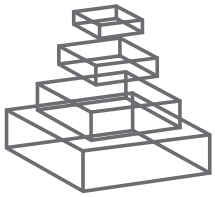
frontiers
RESEARCH TOPICS

NEUROIMAGING WORKFLOW DESIGN AND DATA-MINING

Hosted by
Arthur W. Toga and John Van Horn

frontiers in
NEUROINFORMATICS



frontiers

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714
ISBN 978-2-88919-022-5
DOI 10.3389/978-2-88919-022-5

ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.
All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

DEDICATION TO QUALITY

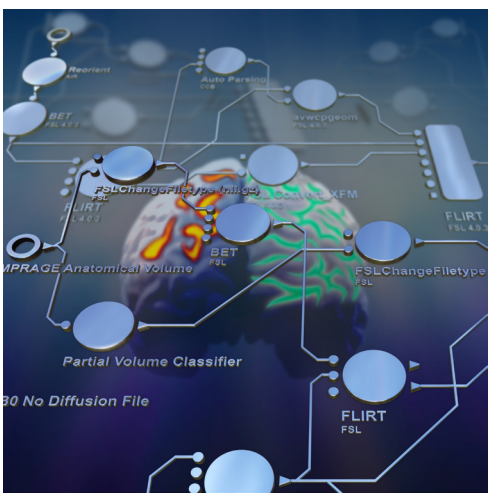
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!
Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

NEUROIMAGING WORKFLOW DESIGN AND DATA-MINING

Hosted By
Arthur W. Toga, UCLA School of Medicine, USA
John Van Horn, University of California at Los Angeles, USA



With the increasing number of neuroimaging studies appearing yearly in the literature, the need to consider the synthesis of the underlying data into new knowledge and research directions has never been more important. The development of large-scale databases and grid-enabled computing has laid the groundwork for mining these rich datasets beyond the scope of their initial collection. Additionally, meta-analyses of the summary results contained in published research articles have provided a powerful way to explore hidden trends in the neuroscience literature. In each case, the processing of data requires a careful consideration of the individual processing steps involved and how they can be assembled into reliable workflows. In results from published studies, the manner in which data were processed may influence meta-analytic results which can have implications on clinical interpretation. Several efforts now exist that provide tools for use in the construction of data processing workflows. However, careful thought must be given to ensuring appropriate, efficient, optimal, and replicable processing. The results obtained from data-mining and meta-analysis must tell a story about a collection of existing data. Also they must suggest novel and testable hypotheses for further investigation with implications for understanding of the brain in health and disease. Where they do, these new results and interpretations often provide fresh insights into the data that extend beyond the rationale for their original collection. In this volume, we have asked leaders in the field of neuroimaging data mining and meta-analysis to provide their thoughts on methods for efficient workflow design, interoperability with large-scale databases, and to discuss their work in exploring the richness of brain imaging data as well as the literature of published research results.

Image caption: "The development of data mining and workflow technologies maximizes opportunities for neuroimaging data analysis and re-use."
Image credit: The Laboratory of Neuro Imaging (LONI)

Table of Contents

frontiers in
NEUROINFORMATICS

EDITORIAL
published: 25 September 2009
doi: 10.3389/neuro.11.031.2009

Neuroimaging workflow design and data-mining: a Frontiers in Neuroinformatics special issue

John Darrell Van Horn* and Arthur W. Toga

Laboratory of Neuro Imaging, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
** Correspondence: jvanhorn@loni.ucla.edu*

The development of sophisticated neuroimaging data processing tools has been of major importance for distilling the large amount of information present in brain imaging data sets into useful and enlightening results. Neuroinformatics-based algorithms, in particular, have been instrumental in analyzing population level cortical anatomy, changes in BOLD activity, and, more recently, the rapid processing of diffusion weighted images (DTI/HARDI). Several notable examples include Statistical Parametric Mapping (Friston, 2006), FSL (Smith et al., 2004), FreeSurfer (surfer.nmr.mgh.harvard.edu), AFNI (Cox, 1996), and BrainVoyager (Goebel et al., 2006), among other analysis packages. The wide availability of neuroinformatics tools has helped to significantly spur growth in cognitive and clinical neuroscience, as well as permitted the efficient re-analysis of data contained in large-scale data archives (Kennedy and Haselgrove, 2006).

Within any of the aforementioned software packages it is possible to find the majority of individual steps needed for processing the most common types of brain imaging data. With these individual operations, accompanied by various inputs, parameters, and other options, investigators frequently link executable programs together as "scripts" or batch processes in which inputs are passed to one executable and the resulting outputs become the input to the next processing executable, and so on. In so doing, many laboratories have found it possible to create efficient yet flexible data processing streams to not only process data within modality but also between modalities. The notion of scientific workflows has now taken on its own formalism, moving from beyond custom-built scripts toward fully-fledged software environments with several available software platforms available to construct neuroimaging workflows, optimize their performance, and that take advantage of super-computing and grid infrastructures to expedite data processing throughput (Romano et al., 2005; Oinn et al., 2006; Van Horn et al., 2006; Ruping et al., 2007; Verdi et al., 2007). With a fully encompassing workflow platform it is also possible to break out of a "package-centric" view of neuroimage data processing and toward an informatics model that draws processing capabilities from across existing software suites as well as the incorporation of local informatics tools into heterogeneous analysis workflows. Such workflow descriptions, which themselves are often highly structured file formats describing the executable operations and their various processing choices, can serve to provide needed data provenance ensuring the fidelity of data reanalysis and replication (Mackenzie-Graham et al., 2008).

More than simply processing individual subject datasets or even the data from complete neuroimaging studies, the notion of workflows has permeated the next level of neuroimaging analysis beyond subject or study-based processing: that of data mining and meta-analysis. Data mining is a process of exploring data to identify potentially interesting patterns in the data that might not have been examined in the original research studies in question or perhaps were not detected by traditional statistical methods. These approaches to sifting through large archives of data to extract potentially useful patterns and relationships has been most evident in the genomic sciences although neuroimagers have explored these methods as well (Mitchell, 1999; Megalooikonomou et al., 2000; Wigle et al., 2001; Anderle et al., 2003). Meta-analysis, on the other hand, first gained the attention of the social sciences and related fields in the late 1970's and 1980's as a means to examine the study-specific and experimental factors that predicted reported effect sizes present in published studies (Glass et al., 1981; Rosenthal, 1984). The notion of performing an "analysis of analyses" to quantitatively critique, explore, and synthesize a literature has proved to be highly compelling and powerful. With the burgeoning growth of neuroimaging studies of brain structural differences between clinical populations and examinations of human cognition using PET and fMRI, the concept of meta-analysis soon found its way into the realm of brain imaging (Van Horn and McManus, 1992; Fox and Woldorff, 1994; Cabeza and Nyberg, 2000). Data mining and meta-analyses permit the exploration of not only the neural structure or patterns of cognitively-induced activity, but these analyses can also provide insights into those study factors that can predict the magnitude of reported effects. These approaches help to synthesize data from across studies, craft general trends in results from across studies, and quantify the effects of predictor variables obtained from the studies themselves that may influence the size and scale of differences.

Data processing workflow concepts have been an important element for meta-analyses and data mining, too, providing the basis for how sufficient summary metrics are obtained, combined with appropriate study meta-data, and then systematically compared and combined from across subjects and studies (**Figure 1**). Visualizing the relationships between subjects and study results has also been an important element for meta-analysis results and workflows are needed to provide graphical representations to still further neuroinformatics tools needed for dynamic and interactive visualization (Toga and Thompson, 2002; Van Essen, 2002; Van Essen and Dierker, 2007).

In this special issue of *Frontiers in Neuroinformatics*, we have invited several leading groups to provide articles focusing on the development of workflow technologies and perspectives for efficient neuroimage data processing and that help to permit subsequent meta-analysis of the results. Articles by Dinov et al., Ooi et al., Kenny et al., and Cheng et al. showcase recent developments in advanced workflow technologies for efficient processing of neuroimaging data. Contributions from Keator and colleagues discuss the use of high-performance computing capabilities upon which workflow environments have been specifically designed to take advantage of for rapid processing, while articles from Costafreda et al., Bockholt et al., Lohrey et al., and Laird et al. discuss the

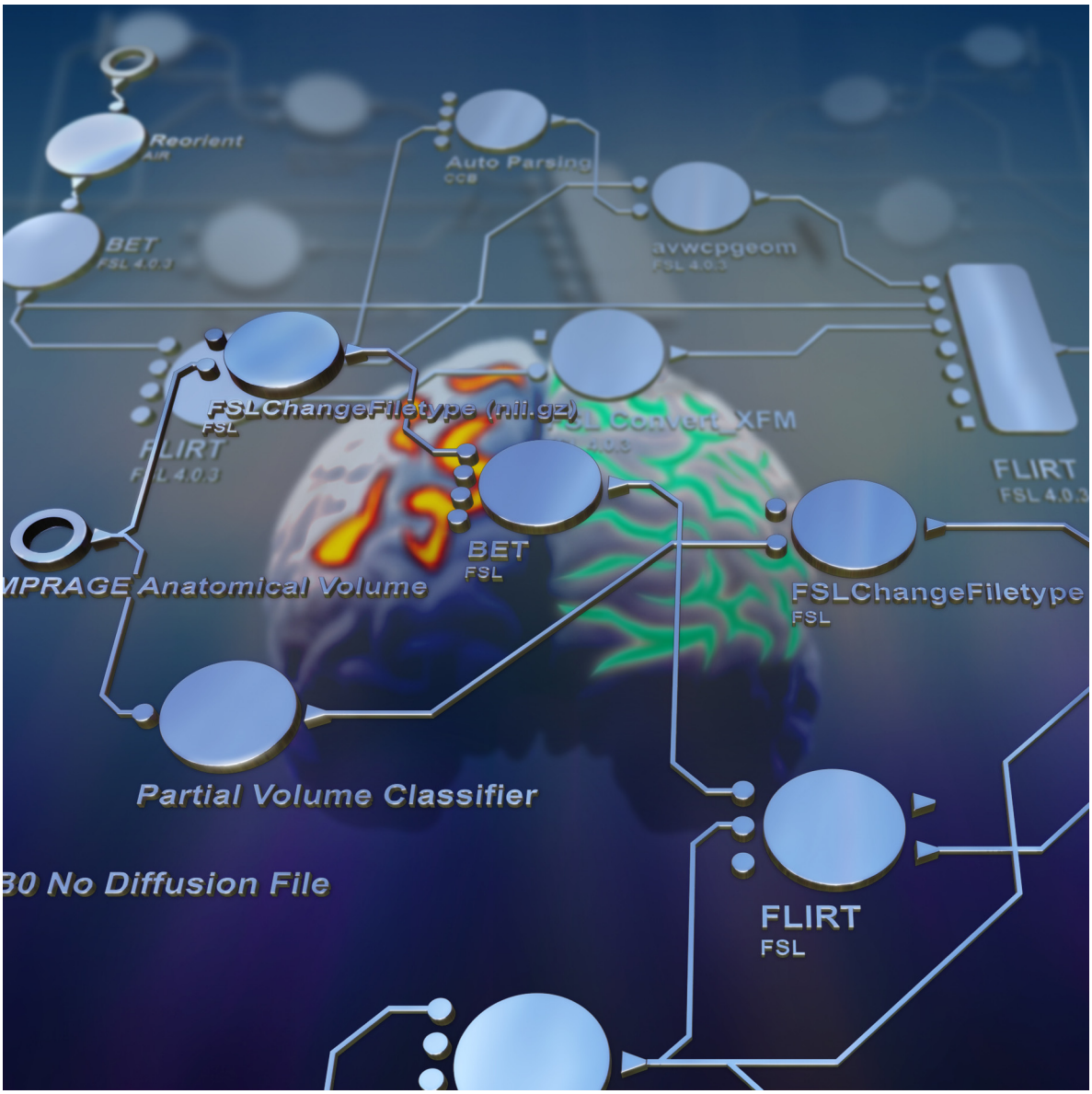


FIGURE 1 | Scientific workflows provide flexible platforms for multimodal neuroimage processing that facilitate high-throughput analysis of individual subjects as well as complete studies. These are also essential software and informatics frameworks for data mining and exploration, meta-analytic consideration of effects from across multiple studies, as well as providing efficient approaches for visualizing synthesized results and the functional/structural relationships that exist between brain imaging data sets.

development of data mining workflows and feature interesting examples of data synthesis. Finally, contributions from Nielsen and from Joshi et al. discuss the important role of visualization in data mining and the workflows necessary to inform novel informatics tools that focus on interactively exploring the relationships amongst large collections of brain data. The quality of these articles is exceptional and provides a broad overview at how workflow concepts have matured for the neuroimaging field, how they are now being used to expedite data mining, meta-analysis, and helping to provide the content needed for graphical data interaction.

Informatics as had a historical foothold in the data-rich field of neuroimaging. However, with *in vivo* datasets continually increasing in size, scope, and complexity, the continued development of efficient processing tools remains necessary to extract the maximal amount of useful information from them. Workflow technologies for data processing design, application, and execution link these tools into high-throughput processing pipelines. Their ongoing development can be expected to greatly enrich the ability of researchers to not only process newly obtained neuroimaging data but also to compare, contrast, and combine results from previous research studies via meta-analytic and data mining approaches and to visualize unique patterns present in neuroimaging results that could only be identified through large-scale informatics approaches.

REFERENCES

Anderle, P., Duval, M., Draghici, S., Kuklin, A., Littlejohn, T. G., Medrano, J. F., Vilanova, D., and Roberts, M. A. (2003). Gene expression databases and data mining. *Biotechniques* Suppl, 36–44.

Cabeza, R., and Nyberg, L. (2000). Imaging cognition II: an empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* 12, 1–47.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.

Fox, P. T., and Woldorff, M. G. (1994). Integrating human brain maps. *Curr. Opin. Neurobiol.* 4, 151–156.

Friston, K. J. (2006). Statistical Parametric Mapping. London, Academic Press.

Glass, G. V., McGaw, B., and Smith, M. L. (1981). Meta-Analysis in Social Research. Beverly Hills, Sage Publications.

Goebel, R., Esposito, F., and Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Hum. Brain Mapp.* 27, 392–401.

Kennedy, D. N., and Haselgrove, C. (2006). The internet analysis tools registry: a public resource for image analysis. *Neuroinformatics* 4, 263–270.

Mackenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage* 42, 178–195.

Megalooikonomou, V., Ford, J., Shen, L., Makedon, F., and Saykin, A. (2000). Data mining in brain imaging. *Stat. Methods Med. Res.* 9, 359–394.

Mitchell, T. (1999). Machine learning and data mining. *Commun. ACM* 42, 30–36.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency Comput. Pract. Exp.* 18, 1067–1100.

Romano, P., Marra, D., and Milanesi, L. (2005). Web services and workflow management for biological resources. *BMC Bioinformatics* 6(Suppl. 4), S24.

Rosenthal, R. (1984). Meta-Analytic Procedures for Social Research. Beverly Hills, Sage.

Ruping, S., Sfakianakis, S., and Tsiknakis, M. (2007). Extending workflow management for knowledge discovery in clinico-genomic data. *Stud. Health Technol. Inform.* 126, 184–193.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23(Suppl. 1), S208–S219.

Toga, A. W., and Thompson, P. M. (2002). New approaches in brain morphometry. *Am. J. Geriatr. Psychiatry* 10, 13–23.

Van Essen, D. C. (2002). Surface-based atlases of cerebellar cortex in the human, macaque, and mouse. *Ann. N. Y. Acad. Sci.* 978, 468–479.

Van Essen, D. C., and Dierker, D. L. (2007). Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* 56, 209–225.

Van Horn, J. D., Dobson, J., Woodward, J., Wilde, M., Zhao, Y., Voeckler, J., and Foster, I. (2006). Grid-based computing and the future of neuroscience computation. In Methods in Mind, C. Senior, T. Russell, and M. S. Gazzaniga, eds (Cambridge, MIT Press), pp. 141–170.

Van Horn, J. D., and McManus, I. C. (1992). Ventricular enlargement in schizophrenia. A meta-analysis of studies of the ventricle:brain ratio (VBR). *Br. J. Psychiatry* 160, 687–697.

Verdi, K. K., Ellis, H. J., and Gryk, M. R. (2007). Conceptual-level workflow modeling of scientific experiments using NMR as a case study. *BMC Bioinformatics* 8, 31.

Wigle, D. A., Rossant, J., and Jurisica, I. (2001). Mining mouse microarray data. *Genome Biol.* 2, REVIEWS1019.

Received: 17 August 2009; published online: 25 September 2009

Citation: Front. Neuroinform. (2009) ***3****:31. doi: 10.3389/neuro.11.031.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 08 July 2009
doi: 10.3389/neuro.11.019.2009

An integrated object model and method framework for subject-centric e-Research applications

***Jason M. Lohrey*[1,2]*, Neil E.B. Killeen*[3]* and Gary F. Egan*[2,3]**

[1] *Arcitecta Pty Ltd, Victoria, Australia*
[2] *Florey Neuroscience Institutes, University of Melbourne, Victoria, Australia*
[3] *Centre for Neuroscience, University of Melbourne, Victoria, Australia*

Edited by:
John Van Horn, University of California, USA

Reviewed by:
Neil R. Smalheiser, University of Illinois at Chicago, USA
Jeffrey S. Grethe, University of California, USA
John Van Horn, University of California, USA

****Correspondence:***
Neil Killeen, Centre for Neuroscience, University of Melbourne, Victoria 3010, Australia.
e-mail: nkilleen@unimelb.edu.au

A framework that integrates an object model, research methods (workflows), the capture of experimental data sets and the provenance of those data sets for subject-centric research is presented. The design of the Framework object model draws on and extends pre-existing object models in the public domain. In particular the Framework tracks the state and life cycle of a subject during an experimental method, provides for reusable subjects, primary, derived and recursive data sets of arbitrary content types, and defines a user-friendly and practical scheme for citably identifying information in a distributed environment. The Framework is currently used to manage neuroscience Magnetic Resonance and microscopy imaging data sets in both clinical and basic neuroscience research environments. The Framework facilitates multi-disciplinary and collaborative subject-based research, and extends earlier object models used in the research imaging domain. Whilst the Framework has been explicitly validated for neuroimaging research applications, it has broader application to other fields of subject-centric research.

Keywords: object model, experimental methods, data repository, subject-centric, e-Research, collaborative research

INTRODUCTION

Research groups worldwide are facing data management challenges[1]. Not only is the volume of data rising dramatically, but also the processes that a researcher follows to analyze and manage research data are increasingly complex. Of crucial importance for a data management system is the way in which information is organized. A common method of data organization is use of an object model that is motivated by the processes and protocols of the specific research domain. These include how the data are acquired, what the relationships between data are, and how the data will be distributed, analyzed and interpreted.

Neuroimaging is a rapidly developing research domain in which enormous quantities of data are acquired. Identification of an appropriate object model for neuroimaging involves firstly identifying the particular class of research that it belongs to. Neuroimaging is an example of "subject-centric" research, which refers to well-defined, persistent subject matter for which data are being acquired over some (perhaps extended) period of time. For example, a subject might be an animal (human, mouse etc.), chemical or mineral sample with a number of data acquisitions undertaken for each subject over time.

A second important aspect of data management is to recognize that research data are often obtained through a well-defined, and sometimes complex workflow. Although organization of information with an object model is an established methodology, the method (or workflow) is less commonly captured along with the data.

An object model captures domain-specific data and metadata. It is a very significant challenge to develop metadata (and data) standards within domains, let alone across domains. Our approach is to develop a framework that represents the essential components of subject-centric research without prescribing particular metadata. The Framework then requires the domain specialists to define the appropriate metadata for their research.

Research is increasingly distributed through collaborations involving researchers at different institutions. The location of objects associated with data and metadata should be largely transparent to the researcher and accessible from anywhere through a number of mechanisms including distributed queries, remote access and replication.

In this paper, we describe a framework that defines an object model to explicitly represent research methods and the resulting acquired and derivative data for subject-centric research. The Framework captures the core relationships required for auditable and reproducible research. The Framework is extended with metadata that is specific to the type of subject and domain of research and it explicitly provides an identification scheme that supports distributed objects. The Framework has been implemented and is used to manage distributed neuroimaging data.

