SUPPORTING SECURITY-ORIENTED, INTER-DISCIPLINARY RESEARCH: CROSSING THE SOCIAL, CLINICAL AND GEOSPATIAL DOMAINS

R.O. Sinnott[1]*, T. Doherty[1], C. Higgins[4], P. Lambert[3], S. McCafferty[1], A. Stell[1], K. J. Turner[2], J.P. Watt[1]

[1]*National e-Science Centre, University of Glasgow, Glasgow, G12 8QQ*
[2]*Department of Computing Science & Mathematics, University of Stirling, Stirling, FK9 4LA*
[3]*Department of Applied Social Science, University of Stirling Stirling, FK9 4LA*
[4]*EDINA, University of Edinburgh, Edinburgh, EH9 1PR*

How many people have had a chronic disease for longer than 5-years in Scotland? How has this impacted upon their choices of employment? Are there any geographical clusters in Scotland where a high-incidence of patients with such long-term illness can be found? How does the life expectancy of such individuals compare with the national averages? Such questions are important to understand the health of nations and the best ways in which health care should be delivered and measured for their impact and success. In tackling such research questions, e-Infrastructures need to provide tailored, secure access to an extensible range of distributed resources including primary and secondary e-Health clinical data; social science data, and geospatial data sets amongst numerous others. In this paper we describe the security models underlying these e-Infrastructures and demonstrate their implementation in supporting secure, federated access to a variety of distributed and heterogeneous data sets exploiting the results of a variety of projects at the National e-Science Centre (NeSC) at the University of Glasgow.

Keywords: *Interdisciplinary e-Research; Virtual Organisations; Security; Authorisation; Single Sign-On.*

1. INTRODUCTION

Much scientific research now crosses the boundaries of individual research disciplines. For example, if one considers studies in particular chronic diseases and the response of specialised study-specific treatments then this can require the interplay between the clinical sciences, the biological sciences, the social sciences and the geospatial sciences amongst others. To accommodate such inter-disciplinary research, e-Research infrastructures need to support the seamless and transparent linkage across disciplines and the resources they offer. This has to be aligned with the way in which the researchers themselves wish to work, and satisfy all concerns associated with the numerous stakeholders in this space, e.g. on access control. This is made especially challenging given the evolving nature of science and the associated resources available and the often dynamic nature of collaborations themselves.

For many disciplines that are primarily computationally-bound, i.e. where access to large high performance computing (HPC) facilities is the primary handicap restricting scientific progress, resources such as the UK e-Science National Grid Service (www.ngs.ac.uk) and ScotGrid (www.scotgrid.ac.uk) are available. However for many would be users, these resources are still largely established for the computationally savvy who are proficient with, or at least happy to deal with the nuances and intricacies of complex Grid middleware. Indeed the initial step in gaining access to such facilities is the

* Author for correspondence (r.sinnott@nesc.gla.ac.uk)

requirement to obtain a UK e-Science X.509 certificate. This step is off-putting for many users especially since it often requires them to refer to registration authorities (which may not exist at their particular institution) and once allocated, convert these certificates to formats that are suitable for accessing and using Grid resources. These issues are described in detail in [1-3]. Furthermore, these HPC facilities are primarily targeted to user communities to compile and run their own simulations although we note that progress in supporting application portfolios - albeit HPC-oriented applications, is now underway through the NGS job submission portal for example (https://portal.ngs.ac.uk/). We claim that the vast majority of researchers that could potentially benefit from research infrastructures are not limited by lack of access to HPC resources. Rather, it is the lack of targeted services and environments that limit their research and, especially with regard to inter-disciplinary research, the interconnectedness of these research environments.

