily identifiable patient data.

5.1 Domain restricted data

The current favoured method of applying anonymisation within NHS-Scotland is to encrypt and de-crypt the data according to a private key only available to privileged users within the NHS. This addresses the inherent security concerns and effectively maintains the privacy and integrity of the patient records within a single domain.

This is an example of domain-specific security. It provides a secure solution for maintaining the privacy and integrity of patient records and keeps the responsibility for security within the administrative domain of the site that generated the data. However, none of the advantages provided by a Grid solution can be applied in this case. Statistics can still be performed on the limited data set but it would require a guarantee that the unique (encrypted) index is indeed unique, something that outside agencies would be unable to verify.

And most importantly, as each unique reference would be encrypted according to an idiosyncratic and (necessarily) unknown algorithm, it would also be impossible to perform joins between data sets across domains using this solution. This is the central reason for proposing a Grid solution and the whole enterprise is redundant if the data cannot be matched in a context beyond the immediate domain boundaries. In terms of the first aim of the VOTES project, it would also be impossible, outside of the original domain, to perform "unblinding" of records, a necessary step during the process of patient recruitment.

As with most Grid solutions, only part of the problem in building inter-institutional solutions is technical. The other part is the human factor of trust establishment between parties. The scenario above would assume a level of trust that was too limited to effectively bring advantage to the field of clinical medicine. However, to develop trust between parties in a VO, or in enabling technologies, requires close collaboration and confidence in the security technology that underpins the infrastructure as a whole.

5.2 Role-based data restriction

This method is that outlined in the implementation of the VOTES portal above. It relies on the application of roles within the authorization mechanism of the framework, restricting the data that the user can view based on that role. An example of the difference in parameters returned, depending on role, is shown in figure 3. The main advantage of this approach is that the data is unencrypted beyond the domain of the site where it originated. As a result, the unique index is available to allow joining between data sets across many domain boundaries. The security paradigm is essentially that responsibility for the data has been delegated to the security administrator at the site node from another site node. It is at this first administrator's discretion what roles can see what data, though this could be established using a pre-defined contract of data usage between all the sites in the VO.

Additional security measures can be taken that allow the privacy and integrity of the data to be maintained against parties that are external to the VO. This would involve the use of a Public Key Infrastructure (PKI) to digitally sign and encrypt messages passed between sites and resources. A reasonably high level of trust between the participating parties is required for this application to work. There is a trust that, once the data has been surrendered, it will be responsibly maintained and not distributed beyond the bounds of the VO.

However, this raises issues of interest. As has been stated previously, virtual organisations are transient entities with degrees of limited trust between participants. It is also possible for sites to be members of multiple VOs simultaneously. While it is necessary for a node not to participate in VOs that would result in a conflict of interest, it is still a possibility that they could become part of a conflicting VO in the future. It is forseeable that legal *static* contracts would have to be defined, in order for participants to protect themselves against adverse action as a result. Therefore this must be a major consideration when releasing unencrypted data within the VO.

As is evident there are both significant advantages and disadvantages to this approach for restricting data. The implementation of this method allows major technological hurdles to be overcome in a quick and easy manner. However, the potential compromises that must be made to achieve this in a production context violate individual security policies to an unacceptable degree. Therefore, it is not obvious that this is the best solution for protecting data in such a flexible and malleable environment.

5.3 Delegated anonymisation service

This method involves using a specifically delegated service to administer the sensitive data being passed between participating sites in the VO. This service acts a source of trust and as a security broker for the VO. The chain of trust is broken down so that now sites can vary the level of trust between each other arbitrarily but, by placing trust in this anonymisation service, are able to make full use of the opportunities provided by Grid technology. The proposed architecture is shown in figure 4.

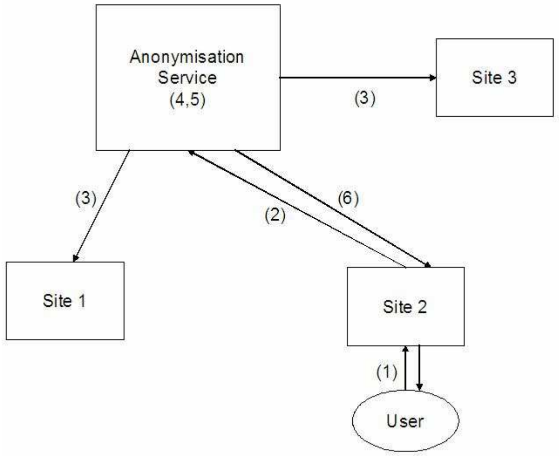


Figure 4: Basic framework for the implementation of an anonymisation service

The method of operation is as follows:

1. A user at site 2 wishes to perform a federated query that requires data to be joined from sites 1 and 3.
2. The query is sent as an SQL query statement to the anonymisation service.
3. The query is executed from the service to the distributed sites, and data returned (with encrypted indices) to the anonymisation service.
4. The anonymisation service decrypts the unique indices for the different sites, as it holds the necessary keys to do this.
5. The data sets are joined on this unique reference.
6. Depending on the policies held at sites 1 and 3 with regard to site 2, the result of that join, with identifying data stripped away is returned to site 2.