MATERIALS AND METHODS

BACKGROUND

Neuroscience research increasingly involves scientific collaborations across sub-domains that acquire, share and analyze multi-modal data (e.g. Gardner et al., 2003; Martone et al., 2004; Toga, 2002). For example, neuroscience research may include types of data such as: magnetic resonance imaging (MR) and spectroscopy (MRS), optical and electron microscopy (OM and EM), positron-electron tomography (PET), computed tomography (CT), electrophysiological, genotype, electroencephalogram (EEG) and event related

[1]http://www.nsf.gov.pubs/nsf0728/nsf0728.pdf

potential (ERP) data types. This list will undoubtedly continue to lengthen, particularly as new forms of collaborative research emerge over time. Various neuroimaging and related groups worldwide have developed applications to provide data management and application capabilities (examples include Keator et al., 2008; Marcus et al., 2007; Marenco et al., 2003; SenseLab[2], LONI Image Data Archive[3] and fMRIDC[4]).

The need in our own research environment to manage many different types of data using a consistent model was the catalyst to seek a generic object model that supports: (i) project-based virtual organizations, (ii) representation of the subject of a study, (iii) recording the state changes in a subject, (iv) representation of the experimental method (process or workflow), (v) participation by subjects in multiple research projects, (vi) disassembling of subjects into constituent parts, (vii) controlled access to all information and especially the identity of a subject, (viii) capture and storage of all types of data, and (ix) the capability to manage raw and processed data.

The requirement to record state arises because the subject may undergo a number of procedures in an experimental process. These state changes might may be transient (e.g. anesthesia) or permanent (e.g. death) and affect the subsequent acquisition of data. A given subject may be disassembled (e.g. removal of the brain) into constituent parts for subsequent study. There may be parallel studies on different "parts", each with a separate procedure and life cycle.

Rather than create yet another object model, we investigated whether an existing model would satisfy our main requirements. Consideration was given to: (i) the Digital Imaging and Communications in Medicine (DICOM[5]) model, (ii) the XML-based Clinical and Experimental Data Exchange (XCEDE[6] and see also Keator et al., 2006) model, (iii) a Project-Subject-Study (PSS) model (our own earlier generation object model) and (iv) the Council for the Central Laboratory of the Research Councils (CCLRC[7]). As will be demonstrated, none of these object models fully met the requirements, but all provided valuable components that have been used and extended.

DICOM Object Model

The DICOM standard includes formatting, communications and object modeling components. DICOM is ubiquitous in medical imaging and was originally created for clinically oriented studies conducted with patients although it can be utilized for other studies. The DICOM object model is complex – the key objects that are relevant to neuroimaging research are shown in a Unified Modeling Language (UML) object diagram[8] (**Figure 1A**). Briefly, the *Patient* represents the subject of the investigation, and may undertake a number of *Visits* over time to imaging facilities. Each *Visit* results in a number of *Studies* that represent a particular imaging setup and procedure. Each *Study* generates a number of actual acquisitions of a particular type (e.g. MR image volumes) that are called *Series*.

The DICOM object model is limited in that it lacks the concept of a project consisting of many subjects, is unable to record the experimental method, nor represent the state of a subject. In addition, the DICOM standard requires the data sets to be encapsulated in the DICOM file format.

Biomedical Informatics Research Network (BIRN) XCEDE Schema

The XCEDE metadata schema (and implicit object model) is intended for the exchange of clinical and research imaging studies. The objects in the XCEDE model are *Project, Subject, Visit, Study* and *Series* (**Figure 1B**), with the XCEDE objects equivalent to the DICOM objects from the *Subject* level. The XCEDE object model, and the associated metadata hierarchy described in the XCEDE XML schema are highly specific to image-based analysis and cannot be easily applied more generally. However, the model contains a number of interesting and useful concepts related to experimental method. For example, the provenance of any object may be used to describe the data processing protocol that was used to generate a sub-set of data. The inclusion of provenance information at any level in the object model hierarchy is an advantage of the XCEDE schema.

PSS Object Model

The project subject study (PSS) object model (**Figure 1C**) was derived from the DICOM object model with two key extensions. Firstly, the *Project* object at the top of the hierarchy (like XCEDE) corresponds to the virtual project team collaborating on a specific scientific experiment. Secondly, the *Subject* object may be decomposed into two parts: the project-specific attributes of the subject, and the project-invariant aspects that are common to all projects. The ability to re-use *Subjects* in multiple *Projects* required a relationship to be specified between the *Project* and *Study* objects. The PSS model also removed the DICOM *Visit* object.

The PSS model was used in research using MR imaging data and although not mandated by the PSS model, only DICOM format data were included. While the PSS model had a number of improvements over the DICOM model, additional key requirements including the ability to capture the experimental method and track subject state were not met.

CCLRC Object Model

The CCLRC model was defined as a generic model for handling e-Science data (**Figure 1D**). This model was examined to establish if it satsified the requirements for representing subject-centric, neuroimaging research studies. The CCLRC *Study* is sometimes referred to as a *Project* and each *Investigation* is directly linked with one *Data Holding* that contains the data generated by the investigation. A *Data Holding* is a hierarchy of *Data Collections* and/or atomic *Data Objects*. The CCLRC model has no concept of the "subject" of an investigation (and associated state), nor the method of research and thus does not meet the requirements for

[2]SenseLab: http://senselab.med.yale.edu/

[3]http://ida.loni.ucla.edu/

[4]The fMRI Data Center's data management tools http://www.fmridc.org/f/fmridc/database/index.html

[5]Digital Imaging and Communications in Medicine (DICOM). http://medical.nema.org

[6]XML-based Clinical and Experimental Data Exchange (XCEDE). http://www.nbirn.net/tools/xcede/index.shtm

[7]Sufi S, Mathews B. Council for the Central Laboratory of Research Councils (CCLRC) Scientific Metadata Model: Version 2. See http://epubs.cclrc.ac.uk

[8]Unified Modeling Language. http://www.uml.org

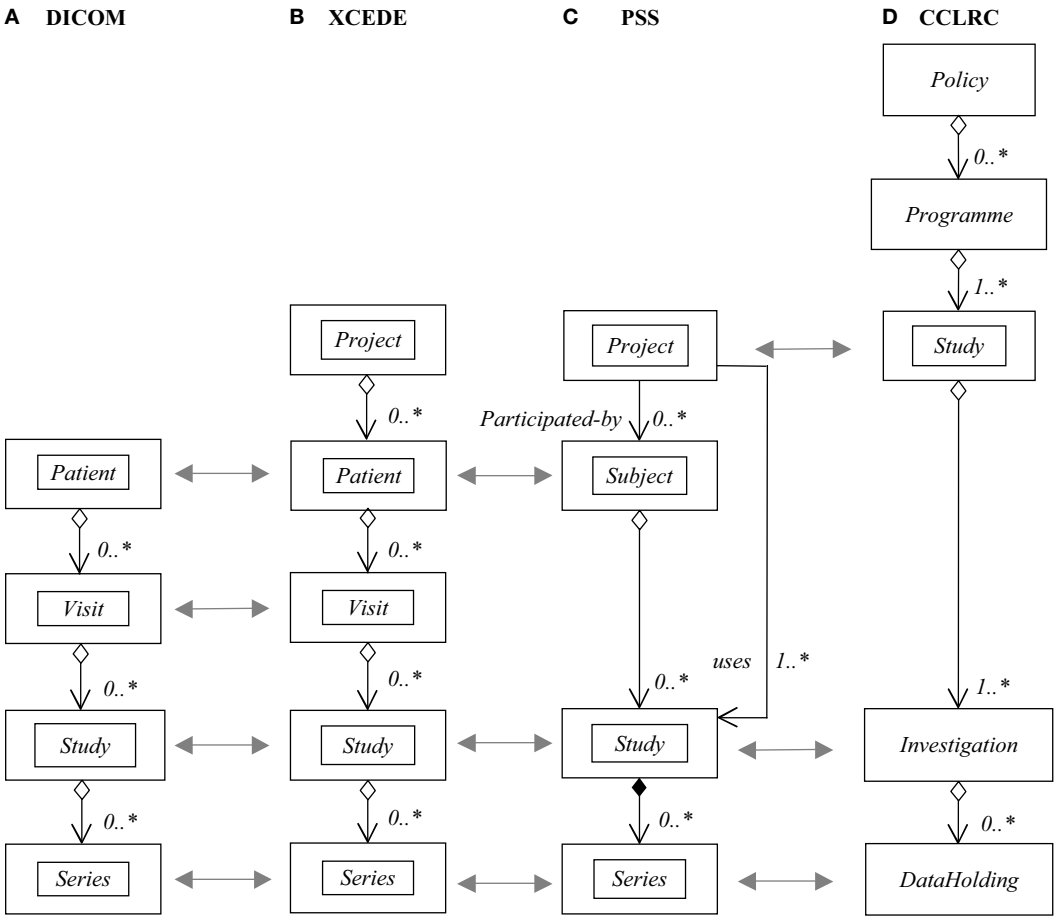


FIGURE 1 | UML object diagrams for the (A) DICOM, (B) XCEDE, (C) PSS and (D) CCLRC object models. In addition to the UML notation the horizontal arrows indicate equivalence of objects between the different models. The UML notation can be summarized as follows: Objects (or classes) are shown in rectangles and named relationships are shown between objects that are qualified by their cardinality (* means infinity, 0..* means 0 to infinity). The relationship direction is indicated via an arrow. Filled diamonds indicate that the relationship is containment (also called composition) and open diamonds indicate an aggregation (has) relationship.

neuroimaging research without extension. However, novel parts of the model, such as the hierarchy of *Data Holding* objects, provide useful elements for inclusion into subject-centric data models.

FRAMEWORK DESCRIPTION

Overview

A subject-centric research object model that includes details of research experimental methods has been developed. The model can be applied to studies involving subjects such as people, animals, plants or minerals. The model does *not prescribe any particular domain-specific metadata*, but instead the domain of research defines specific metadata and semantic interpretation through associated ontologies. The model is independent of a particular implementation technology.

The Framework has a number of characteristics including: (i) objects may have location independent *Citable Identifiers* that allow objects to be referenced in a distributed environment; (ii) objects are primarily organized into a hierarchy of *Project, Subject, ExMethod, Study* and *DataSet* (see below); (iii) the *R-Subject* object allows subjects to be used in multiple projects; (iv) the research *Method* (i.e. the set of steps in a workflow where each step may have meta-data and/or produce data) can be encoded; (v) all state changes for a subject are recorded; any data set produced is a function of the state of the subject at that point in time; and (vi) *DataSet*s may be further organized into a hierarchy of *DataSet(*s*)* and *DataObject(*s*)*.

Citable Identification

The ability to cite research data and data sets is an important part of research publication, allowing peer access, review and reuse of raw and derived data. Citation requires the assignment of unique and long lived identifiers (see Brase, 2004; Klump et al., 2006, 2008) to each citable entity.

In this model, objects are identified using a hierarchical identification scheme that supports unique identity in a distributed environment. The citable identifier scheme is a human-friendly, arbitrary depth hierarchy of positive integer numbers (`NA.ORG.r.`$n_1.n_2 \ldots n_k$). Citable identifiers are used for all objects (see below for an example) within the object model that may be externally cited to allow collections to be distributed across many repositories.

Once assigned, an identifier is immutable although replicas of the same object may exist in multiple locations. These identifiers are compatible with other identification schemes, such as DOI[9] and HANDLE[10] (see also PILIN[11]).

These identifiers should be interpreted as follows: (i) an identifier has depth N (the number of dot characters (".") plus one), (ii) the identifier part at depth 1 is the Naming Authority, (iii) the identifier part at depth 2 is the Organization that can resolve the location of a resource, (iv) the pair (NA.ORG) is unique and (v) the naming authority must be able to reference the organization. The third digit, which follows the NA.ORG part of the identifier provides root namespace separation (e.g. to separate collections of *Projects, R-Subjects* and *Methods)*.

[9]Digital Object Identifier (DOI). http://www.doi.org

[10]Unique persistent identifiers. http://www.handle.net

[11]http://www.pilin.net.au

Objects with the same parent are considered to be in the same collection. These collection semantics allow the members of a collection to be easily located (including any replicas) in a distributed system without requiring more complex centralized registries or cross-repository references.

Object Hierarchy

The object hierarchy (**Figure 2**) and the objects (**Table 1**) can be used in two ways. Firstly, a subject may exist only in a single project (**Figure 2A**). Secondly, a subject may exist in multiple projects (e.g. people, a calibration reference) in which case it may be represented by the (real) *R-Subject* (**Figure 2B**).

A *Subject* is *Project* based and so has attributes of particular interest to that *Project*. The subject matter of an investigation may be disassembled into sub-parts. That is, parts may be removed (e.g. the brain removed from the skull of a mouse) and become independent entities for investigation. When a subject participates in more

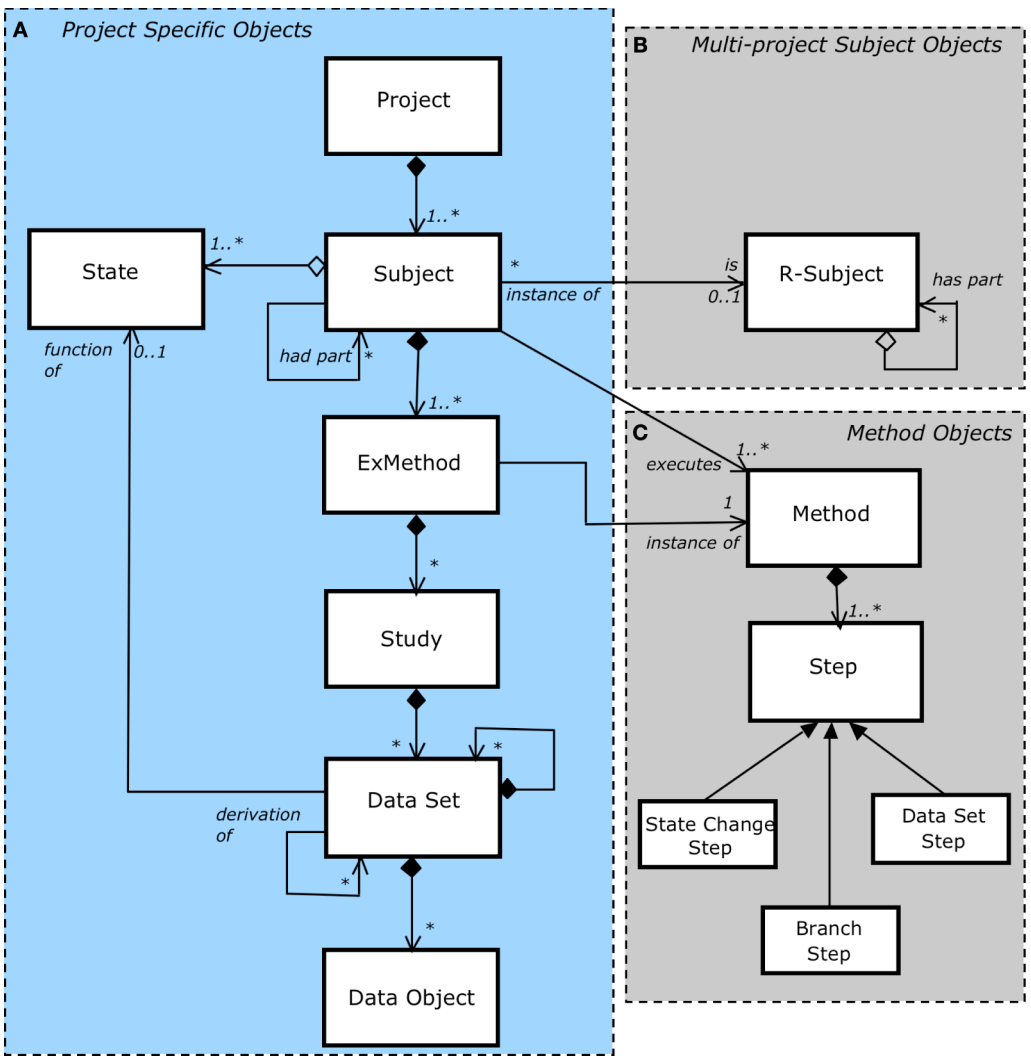


FIGURE 2 | UML object diagram of the Framework object model (see Table 1 for definitions). (A) When subjects are not re-used across multiple projects, only the project specific objects are used. **(B)** If subjects participate in multiple projects then additional objects are required. **(C)** The *Method* object contains *Steps*, each of which is comprised of a possible *State Change*, production of a *DataSet* and a *Branch Point*.

Table 1 | The object definitions for the Framework object model.

Object	Definition
Project	Established by a team to undertake a specific investigation.
Subject	The subject matter (e.g. animal, plant etc.) of a particular *Project*. There are typically many *Subjects* per *Project*.
ExMethod	Container for the execution of a specific *Method;* holds reference to *Method* and the state of execution (e.g. executed *Step*) of the *Method*.
Study	A container for a class of measurements. For example, a neuroscience study might be of type MR, Microscopy, PET or EEG.
DataSet	A set of acquired or processed data that may take any form (e.g. an MR volume)
State	The state (changes may be transient or permanent) of the subject at a point in time.
Method	The specification of a research process. Methods are applied to *Subject* objects.
Step	A single step in a *Method*. A *Method* may have one or more *Step*s to be performed. *Method*s may allow *Step*s to be performed sequentially or in any order
State Change	A specialized *Step* in a *Method* that results in recording a state change for the *Subject*. The state change will be recorded using the metadata specified for the step,
Data Set Step	A specialized *Step* in a *Method* that produces one or more *Data Set*s. The *Data Set Step* details the metadata to be generated for the acquired or derived *Data Set*s.
Branch Step	A conditional branch that refers to one or more other *Method*s. The branch may require one or all of the specified sub-*Method*s be performed.
R-Subject	An *R-Subject (R* for "re-usable" or "real") is used when the subject matter participates in multiple *Projects* (e.g. a person).

than one project, both *R-Subject* and *Subject* objects will represent it. The *R-Subject* captures time invariant characteristics, and, like the *Subject*, which is the subject's manifestation in a project, an *R-Subject* may be an assembly of discrete parts.

Where ethics requirements allow, an *R-Subject* can be used to identify all of the *Projects* in which a subject has participated. The discovery or measurement of new time-invariant characteristics, or recognition of existing and potentially significant characteristics, may be retrospectively important and inform any of the projects in which the subject has participated.

A subject will have one or more identities. Access to the identity (and other attributes) may be restricted by the implementation.

Subjects need not have a direct physical manifestation. They may represent derived entities, such as a probabilistic calculation from multiple input subjects (e.g. an atlas) or a computed model based on data sets from other subjects.

The *ExMethod* object represents the execution of a specific *Method (*which codifies a workflow and is discussed further below*)*. The *ExMethod* object contains a reference to the specific *Method* that is being executed and specifies the state (e.g. "incomplete", "complete") of each step of the *Method* being executed. *Subjects* may have multiple *Methods* executed on them, and therefore may have multiple *ExMethod* objects. *DataSets* may be original (measured) or computed (processed). A computed data set may be derived from one or more other data sets.

The object model indicates containment by the filled diamonds. Therefore deleting a parent object will also delete all children objects. For example, deleting a *Project* will delete all contained *Subject, ExMethod, Study, DataSet* and *DataObject* objects. However, deleting a *Subject* does not delete any disassembled *Subjects* that were previously part of that *Subject*, since they are autonomous objects, nor would it delete any associated *R-Subjects*.

The following objects typically have citable identification: *Project, Subject, ExMethod, Study, DataSet*, *R-Subject* and *Method*. Although a *DataSet* is a member of the *Subject* collection based on the semantics of the assigned citable identifiers, there is an explicit relationship to the *Subject* to identify the state of the subject at the time of acquisition. Note that the identifier scheme can be used to allow for different identifier roots. For example, using `r = 1` (see above) for collections of *Project*s and `r = 2` for collections of *R-Subject*s results in `NA.ORG.1.10.23.2.12` referring to *Study* 12 of *ExMethod* 2 of *Subject* 23 of *Proje*ct 10, whereas `NA.ORG.2.17` refers to *R-Subject* 17.

Methods

A *Method* is comprised of a number of Steps (**Figure 2C**), with each step uniquely identified within the scope of the *Method*. A *Method* can utilize a specialized step to prescribe the metadata required to create a *Subject* (and optionally *R-Subject*) as well as the metadata for each workflow step. A *Method* object should not be confused with an *ExMethod*. A *Method* is simply the specification of a process. When a *Method* is actually executed, then an *ExMethod* object is instantiated for the *Subject* executing the *Method*. This object holds the citable identifier of the *Method*, the number of the current step, as well as containing the *Studies* generated as a result of executing certain steps.

