GLAS-PPE/2010-07
26th October 2010

School of Physics and Astronomy
Experimental Particle Physics Group
Kelvin Building, University of Glasgow
Glasgow, G12 8QQ, Scotland
Telephone: +44 (0)141 330 2000 Fax: +44 (0)141 330 5881

e-Infrastructures supporting research into depression, self-harm and suicide

S.McCafferty (1), R.O. Sinnott (1), T. Doherty (1), J.P. Watt (1).

1 National e-Science Centre, University of Glasgow, Glasgow, G12 8QQ

Email: s.andrews@nesc.gla.ac.uk or t.doherty@physics.gla.ac.uk

Abstract

The Economic and Social Research Council (ESRC)-funded Data Management through e-Social Sciences (DAMES) project is investigating, as one of its four research themes, how research into depression, self-harm and suicide may be enhanced through the adoption of e-Science infrastructures and techniques. In this paper, we explore the challenges in supporting such research infrastructures and describe the distributed and heterogeneous datasets that need to be provisioned to support such research. We describe and demonstrate the application of an advanced user and security-driven infrastructure that has been developed specifically to meet these challenges in an on-going study into depression, self-harm and suicide.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

e-Infrastructures supporting research into depression, self-harm and suicide

S. McCafferty, T. Doherty, R. O. Sinnott and J. Watt

Phil. Trans. R. Soc. A 2010 **368**, 3845-3858
doi: 10.1098/rsta.2010.0142

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2010) **368**, 3845–3858
doi:10.1098/rsta.2010.0142

e-Infrastructures supporting research into depression, self-harm and suicide

By S. McCafferty, T. Doherty, R. O. Sinnott and J. Watt*

National e-Science Centre, University of Glasgow, Glasgow G12 8QQ, UK

The Economic and Social Research Council (ESRC)-funded Data Management through e-Social Sciences (DAMES) project is investigating, as one of its four research themes, how research into depression, self-harm and suicide may be enhanced through the adoption of e-Science infrastructures and techniques. In this paper, we explore the challenges in supporting such research infrastructures and describe the distributed and heterogeneous datasets that need to be provisioned to support such research. We describe and demonstrate the application of an advanced user and security-driven infrastructure that has been developed specifically to meet these challenges in an on-going study into depression, self-harm and suicide.

Keywords: Data Management through e-Social Sciences (DAMES); e-Science; National e-Infrastructure for Social Simulation (NeISS); Shibboleth; portal

1. Introduction

Depression will affect one in five persons in Scotland at some stage in their lives and is treatable in most cases (DAS 2009). Every day, around two persons die of suicide in Scotland (ChooseLife 2003). Most people contemplating suicide do not want to die; they simply want to stop the pain and difficulties they are suffering. In 2002, the Scottish Executive launched a 10 year strategy and action plan to prevent suicide in Scotland. Among the aims of the plan are to develop a framework to ensure that action is taken both locally and nationally to encourage people to seek help early and improve knowledge and awareness of 'what works' to prevent suicide. To try to understand what predisposes individuals to suffer from depression or to try to take their lives, researchers will typically want to look into those individuals' backgrounds (Borrell *et al.* 2002). For example, did individuals seek medical help prior to an episode or before being admitted for self-harming or indeed prior to taking their life? Were individuals previously in psychiatric care? What was their household composition? Were they married? What was their occupation and average household income (McLoone 1996; Rezaeian *et al.* 2006)? Did they excel in secondary education? Did they live with both parents (Boyle *et al.* 2005; Boyle & Exeter 2007)? Did they suffer from any ailments or have a history of drug taking? Is there a history of mental-health-related problems in their family? Did they live within the vicinity of a park or other recreational space? Are there any correlations between mental-health problems and ethnicity

*Author for correspondence (j.watt@nesc.gla.ac.uk).

One contribution of 16 to a Theme Issue 'e-Science: past, present and future I'.

or between mental-health problems and obesity (Gunnell *et al.* 2007)? All of these and many more factors are potentially significant and may well have a direct impact on depression, self-harm and suicide that, in turn, can subsequently be used to identify risk factors and hence to mitigate against these risks.

However, the datasets that allow such questions to be answered exist across many locations and are typically held by many data providers, each with their own strict access and usage policies and often serving different communities, for example, the clinical sciences, the social sciences, the environmental and geospatial sciences. The e-Health component of the Economic and Social Research Council (ESRC)-funded Data Management through e-Social Sciences (DAMES 2008) project aims to provide a virtual research environment through which tailored, secure access to a variety of distributed e-Health data resources is supported in a seamless and secure manner, where all data providers are themselves autonomous, and data access, usage and linkage by authorized individuals are directly controlled throughout.

This paper gives an overview of the key datasets required for such research and the challenges interfacing with the associated data providers. We also describe the technologies and infrastructures we are using to tackle these challenges and demonstrate their application in a case study crossing the clinical, social and geospatial data divide.