The service-oriented architecture paradigm where Grid services provide access to resources is *potentially* more aligned with the actual needs of research communities. In this model, researchers should, at least in principle, be able to make use of distributed resources without necessarily being savvy computational scientists. Instead, tailored research environments should allow the coupling of collections of these services together for specific research purposes, e.g. through portals or workflow environments, as part of a user-oriented research framework. A fundamental requirement in the realisation of this for many disciplines is ensuring that these services and the resources they make available are only accessible to legitimate individuals. The framework of rules and regulation by which the legitimacy of an individual or collaboration more generally can be defined will vary from project to project and more generally from domain to domain. In e-Research and Grid parlance, such a framework is typically expressed through *virtual organisations* (VO) which identify the resources to be shared and the terms and agreements by which these resources can be used by researchers. Key to the success of any service-oriented architecture-based VO model is the delivery of the environments aligned with, and driven by, research needs and working practices. These environments should allow the various stakeholders in this space, e.g. data providers, to define and enforce their own policies on access control for example.

In this paper we outline security-driven e-Infrastructure solutions which are aligned with the way research communities are comfortable in accessing internet resources more generally. Through exploiting the Internet2 Shibboleth technologies [4-5] and the UK Access Management Federation (www.ukfederation.org.uk), and offering supporting tools that allow for a range of security models for a variety of distributed and heterogeneous services we demonstrate how one of the fundamental tenets of the Grid model, namely single sign-on, can be supported to enable user-oriented and truly interdisciplinary research. We emphasise that this single sign-on model goes beyond the X509 authentication-based models as typified with HPC access control through Globus grid mapfiles for example, to include single sign-on models through fine grained authorisation infrastructures. To demonstrate this, we show how the seamless interplay of collaborative research environments can support research into chronic diseases such as diabetes across the Scottish research landscape incorporating clinical, social and geospatial resources – each of which has their own particular access control policies.

The rest of the paper is structured as follows. We first begin with a synopsis of the Grid security models existing today including the UK Access Management Federation and outline results from the OMII-UK funded Security Portals project (www.nesc.ac.uk/hub/projects/omii-sp) and the JISC funded VPman project (http://sec.cs.kent.ac.uk/vpman/) which demonstrate how a variety of fine grained authorisation-based security solutions are now. Based upon this, in section 3 we outline the clinical data landscape across Scotland building upon the results of the MRC funded Virtual Organisations for Trials and Epidemiological Studies (VOTES) project (www.nesc.ac.uk/hub/projects/votes). We describe the challenges in accessing and using

clinical data sets and show how this is currently supported. In section 4, we describe the social data environment and outline how a variety of social data resources can be seamlessly accessed through these VO models drawing upon results of the ESRC funded Data Management through e-Social Sciences (DAMES) project (www.dames.org.uk). In section 5, we describe how geospatial data sets can be accessed and used through building upon results of the JISC funded Secure Access to GeoSpatial Services (SeeGEO) project (http://edina.ac.uk/projects/seesaw/seegeo). Finally in section 6 we draw conclusions on the work and outline areas of future research in supporting inter-disciplinary research.

2. VIRTUAL ORGANISATION SECURITY INFRASTRUCTURES

Historically, much of the focus and effort of Grid computing was based upon addressing access to and usage of large scale HPC resources such as cluster computers. These access models are typified by their predominantly authentication-only based approaches which support secure access to an account on a cluster where domain specific programs can be compiled and/or executed. These approaches are based upon X.509 based public key infrastructures (PKI) [4] where the public key certificates (PKCs) that bind the identities of users to their public keys are issued by trusted third parties called certification authorities (CAs). Through trusting a CA, sites can validate the identity of the individual in possession of the corresponding private key. This PKI based approach has been adopted in the UK by the NGS. In this model, users specifically request access to individual NGS resources by quoting their Distinguished Name (DN) which is embedded in their X.509 PKC. Their DN is registered in a resource maintained grid mapfile which associates their DN with a local account on that resource. If a DN does not have an associated local account then the local gatekeeper will decide that the user does not have privileges to run the job.

