Portal-based Support for Mental Health Research

David Paul[1], Frans Henskens[1] Patrick Johnston[2] and Michael Hannaford[1]

[1] School of Electrical Engineering & Computer Science, The University of Newcastle, N.S.W. 2308, Australia
[2] Centre for Mental Health Studies, The University of Newcastle, N.S.W. 2308, Australia

Abstract. This paper describes experiences with the use of the Globus toolkit and related technologies for development of a secure portal that allows nationally-distributed Australian researchers to share data and application programs. The portal allows researchers to access infrastructure that will be used to enhance understanding of the causes of schizophrenia and advance its treatment, and aims to provide access to a resource that can expand into the world's largest on-line collaborative mental health research facility. Since access to patient data is controlled by local ethics approvals, the portal must transparently both provide and deny access to patient data in accordance with the fine-grained access permissions afforded individual researchers. Interestingly, the access protocols are able to provide researchers with hints about currently inaccessible data that may be of interest to them, providing them the impetus to gain further access permissions.

1 Introduction

Schizophrenia is a brain disease that affects approximately 0.6-1.5% of the population, with an incidence of 18 - 20 cases per 100,000 per year [9]. Although prevalence is low, the burden of the illness upon society and upon sufferers and their families is extremely high. The World Health Organisation, for example, rates schizophrenia amongst the ten leading causes of disease burden. The disorder involves severe cognitive, affective and perceptual dysfunctions, which, at an overt behavioural level, manifest themselves in terms of delusional beliefs and disorganised behaviours; perceptual disturbances including, particularly, auditory hallucinations; and lack of motivation, and general decline in personal and social functioning. Consequently, it is a disease associated with very high costs to government (AUD35,000 per patient per year) [1] and extremes of social impoverishment and economic disadvantage [10].

Recent scientific advances have led to a model of schizophrenia that recognises the role of abnormal neuro-developmental and/or neurodegenerative processes in altering the structure and function of the brain. Until relatively recently detailed images of cerebral morphology could only be obtained from post-mortem tissue. The limitations of the traditional tissue-based approach to neuropathology can potentially be overcome through the use of neuroimaging technologies. Neuroimaging techniques offer the potential for *in vivo* studies of brain structure as well as function, thus overcoming problems relating to tissue degeneration postmortem, invariably small samples of post-mortem brains and, of course, the obvious fact that the tissue is derived from deceased persons. Moreover, techniques such as magnetic resonance imaging (MRI) allow for repeated testing of the same individuals, and thus longitudinal studies may be undertaken. A further advantage of MRI is that it may be employed to produce high-resolution three-dimensional digital representations of brain structure. This approach lends itself more easily to sharing and distribution of the primary source data (i.e. digital images) among research teams than does traditional approaches in neuropathology (i.e. where the brain tissue itself is the primary source data). It also supports the

This research is supported by The Australian Research Council (ARC) grant SR0566756 (2005-2006). On-going work is supported by the National Health & Medical Research Council (NHMRC) grant AIP/ERP #1679 (2006-2010), and by a grant from the Pratt Foundation (2007-2011).

application of computational image processing techniques for the precise definition, localisation and measurement of brain structures.

The heritability of schizophrenia is of the order of 70-80%. However, the inheritance pattern is not the classical Mendelian type. As with other complex diseases (eg, diabetes, cardiovascular disease), it is believed to involve a number of contributing genes, each of small effect, interacting with each other and with environmental factors. With this in mind, traditional genetic research approaches based on the diagnostic category of schizophrenia need to be modified if we are to further our understanding of the genetic basis of this disease. A more recent approach in schizophrenia research has been to investigate discrete neurobiological or neurocognitive characteristics that may be more closely linked to a particular gene [8, 12] rather than the clinical syndrome diagnosed as schizophrenia. These characteristics, known as endophenotypes, can assist researchers in unravelling the complex genetic causality of schizophrenia and help to identify individuals who carry the genetic trait for these discrete deficits [20].

