Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data

Professor R. Sinnott[1], O. Ajayi[1], A. Stell[1], A. Young[2]
[1]*National e-Science Centre*
University of Glasgow
United Kingdom
r.sinnott@nesc.gla.ac.uk

[2]*Clinical Trials Service Unit & Epidemiological Studies Unit,*
Richard Doll Building
Old Road Campus, Oxford
United Kingdom

Abstract. Grid technologies provide an infrastructure through which, amongst other things, data access and integration is facilitated across highly distributed and heterogeneous resources. Different domains have their own requirements on the nature of this data access and integration. The clinical domain offers arguably the greatest challenges facing the roll-out and adoption of Grid technologies to meet the changing face of post-genomic clinical research, especially with regard to information governance, ethics and hence security solutions. This paper outlines a novel system design for secure anonymous data access and linkage that meets the needs of key stakeholders in this space including end user researchers, data providers and owners and ethical oversight bodies amongst others. We identify how existing solutions developed within the Medical Research Council funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project are being re-factored to meet the needs of these players and to address information governance criteria.

Keywords: Grid security, data linkage, anonymisation, virtual organizations.

1. Introduction

The vision of the Grid in providing seamless access to a range of digital resources, be they computers, data sets or other resources, is a compelling one. It is the case that different interpretations of the Grid have evolved however, reflecting the different needs and requirements of the different communities and the changing landscape of standards and technologies. In this paper we focus in particular on the needs and requirements of the clinical sciences, and especially with regard to the importance they place upon infrastructures that meet the strict demands on information governance. In previous papers, we have described and demonstrated the numerous ways in which secure access to clinical and biological data sets has been achieved [1-10] in a single unified framework. These approaches have been based primarily upon security-oriented role based access control [11] where the virtual organization (VO) specific attributes need for authorization decisions on VO resources have been securely pushed out to users or their local administrators exploiting capabilities such as delegation of authority [12], before being subsequently delivered in a user-oriented framework based for example upon the Internet2 Shibboleth-based access management framework (shibboleth.internet2.edu) that has been successfully rolled out across the UK (www.sdss.ac.uk). In addition to this federated attribute model, we have also demonstrated how attribute pull models have been supported and compared centralized versus decentralized attribute authority models [13] for fine grained security.

Despite these systems and the detailed experiences garnered in their development within projects such as the Medical Research Council (MRC) funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project (www.nesc.ac.uk/hub/projects/votes) and various others, we now recognized that irrespective of the technological solutions put forward by the Grid community, the issues of information governance and supporting the various players involved in this space has shown that data access and integration in the Grid sense is simply often not tenable. Data owners, data providers, ethical bodies amongst other stakeholders are extremely wary of any new middleware solutions which provide or *potentially* provide direct access to their data sets, i.e. requiring access

through their firewalls. Instead, new models of data usage are required which meet the stringent requirements of the stakeholders in this domain. Key questions raised throughout the course of the VOTES project include: *how can it be ensured that these data sets will not be disclosed to others?* Or in a similar vein, *how can it be ensured that this data will not be linked with other data resources which may include identifying information?* Anonymisation or pseudo-anonymisation models help in this regard, but it is notoriously difficult to truly anonymise clinical data whilst making it still useful for the end user scientists [14]. Moving towards genetic information of individuals these anonymisation processes raise further issues since the data by its very nature is identifying.

Irrespective of the trust models or security infrastructures in place or guarantees that any middleware developers might make about the robustness and usability of their security solutions, it is the case that clinical data providers simply will not risk *direct* access to their data sets for research purposes. This despite the fact that existing data sharing models typified by practices such as sending CDs through the post containing unencrypted clinical data is a far greater and well documented security risk. To counter such – perhaps understandable positions of data guardians - requires several considerations to be taken on board. Firstly, any solutions have to empower the stakeholders and not remove their essential roles in the access to and management of their data sets. Thus site autonomy is not just a buzzword but a fact that if violated will result in potential legal consequences. We note also that this autonomy has to be beyond simply installing Grid middleware and security software, and managing it locally since few clinical data providers are likely to wish to explore the currently complex offerings of Grid technology providers. Secondly and following on from the first consideration, any solutions have to be based upon pragmatic considerations of usability and accept that any developed systems must fit in to existing clinical systems and practices. Rolling out Grid based X509-based public key infrastructures for authentication systems incorporating advanced authorization infrastructures has to be considered from clinical provider perspectives who, experience in the VOTES project in the UK at least has shown, are busy working with a range of legacy systems far removed from the Grid vision and associated middleware. Thirdly, and something that the Grid community have not satisfactorily addressed, solutions have to be designed with a clear understanding of risk and worse case scenario in mind. The worst case scenario here might imply that the clinical data is accessed and used without express permission, or that it is linked with identifying data that results in disclosure of patient data.