It is often the case that research domains and resource providers require more information than simply the identity of the individual in order to grant access to use their resources. The same individual can be in multiple VOs each of which is based upon a common shared infrastructure. Knowing in what context a user is requesting access to a particular resource is essential information for a resource provider to decide whether the access request should be granted or not. There are numerous technologies and standards that have been put forward for defining and enforcing authorization policies for access to and usage of Grid resources [8]. Role based access control (RBAC) is one of the more established models for describing such policies [26]. In the RBAC model, VO specific roles are assigned to individuals as part of their membership of a particular VO. Possession of a particular role, combined with other policy-specific context information, e.g. the time of day, the amount of resource being requested etc, can then be used by a resource gatekeeper to decide whether an access request is allowed or not. However RBAC models typically assume that a single domain with centralized role management exists so that conflicting roles cannot be issued to users and all systems know which roles a user is a member of. These assumptions do not necessarily hold true for e-Research VOs. Instead different VOs will involve different sites, different individuals, different resources where a variety of roles are required signed by different authorities reflecting the different trust relationships. In this context, it is essential that roles specific to particular VOs can be defined and subsequently used for different access control scenarios where the access control decisions themselves are made at local resource providers.

Several implementations of RBAC models have been implemented and adopted by the Grid community. The two most prominent of these include the Virtual Organisation Membership Service (VOMS) [ref] which provides a centralised VO role definition model, and the Privilege and Role Management Infrastructure Structure Validation

(PERMIS) model [ref]. Furthermore, numerous standards have been defined that support a variety of access control models. The X.812 | ISO 10181-3 Access Control Framework standard [X812] defines a generic framework and terminology for secure access to a given domain resource by a remote requestor. These include the concept of a Policy Enforcement Point (PEP) and a Policy Decision Point (PDP).

It is important that these technologies are aligned with the middleware and the way in which the researchers themselves wish to work.

MORE…

3. E-EPIDEMIOLOGY RELATED PROJECTS AT NESC-GLASGOW

A variety of projects have been carried out at the NeSC in Glasgow over the past six years, largely focused on the development of distributed applications to support e-Science. In essence, e-Science or e-Research as it is often known is focused upon supporting collaborative research often crossing institutional boundaries. Fundamental to this is overcoming, i.e. making transparent to the end user, the heterogeneity of systems, e.g. different platforms used, different database technologies adopted, different schemata etc. To support this vision, different domains of research have their own requirements. The NeSC have focused in particular upon fine grained security and its application to the clinical and e-Health research domains, however we note that we are involved in projects in many other domains, e.g. electronics. We also note that we regard "e-" as enabling/empowering science and not simply electronic – an important distinction. Three projects in particular of relevance to this paper include: VOTES, SeeGEO and DAMES.

The Virtual Organisations for Trials and Epidemiological Studies (VOTES) project (www.nesc.ac.uk/hub/projects/votes) is a three-year, £2.8m project funded by the Medical Research Council (MRC) which has been investigating how best to leverage national and a range of other clinical data-sets to aid with the development and conduct of clinical trials and epidemiological studies throughout the UK. Focusing on three particular areas – patient recruitment, follow-up data collection and study management, a large number of resources, including the UK Biobank project (www.ukbiobank.ac.uk) have contributed to the facilitation of a reusable framework supporting a range of distributed applications involving all stages involved in a typical trial.

The SeeGEO project (http://www.edina.ac.uk/projects/seesaw/seegeo/) is funded by the Joint Information Systems Committee (JISC) and involves collaboration with the EDINA at the University of Edinburgh who provide a national geo-spatial data centre including offering services such as DigiMap which allows secure access to Ordnance Survey data which itself under copyright, and to the UK Borders data which provides access to 300 digitized boundary sets of the UK. Incorporating geo-spatial data is a common and essential aspect of epidemiological studies.