A step may affect a change of state in the subject, or result in the generation of a *Study*, or branch to another step or method. Branching may be qualified as "any" or "all" if there are multiple options. A step may pre-define metadata or define metadata that must be entered by the researcher. An example of a multi-step *Method* that acquires MR and Microscopy*Studies* is show in the Section "Results". Note that the *Method* and definition of metadata can be used to dynamically drive user interfaces.

A *Project* may have one or more prescribed *Methods* (selectable by the researcher) which are applied to a *Subject* and which may result in the generation of *Studies*. All subjects may require the same *Method*, or there may be different *Methods* for different subjects. For example, there could be *N* control subjects, and *M* non-control subjects each with different research *Methods*. In addition, **Figure 2**

shows that *Subject*s may contain one or more *ExMethod*s providing research flexibility. For example, subsequent *Method*s may refine an experimental process, or allow simple ad-hoc capture of data without prescriptive specification of process or metadata.

Methods are identified using citable identifiers so they may be referenced and re-used within a distributed environment. For example, an organization may have "standard" *Methods* that can be used directly or incorporated into more complex methods.

Life-Cycle and State

A subject's state may be altered (transiently or permanently; e.g. application of chemicals, death, etc.) prior to the acquisition of data. An acquisition of data at a point in time reflects the state of the subject at that point in time. The conditions that cause a state change are fully recorded in metadata associated with the *Subject*. A state change is uniquely identified within the context of a *Subject* and the pair (*Subject, State*) is unique. Permanent changes should be recorded with the *R-Subject*, if there is one, or the *Subject* otherwise.

DataSets and DataObjects

A *DataSet* contains the acquired or derived data and may hold data directly or be comprised of one or more *DataSets* and/or *DataObjects* (the smallest addressable item in our object model). We have made use of concepts in the CCLRC's *DataHolding* object model in this design. The definition of "small" is a matter of agreement, since, for example, the smallest unit of data might be a pixel within an image rather than an image.

DataSets may hold content directly, or they may be comprised of a number of smaller *DataSets* as well as zero or more *DataObjects* (**Figure 2A**). For example, many measurements involve the acquisition of calibration data followed by a series of measurements. The calibration data constitute a *DataSet* in their own right, but they are also directly associated with the subsequent measurement *DataSets*. As well as storing primary data, as in the above example, the object model provides for derived *DataSets* that are the transformation of one or more other *DataSets*. The method of transformation (e.g. a series or analysis applications) must be recorded in metadata attached to the *DataSet*. The *DataSet* object may store the transformed data, or may simply maintain the method for the generation of the data, which may be computed dynamically. The ability to precisely record the method for generating a *DataSet* then allows the method of construction to be peer reviewed, and the data can be discarded (e.g. to release storage resources) and re-created on demand.

DataSet identifiers are of two types either with all or none of the members having citable identifiers. A *DataSet* that contains members with citable identifiers (and can return the list of members upon request) is unordered and mutable. A *DataSet* that contains members that have no citable identification can identify the number of members and return the metadata and/or data for any member based on the ordinal position of that member.

A *DataSet* that is accessed by ordinal position must guarantee that the ordinal position of every member is immutable; members may only be appended. For example, a *DataSet* that contains other *DataSets* is unordered. Therefore, the members therein must also have citable identification. A DICOM *Series* is an example of an ordered *DataSet* with no requirement to cite individual members since it contains one or more images, each addressable by an ordinal (slice) position.

Metadata

The object model prescribes a minimum set of metadata elements for each object (**Table 2**).

These are then extended with domain-specific metadata to fully describe the objects and the research being undertaken. For the purpose of hierarchical presentation, identifying metadata must be attached to each *Project, Subject, R-Subject, ExMethod, Study* and *DataSet* object. This will allow type independent presentation of each collection. The "type" is important for semantic interpretation and the "name" provides identifying information for users.

If the *DataSet* is derived from one or more other *DataSets*, then the provenance of the *DataSet* must be identified. In addition, the nature of the derivation should be defined, ideally using structured metadata (when that metadata can easily be captured). A precise description is required if the *DataSet* is to be computed/recomputed at any time. The definition of other provenance metadata is domain specific.

Augmenting the generic prescribed metadata, domain-specific metadata is placed on the objects according to the concept that they represent and the temporal scope of the object (**Table 3** and see Results). For example, a *Project* object may hold metadata

Table 2 | The required minimum metadata for specific objects in the Framework object model. Elements are mandatory unless otherwise specified.

Object	Element	Description
All	*type*	One of [project, subject, r-subject, ex-method, study, dataset].
	name	The name of the collection.
	description	Arbitrary description (optional).
ExMethod	*method*	The citable identifier of the method being executed.
	context	The current execution context (method, sub-method, step).
Study	*type*	An extensible set of study types. In a neuroimaging implementation, the set might include values such as [mr,pet,om,em,eeg].
DataSet (primary)	*subject*	The citable identifier of the *Subject*.
	state	The state identifier of the *Subject*.
DataSet (derived)	*input*	A citable identifier for an input *DataSet*. There may be zero or more input elements. Not set if the *DataSet* is primary acquisition data.

Table 3 | Placement of domain-specific metadata on Framework objects.

Object	Metadata
Project	Details of the objectives, standard methods, investigators, organizations, etc.
Subject	Attributes of the *Subject* that are relevant to the project and which will be constant during the lifetime of the project.
State	Metadata describing the state of each *Method/step*.
Study	Metadata that is common to all contained *DataSet*s. Could also describe relevant information about the subject at the time of acquisition, rather than placing as time-dependent metadata on the *Subject*.
DataSet	Metadata specific to the acquisition or computation itself. For example, this might include method/protocol, the ambient air temperature etc.
R-Subject	Time invariant attributes of the subject. For example, in the case of an animal, the date of birth or date of death will not change.

describing the project team accessing it as well as hold identifiers for Ethics documents. A *Subject* may hold demographical and identity information, medical and educational history (for humans), genetic breeding details (for animals) and so on. These choices are entirely driven by the needs of the research.

Where metadata standards are available for a domain, it is advantageous to follow those standards, or at least provide a means to transform metadata to those standards.

The *Method* may be used to define much of this metadata (it may utilize a specialized step to prescribe metadata needed to create a subject as well as that for workflow) but other agents (e.g. a DICOM server) may also add metadata to objects (e.g. *Study* and *DataSet*).

Controlled Access

Data in a repository must have controlled access. Explicit control over access to metadata and content is best provided by role-based authorization and we have defined four project-specific, hierarchical roles where each role inherits the rights of the subordinate roles. The roles are `ProjectAdministrator` ("super-user" project permissions), `SubjectAdministrator` (administer subjects within the project), `Member` (read access to all research data and metadata generated by the project except protected identity information and `Guest` (can search the metadata only to find out what types of information are available). When an *R-Subject* is created, the `Administrator` roles have the ability to view the identity and update the details of the *R-Subject*. Alternatively, if an *R-Subject* is not utilized, the visibility of any sensitive identity information located on the *Subject* could be controlled via this role.

These roles are further qualified by the citable identifier of the project to provide project-specific access control. For example, for the project with citable identifier `1.1.1.2`, the `ProjectAdministrator` role would be named `ProjectAdministrator_1.1.1.2`.

RESULTS

The Framework has been extensively tested through a functioning reference implementation applied to the neuroimaging research domain to manage research data.

REFERENCE IMPLEMENTATION

A data repository has been built with a service-oriented Digital Asset Management system (Mediaflux™,[12]). A package of Mediaflux™ services implementing the Framework object model has been created. These services provide the basic interface to the data repository and allow a user to create, access and manage the objects of the model. As well as enabling the creation of the generic objects and metadata, the services also provide for the addition of domain-specific metadata and content, and the creation and use of *Methods* to manage experimental process and state.

The implementation uses the citable identifiers described above as arguments to many services to identify specific objects. The implementation does not explicitly create a *State* object. Instead, the state is contained within the *Subject* object. The implementation uses a well-defined XML metadata structure for each object. For example, on *Subject* and *R-Subject* objects, the implementation allows *public* and *private* metadata. The visibility of the metadata contained within these elements then depends upon the user's role (e.g. `ProjectAdministrator` [can see *private*] or `Member` [cannot see *private*]) and their semantic interpretation.

Sophisticated adaptive (to the metadata) graphical ("Web 2.0" and Java) interfaces that are driven by the object model (and especially the *Method*) have also been created (see below). These interfaces (which in turn use the above Mediaflux™ package) provide the primary interface to the system for research scientists. These interfaces are generic and domain independent.

SPECIFIC NEUROIMAGING IMPLEMENTATION

The Framework object model and implementation is currently being used to manage a data repository in the Neuroimaging domain. Services that are not explicitly part of the Framework implementation are used to upload the data (and some associated metadata) into the repository (e.g. a DICOM client). The repository manages over 60 projects that contain mainly MR data (human and small animal) in DICOM (and proprietary formats) and optical microscopy data in TIFF format. Thus we have defined modular (reusable) XML metadata documents and *Method*s specifically to handle these kinds of data in a neuroimaging research environment.

In this implementation, an authorized user first creates a *Project* object, defining the project goals, project context and the team members (and their roles). When the *Project* is created, pre-existing *Method* objects (one or more) are also registered for use with that *Project*. Subsequently, team members with the `SubjectAdministrator` role for this project create *Subject*s (and possibly *R-Subject*s) as needed (*ExMethod* objects are auto-created in this process). *Study* objects are generally created as needed by the agents that upload data (although they can pre-created).

A design principle of the implementation has been to enable the creation of adaptive user interfaces by providing services

[12]Mediaflux™ digital asset management platform. http://www.arcitecta.com

that: (i) retrieve the metadata required to create objects and (ii) retrieve metadata and data on existing objects for subsequent presentation. The implementation makes heavy use of *Method* objects. In particular, a *Method* object defines the metadata required to create *Subject* (and possibly *R-Subject*) objects; this can be thought of as a specialized *Method* step. The *Method* object also defines the metadata required per step of the *Method* during execution and this may include metadata for *Study* objects. The *Method* may pre-specify metadata values and whether it is immutable or not.

As an example, **Figure 3** shows the metadata required to create a *Subject* for a specialized *Method* that combines MR, optical microscopy and electron microscopy image data acquired in translational research of mice (Wu et al., 2007).

This *Method* specifies that the subjects are a particular strain of mouse targeting a specific disease (and these metadata are immutable). Details such as birth date are entered by the user to complete the *Subject* creation. Other *Method*s may specify the use of an *R-Subject*, or different metadata for the creation of the *Subject*/*R-Subject*.

The *ExMethod* (the instantiation of the *Method*) object that was (auto) created for the above *Subject* is show in **Figure 4**. This *Method* acquires MR (of the whole brain) and optical microscopy (of the removed optic nerve) images for mouse subjects. Each numbered step has a name and specifies metadata, the state, and whether a *Study* is created or not. The inset shows the metadata for the *Perfusion* step. These metadata are immutable and pre-specified by the *Method* so that entry by the user is not required.

The subject undergoes distinct (permanent) state changes during the execution of the *Method*. When the imaging data are uploaded and the *Study* objects created, each *Study* is tagged with the relevant step of the *Method*. The *Method* branches can be executed in parallel or serially as the tissue specimens are imaged. Each removed tissue specimen could be represented as a new (disassembled) *Subject*.

Substantial effort from a number of groups has begun the development of biomedical ontological frameworks (e.g. the Unified Medical Language System[13] and the Open Biomedical Ontology[14] (Smith et al., 2007). Specification of metadata in the system could adhere to existing domain standards either by direct use of metadata definitions, or by the ability to inter-operate through exchange processes (e.g. utilizing XSL and XSL Transformations[15]). The implementation of metadata also needs to remain flexible so that scientists can incorporate any metadata that they need, whilst still retaining standard components.

Because the PSSD framework enables project-specific *Method* specification, and because each *Method* specifies metadata independently, the system provides for flexibility and the adherence to standards.

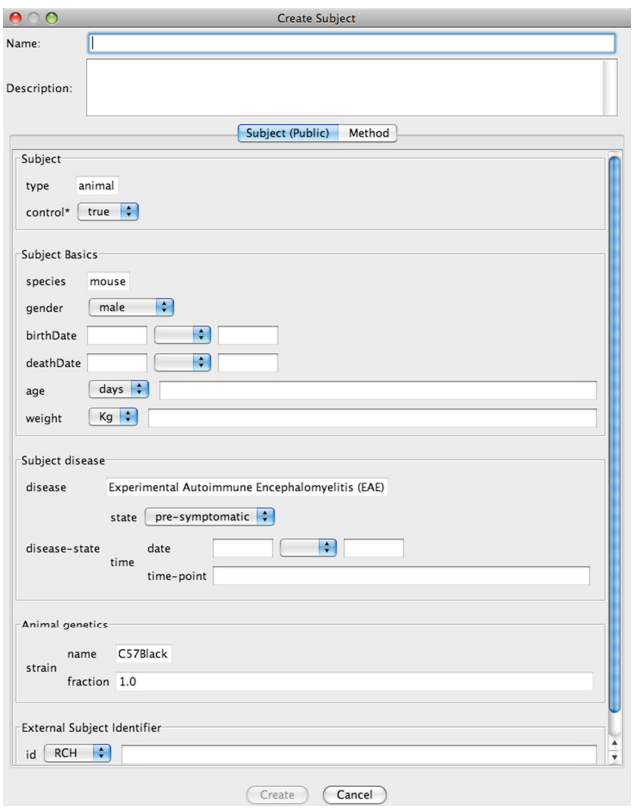


FIGURE 3 |The metadata specified by a particular *Method* (developed for a particular *Project*) that is required to create a *Subject*. The adaptive graphical interface interrogates the *Method* to discover the required metadata. Metadata are presented in XML fragments. Some metadata are predefined and immutable (e.g. *species*) whereas other metadata requires entry.

DISCUSSION

SIGNIFICANCE

Modern scientific research involves distributed collaborative teams, distributed data with distributed processing[16,17]; these are aspects of the e-Research paradigm. Whilst the need to organize information via an object model and the ability to federate information is of course not new, the Framework and methodology described in this paper have a number of significant advantages for e-Research applications. Firstly, the use of a distributed object model enables project teams to participate in a collaborative research project whilst using distributed data repositories and interfaces. Distributed object collections can be managed using the semantics of the citable identification scheme without requiring costly and potentially error prone distributed or centralized registries.

Secondly, codifying research processes into a *Method* means that: (i) *Methods* can be presented unambiguously and reviewed using simple diagrams, (ii) *Methods* can be re-used, (iii) application interfaces can be automatically constructed, (iv) researchers can define new research method(s) without requiring the development of new application interfaces to support the execution of those methods, and (v) the metadata for each class of experiments is derived from the relevant *Method*(s). Note that a *Method* can contain a super-set of any existing metadata standard. Importantly, by recording all state changes for a subject regardless of whether they are transient or permanent, the conditions

[13]http://www.nlm.nih.gov/research/umls

[14]http://www.obofoundry.org

[15]http://www.w3.org/Style/XSL/

[16]http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

[17]http://www.jisc.ac.uk

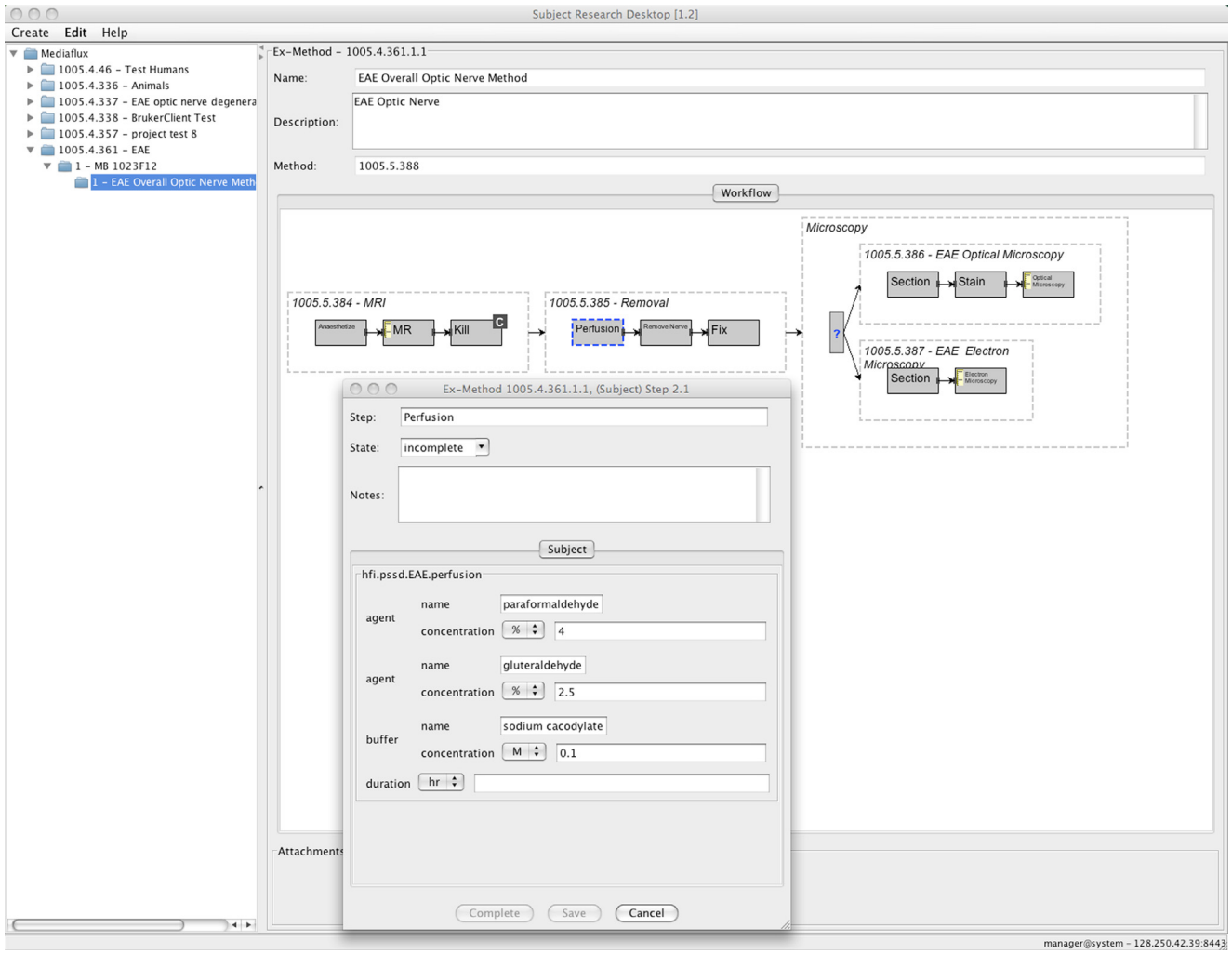


FIGURE 4 |The adaptive interface shows the object trees for the projects that the user is authorized to access. The *Project* with citable ID `1005.4.361` is opened and the *ExMethod* object `1005.4.361.1.1` is displayed. For presentation, this figure shows a simplified version of the *ExMethod* object (it has more steps in reality). The inset shows the (immutable) metadata for the Perfusion step. It can be seen that the overall *Method* (`1005.5.388`), from which this *ExMethod* is instantiated, was built from a number of *Method* fragments (`1005.5.[384,385,386]`).

the led to the acquisition of data can be identified, reviewed and reconstructed.

Thirdly, identification of "real" subjects (*R-Subject*) enables identification of all projects in which a particular subject has participated. For example, a genetic sequence may be identified in a subject that was not previously known. The state of the *R-Subject* could then be updated, with prior research conducted using that subject re-analyzed.

Finally, the Framework object model is extensible to accommodate new relevant information. For example, a human subject may enter into an agreement defining the terms and conditions under which their data may be used. That agreement may apply to all projects in which they have participated or alternatively may be project specific. The agreement may be scanned and associated with either the *R-Subject* or *Subject* objects, depending on the scope of the agreement. Similarly, a researcher may associate other information (via new objects) such as documents or data with any object.

IMPLEMENTATION CONSIDERATIONS

Our implementation of the Framework utilizes a service-oriented digital asset management platform which supports distributed citable identification and distributed repositories. All metadata are encoded using XML. Depending on the type of research, XML schemas for metadata are defined using existing standards where they exist, or defined specifically for the research method, or a combination of both. The Framework may be implemented with any service-oriented system utilizing most database technologies. A service-oriented approach, such as web-services, ensures user interfaces and other systems interact with the Framework's interface, hiding the underlying method of implementation. The key capabilities supported are: (i) citable identifier allocation, (ii) object creation with the ability to associate metadata and arbitrary data with an object, (iii) metadata definitions (e.g. XML Schema) so that domain-specific metadata can be created for any type of object, and (iv) distributed data repositories where distributed projects are undertaken.