The proposition is that regional clinical domains would delegate trust to the anonymisation service centralized at a say, national level. This would require a highly secure back-end data repository to store information retrieved from the various sites.

Potential also exists for this service to be extended to simultaneously address the issue of statistical inference. As the full data set is residing within the anonymisation service repository before being passed on to the requesting site, any policy that highlights where incidences of patient identification *could* occur would be enforced at this point. The offending data combinations could then be either stripped from the data returned or the query itself could be halted before returning the results to the requesting site. This is where an aggregation environment could be built up for a given CVO so that the pre-defined security policies could be imposed, which remove the possibility of statistical inference.

An unsolved issue with the implementation of this service is derived from the necessarily transient nature of virtual organisations. The anonymisation service would have to be created from a root of trust to which participating members subscribe and this would be difficult to achieve in a dynamic manner. To establish and maintain this trust, a necessity for the enterprise to succeed, there will need to be some static implementation of trust, possibly an entity that is analogous to the concept of Certificate Authorities in public key infrastructures.

The service proposed in this paper, is a potential direction for the VOTES project. This does depend on understanding the issues involved in bringing data back from across boundaries in a ‘live’ context and, due to the time it naturally takes to establish the trust required for such an endeavour, this is an area that is currently unavailable.

6 Other Models

There are other methods currently in the Grid community for performing data federation. One is the “push” model of data retrieval, which involves the data from a particular site being sent to the requester, rather than having an active query executing on the repository node.

It is feasible that this could be worked into the solution proposed in section 5.3, with a query being analysed by the data node then, in consultation with a release policy, the data that that site is willing to push out could be sent to the central service. However, this is essentially replicating the function of the central anonymisation service, which, in the model in section 5.3, has already been delegated trust from the repository site.

Security models other than the access matrix approach described should also be considered. For instance, a rule-based model could be used where a final “authorization string” could be built up from the access requests and evaluated. The difference with the current approach is that with the rule-based approach, the authorisation decision would only be made at the end of the process. With the access matrix, the querying string is built up after the authorization for different components has been made, therefore when the final query is run, it is known that it will not fail due to lack of privileges.

Another consideration is that of distributed access. A Grid application that is gaining ground in the academic community is that of Shibboleth [24], which provides a mechanism by which attributes can be exchanged. This is useful in particular, for exchanging security tokens for the purposes of authentication and authorization.

Using a “shibbolized” portal, once the user is authenticated to their home institution, they are then automatically authenticated to the wider federation that their home institution is part of (if the credentials exchanged are authentic and valid). This allows the front-end portal access to be distributed, thereby allowing distribution of the application implementation, in turn distributing load and allowing a measure of failover redundancy to be incorporated in the portal design.

7 Conclusion

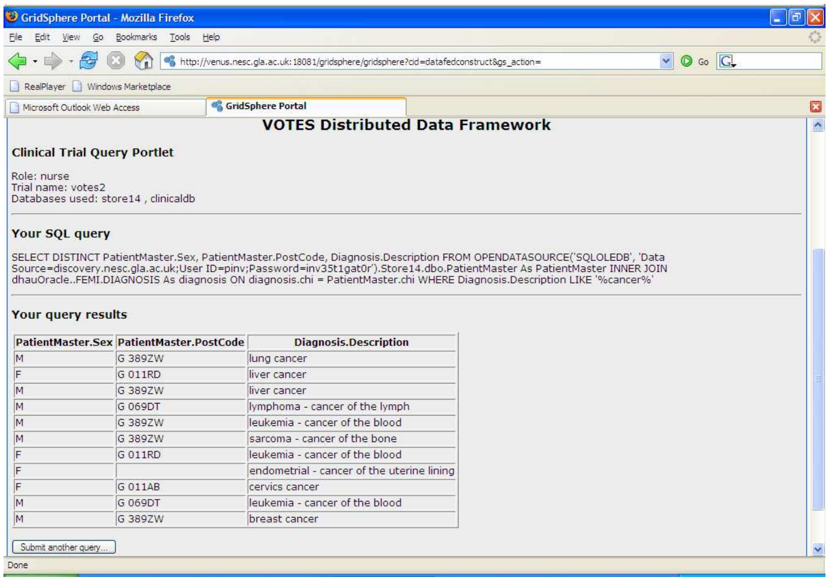
Security is a major issue in the establishment of a distributed framework that will federate clinical data in a context that is both effective and meaningful. As discussed, there are much greater security needs in the clinical trial domain than in other fields that have previously been addressed by Grid technologies (e.g. bioinformatics, particle physics). And in particular the issue of anonymisation and the protection of patient identity outwith the originating domain is of paramount importance.

