

User Oriented Access to Secure Biomedical Resources through the Grid

Richard Sinnott¹, Oluwafemi Ajayi¹, Jipu Jiang¹,
Anthony Stell¹, and John Watt¹

¹National e-Science Centre, University of Glasgow
Glasgow G12 8QQ r.sinnott@nesc.gla.ac.uk

Abstract. The life science domain is typified by heterogeneous data sets that are evolving at an exponential rate. Numerous post-genomic databases and areas of post-genomic life science research have been established and are being actively explored. Whilst many of these databases are public and freely accessible, it is often the case that researchers have data that is not so freely available and access to this data needs to be strictly controlled when distributed collaborative research is undertaken. Grid technologies provide one mechanism by which access to and integration of federated data sets is possible. Combining such data access and integration technologies with fine grained security infrastructures facilitates the establishment of virtual organisations (VO). However experience has shown that the general research (non-Grid) community are not comfortable with the Grid and its associated security models based upon public key infrastructures (PKIs). The Internet2 Shibboleth technology helps to overcome this through users only having to log in to their home site to gain access to resources across a VO – or in Shibboleth terminology a federation. In this paper we outline how we have applied the combination of Grid technologies, advanced security infrastructures and the Internet2 Shibboleth technology in several biomedical projects to provide a user-oriented model for secure access to and usage of Grid resources. We believe that this model may well become the *de facto* mechanism for undertaking e-Research on the Grid across numerous domains including the life sciences.

Keywords: Grid, security, PKI, Shibboleth, clinical trials, epidemiological studies

1 Introduction

The life science domain is booming! The explosion of research areas, exponential growth of the associated data sets and the proliferation of new discoveries across and between disciplines is unparalleled. Building infrastructures to support such a highly volatile area is a fraught process [1]. As new insights and discoveries are made, existing research models and associated data sets have to be augmented, refined and extended to incorporate such new knowledge. This poses a fundamental challenge to infrastructure providers: how to build an infrastructure that has any form of longevity [2]?

One thing which is clear is that scientists need to be able to collaborate with one another. However scientists in the post-genomic era are wary and protective of their own research. Collaborators in grants are also potential competitors in future funding proposals. Being the first scientist to make a major discovery is a strong driver for many leading researchers both in the present and the past. Crick and Watson and their identification of the double helix structure of DNA were driven by their knowledge that competitors such as Linus Pauling were also trying to identify this structure [3]. Similarly the direct financial benefits from given lines of research can pay huge dividends with interest from the multi-billion dollar pharmaceutical industry given the costs they incur in developing new drugs and evaluating their effectiveness.

In this context scientists need to be ensured that they access and share trusted data from collaborators and that this is in accordance with the commonly agreed terms and goals of the collaboration. In the context of the Grid, such co-operations are often termed virtual organisations (VO). From past experiences [4] it is clear that vast amounts of the non-Grid community are uncomfortable with the Grid. System administrators regard it as a possible security threat whilst potential Grid end users are mystified and put off with the common security models needed to access and use a Grid. If the Grid is to ever gain the widespread acceptance envisaged, it needs to be as simple to use as the internet, with front end tools no more complex than existing browsers.

The Shibboleth technology from the Internet2 project [5] has put forward software architecture [6] and associated protocols [7] for new models of security. Rather than a user being required to remember numerous usernames and passwords required to access resources across the internet, with the Shibboleth trust model, a user is able to access resources across a federation through signing in (authenticating) at their own local site. In this paper we present the Shibboleth model of security and outline how it has been applied across a range of biomedical projects to simplify the access and usage of Grid services and data. In all of this, a finer grained model of security is achievable.

The rest of the paper is structured as follows. In section 2 we provide an overview of Shibboleth and the challenges that are needed for this technology to be accepted by the scientific community, especially what it means to trust. In section 3, we present various case studies illustrating how we have successfully combined Shibboleth. Finally in section 4, we draw conclusions and outline plans for the future.

2 Overview of Grid Security and Shibboleth

The majority of Grid solutions today are based upon X.509 digital certificates [8] to support public key infrastructures (PKI) [9]. PKIs are based upon cryptographic technology where public and private keys are used to digitally sign (encrypt) and decrypt messages and information. Messages encrypted with a public key can only be read by an individual who possesses the associated private key. Through this mechanism, a given user can direct a message to a known destination, knowing that it can't be read by anyone

else, simply by encrypting it using the public key of that destination. The owner of the private key can encrypt messages with that key, and the receiver of the message can be sure that it was sent by the owner of the private key. Both public key agreement and public key transport need to know who the remote public key belongs to, i.e. who has associated private key. The public key certificate is the mechanism used for connecting the public key to the user with the corresponding private key. Public key certificates include a Distinguished Name (DN) which can be used for identifying a given user. The establishment of the identity of a given user (or machine) within a public key certificate is ultimately based upon trust, or more precisely trust by a community of one or more Certificate Authorities (CAs).

