Archives Italiennes de Biologie, 148: 207-217, 2010.

Data warehousing methods and processing infrastructure for brain recovery research

T. GEE[1], S. KENNY[2], C.J. PRICE[3], M.L. SEGHIER[3], S.L. SMALL[2],
A.P. LEFF[3,4], A. PACURAR[1], S.C. STROTHER[1,5]

*[1] Rotman Research Institute and Centre for Stroke Recovery, Baycrest, Toronto, Canada;
[2] Computation Institute, University of Chicago, USA; [3] Wellcome Trust Centre for Neuro-Imaging,
Institute of Neurology, University College London, UK; [4] Institute of Cognitive Neuroscience,
University College London, UK; [5] Medical Biophysics Department and Institute
of Medical Sciences, University of Toronto, Canada*

ABSTRACT

In order to accelerate translational neuroscience with the goal of improving clinical care it has become important to support rapid accumulation and analysis of large, heterogeneous neuroimaging samples and their metadata from both normal control and patient groups. We propose a multi-centre, multinational approach to accelerate the data mining of large samples and facilitate data-led clinical translation of neuroimaging results in stroke. Such data-driven approaches are likely to have an early impact on clinically relevant brain recovery while we simultaneously pursue the much more challenging model-based approaches that depend on a deep understanding of the complex neural circuitry and physiological processes that support brain function and recovery. We present a brief overview of three (potentially converging) approaches to neuroimaging data warehousing and processing that aim to support these diverse methods for facilitating prediction of cognitive and behavioral recovery after stroke, or other types of brain injury or disease.

Key words
Neuroinformatics • Database • Stroke • Research • Data mining

Introduction

The task of managing, storing and maintaining large datasets for stroke recovery research requires a combination of data-management techniques. Many of these approaches are also being intensively developed in the context of neuroinformatics infrastructure for multi-centre clinical trials for other brain disorders (Van Horn and Toga, 2009a), although there has been some resistance among the brain imaging science community to adopting the large-scale neuroinformatics infrastructures now available (Van Horn and Toga, 2009b). This trend is part of a larger movement in the biological sciences called “bio-imaging informatics” (Peng, 2008). We present three types of database-centered infrastructure techniques that address different aspects of the stroke research community’s needs, and briefly discuss the idea that some combination of all these approaches is necessary for answering important questions about how the brain recovers from injury or illness. Our overall aim is to integrate these three database and processing frameworks across our multi-national centres to accelerate translational neuroscience with the goal of directly improving clinical care and recovery following stroke.

First we present the Predicting Language Outcome and Recovery after Stroke System developed in Britain (Price et al., 2010). The PLORAS system is used to make predictions on the basis of the recov-

Corresponding Author: Tom Gee, 3560 Bathurst Street, Toronto, ON, Canada M6A 2E1 - Email: tgee@rotman-baycrest.on.ca

ery of previous patients with similar lesions, using a database that records the speech and language deficits, and recovery of a broad range of patients who suffered focal brain lesions producing aphasia. The system quantitatively categorizes a lesion and compares it against all others in the database to predict longitudinal language outcomes based on similar patients in the database.

The second uses the Extensible Neuroimaging Analysis Toolkit (XNAT) as a core database framework (Marcus et al., 2007). This has been extended to support and address the multiple problems of sharing a broad range of heterogeneous data sources and/or preexisting databases in the Stroke Patient Research Recovery Database (SPReD) as part of the multi-institution Centre for Stroke Recovery in Canada. Input data sources include direct data input and databases of stroke cognition testing, motor rehabilitation testing, the separate XNAT-based Rotman Research Institute's neuroimaging database (RRINiD), and the Prospective Urban Rural Epidemiologic MRI study designed to determine the prevalence of covert cerebral ischemia in urban and rural settings in Canada. Through the RRINiD, SPReD is coupled to predictive modeling of systems level functional connectivity for large amounts of fMRI and MRI data using the workflow management portal for high-performance computing provided by the Canadian Brain Imaging Network (CBRAIN) (Rousseau et al., 2009).

Finally, section three presents a natural extension to the aforementioned file-based database systems: the time series database approach implemented by the Computational Neuroscience Applications Research Infrastructure (CNARI) in the USA. This incorporates novel methods for maintaining, serving, and analyzing large amounts of fMRI data with a focus on workflow management and parallel computing coupled to a database infrastructure (Small, 2009). A unique feature is the storage of fine-grained neuroimaging features (e.g., voxels) as elements in the database that may be manipulated in highly parallel processing pipelines using SQL commands in conjunction with other analysis tools. An initial focus has been on neural network modeling and effective connectivity using exhaustive searches of the model space for restricted sets of network nodes (Kenny, 2009).

Each of these approaches has a central focus on enhancing stroke data sharing, and the coupling of this with different innovative capabilities for neuroimage data analysis that will accelerate translational neuroscience for better prediction and understanding of brain networks during recovery within the Brain Network Recovery Group (http://www.brainnrg.org/).

