FP7-212214

Title	**Final Report and Recommendations (D11.1a)**
Work Package	WP11
Authors	Professor R. Sinnott (National e-Science Centre, Glasgow, UK)
Dissemination Level	PU (Public)

Summary/abstract

Part A of the tendered report entitled “Possibilities and Implications of Grid-enabling Social Science and Humanities Data Collections in the Context of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure”.

Possibilities and Implications of Grid-enabling Social Science and Humanities Data Collections in the Context of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure

Report Drafted as part of the CESSDA Preparatory Phase Project (CESSDA PPP)

Professor Richard O. Sinnott
National e-Science Centre
University of Glasgow
Kelvin Building
Glasgow, G12 8QQ
United Kingdom

Tel: 0141 330 8606
Fax: 0141 330 8625
Email: r.sinnott@nesc.gla.ac.uk

Document Control

Details

Document Title	Possibilities and Implications of Grid-enabling Social Science and Humanities Data Collections in the Context of the CESSDA RI
Funded By	University of Essex, UK Data Archive
Circulation	CESSDA PPP
Author	Prof. Richard O. Sinnott
Version	1.0
Date	28th February 2009
Authorised By	Hilary Beedham, Ken Miller, Kevin Schurer

Document History

Version	Editor	Date	Description
1.0	Richard Sinnott	28th February 2009	Version 1

Comments and feedback from CESSDA PPP members are welcomed on this draft including any areas where more information or clarification might be useful.

This work has been funded by the UK Data Archive, University of Essex, December 2008.

Contents

Table of Figures

Executive Summary

The central task of this report is to advise on the theoretical and practical consequences of Grid-enabling social science and humanities resources existing in the CESSDA Research Infrastructure (RI). The CESSDA RI is planning a major upgrade in order to ensure that European social science and humanities researchers (SSH) have access to, and gain support for, data resources they require to conduct research of the highest quality, irrespective of the location of either researcher or data within the European Research Area. In addressing these concerns, the planned upgrade will develop CESSDA from the current situation in which the member organisations work with limited national resources, to create a common platform, sharing a common mission, with a stronger form of integration in which expertise is genuinely pooled, shared and applied in a co-ordinated pan-European experience. This will facilitate the delivery of a fully-integrated data archive infrastructure for the SSH, allowing seamless, permanent access to as many data holdings across Europe as possible.

The Grid vision and its aims to support seamless access to distributed resources through establishment and support of e-Infrastructures (also referred to cyber-infrastructures) offers, at least in principle, a paradigm that meets many of the objectives of CESSDA RI. This report introduces the basic principles of Grids and Grid-based e-Infrastructures and the capabilities they provide. It outlines a variety of Grid initiatives and existing e-Infrastructures and technologies, highlighting the advantages and limitations with regard to the current and importantly the future development of the CESSDA RI and the research environment in which it might exist. The report also describes Grid standards and middleware software solutions that are of direct relevance to the CESSDA RI focusing in particular upon areas related to security, data/meta-data management and the way in which user-oriented research infrastructures can be delivered to potentially non-Grid savvy communities.

The report is based upon experiences gained in development of numerous Grid-based e-Infrastructures at the National e-Science Centre (NeSC – www.nesc.ac.uk) at the University of Glasgow to support a multitude of researchers in different research domains exploiting a wide range of Grid middleware. These research domains include the clinical sciences, biological sciences, geospatial sciences, the electronics domain and the social sciences amongst others. The particular research focus of NeSC is in security-oriented application domains. We emphasise that the NeSC in itself does not produce its own Grid middleware and can thus be regarded as impartial in this regard.

The rest of this report is structured as follows. Section 1 begins with an overview and background information on Grids and their application to develop and support e-Infrastructures. We identify the key components that a Grid-based e-Infrastructure should support in order to be classified as a Grid infrastructure (as opposed to more general internet-based infrastructure). We outline major international efforts in the Grid/e-Infrastructure-space that could have an impact upon the future CESSDA RI. Section 2 focuses upon Grid-related standards and technologies that could be important to the future CESSDA RI. In particular we focus upon standards and technologies that have been applied to establish and maintain Grid-based e-Infrastructures dealing with secure access to distributed data resources, security and portals and related delivery mechanisms. Where necessary we highlight examples of these solutions based upon case studies/projects undertaken at NeSC Glasgow. We also briefly cover relevant technologies and efforts including Web Services and service-oriented architecture based solutions more generally, Web 2.0 based solutions and the more recent move to Cloud Computing and what this might mean for the future CESSDA RI. Section 3 focuses upon key use cases that a future CESSDA RI should support (as outlined in the initial tender document) and we outline how the Grid standards and technologies outlined in section 2 could be applied to support these use cases. We focus in particular upon how virtual organisations can be established for CESSDA RI and the way in which a one-stop portal based solution could be supported. Given that CESSDA wishes to explore secure data enclaves where “sensitive” data can be accessed and used without potentially ever being seen or disclosed, we outline potential solutions to this that have been produced at NeSC in Glasgow. Finally section 4 draws some conclusions on the Grid and its impact upon CESSDA RI.

This report will be augmented with an additional extension covering the resourcing and sustainability issues relating to implementation of a future Grid-based CESSDA e-Infrastructure.

1 Introduction to Grids and e-Infrastructures

Fundamentally Grids are used to support sharing of resources to support research communities. This is typically achieved through development and support of e-Infrastructures often referred to as cyber-infrastructures[1]. Precisely what resources are shared is often not specified or restricted and numerous flavours of Grids and e-Infrastructures exist. These can include shared access to and use of High Performance Computing (HPC) facilities, data repositories, data archives, visualisation facilities, sensor networks, software or indeed specialised resources such as electron microscopes or astronomic telescopes amongst many other possibilities. Grids and e-Infrastructures have been explored in many research domains from the physical sciences, the clinical sciences, through to the humanities and social sciences. Indeed there are few research areas where Grid-based technologies have not been applied in some manner. Given this, it is fair to say that there are a multitude of interpretations of Grids and e-Infrastructures that are used (or have been used): Compute Grids, Data Grids, Information Grids, Campus Grids, Enterprise Grids, Semantic Grids, Knowledge Grids are just some of the terms that are used to describe the various forms of collaborative-driven e-Infrastructures each with different capabilities, exploiting different technologies and supporting or used by a range of user communities. This in turn has resulted in a huge range of middleware and software systems that have been developed to support the different flavours of Grids and e-Infrastructures.

For the vast majority of people outwith the Grid space, the most commonly understood and deployed Grid kinds are Compute Grids. Compute Grids tend to focus upon HPC-oriented computationally-bounded application domains where Grid-technologies are applied to support e-Infrastructures that allow (or should allow) seamless access to distributed, heterogeneous HPC facilities. Such Grids are primarily used to address research challenges that can be tackled by being able to run larger scale simulations. There are numerous examples of such Grids. The UK e-Science National Grid Service (www.ngs.ac.uk) is a typical example of such a Compute Grid. The Enabling Grids for e-Science (EGEE - www.eu-egee.org) is another European-wide Compute Grid example. This is not to say that a Compute Grid does not have to deal with or manage data. They do, but often the data management is often left to the end user scientists/researchers or the solutions that are put forward are targeted at demands from specific research communities (such as high energy particle physicists in the case of EGEE) which don't meet the requirements from other domains. There are several reasons for this. The most important one is that data management is often (indeed nearly always!) domain specific. Whilst resource providers such as the UK e-Science NGS can be used to run a variety of simulation codes from a variety of research disciplines, it is much harder to manage data sets from those different disciplines since it requires much more domain knowledge. For Compute Grids that are used by biologists, physicists, chemists, and geographers etc the multi- and inter-disciplinary domain knowledge required to support domain-specific data management does not exist. Instead, many research communities use Compute Grids to run simulations and subsequently undertake their own data management. Typically this means pulling results of simulations from HPC facilities and storing them locally. Alternatively, Compute Grid providers will offer generic tools for data management, but these do not address many of the key challenges facing research communities: how to find, access, use, share and annotate data and/or meta-data, or deal with the issues associated with data quality, data security, or longer-term data management challenges such as data provenance and data curation.

Instead, for many domains (including the CESSDA), it is precisely these kinds of data challenges that need to be addressed. Data Grids offer one approach that allows many of the issues associated with domain-specific data challenges to be addressed. In this report, our focus is primarily upon Data Grids and the offerings that the Grid/e-Infrastructure community have that can/should impact upon the CESSDA RI.

The technological landscape associated with Grids and e-Infrastructures is especially complex however with numerous standards, software solutions and approaches that exist. For example, many approaches have adopted approaches based upon various flavours of web services (WS) to support service-oriented architectures. Various flavours of Grid services and related technologies have evolved

[1] In the rest of this document we refer to e-Infrastructures only.

from different communities, often resulting in complex Grid software stacks. More recently, other communities have adopted lighter-weight solutions based upon Web2.0 technologies.

To provide a better understanding of the technological landscape for CESSDA RI and to support one of the basic tenets of the Grid, we believe that any infrastructure that is ultimately supported in the CESSDA RI must support *single sign-on* to distributed, heterogeneous and ultimately autonomous resources. That is, depending upon the privileges that a researcher possesses, they should be able to access and use a rich variety of social science data sets and tools without having to authenticate themselves repeatedly at each remote resource provider. Instead, once authenticated once they can simply roam and access resources offered by numerous providers with no further username/password challenge responses for example provided their privileges allow.

Furthermore given the sensitive nature of some of the SSH resources that exist across CESSDA sites, it is essential that security is ensured across CESSDA as a whole. Thus it is the case in computer security that the weakest link rule applies; this fact is often magnified by Grid infrastructures and Compute Grids in particular due to their openness. Highly secure multi-million pound compute facilities can be compromised by inadequately secured remote laptops. Rigorous security procedures at one site can be made redundant through inadequate procedures at another collaborating site. This problem is magnified in many Compute Grids due the lack of granularity in how security is currently considered. Grid security as typified by Compute Grids is primarily based around Public Key Infrastructures (PKIs) which support validation of the identity of a given user requesting access to a given resource – so called *authentication*. There are several key limitations with authentication based approaches to security. Most importantly, the level of granularity of security is limited. There is no mention of what the user is allowed to do once they have gained access to the resource. With Compute Grids such as the UK NGS for example, users can in principle run arbitrary applications, starting a variety of local processes. In reality, a set of existing applications and infrastructure are often pre-deployed across these resources, hence the issue and risks of uploading executables is diminished. However, given the fact that common compilers for C^{++} etc are commonly available, the possibility to upload and compile arbitrary code and run arbitrary executables spawning arbitrary processes exists. There is typically no security middleware enforcement on what processes can be started, by whom and in what context, other than the local enforcement given by the privilege associated with the local account. These issues with PKI-based authentication only models of security deter large sets of the research community from engaging with Grids and are naturally not suited to more sensitive data sets.

Instead, finer grained security models are required where the policies and decisions on what a researcher is allowed to do need to be specified and subsequently enforced by distributed and autonomous providers – so called *authorisation*. We review the state of the art in this area in the Grid domain and highlight key standards and technologies that could be applied in the context of the CESSDA RI. We also highlight how these technologies can be seamlessly linked to Grid resources. Key to this is the concept of supporting Virtual Organisations (VO). One of the key challenges to supporting Grid-based e-Infrastructures is in supporting dynamic VOs where collections of resources may be brought together for a particular time period for a collection of researchers. These resources themselves may change over time, the end user research community may change over time, and the privileges, roles and responsibilities associated with that VO may change over time. We highlight the variety of opportunities that VOs may be established and how this can impact upon the future CESSDA RI.

A fundamental property of any future CESSDA RI infrastructure is that it has to be simple to access and use from the end user researcher perspective but also from the data provider perspective, e.g. with tools that allow for definition and enforcement of local access control policies on access to and usage of local data resources. Social scientists should certainly not need to become Grid-experts or ideally be exposed to any of the underlying technologies that are used to support e-Infrastructures. Since many current Grid-based solutions or indeed accessing and using resources such as the NGS begin with end users having to acquire and subsequently manage their own X509 digital certificates, i.e. the underlying technologies are very much exposed to the researchers, alternative solutions are required. Furthermore given that many countries internationally are moving to federated nationwide access control systems based upon the Internet2 Shibboleth technologies as the model for secure access to resources, e.g. the UK Access Management Federation (www.ukfederation.org.uk) an opportunity

exists to harmonise access to Grid and non-Grid resources. We outline the practical ramifications of Shibboleth and related technologies in an international context such as a future CESSDA RI.

Ideally CESSDA should be interoperable with national and international Grid and e-Infrastructure efforts in this space. There are a multitude of Grid based systems that are deployed and used across Europe for a variety of research purposes. Amongst many examples, these include:

- Enabling Grids for E-Science (EGEE - www.eu-egee.org);
- Distributed European Infrastructure for Supercomputing Applications (DEISA – www.deisa.org);
- Enabling Grids for E-Science South East Europe (EGEE-SEE - http://www.egee-see.org/);
- UK e-Science National Grid Service (NGS – www.ngs.ac.uk);
- Deutsche-Grid Initiative (D-Grid - http://www.d-grid.de/);
- NorduGrid - http://www.nordugrid.org/;

It is also worth noting that these Grid-based e-Infrastructures have primarily adopted a Compute Grid flavour. Each of these e-Infrastructures supports some degree of interoperability, e.g. in recognising the certificates that are used by international collaborators. This interoperability does not scale to more detailed levels of interoperability however, e.g. where data sets and resources might be found on D-Grid, processed on EGEE for analysis by users on the NGS. This is not to say that such interoperability could not be engineered and delivered, but that this is not an artefact of simply using these e-Infrastructures themselves.

Perhaps the primary focal point to demonstrate interoperability of Grid middleware has been in the Open Grid Forum (OGF) effort Grid Interoperability Now (GIN - http://forge.ggf.org/sf/wiki/do/viewPage/projects.gin/wiki/HomePage). This project showed how major Grid efforts (including numerous of those bulleted above), can support a degree of interoperability including recognising certificate authorities, basic compute-oriented job submission.