A CA is a root of trust which holders of public and private keys agree upon. CAs have numerous responsibilities including issuing of certificates, issuing Certificate Revocation Lists (CRL) and they need to have well documented processes and practices which must be followed to ensure identity management. Various PKI architectures are possible and the selection of which depends upon numerous factors. Whether numerous CAs are to be trusted? How important to be able to add new CAs? What kind of trust relationships exist between CAs? The simplest PKI involves a single CA which is trusted by all users. With this model, users only accept certificates and certificate revocation lists issued by this CA. This model makes certificate path analysis easy since there is a single step from a certificate to the CA who issued it. One danger of this PKI infrastructure is that the CA is single point of failure. Thus if it is compromised, then potentially all certificates that have been issued are compromised, requiring all users to be contacted and certificates revoked. The ramifications of such a compromise would be catastrophic with potentially all resources that had been accessed using certificates issued by this CA having to be completely reinstalled (in case backdoor software solutions had been installed). Perhaps more of an issue would be the level of trust and how Grids using PKIs were perceived by the wider community.

Other more complex PKI architectures also exist. For example, users may keep a list of trusted CAs. However, issues such as how to tell trustworthy one from untrustworthy one arise? Hierarchical PKIs where there are chains of trust between the CA, sub-ordinate CAs and users may also exist. This model allows limiting the damage caused by a compromised subordinate CAs. Thus if a subordinate CA is compromised then only the certificates issued by them (or their subordinate CAs) need to be revoked. Other more complex architectures exist again, such as meshes of PKIs where trust relationships (webs of trust) are established on a peer-peer basis. This model often requires bridging solutions [10,11] between CAs and results in certificate paths that are harder to establish – potentially containing loops.

The PKI architecture chosen for UK e-Science is based on a statically defined centralised CA with direct single hierarchy to users. The typical scenario for getting a certificate is as follows. Researchers wishing to gain access to Grid resources such as the NGS (www.ngs.ac.uk) in the first instance have to acquire a UK e-Science X.509 certificate issued by the centralised Certification Authority (CA) at Rutherford Appleton Laboratory (RAL) (www.grid-support.ac.uk/ca). They will thus apply for a certificate via

the Grid Support web site (www.grid-support.ac.uk). The CA will then contact their local Registration Authority (RA) – hopefully at their institution!, who will in turn contact the user and request some form of photographic identification (such as a passport photo or university card). Once the identity of the user has been ratified, the RA contacts the CA who subsequently informs the user (via email) that their certificate is available for download. The user downloads the certificate and associated certificate revocation lists into their browser. Once in their browser they are required to export it to forms appropriate to the Grid middleware. Certificates can be acquired for both users and servers/machines.

The main benefit and reason for the widespread acceptance of PKIs within the Grid community is their support for single-sign on. Since all Grid sites in the UK trust the central CA at RAL, a user in possession of an X.509 certificate issued by RAL can send jobs to all sites, or rather to all sites where a user has requested and been granted access to those sites. Typically with Globus based solutions [12], gatekeepers are used to ensure that signed Grid requests are valid, i.e. from known collaborators. When this is so, i.e. the DN of the requestor is in a locally stored and managed grid-mapfile, then the user is typically given access to the locally set up account as defined in the grid-mapfile.

For the *existing* Grid community, PKIs are a widely accepted and common model for how to support a basic level of security (authentication). The problem is however that the existing Grid community are only a small fraction of the wider e-Science and e-Research community more generally, e.g., in the UK ~3500 UK e-Science X.509 certificates have been issued by RAL, but there are over 3 million academics across higher and further education colleges in the UK registered to use the Athens authentication infrastructure. Thus it could be claimed with some justification that the Grid has only touched the tip of the iceberg in terms of the wider research community.

There are many possible reasons for this as identified in [4]. These include the learning curve in accessing and using Grids. Most scientists do not want to gain access to a user account on a HPC resource but want instead to access a service which performs some function, e.g. BLAST in the case of the bioinformatics community. Why should a biologist go on a training course on Grid technology when all they require is access to a BLAST service on a free national HPC resource for example? Furthermore, the initial hurdles that have to be overcome in getting “on the Grid” in terms of acquiring and using an X.509 certificate are non-trivial for less IT-oriented researchers. For example, users are expected to convert the certificate from their CA which they initially install in their browser into appropriate formats understandable by Grid middleware. This requires them to run obscure *openssl* commands, and since *openSSL* is not commonly available on platforms such as Windows they are then often required to install and configure additional software. In some circumstances this is also not possible, e.g. if they do not have sufficient privileges on their PC (root access etc). In this case the researchers will instead have to refer to a local system administrator to help with the installation and configuration.

Assuming researchers have managed to obtain a certificate which they have converted into the appropriate format, they are then expected to remember necessarily

strong passwords for their private keys with the recommendation to use upper and lower case alphanumeric characters. The temptation to write down such passwords is obvious and an immediate potential security weakness.