LIMITATIONS

The Framework has been developed for subject-centric research and thus is not necessarily optimal for other research domains. The number of objects in the object model has been minimized in order to improve accessibility of the model by researchers. However, a number of important aspects of information management are not included in the Framework. For example, many information models and metadata schema have been developed for the preservation of digital data (see OCLC working group report[18]). The development of a long-term information management capability requires the incorporation of aspects of these models and schemas. Since the Framework object model is extensible, future integration with other information object model components is possible.

The Framework includes the ability to notate and track subject state. In neuroimaging research the subject state changes slowly. However, this limitation could be overcome by acquiring vectors of metadata during the data acquisition process in order to measure rapid state changes. Whilst the Framework has broad applicability, limitations may arise from wider application of it to other domains of subject-centric research.

FUTURE WORK

Future developments of the Framework in the neuroimaging domain will include acquisition of data from different imaging modalities as well as increasingly complex workflows in distributed projects. Research outcomes should be enhanced by integration of the Framework with other resources such as application processing pipelines, brain atlases and publication portals. Finally, tools that support research uses of the model are being developed including a graphical user interface application to enable researchers to create *Methods* and define metadata themselves. The Framework will promote modularization of research processes and associated metadata, which in turn promote re-use and standardization. The unpredictable path of future research provides a significant challenge for identifying re-usable research specific metadata, but is important for interoperability and retrospective interpretation.

CONCLUSIONS

A Framework that incorporates an object model and research methods for distributed subject-centric research has been developed. The Framework facilitates multi-disciplinary and collaborative subject-based research, and extends earlier object models used in the research imaging domain. Whilst the Framework has been explicitly validated for neuroimaging research applications, it has broader applications to other fields of subject-centric research.

ACKNOWLEDGEMENTS

We thank Gavan McCarthy, Steve Melnikoff, Anna Shadbolt, Lyle Winton, Wilson Liu and Wee Siong-Soh for discussions and development that have helped us to validate and refine this work. We also acknowledge grants from the Australian Research Council (grants LE0561231, SR0564829) and the University of Melbourne (Cross-Faculty Fund 2006) that have in part supported this work.

[18]The Online Computer Library Centre & Research Libraries Group (OCLC/RLG) Working Group on Preservation Metadata. Preservation Metadata and the Open Archival Information System (OAIS) Information Model, 2002. http://www.oclc.org.research/pmwg.

REFERENCES

Brase, J. (2004). Using digital library techniques – registration of scientific primary data. *Lect. Notes Comput. Sci.* 3232, 488–494. doi: 10.1007/b100389.

Gardner, D., Toga, A. W., Ascoli, G. A., Beatty, J. T., Brinkley, J. F., Dale, A. M., Fox, P. T., Gardner, E. P., George, J. S., Goddard, N., Harris, K. M., Herskovits, E. H., Hines, M. L., Jacobs, G. A., Jacobs, R. E., Jones, E. G., Kennedy, D. N., Kimberg, D. Y., Mazziotta, J. C., Miller, P. L., Mori, S., Mountain, D. C., Reiss, A. L., Rosen, G. D., Rottenberg, D. A., Shepherd, G. M., Smalheiser, N. R., Smith, K. P., Strachan, T., Van Essen, D. C., Williams, R. W., and Wong, S. T. (2003). Towards effective and rewarding data sharing. *Neuroinformatics* 1, 289–296.

Keator, D. B., Gadde, S., Grethe, J. S., Taylor, D. V., and Potkin, S. G. (2006). A general XML schema and SPM toolbox for storage of neuroimaging results and anatomical labels. *Neuroinformatics* 4, 199–212.

Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H. J., Papadopoulos P, BIRN Function; BIRN Morphometry; and BIRN-Coordinating. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., and Wächter, J. (2006). Data publication in the open access initiative. *Data Sci. J.* 5, 79–83.

Klump, J., Brase, J., Diepenbroek, M., Grobe, H., Hildenbrandt, B., Höck, H., Lautenschlager, M., and Sens, I. (2008). Use of Persistent Identifiers in the Publication and Citation of Scientific Data. AGU Fall Meeting, 5–19 December 2008, San Francisco, CA, USA. Available at: http://epic.awi.de/epic/Main?puid=31047&lang=en.

Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34.

Marenco, L., Tosches, N., Crasto, C., Shepherd, G., Miller, P. L., and Nadkarni, P. M. (2003). Achieving evolvable Web-database bioscience applications using the EAV/CR framework: recent advances. *J. Am. Med. Inform. Assoc.* 10:444–453.

Martone, M. E., Gupta, A., and Ellisman, M. H. (2004). E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nat. Neurosci.* 7, 467–472.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone S-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255.

Toga, A. (2002). Neuroimage databases: the good, the bad and the ugly. *Nat. Rev. Neurosci.* 3, 302–309.

Wu, Q., Butzkueven, H., Gresle, M., Kirchhoff, F., Friedhuber, A., Yang, Q., Wang, H., Fang, K., Lei, H., Egan, G. F., and Kilpatrick, T. J. (2007). MR diffusion changes correlate with ultra-structurally defined axonal degeneration in murine optic nerve. *Neuroimage* 37, 1138–1147.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 February 2009; paper pending published: 04 April 2009; accepted: 23 June 2009; published online: 08 July 2009.

*Citation: Lohrey JM, Killeen NEB and Egan GF (2009) An integrated object model and method framework for subject-centric e-Research applications. Front. Neuroinform. (2009) **3**:19. doi:10.3389/neuro.11.019.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 28 August 2009
doi: 10.3389/neuro.11.027.2009

CamBAfx: workflow design, implementation and application for neuroimaging

Cinly Ooi[1,2]*, Edward T. Bullmore[1,2], Alle-Meije Wink[3], Levent Sendur[1,2], Anna Barnes[1,2], Sophie Achard[1,2], John Aspden[4], Sanja Abbott[2], Shigang Yue[5], Manfred Kitzbichler[2], David Meunier[1,2], Voichita Maxim[1,2], Raymond Salvador[2], Julian Henty[1,2], Roger Tait[1,2], Naresh Subramaniam[2] and John Suckling[1,2]

[1] *Brain Mapping Unit, Department of Psychiatry, University of Cambridge, Cambridge, UK*
[2] *Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge, UK*
[3] *Imaging Sciences Department, MRC Clinical Sciences Centre, Imperial College, London, UK*
[4] *J.L. Aspden Ltd, Cambridge, UK*
[5] *Department of Computing and Informatics, University of Lincoln, Lincoln, UK*

Edited by:
John Van Horn, University of California, USA

Reviewed by:
Ivo Dinov, University of California, USA
John Van Horn, University of California, USA

Correspondence:
Cinly Ooi, Brain Mapping Unit, Herchel Smith Building for Brain And Mind Sciences, University Forvie Site, Robinson Way, Cambridge CB2 0SZ, UK.
e-mail: co224@cam.ac.uk

CamBAfx is a workflow application designed for both researchers who use workflows to process data (consumers) and those who design them (designers). It provides a front-end (user interface) optimized for data processing designed in a way familiar to consumers. The back-end uses a pipeline model to represent workflows since this is a common and useful metaphor used by designers and is easy to manipulate compared to other representations like programming scripts. As an Eclipse Rich Client Platform application, CamBAfx's pipelines and functions can be bundled with the software or downloaded post-installation. The user interface contains all the workflow facilities expected by consumers. Using the Eclipse Extension Mechanism designers are encouraged to customize CamBAfx for their own pipelines. CamBAfx wraps a workflow facility around neuroinformatics software without modification. CamBAfx's design, licensing and Eclipse Branding Mechanism allow it to be used as the user interface for other software, facilitating exchange of innovative computational tools between originating labs.

Keywords: workflow, eclipse, rich client platform, batch processing, camba, open source, pipeline

INTRODUCTION

Workflows are the combination of pipelines (i.e. modules representing individual programs with connecting pipes representing data transfer from one module to another) and data control systems that coordinate data processing on local or distributed computer architectures. Neuroimaging brings together two broad scientific constituencies: the design and implementation of workflows and the application of these workflows to brain imaging datasets. Correspondingly, the demands made upon workflow-based software change according to circumstances.

Conceptually, workflows are a useful way to gain traction over complex data analysis tasks. By decomposing the workflow into constituent parts, the problem is reduced to the creation and maintenance of small, simple programs that can be reused across workflows. To workflow designers (*designers*), the development environment should offer uncomplicated integration of their programs into existing pipelines, quick construction of new pipelines from existing modules and facilities for rapid testing, validation and deployment of workflows.

For those who apply workflows (*consumers*), the small effect sizes and large between-subject variance associated with most neuroimaging techniques call for a simple system for entering data into the workflow at low error rates and data control systems that emphasize high dataset throughput. Ideally, all workflows should follow a common ontology and it should be possible to use the same workflows with different data control systems without modification.

Whilst workflows are important tools for both designers and consumers in their own right, they also form a vital line of communication within the neuroimaging community to disseminate new algorithms or pipelines, optimised module parameters and standardised procedures. A workflow represents the collective wisdom on how to perform a data analysis task and documents the process allowing experience to be reused, transferred, and consolidated.

A review of workflow applications reveals that the majority of existing workflow environments are effective at modifying pipelines, but not optimised for processing large amounts of data. CamBAfx is an Eclipse (International Business Machines, 2006) Rich Client Platform (RCP, McAffer and Lemieux, 2005) based workflow application that provides both a front-end (user interface) optimized for data processing and a back-end pipeline model to facilitate creation and manipulation of pipelines. Additionally, it offers the flexibility of using different process strategies (for example, single machine scripting or grid-based computing) within the same pipeline description.

We begin with a brief overview of alternative workflow applications providing context for the objectives of CamBAfx. The operation of CamBAfx from the viewpoint of consumers is then considered showing how the user interface helps in the analysis of their data. For designers, discussion is orientated towards the delivery of the pipelines and workflows to consumers through the deployment of Eclipse-based facilities. Examples of delivering pipelines from a variety of neuroinformatics packages are given and concluding remarks made on future directions.

WORKFLOW ENVIRONMENTS

Workflows are normally visualised as pipelines, i.e., a collection of modules with pipes to represent the data flow from output ports of one module to the input ports of another. Traditionally, Visual Pipeline Editors (VPEs) are used to manipulate pipelines. VPEs represent pipelines graphically, usually with boxes as modules, lines as pipes, and small shapes inside the module box as input and output ports. Users modify workflows by manipulating this graphical representation, such as adding modules or re-routing pipes. Commercially available software offers workflow capability in two different ways: either specialised for workflow operations (National Instruments' LabView)[1], with a VPE as the main user interface and programming interface for module creation and different data processing strategies, or as extensions to existing programming languages (Simulink)[2] that provide VPEs and programming interfaces for modules to accommodate pipelines.

The LONI Pipeline (Rex et al., 2000) looks and behaves like a traditional Visual Pipeline Editor. To enter data, consumers click on input ports which then request single values or a list of values. Batch processing is achieved by asking the input port of a module to interpret a list of values one-at-a-time instead of all-at-once. For batch-processing, LONI Pipeline offers the run-on-machine method, including via a script containing the individual processing instructions, as well as grid processing. It uses Extensible Markup Language (XML, Bray et al., 2008) to describe the pipeline as a combination of modules, connections, ports and data. Conveniently, meta-data about modules such as their creators and the software suite to which the modules belong can also be stored. LONI Pipeline modules may be downloaded separately to augment the main package.

Fiswidgets (Fissell et al., 2003) visualizes its pipeline as a linear stack without pipes or ports. Modules need not be activated in the order the visual representation implies. Clicking on modules brings up a module window that asks for data and parameters. Fiswidgets' modules are defined in Java or in XML and describe the layout of the module window. For batch-processing a visual programming approach is adopted with loop structures for iteration within the pipeline. Inside the module windows, symbols define inputs and outputs. During data processing the symbols are substituted with the corresponding values from a lookup table. Fiswidgets distributes modules as part of the main software.

BrainVISA (Cointepas et al., 2001) has a collection of workflows each with a "configuration page" that presents a workflow as a tree of modules. Important module parameters can be attached as leaves to the module in the tree, others in an associated detail page. Batch processing is initiated by duplicating the "configuration page" for each dataset. BrainVISA's pipeline is implemented in the form of Python scripts. Pipelines are delivered in toolboxes bundled with the software or downloaded post-installation. The toolbox itself is a directory of configuration files, binary files, text files, help files and python scripts. BrainVISA has an optional database for managing datasets that uses a data ontology and provides software for conversion between images file formats.

In summary, based on the applications' look-and-feel both LONI and FisWidgets give strong emphasis to pipeline manipulation while softwares like BrainVISA prioritise clear data entry. Finally, an established way to deliver workflow-based software is to write a custom user interface for each workflow; some program interfaces in FSL (Smith et al., 2004) and SPM (Friston et al., 1995) fall into this category.

CamBAfx is a user interface for neuroinformatics software designed to support multiple pipelines and to provide the facilities needed to support workflow operation; namely, data management and batch processing. The philosophy is to provide the shortest possible bridge between designers and consumers, iteratively improving processing with pipelines via software development and practical experience. CamBAfx aims to provide resource in equal measure to both constituencies.

CamBAfx

OBJECTIVES

Workflows evolve as algorithms are developed and applications become more demanding. A Workflow environment must therefore be able to maintain flexibility for development while being able to include new applications without modification and maintain a consistent user interface across all pipelines. Thus, a major objective in the design of CamBAfx is to provide for consumers' needs at the front-end, while exploiting the flexibility of workflows at the back-end in order to deliver the pipeline assembly capability for designers. As expectations change, the environment should be flexible enough to refocus these different aspects from front-end to back-end and vice-versa.

The user interface practices a minimalist philosophy: the initial download is a complete, ready-to-use package but only contains those functions that are needed immediately to get started. Consumers customise the interface as dictated by their needs.

Generic functions to manage pipelines and data are provided. Designers are encouraged to make their pipelines more attractive by adding supporting functions.

The environment should reuse existing industrial-grade software and follow existing and de facto standards and practices. Availability of an Integrated Development Environment (IDE) that supports day-to-day programming work such as debugging, version control and automation of mundane tasks greatly improves developers' productivity.

FRONT-END: RESOURCES FOR WORKFLOW CONSUMERS

Our observations indicate that normal practice for workflow consumers is to maintain a library of workflows. Once a workflow has been demonstrated as robust and capable, its composition and parameters are infrequently reconfigured suggesting that it would not be appropriate to focus on workflow manipulation capability for these users. Instead, the biggest workload undertaken by consumers is to enter specific data instances into the workflow and to ensure the data is valid to maximize the success rate of processing. Thus, the front-end of CamBAfx has as its most important undertaking the acceptance and validation of data entered by consumers. Careful validation of the data reduces the number of problematic datasets in a multi-subject dataset, but cannot completely eliminate them. The problems that then arise are corrected between repeats of batch processing. The challenge is to design a system that accommodates multiple repeats, but

[1]http://www.ni.com/labview/

[2]http://www.mathworks.co.uk/products/simulink/

reduces unnecessary reprocessing of datasets already successfully processed. This overall process maps well onto a traditional software usage pattern:

(1) select a workflow and configure it
(2) enter the data into the interface
(3) run the processing in batch mode

Selecting and configuring workflows

In CamBAfx, the process starts by selecting a pipeline from a library of pipelines using a New Wizard (**Figures 1A,B**). A pipeline-specific wizard (**Figure 1C**) is then used to guide the configuration of the pipeline, including a review of the important module parameters and requests to supply values to parameters that cannot have default values. CamBAfx requires pipeline designers to guarantee that the

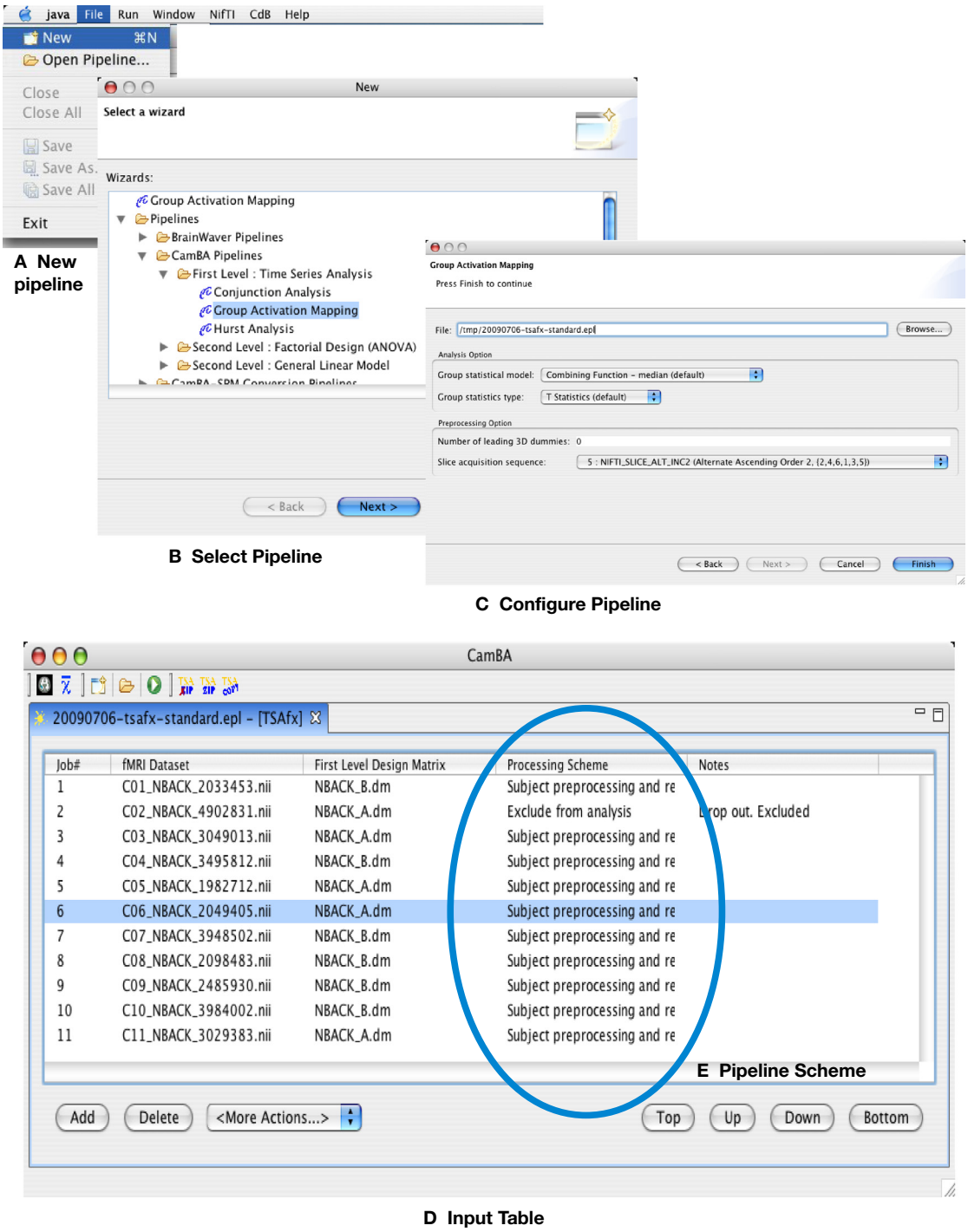


FIGURE 1 | Steps in operation by consumers to select, create and modify workflows. See text for details.

pipeline created at the end of this process is valid and immediately useable.

Data entry via the interface

The pipeline itself is not graphically represented. Instead an Input Table (**Figure 1D**) is presented where all the data necessary for batch processing is specified. The Input Table is customised to the workflow, although there is consistency across the instances of the interface for each pipeline. In general, each row refers to the data for a particular imaging dataset. A table cell only displays the appropriate interactive element determined by the pipeline to solicit data (e.g. text boxes, drop-down lists of choices, file and directory selection dialogs). If the data required is a list, then a new table with one column is presented with the same interactive element facilities as the Input Table. If there are two or more list-based data required, they can each use a separate table or share a multi-column table.