In addition to these e-Infrastructures, a wide variety of projects from national and international bodies have been established to explore a wider variety of e-Science challenges. Taking the UK as an example, the UK e-Science Core Programme, funded a wide variety of projects in a wide variety of application domains of total value over £250million. It is the case that the vast majority of these projects have developed their own software solutions with little direct interoperability.

In this context it is difficult to be overly prescriptive on the CESSDA RI and which Grid middleware or e-Infrastructure it needs to be aligned with. There are multiple flavours of middleware and e-Infrastructures that have been developed. For the most part these are based upon supporting Compute Grid infrastructures.

Rather than focus upon a particular e-Infrastructure, we outline some of the core features that a CESSDA RI might be expected to support and then highlight offerings from the Grid community that could be applied in this space.

2 Grid Standards and Technologies of Relevance to the Future CESSDA RI

In this section we outline various standards and technologies that could be applied to support a future CESSDA RI. We note that much of the work on standardisation in the Grid community, especially with the alignment of Grid-based approaches with web services and service-oriented architecture based approaches has seen much effort on Grid-standards by organisation such as Open Grid Forum (OGF – www.ogf.org) now overtaken by wider initiatives and organisations such as Internet Engineering Task Force (IETF - www.ietf.org) and the Organization for the Advancement of Structured Information Standards (OASIS - www.oasis-open.org/).

2.1 Grid Data Standards and Technologies

A variety of Grid-based standards and technologies have been developed to allow seamless access to distributed data resources. Historically much effort in the Grid-domain has been in supporting scientific disciplines where the vast amount of data existed in files. A typical example of this is in projects such as the Large Hadron Collider (LHC) where petabytes of file-based data is generated from particle detectors and distributed around Europe for analysis on HPC facilities. In the last few years however a considerable momentum shift has occurred in supporting research domains where distributed data exists in heterogeneous formats on heterogeneous storage facilities including files, relational databases, XML databases amongst others. To deal with such variety, the Grid community has put forward a range of Grid standards and associated technologies to meet the needs of these other communities.

Some of the key standards that have been put forward by the Grid community in this space include:

- **Data Format Description Language** (DFDL[2]) which defines an XML-based language for describing the structure of binary and textual files and data streams so that their format, structure, and metadata can be exposed.
- **Database Access and Integration Services** which has developed standards for grid data services, focusing principally on supporting consistent access to existing, autonomously managed databases through web services.
- **Grid File System Standards** which offers standard service interface(s) and architecture of a logical file system that can be used in data grid management systems. This work leverages other efforts in this space including amongst others, the Storage Networking Industry Association (SNIA – www.snia.org) Information Lifecycle Management Initiative and similar efforts.
- **Grid Storage Management** which has specified the functionality of a standard Storage Resource Manager which provide dynamic space allocation and file management of shared storage components on the Grid.
- **GridFTP** standards which have specified improvements and extensions to the File Transfer Protocol (FTP) to allow for example, optimised, parallel data transfer and support for Grid-based authentication.
- **OGSA ByteIO** which has defined a minimal Web Service interface for providing "POSIX-like" file functionality thereby allowing services which implement the interface to be accessed in a file-like way.
- **OGSA Data Movement Interface** tackles the problems of discovering of data transport protocols available at the data's source and destination location and agreeing on one of them, and the actual invocation of the agreed data movement. This includes direct data movements and 3rd party data movements.

The most relevant of these Grid-standards to the future CESSDA RI is the standards to support Database Access and Integration (DAIS). These standards offer a common framework through which data services can be specified and accessed. Key to the idea behind this work is that through

[2] Commonly referred to as daffodil.

development of web services it should be possible to build implementations of services that can be composed in various ways.

The standards themselves include a core Web Service Data Access and Integration (WS-DAI) specification [WS-DAI] which identifies a collection of generic data interfaces that can be extended to support a variety of other data resources including relational databases, XML repositories or files. Two standards which are based upon this work include WS-DAIR [WS-DAIR] which is an extension of WS-DAI targeted specifically to support of relational data resources, and WS-DAIX [WS-DAIX] which is an extension of WS-DAI targeted specifically to support of XML-based data resources.

The primary technology that has been based upon the implementation of these standards is the Open Grid Service Architecture Data Access and Integration (OGSA-DAI) solution (www.ogsadai.org.uk). This software has been developed throughout the course of the WS-DAI and related standards work. The software has become part of numerous mainstream Grid middleware offerings including for example the Globus-based solution (www.globus.org) which is currently at version 4 (GT4), and the Open Middleware Infrastructure Initiative (OMII-UK – www.omii.ac.uk).

OGSA-DAI services are essentially web services that implement one or more of the WS-DAI specified interfaces to provide access to data resources. However, rather than simply being a uniform way in which data can be accessed, OGSA-DAI also allows for a rich variety of interaction patterns to be supported. Thus for example, it is possible to use OGSA-DAI to support data movements between multiple different services through specification and enactment of activities. In the context of CESSDA RI this might be allow a user to specify a collection of data services that they wish to access and use in a particular order and where particular processing or analysis of data occurs between the different data services – including third party file transfers. This might include data formatting manipulations or compression of data between services before final delivery of the data to the end users themselves.

These data standards and their implementations have not in themselves put forward solutions to many of the domain specific challenges that exist with data management. As one example, meta-data is not something that these standards prescribe since it is domain specific. That is not to say, that the standards and technologies cannot be used to capture meta-data associated with services and the data sets they give access to. Rather the solutions are generic and can be applied to exploit existing meta-data standards and tools such as Data Documentation Initiative (DDI - http://www.ddialliance.org/). Indeed the ESRC funded Grid Enabled Occupational Data Environment (GEODE) project (www.geode.stir.ac.uk/) used OGSA-DAI and DDI to capture data and meta-data associated with social science occupational classifications.

We also note that the non-Grid community have also focused upon the challenges of data access and integration. Numerous software products are available that allow for access to and use of remote distributed, heterogeneous data resources. These are often tied to particular domains of application. As one example, IBM Information Integrator (IBM II) has evolved to address the needs of post-genomic researchers. It allows seamless access to distributed genomic data resources. A comparison of IBM II and Grid-based offerings was conducted as part of the Department of Trade and Industry funded Biomedical Research Informatics Delivered by Grid Enabled Services (BRIDGES – www.nesc.ac.uk/hub/projects/bridges) project [ODvII]

2.2 Grid Security Standards and Technologies

2.2.1 Grid Authentication and the move to Shibboleth

A variety of Grid-based standards and technologies have been developed to allow seamless access to distributed data resources on the Grid. Many early Grid-efforts especially oriented towards Compute Grids are based upon X.509-based PKIs for primarily authentication-driven security models. A body of expertise in the establishment and management of PKIs now exist. However, PKIs also have various limitations and issues with regard to their usage in the Grid community. These issues are described in Annex 2.

To address some of the primary problems with PKI-based authentication, much effort has been focused upon the Internet2 Shibboleth technologies (http://shibboleth.internet2.edu) and their use to support federated access control. Figure 1 typifies a typical scenario in applying Shibboleth.

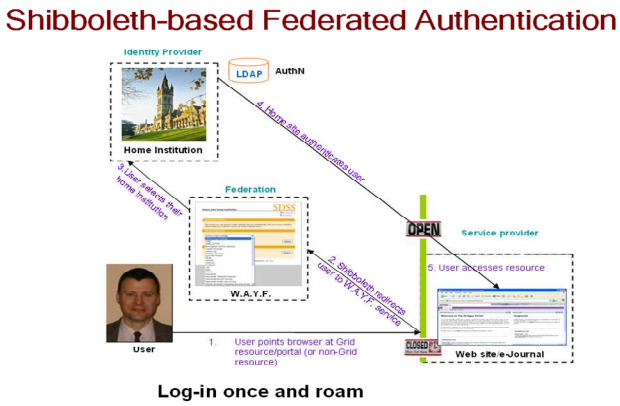

Fig 1. Typical Scenario of Shibboleth Usage

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 exists as part of the federation that asks the user to pick their home Identity Provider (IdP) from a list of known and trusted sites. The service provider site already has a pre-established trust relationship with each home site, and trusts the home site to authenticate its users properly.

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 Security Assertion Markup Language (SAML) authentication assertion message from the home site, asserting that the user has been successfully authenticated (or not!) by a particular means. The actual authentication mechanism used is specific to the IdP.

If 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 trusted message providing it with 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. The Shibboleth trust model is 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.

We note that later versions of the Shibboleth specification have introduced a performance improvement over the earlier versions, by allowing the initial digitally signed SAML message to contain the user's attributes as well as the authentication assertion. Thus the two stages of authentication and attribute retrieval can be combined.

The connection from the IdP to the service provider can also be optionally protected by SSL in Shibboleth. Here SSL is used to provide confidentiality of the connection rather than message origin authentication. In many cases a confidential SSL connection between the IdP and SP will not be required, since the handle can be opaque/obscure enough to stop an intruder from finding anything out about the user, whilst the SAML signature makes the message exchange authentic. However the message exchange should be protected by SSL if confidentiality/privacy of the returned attributes is required. The attributes in this assertion may then be used to authorise the user to access particular areas of the resource site, without the service provider ever being told the user's identity. Shibboleth has two mechanisms to ensure user privacy. Firstly it allows a different pseudonym for the user's identity (the handle) to be returned each time, and secondly it requires that the attribute authorities provide some form of control over the release of user attributes to resource sites, which they term an attribute release policy. Both users and administrators should have some say over the contents of their attribute release policies.

Once authenticated through Shibboleth, the notion of single sign-on is supported whereby a user may redirect their browser to other protected Shibboleth resources with no need for re-authentication.

Underlying Shibboleth-based SAML token exchanges are a core set of eduPerson attributes (www.educause.edu/eduperson/) that are pre-agreed across the federation so that an SP can make its own local access control decision. It is essential that interoperability exists between attribute authorities issuing attribute assertions, policy writers defining access policies, and access decision functions that make decisions based on the initiator's attributes and sites target and resource policy.

A small subset of *eduPerson* attributes has been recognised as providing the necessary core functionality for IdPs and SPs in the UK academic community. These are:

- *eduPersonScopedAffiliation*: which indicates the user's relationship (e.g., staff, student, etc.) with the institution.
- *eduPersonTargetedID*: is needed when an SP is presented with an anonymous assertion only, as provided by *eduPersonScopedAffiliation*. In this situation it cannot for example provide usage monitoring across sessions. The *eduPersonTargetedID* attribute provides a persistent user pseudonym.
- *eduPersonPrincipalName*: is used where a persistent user identifier, consistent across different services, is needed.
- *eduPersonEntitlement*: 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.

Each of these attributes can be used to provide the necessary information to SPs to make authorisation decisions. These attributes are versatile and likely to be sufficient for the great majority of applications. It would be expected that the CESSDA RI would also adopt a small set of agreed attributes for federated authentication.

However given the fact that Grids can be used to establish e-Infrastructures and more security-oriented VOs, the requirement to have VO specific attributes defined and embedded in core eduPerson attributes are highly desirable. The most likely attribute for this purpose is the *eduPersonEntitlement* attribute. The *eduPersonEntitlement* attribute can utilise structured XML data representative of large scale Grid infrastructure users and IdPs. This might include the VO they are involved in, the roles that they might have in that VO etc.

Examples of Shibboleth-based federations are InCommon (http://www.incommonfederation.org), the federation formed by the Internet2 community in the United States, InQueue (http://inqueue.internet2.edu/) for sites wishing to test and explore the Shibboleth federated trust model, the SWITCHaai federation of the higher education system in Switzerland (http://www.switch.ch/aai/), the HAKA federation developed by the Finnish universities and polytechnics (http://www.csc.fi/suomi/funet/middleware/english/) with more in the pipeline such as the Meta Access Management System (MAMS) in Australia (https://mams.melcoe.mq.edu.au/zope/mams/kb/shibboleth/) and the UK Access Management Federation (http://www.ukfederation.org.uk).

Whilst Shibboleth offers numerous possibilities and potential advantages in the context of the Grid and indeed for the CESSDA RI, it is not without potential drawbacks (or at least ramifications that need to be understood). Single sign-on via authentication at a home site and subsequent acceptance and recognition of the authentication and associated attributes released to remote sites is the most obvious advantage. Thus users need not remember X.509 certificate passwords but require only their own institutional usernames and passwords. Institutions can establish their own trust federations and agree and define their own policies on attribute release, and importantly SPs can decide upon what attributes and attribute values are needed for authorisation decisions.

The uptake and adoption of Shibboleth technologies within the CESSDA RI context is not without potential concerns however. 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 may well have numerous usernames and associated passwords that are used to access a variety of services. Until 2006, this was the case at the University of Glasgow. However through roll-out of a centralised active directory based solution,

these issues have now been resolved. This system provides a one to one representation between each user and their corresponding entry in the Human Resource/Registry database – the definitive sources for data. There is an agreed standard for unique identifiers for each user account and an agreed password policy. Thus when a student or staff member leaves the university then they and all of the user accounts and passwords that they had are removed.

With this system, sites collaborating with Glasgow University can be assured that when Glasgow authenticates and releases attributes for a particular individual, then they are actual current members of the university, and not authenticated on some older and overlooked username and password. To make Shibboleth a success, all sites should ideally follow similar practices. Time will tell if this is the case.

The international nature of the CESSDA RI directly impacts upon the federation model that is adopted. Underlying Shibboleth is an associated PKI. That is, each server that is used as an IdP or SP is issued with an X509 certificate that is used for signing and recognising credentials in making access control decisions as part of the access management federation. In the UK, the access management federation is responsible for issuing these certificates. Recognising the credentials and the associated issuing authorities across the CESSDA RI would be a key requirement for exploitation of Shibboleth in this context. There are many ways in which this could be achieved depending upon the nature of the PKI and international collaborations themselves. Bridging between certification authorities is one mechanism that has been supported [B-PKI]. Defining hierarchical PKIs with the root of authority coupled with the CESSDA RI itself is another, although this would directly impact upon the exploitation of existing federations.