This whole process does not lend itself to the wider research community which the e-Science and Grid community needs to reach out to and engage with. It is a well known adage that the customer is always right. Usability and addressing researcher requirements is crucial to the uptake and success of Grid technology. End user scientists require software which simplifies their daily research and not make this more complex. Given the fact that the initial user experience of the Grid currently begins with application for an e-Science certificate, this needs to be made as simple as possible, or potentially removed completely.

PKIs support authentication, however it is clear that the vast majority of researchers require much finer grained security infrastructures which support *authorisation*. Not just establishing the identity of a given user at a resource, but in defining and enforcing how they might access and use a given resource, or if sufficient information is not given, rejecting the request and subsequently logging the information. As a concrete example, it will never be the case that someone is allowed to access a UK National Health Service (NHS) resource to run arbitrary code. It may however be possible for a user to access a given resource to read a given anonymised database provided they have been given the privileges to do so in advance.

Despite these limitations, single sign-on is a compelling model and any refinements, extensions or new solutions for Grid security must provide similar capabilities. Such models should also be targeted to, and at the complete discretion of the resource provider to provide site autonomy. The Internet2 Shibboleth technology provides one way in which many of these issues can be resolved.

Shibboleth introduces several concepts which are fundamental to access and use Grid resources. These include an Identity Provider (IdP), a Service Provider (SP) and optionally a Where Are You From Service (WAYF). The basic scenario by which these components interwork is depicted in Figure 1.

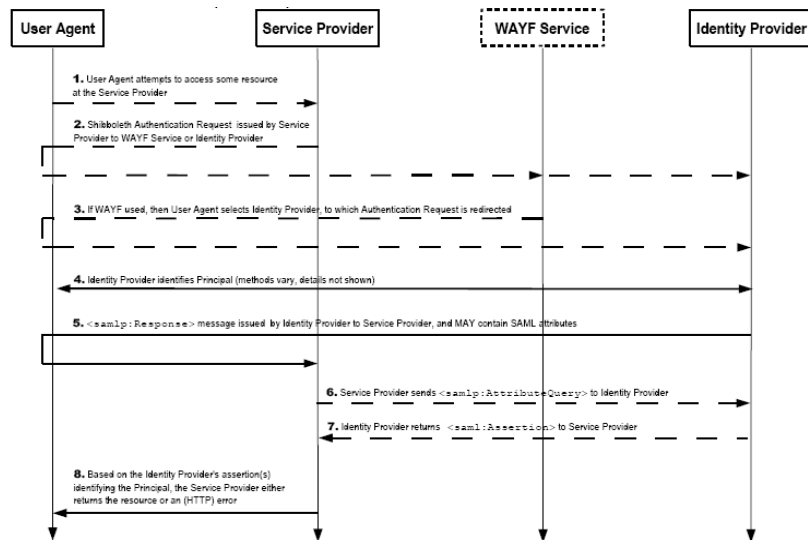


Figure 1: Basic Shibboleth Interactions for Accessing a Resource

When a user attempts to access a Shibboleth protected service or Service Provider (SP) more generally, they are typically redirected to a WAYF server that asks the user to pick their home Identity Provider (IdP) from a list of known and trusted sites. The service provider site has a *pre-established trust relationship* with each home site, and trusts the home site to authenticate its users properly. In the UK a single federation is planned [13]. Other international federations have also been put forward and established [14-17].

After the user has picked their home site, their browser is redirected to their site's authentication server, e.g. an LDAP repository, and the user is invited to log in. After successful authentication, the home site redirects the user back to the SP and the message carries a digitally signed SAML [18] authentication assertion message from the home site, asserting that the user has been successfully authenticated (or not!) by a particular means using an authentication mechanism specific to the IdP. Assuming the digital signature on the SAML authentication assertion is verified and the user has successfully authenticated themselves at their home site, then the SP has a temporary pseudonym for the user (the handle), the location of the attribute authority at the IdP site and the service provider URL that the user was previously trying to access. The resource site then returns the handle to the IdP's attribute authority in a SAML attribute query message and is returned a signed SAML attribute assertion message.

In Shibboleth a trust model exists in that the target site trusts the IdP to manage each user's attributes correctly, in whatever way it wishes. So the returned SAML attribute assertion message, digitally signed by the origin, provides proof to the target that the authenticated user does have these attributes.

This security model offers several direct benefits over PKIs for dynamic establishment of VOs in that users are no longer trusted to manage their X509 certificates and remember complex passwords. Instead institutions within a federation have a degree of trust with one another. Sites/IdPs and SPs are still autonomous and are able to decide for themselves whether the provided attributes are sufficient for access to the resources and which attributes they are prepared to release to which SP. Another key benefit of Shibboleth for VO establishment and management is that users are only required to remember their own usernames and passwords at their home institutions.

Provided a common understanding of the roles and security attributes across the sites comprising the federation exists, single sign-on can be achieved. Thus if a SP trusts a given site for authenticating a user requesting access to its own resource, and also an agreement on the attributes which are to be exchanged between the sites exists, then the SP can authorize/restrict access to its resources from those sites that are within the federation provided the necessary attributes and values are presented by the IdP.