PLORAS: a database for categorizing brain lesions

The aim of the PLORAS database (Price et al., 2010) is to Predict Language Outcome and Recovery After Stroke but the principles behind the procedures could be applied to any other cognitive and sensormotor skills that can be impaired after stroke. The ability to predict how a patient recovers is important because impairment of any key neurological domain can have a devastating effect on carrying out activities of everyday life. Patients and their clinicians need realistic expectations as to how they will recover and what the most effective treatment will be. However, predicting outcome after stroke is notoriously challenging because there is wide variability in how patients recover and a lack of understanding of how the lesion site predicts speech (and other) difficulties. Nevertheless, there is already a long list of potential factors that influence recovery other than lesion site, such as the age of the participant, the time post stroke, medical co-morbidities, educational level, and their motivation and ability to attend. To understand how all these factors integrate together, we need to combine data from multiple patients and multiple sources.

The PLORAS database consists of structural, behavioural and demographic data from a wide range of stroke patients. Information on the lesion site is determined by high-resolution structural MRI; cognitive and language abilities (determined by multiple behavioural tests), and key demographics: age, handedness, nationality, languages spoken, time since stroke, and others. A subset of patients have undergone functional MRI) and this allows us to include information about the patients' ability to activate different brain areas during a range of perceptual, language and motor tasks. All data are fully anonymized. The database has a number of different uses because it can be searched by different criteria, e.g., by behavioral deficits to find the associated lesion site, or by lesion site to predict the behaviour (see Price et al., this issue for how these searches are integrated). Here we will focus on how the database can be developed for use in the clinical environment. From a practical perspective, outcome and recovery predictions after stroke can be provided by clinicians at distal sites uploading a single structural MRI to the PLORAS website (currently under development) and including a few key variables, e.g. neurological domain(s) affected, age. The structural MRI is converted into a 3D "lesion image", which indexes the degree of abnormality at each voxel of the brain in a standard space. The details from the lesion image are then used to search the PLORAS database for patients with similar lesions. The details of all patients with corresponding lesions are then summarised into graphical outputs that illustrate how these patients recovered over time (see Price et al., 2010). The predictions are probabilitistic (mean and standard deviation), and thus their precision can be immediately appreciated. This is important because outcome and recovery are much more consistent after some lesions than others.

From a technical perspective, the success of the predictions depends on a number of factors. The two most obvious factors are the number of patients with similar lesions in the database; and the precision in the lesion similarity measures. More specifically, if the number of "corresponding" patients in the database is small, then the estimation of the mean and standard deviation of their recovery will be imprecise and the influence of other factors (age, time post stroke etc.) may not be available. The precision of the similarity measures depends on the quality of the original structural MRI, the success of the spatial normalisation procedures and the estimation of the structural abnormality at each voxel. The accuracy of the predictions can, nevertheless, be constrained by prior knowledge using Bayesian approaches. The PLORAS procedures can therefore capitalize on recent developments in the use of predictive modelling, using, for example, support vector machines and multivariate lesion analyses. These procedures allow us to bias the whole brain lesion descriptions, towards the parts of the lesion that we expect to have greatest predictive validity. The degree to which our language recovery predictions are improved by the inclusion of prior knowledge can be assessed using Bayesian model comparison.

Our understanding of which brain areas and white matter connections are likely to be important for our structural predictions is developing at an exponential rate. More specifically, we are using the PLORAS database to generate the "structural features" that will be embedded into the "lesion similarity measures". For example, in Price et al. (this issue), we correlated lesion data with behavioural outcome to identify those lesion sites that are most strongly associated with difficulty gesturing the use of an object (i.e. lesion-symptom mapping). We are also correlating structural MRI data with functional MRI data to ascertain how patients recover specific language skills following damage to key language areas. In summary, the key data features for the lesion similarity measures include specific cortical areas that have been functionally defined by functional MRI or lesion-symptom mapping; white matter tracts that have been identified by lesion-symptom mapping; and higher order combinations of cortical areas and white matter tracts that cause more damage than would be expected from the sum of damage to each component alone (the principle of degeneracy).

The PLORAS database can also be used to investigate the non-lesion factors that affect prediction. For some lesions, we expect the variability between patients to be small. In these cases, the precision of our predictions will be good. For other lesions, there may be considerable variability in the course of recovery. This is being investigated by examining the effects of non-lesion factors such as age, handedness, hours of speech and language therapy, educational attainment, motivation, vision, hearing, and attention. Again, the predictive value of these measures will be based on Bayesian model comparison and assessment of generalisation error by splitting extant patients into test and training groups.

