IT SUPPORT FOR THE AUSTRALIAN SCHIZOPHRENIA RESEARCH BANK

Henskens FA[1,2], Loughland CM[1,2], Aphale MS[1,2], Paul D[1], Richards JM[2], Rasser P[2], Carr VJ[1,2], Catts SV[2,4], Jablensky A[2,3], Michie PT[1,2], Mowry BJ[2,4], Pantelis C[2,5], Schall U[1,2], Scott RJ[1,2].

[1]University of Newcastle, NSW 2308, Australia; [2]Schizophrenia Research Institute, NSW 2010, Australia; [3]University of Western Australia, WA 6009, Australia; [4]University of Queensland, Queensland 4072, Australia; [5]University of Melbourne, Victoria 3010, Australia.

Corresponding Author: frans.henskens@newcastle.edu.au

Keywords: Globus Grid, Web services, certificate-based security, JSP, ASRB, schizophrenia.

Abstract: Schizophrenia represents one of the most perplexing and challenging problems confronting both researchers and health-care providers today. An Australian coalition of researchers has commenced construction of a resource termed the Australian Schizophrenia Research Bank (ASRB), which intends to support a wide spectrum of schizophrenia research studies. Software has been written to implement of a series of clinical assessment instruments, and a purpose-built Globus Grid constructed to provide secure aggregation and storage of the collected data. This paper describes the design, technology selections and implementation of the extant computer infrastructure, and discusses planned extensions and enhancements.

1 INTRODUCTION

Schizophrenia represents one of the most perplexing and challenging problems confronting both researchers and health-care providers today. With a prevalence of 4 per 1,000 (Saha et al., 2005), significant advances in schizophrenia research have been limited by the difficulty in achieving sufficiently large samples of probands and unaffected first degree relatives, in order to study the causal role of multiple genetic factors each of small effect. To overcome this limitation, a national (to Australia) coalition of researchers has commenced construction of a resource, termed the Australian Schizophrenia Research Bank (ASRB). The National Health and Medical Research Council (NHMRC), the Schizophrenia Research Institute (SRI), The Pratt Foundation, and other benefactors fund the ASRB, which intends to support a wide spectrum of schizophrenia research studies for Australian and international investigators.

The specific aims of the ASRB include:

1. To develop a research bank with comprehensive, cross-referenced data from a large sample (initially 2,000) of volunteers with schizophrenia and healthy controls (initially 2,000). The data collected includes clinical and cognitive characterisation, blood banking involving DNA extraction and establishment of immortal cell lines and MRI brain scans of schizophrenia using standardised imaging and analysis methodology
2. To establish infrastructure that makes the data set available to Australian and international researchers in a form that is comprehensive, cross-referenced, and able to support schizophrenia research across the clinical, cognitive, genetic, brain imaging and linked domains.

Funding for the establishment of the ASRB commenced in 2006, and following the official launch in 2007, the project commenced recruitment of sufferers and healthy controls. Construction of the IT support infrastructure commenced in 2006, and is currently supporting the collection and storage of data. This paper describes the design, technology selections and implementation of the extant computer infrastructure, and discusses planned extensions and enhancements.

2 REQUIREMENTS

The principles underpinning construction of the ASRB are: useful access to the data set will be provided to all authorised Internet-connected researchers without a requirement that they possess local high performance equipment; access will strictly accord with ethics approvals, local and national; collection of the data will be efficient, and will avoid double-handling as far as possible; collection of data will be ethical, respectful and considerate of all participants.

To this end, it was decided that the electronic dataset should be centrally stored, and it should include metadata about the tissue samples. Gathering of the data, and subsequent access to it would be best facilitated using the Internet. It became apparent that Grid technology (Foster & Kesselman, 1999) could be used in a rather non-conventional way, to provide a web-based secure, single entry-point environment for interaction with the data. In accordance with a requirement of the Australian Research Council, the Globus 4 Grid system (Foster, 2005) is used.

Each volunteer providing data to the Bank meets with a Clinical Assessment Office (CAO) and undergoes a comprehensive assessment process, involving multiple instruments and lasting for some three and a half hours. Traditionally, trained psychologists using paper and pen collect clinical and neuropsychological assessment data; the data is hand-scored, and checked for errors, prior to entering into a software tool for analysis. It was decided that the ASRB assessment data should be collected directly into a purpose-built software package.