The NISAD/LONI Virtual Brain Bank [14] primarily consists of a large distributed database of high resolution 3D computer representations of the brains of approximately 250 schizophrenia patients and age/gender-matched healthy control subjects, derived from structural MRI images and transformed into a standardised spatial coordinate system. The purpose of this bank is to provide a resource for the analysis of subtle structural variations between the brains of schizophrenia patients and healthy controls, and to map brain changes that occur as a result of variables such as age, gender, duration of illness and duration of untreated psychosis. The brain bank also provides the opportunity to explore associations between brain structure and clinical or neurocognitive measures, gene expression or genetic linkage data, and functional measures of brain activity such as functional MRI (fMRI) or event-related potentials (ERPs).

A further example of the significant impact of data-access-enabling infrastructure on research was the National Institute for Schizophrenia and Allied Disorders (NISAD) [15] Schizophrenia Research Register. It was intended that the Virtual Brain Bank would act as the foundation to which could later be added putative endophenotype measurements derived from the Schizophrenia Register participants and other neurocognitive studies of schizophrenia as well as genetic information derived from the DNA Bank and the Laboratory of Neuro Imaging (LONI) [13]. Such integrative strategies that combine various methodological approaches have been shown to considerably further the understanding of the pathology of schizophrenia. The recently established Australian Schizophrenia Research Bank (ASRB) builds on and extends the ideas of such previous facilities to create a nationally accessible resource for schizophrenia researchers in Australia and beyond.

In this paper we describe and discuss issues in the use of primarily Globus-based [4] technology to build a grid [3] that allows geographically distributed researchers to contribute to initially the NISAD/LONI Virtual Brain Bank, and now the encompassing ASRB's collection of schizophrenia-related data and software resources in the quest for knowledge on the reasons for and treatment of schizophrenia.

2 The ASRB Grid

A major issue for schizophrenia research is the expense of the collection of patient data (e.g. MRI brain scans, tissue samples) needed for analysis. The ASRB will have a major impact on schizophrenia research in Australia because it will amortise the high cost and the significant time involved in obtaining data across the

national body of researchers. As schizophrenia is likely to involve multiple genes of small effect, access to large sample sizes is a key to undertaking studies of sufficient statistical power. With its cross-referenced data in clinical, cognitive, neuroanatomical and genetic domains, the ASRB will make a huge contribution to schizophrenia research on a national scale, enabling multiple research questions to be addressed relatively easily in a large sample that would otherwise be inaccessible or prohibitively expensive for independent investigators to acquire. This large data set will be formed by merging existing data held by groups around the country, and supplementing it with data obtained by a concerted recruitment and collection process.

Ethics approvals are necessarily associated with the collection of data and samples from live patients. Such ethics approvals typically specify the project for which data is to be used, and limit the group of researchers who can access the data to, for example, those at a particular institution, or in a particular research group. It is also common that most researchers permitted to use and analyse patient data are prevented from being able to identify patients from their data (i.e. the data is de-identified). The extant data collections currently held at the disparate Australian member sites are all subject to existing ethics approvals. Access to the new patient data for which collection has been funded by the NHMRC, is similarly controlled. 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 newly collected data, but it is also important to allow access to any other data collections 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.

As the ASRB Grid contains personal patient information, security is of vital importance. 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. Further, a researcher should be able to perform a preliminary query on data for which they are not currently authorised, allowing them to identify data of interest as a pre-cursor to a request for access to it. For example, it should be possible for the researcher to search for scans exhibiting particular features to determine if there are sufficient samples to justify their requesting access to them. If there were insufficient data items that match their query, it would be a waste of time and resources to request access to the data. If, on the other hand, it was found that there was a sufficiently large extant data set (albeit currently unavailable to the individual researcher), it is likely that a request for access to that existing data would be significantly easier (and less expensive) to achieve than collection of new data. Notwithstanding, it is essential that certain aspects of the data, especially information that can identify patients, be inaccessible to any user who has not been given specific rights to access it.