2. Depression, self-harm and suicide-related data in Scotland

The datasets involved in undertaking e-Health research are extremely heterogeneous and often cross multiple research domains. Research into depression, self-harm and suicide ideally demands seamless linkage and access to datasets crossing multiple organizational boundaries (Sinnott *et al.* 2009). The Grid and e-Science models of dynamic provisioning and support of ad hoc virtual organizations (VO) are compelling requirements in many domains, but when dealing with organizations such as the UK National Health Service (NHS), such dynamism where resources are discovered automatically is simply not tenable. Rather, access to and usage of these datasets demands that rigorous security mechanisms and ethical agreements are in place before data can be accessed and used. This, in turn, requires that prior agreements are in place with the data providers before permission to access the data can be granted. Typically, this is achieved through an independent ethic review where Caldicott Guardians and Patient Advisory Groups review the justification for data access and usage. As well as the scientific justification, they also need to be convinced of data privacy, confidentiality and long-term information governance.

e-Science fundamentals elicited from many e-Science projects have provided an understanding of common problems faced within different disciplines, and by applying this knowledge to the health sciences, we can address data support and infrastructure issues that have tormented the healthcare profession for decades. The challenges of accessing and using ‘live’ clinical data for research purposes has been considered extensively by many groups and research projects (VOTES 2008; AvertIT 2010; CLEF 2010; NeuroGrid 2010; SHIP 2010) and is an integral part of the UK Connecting for Health Research Capability Programme (NHS Connecting for Health 2006). Key to successful collaboration with organizations such as the NHS is trust: trust of software, people and processes of all aspects of data management more generally. This trust has to extend across all organizations involved in collaboration. Defining and enforcing trust relationships in an e-Health context has been considered by Ajayi *et al.* (2007).

For the DAMES e-Health research environment, we have identified several key data providers and data resources that need to be provisioned for depression, self-harm and suicide-related research.

(*a*) *Clinical data relevant for depression-related research in Scotland*

Scottish Morbidity Records (SMRs) from the NHS Information Services Division (NHS National Services Scotland 2009*a*) form one of the most comprehensive clinical-data repositories in the UK. The datasets are constructed in conjunction with the General Register Office for Scotland (Directgov 2010). The datasets capture an almost complete medical history of the Scottish population, often going back over 40 years.

SMR datasets are maintained and regularly updated from Scottish-wide hospitals with feeds from primary-care sites. The data itself is structured and catalogued in a secure and centralized data warehouse maintained by the NHS. Key to DAMES e-Health research are: the hospital admissions datasets (SRM01); mental health-related datasets (SMR04); and death-related datasets (SMR99). Other datasets also exist, however, for example, cancer-related datasets (SMR06).

The SMR datasets contain, among other things, patient identification and demographic data, episode-management data, general clinical data and development data on all hospital admissions and registered deaths. Gaps in suicide research were set out in a report commissioned by the Scottish Government (McLean *et al.* 2008). In collaboration with DAMES public-health researchers, we have identified key variables relevant for depression, self-harm and suicide research in the SMR01 patient admissions. These include sex, marital status, type of admission, discharge code, diagnosis, ethnic group, discharge type, discharge to, main conditions and operations. For the SMR04 and SMR99 datasets, key variables include occupation, previous psychiatric care, immediate source of referral, residency immediately prior to admission/discharge, injuries and/or poisoning precipitating admission, discharge type, disposal on discharge, after care, length of stay, post mortem and cause of death.

The main variables used to ascertain if an individual has suffered from depression, self-harm or indeed committed suicide are the diagnosis on admission and the cause of death. These variables are coded using the International Classification of Diseases (ICD) code, as compiled by the World Health Organization, which provides the international standard diagnostic classification for all general epidemiological, and many health management purposes and clinical use terms. Its current version is ICD-10 (World Health Organization 2003).

Self-harm is defined as the intentional infliction of tissue harm, modification or poisoning without suicidal intent. Drug and alcohol abuse are classified as self-harm along with, perhaps more obvious, methods such as cutting or eating disorders. Drug and alcohol abuse are often used as methods of escaping reality and can be a consequence of underlying mental-health difficulties. The list of ICD codes relevant to depression, self-harm and suicide is complex and complicated further by different versions of the ICD coding present in SMRs and other clinical records. A major revision of the SMRs took place in 1996 when ICD-10 was introduced to replace ICD-9. Further complications arise in that suicide is often not recorded (coded) as the cause of death if there is any doubt as to what caused an individual's death and/or if it was intentional due to the implications this can cause. For instance, many insurance companies will often not pay out after a suicide. This means that the list of codes required for researchers must include things such as accidental poisoning as well as intentional poisoning. ICD codes are, in themselves, typically not self-explanatory and support is often needed for researchers, for example, metadata and descriptions of codes. To support this process, tools supporting the translation of ICD-9 to ICD-10 and vice versa are highly desirable; however, we note that it is often the case that this translation is not straightforward because it can require clinical expertise and interpretation.

Across Scotland, numerous other clinical datasets exist, including Prescribing Information System (PIS; NHS National Services Scotland 2004), which contains prescription data issued across Scotland. Of particular relevance to mental-health-related research is prescription information related to anti-depressants and benzodiazepines used to treat anxiety or induce sleep. Primary care (General Practitioner; GP) datasets are essential for mental-health research. In Scotland, the vast majority of these datasets are held in the General Practice Administration System for Scotland (GPASS; NHS Scotland 1984). GPASS is used by over 800 general practices across Scotland and is currently used to manage four million Scottish patients' primary-care records.

The Scottish Care Information Store (SCIStore; NHS National Services Scotland 2009*b*) is a data repository that stores patient information at the Health Board level. SCIStore contains information such as hospital admissions, hospital discharges and also laboratory tests and reports.

The integration and usage of primary-care and secondary-care software systems and data resources across Scotland for a range of clinical trials and epidemiological studies were undertaken as part of the Medical Research Council (MRC)-funded Virtual Organizations for Trials and Epidemiological Studies project (Stell *et al.* 2007; Sinnott *et al.* 2008). Key to this is the existence of the Community Health Index (CHI) number. This unique identifier is composed, in part of the individual's date of birth and sex. Since 2006, the CHI number has been rolled out across Scotland and is used to identify, track and link patient records, and more generally to support patient health.