The Data Management through e-Social Science (DAMES) project (www.nesc.ac.uk/hub/projects/dames) has recently been funded by the Economic and Social Sciences Research Council (ESRC) to look at how a range of social science data can be seamlessly accessed, integrated and managed more effectively. One particular strand of DAMES that is relevant to this work is in the e-Health area. Differences and similarities between individuals, societal groups or populations more generally that are traditionally studied in social science, e.g. educational, occupational circumstances, ethnicity, lifestyle - are essential factors that need to be incorporated into future e-Health research. Relating e-Health resources to other social science data is essential to understand social factors and for example, their impact on the health of the nation.

2.1 Clinical Data-sets in Scotland

For the VOTES project referred to above, it has been necessary to source a variety of data-sets from the clinical infrastructure based in Scotland. These include the Scottish Care Information (SCI) Store (www.sci.scot.nhs.uk) and General Practice Administration System for Scotland (GPASS - www.gpass.scot.nhs.uk) which are widely deployed software solutions within the NHS for secondary care and primary care data sets respectively. GPASS is the predominant application used by over 85% of GPs throughout Scotland. The application has an offline mode that allows patient data to be input to the GPs laptop and is not dependent on location or connectivity. This primarily serves as a way for the GP to effectively organise their data electronically, but with it they also have the possibility to upload a subset of their patient data to an aggregated repository managed by their local authorities. SCI store is a regional clinical repository that holds details of many patient incidents, usually through primary or secondary care treatments. The repository is updated with patient records from individual GPASS systems and a range of associated secondary care data. For example, SCI store supports storage of a range of hospital data such as lab results, inpatient/outpatient summaries.

Both SCI store and GPASS provide amongst other things, front-end software to specific clinical databases. Following a prescribed schema definition from the Information Services Division (ISD – www.isdscotland.org) of the Scottish NHS, GPs build up individual data-sets about their own particular practice areas including their patient data and their administration for example. Subsets of this information are then uploaded to a central repository hosted by SCI Store. The data from the numerous SCI store installations across Scotland is then placed into centralised clinical data warehouses. In theory, this centralised regional model would seem ideal for providing a national database that only privileged health-care professionals can access. However in practice, the development and roll-out of the software across different regions of Scotland has resulted in different schemas being adopted reflecting local IT systems/practices and hence different “strains” of this technology now exist. The result of this is that the aggregated national data warehouse has to overcome issues such as the harmonisation of the various regional repositories. We note that the national central warehouses are currently under development. We also note that there are a range of other new technologies currently being rolled out across Scotland hence the heterogeneity of infrastructures and data resource will likely persist for the foreseeable future.

Scottish Morbidity Records (SMR) from ISD forms one of the most comprehensive clinical data repositories in the UK. The data sets are constructed in conjunction with the General Register Office (GRO) for Scotland. They consist of many data-sets, and for the purposes of the VOTES project, a subset of over 4 million anonymised records of the data has been provided. These have included:

- GRO Deaths, Jan 1996 onwards
- GRO Deaths, Jan 1980 – Dec 1995
- COPPISH SMR 01 Admissions, Apr 1997 onwards
- COPPISH SMR 04 Admissions, Apr 1996 onwards
- Historic SMR1, Jan 1981 – Mar 1997
- Historic SMR4 Discharges, Jan 1981 – Mar 1997
- SOCRATES Jan 1980 onwards

These broadly translate to information about the following clinical areas for hospitals throughout Scotland:

- Hospital discharges
- Psychiatric admissions and discharges
- Cancer registrations
- Deaths

These data sets include a variety of clinical information. As an example in Figure 1 we show the schema associated with the registry of deaths. This includes geographical

information, the cause of death and, if resulting from a disease, the associated disease codes. In the anonymised data sets that we have been provided, we have not been given with individual patient details, e.g. their name, however other information has been provided. Despite the fact that we do not have individual identified patient information, this kind of information is still extremely useful for developing a range of longitudinal and epidemiological studies as well as other wider questions. Is there a correlation between deaths due to cancer, ages and postcodes? Are there trends developing over geographical areas in different disease areas? Of course, these and other national statistics data sets such as Census data have been specifically established to help answer such questions. For many researchers, however a range of data linkage questions need to be answered which cannot often be addressed through national-level data resources – which provide aggregated statistical data sets. In such scenarios, linkage with other "silos" is needed. It is this niche which we have focused upon at NeSC.