SPReD: a web-based database for heterogeneous data-sharing using XNAT

The Stroke Patient Research Recovery Database (SPReD) within the Centre for Stroke Recovery (http://heartandstroke-centrestrokerecovery.ca/our-research/spred) acts as a flexible hub that receives heterogeneous stroke-related data from diverse clinical and research projects and integrates the results into a navigable data set that may be shared, augmented and redistributed. The core of SPReD leverages the XNAT neuroinformatics platform (Marcus et al., 2007) that provides a flexible data model, a rich web-based user interface, and an expressive low-level interface for use by auxiliary software. In addition to XNAT, a set of modular interface programs have been developed to adapt SPReD to the unique requirements of each of the projects and processing pipelines with which SPReD interacts and to facilitate incorporation of new projects. Existing and potential project and resource connections to SPReD are illustrated in Fig. 1.

There are four primary functions within SPReD: (1) adaptation, whereby the system conforms to the requirements of a particular research project or database-processing resource such as PLORAS (see above), CNARI or the Virtual Brain (see below and McIntosh et al. this issue); (2) integration, where data of diverse organization and content from multiple sources may be unified along common factors using ontologies (Bug et al., 2008; Bilder et al., 2009; Konstantinos et al., 2009); (3) sharing, so that the integrated data may be made available to a wide audience of researchers for collaboration and data mining while satisfying privacy constraints; (4) navigation, which provides the searching, browsing and export functions required to identify and extract desired data.

As shown in Fig. 1, SPReD is currently connected with or being adapted to the:

– *Rehabilitation Affiliates*, a multi-centre rehabilitation project distributed across six hospital sites in Ontario, and the data collected emphasizes cognitive, physical and emotional assessments, with current uploads via local Access databases and Excel spreadsheets.

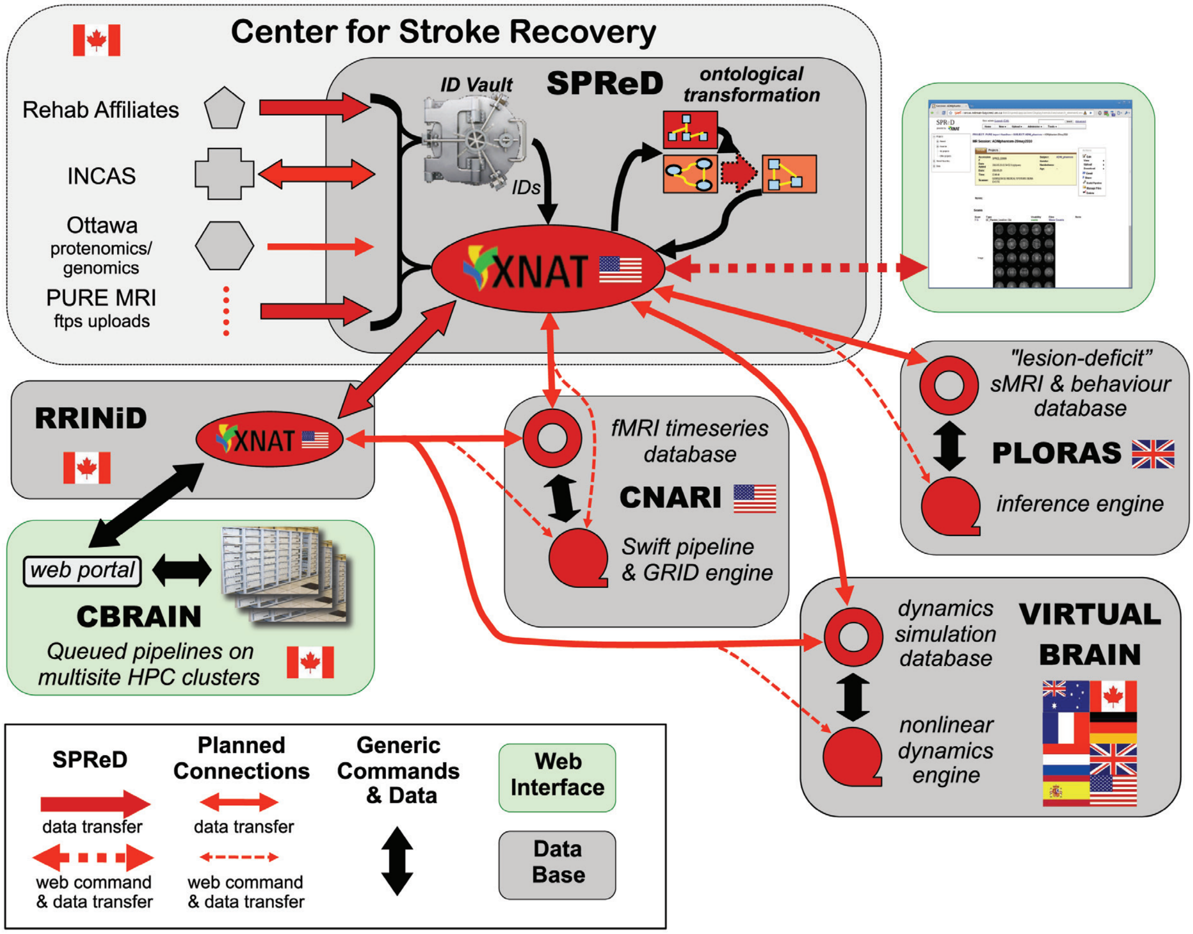

Fig. 1. - Schematic outline of the SPReD database with current and proposed links to other data repositories and processing resources.