To improve the chances for successful data processing the table cells accept or reject data following input. This can be as simple as rejecting letters when numbers are expected or enforcing specific restrictions imposed by the pipeline, such as minimum and maximum values or lengths. Error messages, possibly containing a message from the pipeline designer, are displayed to the user where available. The Input Table additionally contains a free-text cell entitled "Notes" where annotations can be made about the dataset.

Associated with each dataset is a "Pipeline Schemes" (**Figure 1E**). This is a drop-down list with preconfigured schemes that define the precise list of modules activated in the processing of that dataset. By default, the two schemes that bound the possible processing are available; namely, one that activates all modules and another that entirely bypasses all the modules. Pipeline designers can add new schemes that activate only part of the pipeline and in doing so lead to more efficient analyses of datasets that have been partially processed previously.

A drop-down box below the Input Table (**Figure 2A**) is used to host functions that work on the Input Table as a whole. A function to copy data from another instance of the same pipeline is available. Pipeline designers can add pipeline-specific functions into this drop-down box. The table of parameters (**Figure 2B**) can also be invoked from here. Parameters are variables for modules that remain constant throughout processing of the datasets (e.g. a spatial smoothing kernel). In keeping with the philosophy of a pipeline-centric view, this table shows all parameters for all modules. It uses a two column format with one parameter per row. The first column contains the parameter name and the second its value. The table offers the same interactive elements and validation facilities as the Input Table. For parameters that must share the same value, only one will be listed and any modification here is propagated to all parameters.

Batch mode processing

Once data entry is complete, the workflow is initiated via the "Run Wizard" (**Figures 3A,B,C**). Here additional information required by the data processing engine, such as the summary output directory name, will be requested. Currently, the data processing engine operates by script generation and execution.

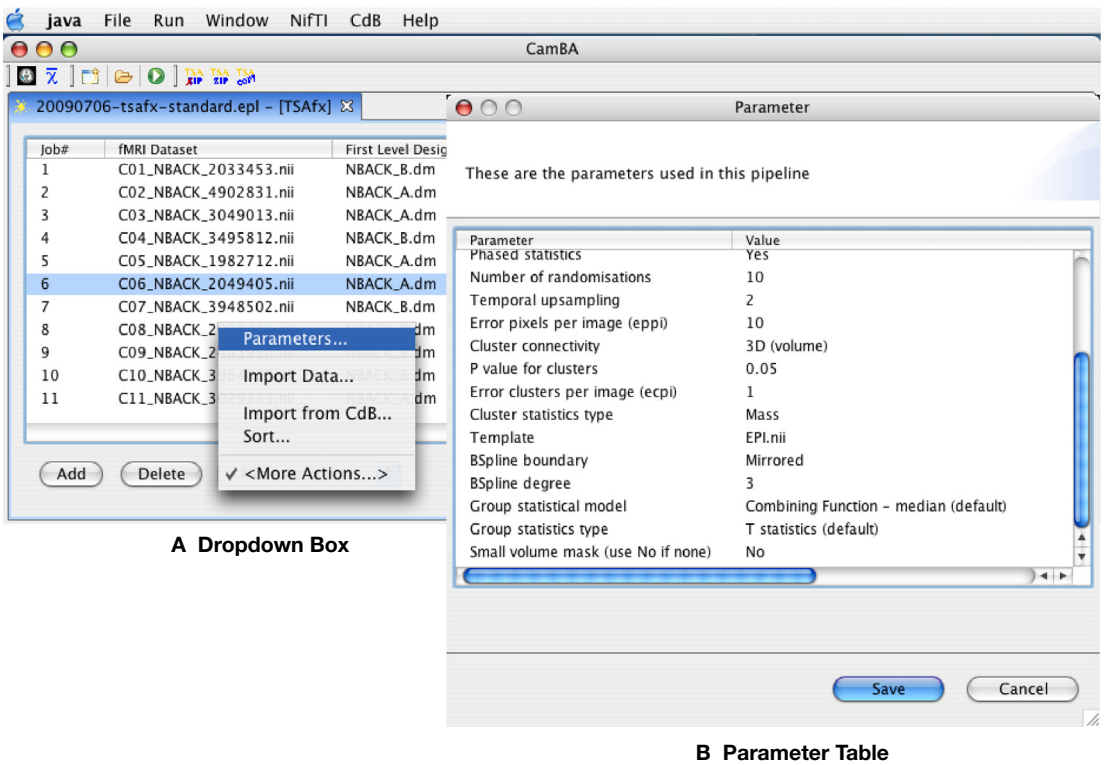


FIGURE 2 | Steps in operation by consumers to modify parameters for the workflow. See text for details.

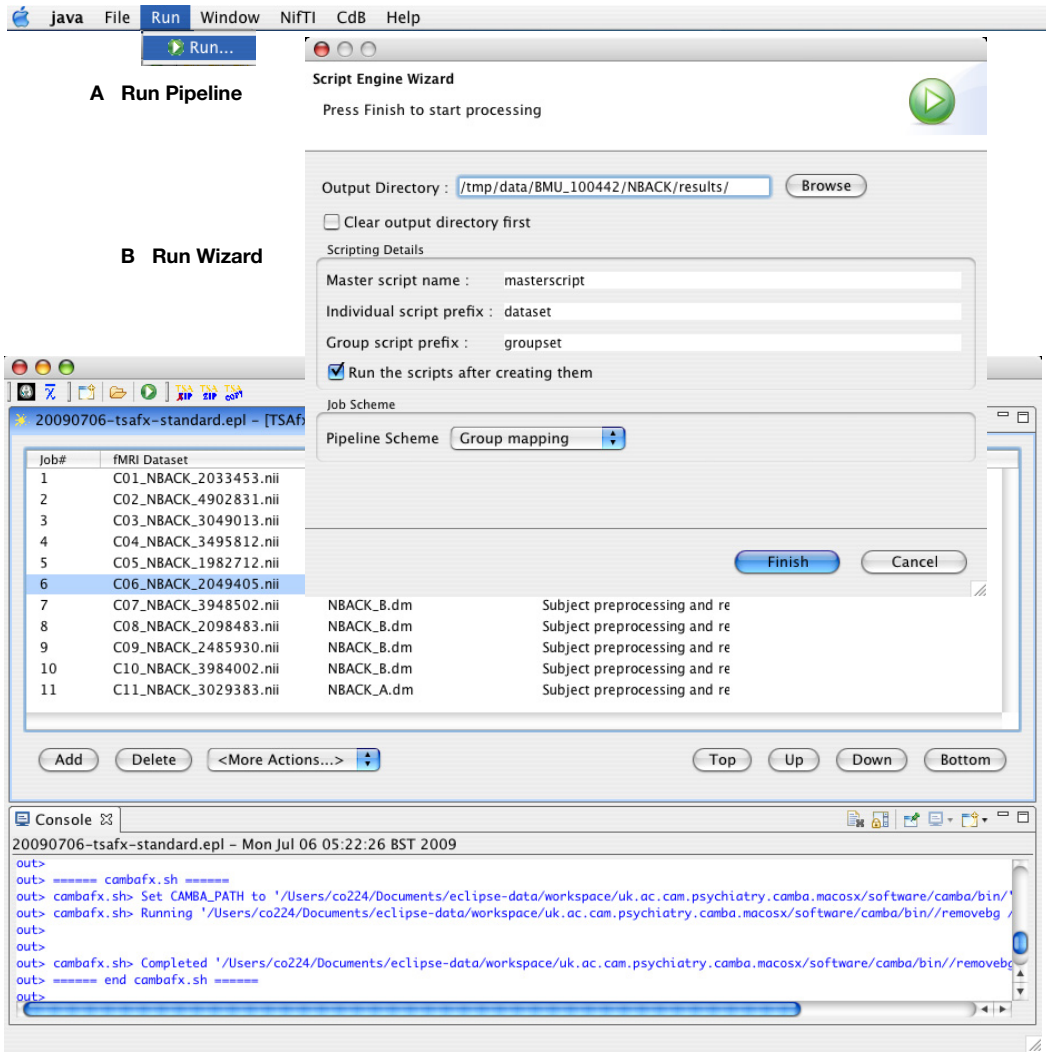


FIGURE 3 | Steps in operation by consumers to run the workflow. See text for details.

Other practical issues

CamBAfx is a self-extracting archive available for download[3] containing both CamBAfx, the workflow environment, and a set of pipelines based on modules of the CamBA software (Suckling and Bullmore, 2004; Suckling et al., 2006). Also included are supporting functions such as functions to copy the results of one pipeline as the input to another. New pipelines and functions are delivered post-installation as plug-ins that are downloaded, dropped into the original installation and included into the distribution following a restart of the software. Most plug-ins orientated towards consumers modify the user interface to advertise their availability.

BACK-END: RESOURCES FOR WORKFLOW DESIGNERS

Out of the box, CamBAfx has all the generic facilities needed to manage workflows. For all pipelines, CamBAfx provides all the expected facilities to configure pipelines as well as collect, collate and batch-process datasets that together form the workflow. However, since the Eclipse Extension Mechanism (EEM, Bolour, 2003) gives access to the user interface and allows them to contribute new functions, CamBAfx plug-ins customize the user interface to support the specific processing requirements of each pipeline and implement support facilities such as data imports from other pipelines.

Pipeline features

All information CamBAfx needs is contained in the pipeline file, written in XML, with three sections: Pipeline, Input Data and Preferences. The Pipeline section represents the pipeline as a collection of modules and connections. The modules are further decomposed into variables (i.e. installation specific values), parameters, input and output ports, and how to invoke the program. Almost everything describing the pipeline is in XML except for complex data manipulation, such as generating the command

[3]http://www-bmu.psychiatry.cam.ac.uk/software/

line instructions, where Java program code is used in the form of a BeanShell Script[4]. Variables, parameters, input and output ports all carry datatype information (e.g., integer or string) and include restrictions on the data. All pipeline components can have variations on, for example, datatypes, modules (input or standard) and ports (data or signals). They start with a XML element with the same name, but with an attribute that identifies the variant. The attached XML leaf elements change according to the variant. The Input Data section simply contains a description of each dataset as displayed by the Input Table. The Preference section contains optional information about the pipeline such as pipeline schemes and a list of linked parameters that should share the same value.

Steps are taken to make the pipeline simpler and easier to understand: First, looping constructs, normally used to effect batch processing, but complicating data flow, are eliminated by insisting that each dataset is processed through the complete workflow from beginning to end and that each input port can only have one connection. Second, uncertainty about whether an input port needs to be connected is removed by insisting that all ports need to be connected. To satisfy this, and to show where datasets enter the pipeline, each pipeline has one (and only one) input module responsible for communication with the outside world.

Data standard and datatype hierarchy

For effective data exchange between modules, CamBAfx has a Data Standard for all datatypes it uses that defines the file format and the meta-data it must provide. For example, functional magnetic resonance imaging data (fMRI) is in 4D NifTI (Cox et al., 2004) single file format and must carry the sequence in which the slices of the three-dimensional volume were acquired, which is encoded as the `slice_code` meta-data. This approach guarantees the exact content available to designers for writing modules. In return, the output from a module should also satisfy this standard and the designer is responsible for converting data to and from the data format their program expects. Adhering to this data standard means data can be easily exchanged between modules. Designers only have to convert their data to one other format, i.e. to the data standard only and not all possible data formats they might encounter. Although CamBAfx is organised to validate data against the data standard following input, this is postponed until CamBAfx develops the appropriate editors to edit the data in situ as consumers prefer to be able to do this if their data fail validation.

Datatypes are organized into a hierarchy, with each datatype having only one parent and children must carry all data inherited from its parent as well as optional data of its own. A special equivalence is used to define a unidirectional relationship between two datatypes that do not share a common ancestry. This data hierarchy tree is used to prevent incompatible data transfer between modules in the VPE by restricting connection of output ports to input ports that expect the same datatype or its parents.

New pipeline wizards

A new pipeline can be created by cloning, i.e., loading the pipeline into the user interface and then saving it under a new name. This approach may, however, also copy unwanted details from the old pipelines, such as the specific dataset names and modifications to the pipeline. Therefore in CamBAfx, the preferred approach is to create pipelines using a New Pipeline Wizard where the new pipeline is cloned from a clean copy of the parent pipeline and can be manipulated if necessary before being presented to consumers.

CamBAfx DESIGN AND ARCHITECTURE

Eclipse and eclipse rich client platform

CamBAfx is an Eclipse Rich Client Platform (RCP, McAffer and Lemieux, 2005) application. Eclipse[5] (International Business Machines, 2006) was originally created as an IDE with an extension mechanism (Eclipse Extension Mechanism, EEM, also known as Eclipse Plug-in Architecture, Bolour, 2003) designed to integrate development tools. The EEM is a way of extending an Eclipse-aware program. A program that supports extensions publishes an extension point and its expectation. Interested parties then provide extension(s) that latch on to this extension point. Extensions can provide configuration information or program code or both and together with their supporting data, such as icons and programs, are packaged into plug-ins.

Eclipse itself is designed as a collection of plug-ins, with the exception of a small kernel that starts up and bootstraps the EEM. After bootstrapping, the EEM discovers and manages all the installed plug-ins. It then searches the command that invoked it, and if necessary a configuration file, to find the master application. This is read through the EEM and executed. In the original design there was only one master application: the Eclipse IDE. However, the Eclipse Extension Mechanism proved sufficiently useful as a platform for development of standard programs that it was exploited by the Rich Client Platform (RCP) project. The RCP project allows other applications, such as CamBAfx, to be the master application.

All RCP applications are programs built using the EEM, and all share a common architecture and plumbing. RCP developers simply write the missing part, i.e. the program code specific to their project and insert it into the RCP framework.

CamBAfx as a RCP application

CamBAfx, like all RCP applications, is actually a collection of plug-ins. For example, all CamBA's command line programs, pipelines and supporting functions are encapsulated into Eclipse plug-ins and managed through the EEM. Tasks such as creating a New Pipeline Wizard are performed by extending CamBAfx using EEM.

The Eclipse extension point *org.eclipse.core.runtime.applications*, is the only mandatory extension point allowing CamBAfx to be invoked as a master application. CamBAfx also uses other Eclipse extension points such as *org.eclipse.ui.editors* for the main Input Table and *org.eclipse.ui.actionSets* to add menu and toolbars items. CamBAfx also defines its own extension points including the *org.genericfx.ui.inputtable.taskagents* extension point which adds items to the Input Table's drop-down box. Extension points, such as *org.genericfx.data.hierarchy* which define the datatype hierarchy customize CamBAfx for designers. CamBAfx also provides a special generic New Pipeline Wizard for the *org.eclipse.ui.newWizards* extension point eliminating the need to write generic wizards for pipelines by using instead CamBAfx's *org.genericfx.ui.base.newWizards* extension

[4]http://www.beanshell.org/

[5]http://www.eclipse.org/

point to read in the pipeline from a file. Part of CamBAfx, such as new items for the Input Table drop-down box, are constructed by extending its own extension points. All extension points, either those of Eclipse or CamBAfx, are available to downstream developers who can also define their own.

Developing for CamBAfx

As standard Eclipse plug-ins, CamBAfx and its plug-ins are developed using Eclipse's Plug-in Development Environment (PDE, Melhem and Glozic, 2003) that is designed specifically to develop, test and integrate plug-ins with their intended application. CamBAfx provides an editor, integrated into the IDE, for development and testing of pipelines. This editor has the Input Table and a rudimentary VPE. CamBAfx has two data processing engines: traditional batch processing controlled directly by the program itself and a version that writes and then executes the processing steps via scripts. Both are callable from the IDE via its Run Wizard.

Eclipse also makes available supporting software facilities, such as an update mechanism and help browser. It provides tools for CamBAfx such as the Graphical Editor Framework[6] (GEF, Hudson, 2004) which is the basis of CamBAfx's VPE.

Developers "pick-and-mix" CamBAfx plug-ins for their applications. Architecturally, there are three major parts: Pipeline, Input Table and Data Processing Engine (**Figure 4**). These three parts are kept independent of each other with minimum communication between them. Conceptually, the software is developed in three layers (**Figure 5**): At the bottom is GenericFX, a complete generic pipeline application; BrainFX is the middle layer that customizes GenericFX for neuroinformatics applications by defining the data hierarchy, data standard and some commonly used routines, such as NifTI data conversion. CamBAfx is the top layer and contains only CamBA-specific pipelines and functionalities. Third party developers who do not need CamBA can create their applications from either GenericFX or BrainFX. The same Eclipse Branding Mechanism (Eidsness and Rapicault, 2004) that defines CamBA's own About Dialog, splash screen and icons can be used to brand other applications.

[6]http://www.eclipse.org/gef/

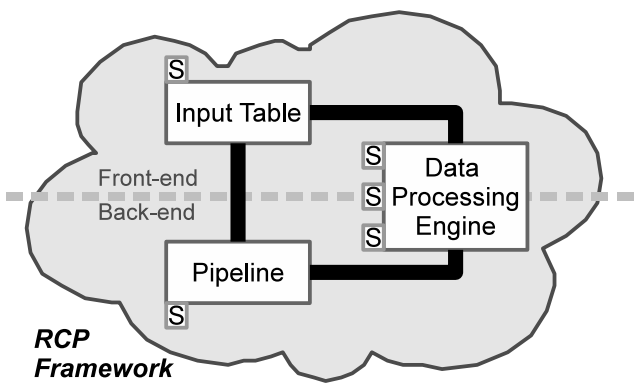


FIGURE 4 | Conceptually, CamBAfx application is constructed from of three components: Pipeline, Input Table and Data Processing Engine, integrated inside the Eclipse Rich Client Platform (RCP) framework. Minimal communication between each component allows components to be replaced or removed. The separation between front-end (user interface) and back-end can be reconfigured as needs change. "S" denotes supporting functions which can be attached to the application as plug-ins.

IMPLEMENTATION OF PIPELINES

CamBA ANALYSIS PIPELINES

CamBA is software for the analysis of neuroimaging data. The initial download contains a number of pipelines available for first-level (within-subject) and second-level (between-subject) analysis for which CamBAfx provides customised interfaces. The CamBAfx application running CamBA pipelines has been widely used in the analysis of functional and structural MRI (examples include: Chamberlain et al., 2008, 2009; Habets et al., 2008; Menzies et al., 2008; Wink et al., 2008).

CamBA's first-level analysis pipelines' main purpose is to generate maps that summarise responses or signal properties from raw 4D fMRI. For example, a "time-series analysis pipeline" preprocesses the data removing subject movement related artefacts followed by response estimation with the general linear model. The resulting effect maps are mapped into a standard stereotactic space in readiness for second-level pipelines.

Consumers start by choosing the "group activation mapping" pipeline from the library of pipelines (**Figures 1A,B**). Its pipeline wizard (**Figure 1C**) can configure the pipeline to perform housekeeping tasks to meet the Data Standard, such as inserting the correct `slice_code` into the fMRI 4D data and removing unwanted 3D scans from the start of the data. The Input Table (**Figure 1D**) asks for the fMRI data and the design matrix file. Its Pipeline Schemes are carefully selected to activate parts of the pipeline according to the specified usage of the pipeline.

At the second level, pipelines that offer flexibility in choosing different statistical models present a more difficult challenge for parameter configuration, with many parameters dependent on others. The pipeline can be invalidated if the wrong combination of parameter values is chosen. The corresponding New Pipeline Wizard therefore guides consumers by changing the display according to the model required. At pipeline creation, the available parameter values are screened to remove incompatibilities. The Wizard adds,

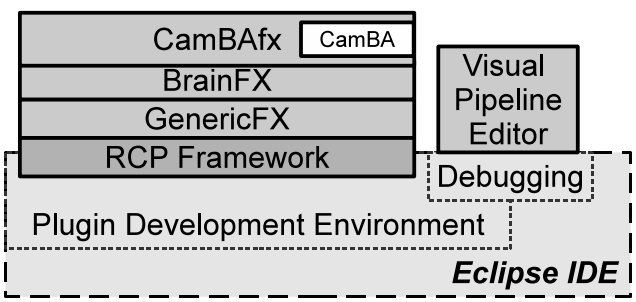


FIGURE 5 | Architecturally, CamBAfx is organized in three layers (Solid boxes): CamBAfx, BrainFX and GenericFX. CamBA is unmodified and completely contained inside a plug-in in the CamBAfx layer. They are all built on top of the Eclipse RCP Framework and developed inside Eclipse Integrated Development Environment's (IDE) Plug-in Development Environment (dashed boxes). CamBAfx extends the Eclipse IDE with a Visual Pipeline Editor to visualize the pipelines during their development and testing.

on request, new ports and connections to the pipeline that represent additional variables. These variables also appear on the Input Table as additional columns. The majority of the Input Table columns are programmed to accept numbers only and where appropriate are further restricted to a small range of values. In effect, the wizard creates different variations of pipelines for the consumers. All second-level pipelines insert an item into the drop-down box below the Input Table that can import results from first-level pipelines.