In Australia, research that involves humans is subjected to close scrutiny, under conditions specified in per-project ethics approval documents. Protection of the privacy of the volunteers is of paramount importance. While the ultimate objective of the ASRB is provision of de-identified data of high usefulness for schizophrenia research, it is necessary, particularly in the collection phase, that identifying data is included and accessible to authorised persons. This fine-grained control over access to data is not typically required in Grid environments, and presented a challenge to the project.

The remainder of this paper discusses the techniques used to achieve these requirements.

3 GRID-BASED SECURITY

Ethics approvals are necessarily associated with the collection of data and samples from live patients. These approvals typically specify the project for which the data is to be used, and limit the group of researchers permitted to access the data, for example, to those at a particular institution, or in a particular research group. It is also common that researchers permitted to use and analyse patient data are typically prevented from being able to identify patients from their data (i.e. the data is de-identified). Whilst the ASRB data ultimately belongs to a single entity, it is collected at five principal locations spread across four Australian states, and each of the collection sites is covered by a local ethics approval. Thus, a major and important aim of the ASRB Grid is to provide controlled access to the data available to each particular user of the Grid. The most obvious need is to allow all authorized users to access the de-identified data, but it is also important to allow access to any other data (e.g. identifying data) for which the user has approval, either through their institution, research group, or personally. A further consideration is that it should be possible for selected personnel to identify patients from their data in the circumstance that analysis has discovered potentially beneficial treatments for those patients.

Typical Grids require strong security to determine whether a user should have access to a given system, or set of systems, without the need for any fine-grained security; a user is either allowed to access the system, or they are not. The ASRB Grid is different because users have different access rights to the resources provided by the Grid, even those on an individual component system.

As version 4 of the Globus toolkit is mainly built on the Web Services Resource Framework (WSRF), it can be easily used over the Web. As more fully discussed in (Paul et al., 2006), a Web portal is used to access the Grid systems, eliminating the need for researchers to install special software on their machines, and providing flexibility with respect to client location and host computer. The portal framework is Gridsphere (Novotny et al.), with GridPortlets (Russell et al., Accessed 2006) used to access the Grid. Gridsphere is an open-source portal framework completely compliant with the JSR 168 specifications, so that any standards-compliant portlet can be used. GridPortlets are a set of portlets for Gridsphere that allow access to Grid resource and user credential management, as well as GridFTP operations, and many other useful Grid activities. The GT4Portlets extension to this allows the execution of jobs on remote Globus Toolkit 4

systems, and further enhances GridPortlet's compatibility with the newest version of Globus.

A SimpleCA (SimpleCA, Accessed 2008) certificate authority supplies users with credentials to access ASRB Grid resources. To further facilitate the researcher's use of the system, PURSe portlets (Christie, 2007) are used to eliminate the user's need to knowingly interact with this system. Using these portlets, a user fills in a Web-based form to request an account. The user is then sent an email to verify their request and an administrator is informed of the request. The administrator can accept or reject the user, and has the capability to provide the user with access to an account on the Grid; ultimately the user is informed by email of the result. When a user is accepted, appropriate Grid credentials are automatically created for him/her and a personal proxy certificate stored the MyProxy server. The user can then log in to the Web portal, using the single password supplied by them in their initial request, and a proxy certificate is automatically retrieved from the MyProxy server. This proxy certificate is then available for access by the portlets in the Web portal; the portlets use these credentials to authenticate with any Grid resources in a manner that is completely transparent to the user.

As previously described, it is vitally important that researchers are restricted to access only that data for which they are approved. The Globus Toolkit includes a component that can be used for this purpose: the Community Authorization Service (CAS) (Globus, Accessed 2008) (not to be confused with JA-SIG's Central Authentication Service (JA-SIG, Accessed 2006)). CAS allows resource providers to give course-grained access to various systems, handing finer-grained access control management to the community of users. This is important for the ASRB Grid because there are very complex levels of access for different data resources, so fine-grained control is needed, and the users themselves (or the administrator, on their behalf) can best handle the complexities of these relationships.

CAS is based on the notion of assigning users to roles, which then have different sets of permissions associated with them. Each user can have multiple roles, and his or her access to a particular resource is only allowed if at least one of the roles allows access to the resource. Support for CAS roles has been added to the PostgreSQL database (PostgreSQL, Accessed 2006) used to store volunteers' responses to the clinical assessment battery.

Whenever a new role is needed (for example when a new ethics approval is granted), the system administrator creates a database view that is accessible only to users in the new role. The view can provide fine-grained control, specifying whether the users with a role have permission only to read data, or to insert new data, update existing data or delete data.