In this paper, we describe the virtual anonymisation Grid for unified access to remote data (Vanguard) system under development within the context of the VOTES project that is being designed with these three factors at the forefront of the design process. The VOTES project itself is a three year project with a final 9-months remaining. Its focus is upon supporting the various stages involved in the conduct of clinical trials and epidemiological studies, namely: patient recruitment including feasibility studies of whether a trial/study has sufficient patient numbers meeting the given criteria for that particular trial; collection of data throughout the course of the clinical trial/study as well as supporting the overall study management. Throughout each of these stages it is paramount that the right information is made available to the right individuals (and only those individuals) to ensure both information governance and ethical considerations are strictly adhered to, and that the results of any trials can be independently validated according to strict and measurable criteria. One aspect which we emphasize within VOTES is that the goal was to develop a framework that could be applied for a range of clinical trials and studies and not simply develop a single bespoke system for a particular trial say. This has been achieved through the definition and implementation of a variety of clinical virtual organizations (CVOs) offering capabilities for data access and integration. These systems were designed upon user-oriented role based access control where services and data sets were made available through portals according to user privileges. Example trials undertaken within the existing VOTES systems include brain trauma trials [15], paediatric endocrinology trials with specific focus on congenital anomalies [16], primary care and secondary care trials [17] and more recently in establishing breast cancer tissue biobanks [18]. Large scale patient recruitment has been undertaken with the UK Biobank project (www.ukbiobank.ac.uk) which aims to recruit a cohort of 500,000 individuals for a range of studies, and where the VOTES partners are actively driving the software development activities.

A key aspect of the Vanguard design process which we outline in this paper is how the re-factoring of the numerous systems already in place within VOTES supports the previous pragmatic considerations to be recognized. The rest of the paper is structured as followed. Section 2 outlines the detailed requirements and initial designs put forward for the Vanguard system. Representative data linkage scenarios are provided to demonstrate the basic principles by which anonymous data linkage can be achieved. Section 3 outlines the various systems already in place in VOTES and how these are being adapted for use within the Vanguard system. Finally section 4 draws some conclusions and outlines further work and initial clinical trials under consideration to exploit the Vanguard system.

2. Design of Vanguard

The design of Vanguard is based upon a range of principals that must be strictly adhered to. The system is expected to have a design-lifetime of the order of decades, i.e. this is not just to develop yet another proof of concept system that ultimately is not hardened by active end user experience or does not have the recommendations of clinical data providers. It is essential that the Vanguard system should run on a variety of platforms. This means that proprietary software and protocols must be avoided. Furthermore the system must be able to survive network outages and hardware failure with minimum disruption to end-users.

Perhaps the most important principal that we are focused upon in the design and implementation of the Vanguard system is with regard to information governance. Specifically, we recognize that information must be exposed to the minimum extent possible or in many circumstances not at all – this implies that strong encryption must be used whenever data is exchanged between systems or temporarily stored outside of memory, and that datasets should be trimmed at source before transmission rather than on receipt, and of course that ultimate control of access to datasets must reside locally with their owners. A key consideration of the Vanguard system is with regard to the acknowledgment of the natural wariness (skepticism) of data providers. Experience from several years of working with clinical data providers is that they simply will not allow direct access through their firewalls to their data. There are many good reasons for this. For example, they are simply not prepared for finer grained authorization on access to resources to distinguish legitimate vs illegitimate requests. Indeed, this issue stems beyond research access to clinical data but also to access to clinical data by healthcare providers more generally. The furor associated with the establishment of a national consent database in the UK by the NHS connecting for health initiative is testament to this [19].

Instead of direct connections through healthcare provider firewalls, a variety of other scenarios have been explored. NHS-Higher Education gateways have been established and are currently being tested in the UK [20]. The establishment of demilitarized zones is another possibility which meets some of the reservations associated with clinical data providers on access to and usage of the data sets they maintain for healthcare purposes.

A new paradigm which is being explored within the development of the Vanguard system is based upon anonymous pull models of data linkage. Thus, rather than clinical data systems being queried by Grid based systems directly, these queries are defined based on an understanding of the data models of the different systems, i.e. knowledge of the schema of the data that may be available. If a given site has registered itself for participation in a given study, it may subsequently pull the generated queries into their clinical systems. Depending upon local security policies, these queries are validated and authorized, and if valid, will result in their execution. In short, the clinical systems are completely protected from inbound internet connections but rather are based upon a model only allowing outbound connections to be established. The Vanguard system itself is being designed based upon this pull model. However the question of security must still be explicitly satisfied, i.e. what queries are being defined by whom and what artifacts are coordinating the access to and usage of clinical data resources to users with particular privileges.