It is noted that all clinical-data resources across Scotland (including SMRs) incorporate some form of geospatial information, whether it is referring to GP practices, hospitals and/or for the patients themselves. The geospatial data not only allow us to understand local, regional and national level datasets, but also provide an opportunity to link and display data using the geospatial aspect, that is, on maps.

At present, the National e-Science Centre at the University of Glasgow has direct access to a dataset containing 3 719 206 SMR01 hospital-admission records, 241 599 SMR04 mental-health discharge records, 171 167 SMR06 cancer-registration records and 173 616 SMR99 death records. Through the on-going work of the Wellcome Trust-funded Scottish Health Informatics Platform

(SHIP) project, live access to data maintained by the NHS is currently being pursued. This paper is based primarily on datasets provided by the NHS Information Services Division (ISD) to be used for research purposes.

(*b*) *Social science data relevant for depression-related research in Scotland*

There are a wealth of social science datasets currently in the UK (and internationally) that have a bearing on mental-health-related research. The UK Census dataset and the British Household Panel Survey (BHPS; UK Data Archive 2003) are two of the largest and most important data resources in the UK. The BHPS is funded by the ESRC and provides an annual survey of around 10 000 households across England, Scotland and Wales. As a panel survey, the same individuals are surveyed yearly and BHPS is thus suited to longitudinal studies. The first survey took place in England in 1991. Since then, the BHPS has grown and now incorporates Scotland and Wales. The core questionnaire covers a broad range of social science and policy interests of particular interest to mental-health research. These include household composition, education/training, general health and the usage of health services. The data is itself available from the UK Data Archive to registered users, that is, those users who have signed up to special conditions on access and usage. Different licences exist that give access to datasets at different levels of geospatial resolution.

The UK Census is a count of all people in the UK and takes place every 10 years, with the next Census scheduled for 2011. The Census provides unique population statistics. Census records for Scotland are available from the General Register Office for Scotland (GROS) and from Mimas (University of Manchester 2010) at the University of Manchester via CASWEB (Mimas 2010). CASWEB also provides access to Census data for England and Wales.

In terms of mental-health-related research, both the Census and the BHPS include numerous tables and variables of direct relevance. These include household composition, for example, whether a household was made up of pensioners, married couples, married couples with children, co-habiting couples, lone parents or other household composition; general health variables; occupational variables and lifestyle variables.

Access to both the Census and BHPS is available to registered users via the UK Access Management Federation (see JANET 2010).

(*c*) *Geospatial data relevant for depression-related research in Scotland*

Geospatial information is key to both understanding and visualizing trends in data. Identification of clusters or patterns in data over particular time periods is crucial to mental-health-related research. The EDINA (University of Edinburgh 2010) data provider makes available a wide variety of geospatial data resources and services to researchers. The UKBORDERS resource provides digitized boundary datasets and geographical look-up tables for the UK offering local-authority-level datasets and allows us to explore the way these boundaries have changed over time. The data itself is under license from the UK Ordnance Survey and, as with the Census and BHPS datasets, can be downloaded subject to licensing agreements being in place on use and re-use of the data. The data can be visualized and used through a variety of geographical information system (GIS) software packages. Of particular importance are geospatial vector data formats (ESRI 2010), for example, 'shapefiles'.

A shapefile refers to a collection of files associated with a main shapefile, representing geometric points, lines and polygons. A shapefile representing UK Census output areas will provide geometric information that allows us to plot individual output areas as polygons on a map or image. The shapefiles provided by UKBORDERS have additional attributes that support identification of polygons as Census output areas or other geographical identifiers if selected. By combining the data we want to associate to known output areas, it is possible to use a shapefile to visualize the data on maps.

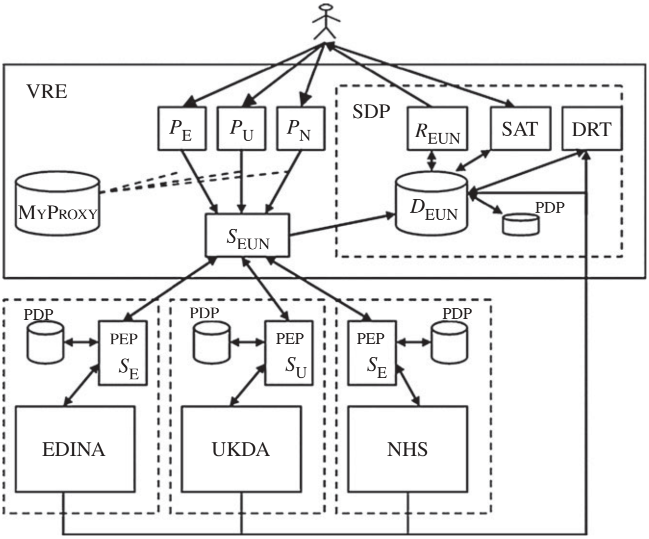
This data resource can also be accessed through the UK Access Management Federation.

3. Security-oriented Data Management through e-Social Sciences e-Infrastructure

The architecture currently realized to support the e-Health research of DAMES is outlined in figure 1. This model provides a secure environment through which a variety of distributed data services and associated datasets from a known set of autonomous providers to a collection of known and trusted (authorized) users potentially from multiple sites with their own specific access and usage privileges are realized. These datasets are brought back (subject to security constraints being satisfied) and placed in a secure data-storage environment, often referred to as a secure data enclave or a virtual safe setting.