In general, data generated by one software suite cannot be used by another because the data are stored as a different data type. The most common data type mismatch is 32 bit and 64 bit floating-point data and therefore CamBAfx provides a pipeline to convert data between these formats. Additional information for performing data type conversions from specific software suite is available inside the Help Browser bundled with the core CamBAfx download.

For first-level pipelines, the repetitive entering of data is assisted by a supporting function for automatically reading data into the Input Table from a directory-based data organization. Following download and installation, it adds itself to the drop-down box of the Input Table. Another download adds a menu item to extract statistics from data in predefined regions-of-interest (anatomical or identified by statistical testing). Finally, users can download a menu item that modifies the NifTI header data in batch mode and checks that the modification satisfies the data standard.

IMPLEMENTATION OF FSL TRACK-BASED SPATIAL STATISTICS

To illustrate the flexibility of the CamBAfx approach, a plug-in (TBSSfx) is available which repackages the tract-based spatial statistics (TBSS, Smith et al., 2006) software for diffusion tensor image analysis, available as part of the FSL package. Since TBSS is part of the FSL pipeline, licensing restrictions require a separate download of FSL[7]. In brief, TBSS is a five step process:

(1) Input data is organised into a directory. Pre-processing software relocates input data into a subdirectory.
(2) If there is a target image that defines the stereotactic space of the analysis, copy and rename into the subdirectory. The target image cannot be copied until step 1 is completed.
(3) The analysis software is executed.
(4) A design matrix and a contrast file are created and further analysis takes place.
(5) Call a collection of programs to perform voxel-wise statistical analysis.

There are a number of restrictions on these steps, particularly with regard to the order in which they are conducted. Furthermore, construction of the design matrix is interactive and unconstrained.

TBSSfx is a collection of plug-ins with a plug-in used to host the FSL archive, which the consumers download separately. TBSSfx simplifies data entry and automates the processing ensuring compliance with the restrictions on the processing steps. For example, during pipeline creation, TBSSfx asks the user to name the number of conditions (columns) for the design matrix and to specify the contrast file and then validates this against the format of the design matrix. The contrast file is defined at this stage (and not later in the pipeline) to guarantee that the pipeline created is configured correctly. In the Input Table consumers enter the image data filename in the first row with subsequent columns only accepting numerical data corresponding to the design matrix.

The Run Wizard asks for an output directory, which is cleaned and populated with hard links to the actual data for speed and economy of resources as well as ensuring that the original data are preserved. The design matrix file and contrast file are then created with filenames constructed to maintain the list orders from the Input Table. The processing script manages data processing in a way consistent with the original TBSS process.

DISCUSSION

CamBAfx is an application that presents workflows according the needs of users: designers or consumers. The initial download consists of the basic program only. New functionalities and pipelines can be added post-installation maintaining the installation to a size adequate for local needs. This is made possible by the EEM which manages plug-ins for consumers.

The overall organisation is as Input Table, Pipeline Configuration and Data Processing Engine. The Input Table presents the full view of the datasets, allows users to take notes and fine tune the actual processing of individual datasets. Both Input Table and Parameter Table validate and reject invalid data. These are all designed to improve the chances of successful data processing.

CamBAfx packages neuroinformatics software, without modification, inside plug-ins. Other CamBAfx plug-ins provide the branding, the pipelines and their New Pipeline Wizards as well as supporting functions. Pipelines are organized into directories and each pipeline comes with its own customized wizards.

The back-end's aim is to deliver workflows to the user. It uses the traditional pipeline view of the workflow making modifications straightforward. Facilities like data hierarchy, data standards and pipeline simplification strategies are designed to assist pipeline construction and improve readability. Pipelines are written in XML for human-readability and can be manipulated programmatically.

For developers, CamBAfx supplies a generic set of functions for their pipelines. However, customization of CamBAfx is encouraged by developing supporting facilities. These supporting functions have access to the user interface via Eclipse or CamBAfx extension points.

Organising software in a consistent manner facilitates construction of new pipelines from modules originating from different software packages and is an important design objective for CamBAfx. Analysis software is not merely repackaged, rather consumers and designers can integrate tools to generate custom workflows or undertake optimisation of pipelines through systematic comparison of modules.

Using Eclipse RCP technology means that CamBAfx uses industrial standard architecture reducing development time and ensuring that the underlying technology is constantly updated and improved. Eclipse-based tools can be incorporated easily and CamBAfx can integrate with other Eclipse programs. Eclipse's PDE is a useful aid for developing CamBAfx and its plug-ins. CamBAfx's extensions for Eclipse IDE allows plug-in integration to be debugged and tested using PDE. The source code is organized in a logical and flexible manner to maximize reuse

[7]http://www.fmrib.ox.ac.uk/fsl/

potential. Workflow applications can be developed from BrainFX or GenericFX if CamBA is not needed.

CamBAfx is released under the terms of General Public License (GPL, Free Software Foundation, 2007) and specifically allows designers to integrate their pipelines before shipping. This removes problems associated with consumers having to download pipelines and workflow applications separately and following instructions to integrate them to form the final application.

GenericFX, BrainFX, CamBAfx and TBSSfx can be downloaded from SourceForge.net[8] or NITRC[9].

CONCLUSION

CamBAfx is a workflow application designed to be the user interface that services consumers' needs in the front-end by guiding them throughout the whole process from pipeline creation, through data entry and validation, to data processing. At the back-end, workflow creation and manipulation are made easier by adopting a pipeline model complete with a strategy to understand and use a data standard and data hierarchy as well as facilities to manipulate these pipelines. Out of the box, CamBAfx provides all the generic facilities expected of a workflow application for any pipeline although, uniquely, designers are encouraged to customize CamBAfx for their own pipelines. CamBAfx is built as an Eclipse RCP application and benefits from industrial standard architecture and modern software facilities, such as supporting post-installation modification. EEM makes CamBAfx highly flexible, configurable and extensible. Designers use it to customise CamBAfx for their pipelines, to insert supporting functions and to access the user interface. Moreover, by selecting components from CamBAfx and with the help of Eclipse Branding Mechanism, new workflow applications can be created. The availability of PDE, designed to support Eclipse plug-in developments, improves CamBAfx designers' productivity.

FUTURE WORK

New versions of CamBAfx will use EEM more extensively. Small utility programs are being developed to check that the CamBAfx instance is error free. The current XML pipeline descriptor can contain two or more ways to describe the same data. This will be reduced to one as part of the effort to rationalise the XML descriptors. The new XML will use XML Namespace (Bray et al., 2006) and support XML Schema (Fallside and Walmsley, 2004) validation. Meta-data such as the author's name and email, are managed centrally using the Resource Description Framework (RDF, Beckett, 2004), removing duplication and simplifying updates. RDF also stores the relationship between meta-data.

ACKNOWLEDGEMENTS

This neuroinformatics research was supported by a Human Brain Project grant from the National Institute of Mental Health and the National Institute of Biomedical Imaging and Bioengineering. EB is employed 50% by GlaxoSmithKline and 50% by the University of Cambridge. The work was conducted in the MRC/Wellcome Trust Behavioural & Clinical Neurosciences Institute, Cambridge UK. This project was awarded an IBM Eclipse Innovation Award 2003.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/neuroinformatics/paper/10.3389/neuro.11/027.2009/

[8]http://sourceforge.net/projects/camba/

[9]http://www.nitrc.org/projects/camba/

REFERENCES

Beckett, D. (ed.) (2004). RDF/XML Syntax Specification (Revised). W3C Recommendation 10 February 2004. Available at: http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/. RDF Interest Group, http://www.w3.org/RDF/.

Bolour, A. (2003). Notes on the Eclipse Plug-in Architecture (Eclipse Corner Article). Available at: http://www.eclipse.org/articles/Article-Plug-in-architecture/plugin_architecture.html.

Bray, T., Hollander, D., Layman, A., and Tobin, R. (eds) (2006). Namespaces in XML 1.0, 2nd Edn. W3C Recommendation 16 August 2006. Available at: http://www.w3.org/TR/2006/RECxml-names-20060816/. XML Core Working Group, http://www.w3.org/XML/Core/.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F. (2008). Extensible Markup Language (XML) 1.0, 5th Edn. W3C Recommendation 26 November 2008. Available at: http://www.w3.org/TR/REC-xml/. XML Core Working Group, http://www.w3.org/XML/Core/.

Chamberlain, S. R., Hampshire, A., Müller, U., Rubia, K., Del Campo, N., Craig, K., Regenthal, R., Suckling, J., Roiser, J. P., Grant, J. E., Bullmore, E. T., Robbins, T. W., and Sahakian, B. J. (2009). Atomoxetine modulates right inferior frontal activation during inhibitory control: a pharmacological functional magnetic resonance imaging study. *Biol. Psychiatry* 65, 550–555.

Chamberlain, S. R., Menzies, L., Hampshire, A., Suckling, J., Fineberg, N. A., del Campo, N., Aitken, M., Craig, K., Owen, A. M., Bullmore, E. T., Robbins, T. W., and Sahakian, B. J. (2008). Orbitofrontal dysfunction in patients with obsessive-compulsive disorder and their unaffected relatives. *Science* 321, 421–422.

Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-B., and Benali, H. (2001). BrainVISA: software platform for visualization and analysis of multi-modality brain data. *Neuroimage* 13, S98. Available at: http://www.brainvisa.info/.

Cox, R. W., Ashbourner, J., Breman, H., Fissell, K., Haselgrove, C., Holmes, C. J., Lancaster, J. L., Rex, D. E., Smith, S. M., Woodward, J. B., and Strother, S. C. (2004). A (Sort of) New Image Data Format Standard: NIfTI-1. 10th Annual Meeting of the Organization for Human Brain Mapping (OHBM 2004), Budapest, Hungary, June 13–17. Available at: http://nifti.nimh.nih.gov/nifti-1/documentation/hbm_nifti_2004.pdf.

Eidsness, A., and Rapicault, P. (2004). Branding Your Application (Eclipse Corner Article). Available at: http://www.eclipse.org/articles/Article-Branding/branding-your-application.html.

Fallside, D. C., and Walmsley, P. (eds) (2004). XML Schema Part 0, Primer, 2nd Edn. W3C Recommendation 28 October 2004. Available at: http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/. XML Schema Working Group, http://www.w3.org/XML/Schema/.

Fissell, K., Tseytlin, E., Cunningham, D., Iyer, K., Carter, C. S., Schneider, W., and Cohen, J. D. (2003). Fiswidgets: a graphical computing environment for neuroimaging analysis. *Neuroinformatics* 1, 111–125. Available at: http://grommit.lrdc.pitt.edu/.

Free Software Foundation (2007). GNU General Public License. Available at: http://www.gnu.org/licenses/gpl.html.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. B., Frith, C., and Frackowia, R. S. J. (1995). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 1995, 189–210. Available at: http://www.fil.ion.ucl.ac.uk/spm/.

Habets, P., Krabbendam, L., Hofman, P., Suckling, J., Oderwald, F., Bullmore, E., Woodruff, P., Van Os, J., and Marcelis, M. (2008). Cognitive performance and grey matter density in psychosis: functional relevance of a structural endophenotype. *Neuropsychobiology* 58, 128–137.

Hudson, R. (2004). The Graphical Editing Framework. EclipseCon 2–5 February 2004, Anaheim, CA, USA. Available at: http://www.eclipsecon.org/2004/EclipseCon_2004_TechnicalTrackPresentations/47_Hudson.pdf.

International Business Machines (2006). Eclipse Platform Technical Overview (Eclipse White Paper). Available at: http://www.eclipse.org/articles/Whitepaper-Platform-3.1/eclipse-platform-whitepaper.pdf.

McAffer, J., and Lemieux, J.-M. (2005). Eclipse Rich Client Platform – Designing, Coding and Packaging Java Applications (Addison-Wesley Professional). RCP website, Available at: http://wiki.eclipse.org/index.php/Rich_Client_Platform.

Melhem, W., and Glozic, D. (2003). PDE Does Plug-ins (Eclipse Corner Article). Available at: http://www.eclipse.org/articles/Article-PDE-does-plugins/PDE-intro.html. PDE project website at http://www.eclipse.org/pde/.

Menzies, L., Williams, G., Chamberlain, S., Ooi, C., Fineberg, N., Suckling, J., Sahakian, B., Robbins, T., and Bullmore, E. (2008). White matter abnormalities in patients with obsessive-compulsive disorder and their first-degree relatives. *Am. J. Psychiatry* 165, 1308–1315.

Rex, D. E., Ma, J. Q., and Toga, A. W. (2000). The LONI pipeline processing environment. *Neuroimage* 19, 1033–1048. Available at: http://pipeline.loni.ucla.edu/.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., and Behrens, T. E. J. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* 31, 1487–1505. Available at: http://www.fmrib.ox.ac.uk/fsl/tbss/index.html. Implementation described her is based on the version in 2007: http://web.archive.org/web/20070703045209/http://www.fmrib.ox.ac.uk/fsl/tbss/index.html.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, 208–219. Available at: http://www.fmrib.ox.ac.uk/fsl/.

Suckling, J., and Bullmore, E. (2004). Permutation tests for factorially designed neuroimaging experiments. *Hum. Brain Mapp.* 22, 193–205.

Suckling, M., Davis, M., Ooi, C., Wink, A. M., Fadili, J., Salvador, R., Welchew, D., Sendur, L., Maxim, V., and Bullmore, E. (2006). Permutation testing of orthogonal, factorial effects in a language processing experiment using fMRI. *Hum. Brain Mapp.* 27, 425–433.

Wink, A. M., Bullmore, E., Barnes, A., Bernard, F., and Suckling, J. (2008). Monofractal and multifractal dynamics of low frequency endogenous brain oscillations in functional MRI. *Hum. Brain Mapp.* 29, 791–801.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 March 2009; paper pending published: 26 June 2009; accepted: 02 August 2009; published online: 28 August 2009.

*Citation: Ooi C, Bullmore ET, Wink A-M, Sendur L, Barnes A, Achard S, Aspden J, Abbott S, Yue S, Kitzbichler M, Meunier D, Maxim V, Salvador R, Henty J, Tait R, Subramaniam N and Suckling J (2009) CamBAfx: workflow design, implementation and application for neuroimaging. Front. Neuroinform. **3**:27. doi: 10.3389/neuro.11.027.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 07 September 2009
doi: 10.3389/neuro.11.030.2009

Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid

David B. Keator[1]*, Dingying Wei[1], Syam Gadde[2], Jeremy Bockholt[3], Jeffrey S. Grethe[4], Daniel Marcus[5], Nicole Aucoin[6] and Ibrahim B. Ozyurt[7]

[1] *Psychiatry and Human Behavior, College of Medicine, University of California, Irvine, CA, USA*
[2] *Brain Imaging and Analysis Center, Duke University, Durham, NC, USA*
[3] *MIND Research Network, Albuquerque, NM, USA*
[4] *Center for Research on Biological Systems, University of California San Diego, San Diego, CA, USA*
[5] *Neuroinformatics Research Group, Washington University, Saint Louis, MO, USA*
[6] *Brigham and Women's Hospital, Harvard University, Boston, MA, USA*
[7] *Department of Psychiatry, Duke University, Durham, NC, USA*

Edited by:
John Van Horn,
University of California, USA

Reviewed by:
Michael Wilde, University of Chicago and Argonne National Laboratory, USA
Rico Magsipoc,
University of California, USA
John Van Horn,
University of California, USA

****Correspondence:***
David B. Keator, Psychiatry and Human Behavior, Brain Imaging Center, University of California, Irvine, Irvine Hall, Room 163, Irvine, CA 92697, USA.
e-mail: dbkeator@uci.edu

Organizing and annotating biomedical data in structured ways has gained much interest and focus in the last 30 years. Driven by decreases in digital storage costs and advances in genetics sequencing, imaging, electronic data collection, and microarray technologies, data is being collected at an ever increasing rate. The need to store and exchange data in meaningful ways in support of data analysis, hypothesis testing and future collaborative use is pervasive. Because trans-disciplinary projects rely on effective use of data from many domains, there is a genuine interest in informatics community on how best to store and combine this data while maintaining a high level of data quality and documentation. The difficulties in sharing and combining raw data become amplified after post-processing and/or data analysis in which the new dataset of interest is a function of the original data and may have been collected by multiple collaborating sites. Simple meta-data, documenting which subject and version of data were used for a particular analysis, becomes complicated by the heterogeneity of the collecting sites yet is critically important to the interpretation and reuse of derived results. This manuscript will present a case study of using the XML-Based Clinical Experiment Data Exchange (XCEDE) schema and the Human Imaging Database (HID) in the Biomedical Informatics Research Network's (BIRN) distributed environment to document and exchange derived data. The discussion includes an overview of the data structures used in both the XML and the database representations, insight into the design considerations, and the extensibility of the design to support additional analysis streams.

Keywords: MRI, medical imaging, analysis, database, XML, XCEDE, HID, BIRN

INTRODUCTION

The biomedical science community has seen increased numbers of multi-site consortia driven in part by advances in speed and robustness of internet technologies, the demand for cross-scale data to understand fundamental disease processes, the need for experts from diverse domains to integrate and interpret the data, and the movement of science in general toward freely available information (Arzberger and Finholt, 2002). The Science of Collaboratories website[1] lists 213 collaboratories since 1993. These consortia face increased challenges in managing, interpreting, and sharing data without informatics methods to clearly document necessary metadata at both the time of data collection and subsequent data processing and analysis.(Olsen et al., 2008; Paton, 2008) The difficulties in sharing and combining raw data become amplified after post-processing and/or data analysis in which the new dataset of interest is a function of the original data and may have been collected by multiple collaborating sites. Simple metadata, documenting which subject and version of data were used for a particular analysis, becomes complicated by the heterogeneity of the collecting sites yet is critically important to the interpretation and reuse of derived results. Numerous recent publications have discussed the benefits of documenting the origin and steps by which data were collected and derived (Foster et al., 2003; Simmhan et al., 2005; Zhao, et al., 2006; MacKenzie-Graham et al., 2008; Moreau et al. 2008). Provenance, as defined by the Oxford English Dictionary, is "the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners" (Freire et al., 2008). MacKenzie-Graham et al. (2008) make a distinction between data provenance and processing provenance where the former refers to metadata describing how the original data was collected and the later referring to the processing original data undergoes after the initial collection. Both types of metadata are crucially important for subsequent use of the data by a single laboratory and the scientific community. In multi-site, distributed, collaboratories where information is dynamic in nature and not centrally managed, robust, scalable metadata management tools are essential (Moreau et al., 2008).

[1]www.scienceofcollaboratories.org

The Biomedical Informatics Research Network (BIRN)[2] is a large multi-site consortia of individual test beds coalesced around a shared set of resources, developing standards, methods, and processing tools in a distributed, grid-enabled environment (Grethe et al., 2005; Keator et al., 2006). The BIRN enables scientists across disparate domains to securely and transparently share data and tools. The Function BIRN test bed (Keator et al., 2006) brings together investigators developing data sharing standards, instrument calibration methods in the context of functional MRI (fMRI), novel statistical models, and advanced clinical/cognitive paradigms necessary to study the neural substrates of schizophrenia in a collaborative setting. Since its inception in 2002, FBIRN has prospectively collected over 400 fMRI human datasets collected during the protocol design and execution of four separate studies and thousands of agar phantom calibration datasets across the 11 participating sites. The datasets generally consisted of a minimum of five functional acquisitions and at least a T1-weighted structural acquisition. Details about the publically available data can be found at http://nbirn.org/bdr.

Beyond prospective data collection, the FBIRN neuroinformatics working group, in collaboration with other BIRN test bed informatics groups, has developed data structures and software to dynamically track and document data acquired and analyzed as part of human imaging studies. The suite of tools forms a cooperative system for managing and documenting acquired and derived data entitled the FBIRN Federated Informatics Research Environment (FIRE)[3]. Data management in the federated environment of both the original and derived data is supported through three core components: the Human Imaging Database (HID)[4] for distributed/federated relational database support and web-enabled graphical user interface, the XML-Based Clinical Experiment Data Exchange (XCEDE2)[5] schema used to define valid XML documents for structured data/metadata storage and exchange, and data publication scripts to organize and transfer data to the distributed file system and send appropriate uniform resource locator (URL) links to the HID database. In this manuscript we introduce tools from the BIRN software suite used for documenting multi-site functional and structural neuroimaging analyses in a federated database and distributed data handling environment. The discussion centers around two data processing pipelines, one designed for multi-site preprocessing of fMRI data and the other, a structural analysis of schizophrenia in humans. Our intention is to provide the informatics community with insights into the data structures used and our view of the extensibility of this system.