One of the key issues with Shibboleth that have still to be resolved for the more security-oriented Grid community is related to attribute release policy. At present an SP will request the attributes associated with the potentially opaque identifier (handle) that is returned from an IdP. If a user from the University of Glasgow is involved in numerous Grid projects and VOs however, and all of this information on what VOs this person is involved in, and what their role is in that VO etc are encoded in the core set of attributes, then it is difficult to restrict the information being released. Thus the *eduPersonEntitlement* attribute might encode much of the information on VO membership and roles etc. If an SP requests the attributes for a given user, and receives this *eduPersonEntitlement* attribute then they will receive more information than they might actually need to make an authorisation decision, e.g. if this SP was just one of the many VOs that the user was involved in, then this SP would know more about all VOs the user was involved in. Of course these attributes will be encoded, however, the SP will be able to decode the attributes due to the trust relationships and certificates previously put in place.

It is of course possible to have a richer array of attributes other than the core set of eduPerson attributes identified previously, but for interoperability and simplicity, having a core set is beneficial. Given that the focus of much of the Grid community as being represented by Compute Grid efforts which do not focus upon privacy or confidentiality, such issues are not immediately important. For more security focused domains however such as CESSDA RI, attribute release policies will become more important and only those attributes absolutely needed, should be released.

Another potential solution to this situation is to have a proliferation of IdPs. Thus each individual virtual organisation might have their own IdP and be associated with different WAYF servers. This would allow for those sets of attributes to be released deemed necessary for particular SPs, however the more IdPs that exist requires more trust relationships to be put into place, thereby weakening the overall security. Having multiple WAYF services and IdPs and SPs being involved in more than one trust federation also brings with it potential difficulties. Do we trust all federations equally? Do some treat authentication and identity management more stringently? If there are differences between the assurance levels, then multiple memberships will be problematic.

Issues also arise when dealing with institutions that do not themselves have their own IdP and are not part of a national federation. To address this, it is possible to establish a virtual home for individuals at specific IdPs, i.e. establish or use an IdP at a remote location which is part of the federation, however this model has several drawbacks. One the most obvious drawback is that authentication assertions are made to individuals not at those institutions. For specific virtual organisations where collaborators are known, this model does offer a practical solution.

2.2.2 Grid Authorisation

Given that the CESSDA RI will likely comprise access to sensitive materials, fine-grained security is required which goes beyond Shibboleth-based authentication. Thus, knowing that someone has authenticated at the University of Essex is unlikely to be sufficient information for a CESSDA RI data provider to allow or deny access to that sensitive data sets. Instead, authorisation capabilities are required. We briefly cover the background to authorisation and highlight key technologies in this space that might impact upon the CESSDA RI. We note that this review is not exhaustive and numerous other technical offerings exist in the Grid and web services space (we highlight some of the web service offerings in Annex 3). There are also many technologies and standards in this space covering different approaches to authorisation such as role based access control, identity based access control and process based access control amongst others. The technologies outlined represent the leading solutions that have been successfully applied in various projects at NeSC in Glasgow and could address many of the security authorisation challenges facing the CESSDA RI.

Authorisation is closely linked to authentication. Once a user has had their identity validated at a remote resource, it is essential that users actions are restricted based on who they are, what they are trying to do, and in what context etc. There are various methods of enforcing this restriction, the simplest method being the use of an Access Control List (ACL), which lists what users have access to a privilege. Essentially, a user presents their credentials at the gatekeeper to a resource, which consults a list of users. This basic authorisation structure extends the concept of authentication and no more. If the user cannot authenticate to the satisfaction of the gatekeeper then the resource request will be denied. A problem that arises when trying to apply this method to a dynamic Grid environment is that only one list exists, where there could be many privileges that require different ACLs. For example, a user might need access to a given resource for different purposes within a given VO. Having a single list with a predefined set of accounts and user (X.509) Distinguished Names (DN) does not support this multi-role approach. This is a solution that would not scale well in a large VO. A more sophisticated method of applying authorisation controls is through use of Role-Based Access Control (RBAC) mechanisms, which allow Privilege Management Infrastructures (PMI).

The relationship between a PMI and authorisation is similar to the relationship between a PKI and authentication. Consequently, there are many similar concepts in the two types of infrastructure. Central to a PMI is the idea of the attribute certificate (AC), which maintains a binding between the user and their privilege attributes. It is similar in notion to the public key certificate in a PKI. The entity that signs a public key certificate is a CA; the entity that signs attribute certificates is called an Attribute Authority (AA). The root of trust of a PKI is often called the root CA, which can delegate this trust to a subordinate CA; the root of trust of a PMI is called the Source of Authority (SOA). The SOA may have subordinate authorities to which it can delegate powers of authorisation. Certificate Revocation Lists (CRLs), which show a list of certificates that should no longer be accepted as valid, exist in a PKI; Attribute Certificate Revocation Lists (ACRLs) exist in a PMI.

The critical idea in a PMI is that the access rights of a user are not held in an ACL but in the privilege attributes of the ACs that are issued to the users. This is the central idea behind RBAC – the privilege attribute will describe one or more of the user's rights and the target resource will then read a user's AC to see if they are allowed to perform the action being requested. This de-couples the user's privileges from their local identity and allows a more dynamic and flexible approach to access control.

The X.812 | ISO 10181-3 Access Control Framework standard [X812] defines a generic framework to support this type of authorisation, depicted in Figure 1.

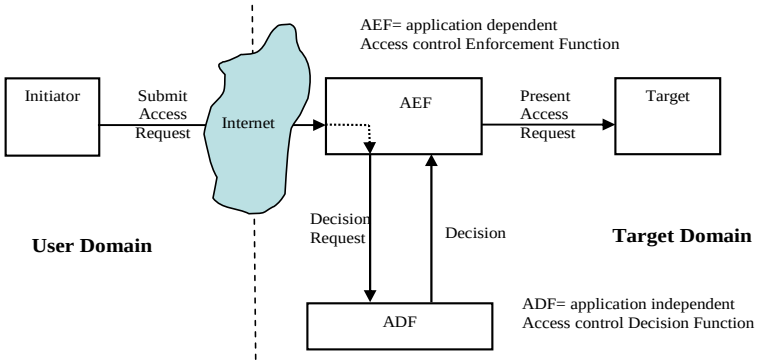

Fig 2. X.812 Access Control Framework

In this model, the initiator attempts to access a target in a remote domain (this might for example be a remote service or survey existing in a remote CESSDA RI archive). Two key components support authorised access to the target: a Policy Enforcement Point (PEP), described in the figure as the Access control Enforcement Point (AEF), and a Policy Decision Point (PDP), described as the Access control Decision Function (ADF). The PEP ensures that all requests to access the target are run through the PDP and the PDP casts the authorisation decision on the request based on a collection of rules (policies). To make this structure scalable and easily applicable within a Grid environment, a generic API to model the PEP has been proposed and created by the Authorisation Working Group of the Open Grid Forum (OGF) (www.ogf.org).

The OGF have put forward an API that provides a generic PEP, which can be associated with different authorisation infrastructures. The specification for Grid technologies is an enhanced profile of the OASIS Security Assertion Markup Language (SAML) v1.1 [SAML1-1]. The OASIS SAML AuthZ specification defines a message exchange between a PEP and PDP consisting of an *AuthorizationDecisionQuery* (which contains a *subject*, a *resource* and an *action*) going from PEP to PDP, and an assertion returned containing a number of *AuthorizationDecisionStatements*.

The OGF SAML AuthZ specification [WSCMP] defines a *SimpleAuthorizationDecisionStatement* (a boolean stating "granted/denied") and an *ExtendedAuthorisationDecisionQuery* that allows the PEP to specify whether the simple or full authorisation decision is to be returned. Figure 2 shows the interactions supported by this API.

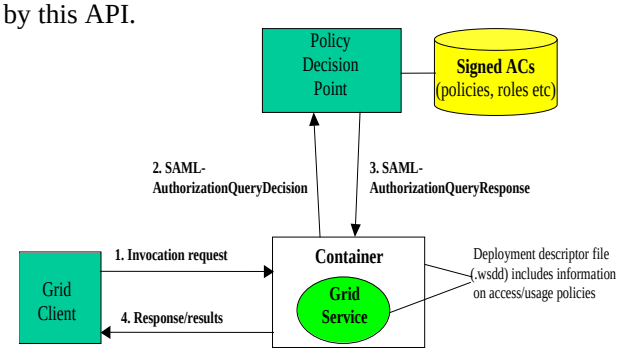

Fig 3. Open Grid Forum SAML AuthZ API

Through this SAML AuthZ API, a generic PEP can be achieved which can be associated with arbitrary Grid services. Thus rather than developers having to explicitly engineer a PEP on a per application basis, the information contained within the deployment descriptor file (.wsdd) when the service is deployed within the its hosting environment (container), is used. Authorisation checks on users attempting to invoke "methods" associated with this service are then made using the information in the .wsdd file and the contents of the LDAP repository (PDP) together with the DN of the user themselves.

Various authorization infrastructures support this PDP and indeed have been put forward for finer grained authorization in a Grid environment. We review some of these and highlight and their pros and cons in the context of the CESSDA RI.

2.2.2.1 Privilege and Role Management Infrastructure Standards Validation (PERMIS)

The Privilege and Role Management Infrastructure Standards Validation (PERMIS) project (www.permis.org) [COB,CO] was an EC project that built an authorisation infrastructure to realise a scalable X.509 AC based PMI. Through PERMIS, an alternative and more scalable approach to centrally allocated X.509 public key certificates can be achieved through the issuance of locally allocated X.509 ACs.

The PERMIS software realises a RBAC authorisation infrastructure. It offers a standards-based Java API that allows developers of resource gateways (gatekeepers) to enquire if a particular access to a resource should be allowed. The PERMIS RBAC system uses XML based policies defining rules, specifying which access control decisions are to be made for given VO resources. These rules include definitions of: subjects that can be assigned roles; SOAs, e.g. local managers trusted to assign roles to subjects; roles and their hierarchical relationships; what roles can be assigned to which subjects by which SOAs; target resources, and the actions that can be applied to them; which roles are allowed to perform which actions on which targets, and the conditions under which access can be granted to roles.

Roles are assigned to subjects by issuing them with X.509 Attribute Certificate(s). Various tools have been developed to support this process. These can support local assignment (using a Privilege Allocator) or support remote assignment of ACs to individuals based upon delegation of authority. In this latter case, a resource provider will delegate credentials that can subsequently be assigned to remote users by a trusted authority. Depending upon the authorisation policy these credential can be further delegated to other remote trusted authorities for subsequent assignment to users. Once roles are assigned, and policies developed, they are digitally signed by a manager and stored in one or more LDAP repositories. When a user attempts to access a PERMIS protected resource, their credentials and potentially the hierarchy of trust relationships between SOAs are either pushed to the service provider or pulled from trusted sources of authority to make access control decisions. The standards and protocols for achieving this have also been standardised by the OGF (https://forge.gridforum.org/sf/projects/ogsa-authz).

The process to set up and use PERMIS can be split into two parts: *Administration* and *Use*. To set up and administer PERMIS requires the use of a LDAP server to store the attribute certificates and reference the SOA root certificate. A local CA is required to be set up using OpenSSL – this designates the SOA and all user certificates created from this CA must have a DN that matches the structure of the LDAP server. The DN of the user certificate is what is used to identify the client making the call on the Grid service.

From the user's perspective, once the administrator has set up the infrastructure, the PERMIS service is relatively easy to use. Unique identifiers are placed as parameters into the user's Grid service deployment descriptor (.wsdd file). These are the Object Identification (OID) number of the policy in the repository, the URI of the LDAP server where the policies are held and the SOA associated with the policy being implemented. Once these parameters are input and the service is deployed, the user creates a proxy certificate with the user certificate created by the local CA to perform strong authentication. The client is run and the authorisation process allows or disallows the intended action.

The PERMIS infrastructure offers very fine grained authorisation capabilities both in terms of policy expression and enforcement. The policy editing tools allow for easy development of the XML based policies. The NeSC at Glasgow have put a variety of user guides on how to set up PERMIS and integrate it with Grid services (see www.nesc.ac.uk/hub/projects/etf).

For the CESSDA RI, it might well be the case that PERMIS is explored as one of the key technologies that can be used to define and enforce access control authorisation policies on CESSDA RI resources.

2.2.2.2 Globus Security Infrastructure (GSI)

GSI (www.globus.org/security) is an example of the classic Access Control List (ACL) used to enforce authorisation and provides a relatively coarse-grained approach to implementing security. A

list is compiled that maps each user's local account name to the DN that appears on their user certificates. When a user makes a method call on a service, this list is consulted and access is granted or denied depending on whether they appear on the list with the correct credentials. Rather than distinguishing between methods this restriction applies to that user for all secured services across the container.

To run the Globus container requires an administrative user (usually 'globus') to set up the container. Each user that wishes to run secure services within this container must have a user certificate located in their home directory. The machine upon which the container is running must also have a host certificate installed by 'root'. Once the container is running, any user should be able to run an unsecured service, with or without a certificate. However, using GSI, a measure of security can be introduced on the service that allows only those with the necessary credentials to run it, typically through a proxy certificate generated from their user certificate.

To use GSI, Grid clients must normally be in possession of a Grid (X.509) certificate which is used to encrypt the communication between client and Grid service. The Grid service is then able to check the identity of the user invoking the service against the local ACL (*grid-mapfile*) that an authorised client is invoking the service.

The latest release of the Globus toolkit (www.globus.org) supports GSI-based authentication and authorization. This includes:

- WS Authentication with support for both message level and transport level security. Message level security is achieved through an implementation of the WS-Security standard that supports message protection at the Simple Object Access Protocol (SOAP) message level. Transport level security is achieved through use of X.509 certificates to establish Transport Layer Security (TLS) connections.
- WS Authorization through an authorisation framework based upon the SAML AuthZ api defined previously.
- Credential Management through MyProxy (a credential storage and management system) and SimpleCA (which as its name implies provides a simple CA).

The MyProxy solution [MyProxy] in particular should be mentioned since this is gaining widespread acceptance as the way in which credentials should be managed within a Grid environment. Instead of users managing their own private keys and credentials, they can delegate them to a MyProxy repository. Through username and password access to MyProxy repositories, short lived proxy certificates can be created. MyProxy also allows for the creation of PKI credentials since later releases now include a CA.

MyProxy solutions are now being used in combination with portals for example, where users accessing a portal through a username and password will automatically have short lived proxy certificates created which can subsequently be used for Grid based job submission. This capability exists for example on the NGS (http://portal.ngs.ac.uk).