To support the user-driven access and usage of the DAMES e-Health environment, it has been provisioned as a secure resource offered as part of the UK's Access Management Federation, that is, it is made available as a SHIBBOLETH-protected resource. An underlying public key infrastructure (PKI) associated with the UK Federation is used to establish a secure connection between the end user and the portal itself. However, as discussed previously, it is essential that not all users or sites from the UK Access Management Federation are able to access the resources made available. Rather only recognized sites and users with specific privileges as agreed as part of the VO should be able to access the DAMES e-Health resources.

The DAMES services themselves are exposed to end users through Java Specification Request (JSR)-168 compliant (Abdelnur & Hepper 2003) portlets (P_{E}, P_{U}, P_{N}) within a Liferay web portal, which in turn is protected by the INTERNET2 SHIBBOLETH software (Cantor 2005) configured for the UK Access Management Federation. Using the Open Middleware Infrastructure Institute (OMII)-UK Security Portlets simplifying Access to and Management of Grid Portals (SPAM-GP) suite of tools (Watt *et al.* 2009), user attributes from the Security Assertion Markup Language (SAML) assertion provided by a SHIBBOLETH Identity Provider (IdP) are used to filter the user's view of the portal, allowing access to a particular portlet based on the correct security attribute being received. These attributes are text strings that may be stored in the portal framework and briefly describe the functionality they enable (e.g. DAMES_wikiAdmin). A variety of portlets have been created for each of the services accessible through the portal (S_{E}, S_{U}, S_{N}).



P_E	portlet for EDINA
P_U	portlet for UKDA
P_N	portlet for NHS
S_E	service for EDINA
S_U	service for UKDA
S_N	service for NHS
S_{EUN}	secure data merging/anonymous service
D_{EUN}	secure data-storage environment
R_{EUN}	secure data results service
PEP	policy enforcement point
PDP	policy decision point
SAT	statistical analysis tools
DRT	disclosure risk-analysis tools
SDP	secure data playground
VRE	virtual research environment

Figure 1. DAMES e-Health architecture.

To support this filtering, the Content Configuration Portlet (CCP) of the SPAM-GP suite of tools has been developed. The CCP is a portal framework module that scans the SAML assertion for user identifiers which can be used to automatically create or locate a portal account, noting that these user identifiers are a requirement of the portal framework and not of the services. For example, when creating an account, the Liferay Portal framework demands that the user's first name, surname and email account are available within the SAML assertion in order for it to perform the account discovery or initialization. The ability of an IdP to provide this information cannot always be relied on, which is discussed in the final paragraph of this section.

All the DAMES portlets require access to remote services and data resources, and each of these in turn may have their own specific security requirements and policies in place (policy enforcement points (PEPs) and policy decision points (PDPs) in figure 1). A PEP is an application that intercepts calls to protected functions in order to enforce some security policy on it, and the PDP is the application that informs the PEP if a user has satisfied the appropriate access conditions. These will typically be in the form of authorization checks on access and usage, as well as on data-release policies. Using the Privilege and Role Management Infrastructure Standards (PERMIS) infrastructure (Chadwick & Otenko 2002), licences may be issued to users in the form of X.509 Attribute Certificates allowing fine-grained, policy-controlled access to particular sets of data. For databases without their own external authorization capabilities, a security gateway web service realized using the Globus Toolkit 4 was created to allow the PERMIS infrastructure to protect the existing data resources without any changes being needed to their own local functionality. In this case, portlets contact this service directly instead of the remote data-service database, and the data provider is thus free to configure the policy of the gateway service to meet their own requirements, and through the SPAM-GP Attribute Certificate

Portlet (ACP), they have direct control over the issuance of licences to users. Such a scheme was first tested as part of the Secure Access to Geospatial Services (SEE-GEO) project (Higgins *et al.* 2009), which provided services linking UK Census data provided by Mimas with geospatial boundary data from EDINA.

However, even with these authorization solutions in place, we recognize that data providers need to be aware of the context of their data and how it is being used/analysed and/or linked with other datasets. To allow all providers to be in complete control of their datasets and their derived (linked) datasets, the secure data playground (SDP) in figure 1 allows for a range of statistical analysis and data disclosure tools to be incorporated. These can be used to ensure that statistical data disclosure risks are minimized, for example. It is only when all providers have signed off on the derived dataset, following whatever data disclosure protocol is valid for the particular project, that the final results of the combined EDINA, UK Data Archive (UKDA) and NHS-related datasets are available for usage by authorized individuals.

Where possible, the flow of user information from Shibboleth, through the portal framework, and out to the remote database services and the secure storage environment has been implemented to avoid any user interaction beyond the initial Shibboleth login. An exception being the authentication to the Globus gateway service, which requires a proxy credential based on a National Grid Service (NGS) UK e-Science certificate (Jensen 2003). The overall experience of non-information technology savvy end users with digital certificates is not a positive one (Beckles *et al.* 2006). Effective use of PKIs requires a non-trivial degree of understanding in order for their use not to pose a security risk. Examples from the past include the sharing of private keys, storing of unencrypted keys and sharing of encryption passphrases. A temporary solution to this is the issuance of long-life proxies, which itself is not a recommended practice. However, through the use of a well-controlled MyProxy server (Novotny *et al.* 2001), the risk of this practice is mitigated by securely storing the long-lived proxy and using it to generate the safer short-lived proxies. A portlet interface to a MyProxy server was created to allow a user to upload their proxy, but other infrastructures such as Shibboleth Access to Resources on the NGS (SARoNGS; JISC 2008) have the potential to provide a more automated approach to grid credential management in the future (SARoNGS currently only allows creation of credentials for job submission through the NGS portal).