2.2 Other Data Sets

The social sciences are awash with data. Innumerable surveys have been conducted by a multitude of social science researchers and centres. Some of the more prominent resources available to the social scientist in the UK include the UK Data Archive (www.data-archive.ac.uk), Economic and Social Data Service (www.esds.ac.uk), MIMAS (www.mimas.ac.uk), the Office of National Statistics (ONS - www.ons.gov.uk) and the Scottish Census Results On-Line (SCROL - www.scrol.gov.uk/). The SCROL and ONS resources incorporate the major source of statistical information in the UK related to a broad range of areas: from households, to health to occupational data sets amongst numerous others.

Numerous other international resources also exist such as the Council of European Social Science Data Services (CESSDA - www.nsd.uib.no/cessda) which offers a co-ordinated gateway to various national social survey micro-data data archives. Other more specialised social science resources also exist. Examples of these in the occupational research domain (as one of the many examples that might be selected) include: IPUMS (www.ipums.org/) which provides access to extensive volumes of census micro-data and specialist information associated with it, such as occupational information; EurOccupations (www.euroccupations.org/main/); the Grid-Enabled Occupational Data Environment (GEODE) resource (www.geode.stir.ac.uk) and the History of Work Information System (historyofwork.iisg.nl/).

Linking these data sets with other resources through any infrastructure, irrespective of the security considerations demands an awareness of the data standards in use. This in turn is greatly benefited from harmonisation of the data resources and standards adopted. The ONS is currently undertaking a data harmonisation project (www.statistics.gov.uk/about/data/harmonisation/) aimed at defining precisely such work within the UK. Numerous other complementary efforts also exist such as the Data Documentation Initiative (DDI) metadata standard (www.icpsr.umich.edu/DDI/) which provides a standard metadata framework widely used in the social sciences.

As noted previously, geo-spatial data is also fundamental to the success of many epidemiological studies. The EDINA geo-spatial data services provide one key resource used within the UK for a wide range of research. Other software solutions and associated data sets also now exist with Web 2.0 technologies and solutions such as GoogleMaps now with widespread adoption. One key aspect of the use of geo-spatial data in studies is the ability to layer information over geographical information. Technologies such as GoogleMaps provide simple lightweight solutions that allow a range of information to be overlaid onto a given map, e.g. APIs exist that allow a variety of information associated with particular postcodes to be presented. The EDINA DigiMap resource provides a variety of ways in which other information can be overlaid onto a given set of maps or geo-spatial co-ordinates. This includes the support for Web Map Service (WMS) which

responds to requests by creating map images of spatial data; Web Coverage Services (WCS) which allow access to the raw data which can then be used for further analysis or for portrayal if required, and Web Feature Sets (WFS) which allow to add a range of features over a given map set.

3. E-INFRASTRUCURES FOR EPIDEMIOLOGICAL STUDIES ACROSS SCOTLAND

To achieve a reusable infrastructure supporting a range of data federation scenarios in the e-Health domain demands simple (from the end user perspective) fine-grained security. The way in which the overall simple, user-oriented requirements have been addressed has been through the exploitation of the Internet2 Shibboleth technologies (shibboleth.internet2.edu). With this model end users requesting access to a specific resource, typically a portal through which clinical trials or epidemiological studies can be supported, is redirected to their home institution to authenticate. Once logged in to their home site, signed assertions confirming the user authentication at their home institution along with zero or more sets of digitally signed security tokens which the home institution is prepared to realise to that provider are sent. These tokens are then used to make fine grained authorisation decisions on the access to and usage of the contents of the portal. More details on these kinds of interactions, the associated protocols and how they can be supported for a range of e-Research areas are outlined in detail in [3-8]. The completed projects DyVOSE and GLASS and the on-going project SPAM-GP and VPman (see www.nesc.ac.uk/hub/projects for details) are also delivering technical solutions to support this process.