Of all of the authorisation infrastructures, GSI is arguably the most straightforward to establish and use. Unsurprising since GSI has been developed as an integral part of the Globus development. That said, the ACL based approach offered by *grid-mapfile*s is a limited form of authorisation however.

For the CESSDA RI it may well be the case that some form of GSI support is needed, e.g. when large scale statistical analysis on HPC/e-Infrastructure facilities is required.

2.2.2.3 Virtual Organization Membership Service (VOMS)

VOMS [VOMS] is a system for managing authorisation data within VOs. It was developed as part of the European DataGrid project (edg-wp2.web.cern.ch/edg-wp2) VOMS has gained widespread acceptance across the Grid community, in part due to the simple model for defining the roles specific to a particular VO and how they can be used/enforced. Sites themselves are responsible for configuring their resources to use these roles. With VOMS, this is implemented with tools such as the Local Centre Authorization Service (LCAS) and the Local Credential Mapping Service (LCMAPS) [LCAS/LCMAPS] which map the user role information into group identities (gid), user identities (uid) and associated local pool accounts established on the local cluster for that particular VO. Refinements can be made to this model in order to allow more local control over the use of resources, e.g. applying file store limits to a particular VO. We note that this local enforcement is not explicitly defined within

the VO policy (given by the definition of the roles in the VOMS server). Rather, this is left up to local administrators to decide how the particular roles and privileges associated with that VO should be interpreted when accessing the resource. Combining VOMS attributes with other authorisation infrastructures has also been explored in projects such as VPman (www.nesc.ac.uk/hub/projects/vpman).

In terms of CESSDA RI, VOMS has several advantages. Firstly it is widely accepted across the Grid community e.g. VOMS has been accepted by many large scale mainstream Grid communities. Consequently good tool support exists for the central management of these roles, such as VOMRS [VOMRS] which allows multiple managers to assign roles to members of the VO, and for end users to see which roles they have been allocated. Furthermore, tools such as *voms_proxy_init* exist for embedding these roles into proxy certificates and for pushing them to the resource sites. Other complementary tools exist for extracting the roles from the proxy certificates at the resource site.

VOMS is ideally suited when large scale, primarily static VOs are needed. Here static implies that the roles and end users with those roles do not change rapidly across the VO. The interpretation and mapping of those roles to local resources may well change more frequently however. If a user's privileges are to be revoked, then the VO administrator can simply remove the roles assigned to this user in the VOMS server, with the consequence that the user's roles are no longer recognized across the whole VO.

Given that the VO roles are agreed by all sites up front when establishing the VO, the VOMS model is simpler to define and agree upon. This model does not depend on the aggregation of numerous bilateral agreements between VO partners where roles and associated trust levels are defined. Rather roles are defined globally across the VO, based upon a VO-wide collaborative agreement. The assignment of these roles to individuals is then made by a designated VO-manager – typically the VOMS administrator (although the manager role can be shared by several people). This super-role is responsible for deciding which users can be assigned which roles across the VO.

The VOMS model, or more precisely agreement on a core set of roles, is also aligned with the principle behind the definition of the eduPerson attribute set for use with technologies such as the Internet2 Shibboleth solutions. Through widespread definition and agreement of the roles to be used across a federation, these may then be delivered and used in a variety of ways.

In the context of the CESSDA RI, VOMS could well be used as a centralised attribute authority across the whole CESSDA RI. However, having a single VOMS authority is potentially dangerous in that it is a single point of failure.

2.2.2.4 Extensible Access Control Markup Language (XACML)

XACML [XACML] is an OASIS standard that describes both a policy language and an access control decision request/response language (both written in XML). XACML version 2.0 was published in 2005. The policy language associated with XACML is used to describe general access control requirements, and has standard extension points for defining new functions, data types, combining logic, etc. The request/response language allows formation of queries to ask whether or not a given action should be allowed, and interpret the result. The response always includes one of four values: Permit, Deny, Indeterminate (an error occurred or some required value was missing, so a decision cannot be made) or Not Applicable (the request can't be answered by this service).

The typical setup is that someone wants to take some action on a resource. They will make a request to a PEP protecting a resource. The PEP will form a request based on the requester's attributes, the resource in question, the action, and other information pertaining to the request. The PEP will then send this request to a PDP, which will look at the request and some policy that applies to the request, and come up with an answer about whether access should be granted. That answer is returned to the PEP, which can then allow or deny access to the requester. In addition to providing request/response and policy languages, XACML also supports finding policies that apply to a given request and subsequent evaluation of requests against that policy. XACML also allows for generic, distributed policies to be supported. Thus a policy can be written which refers to other policies kept in various remote locations. Hence rather than having to manage a single monolithic policy, different people or groups can manage sub-pieces of policies as appropriate, and XACML supports combination of the results from these different policies into one decision.

XACML comes with a core base language which can be extended. The core language supports a wide variety of data types, functions, and rules about combining results of different policies. In addition, standards groups are working on extensions and profiles that will hook XACML into other standards like SAML and LDAP, which will increase the number of ways that XACML can be used.

XACML represents both an alternative to mainstream Grid-based authorisation technologies and a complementary technology. XACML can be used to establish standalone authorisation policies much aligned with PERMIS for example. XACML can also be used to with PERMIS, e.g. to acquire and establish a security context upon which a PERMIS-based access control decision can be based for example.

XACML is also being explored and extended by numerous other research domains in which the CESSDA RI might ultimately need to interoperate with. As one example, the Open Geospatial Consortium (OGC) [OGC] are exploring and enhancing XACML in the geo-spatial domain through a geoXACML refinement of the core XACML specification [geoXACML]. Linkage of CESSDA RI data sets that are spatially referenced with maps/geospatial coordinates may indeed require XACML based interoperability for security. We note that Grid-based systems allow for a multitude of authorisation infrastructures and this in itself is not an issue. More technologies and alternative approaches make building inter-operable security infrastructures more difficult however.

2.3 Grid Portals Standards and Technologies

Web portals provide a single point of access where a variety of information is aggregated and personalised to individuals to improve their experience in accessing and using a range of Internet resources. Common features of web portals include support for categorization of web content and advanced search facilities. Grid portals build upon the general web portal model to deliver the benefits of Grid computing to virtual communities of users, providing a single access point to Grid services and resources. Web 2.0 based solutions whether this be wikis, social networking capabilities, lightweight tools, e.g. for visualisation or mash-ups, can also be made available through portals.

The major difference between a Web portal and a Grid portal is that Grid portals provide a single point of access for Grid resources specific to a given domain, rather than more general Internet-based web pages or content. Grid portals provide end users with a customized view of software and hardware resources specific to their particular problem domain. This customisation can be based upon the privileges that end users have. This can be used to restrict or authorise access to collections of remote services and data sets. Grid portals should ideally allow researchers to focus on their research problems by making the Grid a transparent extension of their desktop computing environment.

The development of targeted portals for the CESSDA RI offers a direct way in which a rich variety of applications and resources can be made available in a transparent manner to users who do not wish to become Grid experts. Should a one-stop CESSDA RI portal be established, it is essential that common approaches be taken to support interoperability, overall manageability of the CESSDA RI and a CESSDA RI branding (common interface).

A CESSDA RI portal-based solution should meet the following requirements:

- *Usability* – the portal should be developed with both the experienced and inexperienced Grid communities in mind. This might benefit from use of backend MyProxy servers to manage user certificates and proxy credentials across the CESSDA resources.
- *Single sign-on* – secure access to a CESSDA portal should allow seamless access to a range of CESSDA-wide resources without the need for multiple authentications. Access should, of course, depend on user privileges.
- *Interoperability* – it should be possible for research communities to develop their own services using potentially different middleware on their own local resources, but be able to make these available to remote researchers through portal technologies.
- *Support for research* – it is essential that the services and data sets made available through the portals meet the real needs of CESSDA researchers. Their input and feedback should drive the design and development of these portals.
- *Support for collaboration* – the portal environments must facilitate collaboration between researchers at all levels – within an institution, between institutions, across national and international levels.

- *Portal administration and management* – user communities should be able to establish and ultimately manage their own services and their own user base. The shared resources underlying these communities need to be autonomous, however, and under the control of these communities.
- *Monitoring* – administrators and users should have direct access to monitoring information about various aspects of the Grid for their virtual organisation, for their institution, and for those using the shared resources. For CESSDA this might for example offer functionality including notifications of new data sets or new tools of interest to communities.

In addition to these, numerous other criteria may be important for the future CESSDA RI including:

- *Security-oriented* – certain CESSDA communities and data providers may wish to use fine-grained security for tailoring access and usage of Grid resources. This might be based on specific roles particular to a virtual organisation.
- *Workflow definition and enactment* – for some CESSDA research activities, it will be necessary to compose Grid services and data movement between those services in a variety of ways on the fly, reflecting the demands of the particular applications.
- *Legacy application support* – portal-based solutions should support the simple upload and execution of existing legacy code/applications, e.g. SAS, STATA, SPSS, R scripts etc;
- *Visualization* – portals can be used to host shared visualisation facilities, e.g. for visualisation of geo-spatial data sets overlaid with social science related data.

Rather than each CESSDA member organisation developing its own Grid portal, sharing expertise and development effort across these portals is essential. Grids require interdisciplinary research techniques and the ability to seamlessly move across and between Grid portals. A common approach to the specification, implementation and management of portal content will contribute to seamless usage.

Whilst it is possible to develop hand-crafted portals, recent advances in this area have resulted in Grid portal frameworks which facilitate re-use of code and support various forms of structuring portal pages. Grid portal frameworks provide a set of basic functionalities and infrastructure for developing further portal components as plug-ins. Common components are offered for security (e.g. access management), for personalisation (e.g. user/group profiles), and for different presentation capabilities (e.g. JSP, XSP, XML/XSLT).

Portals themselves provide access to families of portlets or other hosted applications. Portlets are typically Java-based web components managed by a portlet container that processes requests and generates dynamic content. Portals use portlets as pluggable user interface components, providing a presentation or access layer to systems. Portlets support modular and user centric web applications. Portlets are the building blocks of portals and are typically small units of functionality within a portal. Each portlet typically provides an interface to a Grid service offering some well defined functionality. Users and administrators of communities or virtual organisations more generally can build customized environments by adding portlets. For advanced scenarios, security techniques can be used to authorize use of particular portlets or the resources available to the services accessible via those portlets.

To support portlet and portal interoperability, the portal community and wider industry have developed two key standards of relevance to the Grid community: the Java Portlet Specification (JSR-168) and the Web Service for Remote Portlets (WSRP). JSR-168 enables interoperability among portlets and portals. The specification defines the contract between a portlet and portlet container, and a set of portlet APIs that address personalization, presentation, and security. The specification also defines how to package portlets in portal applications. WSRP allows plug-and-play of content sources (portlets) within portals and other aggregating web applications. WSRP standardizes the consumption of web services in portal front-ends, and the way in which content providers write web services for portals. This allows content producers to maintain control over the code that formats the presentation of their content. By reducing the cost for aggregators to access their content, WSRP improves the integration of content sources into pages for end users.

WSRP and JSR-168 are complementary specifications. JSR-168 defines a standard portlet API for Java-based portals. WSRP allows content to be hosted in the environment most suitable for its execution, while still being easily accessed by content aggregators. Second generation Grid portals can be produced from pluggable (JSR-168 compliant) Grid portlets. Running inside a portlet container, portlets can be added or removed, thus providing administrators with the ability to customize access

and usage of Grid services at portal level. A portal built from Grid portlets can provide users with the ability to integrate services provided by different Grid-enabling technologies. This aspect is critical to the success of CESSDA since a range of distributed services will likely be developed by different communities and institutions, and subsequently made accessible through common research specific portals (VREs).

The Open Middleware Infrastructure Institute (OMII-UK) Security Portlets project at NeSC Glasgow, developed a family of portlets that make secure access through Shibboleth, content configuration and secure access to remote services possible. These JSR-168 portlets support:

- A portlet for scoped attribute management (SCAMP) which allows restricted and syntactically correct manipulation of the Shibboleth attribute acceptance policy, streamlining the subset of IdPs from whom a portal will accept user attributes across the federation. This portlet is generic and can be applied with any portal framework.
- A portlet for creation and usage of X509 attribute certificates (ACP) to allow distributed service providers to make their own local authorisation decisions when users attempt to invoke remote (protected) services. This portlet is generic and can be applied with any portal framework.
- A portlet for content configuration which will support dynamic configurability of portal content based on Shibboleth attributes and knowledge of available services. Once authenticated to a portal via Shibboleth, users are presented with a filtered view of available portlets (and hence access to a restricted set of services). This portlet has been targeted specifically to extend the GridSphere portal framework.

There are a multitude of projects and efforts that are developing portal based frameworks. Some of the more prominent of these include: WebSphere, Java CoG kit, Sakai, uPortal, LifeRay, GridSphere, StringBeans, Jetspeed, OGCE, LifeRay, eXo, GPDK and Pluto. A summary of some of these is included in (http://archive.niees.ac.uk/documents/AH06_Portals_2006.pdf)

There are also rich selections of projects that have developed portal based solutions and rolled them out to wider applications-oriented research communities. If the CESSDA RI is to be fully integrated into wider international efforts crossing social science and other related disciplines, e.g. e-Health, then harmonisation with these efforts through standardised portal solutions such as WSRP and JSR-168 would be highly beneficial.

3 Exploration of Case Studies

In the previous section we have provided an overview of Grid related efforts related to data management, security and portal based solutions. In this section we outline how these standards and technologies could be used to support the use cases described in the original tender document as examples of core capabilities that a future CESSDA RI might be expected to support.

3.1 Scenario 1

A social science researcher wishing to perform European cross-national comparative research needs to efficiently conduct several operations within the research life-cycle.

- *First, the identification of suitable datasets and variables.*
- *Second, seamless access to both data and metadata.*
- *Third, a test bed to determine whether data harmonisation is possible / practical.*
- *Fourth, access to data harmonisation tools and methods.*
- *Fifth, the application of middleware tools to conduct the complex analysis of the resulting harmonised data.*

The researcher would need to select, comment on and compare datasets to be harmonised. This process would involve the standardisation of variables, the editing of questions and the actual harmonisation of variables. They would want to be able to compare terms and question texts, along with the sequence of questions in each case, in order to determine the degree of similarity between variables. The process might also include classification of a dataset within a standard category, the grouping of variables by linking the group to a Classification within a scheme and the addition of new classifications to the system and the editing existing ones.