MATERIAL AND METHODS

FBIRN NEUROIMAGING DATA MANAGEMENT AND WORKFLOWS OVERVIEW

Scientific data management systems generally consist of at least a few core components: a back-end database for permanent, structured, data storage and efficient query, a front-end graphical user interface for client interaction, and an import/export mechanism to get data into and out of the database and share with collaborators (Keator et al., 2008). These systems can exist entirely at a single site or be distributed geographically. FBIRN operates in a completely distributed environment. The suite of tools developed by FBIRN form the FIRE, providing management support of clinical, behavioral, and imaging data in a decentralized way using federated databases and a distributed file system (**Figure 1**). Each site maintains its own HID database back-end and graphical user interface (Ozyurt et al., 2004a,b, 2006; Keator et al., 2006; Keator, 2009). The HID is

[2]www.nbirn.net

[3]www.nitrc.org/projects/fbirn/

[4]www.nitrc.org/projects/hid

[5]www.xcede.org

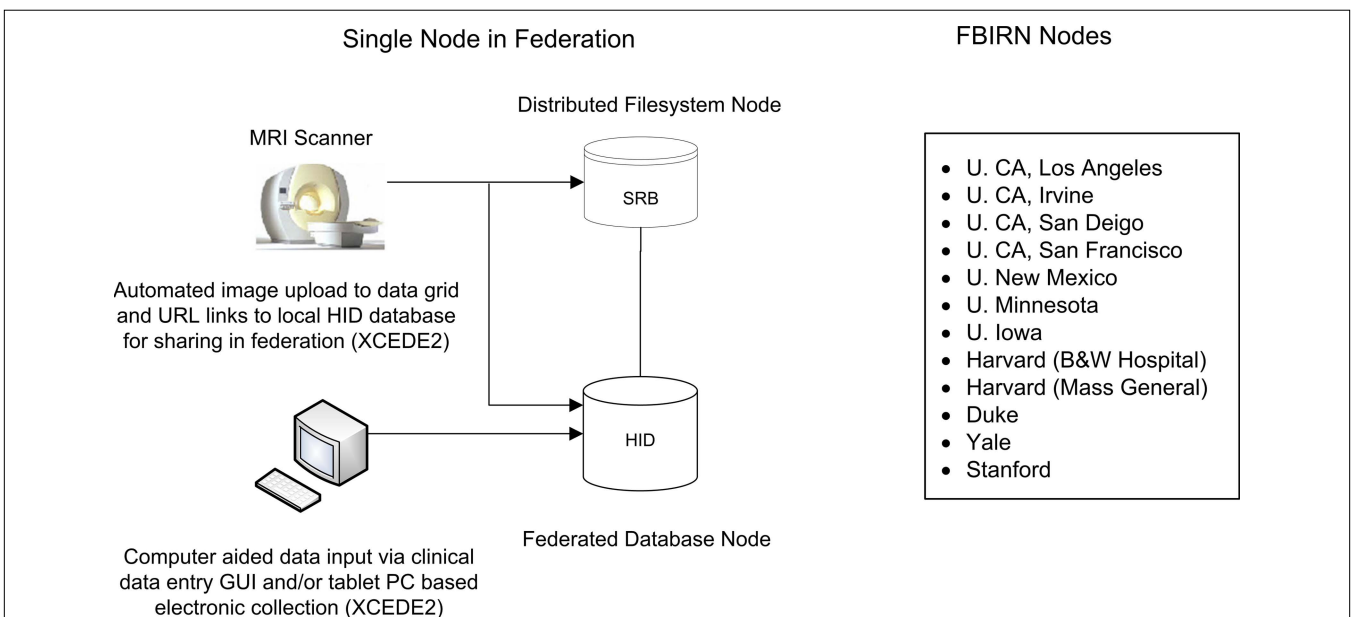


FIGURE 1 | Schematic of original (raw) data entry and file registration process at a node in the federation and FBIRN node locations.

an open-source extensible database schema designed to support multi-site, federated, installations and inclusion of new data types without changing the core table space. The graphical user interface is a three-tier J2EE application supporting data input, single-site and multi-site query, data export, and core system administration tasks. More detailed information about HID can be found in the references and the software is available through the NITRC website[6]. Currently, within FBIRN, there are 11 federated installations managing 790 imaging visits and 4239 clinical assessments as collected across four prospective FBIRN studies and retrospective data contributed by the Brainscape repository of Washington University, St. Louis[7]. Clinical assessments collected are those common in studies of Schizophrenia such as SCID, Beckman Depression Inventory (BDI), North American Adult Reading Test (NAART), InterSePT scale, and many others. Details of publically available data can be found at http://nbirn.org/bdr. Data files that are part of an imaging study are published to the Storage Resource Broker (SRB) distributed file system and cross-linked in the database using a URL string (Rajasekar et al., 2003). The data publication process involves data reorganization into a standardized directory hierarchy, format conversions, and the creation of XCEDE2 XML (eXtensible Markup Language)[8] files containing minimal metadata about the experiment stored with the imaging files on the SRB. This process is facilitated by data publication scripts. The scripts use an XML formatted template which a site can configure using an XML editor or a provided GUI. The upload template consists of metadata describing the imaging series, visit, and project information. When available the information is automatically extracted from DICOM image headers. Information that is not available in the DICOM headers is input manually. The data publication scripts include schematron validation definitions which are prepared during study design to validate the data publication XML templates. Once the templates are created they can be reused with minor modifications to visit dates and subject IDs using the GUI provided with the publication scripts. The bulk of metadata describing the subject visit is stored in the database. Additional details about data provenance and management of the original collected data can be found in publications by Ozyurt et al. (2006) and Keator et al. (2006).

Once the data has been published into the federated system, it is available for processing. The FBIRN has developed quality assurance and image processing utilities optimized to work with data from the federated system. Data analysis and/or post-processing workflows currently instantiated in FBIRN share a few common steps. First, the datasets are located in the federation, either by browsing the low-level distributed file system or interacting with the HID graphical user interfaces to query and filter data collected in the federation. Once datasets of interest are identified, they are downloaded to the local system for computation (**Figure 2**). The downloaded datasets contain both imaging data files and the XML metadata files stored with the dataset. Additional metadata exports from the HID database are also available during the downloading process if one is using the graphical user interface. Once data is downloaded, any number of analysis algorithms could be run and

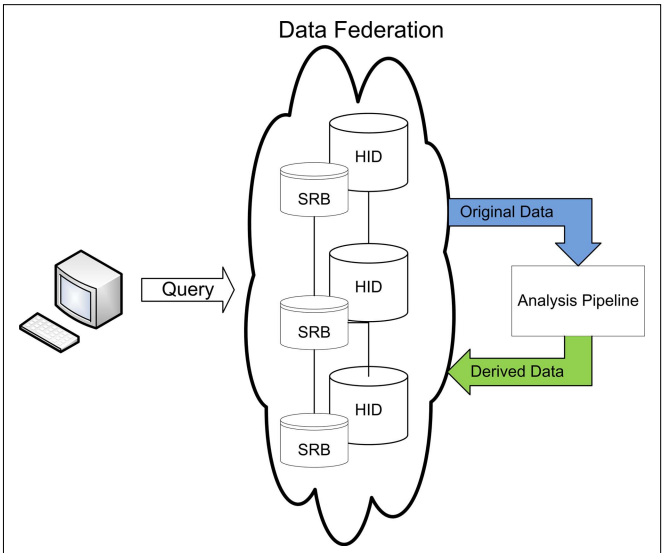


FIGURE 2 | FBIRN data management discovery and analysis workflow.

a new derived dataset created. If an investigator feels the derived dataset is of sufficient technical quality and scientific interest to others in the collaboratory, it should be published to the federation with sufficient processing provenance and searchable metadata such that others can effectively interpret and reuse the derived data. This overall process of documenting steps in an analysis pipeline, representing the provenance in a consistent and well documented way, and providing a means of querying derived data which references original subjects collected at geographically distributed sites in a robust and extensible manner were the motivations driving the informatics components presented here.

CASE STUDIES

Two analysis workflows will be referred to throughout the following sections, giving substantive context to the abstract informatics structures discussed. Each workflow has slightly different requirements for processing provenance and metadata storage. Together the case studies illustrate the robustness of the informatics structures.

Structural MRI analysis workflow

This workflow consists of a multi-site structural MRI analysis of schizophrenia. The imaging data consisted of 3D T1-weighted MRI images collected across consortium sites. The original images were shared using the data management components described in Section "FBIRN Neuroimaging Data Management and Workflows Overview." The structural morphometric (StructMorph) analysis was performed across two participating sites. Data were analyzed with the FreeSurfer software[9] using a single program "autorecon-all". The "autorecon-all" script calculates cortical and sub-cortical thickness statistics in two stages: a volumetric processing stage which includes noise correction, volumetric registration, and white matter segmentation, and a surface processing stage for cortical parcellation and thickness measurements. The "autorecon-all" is

[6]http://www.nitrc.org/projects/hid/

[7]www.brainscape.org

[8]www.w3.org/XML

[9]http://surfer.nmr.mgh.harvard.edu

a black box processing script. Provenance documentation about which FreeSurfer binaries are called by "autorecon-all" were not provided with the analysis. It has a version number and compilation date that uniquely identifies the script but the details about what other modules it calls during the course of execution is hidden from the user. Cortical and sub-cortical thickness estimates from the structural processing pipeline were chosen by study investigators as metadata to make available for query in the database federation. All other images, intermediate files, and program specific outputs were made available on the distributed file system. Cortical thickness measurements are extracted from output files using the script "fsstats2xcede.pl". The overall workflow is shown in **Figure 3**. This case study is used to illustrate the process of extracting relevant analysis specific metadata, encapsulating it in XML, loading it into database tables, and making it available for query in the federated data management system in a generic way.

fMRI data preprocessing workflow

The fMRI data preprocessing (PreProc) workflow consists of a multi-level pipeline with numerous intermediate derived results combined with original data inputs at various points in the workflow (**Figure 4**). The complex nature of the workflow makes it an ideal test case for the informatics structures. This workflow was designed to provide an automated and consistent pre-processing pipeline for FBIRN studies. Preprocessing in fMRI is a general term describing any processing done after image reconstruction prior to statistical analysis of brain activation (Strother, 2006). The PreProc pipeline consists of motion correction, slice timing correction, magnetic field inhomogeneity correction (B_0), and spatial smoothing. For additional information on the FBIRN imaging processing pipeline used for the PreProc analysis, please visit www.nitrc.org/projects/fips/. For this workflow, investigators were most interested in documenting the processing provenance. Unlike the StructMorph analysis discussed in Section "Structural MRI Analysis Workflow" in which the processing is treated as a single black box script, this workflow has many separate programs put together in a specific order. Changing the order and/or any of the parameter settings potentially alter the derived results. Investigators were most interested in carefully documenting the ordering of steps and the parameters used. Proper documentation of the PreProc workflow enables its use in higher order analyses without duplicating work. As the data federation grows, original data may be processed numerous times with slightly different steps or with different parameter settings and made available through the data management systems. It is therefore critically important to document the workflow as completely as possible given limited time and resources of investigators to enable maximum derived data reusability.

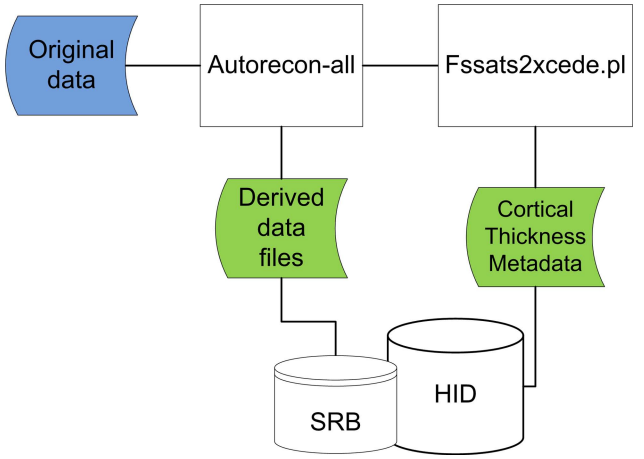


FIGURE 3 | Original data files (blue) downloaded from data federation are processed using autorecon-all and cortical thickness data extracted (green). Resulting data files are loaded back into data federation.

DERIVED DATA EXCHANGE SCHEMA

The XML-Based Clinical Experiment Data Exchange (XCEDE2)[10] schema was designed for documenting research and clinical studies (Keator et al., 2006). The schema defines components and constraints on those components required to form a valid XCEDE2 compliant XML document. Initially the focus of XCEDE2 was on human imaging studies but the schema contains many generic and extensible structures useful for a wider range of scientific domains. Development of the schema was a joint effort within BIRN and is the exchange medium for many database web services currently

[10]www.xcede.org

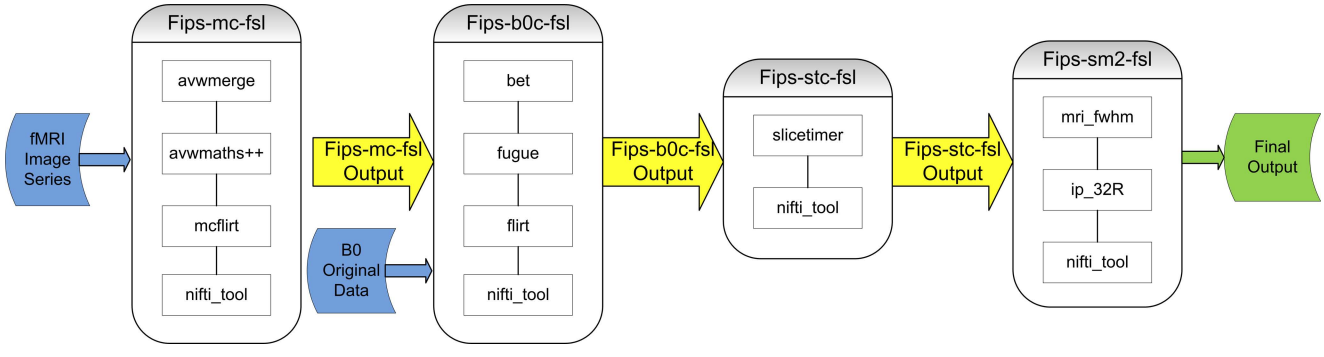


FIGURE 4 | fMRI PreProc data preprocessing workflow. Original data inputs colored blue, intermediate derived data colored yellow, final derived data output colored green. Workflow transform modules Fips-mc-fsl, Fips-b0c-fsl, Fips-stc-fsl, and Fips-sm2-fsl contains variable numbers of sub-transform steps to produced intermediate and final derived results.

in use. The schema is flexible, providing mechanisms for linking to external output files and for storing analysis data directly in the XML document. XCEDE2 documents can be split into sub-documents and linked together using constructs of the schema. The data analysis portion of the XCEDE2 schema is the most relevant to the case studies and will be presented in more detail. For complete documentation of the schema readers are encouraged to visit the website. The analysis component of the XCEDE2 schema was designed as a generic container used for documenting results of analyses. An analysis in this context is composed of the "inputs" (i.e., the files and parameters used in an analysis or processing of data), a list of the application(s) or method(s) used in the analysis (provenance), and the resultant data (i.e., values and output files) (**Figure 5**).

The format of the <*input*> and <*output*> components (**Figure 6A**) are essentially identical. Using the ID attributes <*dataID*> and <*analysisID*>, they serve as pointers to other portions of the XCEDE dataset (in the same XML document or another XCEDE2-compliant XML file) that more fully describe the analysis or data consumed or written by this processing step.

The <*measurementGroup*> component is used to store information and data related to the outcome of analyses (**Figure 6B**). Each measurement group contains observations on an entity. Entities are used to give meaning to the measurements being stored. The entity element can reference any number of terminology sources and is composed of multiple nomenclature/termID pairs. The observation element of a measurementGroup contains the actual measurement values for the particular entity along with attributes defining the data type and units of the measurement. An example of the <*measurementGroup*>entry for the StructMorph analysis is shown in **Figure 7**. The measurements for this analysis are related to curvature and thickness of particular anatomical parcellations of the cortex. The <*measurementGroup*> component is extensible in that any number of self-describing observations can be

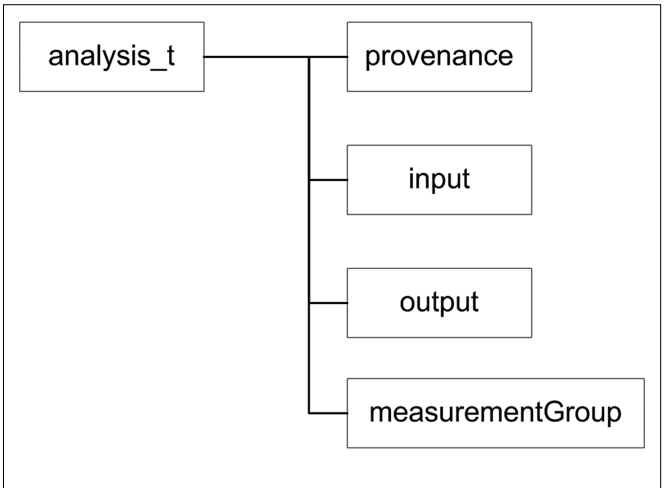


FIGURE 5 | Base <*analysis*> component.