To understand the way in which e-Infrastructures for epidemiological studies can be supported we consider a typical scenario that builds upon a range of heterogeneous data sets and services crossing multiple domains. In this case we consider the slightly contrived research question: *what is the correlation between living adult males over 50 years of age who have had type-2 diabetes for 5 years and those employed in manual versus office jobs, i.e. does having the type-2 diabetes condition imply that those afflicted are more likely to be employed in manual or office jobs?* This kind of query can be used as the basis for a wide range of public health and epidemiological questions. Examples might be, is home care, primary care or secondary care hospital visits the best way to treat these patients so that they best recover? Where in Scotland are most patients with this condition found and how does this impact upon local policies on social expenditure and of course where should public money be best targeted to treat these men.

This kind of survey requires a body of information to be accessed, collected and filtered across many sites. The way in which the e-Infrastructure supports such kinds of scenarios is depicted in Figure 2.

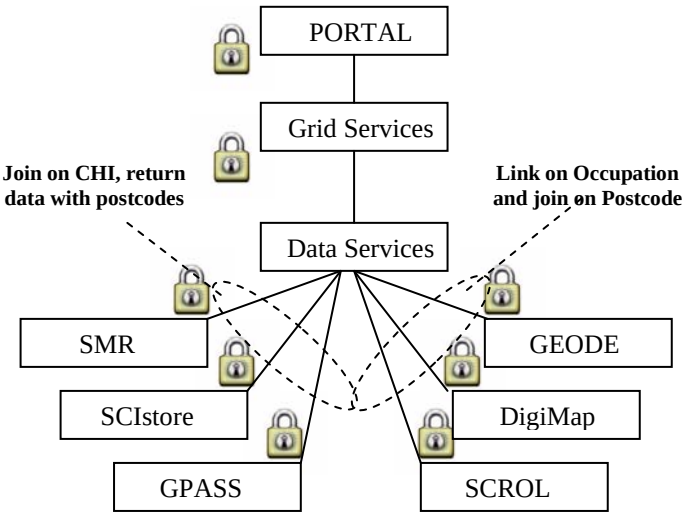

Figure 2: Architecture for Typical Epidemiological Study in Scotland

Here we see that the various data resources that need to be access are available from within a portal. This portal is itself implemented using GridSphere (www.gridsphere.org) a technology specifically designed to give user-friendly and lightweight access to distributed resources. The contents of the portal, i.e. the services and the data that they provide access to, exploit the digital credentials delivered through Shibboleth to enforce local authorisation decisions. The clinical and other data services themselves have been developed with a range of middleware including Grid technologies such as the Globus toolkit version 4.0 (www.globus.org) and the OGSA-DAI software version 3.0 (www.ogsadai.org.uk) which now supports distributed joins.

Each resource in the infrastructure has some form of trust relationship/policy with the portal and hence with the end user and the privileges that they present to the portal (via Shibboleth) - as indicated by the locks in the diagram. We note that this does not have to be explicit. Specific resources accessed via the portal can if needs be, pull other security tokens needed to make authorisation decisions. The above query itself is formulated through predefined client interfaces (portlets in the portal) – however we note that these can be parameterised. Thus for example, age ranges can be included or pull down lists used for queries for specific diseases, e.g. cancer, etc. A key part of our focus here is that the query itself cannot be extended, i.e. the roles presented to the portal are used to configure the query interface that the user is provided with. Hence there is no possibility to run any other queries.