- *Integrated Neurocognitive assessment system* (INCAS), which provides clinical management and assessment of stroke patients at three Center for Stroke Recovery hospitals using a behavioural battery based on the NIH consensus battery (Hachinski et al., 2006), and tested using a computer tablet that feeds results to a MySQL database and a DICOM server.
- *Prospective Urban Rural Epidemiologic* (PURE)-MRI project, which is collecting approximately 1000 subjects' MRI scans to study covert stroke using the Canadian cohort within the PURE study, a multinational study examining determinants of cardiovascular disease coordinated by the Population Health Research Institute (http://www.phri.ca/body.cfm?id=514) – SPReD provides the neuroimaging database facility with DICOM data uploads from four Canadian collection sites, download and analysis at the Seaman MRI Centre in Calgary.
- *Rotman Research Institute's Neuroimaging Database* (RRINiD) which manages all neuroimaging studies from research projects at the Rotman Research Institute (http://www.rotman-baycrest.on.ca/index.php?section=532) with data flows to SPReD through a Representational State Transfer (REST) interface. The RRINiD is connected to Canadian high-performance computing sites through the Canadian Brain Imaging Network (CBRAIN) web portal (http://cbrain.mcgill.ca/), a platform currently supported by five neuroimaging centres that provides distributed processing, analysis, exchange and visualisation with the goal of rendering the processing environment transparent to a remote user (Rousseau et al., 2009).
- *Genomic and Proteomic* data sequences generated by the Ontario Hospital Research Institute in Ottawa on Centre for Stroke Recovery subjects who may also be included in the Rehabilitation Affiliates or other SPReD project collections.

Each input site stores a variety of data in a variety of structures. To adapt to these, SPReD makes use of adaptable formats such as the extensible markup language (XML) and XML schema description (XSD), the latter largely implemented by XNAT. SPReD also adapts its communication structure using modular interface programs to obtain and store the data from each site. In addition, by extending the XNAT pipeline architecture, we are currently testing an interface between the CBRAIN portal and the RRINiD. We plan to use this interface to provide the capability for SPReD to also initiate and receive results from high performance computing environments via CBRAIN (Rousseau et al., 2009) and CNARI (see below). For example, we plan to utilize this high-performance computing interface via the CBRAIN portal to extensively test the relationship between complexity measures based on fMRI functional connectivity and behavioural recovery following stroke (see Yourganov et al., this issue).

Integration of the data involves two main factors, identification of common subjects and unification of common categories. The input sites practice only internal subject coordination (e.g., only project- or site-specific subject identifiers), thus preventing aggregation of an individual subject's data should the person move among projects. SPReD has the ability to identify and integrate diverse project data from a single common subject with a high degree of certainty in the identification, while satisfying privacy concerns and legislation (e.g., Canadian Panel on Research Ethics, 1998) by not exposing personally-identifying information. Identifying information such as name and date of birth are replaced with unique cryptographic hashes of those values, which enables common subjects to be identified but does not permit the original values to be retrieved, thus preventing personal identification. By adding a random numeric or salt value to the calculation and protecting the values in a secure location, we avoid the small risk of re-identification through dictionary attack (Mironov, 2005). In Fig. 1 this secure location is represented by the ID Vault, a tightly-secured computer with a hardware-encrypted disk connected only to the SPReD server via an encrypted private network.

Data sharing in SPReD is implemented through XNAT combined with custom extensions. Data is organized into individual projects, where each project has any number of users who may be designated as owners, collaborators or members, each with differing roles and permissions. SPReD adds to XNAT the ability to anonymize data – stripping DICOM header fields and removing facial MRI features – and to export limited subsets of the data set. Ownership and sharing of the original data remains with the principal investigator of each contributing project, but new projects accessible to other members (e.g. all Centre for Stroke Recovery and Brain Network Recovery Group members) may be defined to contain selected subsets of data from multiple projects (e.g. resting state data from multiple stroke projects).

