The Challenges of developing a collaborative data and compute grid for Neurosciences

John Geddes [1], Clare Mackay [1], Sharon Lloyd [2], Andrew Simpson [2], David Power [2], Douglas Russell[2], Mila Katzarova[2], Martin Rossor [3], Nick Fox [3], Jonathon Fletcher[3], Derek Hill [4], Kate McLeish[4], Joseph V Hajnal [5], Stephen Lawrie[6], Dominic Job[6], Andrew McIntosh [6], Joanna Wardlaw [7], Peter Sandercock [7], Jeb Palmer [7], Dave Perry [7], Rob Procter[8], Jenny Ure[8], Philip Bath [9], Graham Watson [9]

[1]University of Oxford, Department of Psychiatry, [2]University of Oxford, Computing Laboratory, [3]University College London, Institute of Neurology, [4]University College London, Centre for Medical Image Computing (MedIC), [5]Imaging Sciences Department, Imperial College London, [6] University of Edinburgh, Department of Psychiatry, [7] University of Edinburgh, Department of Clinical NeuroSciences, [8]School of Informatics, University of Edinburgh, [9]The University of Nottingham

John.Geddes@psych.ox.ac.uk

Technical Areas covered in the paper: large scale neurosciences clinical trials, data aggregation, Grid technology, collaborative research,

Author Details:

Professor John Geddes
Department of Psychiatry
University of Oxford
Warneford Hospital
Oxford
OX3 7JX
United Kingdom

Tel: +44 (0)1865 226480
Fax: +44 (0)1865 793101
Email: John.Geddes@pysch.ox.ac.uk

Preference: Oral presentation

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 **IEEE**

The Challenges of developing a collaborative data and compute grid for Neurosciences

John Geddes [1], Clare Mackay [1], Sharon Lloyd [2], Andrew Simpson [2], David Power [2], Douglas Russell[2], Mila Katzarova[2], Martin Rossor [3], Nick Fox [3], Jonathon Fletcher[3], Derek Hill [4], Kate McLeish[4], Joseph V Hajnal [5], Stephen Lawrie[6], Dominic Job[6], Andrew McIntosh [6], Joanna Wardlaw [7], Peter Sandercock [7], Jeb Palmer [7], Dave Perry [7], Rob Procter[8], Jenny Ure[8], Philip Bath [9], Graham Watson [9]

[1]University of Oxford, Department of Psychiatry, [2]University of Oxford, Computing Laboratory, [3]University College London, Institute of Neurology, [4]University College London, Centre for Medical Image Computing (MedIC), [5]Imaging Sciences Department, Imperial College London, [6] University of Edinburgh, Department of Psychiatry, [7] University of Edinburgh, Department of Clinical NeuroSciences, [8]School of Informatics, University of Edinburgh, [9]The University of Nottingham

John.Geddes@psych.ox.ac.uk

Abstract

The three-year UK NeuroGrid project aims to develop a Grid-based collaborative research environment to support the data and compute needs for a UK Neurosciences community. This paper describes the challenges in developing this architecture and details initial results from the development of its first prototype to support psychosis, dementia and stroke research and the social challenges of such a collaborative research project. The paper discusses approaches being taken to explore the collaborative science process to inform the requirements for follow on prototypes and methods utilized to develop an effective project team.

1. Introduction

There is an increasing drive within the UK to maximize the research productivity of the NHS – particularly in terms of realizing the potential of the unique selling points of the clinical research environment, including clinical academic strengths in imaging and early phase drug development (McKinsey report www.ukcrc.org). Within this context, NeuroGrid [1] is a three-year, £2.1million project funded through the Medical Research Council in the UK to develop a Grid-based collaborative research environment for imaging in large scale studies in neuropsychiatric disorders in the UK. NeuroGrid will be developed around three component clinical exemplars in stroke, dementia and psychosis, and complex services for quantitative and qualitative image analysis. This project, which started in March 2005, has a project team distributed across 11 sites in the UK and includes both clinical and technical staff. This paper aims to describe the challenges identified to date and progress made in developing a collaborative data and compute Grid for neurosciences in the UK, and in the development of the clinical usage exemplars. The paper starts by recapping the motivation and objectives for the project, discusses the initial requirements which have been determined for an early prototype data grid between the clinical sites, describes the architecture that has been developed and discusses the challenges posed in developing such an infrastructure. The paper concludes with a discussion on the lessons learned in developing NeuroGrid to date and the activity to be carried out over the next 12 months.