When a user (or group of users) is provided with a new role, the system checks to see whether the view allowed by the user's combination of roles currently exists. If it does not, a new view is automatically created, including rules that allow database operations such as insertion of new data or updating of existing data, but only if those permissions are included in one or more of the roles assigned to the user.

Users are then able to access the database using the Web portal, and any of their queries are automatically restricted to the views permitted by their roles. For example, the role of clinical assessment officer (CAO) allows the user access to complete volunteer information collected at the officer's site. Thus, if a Newcastle CAO performs the same database search as a Sydney CAO, they are returned different results; the Newcastle CAO will 'see' identifying data for Newcastle volunteers, while the Sydney CAO will be returned similar data about Sydney volunteers.

4 CLINICAL ASSESSMENT

After considering various structured numbering schemes for volunteer identification, it was decided that any such scheme could potentially breach privacy. For example, representation of familial relationships, as occurs in some numbering schemes, could allow a CAO with access to a volunteer's identity to discern and possibly identify that person's relation(s) for whom the CAO does not have identification permission. Accordingly a simple scheme was adopted whereby numbers are centrally and sequentially allocated.

Clinical Assessments (CAs) are performed at locations chosen in part with regard to the convenience of volunteers, so CAOs may, or may not be Internet connected while performing the assessment. Each CAO has been provided with a notebook computer, with installed CA software. To minimise the possibility of discomfort to volunteers who may feel threatened by the sight of the back of a screen, the tablet form of computer was chosen, so that the CA can be conducted with direct computer input using a stylus (very similar to use of a pen and paper). The lack of Internet connectivity during CA requires initial storage of the collected data on the notebook computer. It also requires allocation of the identification numbers prior to commencement of the CA. The process is that the CAO, in preparation for a CA, logs on to the Grid, selects the 'number allocation' portlet that accepts basic volunteer

identification and contact information (obtained when the appointment was made) and returns that volunteer's identification number. Centralised allocation ensures uniqueness of volunteer identifiers.

Another major benefit of this allocation technique is that accidental (or intentional) duplicate assessment of volunteers can be avoided. When a CAO requests allocation of a new identifier, the system checks the database to determine whether the provided volunteer information is substantially similar to that of an existing participant. In some cases (if access to the potential duplicate entry is permitted for the CAO) the issue can be resolved straight away; otherwise the situation is referred to the Project Manager, whose 'super user' permissions allow a determination to be made.

The CA software is written in the object-oriented Java programming language (Arnold & Gosling, 2000), chosen in part because the compiled code will execute on any computer with an implemented Java Virtual Machine (JVM), most significantly the CA notebooks and within the Unix-based central server's web site. The data collected during assessment, and its structure, is described in XML (Bray et al., 2006). JSP technology (Sun Developer Network, 2008) is used to facilitate alignment of software screens with the database fields. This architecture has already proven its value as user requirements evolved during the software development.

The assessments performed in a CA were determined after consultation with the principal schizophrenia researchers in Australia and abroad, and include the entire Diagnostic Interview for Psychoses (DIP) (Castle et al., 2006). Scoring of the DIP is achieved using the OPCRIT dynamic link library (Craddock et al., 2006) on the Windows-based notebooks, however the OPCRIT functionality will require re-coding for the server implementation. Also included in the CA are: Socio-demographic and Clinical History Schedule; Neurological, Personality, and Cognitive Functioning evaluations, and a psychosis screening tool.

Whilst clinical assessment has been automated previously, for example (Peters et al., 1998), this has largely been for one instrument only, not for a complete series as conducted by ASRB, making this process unique.

The benefits of an electronic clinical assessment system are that, by and large, it reduces the need for paper copies. Because assessments can be scored and the data checked automatically, this reduces the amount of time assessors spend second-handling the data, and the one data input occasion (at the time of assessment) reduces human error due to data scoring and entry. In addition, data is automatically checked to ensure all questions have been answered before the assessment is uploaded to the servers at the ASRB website. This automated data transfer allows for the quick and efficient aggregation of the scored dataset.

Security of data transfer is ensured by the use of Transport Layer Security (Freier et al., 1996), supported by a commercially provided Secure Socket Layer (SSL) certificate. This system uses key-based encryption (National Institute of Standards and Technology, 2006), with the ASRB server having its own private key, and client browsers using the matching public key (certificate) that is certified as belonging to the ASRB by an international authority trusted by all modern Web browsers. In this way, clients can be assured that all communication is encrypted and that only the ASRB system can decipher/understand it.

Computer usage is now commonplace, so the use of electronic means to collect data no longer intimidates participants or assessors. However, when required, either because of equipment malfunction or volunteer discomfort, assessors can revert to using traditional paper and pen versions by printing a copy from the electronic version.