The Vanguard system architecture is shown in Figure 1 and shows the following principal components:

- *Viewer* – which is used by researchers who require access to data;
- *Agent* – which is the intermediary between other components;
- *Guardian* – which manages access to and data release from local resources;
- *Banker* – which logs usage and maintains use accounts for the clinical data access and usage;

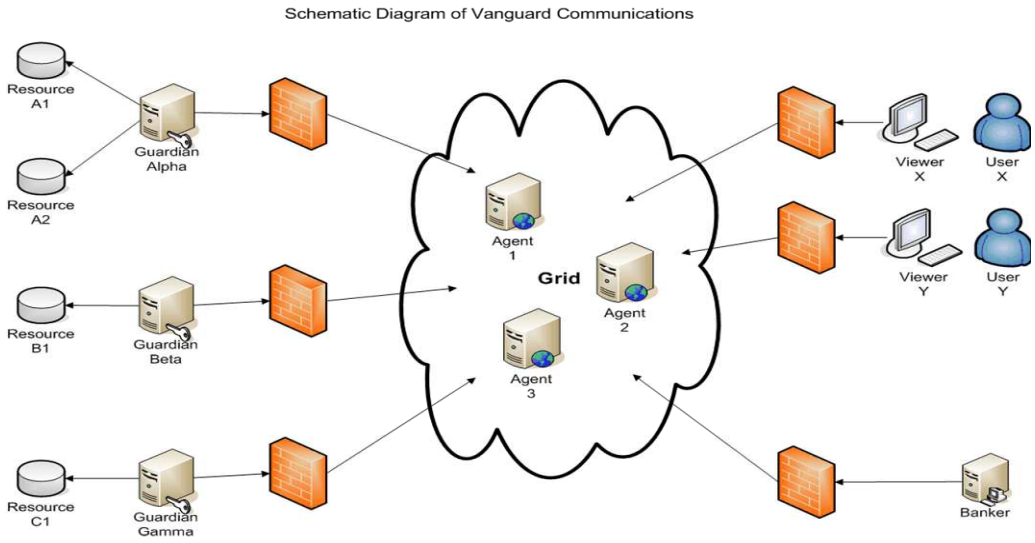

Figure 1: Vanguard Architecture

The roles of these components are defined as follows.

2.1. Viewer

The Viewer is an application run by the end-users of Vanguard. The viewer provides users with an interface to perform the following key functions:

- Display the different clinical data resources available to a particular user;
- Facilitate construction of data requests from these resources;
- Handle the datasets returned as a result of successful requests.

As described in section 3, we are primarily focusing upon the viewers being web based browser interfaces to portals. These portals provide an environment where different agents can be accessed and used depending upon user privileges. It is quite possible to have more than one agent involved in a given trial and study, however to begin with and to minimize the security risk we are focusing upon a model where each agent provides the data linkage possibilities for a single study. It may well be the case for software re-use for example, that re-factoring of agents is undertaken so that they can be combined in a variety of ways. The data available to a user within a given viewer is dependent upon the privileges that they possess. To support this model, we exploit digitally signed security attributes incorporating role based access control models. These attributes are specific to a given study or trial and allow access to one or more data resources, where data providers agree how their combined data can be linked in particular ways. Once defined, the security attributes are then used to enforce local access decisions – or more precisely, they are presented to the local data providers along with the queries that are generated. The combined roles and queries are then used to determine the authorization decision on access to the local data.

2.2. Agent

Agents play a pivotal part in the Vanguard system and typically perform the following roles:

- They enable communication between other system components including viewers, guardians, bankers and potentially other agents;
- They accept the generated user queries and manage the query requests themselves including their transfer, delivery and dealing with the associated security;
- They collate and manage the results of the submitted queries.

Agents are at the fulcrum of the Vanguard system in that they are responsible for ensuring both the secure communications between the different system artifacts and their coordination. A key role of the Agents is in the generation of hash keys for use within secure communications. These are passed

through to the various Guardians that they help to coordinate. Through hashing of data items that should remain anonymous, data can be securely linked across sites without direct data disclosure being made.

Agents also provide capabilities for defining the plans by which queries generated via a user through a viewer can be both formulated and subsequently enacted. Thus for example, if a particular data provider which a certain data element to be hashed or closed, then the Agent is responsible for ensuring that the appropriate hash keys are distributed and for removing the final hashed values once data has been linked across sites.