2. The NeuroGrid project

2.1 Motivation

Neuro-imaging in large–scale clinical trials is promising huge benefits in the assessment of specific treatment effects on neuropathological processes. The efficiency and power of neuroimaging is, however, hampered by several factors including variances in acquisition techniques, quality assurance and access to remote datasets. Neuroimaging research is still typified by small studies carried out in single research centres. Data is

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

rarely shared and the sharing of image analysis tools is made available through the web but this is limited to those groups prepared to publish their methods. When data is shared, subtle differences between centres in the way that the images acquired normally inhibit reliable quantitative analysis of aggregated data. Furthermore, data curation in neuroimaging research tends to be poor, making aggregation of data between or within sites difficult, if not impossible.

Researchers use innovative imaging techniques to detect features that can refine a diagnosis, phenotype subjects, track normal or often subtle pathophysiological changes over time and/or improve our understanding of the structural correlates of the clinical features. The identification of true disease-related effects is obviously crucial and problems are caused by confounding and artefactual changes in the complex procedures involved in image acquisition, transfer and storage. There are two basic approaches to the extraction of detailed information from imaging data – quantitative assessment and qualitative assessment – both of which pose key challenges. In quantitative assessment, sophisticated, largely automated and computationally intensive image analysis algorithms have recently been developed that offer great promise in terms of quantification and localization of signal differences. Current practice relies on these algorithms being locally implemented, which can lead to a lack of standardization and these algorithms are only shared if the research group decides to publish the software. In qualitative assessment, many large randomised controlled trials or observational studies use imaging to phenotype, to assess severity, and may use follow-up scans to assess disease progression or response to treatment. Such studies may enrol thousands of patients from hundreds of centres using a large variety of imaging equipment. A reliable system is required for managing the scans (collection, storage and dissemination), and the results from image raters and the study metadata.

There are, of course, other challenges associated with managing imaging data for clinical trials.

- Multiple and ever changing technologies including scanners and acquisition media.
- Secure long term data storage of large datasets and effective use and integration of data
- Efficient observer rating for large imaging studies, which is essential in multicentre trials and observational studies to improve consistency of diagnosis
- Image data quality and consistency

The NeuroGrid project is motivated, in part, by a desire to address all of the above problems.

2.2 Objectives

The principal objective of the NeuroGrid consortium is to enhance collaboration within and between clinical researchers in different domains, and between clinical researchers and e-scientists. Sharing data, experience and expertise will facilitate the archiving, curation, retrieval and analysis of imaging data from multiple sites and enable large-scale clinical studies. To achieve this, NeuroGrid aims to build upon Grid technologies and tools developed within the UK e-Science programme to integrate image acquisition, storage and analysis, and to support collaborative working within and between neuro-imaging centres.

There are three main elements to NeuroGrid:

1. NeuroGrid will create a Grid-based infrastructure to connect neuro-imaging centres, thereby providing rapid and reliable flow of data, and facilitating easy but secure data sharing through interoperable databases, and sophisticated access control management and authentication mechanisms.
2. NeuroGrid will develop distributed, Grid-based data analysis tools and services, including a neuroimaging toolkit for image analysis, image normalization, anonymisation and real-time acquisition error trapping..
3. NeuroGrid will iteratively develop and deploy the tools and techniques it creates in three clinical exemplar projects in stroke, psychosis, and dementia to explore real world problems and solutions. NeuroGrid will use the clinical exemplars both to derive detailed requirements and to validate them.