To achieve further integration by unification of common categories, SPReD extends the schema functionality of XNAT to capture high-level information about the relationships between the data elements (an ontology.) This information is processed via an inferential query engine to identify and integrate common categories of data, making them amenable to searching. For example, the integrated neurocognitive assessment system may record that a subject was prescribed a specific dose of warfarin, while Rehabilitation Affiliates simply notes whether the subject was taking anticoagulants; SPReD would allow the former subject to be retrieved even while searching for subjects via the latter description ("show subjects taking anticoagulants"). SPReD preserves the original structure and content of the input site's data, and allows complex integration ontologies to be developed through stepwise refinement, as time, motivation and expertise are available. This process is facilitated through the use of existing technologies developed for the Semantic Web (Berners-Lee et al., 2001), including the resource description framework (RDF) data model (RDF, W3C, 2004), ontological formalization through the OWL web ontological language (OWL, W3C, 2009) and the RDF query language SPARQL (SPARQL, W3C, 2008). We are starting to employ these technologies through the integrated knowledge-base system Protégé (Gennari et al., 2003). The development of standardized neuroinformatics ontologies has been underway for several years (Bug et al., 2008) and SPReD benefits from that work by employing the same foundational technology.The flexible search and browsing features of XNAT enable navigation of data in SPReD, with a small extension for generating aggregate statistics. When custom types are defined within XNAT, they can include specific search and browsing web page definitions, which allows comprehensible displays to be created. To assist in data mining while preserving ethics-mandated requirements for sharing only consented information, SPReD extends that search capability to allow users to search on data that is not otherwise accessible to them, returning simply the number of matches along with the source of the matching data but without additional details, i.e., the actual data is not shared. The searching user may then contact the owner of the matching data and request access. This enables a researcher to determine if there are sufficient subjects matching specific criteria to justify approaching other researchers and arranging ethics amendments and other regulatory requirements. Finally XNAT allows the accessible information to be downloaded in several different formats, including XML, zip and tar archives.

Planned integration: PLORAS + SPReD

The planned integration between PLORAS and SPReD will operate in several phases. Our initial goal is to develop and test a simple SPReD pipeline that sends one or more, de-identified aphasic stroke MRIs from SPReD or CNARI (see below) to query PLORAS' inference engine for recovery probabilities, and returns these probabilities and associated measures to SPReD for incorporation into the patient's record. This will require testing of the lesion segmentation tools available within PLORAS (Seghier et al., 2008) on MRIs from SPReD and CNARI, and development of metrics that measure if the tools have performed sufficiently well to allow comparison with the existing database. We will facilitate this testing by implementing the tools as pipelines that may be run directly on MRIs within SPReD. Such testing and querying of the PLORAS database and inference engine will help to augment the PLORAS sample size and prediction accuracy for MRIs that pass the quality metrics and can be shared. Second, we will develop a joint aphasic-stroke schema modeled on that used by PLORAS and in coordination with existing stroke and behavioural schemas in SPReD and CNARI. Our goal here is to create a merged schema for aphasic stroke, which can form the basis of a future stroke ontology, and be used to enable bidirectional sharing of full data sets, including structural and functional neuroimaging and all associated meta data. Such sharing of complete data sets between our three sites with large research efforts directed at stroke recovery will facilitate targeted testing of the additional predictive power of site-specific measures available from Integrated Neurocognitive Assessment System and Rehabilitation Affiliates in SPReD, and CNARI. We believe that such sharing, augmented by state-of-the-art data warehousing, and processing pipelines will help to rapidly narrow the range of imaging tests and measures most effective in prediction of stroke recovery, initially in aphasia, and eventually in other types of behavioral recovery outcomes.

Computational Neuroscience Applications Research Infrastructure (CNARI): time series data mining on the Grid

XNAT-enabled neuroscience portal

Science gateways are a means of allowing groups of researchers and large collaborations to share processing power, workflow management systems, and analysis techniques. Gateways are well-established in the field of Grid computing (Welch, 2006; Adolphs 2007; Scavo, 2007) and CNARI is currently part of a science gateway to the TeraGrid (Catlett, 2007). As noted above a primary goal of integrating our three database frameworks is to develop an infrastructure that will allow multiple, international collaborating institutions to manage and execute processing workflows from an XNAT repository maintained as part of the SPReD collaboration, and feed analysis results back into the repository for visualization and sharing.

A central component of CNARI is the Swift workflow management system (Zhao, 2007). XNAT lends itself well to integration with Swift, as well as with CBRAIN, in that there is existing support for cluster execution via XNAT's pipeline engine, which can be extended to execute Swift scripts that are either automatically generated on the portal or constructed by the user. This results in an integrated infrastructure that is both a repository and a portal.

Motivation for time series databases in neuroimaging

A natural supplement to systems for storing and sharing image files in stroke research are time series databases such as used in CNARI. With a time series database researchers can process imaging data using highly-specific search criteria to run analyses directly on the brains, regions, voxels or surface-mesh vertices using a shared data base management system (Small, 2009) without the need for downloading individual image files. In particular, in a processing environment where downloaded files will necessarily undergo transformation and likely be farmed out to a remote cluster for processing, downloading becomes inefficient. As part of our planned integration of PLORAS, SPReD and CNARI, we will utilize the strengths of all three systems by developing both a repository that allows for archiving, and distributed processing of the flat-files in the repository (e.g., using the SPReD-CBRAIN interface), and user-generated time series tables.