3.1.1 Grid Possibilities for Scenario 1

Each of the steps in the scenario outlined above could have numerous possible solutions when supported in a Grid environment. For the first step in identifying suitable data sets and variables, I would suggest that there is no particularly radical/novel Grid approach that would offer any direct advantages over existing solutions already in place, e.g. through harvesting RDF descriptions of the available statistical objects on NESSTAR services. The Grid could be used to improve the speed up time for building of indexes using Lucene (also exploiting the ELSST-thesaurus) on HPC facilities however. Several projects already exploit HPC facilities to build a variety of data indexes including the Terabyte Information Retrieval (Terrier – http://ir.dcs.gla.ac.uk/terrier/) project at the University of Glasgow. Further refinements to indexing of CESSDA data could be achieved using the Hadoop Map Reduce distributed indexing scheme.

Of course, building of indexes assumes that direct access to the data or metadata itself is possible. When this is not the case, e.g. due to sensitivity or security constraints of data, then this *modus operandi* is not possible. Instead alternative models of identifying suitable datasets and variables are required. One way that this can be achieved is through adopting a service-oriented architecture where the lower level data variable level is abstracted up to a service level. As a simple example of this, one could imagine that Grid services exist which allow access to particular data sets, e.g. UK Census data sets. If access to these services was made through a single centralised CESSDA RI portal, then portlets could be developed which allow researchers to directly select the variables that they are interested in for the UK Census data set for example. By selection of the variables of interest and submitting the query to the remote Grid services (in this case a remote service that allows access to the UK Census data set), access control to the data can be defined and enforced by that service provider.

With this model, the variables are in effect raised to the level of portlets. The simplest model is where a single portlet provides access to the variables associated with a particular data set (via a single service) of interest. Multiple portlets for multiple services can be supported for social science researchers to access a wide variety of services and data sets. Examples of how single portlets can be used to access individual remote services is shown in Figure 4. In this portal (which is based upon on-going work in the ESRC funded DAMES project), portlets are developed and targeted for individual data sets and variables. In this specific example, a UK Census data portlet has been created which

allows researchers to select the particular variables that they are interested in. Once selected through the portlet interface, the user can subsequently submit the query to the remote service for access to the data/variables of interest.

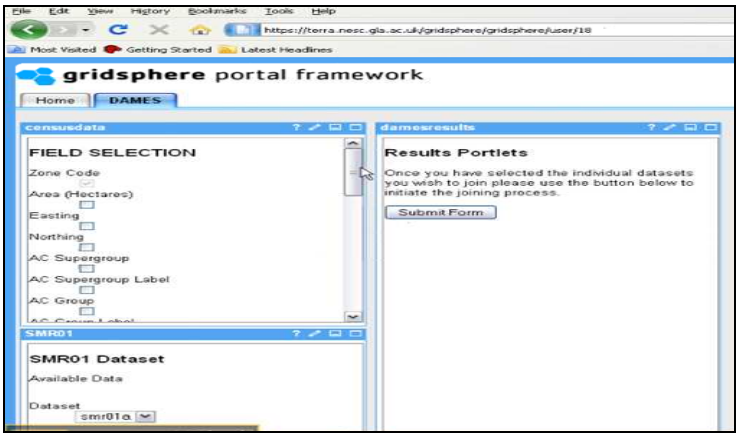

Fig 4. Single Portlets for Accessing Single Services

In addition to UK Census data, researchers are also interested in linkage of UK Census data with health related data sets. In particular, the Scottish Morbidity Records which provide historic clinical information on all hospitals admissions across Scotland for over 30 years; mental health/psychosis data; cancer registrations and death data sets. An SMR portlet for hospital admissions is shown at the bottom of Figure 4. (The above portal also supports access to portlets for mental health and death related data resources also – not shown here). Through this portlet-service related model data can be accessed and used directly, i.e. portlet related requests can return data directly to the portal for the researchers. Alternatively, as is the case in the DAMES project, each of these portlets allows storage of result data sets in a temporary store where they can be linked and processed using further statistical tools. Currently the DAMES project is focused upon STATA, SPSS and R. Portlets for each of these statistical software packages are currently being explored. Richer models of data transfer are possible also, e.g. where portlets interact with services and data sets are themselves transferred to other locations for access by other researchers/research communities. Extending the capabilities and functionality through portals is directly possible. This might through services that can be subscribed to, to provide notifications to researcher communities of the availability of particular data sets or results of particular analyses or tools.

This single portlet-service model works but is likely to have scalability issues when dealing with the full extent of the CESSDA data sets and surveys. It is the case that more complex federated scenarios are also possible where single portlets can provide access to services which in turn support federated queries to one or more remote services. As one example of this, one could have a portlet which invoked a service which accessed geospatial data sets **and** Census data sets for some geospatial analysis. This model was supported in the NeSC Glasgow SeeGEO project. In this model a single portlet interacts with a single remote Grid service which, after authorisation has taken place, submits queries to a remote geospatial web feature service and a Census data provider service (indicated by red circle on the pull down list of variables indicated in Figure 5). The resultant data sets can then be rendered and displayed in the portal (through a suitable geospatial linkage service) to the end user. One interface which describes this is shown in Figure 5.

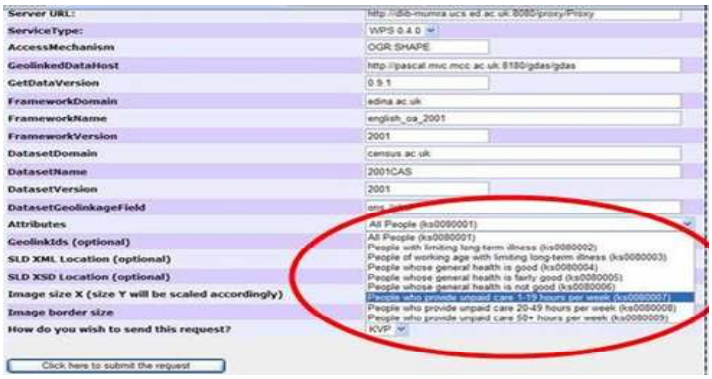

Fig 5. Single Portlet for Multiple Services

The question of scalability still exists however. Is it the case that the CESSDA RI should develop services which provide access to all existing and future data sets? How does this relate to the existing NESSTAR solutions?

My own opinion is that the CESSDA RI might become an infrastructure where multiple services and data sets are made available to *targeted* communities. In Grid parlance these targeted communities might be virtual organisations. Thus a CESSDA RI Grid infrastructure might be one that supports a variety of virtual organisations as opposed to an infrastructure that makes all data accessible to all researchers. (I am not sure how the Grid could actually be applied to achieve this without radically breaking existing models which are in place and seem to perform very well to a wide research community!)

3.2 Scenario 2

Data archivists wish to produce a harmonised dataset from existing CESSDA data resources. The requirement here is for a virtual organisational laboratory where experts from the various CESSDA members can undertake the production of the harmonised resource and make it available to researchers via the existing CESSDA infrastructure for simple analysis and additional tools that allow for more complex analysis. The process involved are the same as those outline in Scenario 1 above but involve a collaborative environment.

3.2.1 Grid Possibilities for Scenario 2

To support a virtual organisation across CESSDA resources involving CESSDA members there are a multitude of possibilities that exist. Will there be a single one-stop shop CESSDA RI portal? Will there be multiple national portals? Will there be a single coordinating centre for the CESSDA RI or will there be multiple coordinating national centres?

To give an example of some of the ways in which Scenario 2 can be supported I use the examples shown in Figures 4 and 5. We assume that a virtual organisation needs to be formed which provides access to a range of distributed social science data sets. A portal is developed which will provide access to the relevant services which in turn will give access to the data sets/variables of interest to that community. This portal is to be made accessible as part of an international Shibboleth federation and we assume that the underlying PKI supports the necessary trust relationships and signing of credentials.

The roles and the privileges can be assigned to that community through either a centralised or decentralised virtual organisation model (or a hybrid combination of them). This is illustrated in Figure 6 where federated IdPs are used for authentication and authorisation information (left of Figure 6) or a centralised attribute authority is used for authorisation information (right of Figure 6). In this scenario we assume that a decentralised model of attributes is used only, i.e. where roles specific to that virtual organisation are provided from the IdPs. Let us assume two basic roles (reflecting advanced/basic roles related to that particular virtual organisation) and a particular end user license agreement is needed for accessing particular resources in that virtual organisation.

Users who attempt to access the portal are redirected to their home site to authenticate and a signed SAML assertion is returned including the role (or roles) and licenses that this individual possesses that the IdP is prepared to release.

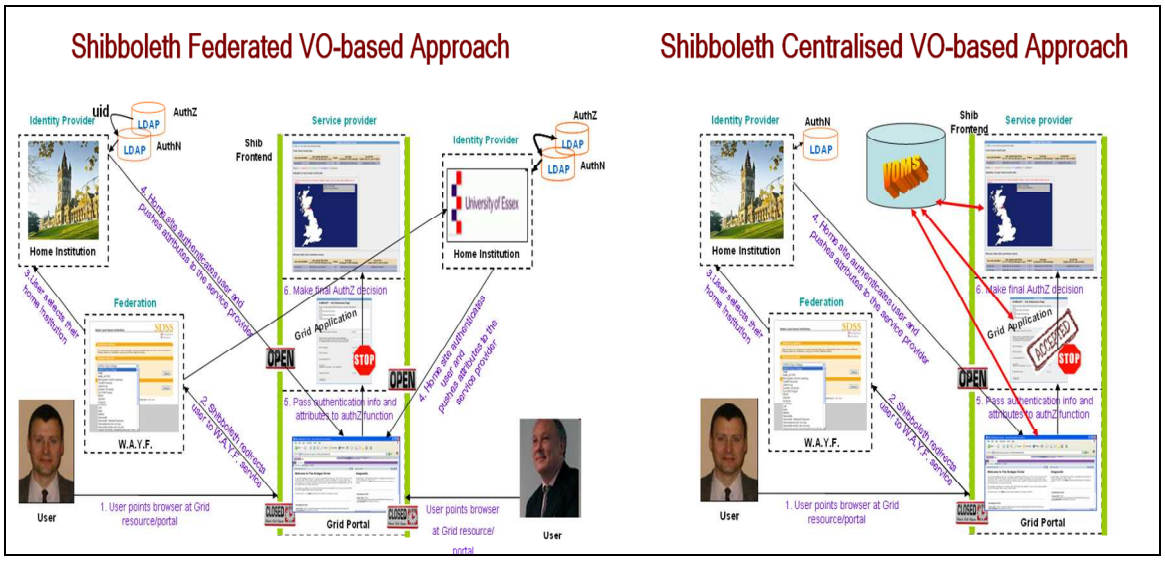

Fig 6. Shibboleth-based Centralised and Decentralised Virtual Organisations

The SAML assertion is checked for validity and based upon the roles that are presented the contents of the portal are configured appropriately. Considering Figure 5, a user with a basic role for access to the UK Census data for example might have a reduced set of variables that they are allowed to access. This is shown on the right of Figure 7 (also based upon work undertaken in the SeeGEO project).

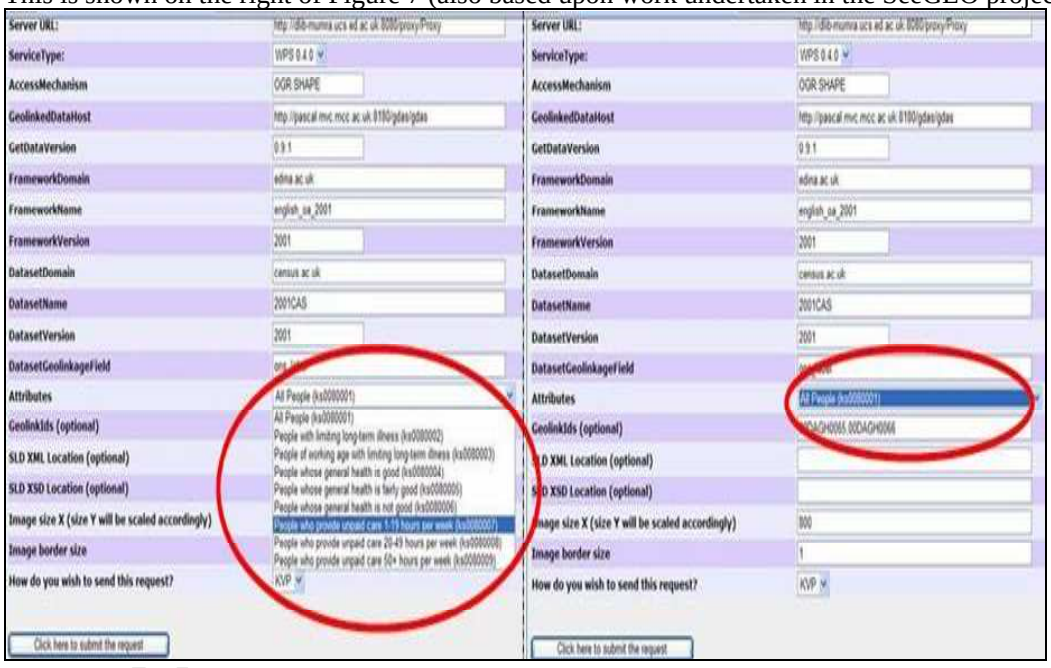

Fig 7. Virtual Organisation-Specific Portal Configuration based upon Different Roles

The queries that are submitted through these portlets to the remote services themselves require to be signed so that the remote provider can determine the validity of the requests. To support this, the above portal exploits a MyProxy service that creates a short term X509 proxy certificate. The service above also uses the SPAM-GP attribute certificate portlet which allows for creation of attribute

certificates (using the role information that was sent through from the IdP). Through extracting the distinguished name of the individual attempting to access the remote resource, e.g. the UK Census data, the service requests the attribute certificates it needs for this individual (from an LDAP server associated with this virtual organisation portal). Through pulling the needed attribute, checking its authenticity and validity, a local authorisation decision can be made and data potentially released for linkage and/or joining with other data sets.