3.0 Initial requirements for a first stage prototype

NeuroGrid is taking an iterative approach to development of an architecture to ensure close interaction with the potential users of any system deployed. With this in mind, the project team performed an early analysis of the needs of each of the exemplars through intensive discussions with the clinical investigators on the project. This exercise was conducted to determine requirements whilst developing a working relationship across the disparate teams. The requirements were documented in textual form and were supplemented with data flow diagrams provided by clinical researchers. The results of this exercise are detailed below.

3.1 Generic requirements

Common to all exemplars is the need to determine secure and effective ways to manage the data used in the clinical trials. The data takes the form of medical images in the form of CT scans and MRI volumes and coded or descriptive information from patients who have consented to take part in trials and often is sensitive in nature. Often the case information includes image studies and associated information collected through clinical assessment, visual reading information and acquisition details. MRI volumes could in theory be reconstructed to show a patient's facial features, so the anonymisation of a patient record is often insufficient for collaborative studies where knowledge of the original subject is not necessary. Data curation involves the acquisition, processing, archiving, retrieval and usage of this medical data and failure to perform this function thoroughly could result in inaccurate research results, ethical non-compliance or loss of valuable clinical data that has been costly to acquire. Typically clinical trials have to manage these processes in an ad hoc fashion as standards do not exist for this activity and multiple centre trials add additional complexity with the need to coordinate such issues as naming conventions for files, patient clinical trial ID management and acquisition parameters. The retrieval and access of this data also requires new architectures to support the secure sharing of the data. Other common requirements across the exemplars include desktop tools for image representation, manipulation and annotation and the need ensure that the control for the access rights for specific data sets resides with the owners of that data. One particular requirement that has been expressed is to be able to run algorithms on other datasets that a researcher does not own and retrieve the results of this analysis, but not the original data. Image analysis is a core requirement for these users and with this in mind, we will be deploying dedicated compute facilities for the researchers to utilise to process the images.

3.1 Dementia exemplar specific requirements

The dementia exemplar involves researchers from the Institute of Neurology in London, and from University College London, Imperial Collage, The University of Newcastle, Cambridge University and the Hammersmith Hospital, London. The dementia exemplar differs from the psychosis and stroke exemplars in that one of its primary aims is to collect a new dataset as part of this project from to utilise this data and the process to develop methods of measuring image quality whilst the patient is still in the scanner. This process will require consideration of the challenges of acquiring data from a clinical environment and of processing raw patient information rapidly within these constraints. The data to be collected will include baseline demographic data (age, gender, but not any identifiable data), digital scans for each of the time-steps and outcome information about these cases, associated with each timestep.

The requirements for an infrastructure to support this activity will additionally need to consider means of rapidly determining image quality through the grid to ensure that images may be analysed rapidly whilst a patient is still in the scanner to determine whether images should be retaken before the patient leaves. Key to these requirements will be developing methods of automation for the process include workflow to ensure that repetitive steps may be reproduced without extensive use interaction.

3.2 Psychosis exemplar specific requirements

The psychosis exemplar involves researchers from the University of Oxford and from the University of Edinburgh. Each centre has data collected over many years, as technology has advanced, with the result that key image acquisition parameters have differed and researchers have developed various methods for managing and processing this data over the years. This data resides on fire-walled servers managed by internal staff and is typically stored in directories. The data is viewed using various methods, most commonly SPM99/2 and FMRIB [2] for structural and functional MRI.

Raw data is processed and usually this processed data is kept, but not the data from the intermediary steps as it takes seconds to process. Often this processed data is stored in new directories to keep the original data 'pure'. Image processing techniques like the brain extraction process (BET), are very quick and intermediary data resides temporarily in a scratch area. Due to scanner variations, there is a need to homogenize the data. Through work on 20 healthy controls scanned twice each on two different scanners, as part of the Edinburgh High Risk Study [3], we have been able to develop a number of metrics of image quantities (e.g. signal/contrast to noise, partial volume, entropy) and techniques to improve the comparability of these across scanners.