For example, a set of user-generated time series tables might include signal values for each voxel at each time point along with their associated regions of interest and t-statistics for a given scan or set of subject scans. A workflow can then be constructed to extract the time series of voxels that showed a particular level of activity and are associated with selected regions of interest; Hasson (2008a) gives a detailed example of such a workflow. Coupling archiving and storage with data processing reduces the need for data transfer and, perhaps more significantly, the need for the user to store an entire dataset on his local file system in order to run an analysis on it. For example, in the case of flat files, performing an analysis over all subjects in an experiment would require specifying the processing workflow and its input files directly on the portal and then launching the workflow, to be processed on a remote cluster (e.g. the TeraGrid or CBRAIN). The final stage of the workflow would include uploading the results from the compute resource back into the repository so that they would be available for download, visualization or further processing.

If a user wishes to run an analysis on imaging data stored in a time series database, workflow specification also includes generation of the database tables. Input tables are populated from files within the repository and results tables are populated by the processing jobs themselves. With remote processes operating directly on the tables there is no need to transfer complete image files to and from the compute resource and both cases obviate the need for the user to directly download imaging data. Thus, while flat file repositories are vital for archiving and meta-analysis, time series databases that represent signal values at each time point for each voxel or vertex in a scan are well suited to the mining of patterns making a hybrid system potentially more flexible and powerful than either approach by itself.

Time series databases have been in wide use for some time in other communities, such as financial trading (Chandra, 1993) and weather information systems (Goodall, 2007). Furthermore, they lend themselves well to various web services. For instance, Goodall (2007) demonstrated a suite of tools for mining the national weather information system time series database, which collects 800,000 observations of ground water level on a daily basis. Users can then capture subsets of this based on a web form submitted directly from the site or using a variety of web services. The web form generates queries across a large time series database and returns the results to users in the form of temporary tables, which can then be fed into locally available statistical tools. This provides a reasonable prototype for combining the efforts of SPReD with those of CNARI into a unified resource via a web interface that would be suited to probing issues such as effective connectivity for predefined groups of regions during recovery from stroke (James et al., 2009). For a detailed description of how such structural equation modeling has been implemented within the CNARI framework, please see (Kenny et al., 2009), which is summarized below.

Integration: CNARI + SPReD

Schemas that optimize both data importation and mining have become vital to some important parallelized, database-centric workflows in the neuroimaging community. For example, users can execute a workflow that pulls signal values for batches of vertices, in parallel from a database holding surface data indexed by vertex (Skipper, 2007; Hasson, 2008b) for rapidly running statistical tests. Integrating a time series database with SPReD creates additional possibilities for both single session and longitudinal stroke studies utilizing novel mining techniques and exploratory queries that look for patterns (Andric, 2009) while lending itself to fast parallel analysis workflows (Kenny, 2009; Small, 2009).

The CNARI stroke data ontology includes fMRI, physiological measures, DTI data and stimulus features, with the aim of providing a platform for the fine-grained, exploratory analysis workflows to be coupled with querying across multiple measures simultaneously. Because storage of all of this data in a DBMS requires a great deal of space, we have chosen a hybrid architecture whereby data can be exported to a database on an as-needed basis. Specifically, XNAT allows for users to create customized processing pipelines and our CNARI-SPReD implementation of XNAT will offer, as a standard processing pipeline, the transformation of file data into temporary, compressed, time series database tables that would enable distributed, database-centric processing. Such a pipeline would also include tools for re-importing the data into the repository. If the output from such a workflow results in another table, which can be written to file and stored in the repository (and subsequently transformed back to a table should the user require it). This infrastructure would allow for fluid movement back and forth from time series database tables to XNAT-controlled files such that users can easily specify, via an XNAT pipeline, the input they require for their processing tools.

Currently within CNARI, processing is distributed using the Swift workflow management system where users write and execute Swift scripts, which are used to call arbitrary image processing or statistical software within a single script. Input files to the scripts generally sit on the local file system (though as far as Swift is concerned they can live anywhere). Users execute the script using a simple command-line interface usually running in a screen session where users can detach from the running process if it is a long-running workflow, and periodically check for results. While this method for running is rather simplistic it is useful and well tested. The next step is a more robust user interface such as integration with SPReD/XNAT. In our CNARI-SPReD implementation, both analysis (database-centric) and preprocessing (flat-files) would be run on remote clusters or Grids separate from the XNAT server repository, but launched and monitored from within it. Because CNARI already distributes its workflows over a local cluster and TeraGrid's TACC and NCSA sites, executed from the local file system, initiating execution from the XNAT repository seems a logical progression for that functionality.

Our initial work has specifically focused on fMRI data, in which we have enabled querying CNARI for activation from sets of nodes (or regions of interest) to examine connectivity within networks using structural equation modeling (Kenny, 2009). Applying this method to the analysis of stroke data includes comparing this network connectivity in stroke subjects to that of control subjects as well as comparing connectivity in stroke subjects at different time points for longitudinal analysis. Once a user has exported a time series to a database, CNARI can be used to optimize time series extraction by iterating over a) different criteria for node activation (e.g. using the maximum t-value from the node versus using the average) b) multiple subjects, c) multiple sets of nodes, and d) multiple experimental conditions, all of which would be specified in a single Swift script and run in parallel. This workflow can then be extended to extract behavioral measures (residing in the repository) that are associated with the given subject, scan session and experimental condition to explore relationships between the structural equation models and observed behavior.