As institutional IdPs do not allow arbitrary user roles to be asserted, to retain full control of the SAML assertion required for access to the DAMES infrastructure, a custom IdP was created that was loaded with the DAMES user roles. The National e-Science Centre (NeSC) Glasgow has been involved in a case study exploiting the University of Kent's Shintau (Shib-Grid Integrated Authorization) software (Chadwick & Inman 2009), which is a standards-based extension to the Shibboleth infrastructure using PERMIS and a customized version of the Liberty Alliance Discovery Service. Shintau aims to provide users with more control over the attributes received by Service Providers (SPs) by granting users the ability to link two or more IdPs together using a Linking Service (LS) that merges the attribute assertions from each linked IdP. This allows roles, attributes or licences that cannot be provided by an institutional IdP to be extracted from other IdPs and presented to the SP in one merged attribute set. The LS stores the persistent identifiers asserted by every IdP to link the user's attributes together without retaining any identifying user information. The SHINTAU-enabled IdP is specially configured to release attributes to the LS on demand, and this assertion is forwarded to the PERMIS service running on the SP for authorization decisions. The case study showed how SHINTAU could be used to grant access to the DAMES portal provided the user successfully authenticated against an institutional IdP and used the LS to link two other IdPs representing the UK Data Archive Census Registration Service and the EDINA UKBORDERS user licence database. Based on user feedback from the case study, the SHINTAU software will be refined and submitted to Internet2 as a candidate for inclusion with future SHIBBOLETH versions.

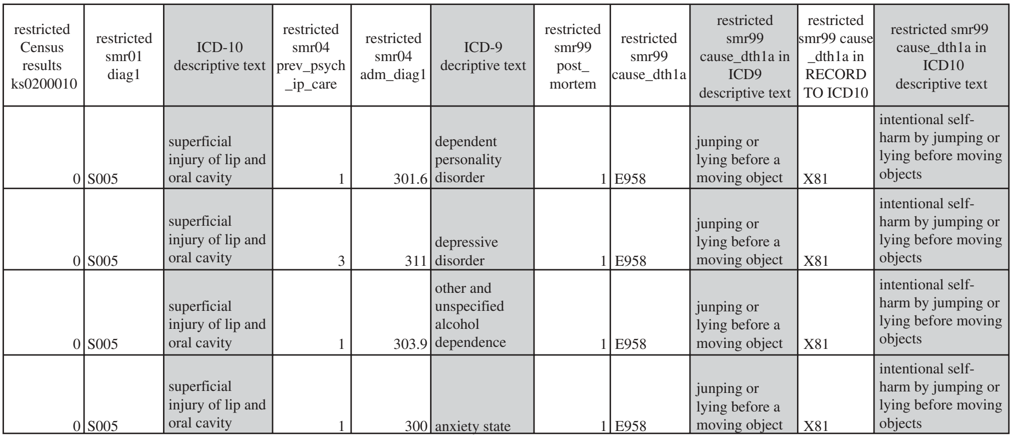
It is important to stress here that none of the *live* data-provider datasets are directly accessible to the end users through the DAMES e-Infrastructure. Rather in the most conservative case, they can only be accessed/analysed through a predefined collection of targeted tools offering specific (fixed) interfaces through which data processing and analysis can take place. Thus, for a particular case study, this might be a predefined library of statistical analysis functions in R, STATA, SPSS, etc. that are offered through a portlet interface which allows a pre-determined set of coding and analysis capabilities, for example, recoding routines to take local geospatial datasets (postcodes, etc.) from results datasets and map these onto larger local-authority level coordinates, or for recoding specific variables according to particular health classifications. This point is key to the work because it ensures that data are kept in a secure setting and targeted tools/interfaces are the only mechanism through which users are able to interact with the data. The users themselves are not able to perform arbitrary analysis of the data, but are restricted to a fixed set of analysis possibilities that have been agreed as part of the process in establishing the virtual collaboration between the end users and the data providers.

4. Depression, self-harm and suicide-related case study

To demonstrate the applicability of the DAMES e-Infrastructure, we show how seamless, secure access to a range of data resources is supported. This includes secure access to a subset of the Census data (available through Mimas CASWEB), a targeted subset of SMR datasets (SMR01, SMR04, SMR99) provisioned by the NHS ISD and exploitation of geospatial datasets (shapefiles) made available through EDINA. The specific case study we present here illustrates the process of secure access to and usage of security-oriented data to support mental-health research.

A key scenario for mental-health research is understanding the history of a particular suicide. In the first instance, it is necessary to identify the particular ICD-9/ICD-10 codes used in the SMR99 datasets to discover those individuals whose death can be attributed to self-harm/suicide. Once identified, knowing which of these individuals had been in some previous form of psychiatric care (SMR04) and/or were admitted to hospital for self-harm and/or attempted suicide is essential.