Within the UK federation, a small set of security attributes based upon a subset of the *eduPerson* specification is being adopted [19]. These attributes include *eduPersonScopedAffiliation* which indicates the user's relationship (e.g., staff, student, etc.) within their home institution; *eduPersonTargetedID* which is needed when an SP is presented with an anonymous assertion only as provided by *eduPersonScopedAffiliation*; *eduPersonTargetedID* attribute which provides a persistent user pseudonym; *eduPersonPrincipalName* which is used where a persistent user identifier, consistent across different services is needed, and *eduPersonEntitlement* which enables an institution to assert that a user satisfies an additional set of specific conditions that apply for access to a particular resource. A user may possess different values of the *eduPersonEntitlement* attribute relevant to different resources.

To support single sign-on across numerous resources, session information is maintained. Thus a user is able to specify whether the WAYF should remember them for the duration of the session, for a week or not at all. In accessing subsequent Shibboleth protected services, the WAYF will automatically recognize which IdP the users are from and redirect them accordingly.

Ensuring that an institution in a Shibboleth federation can guarantee the authenticity of a user when accessing a remote resource is crucial to the overall principles upon which Shibboleth and Shibboleth federations are based. In short, institutions in a federation should *trust* one another. It is the case however, that users at larger institutions will likely have numerous usernames and associated passwords that are used to access a variety of services. This is the case at the University of Glasgow! However activities are currently underway to roll out a unified user account management system based upon *directory* technology.

The directory is the part of any service which retains the authentication data, most commonly a username and a password. Until now this information at Glasgow has primarily been closely linked to specific operating systems or infrastructures. This has resulted in a myriad of solutions holding a variety of authentication information across the university. One of the consequences of this is that the evolution of services can become

tied to the platform which hosts the user identities, rather than the best platform for the job. In most cases these accounts are not necessarily the same - indeed in lots of cases they are very different, and often based on a combination of central and local accounts. Thus users are expected to keep multiple accounts and multiple passwords. Under these circumstances users tend to either leave the password at the value it was when they received it; change it to the same value as their other passwords; they have to remember multiple passwords, or they end up with passwords they can't change because changing it in one place means changing it everywhere. With multiple accounts, across multiple systems with potentially multiple different administrators coordinating changes is almost impossible. Addressing such issues is crucial for the wide scale successful deployment and take-up of Shibboleth.

The above problems are not isolated. Until recently no mechanisms existed to keep the various user accounts synchronised across all of the systems used. This arrangement meant there was a high number of redundant accounts, which has meant that it was very difficult to ensure all access and privileges were removed in a timely fashion. In some circumstances users could retain rights to data and services long after they should. This was possible since different representations for the same users could in principle lead to situations where one account could be disabled, but users could retain access to services and data via a second account. A key challenge is therefore to address the whole user base since there may be no definitive source for authentication data, but rather a collection of sources.

To overcome these issues the University of Glasgow is moving to a system that offers a more consistent representation of staff and students across multiple systems that will allow: timely creation/modification and deletion of accounts; an audit trail against central records; a single authority for services covering the whole university; password synchronisation; and the implementation of a rigorous password policy. To support this, the university is planning: a one to one representation between each user and their corresponding entry in the Human Resource/Registry database – the definitive sources for data; an agreed standard for unique identifiers for each user account; an agreed password policy; an agreed definition of department/faculty codes where user accounts should reside. This system is based upon the Novell nSure technology (www.novell.com/solutions/nsure) and is currently being rolled out by University Services across the university. Thus for federations with the University of Glasgow we hope to state with some confidence that we are able to authenticate users that are members of the university.

To fully understand the benefits of Shibboleth to seamlessly support access to a wider range of resources, both Grid-based and non-Grid based resources, we have applied it across a range of areas and projects and research areas. Details of how these have been successfully applied in the education domain are given in [20,21]. In this paper we focus in particular on how we have applied this technology in the life science domain.

3 Case Studies Applying Shibboleth

The vision of the Grid in seamlessly accessing and using a range of resources is a compelling one, but one that depends on supporting technologies. Single-sign on to resources is one of the fundamental requirements to the realisation of this vision. From the life science community perspective, single sign-on to a whole range of post-genomic resources (both computational and data resources) through to clinical and epidemiological data sets and services is needed.

Several projects at the National e-Science Centre at the University of Glasgow have been used to explore the suitability of Shibboleth for Grid security. These include the UK DTI funded Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES) [22] project and the MRC funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) [23] project. We outline these projects here and present how Shibboleth simplifies the user experience in accessing and using the Grid.

3.1 Shibboleth version of the BRIDGES Grid BLAST Service

The BRIDGES project involved the National e-Science Centre at the University of Glasgow and Edinburgh with industrial involvement from IBM. BRIDGES successfully completed at the end of 2005. BRIDGES remit was to provide a Grid infrastructure to support the Wellcome Trust funded Cardiovascular Functional Genomics (CFG) project [24] who are investigating possible genetic causes of hypertension. Hypertension is one of the main causes of cardiovascular mortality. This consortium which involved five UK sites and one Dutch site pursued a strategy combining studies on rodent models of disease (mouse and rat) contemporaneously with studies of patients and population DNA collections.