At this time two user manuals have been produced. One comprises step-by-step instructions about installing the software, and updates, on the CA notebook computers. The other is a "How To..." manual that instructs CAOs on how to use the software. Experience shows that CAOs require some initial training (for example running through a training assessment with them, for which a practice web portal has been provided) and troubleshooting, but they have quickly adapted to the system and feedback is very positive, as shown in the next section.

5 USER FEEDBACK

Users of the systems have already reported the following benefits of computerisation:

1. It is quick, because assessors only have to handle the data once, at collection.
2. Labour costs are reduced because assessors can complete assessments more quickly. This is because they do not have to spend time scoring or on multiple data entry.
3. Accuracy is enhanced. The system allows consistent data collection across sites. Assessment is structured but flexible, in that it asks questions in the same way and order, but it is flexible enough to allow CAOs to return to questions if new information comes to light throughout the interview.

4. Comment sections are available for each question to record verbatim responses by the participant or any information of relevance. That information can then be checked by a psychiatrist for quality assurance, diagnostic, and staff supervision purposes.
5. Set data parameters reduce input errors during the interview. All assessments are checked for completion at the end of the assessment, reducing missing data. Data is scored automatically, reducing scoring errors.
6. It is efficient. The software is installed on laptop computers, making the assessment fully portable. In addition, the software has inbuilt automated skips for redundant questions so CAOs do not waste time on irrelevant questions.

6 FUTURE WORK

The next phase of development involves management of the growing number of collected MRI scans. Work is in progress on determination of the metadata that will be stored in the database about each scan. Additionally, standard structures are being designed to store the scans, both in original form (in excess of one gigabyte per scan), plus a set of standard mutations of the scan data. For example, prior to release of scans to researchers, the facial features must be removed to prevent identification.

Another aspect of the MRI data is pre-processing in preparation for analysis. Previous research suggests that variations in the volumes of cerebrospinal fluid, grey matter and white matter are related to brains diseases such as schizophrenia (da Silva, 2007). It follows that volumetric analysis of the brain structures is of interest to the research community. In preparation of such measurement, it is necessary to delineate the regions, after which the images can be normalised. At present the delineation is performed manually, since computerised algorithms, e.g. (Hata et al., 2000, Pham & Prince, 1999) do not yet produce the required level of accuracy. This is time-consuming, and not practical when thousands of images are involved. Work is in progress on development of new techniques for computerised segmentation of the brain MRIs.

Ultimately, the ASRB Grid infrastructure aims to provide researchers with the ability to analyse (subsets of) the data collection, leading to advances in the understanding and treatment of schizophrenia. While at present the Grid is primarily used for its security features, some researcher tasks will benefit from access to the parallel resources it makes available. For example, transfer of sets of MRI data over the Internet for local (to the researcher) analysis is expensive with respect to time (noting that some of the member sites are up to 4,000 kilometres apart). It is much more efficient to carry out the analysis by performing the computation close to the data source, with high bandwidth data path(s) joining the storage and compute nodes.

It is not easy to automatically execute a task on a set of remote machines. Projects such as GT4Portlets allow the execution of jobs on a single remote machine, and projects such as the Gridbus Broker (Venugopal et al., 2006) automatically allocate tasks to servers, but the interfaces to these are very general. Thus a further task for this project is to create a portlet wizard that allows the easy creation of a portlet to execute a particular application. These portlets will likely be based on the Gridbus Broker, but will enable researchers to choose input files and set parameters using a simple, easily understandable Web form. Provision of a wizard will make it easy for developers to create portlets for many different programs. By way of support for computations with special needs, more knowledgeable developers (or their programmers) will have access to the full source code so the portlet can be modified as needed. In this way researchers and developers will be able to use the processing capabilities of the distributed compute servers much more easily than would otherwise be possible.

7 CONCLUSIONS

This paper presents the current state of development of IT-based infrastructure to support the Australian Schizophrenia Research Bank. A Globus-based Grid underpins the centralised data storage, providing a secure (using CA certificates) processing and access system that presents a Web browser interface to users. Purpose-built software executing on (optionally non-networked) Windows tablet computers is described. This software automates a comprehensive set of clinical assessment instruments; each collected data item, together with the diagnoses, is later securely uploaded to the central repository using a Web portal.

The use of Grid technology for its security, rather than its resource sharing capabilities, is somewhat unusual. Techniques are described that provide fine-grained control over user access to the centralised dataset. This granularity is essential to achieve compliance with the ethics conditions under which research that involves humans is conducted in Australia.