To enable one or more datasets to be joined on ICD codes or to convert results set to one code version, targeted recoding is required. The translation is typically not straightforward, as there is no one-to-one relationship between



restricted Census results ks0200010	restricted smr01 diag1	ICD-10 descriptive text	restricted smr04 prev_psych _ip_care	restricted smr04 adm_diag1	ICD-9 decriptive text	restricted smr99 post_ mortem	restricted smr99 cause_dth1a	restricted smr99 cause_dth1a in ICD9 descriptive text	restricted smr99 cause _dth1a in RECORD TO ICD10	restricted smr99 cause_dth1a in ICD10 descriptive text
0	S005	superficial injury of lip and oral cavity	1	301.6	dependent personality disorder	1	E958	jumping or lying before a moving object	X81	intentional self-harm by jumping or lying before moving objects
0	S005	superficial injury of lip and oral cavity	3	311	depressive disorder	1	E958	jumping or lying before a moving object	X81	intentional self-harm by jumping or lying before moving objects
0	S005	superficial injury of lip and oral cavity	1	303.9	other and unspecified alcohol dependence	1	E958	jumping or lying before a moving object	X81	intentional self-harm by jumping or lying before moving objects
0	S005	superficial injury of lip and oral cavity	1	300	anxiety state	1	E958	jumping or lying before a moving object	X81	intentional self-harm by jumping or lying before moving objects

Figure 2. A sample of the ICD10 dataset.

different code versions. To address this, we have used internal project expertise to provide mapping and translation rules specific to mental-health and suicide ICD codes. These have been realized through a conversion tool (portlet) made available through the portal. We have stored the ICD9/ICD10 mapping in a relational database and could easily add further mapping to newer versions of the ICD when they become available, provided we still have the expertise on hand to complete the translation. To support this process and since ICD codes are, in themselves, not self-explanatory, we allow textual metadata to be added to descriptions of all the ICD9 and ICD10 codes. A sample of this data can be seen in figure 2.

In figure 2, we see that the first patient was admitted to hospital with a superficial injury to their lip/oral cavity (SMR01 ICD10 coding). This patient was also referred to psychtaric care (SMR04 ICD9 coding) where they were diagnosed with a dependent personality disorder. Finally, this patient committed suicide by intentional self-harm by jumping or lying before moving objects. In this last case, the recoding of the ICD9 code to ICD10 code was necessary. The actual linkage and tracking of patients between different SMR datasets is achieved using the CHI number.

In the current realization, authenticated and authorized individuals (achieved through Shibboleth, the UK Access Management Federation and the DAMES-specific Attribute Authority) are presented with tailored access to SMR01, SMR04, SMR99 and Census data resources via specifically targeted JSR-168 compliant portlets. A fifth portlet has been realized to allow users to join selected data from the different data resources. In the future, we expect this portlet to offer an interface to the statistical linkage and analysis tools identified in figure 1.

The specific portlets and data a user is given access to is determined by the roles they have previously been assigned. In the current realization, we have created a variety of portlets and roles specific to each interface, e.g. *DAMES_SMR01_adv* and *DAMES_SMR01_basic*. These allow users to select from a known subset of variables related with SMR01 hospital datasets

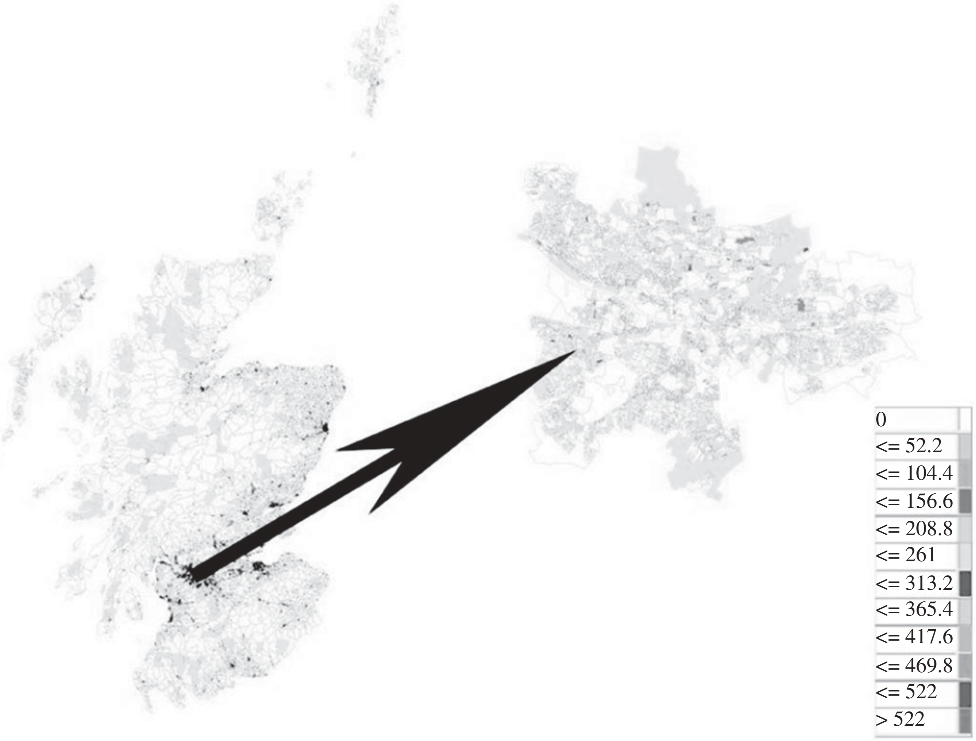


Figure 3. Map returned by the DAMES results. The shading/legend represents number of all mental-health-related hospital admissions across Scotland.

identified as being relevant to mental-health-related research. We use either the patient identifier or the CHI number to join the SMR datasets and then join these to the Census data using the Census output area as determined from the patient's home address. We check that the results set returns more than a minimal number of individuals and subsequentially strip geographical references from the results set returned to the user—this is the most basic way of avoiding disclosure of individuals. The results are then saved to Comma Separated Value (CSV) format in the secure data playground. CSV was chosen since it is compatible with most statistical packages used by social scientists.