The additional requirements for the psychosis exemplar include the need to look at elements of standardisation across studies, including the definition of a psychosis ontology and the determination of other potential datasets which may be utilised to extend the research validation already performed. Ideally, raw datasets will not need to be shared, as it will be possible for distant analysis to be performed on the raw data and only the results returned to the requester; although this requires consideration of what the processing actually returns as this may inadvertently provide a requester with a means of generating the raw data e.g. image reversal. Looking at the needs of this research community for more automated methods of working and potential workflow capability, the user requirement has been articulated as; '*What we could use*

is a piece of software that knows where all the files are - an automated file manager/ labeling tool/ tracker type thing, so that you can run a piece of code on some 'before_data' and the code will keep track of 'before' and 'after' locations and the version of the code that was used to create it. I envisage a piece of software that allows you to select the 'before data', select some processes to run in a sequence and then lets you choose a suitable location for the results, keeping track of the data and process from beginning to end. This would be able to call either Matlab or c executables, etc, and handle the redirection of the output and temporary files.'

3.3 Stroke exemplar specific requirements

The stroke exemplar, undertaken by the University of Edinburgh and The University of Nottingham, will utilise existing evolving datasets from the ENOS [4] and IST3 clinical trials, which are looking at the affects of specific treatments on early stroke cases. Key additional requirements for this exemplar involve the grid enabling of the reading tool that has been developed by Edinburgh to work across the data from both trials, methods for the archiving of the data to ensure continuous availability of service and a means of future proofing the storage and access of the datasets. Workflow needs are believed to be simple at this stage and include DICOM anonymisation (header stripping) and a DICOM quality control step to ensure sensible preset window levels, as often the levels stored in the headers are incorrect This process currently involves a user displaying images using a web application and and ensuring that they can 'see' the images and that the initial window levels allow a suitable starting point for reviewing the images. They can adjust the settings and then store them as future defaults. When the DICOM headers hold strange values for window width and level, the images often render to screen as pure black which can lead the reviewer (remote user) to believe that infact no image is actually there. We would look to automating this process to ensure that window levels are create a good spread across the 8bit grey scale.

4.0 Technology development

The technological development of the NeuroGrid system has three main elements. There is the federated file store, the federated database and the workflow management system. Together they will form a powerful tool for neuroimaging research. An orthogonal set of concerns relate to security and access control without which it would not be possible to share data and algorithms [5, 6]. The system is built using service orientated architecture principles. Each node in the NeuroGrid provides a web service interface both to other nodes in the system and to the client applications which will be produced as part of the project. Access to the data and services on the NeuroGrid is controlled by using X509 certificate based authentication. Certificates are published by the NeuroGrid certificate authority which only grants certificates to trusted individuals. To secure the SOAP messages sent to and from the web services the messages are both signed and encrypted before being sent. The signature insures that the message really was sent by the sender and the encryption ensures the message can only be read by the receiver.

To enable the efficient transfer of large image files each node also hosts a web server which utilises the webDAV extensions to the HTTP protocol. These extensions enable both the reading and writing of data, enabling clients to upload data directly to specially designated areas on the web server. Two methods are used to ensure the security of the files hosted by the web server. First mutual authentication is used to establish a secure connection using the SSL protocol. As the authentication is mutual the server is identified to the client and the client is identified to the server. Again the same X509 certificates are used as was the case for the SOAP messages. Once the connection is established all communication will be encrypted. The second method of securing the files ensures that users cannot read each others files until they have been published to the system. To ensure this, each user can create its own private area on the web server which they can then copy files into from either a client machine or from files they have access to in the system. After processing the files the user can then publish them so that others can use them.

While access to medical images is useful it is only when an image is put into context that you can fully utilise it. To this end it is essential to store information about each of the files, which is why a federated database is needed. Each study will produce a different set of metadata about its images, when the studies are from different clinical areas then the differences will grow. However it is important that a common core of metadata is identified. Once this is done it becomes possible to search all of the data on the NeuroGrid to find cases of interest. In addition exemplar specific information can also be stored for some files. This will allow more specific queries to be made about a subset of the images. Once you have found the sets of images that you are interested in you may wish to analyse them using a range of neuroimaging algorithms. Much of the technical effort of NeuroGrid will be dedicated to producing a workflow management system.

There are three main parts to the workflow management system, they are the web portal, the web service generator and the workflow management engine.