Discussion

While PLORAS, SPReD and CNARI have each developed relatively separately they are all focused on facilitating data sharing and analysis with the common goal of understanding what may improve treatment and prediction of recovery from the devastating consequences of stroke. These three databasing approaches are quite complimentary, and we believe that linking them using SPReD-XNAT tools for managing and processing shared data collections from heterogeneous data-repositories will provide a coordinated sum that will almost immediately improve prediction of recovery, and greatly facilitate development of better prediction and treatment assessment tools. To our knowledge this will form the first such integrated, multi-national stroke database, initially focused on aphasia, but eventually on a broad range of possible outcomes. It is further unique in our ongoing efforts to provide integrated portals and parallelized processing pipelines with a focus on functional metrics (Yourganov et al., this issue) and effective connectivity measures (Kenny et al., 2009) in stroke recovery, which may further augment prediction over the basic measures available within PLORAS using segmented lesions from MRIs.

Finally, these integrated data storage and processing platforms will be linked to the proposed new simulation tool provided by the virtual brain database (See Fig. 1, and McIntosh et al., this issue). The central goal is to allow clinicians and researchers to access not only comprehensive data sets, processing tools and prediction results for recovery of brain function, but to augment these with a state-of-the-art simulation of the brain's systems-level, nonlinear dynamics based on structural and functional connectivity constraints from real data. The ultimate goal is to incorporate simulations of lesion results as part of our armamentarium for enhancing prediction of network recovery after brain damage and testing both virtual and real treatment options. The first step towards this goal is our proposed integration of the existing PLORAS, SPReD and CNARI databases and their associated processing platforms.

Acknowledgements

This work was partly supported by a James S. McDonnell Foundation, USA (Bridging Brain, Mind, and Behavior) grant to Dr. McIntosh. TG, AP and SCS gratefully acknowledge the support of the Heart and Stroke Foundation of Ontario through the Centre for Stroke Recovery for development of SPReD and RRINiD, and the CANARIE, Canada (Network-Enabled Platforms Program) for development of CBRAIN tools. SK and SLS also acknowledge the support of the National Institutes of Health under grant R33-DC008638-04.

References

Adolphs S., Bertenthal B., Boker S., Carter R., Greenhalgh C., Hereld M., Kenny S., Levow G., Papka M.E., Pridmore T. Integrating cyberinfrastructure into existing e-social science research. *Proceedings of the e-Social Science 2007 Conference*, 2007.

Andric M. and Small S.L. Hemodynamic signal fluctuation due to occurrence of cospeech gestures in audiovisual story comprehension. *Presented at the Neurobiology of Language Conference*, Chicago, Illinois, USA, 2009.

Berners-Lee T., Hendler J., Lassila O. The semantic web. *Scientific American*, May, 2001.

Bilder R.M., Sabb F.W., Parker D.S., Kalar D., Chu W.W., Fox J., Freimer N.B., Poldrack R.A. Cognitive ontologies for neuropsychiatric phenomics research. *Cogn. Neuropsychiatry*, **14** (4-5): 419-450, 2009.

Bug W.J., Ascoli G.A., Grethe J.S., Gupta A., Fennema-Notestine C., Laird A.R., Larson S.D., Rubin D., Shepherd G.M., Turner J.A., Martone M.E. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*, **6** (3): 175-194, 2008.

Catlett C. et al. TeraGrid: analysis of organization, system architecture, and middleware enabling new types of application. In: Grandinetti L. (Ed.) *HPC and grids in action*. Amsterdam, IOS Press, 2007.

Chandra R., Segev A., Managing temporal financial data in an extensible database. *Proceedings of the 19th International Conference on Very Large Data Bases*, 302-313, August 24-27, 1993.

Gennari J.H., Musen M.A., Fergerson R.W., Grosso W.E., Crubezy M., Eriksson H., Noy N.F., Tu S.E. The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, **58** (1), 2003.

Goodall J.L., Horsburgh J.S., Whiteaker T.L., Maidment D.R., Zaslavsky I. A first approach to web services for the National Water Information System. *Environmental Modelling and Software*, **23** (4), 404-411, 2008.

Hachinski V., Iadecola C., Petersen R.C., Breteler M.M., Nyenhuis D.L., Black S.E., Powers W.J., DeCarli C., Merino J.G., Kalaria R.N., Vinters H.V., Holtzman D.M., Rosenberg G.A., Dichgans M., Marler J.R., Leblanc G.G. National Institute of Neurological Disorders and Stroke-Canadian Stroke Network vascular cognitive impairment harmonization standards. *Stroke*, **37**: 2220-2241, 2006.

Hasson U., Skipper J.I., Wilde M.J., Nusbaum H.C., Small S.L. Improving the analysis, storage and sharing of neuroimaging data using relational databases and distributed computing. *NeuroImage*, **32**: 693-706, 2008a.