A further portlet has been developed that allows the shapefiles provided by EDINA to be used with a web service developed in the SEE-GEO project to plot the results on maps. An example of the resultant datasets is shown in figure 3. This image depicts the total of all mental-health-related hospital admissions in Scotland (left) and the distribution of mental-health-related hospital admissions in the Greater Glasgow region (right). It is equally possible to overlay multiple datasets across such maps, thus contrasting average income and suicide information, etc. Work is on-going to produce such multi-variate images.

To create the image map, the data from the chosen shapefile is used to create a list of named Census output areas. By marrying up the SMR data values to the Census output areas, with the co-ordinates provided by the shapefile, we are able to plot the image colouring the individual polygons that correspond to the Census output areas, according to the values from the data results. This is a computationally expensive exercise however, and we are currently exploring use of larger scale high-performance computing facilities for this purpose.

Key to the success of this design is that throughout the whole process, the user does not need to know where the data resources lie and is hidden from the complexity of the security mechanisms that are enabling them to access multi-disciplinary research datasets. Our portlet user interfaces are dynamically generated from information provided to us by the data providers. Any future updates/changes to the data schemes can easily be accommodated with minimal effort, therefore making the systems sustainable. The JSR-168 portlets we have developed can easily be configured to access different datasets, and the security model is transferrable to different domains. Indeed, this model is currently being used in numerous security-oriented projects at NeSC Glasgow (NeSC Glasgow 2010), including the National e-Infrastructure for Social Simulation (NeISS 2010).

5. Conclusions

This paper has shown how the Grid and e-Science vision of seamless, secure access to inter-disciplinary research data and resources can be realized. The datasets and resource providers are especially complex, and environments that allow us to tackle this complexity are urgently required. This has to go beyond the individual silo models currently being pursued. e-Research is inter-disciplinary, and multiple organizations and numerous researchers will always need to be involved in understanding and tackling e-Health problems such as mental health and suicide. While this work is still on-going, we believe that the models and solutions put forward offer a step change in such a way that research can be conducted in the future.

One aspect of this work that we are currently refining is the need for push-oriented federated queries. Many organizations, such as the NHS, are loathe to open their firewalls to incoming connections from the internet. To accomodate such scenarios, we are developing a pull-oriented solution (Stell *et al.* 2009) as part of the Scottish Health Informatics Platform. This model is currently being rolled out across the NHS in Scotland.

The DAMES project is funded by the ESRC. We gratefully acknowledge their support as well as the contributions made by the other partners in the DAMES project at the University of Stirling. We also acknowledge the numerous other funders and projects who have contributed to the security aspects of this paper including the EPSRC, JISC and the MRC.

References

Abdelnur, A. & Hepper, S. 2003 Java portlet specification, version 1.0. See http://jcp.org/en/jsr/detail?id=168.

Ajayi, O., Sinnott, R. O. & Stell, A. 2007 Trust realization in multi-domain collaborative environments. In *6th IEEE/ACIS Int. Conf. on Computer Information Science 2007* (*ICIS 2007*), pp. 906–911. Silver Spring, MD: IEEE. (doi:10.1109/ICIS.2007.187)

AvertIT. 2010 Advanced arterial hypotension adverse event prediction through a novel Bayesian neural network. See http://www.avert-it.org/.

Beckles, B., Coveney, P. V., Pickles, S. M., McKeown, M., Brooke, J. M., Ryan, P. Y. A. & Abdallah, A. E. 2006 A user-friendly approach to computational grid security. In *Proc. 5th UK e-Science All Hands Meeting, Nottingham, UK, 18–21 September 2006*, pp. 473–480.

Borrell, C., Rodrguez, M., Ferrando, J., Brugal, M., Pasarín, M., Martnez, V. & Plasncia, A. 2002 Role of individual and contextual effects in injury mortality: new evidence from small area analysis. *Injury Prevention* **8**, 297–302. (doi:10.1136/ip.8.4.297)

Boyle, P. & Exeter, D. 2007 Does young adult suicide cluster geographically in Scotland? *J. Epidemiol. Comm. Health* **61**, 731–736. (doi:10.1136/jech.2006.052365)

Boyle, P., Exeter, D., Feng, Z. & Flowerdew, R. 2005 Suicide gap among young adults in Scotland: population study. *Br. Med. J.* **330**, 175–176. (doi:10.1136/bmj.38328.559572.55)

Cantor, S. 2005 Shibboleth architecture: protocols and profiles. See http://shibboleth.internet2.edu/docs/internet2-mace-shibboleth-arch-protocol s-latest.pdf.

Chadwick, D. W. & Inman, G. 2009 Attribute aggregation in federated identity management. *IEEE Computer* **42**, 33–40. (doi:10.1109/MC.2009.143)

Chadwick, D. W. & Otenko, A. 2002 The PERMIS X.509 role based privilege management infrastructure. *Future Generation Comp. Sys.* **19**, 277–289.

ChooseLife. 2003 The national stategy and action plan to prevent suicide in Scotland. See http://www.chooselife.net/Statistics/Overview.asp.

CLEF. 2010 Clinical e-Science Framework. See http://www.clinical-escience.org/.

DAMES. 2008 Data Management through e-Social Sciences. See http://www.dames.org.uk.