The web portal will provide a user interface to NeuroGrid. It will mask the complexity of the distributed

environment from the users while providing them with a full set of functionality. As was the case with the web server for the federated file system the users will be authenticated using their NeuroGrid X509 certificate.

The web service generator will allow a user to take an algorithm which they have developed and to publish it as a web service. The web service generator will orchestrate the generation, deployment and installation of the web service, as well as providing a web service client that will be able to communicate with the generated web service. While access to all of the algorithms that are published by the users of the NeuroGrid will be invaluable, they become even more useful if they can be joined together. By allowing users to join algorithms together by producing a workflow the portal will provide even more functionality to the users.

Once the user has produced a workflow using the portal, it can then be executed on a specified set of files using the workflow management engine. This engine is responsible for the definition and extraction of the individual work units, their dispatching to the proper components, and also ensuring that the entire set of tasks is properly completed. In the case of failure, error messages will be returned to the user.

As was stated above, much of the data that is being published on NeuroGrid can only be published because access to the data is restricted in a way that is in accordance with the ethical clearance for the data. In addition data owners are free to add their own restrictions as they see fit. Not all restrictions are simple to define.Some users can only allow their data to be used with certain algorithms or workflows. Requirements such as this require a flexible access control system that can support complex policies. To ensure the autonomy of the data owners, the access control policies will be stored with the data they protect. There will not be one central access control system, but rather a federated access control system. When a new node joins the NeuroGrid it brings with it its own access control policies along with its images and metadata.

To date the project has implemented its 8 data nodes across the UK. These nodes are administered remotely and are secured with certificates issued by the project administrator. The core architecture of the server nodes has been designed, developed and tested. This initial prototype supports image storage, image retrieval, cataloguing of images, certificate based authentication and secure data transfer. The core schema deployed allows for all modalities to allow for the CT scans and MRI images acquired to be stored with the relevant metadata. Exemplar specific schemas will supplement this core information and will evolve over time.

We intend developing on this prototype over the coming months and adding to its functionality in an iterative fashion to include publishing and querying of image metadata, federation of image retrieval, federation of metadata queries and user-defined access control policies. The intention is that these features will be incorporated gradually and will build on the stable base provided by the initial prototype

5.0 The social challenges of collaborative research

NeuroGrid brings together disparate groups of clinicians, technologists and researchers from across the UK and in many instances, individuals had no prior working experience of large scale collaborative research or of the individuals collaborating on this grant. This in itself proved to be a challenge, and requires extensive activity and firm, but flexible, project management to bring together these disparate people to agree common goals and develop an effective working relationship. NeuroGrid has attempted to overcome these problems over the first ten months by establishing clear roles and responsibilities, project structure and methods for communication through a steering committee, technical steering committee, regular reviews and workshops, as well as using video conferencing on the desktop through personal 'access grids' to enable busy clinicians to play an active part in these reviews without having to travel. It has been essential to ensure that project members feel part of a 'virtual team' [7] and mechanisms like a project repository, a regular newsletter and group publications are being utilized effectively to enable this. Common problems across the clinical groups are also evolving and further collaboration will be encouraged through the use of workshops and special interest activity to resolve common issues. The project has recently instigated a 'share ideas' area on its collaborative webspace to encourage disparate scientists to suggest tools and tricks in medical imaging and clinical science that could help other collaborators. Many of these resonate with those of other Grid projects in areas such as data quality, security, data ownership, confidentiality, IPR, ethics, and the management of clinical trials.

Collaboration across the very different communities of interest [8] depends on finding ways of ensuring early engagement and dialogue, so that negotiation of diverse aims and requirements can inform the design process as early as possible. The creation of real and virtual 'shared spaces' [9] is intended as means of supporting this new hybrid community develop its own rules of engagement, and start making collective sense of local knowledge and requirements in relation to project goals. Trade-offs may need to be negotiated in areas such as security, data access and data quality, metadata and ontologies, data ownership and IPR, as well as the problems of securing patient consent for multiple uses Special interest groups of a

more transient nature are also being set up where there is a potential for aligning aims and knowledge in relation to a shared problem,or a shared objective. As lines of communication open up, there are new opportunities for brokerage [10] and alliancing between communities that present both potential benefits and challenges.