The solution proposed goes some way to addressing this issue but only looks at the technological aspects of the problem. In practical terms, a lot of the issues to be overcome involve the human factor, where political issues must be resolved and, especially, a chain of trust must be built, not only between the participating domains, but in the technology that is being used to solve these problems. Legal requirements of the potential users of the system must be taken into account and a full risk analysis of the system would need to be done, before conclusions on the potential for uptake can be drawn.

The prototype application described in the VOTES project is still in progress. It does not currently overcome all the obstacles outlined in this paper but does provide a starting point, which, through augmentation with features such as the anonymisation service, could become a greater infrastructure able to serve the needs of the wider clinical trials community. The eventual vision is that this infrastructure will one day be available on a global scale allowing health information to be exchanged across heterogeneous domains in a seamless, robust and secure manner. In this regard, we are currently exploring international collaborative possibilities with the caBIG project in the US [20] and closer to home in genetics and healthcare projects across Scotland [21].

8 References

[1] e-Health Initiative – http://www.ehealthinitiative.org
[2] Virtual Organisations for Trials and Epidemiological Studies (VOTES) – http://www.nesc.ac.uk/hub/projects/votes
[3] PERMIS – http://sec.isi.salford.ac.uk/permis
[4] CAS – http://www.globus.org/toolkit/docs/4.0/security/cas
[5] VOMS – http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms/html
[6] Akenti – http://dsd.lbl.gov/akenti
[7] S. D. Vimercati and P. Samarati, "Access Control in Federated Systems," in NSPW '96:
Proceedings of the 1996 workshop on New Security Paradigms, (New York, USA), pp. 87–99, ACM Press, 1996.
[8] Health-Level 7 (HL7) – http://www.hl7.org
[9] SNOMED-CT – http://www.snomed.org/snomedct
[10] OpenEHR – http://www.openehr.org
[11] International Statistical Classification of Disease and Related Health Problems (ICD-10) – http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd-10
[12] ICD Background, http://www.connectingforhealth.nhs.uk/clinicalcoding/faqs/
[13] GridSphere – http://www.gridsphere.org
[14] Globus – http://www.globus.org
[15] OGSA-DAI – http://www.ogsadai.org.uk
[16] Web Services Resource Framework (WS-RF) – http://www.globus.org/wsrf
[17] OpenSAML – http://www.opensaml.org
[18] OpenSAML Development wiki – https://authdev.it.ohio-state.edu/twiki/bin/view/Shibboleth/OpenSAML
[19] R. S. Sandhu and P. Samarati, "Access Control: Principles and Practice" IEEE Communications Magazine vol. 32, no. 9, pp. 40-48, 1994
[20] National Cancer Institute, cancer Biomedical Informatics Grid, https://cabig.nci.nih.gov/
[21] Generation Scotland Scottish Family Health Study, http://www.innogen.ac.uk/Research/The-Scottish-Family-Health-Study
[22] N. Zhang et al., "A Linkable Identity Privacy Algorithm for HealthGrid", pp. 234–245. From Grid to HealthGrid, IOS Press, 2005
[23] A. P. Sheth and J. A. Larson, "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," ACM Computer Survey, vol. 22, no. 3, pp. 183–236, 1990.
[24] Shibboleth – http://shibboleth.internet2.edu


GridSphere Portal - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://venus.nesc.gla.ac.uk:18081/gridsphere/gridsphere?cid=datafedconstruct&gs_action=
Go
RealPlayer
Windows Marketplace
Microsoft Outlook Web Access
GridSphere Portal

VOTES Distributed Data Framework

Clinical Trial Query Portlet

Role: nurse
Trial name: votes2
Databases used: store14 , clinicaldb

Your SQL query

```
SELECT DISTINCT PatientMaster.Sex, PatientMaster.PostCode, Diagnosis.Description FROM OPENDATASOURCE('SQLOLEDB', 'Data
Source=discovery.nesc.gla.ac.uk;User ID=pinv;Password=inv35t1gat0r').Store14.dbo.PatientMaster As PatientMaster INNER JOIN
dhauOracle..FEMI.DIAGNOSIS As diagnosis ON diagnosis.chi = PatientMaster.chi WHERE Diagnosis.Description LIKE '%cancer%'
```

Your query results

PatientMaster.Sex	PatientMaster.PostCode	Diagnosis.Description
M	G 389ZW	lung cancer
F	G 011RD	liver cancer
M	G 389ZW	liver cancer
M	G 069DT	lymphoma - cancer of the lymph
M	G 389ZW	leukemia - cancer of the blood
M	G 389ZW	sarcoma - cancer of the bone
F	G 011RD	leukemia - cancer of the blood
F		endometrial - cancer of the uterine lining
F	G 011AB	cervics cancer
M	G 069DT	leukemia - cancer of the blood
M	G 389ZW	breast cancer

Figure 5: Results from a query executed by a "nurse" role. The name of the clinical trial context is shown along with the databases that have been queried to bring back this data. The SQL query constructed from the parameters selected is also shown, with the final table of results at the bottom. Note that the results are restricted to non-identifying data because of the limited privileges of this role.