Various extant and future Grid facilities and capabilities are discussed, together with user affirmation of the current system. Experience with the design and implementation of further functionality will be the subject of future publications.

ACKNOWLEDGEMENTS

This work was supported by the Australian Research Council (ARC) grant SR0566756 (2005-2006), the National Health & Medical Research Council (NHMRC) grant AIP/ERP #1679 (2006-2010), the Schizophrenia Research Institute (SRI) utilising infrastructure funding from NSW Health, and a grant from the Pratt Foundation (2007-2011).

REFERENCES

ARNOLD, K. & GOSLING, J. (2000) *The Java Programming Language, 3rd ed.*, Addison Wesley.

BRAY, T., PAOLI, J., SPERBERG-MCQUEEN, C. M., MALER, E. & YERGEAU, F. (2006) Extensible Markup Language (XML) 1.0 (Fourth Edition). http://www.w3.org/TR/2006/REC-xml-20060816.

CASTLE, D. J., JABLENSKY A., MCGRATH J. J., CARR V., MORGAN V., WATERREUS A., VALURIG G., STAIN H., MCGUFFIN P. & FARMER A. (2006) The diagnostic interview for psychoses (DIP) : development, reliability and applications. *Psychological Medicine* 36(1)**,** 69-80.

CHRISTIE, M. (2007) PURSe Portlets Website. http://www.extreme.indiana.edu/portals/purse-portlets.

CRADDOCK, M., ASHERSON, P., OWEN, M. J., WILLIAMS, J., MCGUFFIN P. & FARMER A.E. (2006) The OPCRIT Webpage. *The British Journal of Psychiatry,* 169**,** 58-63.

DA SILVA, F. (2007) A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis,* 11**,** 168-182.

FOSTER, I. (2005) Globus Toolkit Version 4: Software for Service-Oriented Systems. *IFIP International Conference on Network and Parallel Computing.* Springer-Verlag.

FOSTER, I. & KESSELMAN, C. (1999) *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann.

FREIER, A., KARLTON, P. & KOCHER, P. (1996) The SSL Protocol Version 3.0. *http://wp.netscape.com/eng/ssl3/draft302.txt*.

GLOBUS (Accessed 2008) GT 4.0: Security. http://www.globus.org/toolkit/docs/4.0/security/.

HATA, Y., KOBASHI, S., HIRANO, S., KITAGAKI, H. & MORI, E. (2000) Automated segmentation of human brain MR images aided by fuzzy information granulation and fuzzy inference. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews,* 30(3)**,** 381-395.

JA-SIG (Accessed 2006) JA-SIG Central Authentication Service. http://www.ja-sig.org/products/cas.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (2006) Secure Hashing. http://csrc.nist.gov/groups/ST/toolkit/secure_hashing.html.

NOVOTNY, J., RUSSELL, M. & WEHRENS, O. Gridsphere: A Portal Framework for Building Collaborations. Gridsphere Project Website.

PAUL, D., HENSKENS, F. A., JOHNSTON, P. & HANNAFORD, M. R. (2006) Portal-based Support for Mental Health Research. *Grid Computing Environments Workshop, Supercomputing '06.* Tampa, Florida, IEEE Computer Society.

PETERS, L., CLARK, D. & CARROL, F. (1998) Are computerised interviews equivalent to human interviewers?: CIDI-Auto vs. CIDI. *Psychological Medicine,* 28**,** 893-901.

PHAM, D. L. & PRINCE, J. L. (1999) Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Transactions on Medical Imaging,* 18(9).

POSTGRESQL (Accessed 2006) The world's most advanced open source database. *http://www.postgresql.org/.*

RUSSELL, M., NOVOTNY, J. & WEHRENS, O. (Accessed 2006) The Grid Portlets Web Application: A Grid Portal Framework. Gridsphere Project Website.

SAHA, S., CHANT, D., WELHAM, J. & MCGRATH, J. (2005) A systematic review of the prevalence of schizophrenia. *PLoS Medicine,* 2**,** 413-433.

SIMPLECA (Accessed 2008) SimpleCA Instructions . *http://www.vpnc.org/SimpleCA/.*

SUN DEVELOPER NETWORK (2008) JavaServer Pages Technology. http://java.sun.com/products/jsp/.

VENUGOPAL, S., BUYYA, R. & WINTON, L. (2006) A Grid Service Broker for Scheduling e-Science Applications on Global Data Grids. *Concurrency and Computation: Practice and Experience,* 18(6)**,** 685-699.