In the existing Vanguard system we are developing a range of secure protocols reflecting the variety of infrastructures at clinical data provider sites. Examples of these include web security models exploiting X509 credentials for example as well as SSL/TLS for message/channel security respectively.

2.3. Guardian

Guardians act as a controlled secure gateway between the local data at clinical data provider sites and external Agents. One or more Guardian systems will be installed at every site which provides a data resource for the system. The Guardians will typically perform the following roles within the Vanguard system:

- Construction of site specific security policies;
- Describe the local data and security policies to known and trusted Agents;
- Enforce security policies related to incoming (pulled) queries;
- Handle incoming data requests from Agents;
- Export results of requests to Agents according to site specific data release policies.

We expect a range of Guardians to exist within the Vanguard system reflecting the existing systems already in place across the clinical data provider sites. Initially we are focused upon models where Guardians supporting role based access control policy definition and enforcement are in place.

In the Vanguard system Guardians allow for data providers to make available their data models and importantly the way in which the data associated with these data models may be access and/or linked. Specifically, we consider the situation where the data that a given data might have can be assigned access privileges of:

- *Open* – in which case the Guardian is willing to supply the actual value of the data field;
- *Hashed* – in which case the Guardian is willing to supply a hashed (and hence anonymised) value of the data;
- *Closed* – in which case the Guardian will not supply the value, but is willing to run queries for example that involve it as a selector;

We note that the clinical databases accessible within Vanguard may well have other fields which do not form part of this system. The existence of such fields is hidden entirely from the Agents. As a security precaution for data providers, we strongly recommended that Guardian `owners' create a set of read-only views of their data resources which contain the fields they are willing for a Guardian to process, and which do not contain any other fields.

Prior to running any queries, the Guardian must supply a description of the data that it has to the Agent. This will typically contain a list of the names and types of the data resources a Guardian is managing; the version information of the Guardian and a list of the features that a Guardian can support. Initially our focus is upon relational database resources supporting SQL-based queries hence this kind of information includes the list (names) of tables in the database; a textual description of the database contents; for each table a list of names of fields in table; text description of the table contents and the number of rows in each table; for each field in each table information might include the field type (int, string etc); the protection level (Open, Hashed, Closed); a textual description of the field contents; alternative nomenclature(s) for the field, e.g. Snomed codes; nullability; the uniqueness of the field; relationship to other fields, e.g. is-a-foreign-key; and the size of the field itself. In short, the Guardian must provide detailed information on the data model for the databases it makes available so that this can subsequently be used for data access and linkage by the Agents. We emphasize that this is purely the data model that is being made available via the Guardian and not the data itself.

2.4. Banker

The Banker in the Vanguard system is responsible for managing resources across the whole system. Bankers have the following main roles

- Maintain a log of actions taken across given trial systems – specifically through recording the queries generated by the viewers/agents and those sent to the guardians;

- Maintain charging accounts for users – to ensure for example that a single user is not over-utilizing the federated data available through the Vanguard system.

2.5. Vanguard Component Interaction Scenario

The interactions between the previous components in supporting secure anonymous data access and linkage proceeds as follows. In the first instance, we assume that ethical approval for a given trial is applied for and granted as per typical procedures. In the UK this might for example be through applying for Multicentre Research Ethics Committee (MREC) or Local Research Ethics Committee (LREC) approval. In addition, we assume the precondition that an Agent for this particular trial has requested the data-schemas from all Guardians it is aware of. This includes information on the visibility of the data sets themselves, e.g. whether they are open, closed or hashed. The trial coordinator will then use a Viewer to request the data-schemas available from that Agent to construct particular queries. We note that the trial coordinator may construct particular views of data for the different roles of individuals involved in a particular trial, e.g. a nurse may issue the following queries, or an ethical oversight committee member may see all data etc. The Vanguard system extends the existing VOTES systems to support such role-based scenarios. We outline the basic functionality here in terms of creating and executing queries and retrieving query results since they are similar irrespective of the role in the study.

2.5.1. Query Creation

To create a specific query the end-user uses the Viewer to construct a query based on the data-schema available to them. The query is transmitted from the Viewer to an Agent, where it is stored and given a unique ID.

2.5.2. Query Execution

To start executing a query the Viewer transmits a signal to the Agent requesting that a previously-stored query is executed. This is accompanied by a public-key generated by the end-user (PKU). The Agent verifies that the query is permitted and produces an action plan decomposing the query into local-queries across any Guardian system(s) required. Fields are tagged as either being required for external-joining (within the Agent) or purely for returning to the Viewer.