Once the researcher has the data needed for their experiment, they typically would execute computer programs to analyse this data. At present this can involve manually collecting the data into a compressed archive, sending it to, for example, Los Angeles via FTP, and waiting for the results to be returned. At the remote processing site, a user must extract the data,

schedule it for analysis, collect the results and then return them to the initiating researcher. Other less compute-intensive tasks can be controlled by a single user, though these still require manual scheduling on computers in Australia, which can be time consuming, increasing the time needed by the researcher to do their job. It is intended that by utilizing compute servers in the ASRB Grid, this hands-on approach to computer-based analysis can be reduced, with researchers simply submitting the job to the Grid, after which the Grid automatically schedules and runs the job, collects the results, and returns them to the researcher, with no further human interaction required.

A final and important requirement of the ASRB Grid is that it should be easy to use, and provide reasonable performance and feedback. If the user interface to the new infrastructure is too complex, or if the performance is pedestrian, users will prefer to continue using the familiar old methods, with all their problems. Thus, use of the new system must be as intuitive as possible, and should hide or abstract over all unnecessary complexity. This means that sensible defaults should be chosen for all options, and a consistent interface should be provided to enable the researchers to concentrate on their research rather than being caught up dealing with the vagaries of the computer support.

3 Support for Fine-grained Security

To make the ASRB Grid as accessible as possible, it was decided at an early stage that Web services should be used wherever possible. It was also a preference of the Australian Research Council that the Globus Toolkit 4 [4] be used. Thus Globus was chosen as the software to provide the grid framework. Version 4 of the Globus toolkit is mainly built on the Web Service Resource Framework (WSRF), which allows Web services to have state, so that after a request has been made the service can later be queried to obtain updated information about the task.

As data access is a very important part of the ASRB Grid, two important components of the Globus Toolkit for this project are GridFTP [5] and OGSA-DAI (Open Grid Services Architecture Data Access and Integration) [17]. GridFTP is an extension to regular FTP that supports using Globus credentials for authorization and authentication. It has been extended in Globus Toolkit 4 with the Reliable File Transfer service, which is a Web service for managing secure third-party GridFTP transfers. OGSA-DAI is middleware designed to give secure access to data stores such as relational databases, as well as to integrate data from different sources via the Grid. It allows the access of relational databases using the WSRF, giving the ability to securely access them via Web services.

It was decided that a Web portal should be used to access the Grid systems, as this will eliminate the need for researchers to install special software on their machines, providing flexibility with respect to client location and host computer. The portal framework chosen is Gridsphere [16], with GridPortlets [19] 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 by Gridsphere. 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.

In order to supply users with credentials to access ASRB Grid resources, a SimpleCA certificate authority is being established. To further facilitate the researcher's use of the system, PURSe portlets [2] 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. Provided the user is accepted, appropriate Grid credentials are automatically created for the user and a proxy certificate stored for them in a MyProxy server. The user can then log in to the Web portal, using a password supplied by them in their initial request, and a proxy certificate is automatically retrieved from the MyProxy server. This proxy certificate will then be 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.

Since identified patient data will be stored on the ASRB Grid, it is vitally important that researchers are restricted to access only that data for which they are approved (resultant from ethics approval, or otherwise). As a result it is required that users be given different levels of access to resources based on both their own identity, and the groups to which they belong. The Globus Toolkit includes a component that can be used for this purpose: the Community Authorization Service (CAS) [6] (which is not to be confused with JA-SIG's Central Authentication Service [11]). 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 complexities of these relationships can best be handled by the users themselves. GridFTP is the only component of the Globus Toolkit that supports CAS out of the box, though OGSA-DAI can be extended to support CAS with very little impact on performance [18].

Much of the Globus Toolkit is currently accessible only through the use of command-line statements. Technologies such as the CoG Kits [22] and GridPortlets make access to Globus Grids much easier, but the CAS technologies that we have chosen to use have really only been usable from the command line. Thus, one of the first things needed by this project is portlets for accessing CAS. A portlet that allows authorized users to manage CAS entities has been created. With this facility users with the correct CAS permissions are able to view, create, and delete CAS entities, such as groups or service actions. In addition the portlet provides the ability to grant and revoke rights to groups and services. CAS will thus also be used by administrators to grant access to various database tables, through OGSA-DAI.