6.0 Discussion

The NeuroGrid project faced a challenging start with so many disparate groups of clinical researchers and a newly formed technology team, and little or no experience of how each group worked. Key to any early success was the development of a common understanding of the project requirements and objectives and an understanding of the needs of an initial baseline prototype system from which the project could iterate towards the more complex and visionary needs alluded to in the proposal. By developing an early prototype that enables the project centres, or 'nodes', to store and retrieve images and to move data around securely using certificate based authentication, the team aims to progress to developing more complex services for managing image analysis and compute capability, as well as developing workflow techniques and the ability to manage bespoke services for data analysis. This prototype aims to stimulate ideas for enabling technology for collaboration.

The toolkit aspect of the project will be developing a web portal which will be the NeuroGrid user interface. This is required to mask the complexity of the distributed environment from users while providing fully distributed functionality. All the support services will be integrated or accessible through the portal. The team will also implement a web service generator which provides many web related tools to create web services and web service clients for various applications. A workflow management system will be responsible for the definition and extraction of the individual work units, their dispatching to the proper components, and ensuring that the entire set of tasks is properly completed.

The grid connectivity team will be extending the physical architecture and middleware services to encompass the complex data and compute needs of a grid for neurosciences. NeuroGrid aims to build on this early prototype by working with the clinical partners to populate the data nodes with ethically cleared data, and exploring the process for data storage and retrieval. By initially allowing the clinical groups to explore the local use of the data and how this compliments their local research, we will then explore the move to sharing, or federating the data, initial with small amounts of information and finally looking at the deployment of additional infrastructure to support data replication for system failover ensuring resilience for the clinical exemplars. In addition to the data grid development, the project will be implementing compute facilities on this grid in the next 12 months to allow researchers to perform analysis on local, remote or federated datasets by utilising shared compute facilities. Again security will be highly important in this deployment and the team aim to have implemented an infrastructure for neuroscience research to support the partnering organizations by the end of the project in 2008 with the ability to extend to new research groups in the future. Complimenting the technology development will the closer engagement of the clinical research groups who will be able to explore new opportunities through the use of this technology.

7.0 References

[1] Neurogrid website – www.neurogrid.ac.uk

[2] FMRIB libraries - http://www.fmrib.ox.ac.uk/fsl/fdt/

[3] Job DE, Whalley HC, Johnstone EC, Lawrie SM. (2005) Grey matter changes over time in high risk subjects developing schizophrenia. Neuroimage. 2005 May 1;25(4):1023-30.

[4] Bath PM. Major ongoing stroke trials. Efficacy of Nitric Oxide in Stroke (ENOS) trial. *Stroke* 2001;32:2450-2451 (abstract)

[5] D. J. Power, E. A. Politou, M. A. Slaymaker, and A. C. Simpson. Towards secure grid-enabled healthcare. */Software: Practice and Experience/*, 35(9):857-871, 2005.

[6] D. J. Power, E. A. Politou, M. A. Slaymaker, and A. C. Simpson. Securing web services for deployment in health grids. /Accepted for publication in Future Generation Computer Systems/, 2005.

[7] Zakaria, Norhayati, Amelinckx, Andrea & Wilemon, David (2004) Working Together Apart? Building a Knowledge-Sharing Culture for Global Virtual Teams. *Creativity and Innovation Management* 13 (1), 15-29. doi:10.1111/j.1467-8691.2004.00290.x

[8] Wenger E. and Snyder W. 2002, Communities of practice: the organizational frontier, *Harvard business review,* Jan/Feb, 139-145

[9] Nonaka I. & Nishiguchi T. 2001, *Knowledge Emergence: Social,Ttechnical and Evolutionary Dimensions of Knowledge Creation, OUP*

[10] Burt R.S. 2001, Structural holes versus network closure as social capital. In Lin N., Cook., K. and Burt R.S. (eds) *Social Capital, Theory and Research.* Walter, New York