As before the resultant data sets can also be placed in a temporary storage location where other researchers can subsequently access and use them. Further to this, it is possible to define other specific roles that are needed to access and use these data resources, i.e. roles not included in the advanced/basic roles identified in the virtual organisation. The definition of these roles and their subsequent assignment to individuals in a dynamic, distributed manner was supported in the NeSC Glasgow DyVOSE project (www.nesc.ac.uk/hub/projects/dyvose).

Once again, this potentially model has issues with scale. Thus rather than having a single portal comprising access to potentially thousands of services through thousands of portlets, we are assuming that specific virtual organisations only require access to subsets of these services and that subsets of suitable portlets are supported. For example, researchers might only be interested in occupational data resources related to surveys undertaken between a given set of years between a given set of countries. In this case, they should be given access to a greatly reduced set of portlets and services.

It should be noted that this is just one example of a virtual organisation model. A centralised authorisation model can also be supported, e.g. using centralised attribute authorities such as VOMS. Furthermore, many other models can also be defined and implemented. Thus it is quite feasible for workflows to be defined and enacted that allow for the composition of a variety of services in a variety of ways. There is no single Grid model, however as noted, the above models reflect working systems already built at NeSC Glasgow and are thus proven to work.

3.3 Scenario 3

A social science researcher wishes to perform analysis of sensitive potentially disclosive datasets. The requirement here is for a virtual safe setting where access to the data is strictly controlled, with both the proposed analysis and the final results being monitored and approved, and no data actually leaving the safe setting arena. The data are usually very detailed individual and household information collected by national governments. At present this type of data are only accessible to approved researchers in physical safe settings. These safe settings involve physically travelling to a government office, in some cases having to become a temporary member of staff. All code and outputs are checked before analysis can be performed and the results analysed.

The Grid with sufficient security systems in place could become a virtual safe setting, hence offering to the social science community a unique online service that would meet its needs for easier access to this type of detailed data. These virtual environments, also known as data enclaves, have been set up in other countries involving physical safe settings at a researcher's institution accessing data held on a central secure server. However the same processes of control, monitoring and approval are in place with again no data leaving the institutional physical safe setting. This is also more in line with the way the Grid has been used in other disciplines where uptake of the Grid has been much greater.

3.3.1 Grid Possibilities for Scenario 3

The secure data enclave model of data access and usage as supported through the ESRC Secure Data Service and other similar initiatives such as the ONS VML and the NORC Secure Data Enclave has several benefits. The primary benefit of such models is that researchers are offered access to sensitive data sets for research purposes in a strictly controlled and regulated framework. Typically the data enclave model is achieved through a Citrix interface which turns the end user's computer into a 'dumb terminal' giving access to data, statistical software, and collaboratory spaces on a remote secure server. Depending upon the wishes of the data custodians, access can be restricted to particular users and/or particular safe remote locations/machines. Vetting of data analysis outputs for disclosure issues can also be undertaken supported by tools such as τ-ARGUS (http://neon.vb.cbs.nl/casc/tau.htm) which supports algorithms for controlled rounding and cell perturbation of tabular data.

However one of the main issues with such data enclaves is that they do not address the research community requirements for data linkage and analysis. Rather they assume that the data exists in a data enclave and can be analysed and processed there by remote users with dumb terminals. It is often the case however that multiple sensitive data sets need to be brought together for analyses which do not exist in any single data enclave. This might be were a range of primary and secondary care clinical records, longitudinal data need to be linked with individual-level microdata from particular surveys for example. In this case, data enclaves fall short of addressing the needs of the research community since they do not contain all of the necessary data.

To overcome this novel algorithms and e-Infrastructures are required that that can link, analyse and anonymise data yet still meet the needs of the research community. Precisely such algorithms have been implemented at NeSC Glasgow in the MRC funded VOTES project (www.nesc.ac.uk/hub/projects/votes) and they will be applied in the Scottish Health Informatics Platform for Research (www.scot-hip.ac.uk). These algorithms and overall architecture for secure, confidential, anonymising data linkage are incorporated in the Virtual Anonymisation Grid for Unified Access to Remote Clinical Data (Vanguard) system [HG].

The design of Vanguard system is based upon a range of principals that must be strictly adhered to. 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. Vanguard recognizes 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. It is essential 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 (and others with sensitive data sets) is that they simply will not allow direct access through their firewalls to their data. To address this, the Vanguard system is based upon anonymous pull models of data linkage. Thus, rather than data systems being queried directly, i.e. through opening of firewalls, queries are generated based upon a knowledge of the data sets (schemas) that exist at given data provider sites. If a given site has registered itself for participation in a given study or particular virtual organisation, it may subsequently pull the generated queries into their local systems. Depending upon local security policies, these queries are validated and authorized, and if valid, will result in their execution. In short, the systems are completely protected from inbound internet connections (and hence do not have to open their firewalls to the outside world!) but rather are based upon a model only allowing outbound connections to be established. Whilst the Vanguard system itself has been designed based upon this pull model, 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 data resources to users with particular privileges.

The Vanguard system architecture is described in more detail in [HG, OL]. In brief, key components are defined which orchestrate the secure, anonymous interactions. *Viewers* which are used by researchers who require access to data; *Agents* which acts as intermediary between other components; *Guardians* which manage access to and data release from resource providers, and *Bankers* which log usage and maintains use accounts for the data access and usage in the Vanguard system. At the heart of Vanguard is exploitation of public and private keys for encrypting communications between users, between viewers and agents, between agents and guardians.

Through understanding the data models of data providers, i.e. the schemas, and agreements upon visibility of data it is possible to secure link and anonymise data. To achieve this, Vanguard Guardians are able to annotate their data with three access models: *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.

To briefly understand how Vanguard improves upon data enclaves we describe a brief and basic example of Vanguard. The typical scenario involves a user creating a query (through a Viewer) which is encrypted with the Viewers public key and sent to an Agent along with the public key of the individual themselves, e.g. *PKV(Qx,PKUx)* where *PKV* is the public key of the viewer, *Qx* a particular query to be run and *PKU*x the public key of the User. The Agent verifies the Viewer key, checks the

validity of the request, and subsequently defines a data linkage strategy based upon the agreements set out in the particular collaboration, i.e. a federated query needs to be generated which will be pulled down by the Guardians protecting access to the remote (protected) resources. At this point a unique hash key (*HA*) is also generated by the Agent.

At some later time, a Guardian involved in this study will check to see if any queries are generated that it needs to deal with, i.e. Vanguard adopts an asynchronous model of communication. When this is the case the Guardian pulls the query in and checks that it is appropriately signed, i.e. from an Agent it trusts. For a given resource this looks like *PKA(AQuery,HAx,PKUx)* where *PKA* is the Agents public key, *AQuery* the query that is requested to be run against that resource, *HAx* the unique hash key generated and *PKUx* the users public key.

Similar queries are pulled in to and verified by other Guardians involved in the study. Each data provider will assess the query (either through automated RBAC or similar approaches) or through non-automated mechanisms, e.g. discussions with organizational representatives. Assuming that the organization is satisfied with the request, the query is run. The data that can be linked only is hashed with the unique hash key from the Agent. The other contents of the message, i.e. the releasable data are encrypted using the public key of the individual user and the message as a whole encrypted and signed using the Agent's public key. For a given resource this looks like *PKA(PKUx(ARres),HAxAres)* where *HAxAres* is a hashed identifier that cannot be seen directly, but can be linked upon.

After receiving similar encrypted, hashed and encrypted results from all of the Guardians, the Agent can subsequently: decrypt the data using its own private key; join the resultant hashed data sets using the unique hash values that were generated previously, i.e. $PKA^{-1}(Join(HA_x^{-1}(HA_{xAres})\ldots),HAx^{-1}(HA_{xBres})\ldots),HAx^{-1}\ (HA_{xCres})\ldots))$ where the *"…"* represent the other data sets that themselves are encrypted using the users public key. Once joined on the hashed keys, these other data sets are then themselves encrypted using the Viewer public key and released to the end users, i.e. *PKV(PKUx(Joined Linked Anonymised Data)).* The user, i.e. the holder of the private key is thus able to decrypt the joined, linked, anonymised data from the Viewer.

Thus through Vanguard it is possible to link data across multiple secure data resources without directly ever seeing any identifying data. Key to this is that the Agent itself does not have access to identifying data (since it has been hashed) and other non-hashed data has been encrypted using the user's public key.

3.4 Scenario 4

Survey producers wishing to create new instruments based on existing CESSDA metadata elements/questions. The researcher would want to develop a new instrument (survey, questionnaire) based on existing questions that will be comparable, by design, to existing instruments. A question bank would allow for flexible searching/browsing of existing survey questions, with links to the surveys they have been used in, and access to the variations within different waves of the same surveys.

Ideally, the researcher should be able to run different statistical tests on the datasets where the questions have been used/concepts have been measured in order to find questions that have "performed well" earlier.

The researcher should be able to map their new questions to defined concepts or create new ones and to output the final instrument in a format that easily imports into the question bank for the re-use of the questions.

In the creation of new instruments, it is likely that researchers collaborate across sites/geography to design the new surveys, so a collaborative environment may be suitable.

It should also be noted that instruments are often multi-lingual (i.e. surveys are translated into a number of languages), and there is a need to take this perspective into account from the beginning.

The processes involved include the entering of documentation about the new instrument at study level, the addition of information on the operational of the instrument, the preparing and checking of the translation of this metadata and actually translation of the instrument itself.

3.4.1 Grid Possibilities for Scenario 4

Once again, there a variety of potential solutions that could be adopted to tackle this scenario, however from the above text I do not believe that the Grid can be used directly to solve these kinds of questions, i.e. many of these points are domain/survey specific. However as examples of the kinds of approaches that might be taken to tackle this scenario and building upon the previous portal-based examples given in the preceding scenarios, if the new instruments were represented as portlets which define (subsets of) the questions of interest, then the problem becomes one of identifying particular families of portlets[3]. Building up a questionnaire or survey from families of portlets can be undertaken directly in a portal framework. Portal frameworks such as GridSphere allow to define a wide range of portlets and to associate a variety of metadata with them. Definition of generic portlets (or web based forms more generally) for specific research questions that can be refined/extended/adapted and saved for further use can be readily supported.

Questions of scalability are important to address here however. If CESSDA RI is to become the place for all social science and humanities data and services, then the proliferation of portlets would become increasingly difficult to manage and simply will not scale to the full CESSDA RI with the entire research question data bank. This is especially the case when multi-lingual surveys comprised of many thousands of questions exist.

I note that I am not sure how the Grid could be used to tackle this directly.

4 Conclusions and Recommendations

In this report we have explored the Grid and e-Infrastructure landscape and the impact this might have upon a future CESSDA RI. The Grid and development of e-Infrastructures can address many of the needs of the CESSDA RI including seamless access to federated data sets; single sign-on security models and exploitation of wider computational resources for larger scale analysis. It is the case that a multitude of choices and opportunities exist in this space right now. This report has outlined a collection of these choices based upon a given set of use cases that were representative of the domain. All of the scenarios could be supported through a Grid-based e-Infrastructure however numerous issues and questions facing a Grid-based e-Infrastructure for CESSDA RI remain. Chief amongst these questions from my own position is how much change CESSDA RI will accept from the existing models and infrastructure that is already in place? As I state in this report, the Grid does not offer a silver bullet that will make the data, security and delivery challenges facing CESSDA vanish. Rather it offers paradigms that can address key challenges for specific research communities. The model that I would endorse is that the CESSDA RI would support technologies and approaches that allow for the establishment of a wide range of virtual organisations for social science researchers. Thus if I am a social science researcher interested in occupational data sets then should I be offered the complete set of data sets associated with CESSDA or should I be offered a tailored interface to research data sets and tools that meet my own research interests?

There are a multitude of other issues and challenges that remain to be addressed for a Grid-enabled CESSDA RI. These will be outlined in the next report along with a work plan and what it would cost to define and sustain a future Grid-enabled CESSDA RI.

[3] I note that this need not be portlet based solutions but web-based forms, servlets or other web-based applications. The technology is not the issue really.

References

[PKI] R. Housley, T. Polk, *Planning for PKI: Best Practices Guide for Deploying Public Key Infrastructures*, Wiley Computer Publishing, 2001.
[X509] ITU-T Recommendation X.509 (2001) | ISO/IEC 9594-8: 2001, Information technology – Open Systems Interconnection – Public-Key and Attribute Certificate Frameworks.
[TIES] JISC Authentication, Authorisation and Accounting (AAA) Programme Technologies for Information Environment Security (TIES), http://www.edina.ac.uk/projects/ties/ties_23-9.pdf
[ESP] ESP-Grid project, e-science.ox.ac.uk/oesc/projects/index.xml.ID=body.1_div.1
[PH] W. T. Polk and N. E. Hastings, *Bridge Certification Authorities: Connecting B2B Public Key Infrastructures*, http://csrc.nist.gov/pki/documents/B2B-article.doc
[JBH] J. Jokl, J. Basney and M. Humphrey, Experiences using Bridge CAs for Grids, Proceedings of UK Workshop on Grid Security Practice - Oxford, July 2004
[X812] ITU-T Rec X.812 (1995) | ISO/IEC 10181-3:1996, Security Frameworks for open systems: Access control framework.
[WSCMP] V. Welch, F. Siebenlist, D. Chadwick, S. Meder, L. Pearlman, Use of SAML for OGSA Authorization, June 2004, https://forge.gridforum.org/projects/ogsa-authz
[OASIS] Organization for the Advancement of Structured Information Standards (OASIS), http://www.oasis-open.org
[XACML] eXtensible Access Control Markup Language TC v2.0 (XACML), http://www.oasis-open.org/specs/index.php#xacmlv2.0
[WS-S] Web Services Security (WS-Security), version 1.0 5th April 2002, www-106.ibm.com/developerworks/webservices/library/ws-secure
[WS-E] Web Services Eventing (WS-Eventing), www-128.ibm.com/developerworks/webservices/library/specification/ws-eventing
[WS-N] Web Service Notifications (WS-Notifications), http://www-128.ibm.com/developerworks/library/specification/ws-notification/
[WS-RM] Web Services Reliable Messaging (WS-ReliableMessaging), http://www-128.ibm.com/developerworks/library/specification/ws-rm/
[WS-R] Web Services Reliability (WS-Reliability), http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrm
[WS-C] Web Services Co-ordination (WS-Co-ordination), http://www-128.ibm.com/developerworks/library/ws-coor/
[WS-Ch] Web Services Choreography (WS-Choreography), http://www.w3.org/TR/ws-chor-model/
[WS-O] Web Services Orchestration (WS-Orchestration), http://www-128.ibm.com/developerworks/webservices/library/ws-bpelcol2/
[WSS4J] Apache WSS4J, http://www.ws.apache.org/axis.
[WS-Policy] Web Services Policy Framework, September 2004, http://www-128.ibm.com/developerworks/library/specification/ws-polfram/
[WS-Trust] Web Services Trust Language, February 2005, http://www-128.ibm.com/developerworks/library/specification/ws-trust/
[WS-Fed] Web Service Federation Language (WS-Federation), http://www-128.ibm.com/developerworks/webservices/library/ws-fed/
[WS-FW] WS-Federation Passive Requester Profile Interoperability Workshop, http://msdn.microsoft.com/webservices/community/workshops/wsfedprmar2004.aspx
[WS-SC] Web Services Secure Conversation Language, http://www-128.ibm.com/developerworks/library/specification/ws-secon/
[WSE] Microsoft Web Service Enhancements (WSE), http://msdn.microsoft.com/webservices/webservices/building/wse/
[WSW] *Security in a Web Services World: A Proposed Architecture and Roadmap*, A Joint White Paper from IBM Corporation and Microsoft Corporation, April 7, 2002, Version 1.0.
[XMLSig] IETF/W3C XML DSIG Working Group, http://www.w3.org/Signature/
[XMLEnc] W3C XML Encryption Syntax and Processing, W3C Recommendation, December 2002 http://www.w3.org/TR/2002/REC-xmlenc-core-20021210/