The Agent generates the queries and either issues them directly to the Guardian systems (where they allow direct querying), or signs and stores the queries. Each non-directly querying Guardian will then periodically pull these queries down and if valid/authorized, return the results associated with it. The query requests themselves are accompanied by PKU, a public-key for the Agent (PKA), and a unique (per-query) hashing key generated by the Agent (HA).

To execute the parts of this query each Guardian checks the local-query against its local access-schema to verify it is permitted. If satisfying the local policy, the Guardian executes the query. In the Vanguard system we are exploiting both role based access control models based around the PERMIS infrastructure [21] and access control matrices as described in [22]. Fields tagged for external-joining by the Agent are hash-encoded using HA. Fields tagged for returning to the Viewer are encrypted using PKU. The whole datasets are then encrypted using the PKA and returned to the Agent along with the cost for executing the query in resource credits.

On receipt of the local-queries the Agent stores the resultant datasets until all partial queries have returned. Once all queries have returned, the Agent transmits a signal to the Banker with the action taken and the number of resource credits used, e.g. the number of result sets. In addition, the agent joins the partial queries according to any external-joining fields. It will also discard any external-joining fields that the end-user has not requested or that the end user does not have the privilege to view. Finally it sets a query flag to indicate that the query is completed.

2.5.3. Query Retrieval

Within the Vanguard system, whilst a query is being executed the user is able to use the Viewer to see the state of progress of the query. Once the query-complete flag has been set on the Agent, the end-user uses the Viewer to download the results of the query from the Agent. At this point, the Agent deletes all data returned by the query and passes all logging/charging data to the Banker. Once acknowledged by the Banker, the Agent deletes this information from its cache.

2.6. Example of Vanguard System

To understand how these various components can be used for secure data linkage within the Vanguard system we consider the following example. We assume that a range of clinical datasets are distributed across clinical data provider sites *alpha, gamma* and *delta* as depicted in Figure 2 hosting the datasets stay and birth, linkage and disease respectively.

alpha.stay

Field	Type
hospID	Integer
mother	Integer
days	Integer
status	Integer

alpha.birth

Field	Type
nhs	String
mother	Integer
dob	Date
weight	Real
sex	Int

gamma.linkage

Field	Type
nhs	String
chi	Int
Active	Bool

delta.disease

Field	Type
chi	Int
hiv	Bool
hepatitis	Bool

Figure 2: Example Clinical Data Providers and Data Provision

With the above tables we assume that the National Health Service (NHS) number in table alpha.birth, and the community health index number (CHI) number in data resource delta.disease must both be hashed (represented here through different colouring). Similarly, the HIV information in the database delta.disease is closed and hence cannot be disclosed. With this data model, in place we wish to answer the following query: *How many days did mothers with HIV stay in hospital?*

The SQL to run this query *directly* is represented in Figure 3.

```
SELECT alpha.stay.days
    WHERE alpha.stay.mother = alpha.birth.mother
        AND alpha.birth.nhs = gamma.linkage.nhs
        AND gamma.linkage.nhs = gamma.linkage.chi
        AND gamma.linkage.chi = delta.disease.chi
        AND delta.disease.hiv = true
```

Figure 3: SQL for Direct Querying of Clinical Data Sets

However, given that certain information associated with these data resources is not directly visible to the Agents and hence to the end users via the Viewers, the actual SQL plans that is generated by the Agents needs to link the data and remove unnecessary information or data which is flagged as being for restricted disclosure. In this case with the NHS and HIV data information limitations the SQL plan generated by the Agent in this case is shown in Figure 4.

```
SELECT alpha.stay.days,H(alpha.birth.nhs)
    WHERE alpha.stay.mother = alpha.birth.mother
SELECT H(gamma.linkage.nhs),H(gamma.linkage.chi)
SELECT H(delta.disease.chi) WHERE delta.disease.hiv = true;
Join on H(*.nhs) AND H(*.chi), then remove H(*.nhs) and H(*.chi)
```

Figure 4: SQL for Anonymous Linkage of Clinical Data Sets

In this case, the data that is restricted, e.g. the NHS number in alpha.birth, is hashed (and hence anonymised). These hash values are unique since a different hash key is used each time by the Agents and it is guaranteed that the same hash key will not be used by the Agents. Thus through the use of hashed information across the different data resources, data linkage can be made, yet direct data disclosure is avoided. Furthermore, the final line of the query above then removes the hashed values to ensure that the final resultant data set protects the required confidentiality of the data providers.