Before BRIDGES, many of the activities that the CFG scientists undertook in performing their research were done in a time consuming and largely non-automated manner. This was typified through “internet hopping” between numerous life science data sources. To address this, the BRIDGES project developed a security focused data Grid using commercial [25] and open source Grid middleware [26]. Information on the data Grid that was developed within BRIDGES is described in [27-29]. In undertaking their research the CFG scientists also required simple access to large scale HPC resources, to run compute intensive bioinformatics applications such as Basic Local Alignment Search Tool (BLAST) [30].

To simplify the user experience in accessing and using large scale HPC resources, it was decided to remove digital certificates from the end user environment and replace them with simple username and password authentication at a central project web portal. The model assumed throughout the lifetime of BRIDGES was that the end users would only have a web browser through which they would access and use the Grid resources. Given that sites such as the UK National Grid Service require a UK e-Science certificate for a given jobs (following the grid-mapfile approach described previously), the

BRIDGES BLAST service host identity was mapped locally to a project account in the local grid-mapfile on the remote Grid nodes. Thus, all jobs run under the project's identity on the NGS resources, and the logging and monitoring of user activity was maintained by the BRIDGES support staff.

BRIDGES supported a fine grained security infrastructure based upon the Privilege and Role Management Infrastructure Validation Standard (PERMIS) (www.permis.org), whereby distinctions between different privileged and non-privileged users were defined and subsequently enforced. These were:

- If they are *unknown* users the job will only be submitted to the local "free" Condor pool;
- If we recognise the users but they do not have a local account on HPC resources at Glasgow, the job will be submitted to the Condor pool and NGS;
- If we recognise the users and they have an account Glasgow HPC resources then the job will be to the Condor pool, the NGS and to ScotGrid.

In all of these scenarios, the selection of where to submit jobs was based on availability of resources (which was established dynamically).

The Shibboleth version of the Grid BLAST service did not require users to log in to the project portal. Instead when users directed their browsers to the project portal for the first time, i.e. when no security context had been established, they were automatically redirected to the WAYF service as depicted in Figure 2. They would then select their own institution from those that were listed. In this case a user might select the University of Glasgow for example as shown.

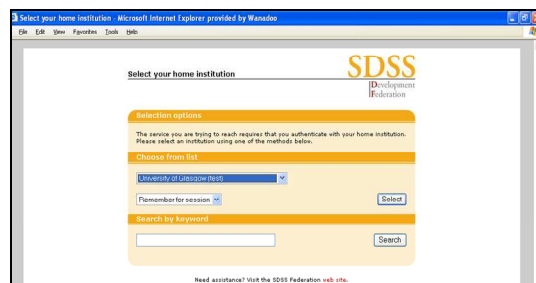


Figure 2: WAYF Used to Select a Home Identity Provider

Once redirected to their home IdP, which at Glasgow was a local LDAP server, users would log-in with their own usernames and passwords. Once authenticated, the attributes that were returned to the SP were used to enforce subsequent authorisation decisions by the service. We note that whilst BRIDGES VO specific attributes could be defined and returned, the DN of the user from the IdP was sufficient to allow an authorisation decision to be made. Once successfully signed in the front end to the Grid BLAST service is accessible as shown on the left of in Figure 3. This provides access to a range of genomic

and microbial data resources which can be BLAST'ed against. The service supports protein and nucleotide sequence searches and allows upload of input sequences or cut/paste of sequences in FASTA format. To support large scale BLAST'ing users can select options to be emailed the results when the jobs are completed, or they can interactively see the status of the jobs across the various Grid resources (whether they are queued, completed, running) shown in the right of Figure 3.

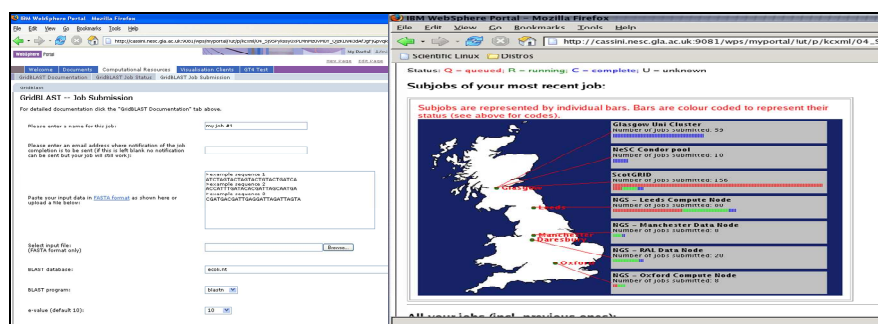


Figure 3: Grid BLAST front end

We note that this model of applying Shibboleth where the user identity (given by the DN) is returned and subsequently used to make authorisation decisions, raises issues in the application of Shibboleth. Shibboleth has been developed to support user anonymisation and privacy in accessing and using resources across a federation. However, with the Grid model, knowing which user is accessing a resource, especially in the biomedical domain is crucial. We also note that whilst Shibboleth supports user anonymisation and privacy it is not mandatory and free text strings containing information such as the DN of the user from an IdP to an SP can be returned. The policies on what information and attributes an SP can ask for and what information an IdP is prepared to release will form part of the overall federation contract. There is no obligation on an IdP releasing potentially sensitive information about a given user. However if an SP requests certain attributes to be returned for example which the IdP refuses to release then the SP is completely free to refuse to grant access to their own resource. SP autonomy is thus assured.