DAS. 2009 Depression Alliance Scotland. See http://www.dascot.org/.

Directgov. 2010 General Register Office for Scotland. See http://www.gro-scotland.gov.uk/.

ESRI. 2010 Environment Systems Research Institute. See http://www.esri.com/.

Gunnell, D., Hart, C. L., Hole, D. J., Lawlor, D. A. & Smith, G. D. 2007 Body mass index in middle life and future risk of hospital admission for psychoses or depression: findings from the Renfrew/Paisley study. *Psychol. Med.* **37**, 1151–1161.

Higgins, C., Koutroumpas, M., Sinnott, R. O., Watt, J., Doherty, T., Hume, A. C., Turner, A. G. D. & Rawnsley, D. 2009 Spatial data e-infrastructure. In *Proc. 5th Int. Conf. on e-Social Science, Maternushaus, Cologne, 24–26 June 2009*, pp. 22–29.

JANET. 2010 The UK Access Management Federation for Education and Research. See http://www.ukfederation.org.uk/.

Jensen, J. 2003 The UK e-Science certification authority. In *Proc. 2nd UK e-Science All Hands Meeting, Nottingham, UK, 2–4 September 2003*, pp. 336–369.

JISC. 2008 Shibboleth access to resources on the National Grid Service. See http://www.jisc.ac.uk/ whatwedo/programmes/einfrastructure/sarongs/.

McLean, J., Maxwell, M., Platt, S., Harris, F. & Jepson, R. 2008 Risk and protective factors for suicide and suicidal behaviour: a literature review. See http://www.scotland.gov.uk/Publications/2008/11/28141444/2.

McLoone, P. 1996 Suicide and deprivation in Scotland. *Br. Med. J.* **312**, 543–544.

Mimas. 2010 Casweb. See http://casweb.mimas.ac.uk/.

NeISS. 2010 National e-Infrastructure for e-Social Science Simulation. See http://www.neiss.org.uk/.

NeSC Glasgow. 2010 Project webpages. See http://www.nesc.ac.uk/hub/projects/.

NeuroGrid. 2010 Grid technology for neuroscience. See http://www.neurogrid.ac.uk/.

NHS Connecting for Health. 2006 UK Connecting for Health Research Capability Programme. See http://www.connectingforhealth.nhs.uk/systemsandservices/research.

NHS National Services Scotland. 2004 ePharmacy. See http://www.psd.scot.nhs.uk/professionals/pharmacy/epharmacy.html.

NHS National Services Scotland. 2009*a* Information Services Division Scotland. See http://www.isdscotland.org/.

NHS National Services Scotland. 2009*b* Scottish Care Information Store. See http://www.sci.scot.nhs.uk/products/store/store_main.htm.

NHS Scotland. 1984 GPASS. See http://www.gpass.scot.nhs.uk/.

Novotny, J., Tuecke, S. & Welch, V. 2001 An online credential repository for the grid: MyProxy. In *Proc. 10th Int. Symp. on High Performance Distributed Computing* (*hpdc-10*). Silver Spring, MD: IEEE Computer Society Press.

Rezaeian, M., Dunn, G., Leger, S. S. & Appleby, L. 2006 Ecological association between suicide rates and indices of deprivation in the north west region of England: the importance of the size of the administrative unit. *J. Epidemiol. Comm. Health* **60**, 956–961. (doi:10.1136/jech.2005.043109)

SHIP. 2010 Scottish Health Informatics Platform for Research. See http://www.scot-hip.ac.uk/.

Sinnott, R. O., Ajayi, O., Jiang, J., Stell, A. J. & Watt, J. 2008 Supporting grid-based clinical trials in Scotland. *Health Inform. J.* **14**, 79–93. (doi:10.1177/1081180X08089317)

Sinnott, R. O., Ajayi, O. & Stell, A. 2009 Data privacy by design: digital infrastructures for clinical collaborations. In *Proc. Int. Conf. on Information, Security and Privacy, Orlando, FL, 13–16 July 2009*, pp. 22–29.

Stell, A., Sinnott, R. O., Ajayi, O. & Jiang, J. 2007 Security oriented e-infrastructures supporting neurological research and clinical trials. In *2nd Int. Conf. on Availability, Reliability and Security 2007 (ARES 2007)*, pp. 629–636. Silver Spring, MD: IEEE Computer Society.

Stell, A., Sinnott, R. O., Ajayi, O. & Jiang, J. 2009 Designing privacy for a scalable electronic healthcare linkage system. In *Proc. IEEE Int. Conf. on Computer Science Engineering*, pp. 330–336. (doi:10.1109/CSE.2009.323)

UK Data Archive. 2003 British Household Panel Survey. See http://www.data-archive.ac.uk/.

University of Edinburgh. 2010 EDINA. See http://www.edina.ac.uk/.

University of Manchester. 2010 Mimas. See http://www.mimas.ac.uk/.

VOTES. 2008 Virtual Organisations for Trials and Epidemiological Studies. See http://www.nesc.ac.uk/hub/projects/votes/.

Watt, J., Sinnott, R. O., Jiang, J., Doherty, T., Higgins, C. & Koutroumpas, M. 2009 Tool support for security-oriented virtual research collaborations. In *Proc. IEEE Int. Symp. on Parallel Distributed Processing Applications, Chengdu, China, 9–11 August 2009*, pp. 419–424. (doi:10.1109/ISPA.2009.49)

World Health Organization. 2003 International classification of diseases. See http://www.who.int/classifications/icd/en/.