It should be emphasized that the primary benefit of this scenario, is that no identifying information is released from the data providers yet data linkage is made across different data provider sites. This model thus satisfies the data providers worries of their data potentially linked in unforeseen ways with other data resources. The result of the query will then simply be the number of days that mothers with HIV stayed in hospital, without any information identifying which mothers for example.

3. Re-use of Existing VOTES Infrastructure

The Vanguard system is the latest incarnation of the VOTES infrastructure. Numerous prototypes of VOTES components have been established previously. In this section we briefly outline how we are adapting the VOTES systems to support Vanguard interaction models.

The overall VOTES architecture is depicted in Figure 6. In this infrastructure, a portal is provider through which a variety of clinical trials systems and studies are available. Access to the portal is achieved through the Internet2 Shibboleth technologies using the UK Access Management Federation augmented with attributes specific to the roles of the participants involved in the particular studies. Through the portal numerous services and data services are available which allow for secure access to the federated data sets. A key aspect of this infrastructure is the trust model whereby each data provider decides itself which data sets it wishes to make available and in turn the local authorization decisions on access requests which must strictly adhere to local policies on access and usage.

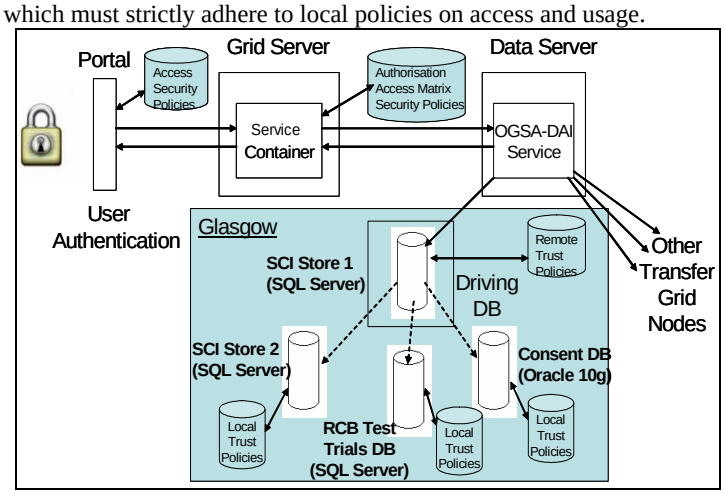

Figure 6: Overall VOTES Architecture

The actual roles and the privileges that are applicable within a given trial can be defined in several ways. Through delegation of authority, these roles can be pushed out to remote collaborators as described in [12]. Alternatively, centralized models for the definition of the security attributes can be achieved as described in [13]. A third model and one which lends itself directly to the Vanguard system is where the roles and are defined expressly with the data models and the openness of the data sets in mind. Through portlets available to privileged users, e.g. trial coordinators, the different views of data can be defined and assigned to specific roles as depicted in Figure 7.

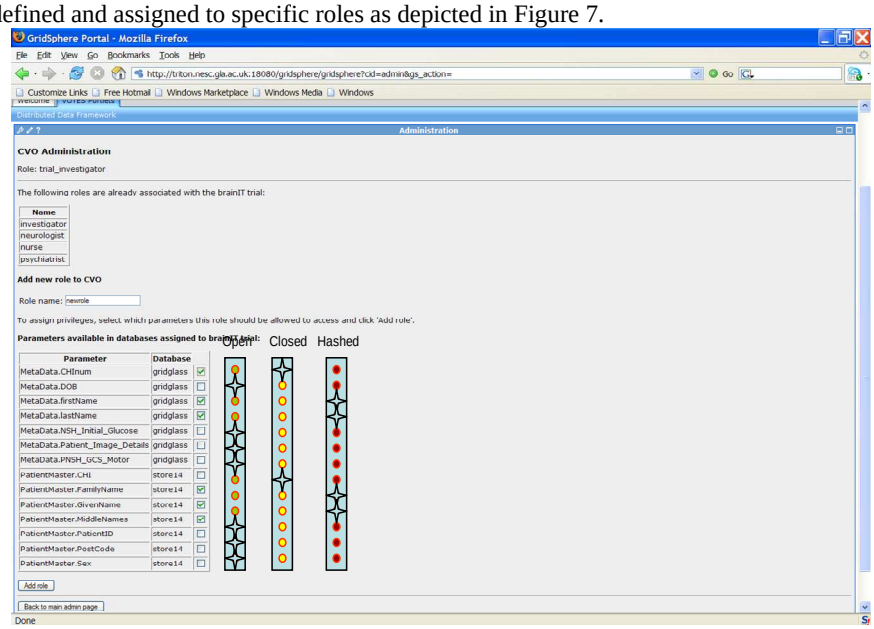

Figure 7: Clinical Virtual Organisation Administration Interface

The definition of the roles identified above is used to configure access to the various services and hence data available within the portal. In the Vanguard system, it is implicit that the trials coordinator is limited to defining roles that are a subset of the roles that they themselves have been allocated. Or put another way, it is not possible for the trials coordinator to offer roles that allow to link data in ways that they themselves are not able to link.