Hasson U., Andric M., Kenny S., Wilde M., Small S.L. An architecture for GRID-based analysis of Neuroimaging data using relational databases and the SWIFT workflow engine. *Frontiers in Neuroinformatics. Conference Abstract: Neuroinformatics 2008*, 2008b.

James G.A., Lu Z.L., VanMeter J.W., Sathian K., Hu X.P., Butler A.J. Changes in resting state effective connectivity in the motor network following rehabilitation of upper extremity poststroke paresis *Top Stroke Rehabil.*, **16**: 270-281, 2009.

Kenny S., Andric M., Boker S.M., Neale M.C., Wilde M., Small S.L. Parallel workflows for data-driven structural equation modeling in functional neuroimaging. *Front. Neuroinform.*, **3**: 34, 2009.

Konstantinos M., Nikos B., Nektarios G., Chrisa T., aStavros C. Towards a Mediator based on OWL and SPARQL. *Visioning and Engineering the Knowledge Society. A Web Science Perspective*, 326-335, 2009.

McIntosh AR. Overview: Integrating computational, cognitive and clinical expertise to understand brain network recovery. *Arch. Ital. Biol.*, 2010 (this issue).

Marcus D.S., Olsen T.R., Ramaratnam M., Buckner R.L. The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*, **5**: 11-33, 2007.

Mironov I. Hash functions: theory, attacks and applications. Technical Report MSR-TR-2005-187, Microsoft Research, Nov 2005.

OWL 2 Web Ontology Language Document Overview. *W3C Owl Working Group*, W3C Recommendation, 27 October 2009, http://www.w3.org/TR/owl2-overview/.

Panel on Research Ethics. Section 3: Privacy and Confidentiality. *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*, 1998, http://www.pre.ethics.gc.ca/eng/policy-politique/tcps-eptc/section3-chapitre3/.

Peng H. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24 (17): 1827-1836, 2008.

Price C.J., Seghier M.L., Leff A.P. Predicting language outcome and recovery after stroke: the PLORAS system. *Nat. Rev. Neurol.*, **6**: 202-210, 2010.

Price C.J., Crinion J.T., Leff A.P., Richardson F.M., Schofield T., Prejawa S., Ramsden S., Gazarian K., Lawrence M., Ambridge L., Andric M., Small S.L., Seghier M.L. Lesion sites that predict the ability to gesture how an object is used. *Arch. Ital. Biol.*, 2010 (this issue).

Resource Description Framework (RDF). Concepts and Abstract Syntax. In: Klyne G. and Carroll J.J. (Eds.), *W3C Recommendation*, 10 February 2004, http://www.w3.org/TR/rdf-concepts/.

Rousseau M., Rioux P., Sherif T., McCloskey A., Adalat R., Evans A. CBRAIN: Canadian Brain Imaging Network. Poster, *IEEE Grid Conference*, Oct 2009.

Scavo T. and Welch V. A grid authorization model for science gateways. *International Workshop on Grid Computing Environments*, 2007. See http://library.rit.edu/oajournals/index.php/gce/article/view/99

Seghier M.L., Ramlackhansingh A., Crinion J., Leff A.P., Price C.J. Lesion identification using unified segmentation-normalisation models and fuzzy clustering. *Neuroimage*, **41** (4): 1253-1266, 2008.

Skipper J.I., Goldin-Meadow S., Nusbaum H.C., Small S.L. Speech associated gestures, Broca's area, and the human mirror system. *Brain Lang.*, **101**: 260-277, 2007.

Small S.L., Wilde M., Kenny S., Andric M., Hasson U. Database-managed grid-enabled analysis of neuroimaging data: the CNARI framework. *Int. J. Psychophysiol.*, **73**: 62-72, 2009.

SPARQL Query Language for RDF. In: Prud'hommeaux E. and Seaborne A. (Eds.), *W3C Recommendation*, 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/.

Van Horn J.D. and Toga A.W. Multisite neuroimaging trials. *Curr. Opin. Neurol.*, **22**: 370-378, 2009a.

Van Horn J.D. and Toga A.W. Is it time to re-prioritize neuroimaging databases and digital repositories? *Neuroimage*, **47**: 1720-1734, 2009b.

Welch V., Barlow J., Basney J., Marcusiu D., Wilkins-Diehr N. A model to support science gateways with community accounts. *Concurrency and Computation: Practice and Experience*, **19** (6): 893-904, 2006. http://dx.doi.org/10.1002/cpe.1081

Yourganov G., Schmah T., Small S.L., Rasmussen P.M., Strother S.C. Functional connectivity metrics during stroke recovery. *Arch. Ital. Biol.*, 2010 (this issue).

Zhao Y., Hategan M., Clifford B., Foster I., von Laszewski G., Nefedova V., Raicu I., Stef-Praun T., Wilde M. Swift: fast, reliable, loosely coupled parallel computation. *IEEE Congress on Services*, 199-206, 2007.