The query generated by the authorised user within the portal is federated across the data resources accessible within the portal exploiting whatever remote interface/resource is offered and agreed within the context of the study. Thus for example, the query can result in a SOAP message used to interact with a remote/secure web service; or the query can result in a direct connection to a remote database if permitted, e.g. a jdbc connection to specific database established for the given study. The resultant clinical data sets from the SCI store, SMR data and the GRO death registry can be linked on the Community Health Index number (a unique identifier for patients in Scotland) where the generated query is coded for type-2 diabetes and the remaining patient information, i.e. those living with type-2 diabetes are recorded. We note also that a range of coding systems are in place in Scotland and beyond including International Classification of Disease version 9 (ICD9) and version 10 (ICD10) [9], as well as older Read coding schemes, HL7 (www.hl7.org) and SNOMED-CT [10] amongst others. Provided that this information is

known then the subsequent queries can be developed by the portal accordingly. These kinds of issues of data classification are described in more detail in [11].

In the case of the above query, the SMR data sets are securely accessed for male patients who have had type-2 diabetes for over 5-years. The resultant data sets themselves are filtered using the associated CHI numbers with the GRO data sets (to remove patients who have died from this or other conditions). The reduced set of patients are then joined through the CHI number with the data held in the SCI store and GPASS databases to extract a range of information including the treatment of the patients, the number of times that they have visited the hospitals and GP clinics, and the postcodes (or partial postcodes) of the individuals themselves. This results in a set of *anonymised* patients and their postcodes being generated. We note that this *anonymised* set raises an important point, since the end user epidemiologist does not need to know the individual patient information or even their CHI numbers. Rather the matching postcodes (or partial postcodes) and number of records will suffice in this case and this matching/joining is done by the back end data services.

In a similar vein, occupational identifiers can be used to link employee status (office or manual workers) from resources such as GEODE which provide such occupational classifications with census data existing within the SCROL data resource or ONS data for example. This results in generation of a collection of postcodes for the regional manual/office employment of individuals across Scottish regions. These postcodes are then joined with the postcodes from the clinical query to show how type-2 diabetes patterns relate to job patterns across Scotland where the linking information is the postcodes. The information on the cases of patients with type-2 diabetes, the employment patterns in those areas can then be overlaid onto the geo-spatial data resource offered for example by the DigiMap at EDINA.

The interface for querying the clinical data is shown on the left of Figure 3. The various security attributes used to restrict access to the various data resources and associated services are shown on the right hand side of Figure 3.