The primary difference between the VOTES architecture outlined in Figure 6 and the Vanguard system is with regard to the pull model versus the push model of the queries. In the existing VOTES infrastructure, numerous proofs of concept prototypes showing how privileged users can generate queries that can be federated and sent to remote databases that have their own local authorization policies to decide on the validity and authorization of the queries. These have, until now primarily been proofs of concept and not live systems, e.g. they have been built on either data provided directly by clinical data providers or for example the training data sets that they use. This has been due to the inability to directly access the clinical data resources through for example penetrating the NHS firewalls. In the Vanguard system with the pull model, it is no longer necessary to directly tunnel through these firewalls. Rather, the queries can be pulled in and hence the clinical data centers remain completely in control of their firewalls.

4. Conclusions

The work on the Vanguard system is on-going however we have clearly defined the components and the interactions between them that will overcome the immediate concerns from the clinical data providers as has arisen throughout the course of the VOTES project. Nevertheless the work we have described here is not complete and numerous challenges remain to be addressed. Some of the most critical challenges that we envisage include scalability. The scenarios outlined in Figures 3 and 4 are at a small and understandable scale. In real clinical systems in Scotland such as SCIstore, GPASS and the Scottish Morbidity Records amongst others however we are often dealing with database models comprising several hundred tables with complex field and primary/foreign structures. Furthermore it is often the case that few clinical data centers are well positioned to disclose what data tables/fields should be made available. In the UK numerous solutions exist and are outsourced to commercial software providers who are often unwilling to disclose their detailed data models. As such, knowing what data can be disclosed and linked is often not a trivial exercise. This is often further complicated through fields that can be used for textual information on a given patient for example and including identifying patient information for example. To address such scenarios, we believe the only way is to develop the systems in close liaison with the clinical providers and only after they are completely satisfied that the systems meet their rigorous information governance policies can they be used in a truly live setting.

A second challenge that we expect to face in the roll-out of the Vanguard system is with regards to data disclosure risks arising due to global data models. Thus it is only when the various data providers have agreed to release their data sets that the issues of identifying data sets can arise. As one example, it would be quite possible for to extend the *alpha.birth* data provider given above with the patient name and data of birth and allow open access to this information, but this may well be opposed to the data disclosure policy of *delta.disease* for example. It is only when considering the joining or union of these data sets on the CHI and NHS numbers in the example above that such policy conflicts can be identified.

We also note that many of the challenges faced in obtaining access to clinical data stem from researchers being considered as external to the clinical and administrative bodies such as the NHS. To overcome these issues, the NeSC team is in the process of being allocated NHS honorary contracts through the work on the breast cancer tissue bank for example.

4.1. Acknowledgements

The work described here was suported by a grant from Medical Research Council in the UK to support the efforts of the Virtual Organisations for Trials and Epidemiological Studies (VOTES) project. The authors thank the partners involved in the project for their inputs.

5. References

[1] R.O. Sinnott, From Data Access and Integration to Mining of Secure Genomic Data Sets, International Journal of Grid Computing: Theory, Methods and Applications, Special Issue on Life Science Grids for Biomedicine and Bioinformatics, pp 447-456, Volume 23, Issue 3, March 2007.

[2] R.O. Sinnott, O. Ajayi, J. Jiang, A. J. Stell, J. Watt, User-oriented Security Supporting Inter-disciplinary Life Science Research across the Grid, New Generation Computing, Special Edition on Life Science Grids, editors A. Konagaya, P. Arzberger, T. W. Tan, R. Sinnott, D. Angulo, pp 339-354, Vol. 25 No. 4, 2007.

[3] R.O. Sinnott, O. Ajayi, A.J. Stell, Supporting Grid Based Clinical Trials in Scotland, Health Informatics Journal Special Issue on Integrated Health Records, Vol. 14 (2), June 2008.

[4] R.O. Sinnott, M. Bayer, D. Berry, M. Atkinson, M. Ferrier, D. Gilbert, E. Hunt, N. Hanlon, Grid Services Supporting the Usage of Secure Federated, Distributed Biomedical Data, Proceedings of UK e-Science All Hands Meeting, September 2004, Nottingham, England.