As stated, the primary benefit of the Shibboleth enabled version of the BRIDGES Grid BLAST service is that the user no longer needs to remember a username and password for a given portal. Instead they only need to know their username and password at their home site. Whilst only needing to know a single username and password is a key benefit in applying Shibboleth, the true benefits arise when the user wishes to access a multitude of different resources across many sites, as is typically the case in supporting systems-biology based research from the genotype to the phenotype and population studies, i.e. supporting single sign-on across a range of resources and sites. Whilst BRIDGES was

targeted to the genetic end of the spectrum, the VOTES project was targeted towards the other extreme and was focused on clinical trials and epidemiological studies.

3.2 Shibboleth version of the VOTES Data Federation Framework

The VOTES project was funded by the MRC for 3-years and began in October 2005. The project involves the universities of Glasgow, Oxford, Imperial College London, Nottingham, Leicester and Manchester. The overall goal of VOTES is to develop a Grid framework through which a multitude of clinical trials and epidemiological studies can be supported. Thus rather than engineering bespoke solutions for a given trial or study, VOTES intends to provide an infrastructure where a multitude of trials and studies can be developed and supported, each with their own particular nuances in terms of the data that is being accessed, the security policies that apply etc.

At the heart of clinical trials are three key processes: patient recruitment; data collection and study management. Recruitment is primarily concerned with identifying the patients that are potentially suitable for a clinical trial. Once identified it is essential that they are advised about all matters related with the clinical trial: the benefits, the potential dangers and more widely how this information might be used in the future. One of the challenges of this is that it is essential that patient consent is obtained before any access to the patient data sets is made. Once a set of patients have been identified, invited to join the trial and accepted, the next phase is typically focused on that actual undertaking of the trial itself. This will be in collecting data on the patients throughout the course of the trial. If the trial is concerned with drug evaluation say, then it is necessary to randomize the patients with some patients being given the drugs to be evaluated and others a placebo. Tracking and monitoring the patients throughout the trial is essential to ensure that all necessary information is collected, e.g. that the patients always visited the doctor or hospital and took the drug or placebo. When this is not the case, the results need to incorporate any discrepancies between the planned and actual trial plan since this may well impact upon overall evaluation of the drugs.

Throughout the recruitment and data collection process it is essential that the study or trial is conducted according to a strictly defined protocol. This will focus on what information is being collected, for what purpose and how it will subsequently be used both within and following the trial. A key element of this is ensuring that the different people with different roles within the trial can only access and use the different data sets and services associated with their particular role in the trial.

The Grid infrastructure developed within VOTES has been described in [31-33]. In brief, the VOTES infrastructure is based upon a GridSphere portal (www.gridsphere.org) which provides access to a range of Globus toolkit version 4 services, which themselves access and use a data federation framework utilising OGSA-DAI technology. The current implementation combines the access to and usage of a range of software and data sets in widespread use across the NHS in Scotland.

The project has defined and implemented a fine grained authorisation infrastructure based upon an access matrix. In this model, very fine grained security is achievable that

allows a clinical trials co-ordinator to define access to and usage of individual tables, rows and columns across the range of distributed clinical data sets and attach these with trials specific roles.

At the time of writing, the infrastructure supports a range of clinical trials to allow exploration of the problem space. Each of these trials has associated specific roles that have been defined. These include *investigator*, *consultant* and *nurse*. We note that whilst there are over 180 different roles in the NHS in the Greater Glasgow region alone, clinical trials will typically not require such a detailed range of roles and privileges.

The Shibboleth version of this infrastructure follows a similar access and usage pattern as described above in section 3.1 for BRIDGES. Namely, that the user attempts to access the VOTES portal and is initially redirected to the WAYF service where they select their home IdP. We note that when a user has signed in already, e.g. they have authenticated themselves to access the BRIDGES portal, provided the user is using the same browser, they will *in principle*, automatically be allowed access to the VOTES portal. Here the term *in principle* depends on whether the attributes necessary for access to the VOTES portal were already released to the BRIDGES portal. In Figure 4 the portlet is shown which details the attributes that are returned from the IdP. As can be seen one of the roles is *investigator* (the other roles are used for a different demonstration in the education domain). The common name (CN) is also returned (CN=Richard Sinnott) along with information on the Shibboleth origin (another name for the IdP).