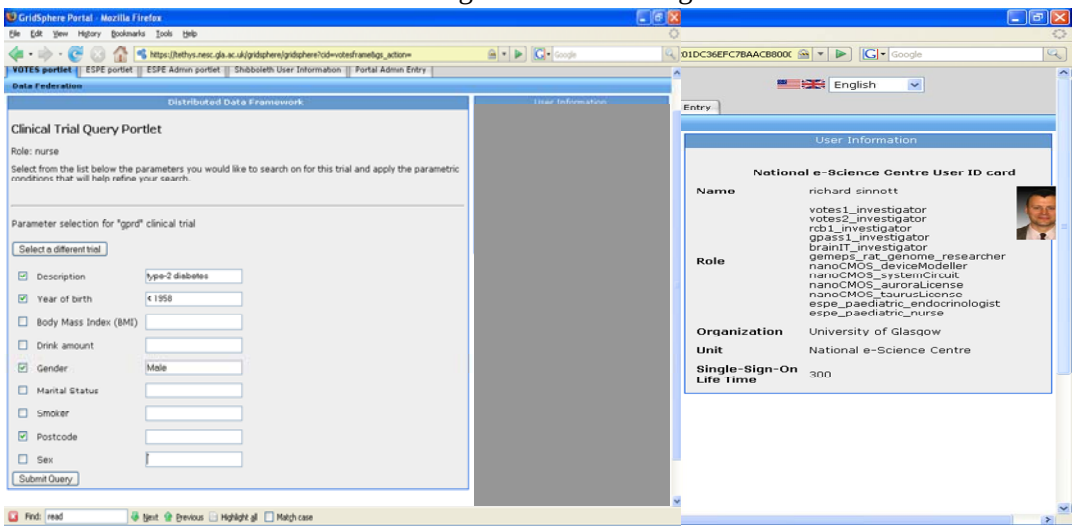

Figure 3: User-oriented Clinical Interface

4. CONCLUSION AND FURTHER WORK

This work is still very much under in progress, especially with regard to the epidemiological studies. We note that the idea of leveraging distributed health data is not a new one. However, as a global vision it is an incredibly powerful one – if clinicians and doctors can have immediate to access to patient records at the touch of a button, the

impact on the level of health-care that can be provided will be on a scale larger than political boundaries. Indeed we are currently involved in new projects [12,13] in the EU which are looking at taking these kinds of approaches into a truly international clinical and genetic data setting. In order to achieve this it is paramount that ethics and information governance issues are identified and rigorously addressed. This in turn demands a build-up of trust, and a common understanding of standard interfaces between the parties involved. Our experiences have shown that this has proven to be the largest obstacle in developing distributed, heterogeneous systems.

We also note that the scenarios outlined above are based upon detailed knowledge of the data sets and the way in which the users should be able to interact with them. Whilst possible for simple queries and studies, it is often the case that more expressive queries and statistical analysis should be supported. Within the DAMES project we wish to look at development and/or adaptation of workflow languages that allow for a variety of services to be accessed and queried in a variety of ways – and not simply be encoded with the queries in the portal directly. Similarly, we expect to incorporate statistical software used within this domain as the basis for analysis of returned results. This will require our addressing the issues of license management and access to commercial software. We are exploring these issues in other on-going projects in a variety of domains since it is widely accepted as one of the key limitations facing collaborative e-research. In particular we are looking at Shibboleth as the basic building block upon which such solutions can be based, e.g. in providing license information from remote identity providers.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the funding bodies for the various projects mentioned in the paper – specifically the UK Medical Research Council (MRC); Joint Information Systems Committee (JISC) and the Economic and Social Research Council (ESRC).

7. REFERENCES

[1] http://www.guardian.co.uk/bigbrother/privacy/

[2] Orwell, George (1949). Nineteen Eighty-Four. A Novel. New York: Harcourt, Brace & Co.

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

[4] Sinnott, R.O., Watt, J., Stell, A., Ajayi, O., Jiang, J., 2006, Single-Sign on and Authorization for Dynamic Virtual Organizations, *International Conference on Virtual Enterprises, (PRO-VE'06), Helsinki.*

[5] Lambert, P., Turner, K., Tan, L., Sinnott, R.O., Gayle, V., 2006, Distributed Occupational Information Resources Using OGSA-DAI, *UK e-Science All Hands Meeting, Nottingham UK.*

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

[7] Watt, J. Sinnott, R.O., Jiang, J., Stewart, G., Stell, A., Martin, D., Doherty, T., 2007, Federated Authentication and Authorisation for e-Science, *in Proceedings of APAC 2007 conference, Perth, Australia*.

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

[9] International Statistical Classification of Disease and Related Health Problems (ICD) - http://www.connectingforhealth.nhs.uk/clinicalcoding/classifications/icd_10, August 2007.

[10] International Health Terminology Standards Development Organisation, www.ihtsdo.org.

[11] Bowker G., Star, S.L., Sorting Things Out: Classification and Its Consequences, *Journal of Computer Supported Cooperative Work (CSCW), Springer Netherlands, Issue Vol. 10, No. 1, March, 2001*.

[12] Advanced Arterial Hypotension Adverse Event prediction through a Novel Bayesian Neural Network (AVERT-IT), EU FW7, January 2008, www.avert-it.org.

[13] Investigation of the molecular pathogenesis and pathophysiology of Disorders of Sex Development, EU FW7, January 2008, www.eurodsd.org.