[5] R.O. Sinnott, D.W. Chadwick. Experiences of Using the GGF SAML AuthZ Interface, Proceedings of UK e-Science All Hands Meeting, September 2004, Nottingham, England.

[6] R.O. Sinnott, A.J. Stell, D.W. Chadwick, O.Otenko, Experiences of Applying Advanced Grid Authorisation Infrastructures, Proceedings of European Grid Conference (EGC), LNCS 3470, pages 265-275, Volume editors: P.M.A. Sloot, A.G. Hoekstra, T. Priol, A. Reinefeld, M. Bubak, June 2005, Amsterdam, Holland.

[7] R.O. Sinnott, A.J. Stell, J. Watt, Dynamic Privilege Management Infrastructures Utilising Secure Attribute Exchange, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.

[8] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Advanced Security on Grid-Enabled Biomedical Services, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.

[9] R. O. Sinnott, M. M. Bayer, J. Koetsier, A. J. Stell, Grid Infrastructures for Secure Access to and Use of Bioinformatics Data: Experiences from the BRIDGES Project, 1st International Conference on Availability, Reliability and Security, (ARES'06), Vienna, Austria, April, 2006.

[10] R.O. Sinnott, J. Watt, O. Ajayi, J. Jiang, Shibboleth-based Access to and Usage of Grid Resources, IEEE International Conference on Grid Computing, Barcelona, Spain, September 2006.

[11] D.W.Chadwick, A. Otenko, E.Ball, Role-based Access Control with X.509 Attribute Certificates, IEEE Internet Computing, March-April 2003, pp. 62-69.

[12] R.O. Sinnott, J. Watt, D.W. Chadwick, J. Koetsier, O. Otenko, T.A. Nguyen, Supporting Decentralized, Security focused Dynamic Virtual Organizations across the Grid, 2nd IEEE International Conference on e-Science and Grid Computing, Amsterdam, December 2006.

[13] R.O. Sinnott, D. Chadwick, T. Doherty, D. Martin, A. Stell, G. Stewart, L. Su, J. Watt, Advanced Security for Virtual Organizations: Exploring the Pros and Cons of Centralized vs Decentralized Security Models, 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2008), May 2008, Lyon, France.

[14] Elliot, M., Purdam, K., Smith, D. Patient Record Data: Disclosure Control for Grid Based Data Access, Proceedings of National Centre for e-Social Science Conference Manchester, 2006.

[15] R.O. Sinnott, A.J. Stell, O. Ajayi, Development of Grid Frameworks for Clinical Trials and Epidemiological Studies, HealthGrid 2006 conference, Valencia, Spain, June 2006.

[16] R.O. Sinnott, O. Ajayi, A.J. Stell, Grid Infrastructures Supporting Paediatric Endocrinology across Europe, UK e-Science All Hands Meeting, Nottingham, UK, September 2007.

[17] R.O. Sinnott, O. Ajayi, A.J. Stell, Secure Federated Data Retrieval in Clinical Trials, Telemedicine 2006 conference, Banff, Canada, July 2006.

[18] Breast Cancer Tissue Bank, www.nesc.ac.uk/hub/projects/bctb

[19] Family doctors to shun national database of patients' records, 20 Nov 2007, records of 50 million NHS patients on a national electronic database, on the database without getting a patient's specific consent, www.guardian.co.uk/society/2007/nov/20/nhs.health

[20] The NHS-HE Connectivity Gateway Project, http://www.nhs-he.org.uk/

[21] D.W.Chadwick, A. Otenko, The PERMIS X.509 Role Based Privilege Management Infrastructure, Future Generation Computer Systems, 936 (2002) 1–13, December 2002. Elsevier Science BV.

[22] R.O. Sinnott, O. Ajayi, A.J. Stell, Secure, Reliable and Dynamic Access to Distributed Clinical Data, Life Science Grid Conference, Yokohama, Japan, October 2006.

Author/s:
Sinnott, R; Ajayi, O; Stell, A; Young, A

Title:
Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data

Date:
2008-01-01

Citation:
Sinnott, R., Ajayi, O., Stell, A. & Young, A. (2008). Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data. Solomonides, T (Ed.) Silverstein, JC (Ed.) Saltz, J (Ed.) Legre, Y (Ed.) Kratz, M (Ed.) Foster, I (Ed.) Breton, V (Ed.) Beck, JR (Ed.) GLOBAL HEALTHGRID: E-SCIENCE MEETS BIOMEDICAL INFORMATICS, 138, pp.90-101. IOS PRESS.

Publication Status:
Published