4 Future Work

Development of the ASRB Grid is very much an on-going project, and there are a number of parallel development tasks in progress, as described in the following sub-sections.

4.1 Description of Patient Data

The above security framework is designed to provide tightly controlled access to resources such as data and computation. To date much of the extant patient data has not been available on-line; rather the data are stored on CDs or DVDs in researcher's offices, and these must be moved to on-line storage subsystems so they can be

accessed using the Grid. A further issue is the existence of aggressive firewalls that have been used to protect confidentiality of patient data at some of the host sites. The recently-funded collection of substantial quantities of new patient data has not yet begun but is imminent, so provision of infrastructure for storage and processing of that new data is a priority. In parallel it is necessary to finalise meta-data description of the heterogeneous extant (and the homogeneous future) data that will be accessible through the ASRB Grid. Until this significant task is completed no specific tools development can take place.

4.2 Extension of Portlet Support

To date, there has been a paucity of reported development of portlets to access OGSA-DAI resources, especially for OGSA-DAI secured by CAS. While some OGSA-DAI portlets have been developed, they currently do not provide the level of support for security required by this implementation, and so must be extended to provide the necessary security. It will also be necessary to create or modify some GridFTP portlets to include CAS functionality so that researchers are able to easily share their data with groups to which they wish to provide such access. It is also planned to create a new PURSe Portlets registration module to automatically enrol users in various CAS groups when their account is created. This will include placing them in a group over which they have complete control, as well as giving them exclusive access to space on a GridFTP server. Users will then be able to create their own self-controlled groups, allowing them to share their data with authorised users while asserting as much fine-grained control as is necessary. There would be no requirement for administrator intervention in the establishment and control of such groups.

4.3 Abstraction over Distributed File Storage

A service that allows users to create logical folders, providing a window onto data on all the different GridFTP servers to which they have access, will also be integrated into the system. Thus, users will simply see a familiar folder-like structure containing sub-folders and files. This is achieved using the Globus Replica Location Service (RLS) [7] and a system to map a set of logical files to a set of logical folders; the actual files in the folders may be stored on any of the GridFTP servers available to the user of the Grid. The various locations of the data available to the user will be abstracted away by this service, allowing users to simply see their data without regard for the location at which it is stored.

4.4 Access to Data Processing and Analysis Facilities

The ultimate aim of the ASRB Grid infrastructure is to provide researchers with the ability to analyse (subsets of) the data collection, leading to advances in the understanding and treatment of schizophrenia. While it will be possible (subject to access rights) for researchers to download data to their own machines to perform analysis, there will be tasks which will benefit from access to the parallel resources of the Grid. For example, the data associated with a single MRI scan can exceed one gigabyte, and transfer of such quantities of data across the Internet is expensive with respect to time (noting that some of the member sites are up to 4,000 kilometres apart). Analysis of such data is more efficiently performed by positioning the computation close to the data source, with high bandwidth data path(s) joining them. Unfortunately, automatically executing a task on a set of remote machines is difficult. Projects such as GT4Portlets allow the execution of jobs on a single remote machine, and projects such as the Gridbus Broker [21] 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. It is envisaged that these portlets will be based on the Gridbus Broker, but will enable researchers to choose input files and set parameters using a simple, easily understandable Web form. The provision of a wizard will make it easy for developers to create portlets for many different programs. If a specific program has special needs, however, developers will still have access to the full source code so that the portlet can be modified as needed. This will enable researchers and developers to use the processing capabilities of the distributed compute servers much more easily than is currently possible.

5 Conclusion

This paper introduces a project that uses the Globus toolkit and related technologies that allows Australian Mental Health researchers to share data and application programs in their quest for understanding of schizophrenia and ultimately improvements in its treatment. A web services portal that provides fine-grained control over user access to resources is described. This portal simultaneously provides simple authentication-based access for users and certificate-based access to sub-sets of the entire resource collection. Users are unaware of host network boundaries and the need for separate authentication for the disparate sites and servers; these requirements are abstracted away by the portal.