[SAML1-1] OASIS, Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) v1.1, 2 September 2003, http://www.oasis-open.org/committees/security/
[SAML2] Security Assertion Markup Language (SAML) version 2.0, March 2005, http://www.oasis-open.org/specs/index.php#samlv2.0
[LibAll] Liberty Alliance, www.projectliberty.org
[LA-IFF] Liberty Alliance Identity Federation Framework, https://www.projectliberty.org/resources/specifications.php
[LA-WSF] Liberty Alliance Identity Web Service Framework version 1.1., https://www.projectliberty.org/resources/specifications.php#box2a
[COB] 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.
[CO] 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.
[OpenSSL] OpenSSL to create certificates, http://www.flatmtn.com/computer/Linux-SSLCertificates.html
[ShibA] Shibboleth Architecture Technical Overview, http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-tech-overview-latest.pdf
[ShibP] Shibboleth Architecture Protocols and Profiles, http://shibboleth.internet2.edu/docs/draft-mace-shibboleth-arch-protocols-latest.pdf
[GT2] Globus toolkit version 2, http://www.globus.org/toolkit/downloads/2.4.3/
[GT4] Globus toolkit version 4, http://www.globus.org/toolkit/downloads/4.0.1/
[EGEE] Enabling Grids for E-sciencE (EGEE) project, public.eu-egee.org
[gLite] gLite software, glite.web.cern.ch/glite
[OMII] Open Middleware Infrastructure Institute (OMII), www.omii.ac.uk
[SSCO] R.O. Sinnott, A.J. Stell, D.W. Chadwick, O.Otenko, Experiences of Applying Advanced Grid Authorisation Infrastructures, Proceedings of European Grid Conference (EGC), pages 265-275, Vol. editors: P.M.A. Sloot, et al June 2005, Amsterdam, Holland.
[SSW] R.O. Sinnott, A.J. Stell, J. Watt, Comparison of Advanced Authorisation Infrastructures for Grid Computing, Proceedings of International Conference on High Performance Computing Systems and Applications, May 2005, Guelph, Canada.
[SC] 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.
[CHAD] D.W Chadwick, *An Authorisation Interface for the Grid*, Proceedings of UK e-Science All Hands Meeting, September 2003, Nottingham, England.
[MyProxy] MyProxy Credential Management Service, myproxy.ncsa.uiuc.edu
[XCO] W. Xu, D. Chadwick, A. Otenko, "Development of a Flexible PERMIS Authorisation Module for Shibboleth and Apache Server", 2nd European PKI Workshop, University of Kent, July 2005.
[eduPerson] eduPerson Specification, http://www.educause.edu/eduperson/
[AuthZ2] Prof David Chadwick, JISC proposal, Authorisation Interface V2 for the Grid, June 2005 – accepted for funding.
[CAS] Community Authorisation Server – http://www.lesc.ic.ac.uk/projects/cas.html
[GSI] Globus Grid Security Infrastructure (GSI), http://www.globus.org/toolkit/docs/4.0/security
[VOMS] R. Alfieri,et al, *Managing Dynamic User Communities in a Grid of Autonomous Resources*, CHEP 2003, La Jolla, San Diego, March, 2003;
[ODvII] 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.
[WS-DAI] OGF Web Service Data Access and Integration Specification https://forge.gridforum.org/sf/projects/dais-wg
[HG] R.O. Sinnott, O. Ajayi, A.J. Stell, A. Young, *Towards a Virtual Anonymisation Grid for Unified Access to Remote Clinical Data*, Proceedings of the 6th International HealthGrid conference, Chicago, USA, June 2008.

[RAS] R.O. Sinnott, O. Ajayi, A.J. Stell. *Data Privacy by Design: Digital Infrastructures for Clinical Collaborations*, to appear in the International Conference on Security and Privacy, Orlando, USA, July 2009.

Annex 1: CESSDA PPP Work Package Context

Task 11.1: A report to consider the possibilities and theoretical implications of grid-enabling social science and humanities (SSH) data collections in the context of the CESSDA RI;

To investigate current developments and applications in Grid technologies in order to propose a strategy for the implementation of a future CESSDA-based SSH cyber-infrastructure. The report should be based on the "Use Cases" above and the specific objectives of the various work packages within the CESSDA-PPP where it envisages that the Infrastructure could utilise Grid technologies and e-social science methodologies to provide pan-European services.

Two objectives **WP4 "Controlled vocabularies"** are:-
To contribute to the harmonisation of datasets by preparing a strategy for the use of controlled vocabularies.
And
To enhance the comparability of datasets by incorporating information on international standard classifications.
At present the controlled vocabularies employed in the Data Portal have only been assigned by a few CESSDA members and then mainly at study level and only to convey subject coverage. Harmonisation aids might require controlled vocabularies on other DDI summary data description elements such as lowest level of geographic aggregation; basic unit of analysis or observation and the type of data. Controlled vocabularies may also be required on some DDI methodology elements such as the time method or time dimension; the type of sample and sample design; the method used to collect the data and the type of data collection instrument. Additionally standard methods of recording temporal and geographic coverage; sample size, response rate and frequency of data collection need to be considered.

All the above would be aids in the discovery of variables that could be harmonized. However these variables might also have categories that use national standard classifications on say education or occupations which would different for each country; hence the need for some mapping between these classifications.

There are two main problems here. Firstly the assigning of controlled vocabularies, especially at variable level which has major resource implications. And secondly when the CESSDA RI starts to include data from resources using metadata schema other than DDI (for example SDMX, ISO 11179, Premis and TripleS) and controlled vocabularies other than ones commonly agreed within CESSDA.

Our initial thoughts are that maybe Grid technologies and e-social science methodologies could be used for automatic indexing, metadata registries and mapping of controlled vocabularies and standard classifications. Another useful tool in the discovery of compatible variables would be a question bank.

Two objectives of **WP5 "Developing the CESSDA RI one-stop-shop Portal"** are:-
To evaluate middleware for a common federated access, authentication and authorisation system.
And
To evaluate persistent common identification systems for data objects in order to maintain strict identification version control.
Our initial thoughts are that maybe Grid technologies and e-social science methodologies could offer different architectures for the Data Portal involving alternatives to Shibboleth authentication and the use of DOI or Handle.
However this workpackage is also investigating the present architecture of the data portal, additional services that might result from the adoption of version 3 of the DDI xml standard and publishing to the data portal of data described with the new xml.

Two objectives each of **WP6 "Strengthening the CESSDA RI" and WP7 "Widening the CESSDA RI"** are:
To support capacity-building through developing the skills, knowledge and abilities of less-developed and less-resourced CESSDA organisations, by means of staff training and exchange programmes.
And
To foster and develop emerging CESSDA organisations through the provision of a complete 'tool kit' of standards, operational tools and expertise, allowing effective knowledge transfer.
And
To extend the existing CESSDA RI and to foster the development of national data archiving initiatives in those countries which are not currently part of the CESSDA network, in order to create and maintain a 'complete' pan-European RI, including representation from emerging and candidate countries.
And
To extend the network to agencies and organisations which remain outside of CESSDA yet continue to host important data collections.
Our initial thoughts are that maybe Grid technologies and e-social science methodologies could offer a virtual organisation setting for the distributed CESSDA members and a virtual safe setting for the more sensitive data that the CESSDA RI would wish to host. Both WPs are concerned with issues such as certifying data archives and professionalising data
archivists, best practices, pooling of distributed expertise and quality assurance. It is hoped that CESSDA archives would become trusted data repositories in all member countries, hence opening up the opportunities for richer data deposits.

An objective of **WP8 "Enhancement of data and metadata infrastructures for the CESSDA RI"** is:
To plan the strategic developments required for metadata, data models and software upgrades for data and metadata capture, management, processing, publishing and access within the CESSDA RI to support more complex dataset types.

An objective of **WP9 "Developing the CESSDA RI by building an infrastructure for content harmonisation and conversion"** is:
To strategically plan for meeting substantive harmonisation demands in the European SSH research community.
Our initial thoughts are that maybe Grid technologies and e-social science methodologies could again be used for metadata registries and the manipulation of the complex data. Further the use of these technologies and methodologies could assist in the creation of actual harmonisation tools.
Both these WPs have strong links with WP4 and WP5 and other issues involved have been discussed above in those two workpackages.

WP 10 "Data collection, dissemination and access issues" main concern is with data collection, data preservation and data accessibility and the issues involved have been discussed above in other workpackages.

Annex 2 : Grid-based Security Practices Today

It is the case that the future CESSDA RI e-Infrastructure may well have to interface with existing Grid-based solutions for security. Given that most Grid solutions today are based upon X.509 certificates to support PKIs it is worth highlighting PKIs, the reason they have been adopted, and their issues and limitations. More information on PKIs is available through the JISC funded TIES project [TIES] or in [POLK].

Public Key Infrastructures (PKI)

Cryptography is one of the main tools available to support secure infrastructures. Using cryptographic technology, confidentiality can be established by encrypting and decrypting messages and their contents. Encryption and decryption are done using keys. When these keys are the same, this is called symmetric-key cryptography.

Public-key cryptography uses different keys: private and public keys. Messages encrypted with a public key can only be read by an individual who possesses the private key. Any 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.

A PKI is responsible for deciding policy, managing, and enforcing certificate validity checks. The central component of a PKI is a Certificate Authority (CA). A CA is a root of trust which holders of public and private keys agree upon. CAs have numerous responsibilities including issuing of certificates, often requiring delegation to a local Registration Authority (RA) used to prove the identity of users requesting certificates. CAs are also required amongst other things to revoke older or compromised certificates through issuing Certificate Revocation Lists (CRL). A CA must 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 [PH,JBH] 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 depicted in Fig 1.

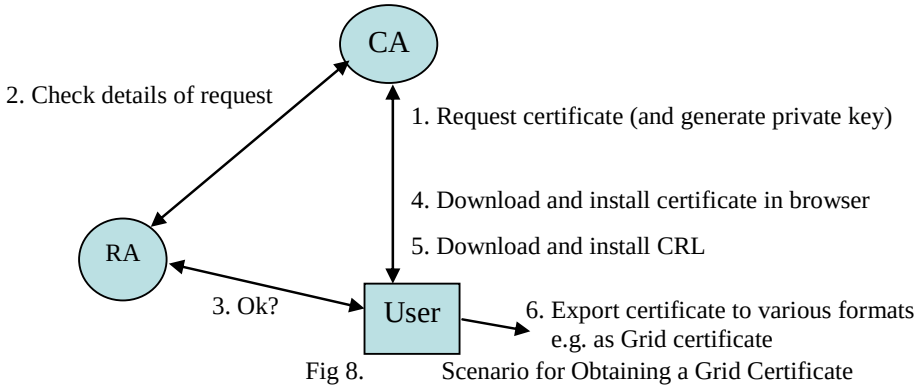

Fig 8. Scenario for Obtaining a Grid Certificate

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) 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.

The main benefit and reason for the widespread acceptance of PKIs within the Grid community is their support for single-sign on. Thus 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 *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*.

Problems with PKIs

As stated, researchers wishing to gain access to Grid resources such as the NGS in the first instance have to acquire a UK e-Science X.509 certificate issued by the centralised CA at RAL. This process itself is off-putting for many of the wider less-IT focused research community since it required them to convert the certificate to appropriate formats understandable by Grid (Globus) middleware, e.g. through running commands such as:

$> openssl pkcs12 -in cert.p12 -clcerts -nokeys -out usercert.pem

Such requirements are likely to dissuade less IT-savvy researchers from engaging – especially as openSSL is not commonly available on platforms such as Windows. We note that the Certification Authority now suggests for researchers with Windows based PCs that they can use a Windows openSSL based solution (http://www.openssl.org/related/binaries.html) but this in turn requires them to install and configure additional software etc. In some circumstances this is not possible, for example if they do not have sufficient privileges on their PC (root access etc) – a not uncommon practice in certain departments and faculties at Glasgow University for example. 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 strong 16-character passwords for their private keys with the recommendation to use upper and lower case alphanumeric characters. The temptation to write down such passwords is apparent and an immediate and obvious potential security weakness.

This process as a whole does not lend itself to the wider research community which the e-Science and Grid community needs to reach out to and engage with. Given the fact that the initial user experience of the Grid currently begins with application for UK e-Science certificates, this needs to be made as simple as possible, or potentially removed completely.