The screenshot shows a web browser window titled "GridSphere Portal - Microsoft Internet Explorer, provided by Wanadoo". The main content area displays "A set of Attributes Portlet" with the heading "The Attributes got from Shibboleth AA are:". Below this, a table lists various attributes and their values. The "User from:" field shows "UNIVERSITY of GLASGOW" with the university's crest. Other attributes include "accept-encoding" (gzip, deflate), "title" (glasgow), "permissions" (studentteam1;studentteam2;investigator), "connection" (Keep-Alive), "Shib-Authentication-Method" (urn:oasis:names:tc:SAML:1.0:om:unspecified), "Shib-Application-ID" (default), "cookie" (a long alphanumeric string), "content-length" (0), "Shib-Origin-Site" (urn:mace:ac.uk:sdss.ac.uk:provider:identity:titania.nesc.gla.ac.uk), "accept-language" (en-gb), "host" (labpc-2.nesc.gla.ac.uk), and "user-agent" (Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)).

Attribute Name	Value
User from:	UNIVERSITY of GLASGOW
accept-encoding	gzip, deflate
title	glasgow
permissions	studentteam1;studentteam2;investigator
connection	Keep-Alive
Shib-Authentication-Method	urn:oasis:names:tc:SAML:1.0:om:unspecified
Shib-Application-ID	default
cookie	...jmmx=232251656.115306093.1130307861.1153740821.1154431509.18; MOODLEID_%25B3%259C%2510C%25A9%25BD%250F%25F0; shibsession_default=f06799ac8c08caca9ef62c90eb01be27
content-length	0
Shib-Origin-Site	urn:mace:ac.uk:sdss.ac.uk:provider:identity:titania.nesc.gla.ac.uk
accept-language	en-gb
host	labpc-2.nesc.gla.ac.uk
user-agent	Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322)

Figure 4: Authenticated User Attributes Delivered via Shibboleth to the VOTES Portal

Based upon this authentication information and the attributes returned, the portal will ensure that the authenticated user is restricted to seeing the trials and associated data associated with their particular role (*privilege*). Figure 5 shows the impact of the different roles and attributes that are returned from the IdP. On the left an authenticated user with a more privileged role (*investigator*) is able to issue richer queries across a range of resources. This includes allowing the user with this particular role to see identifying data, i.e. the patient name and address. An authenticated user with a nurse on the other hand is only able to issue queries returning non-identifying data.

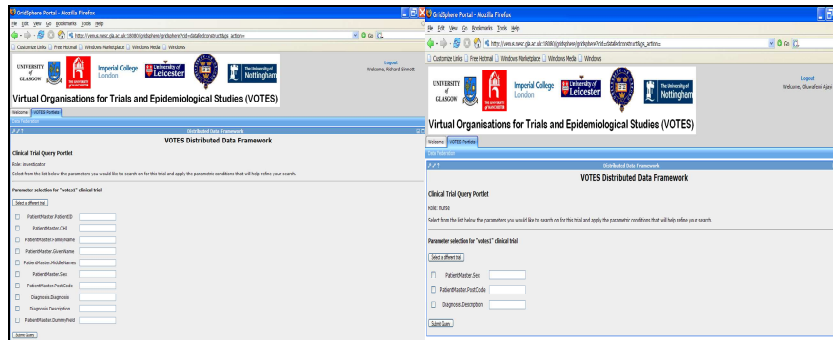


Figure 5: VOTES Data Framework targeted to Different Roles

4 Conclusions and Acknowledgements

The life scientists interesting in tackling the big questions like how does a brain work, why do people who eat less tend to live longer, what genes cause cancer, etc cannot work in isolation. They need to access and use a wide range of information much of which may well have very fine grained security associated with it. The Grid provides an infrastructure whereby heterogeneous, highly distributed data sets can be accessed and integrated. A simple user oriented model of security is essential that should be cognisant of existing security infrastructures and policies, e.g. how local authentication is supported. Shibboleth provides for many direct advantages in this domain: it supports single sign-on to a widespread range of resources across a federation; it recognises and respects local security infrastructures policies and procedures; it supports users only having to know their own local institutional usernames and passwords; it allows for very fine grained authorisation infrastructures to be supported based on returned attribute sets.

Shibboleth on its own offers a largely static mechanism whereby the roles and attributes needed to access resources have to be defined and agreed in advance. Whilst for the non-Grid community where the only authorisation information needed is often of the form, is a member of the University of Glasgow say, this model does not support the more dynamic Grid environment where new virtual organisations need to be created dynamically for evolving Grid environments requiring richer attribute sets are needed specific to the different roles across a given Grid project say. The DyVOSE project [34] has implemented a delegation issuing service [35] which supports precisely this kind of scenario. Our next plans are to consider how this service can be applied in this domain for dynamic attribute creation and recognition across clinical domain boundaries.

4.1. Acknowledgements

This work was supported by grants from the UK Department of Trade and Industry, from the Joint Information Systems Committee and from the Medical Research Council. We gratefully acknowledge their support.