The ASRB Grid is very much a work in progress. On-going development of abstractions over distributed data storage, remote compute services and portal development are also presented. These facilities will result in: nested folders that provide consistent access to locally and remotely stored data; intuitive wizard-based access to distributed compute servers and application programs; the ability for users to provide individuals and/or groups with controlled access to their personal data store.

6 References

1. Carr, V., Lewin, T., Neil, A., Halpin, S., and Holmes, S., *Premorbid, psychosocial and clinical predictors of the costs of schizophrenia and other psychoses.* British Journal of Psychiatry, 2004. **184**: p. 517-525.
2. Christie, M., *PURSe Portlets Website*, http://www.extreme.indiana.edu/portals/purse-portlets.
3. Foster, I. and Kesselman, C., *The Grid: Blueprint for a New Computing Infrastructure.* 1999: Morgan Kaufmann.
4. Foster, I. *Globus Toolkit Version 4: Software for Service-Oriented Systems.* in *IFIP International Conference on Network and Parallel Computing*. 2005: Springer-Verlag.
5. Globus, *GT 4.0 GridFTP*, http://www.globus.org/toolkit/docs/4.0/data/gridftp/.
6. Globus, *GT 4.0: Security*, http://www.globus.org/toolkit/docs/4.0/security/.
7. Globus. RLS: Replica Location Service, http://www.globus.org/rls/.
8. Gottesman, I.I., McGuffin, P., and Farmer, A.E., *Clinical genetics as clues to the real genetics of schizophrenia (a decade of modest gains whilst playing for time).* Schizophrenia Bulletin, 1987. **13**(1): p. 23-47.
9. Gureje, O. and Bamidele, R.W., *Gender and schizophrenia: association of age at onset with antecedent, clinical and outcome features.* Australia and New Zealand Journal of Psychiatry, 1998. **32**(3): p. 415-423.
10. Jablensky, A., *Epidemiology of schizophrenia: the global burden of disease and disability.* European Archives of Psychiatry and Clinical Neuroscience, 2000. **250**(6): p. 274-285.
11. JA-SIG, *JA-SIG Central Authentication Service*, http://www.ja-sig.org/products/cas.
12. Kremen, W.S., Faraone, S.V., and Seidman, L.J., *Neuropsychological risk indicators for schizophrenia: a preliminary study of female relatives of schizophrenic and bipolar probands.* Psychiatric Research, 1998. **79**(3): p. 227-240.
13. LONI, *Laboratory of Neuro Imaging*, http://www.loni.ucla.edu/.

14. NISAD, *The NISAD/LONI Virtual Brain Bank*, http://www.nisad.org.au/newsEvents/resNews/wwwsczres.asp.
15. NISAD, http://www.nisad.org.au/.
16. Novotny, J., Russell, M., and Wehrens, O., *Gridsphere: A Portal Framework for Building Collaborations*, Gridsphere Project Website.
17. OGSA-DAI, *OGSA-DAI Software*, http://www.ogsadai.org.uk/index.php.
18. Pereira, A., Muppavarapu, V., and Chung, C., *Role-Based Access Control for Grid Database Services Using the Community Authorization Service.* IEEE Trans. on Dependable and Secure Computing, 2006. **3**(2): p. 156-166.
19. Russell, M., Novotny, J., and Wehrens, O., *The Grid Portlets Web Application: A Grid Portal Framework*, Gridsphere Project Website.
20. Trillenberg, P., Lencer, R., and Heide, W., *Eye Movements and psychiatric disease.* Current Opinion in Neurology, 2004. **17**(1): p. 43-47.
21. Venugopal, S., Buyya, R., and Winton, L., *A Grid Service Broker for Scheduling e-Science Applications on Global Data Grids.* Concurrency and Computation: Practice and Experience, (accepted Jan 2005).
22. von Laszewski, G., Gawor, J., Lane, P., Rehn, N., Russell, M., and Jackson, K., *Features of the Java Commodity Grid Kit.* Concurrency and Computation: Practice and Experience, 2002. **14**: p. 1045-1055.