There are other issues with PKIs and Grid certificates as currently applied in the UK community. Thus for example *grid-mapfile*s are currently manually updated and managed based upon individual user requests. Solutions such as VOMS offer capabilities for dynamically updating *grid-mapfile*s across multiple Grid resources. The dynamicity of this manual approach is also not conducive to the Grid-idea for establishing new short term VOs. Instead users have to statically have their DNs registered at collaborating sites which have previously made available/allocated local accounts.

The fundamental issue with PKIs however, is trust. Sites trust their users, CAs and other sites. If the trust between any of these is broken, then the impact can be severe, especially since users are currently free to compile and run arbitrary code. With the now global PKI and associated recognition of international CAs through efforts such as the International Global Trust Federation (www.igtf.net), this basic trust model is naïve. Practices and solutions which help make Grid infrastructures safer are thus required.

Annex 3: Service-Oriented Architectures and Security

The development of robust Grid security infrastructures is very much dependent upon agreements on technologies and practices. Standardisation plays an extremely important role in this regard. With the move of the Grid community towards web services and service-oriented architectures, web service security standards and their associated implementations are crucial and could have a major impact upon the future CESSDA RI. Unfortunately it is the case that a multitude of specifications and proposals for web service standards have been promised and put forward. There are often cases of web service standards covering similar topics resulting in multiple competing specifications such as WS-Notifications [WS-N] and WS-Eventing [WS-E]; WS-ReliableMessaging [WS-RM] and WS-Reliability [WS-R]; WS-Orchestration [WS-O], WS-Co-ordination [WS-Co] and WS-Choreography [WS-Ch], along with the many varieties of workflow or business process languages that have been put forward to name but a few examples of the issues in the proliferation of web service standards. It is also the case that many web services standards exist only in working draft status, often with no associated implementations or acknowledged conformance or interoperability definitions. Claiming conformance or compliance to a particular web service standard is thus often not possible (or meaningful!). The structuring and content of SOAP messages used in Web Services leaves open numerous possibilities which complicate interoperability.

It is the case that although many standards use the common prefix "WS-", this in itself does not mean that there is an agreed WS-Architecture. This stems from a variety of reasons: vendor and commercial issues; political aspects and also the different bodies involved. For example the Internet Engineering Task Force (IETF) (www.ietf.org); the World Wide Web Consortium (W3C) (www.w3.org); the Organization for the Advancement of Structured Information Standards (OASIS) (www.oasis-open.org); and the Web Services Interoperability Organization (WS-I) (www.ws-i.org) are some of the most prominent bodies. The consequence of this profusion of standards and standards making bodies, and the lack of consensus on the core web service architecture, impacts directly upon development of Grid standards, architectures and associated implementations and middleware.

In this section we provide a brief overview of these security standards. All of these standards build upon the basic SOAP foundations which include XML Signature [XMLSig] and Encryption [XMLEncrypt] for ensuring the security of messages. The XML Signature specification defines a methodology for cryptographically signing XML. The signatures are defined using a <Signature> element and accompanying sub-elements as part of a security header. The signature itself is computed based on the SOAP message content and a security token.

WS-Security

WS-Security describes enhancements to SOAP messaging to provide security enhancements for message integrity and message confidentiality. WS-Security also defines a general purpose mechanism for how to attach and include security tokens within SOAP messages including binary encoded security tokens such as X.509 certificates. These mechanisms can be used independently or in combination to accommodate a wide variety of security models and encryption technologies.

Message integrity is provided by leveraging XML Signature in conjunction with security tokens to ensure that messages are transmitted and received without modifications. The integrity mechanisms are designed to support multiple signatures, potentially by multiple actors, and to be extensible to support additional signature formats. The signatures may themselves reference security tokens.

Message confidentiality is provided by leveraging XML Encryption in conjunction with security tokens to keep portions of SOAP messages confidential. Any portions of SOAP messages, including headers, body blocks, and substructures, may be encrypted. It should be noted that the encryption mechanisms of XML Encryption are designed to support additional encryption technologies, processes, and operations by multiple actors. The encryption itself can be realized using either symmetric keys shared by the sender and the receiver of the message or a key carried in the message in an encrypted form.

WS-Security defines a framework for securing SOAP messages, with the specifics being defined in profiles determined by the nature of the security token used to carry identity information. There are

for example different profiles of WS-Security for various different security token formats such as X.509 certificates and Kerberos tickets. There is also a SAML token profile of WS-Security that specifies how SAML assertions can be used to provide message security. Additionally, SAML itself points to WS-Security as an approved mechanism for securing SOAP messages carrying SAML protocol messages and assertions.

WS-Security has now been fully implemented by several web service providers and the Grid middleware community. For example, the OMII server and client software stacks provide an implementation of WS-Security based upon Axis and WSS4J [WSS4J].

WS-Policy

WS-Policy [WS-Policy] describes how senders and receivers can specify their security requirements and capabilities. WS-Policy has been designed to be extensible and does not place limits on the types of requirements and capabilities that may be described. However, the specification does identify several basic attributes including privacy attributes, encoding formats, security token requirements, and supported algorithms. WS-Policy thus provides a flexible and extensible grammar for expressing the capabilities, requirements, and general characteristics of web service-based systems. WS-Policy also defines a framework and a model for the expression of these properties as policies. Policy expressions can include both simple declarative assertions as well as more sophisticated assertions. A policy itself can be regarded as a collection of one or more policy assertions. These assertions might include for example the authentication scheme, transport protocol selection, privacy policy, or quality of service characteristics. WS-Policy provides a single policy grammar to allow for such kinds of assertions to be reasoned about in a consistent manner.

It should be noted that WS-Policy stops short of explicitly specifying how policies are discovered or attached to a web service. It is envisaged that subsequent specifications will provide profiles on WS-Policy usage within given web services technologies and domains. For example, specifications for WS-PolicyAttachments, WS-PolicyAssertions, WS-SecureConversation have been put forward already as have various domain-specific assertions such as WS-SecurityPolicy and WS-ReliableMessagingPolicy. (See [WS-Policy] for further information).

WS-Trust

The goal of WS-Trust [WS-Trust] is to enable applications to construct *trusted* SOAP message exchanges. WS-Trust uses the basic mechanisms for secure messaging from WS-Security and defines additional primitives and extensions for security token exchange to enable the issuance and dissemination of credentials within and between different trust domains. Thus for example, to secure a communication between two parties, the two parties must exchange security credentials (either directly or indirectly). However, each party needs to determine if they can *trust* the asserted credentials of the other party. To support such situations, WS-Trust has defined extensions to WS-Security that provide methods for issuing, renewing, and validating security tokens; and ways to establish, assess the presence of, and broker trust relationships. Through these extensions, applications can engage in secure communication designed to work with the general web services framework including WSDL service descriptions, UDDI and SOAP messages.

WS-Privacy

The WS-Privacy specification was outlined in a joint white paper from IBM and Microsoft [WSW]. Here it was presented how the WS-Privacy specification could address how privacy practices could be stated and subsequently implemented and enforced by web services. By using a combination of WS-Policy, WS-Security and WS-Trust, organizations should be able to state and indicate conformance to stated privacy policies. The specification would describe a model for how a privacy language could be embedded into WS-Policy descriptions and how WS-Security may be used to associate privacy claims with a message. In addition, the WS-Privacy specification would describe how WS-Trust mechanisms could be used to evaluate these privacy claims for both user preferences and organizational practice claims.

At the time of writing, the WS-Privacy specification and associated implementation(s) have not materialised, nor is it clear when they will appear.

WS-SecureConversation

The Web Services Secure Conversation Language (WS-SecureConversation) [WS-SC] allows clients and web services to establish a token-based, secure conversation for the duration of a given session. The secure conversation itself is based on security tokens that are procured from a service token provider. Once obtained and a secure channel established, the client and service exchange a lightweight, signed security context token, which optimizes message delivery time compared with using regular security tokens. The security context token enables the same signing and encryption features as other security tokens such as X509 security tokens.

WS-SecureConversation itself is built on top of the WS-Security and WS-Policy models to provide secure communication between services. WS-Security focuses on the message authentication model but not a security context, and thus is subject several forms of security attacks. WS-SecureConversation defines mechanisms for establishing and sharing security contexts, and deriving keys from security contexts, to enable a secure conversation.

It should be noted that WS-SecureConversation by itself does not provide a complete security solution rather WS-SecureConversation is a building block that is used in conjunction with other web service and application-specific protocols such as WS-Security to accommodate a wide variety of security models and technologies. It should also be noted that WS-SecureConversation is designed to operate at the SOAP message layer so that the messages may traverse a variety of transports and intermediaries. This does not preclude its use within other messaging frameworks. In order to further increase the security of the systems, transport level security may be used in conjunction with both WS-Security and WS-SecureConversation across selected links.

Several implementations of WS-SecureConversation are now available for example within Microsoft Web Service Enhancements for the .NET platform [WSE].

WS-Federation

The Web Service Federation Language (WS-Federation) [WS-Fed] defines how to construct federated trust scenarios using the WS-Security, WS-Policy, WS-Trust, and WS-SecureConversation specifications. For example, WS-Federation describes how to federate between Kerberos and PKI infrastructures. The WS-Federation specification defines the model and framework for federation between security domains. Subsequent documents define profiles which detail different ways that the WS-Federation language can be applied.

WS-Federation supports specification of a trust policy to identify and constrain the type of trust that is being brokered. Through this different security realms are able to federate by supporting the brokerage of trust of identities, attributes, and authentication information between participating web services.

Various implementations of WS-Federation have been put forward. For example, Microsoft, IBM, RSA Security Inc. and various other vendors have implemented this specification and demonstrated a degree of interoperability between their implementations, e.g. through passing a particular identity between different exemplar portals [WS-FW].

WS-Authorization

A standard for authorization does not exist for web services. In the Microsoft/IBM roadmap for web services security white paper [WSW], an outline for WS-Authorization was loosely described. This document outlined how the WS-Authorization specification would “describe how access policies for a web service are specified and managed. In particular it will describe how claims may be specified within security tokens and how these claims will be interpreted at the endpoint. This specification will be designed to be flexible and extensible with respect to both authorization format and authorization language. This enables the widest range of scenarios and ensures the long-term viability of the security framework”.

However, the WS-Authorization specification has not (yet?) been published. Since this roadmap document was published, developments within the Grid community regarding authorisation and how such infrastructures can be seamlessly linked to Grid services have matured however (as described in section 2 of this document). As such, from a Grid community perspective, the question may well be asked, what would a WS-Authorization specification offer that can not yet be supported by Grid based solutions and existing authorisation infrastructures?

Security Assertion Markup Language (SAML)

The OASIS SAML specification [SAML1-1] is an XML-based framework for communicating user authentication, entitlement, and attribute information. SAML allows making assertions regarding the identity, attributes, and entitlements of a subject to other entities. SAML has been designed to be a flexible and extensible protocol which can be customised by other standards.

SAMLv1.0 became an OASIS standard in November 2002. SAMLv1.1 followed in September 2003 and has seen significant success, gaining acceptance across a wide range of domains and is supported by numerous security technology providers.

SAML is defined in terms of assertions, protocols, bindings, and profiles. An assertion is a package of information that supplies one or more statements made by a SAML authority. SAML defines three different kinds of assertion statement that can be created by a SAML authority:

- Authentication: which indicates that the specified subject was authenticated by an identity provider through some means at some given time;
- Attribute: the specified subject is associated with the supplied attributes;
- Authorization Decision: a request to allow the specified subject to access the specified resource has been granted or denied.

The outer structure of an assertion is generic, providing information that is common to all of the statements within it. Within an assertion, a series of inner elements describe the authentication, attribute, authorization decision, or user-defined statements containing the specifics.

SAML defines a number of request/response protocols that allow service providers to request various things. For example, to request one or more assertions from given SAML authorities, or to request that an identity provider authenticate a principal and return the corresponding assertion.

Mappings from SAML request-response message exchanges into standard messaging or communication protocols are called SAML protocol bindings. A SAML SOAP Binding has been defined which outlines how SAML protocol messages can be communicated within SOAP messages.

A profile of SAML typically defines constraints and/or extensions in support of the usage of SAML for a particular application. For instance, the Web Browser Single Sign On [WebSSO] profile specifies how SAML authentication assertions are communicated between an identity provider and service provider to enable single sign-on for a browser user. This profile details how to use the SAML Authentication Request/Response protocol in conjunction with different combinations of the HTTP Redirect, HTTP POST, HTTP Artefact, and SOAP bindings.

Other SAML profiles also exist such as attribute profiles which provide specific rules for interpretation of attributes in SAML attribute assertions. For example the X.500/LDAP profile, describing how to carry X.500/LDAP attributes within SAML attribute assertions.

SAMLv2.0 unifies the building blocks of federated identity in SAMLv1.1 with input from the Internet2 Shibboleth initiative and the Liberty Alliance's Identity Federation Framework [LA-IFF]. As such, SAMLv2.0 is a significant step towards convergence for federated identity standards.

SAMLv2.0 includes numerous additional features from v1.1 which could have a direct impact upon the CESSDA RI. These include support for:

- opaque pseudo-random identifiers (pseudonyms) which can be used between providers to represent principals;
- identifier management allowing providers to establish and subsequently manage the pseudonym(s) for the principals for whom they are operating;
- metadata defining how to express configuration and trust related data to make deployment of SAML systems easier;

- attribute statements, name identifiers, or entire assertions may be encrypted in SAMLv2.0. This feature ensures that end-to-end confidentiality of these elements may be supported as needed.
- attribute profiles which simplify the configuration and deployment of systems that exchange attribute data. These include basic attribute profiles for string based attribute names and XML schema primitive type attribute value definitions; X.500/LDAP attribute profiles; and XACML attribute profiles.
- SAMLv2.0 supports situations where authenticated users can be automatically logged out of all service providers in the session at the request of the identity provider.
- SAMLv2.0 includes mechanisms that allow providers to communicate privacy policy and settings. For instance, SAML makes it possible to obtain and express a principal's consent to some operation being performed.
- In scenarios with more than one identity provider, service providers need a means to discover which identity provider(s) a principal uses. The identity provider discovery profile relies on a cookie written in a common domain between identity and service providers.

The UK Access Management Federation now supporting SAMLv2.0 with SAML1.1 slowly being phased out.