References

- [1] R.O. Sinnott, M. Bayer, *Controlling the Chaos: Developing Post-Genomic Grid Infrastructures*, Life Science Grid Conference (LSGrid2005), May 2005, Singapore.
- [2] R.O. Sinnott, P. Lord, A. MacDonald, Large-scale data sharing in the life sciences: Data standards, incentives, barriers and funding models (The “Joint Data Standards Study”), prepared for The Biotechnology and Biological Sciences Research Council, The Department of Trade and Industry, The Joint Information Systems Committee for Support for Research, The Medical Research Council, The Natural Environment Research Council and The Wellcome Trust.
- [3] B. Brunch, *The History of Science and Technology*, Houghton Mifflin Books, ISBN0618221239, 2004.
- [4] R.O. Sinnott, Grid Security: Middleware, Practices and Outlook, prepared for The Joint Information Systems Committee for Support for Research, November 2005.
- [5] Internet2 Shibboleth technology, shibboleth.internet2.edu
- [6] Shibboleth Architecture Technical Overview, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf>
- [7] Shibboleth Architecture Protocols and Profiles, <http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf>
- [8] ITU-T Recommendation X.509 (2001) | ISO/IEC 9594-8: 2001, Information technology – Open Systems Interconnection – Public-Key and Attribute Certificate Frameworks.
- [9] R. Housley, T. Polk, Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures, Wiley Computer Publishing, 2001.
- [10] W. T. Polk and N. E. Hastings, Bridge Certification Authorities: Connecting B2B Public Key Infrastructures, <http://csrc.nist.gov/pki/documents/B2B-article.doc>
- [11] J. Jokl, J. Basney and M. Humphrey, Experiences using Bridge CAs for Grids, Proceedings of UK Workshop on Grid Security Practice - Oxford, July 2004.
- [12] Globus, www.globus.org/toolkit
- [13] UK Federation, www.sdss.ac.uk
- [14] SWISS SWITCHaai federation, <http://www.switch.ch/aai/>
- [15] Finnish HAKA federation, <http://www.csc.fi/suomi/funet/middleware/english/>
- [16] Australian Meta Access Management System (MAMS), <https://mams.melcoe.mq.edu.au/zope/mams/kb/shibboleth/>
- [17] US InCommon federation, <http://www.incommonfederation.org>
- [18] Security Assertion Markup Language (SAML) version 2.0, March 2005, <http://www.oasis-open.org/specs/index.php#samlv2.0>

- [19] A. Robiette, T. Morrow, Blueprint for a JISC Production Federation, JISC Development Group, Version 1.1: issued 27 May 2005, http://www.jisc.ac.uk/index.cfm?name=middleware_documents
- [20] J. Watt, R.O. Sinnott, O. Ajayi, J. Jiang, J. Koetsier, A Shibboleth-Protected Privilege Management Infrastructure for e-Science Education, 6th IEEE International Symposium on Cluster Computing and the Grid, CCGrid2006, May 2006, Singapore.
- [21] 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.
- [22] Biomedical Research Informatics Delivered by Grid Enabled Service (BRIDGES) project, www.nesc.ac.uk/hub/projects/bridges
- [23] Virtual Organisations for Trials and Epidemiological Studies (VOTES) project, www.nesc.ac.uk/hub/projects/votes
- [24] Cardiovascular Functional Genomics (CFG) project, www.brc.dcs.gla.ac.uk/projects/cfg.
- [25] IBM Information Integrator, <http://www-306.ibm.com/software/data/DL>
- [26] Open Grid Service Architecture – Data Access and Integration Two (OGSA-DAIT), www.ogsadai.org
- [27] 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.
- [28] R.O. Sinnott, D. Houghton, Comparison of Data Access and Integration Technologies in the Life Science Domain, Proceedings of UK e-Science All Hands Meeting, September 2005, Nottingham, England.
- [29] 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.
- [30] R.O. Sinnott, M. Bayer, Distributed BLAST in a Grid Computing Context, Proceedings of First International Workshop on Distributed Data Mining in Life Science, Konstanz, Germany, September 2005.
- [31] A.J. Stell, R.O. Sinnott, O. Ajayi, Secure Federated Data Retrieval in Clinical Trials, Telemedicine 2006 conference, Banff, Canada, July 2006.
- [32] R.O. Sinnott, A.J. Stell, O. Ajayi, Development of Grid Frameworks for Clinical Trials and Epidemiological Studies, HealthGrid 2006, Valencia, Spain, June 2006.
- [33] A.J. Stell, R.O. Sinnott, O. Ajayi, Supporting the Clinical Trial Recruitment Process through the Grid, Nottingham UK e-Science All Hands Meeting, September 2006.
- [34] Dynamic Virtual Organisations in e-Science Education project (DyVOSE), www.nesc.ac.uk/hub/projects/dyvose
- [35] 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, submitted to 2nd IEEE International Conference on e-Science and Grid Computing, Amsterdam, December 2006.