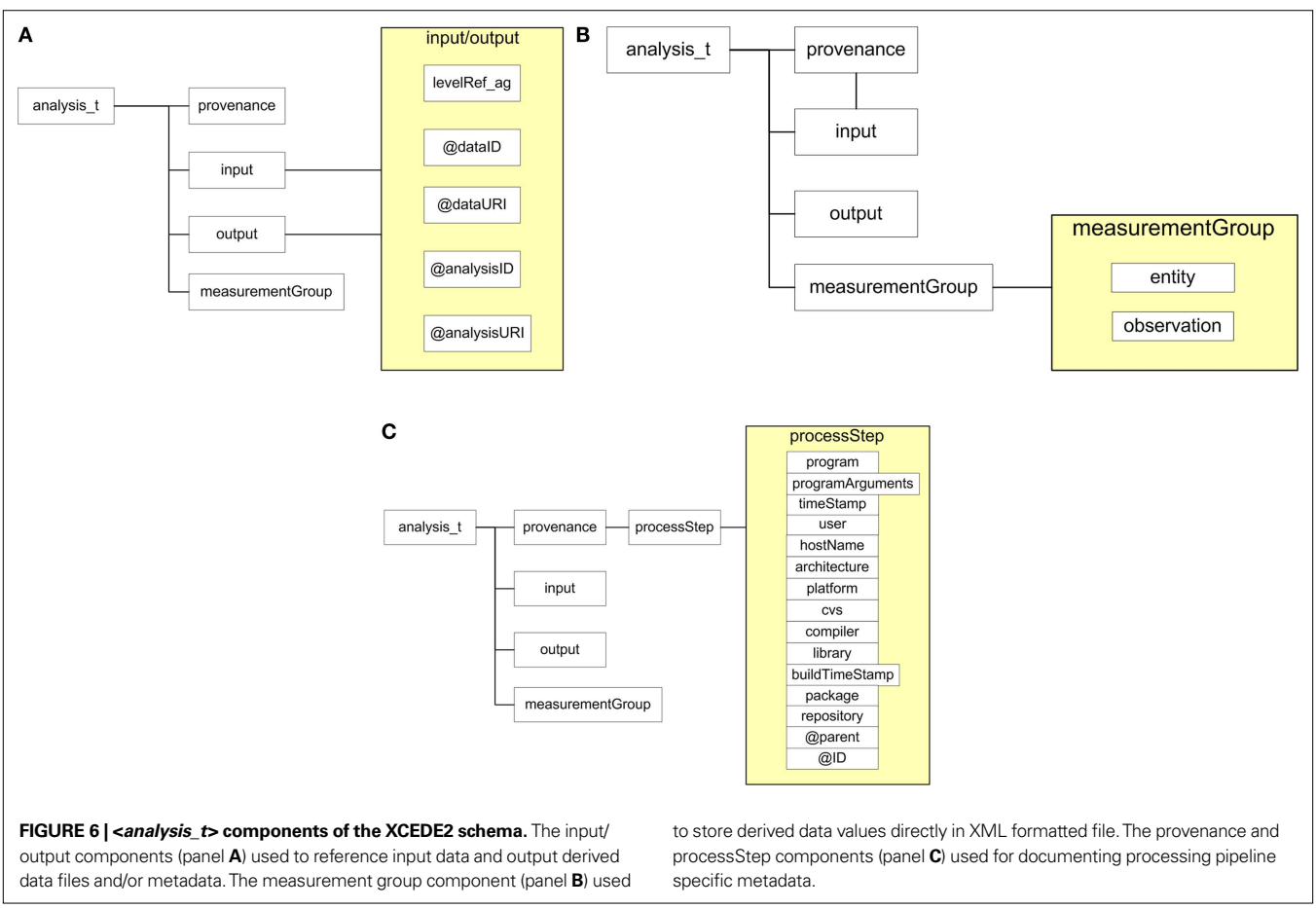


FIGURE 6 | <*analysis_t*> components of the XCEDE2 schema. The input/output components (panel **A**) used to reference input data and output derived data files and/or metadata. The measurement group component (panel **B**) used to store derived data values directly in XML formatted file. The provenance and processStep components (panel **C**) used for documenting processing pipeline specific metadata.

```
<measurementGroup>
  <entity xsi:type="anatomicalEntity_t" laterality="left">
    <label nomenclature="lh.aparc.annot" termID="unknown">unknown</label>
  </entity>
  <observation name="NumVert" type="integer">15448</observation>
  <observation name="SurfArea" type="float" units="mm^2">9998</observation>
  <observation name="GrayVol" type="float" units="mm^3">17416</observation>
  <observation name="ThickAvg" type="float" units="mm">1.671</observation>
  <observation name="ThickStd" type="float" units="mm">1.628</observation>
  <observation name="MeanCurv" type="float" units="mm^-1">0.092</observation>
  <observation name="GausCurv" type="float" units="mm^-2">0.026</observation>
  <observation name="FoldInd" type="float">131.946</observation>
</measurementGroup>
```

FIGURE 7 | XCEDE2 XML entry for thickness and curvature derived data. Entity tags document terminology source "rh.aparc.annot" and term "caudalmiddlefrontal" which is the native term and source within FreeSurfer analysis software.

grouped together to record a derived data output complete with entity information. The nomenclature used in this example is the FreeSurfer native terminology thus giving meaning to an otherwise arbitrary anatomical location identifier. In the StructMorph analysis, there are many <measurementGroup> entries, one for each anatomical region analyzed. Hemispheric analyses are physically separated into different XCEDE2 files but could alternatively be contained within one file. The decision to separate results into multiple XCEDE2 files was to facilitate granularity of analysis summary downloads.

In thinking about how users would interact with the derived results, there were two methods that were most desirable to support in FBIRN. The first method is a direct query of the database, filtering on cortical thickness and/or curvature measurements by anatomical region for the StructMorph analysis shown in **Figure 7**. To facilitate this use case, the parcellation results need to be loaded into the data management system. Web services for the HID database were developed in support of derived data loading using the XCEDE2 format. Effectively any derived result that can be represented using XCEDE2's <*analysis*> component can be directly imported into the HID database without table space changes (see Section "Derived Data Database Schema" for database design). The intermediate representation of derived results in the form of an XCEDE2 file is important for downstream processing tools, data management systems, and structured data exchange. Tools that might otherwise not have access to a processing pipeline's native output file formats can be written to parse XCEDE2 documents and obtain an agnostic view of derived results. For those pipeline stages that don't directly export XCEDE2 data, it is a simple matter to create wrapper scripts that extract relevant summary data into XCEDE2 documents. The second method of derived data use in the FBIRN federation is downloading the entire analysis output and exploring the output within the analysis tool or pipeline itself. For this method of interaction, a user may just need to filter on some aspect of the processing provenance. For example, a user might query on all analyses performed using named pipeline PreProc, version 1.0. Additionally, the user might want to find all analyses that used a particular dataset as input. To support these use cases, structured documentation of original data and processing provenance is needed.

The <*provenance*> element of the <*measurementGroup*> component provides a mechanism for documenting processing provenance in an XCEDE2 compliant XML document (**Figure 6C**). A typical <*provenance*> entry consists of many <*processStep*> blocks used to store metadata about the analysis pipeline itself. The schema provides elements for documenting program arguments, compiler and library information, platform and architecture, time stamping, and user identification. Typically in standalone analysis packages and in arbitrary processing pipelines constructed from multiple standalone applications, rich metadata is difficult to capture. Unless there has been concerted effort by software developers to provide provenance with analysis execution, it is up to the user to maintain accurate records. Workflow environments such as the LONI pipeline[11] and Fiswidgets[12] augment information provided by tool developers with enhanced pipeline metadata easing the burden of provenance documentation (Fissell et al., 2003; MacKenzie-Graham et al., 2008). The XCEDE2 schema provides flexibility in storing pipeline provenance alongside derived data. **Figure 8** shows examples of processing provenance collected for the StructMorph and PreProc use cases. The complete provenance records for the analyses are quite long so selected segments have been extracted. To provide the ability to reconstruct arbitrarily complex pipelines, the data provenance schema in XCEDE2 supports multiple forks, merges, and/or parallel analysis streams. Currently, the XCEDE2 provenance <*processStep*> components have attributes "id" and "parent" that together are used to document complex tree structured processing pipelines. The schema does not put any restrictions on "parent" attributes allowing maximum flexibility at some expense of clarity. In the StructMorph use case, provenance wrapping scripts were written by FBIRN developers working directly with FreeSurfer software developers. In the PreProc use case, provenance was compiled by FBIRN developers using information available from only the standalone tools and linked together using XCEDE2 constructs consistent with the defined the pipeline. The <*provenance*> components in an XCEDE2 compliant export of an analysis are used directly by the HID database web services to store the processing

[11]http://pipeline.loni.ucla.edu

[12]http://grommit.lrdc.pitt.edu/fiswidgets/

```
    <processStep>
      <programName>fips-mc-fsl</programName>
      <programArgument>/data/fBIRN/src/human/fips/ppc-global/fBIRNPhaseII/M_mc-fsl.ppc
        000347539107/scanVisit__0003__0002/MRI__0001/AudOdd1/Native/Original__0001/NIFTI</programArgument>
      <version>v 1.9 2007/01/12 23:41:21 (PPC: v 1.1 2006/11/09 19:09:39)</version>
      <timeStamp>01/14/07- 0-25-21-PST</timeStamp>
      <cvs>$Id: fips-mc-fsl,v 1.9 2007/01/12 23:41:21 rnotestine Exp $</cvs>
      <user>randy</user>
      <machine>x86_64</machine>
      <hostName>intuition</hostName>
      <platform>GNU/Linux</platform>
      <platformVersion>2.6.9-42.0.3.ELsmp</platformVersion>
    </processStep>
     <processStep>
       <programName>avwmerge called from fips-mc-fsl</programName>
       <programArgument>
/raids/raid7A/fBIRN/LOCAL_SRB/fBIRNPhaseII__0010/Data/000347539107/scanVisit__0003__0002/MRI__0001/AudOdd1/Analysis/O
riginal__0001/FIPS_MBTS_preprocs__0008__0001/M_mc-fsl.preproc/func.nii.gz
         f0001.img f0002.img f0003.img f0004.img f0005.img f0006.img f0007.img f0008.img f0009.img ...
         f0136.img f0137.img f0138.img f0139.img f0140.img</programArgument>
       <version>FSL Release 3.3 (64-bit) 2006/04/07</version>
       <timeStamp>01/14/07- 0-25-22-PST</timeStamp>
       <cvs>unknown</cvs>
       <user>randy</user>
       <machine>x86_64</machine>
       <hostName>intuition</hostName>
       <platform>GNU/Linux</platform>
       <platformVersion>2.6.9-42.0.3.ELsmp</platformVersion>
     </processStep>
```

```
     <provenance>
       <processStep>
         <program>/Applications/freesurfer/bin//recon-all</program>
         <programArguments>-autorecon-all -s 001029291693_visit1</programArguments>
         <timeStamp>2006-09-08T17:12:16-07:00</timeStamp>
         <user>jsegall</user>
         <hostName>newton.mind.unm.edu</hostName>
         <architecture>powerpc</architecture>
         <platform>Darwin Kernel Version 8.7.0: Fri May 26 15:20:53 PDT 2006;
 root:xnu-792.6.76.obj~1/RELEASE_PPC</platform>
         <cvs>$Id: recon-all,v 1.17.2.4 2006/05/02 18:28:49 nicks Exp $</cvs>
         <package>FreeSurfer</package>
       </processStep>
       <processStep>
         <program>fsstats2xcede.pl</program>
         <programArguments>--projectid="fBIRNPhaseII__0010" --subjectid="001029291693" --visitid="scanVisit__0002"
 --studyid="MRI__0001" --episodeid="t1" aseg.stats</programArguments>
         <timeStamp>2008-08-08T16:07:00-05:00</timeStamp>
         <user>gadde</user>
         <hostName>varese.dhe.duke.edu</hostName>
         <architecture>i386</architecture>
         <platform>Linux</platform>
       </processStep>
```

FIGURE 8 | Example XCEDE2 provenance blocks from PreProc (top) analysis and StructMorph (bottom) analyses.

pipeline description. The <*measurementGroup*> data is also parsed by the web service layer and loaded into the data management system (see Section "Derived Data Database Schema").

DERIVED DATA DATABASE SCHEMA

Cataloging derived data and metadata in the HID data management system is a vital step in making the analytic results available to BIRN collaborators and ultimately the wider scientific community. Because the BIRN infrastructure is inherently distributed and federated in nature, simple changes to database schema at one site becomes difficult and time consuming in the federation. Therefore, an important requirement for the database schema is a stable set of generic tables capable of storing processing pipeline provenance, interesting analytic results, and metadata about analyses complete with ontology and terminology source references. The table space should not change when presented with new derived data types and/or pipeline definitions. The StructMorph and PreProc analyses are interesting cases to test the stability of the data management schema. The StructMorph use case tests the capability of storing derived data values directly in the database and the automated query interface creation by the web application. The PreProc use case tests the table space for documenting multi-layered processing pipeline provenance. As shown in **Figure 4**, the processing pipeline is complex with transforms composed of sub-transforms hierarchically, with inputs and outputs interleaved along with multiple intermediate states.

The database schema for documenting processing pipeline definitions consists of four core tables: *nc_analysis*, *nc_analysisFlow*, *nc_analysisComponent*, and *nc_transformation* (**Figure 9**). Defining a processing pipeline is differentiated from any particular instantiation of that processing pipeline on actual data. The *nc_transformation* table serves as a generic bag of processes where each entry contains a reference name, reference version, package name, package version, and ontological information. The reference name and version are

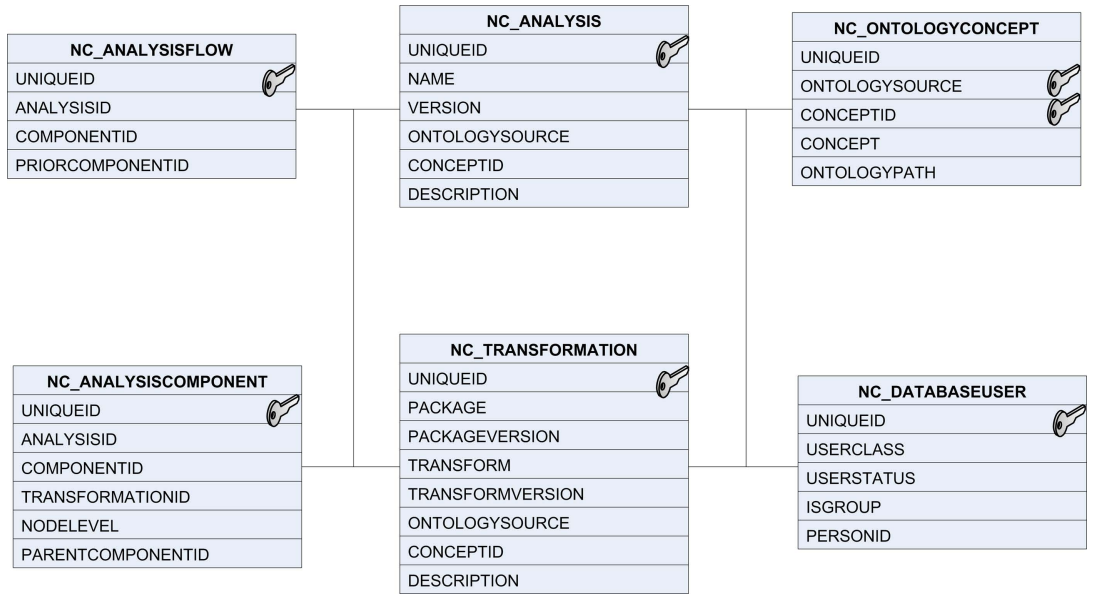


FIGURE 9 | Core HID tables for defining processing pipelines.

user-defined identifiers for the process whereas the package name and version corresponds to the name given by the process developers. The idea is to select processes from the *nc_transformation* table and put them together into pipelines. By adding the processes to the *nc_transformation* table, one can reuse tools in subsequent analytic pipelines. With respect to the use cases, the *nc_transformation* table contains entries for "autorecon-all" and "fsstats2xcede.pl" for the StructMorph analysis and "avwmerge", "avwmaths++", "mcflirt", "nifti_tool", "bet", "fugue", "flirt", "slicetimer", "mri_fwhm", and "ip_32R" for the PreProc analysis. By comparing the list with **Figure 4** there are four occurrences of the "nifti_tool" process in the pipeline but only a single entry in the *nc_transformation* bag of tools table. Next, the processing pipeline is assembled from the tools available in the *nc_transformation* table and the processing flow defined. The *nc_analysisFlow* table defines the flow through the processing tree defined in the *nc_analysisComponent* table.

For the PreProc pipeline, the *nc_analysisFlow* table contains two entries, one for "autorecon-all" and one for "fsstats2xcede.pl". The *analysisid* entry uniquely identifies the processing pipeline as described in the *nc_analysis* table's name, version and ontology source fields. The *componentid* field in the *nc_analysisFlow* table references the component ID stored in the *nc_analysisComponent* table for a process (autorecon-all for example). The *priorcomponentid* field in the *nc_analysisFlow* table defines a component executing immediately prior to the current step of the pipeline. Any number of entries for prior components can be added to the *nc_analysisFlow* table for a given *componentid* providing flexibility in defining complex pipelines. The *nc_analysisComponent* table defines the hierarchical relationship between steps in the pipeline. The *analysisid* and *transformationid* fields reference the pipeline and processing steps. Fields *parentcomponentid* and *nodelevel* reference the parent processing step and the depth within the processing pipeline tree. The *nodelevel* field is used to both identify the first step in the processing pipeline tree (nodelevel = 1) and to group processing tasks into distinct levels (or depths). The *parentcomponentid* identifies the parent node in the pipeline. Cyclic operations in a graph representation of a processing pipeline where there are multiple executions of a particular step are duplicated in the current implementation. Database queries through the HID web interface can be constructed either as simple queries filtering on particular components of the pipeline (*nc_analysisComponent* table), on sequences of tools (*nc_analysisComponent* and *nc_analysisFlow* tables), and by overall pipeline named identifiers (*nc_analysis* table). More advanced concept and ontology based queries are also supported if the ontology fields are populated for the processing pipeline.

Pipeline metadata related to output formats from an analysis are described in a generic way similar to those used in HID for storing new data types (Ozyurt et al., 2004a,b, 2006). The *nc_extendedTuple* table along with a number of accessory tables enables new classes of data to be described in a similar way as one constructs classes in programming languages such as C++ and Java. In the StructMorph use case, the extended tuples functionality is used to describe the anatomical thickness measurements that are loaded into the database from the XCEDE2 document discussed above. The database graphical user interface uses the extended tuples class definition to construct a query interface in the web application that is appropriate for basic logical queries over the results from an instantiation of the pipeline on actual data (**Figure 10**). The mechanisms used by HID to automatically construct web based query forms are in active development and beyond the scope of this manuscript. Interested readers are encouraged to visit the NITRC HID website for further details and documentation.

The instantiation of a processing pipeline and the resulting derived data is stored among a variety of HID tables linking the analysis files deposited in the data grid (SRB) with the pipeline metadata stored using the data class description discussed above.

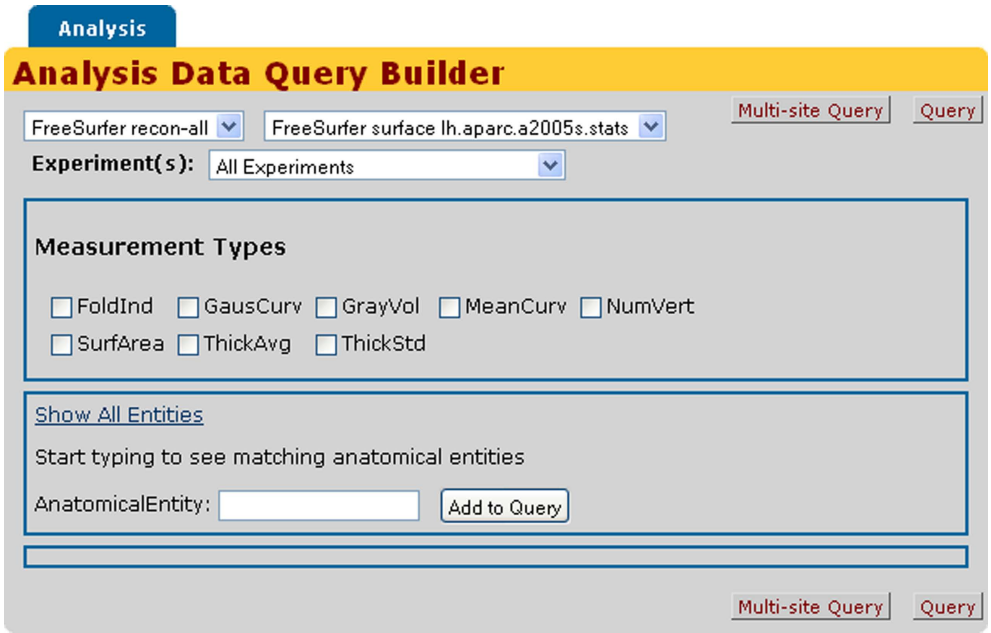


FIGURE 10 | HID web interface derived data query form for StructMorph analysis.

Because the databases are federated, it may not be the case that the original data used to produce a derived result are registered in the database where the pipeline outputs are to be stored. The provenance information stored in the XCEDE2 formatted output files includes information about which original data were used in the processing pipeline. This information is used by the HID import web service to determine whether the original data exists in the particular HID the derived result is being deposited or not. If the original data does not exist in the database, an entry is put into the *nc_externalData* table with information about which HID to contact for more detail about the original input data such as demographics, behavioral assessments, visit dates, etc. The HID federated query mechanism used to find information across the data federation is used in this context to provide additional information about the data included in an analysis pipeline. Interesting queries can be executed to locate all data processed with a particular pipeline and find which pipelines a particular dataset were used in, for example.

RESULTS

The StructMorph analysis was performed on 146 subjects collected across the FBIRN sites. The analysis was performed at two sites and the resulting derived data loaded into the HID systems at those two sites. Beyond the database and schema structures, code was written to convert the FreeSurfer cortical parcellation and volumetric segmentation output measures to XCEDE2 XML files. Instantiated pipeline provenance for each of the 146 runs was more difficult to obtain. Log files extracted from the processing tools were parsed for provenance and in some cases were unsatisfactory depending on the amount of information stored by the applications. The "fsstats2xcede.pl" tool was written within the FBIRN consortium and contained rich provenance information highlighting the need for either provenance wrappers around tools developed elsewhere or advocating the use of workflow environments such as LONI and Fiswidgets. Preliminary testing of metadata queries was successful, identifying the derived data consistently. The design of the derived data query pages required programmer input for clearer organization of form components.

The PreProc analysis was performed at one site after downloading the distributed data sets from the data federation and the resulting derived data loaded into the HID at the site performing the analysis. The database tables and XCEDE schema structures were sufficient in describing the more complicated processing pipeline. Investigators were initially interested in querying the PreProc data by filtering on pipeline provenance therefore the pipeline definition itself was used in test queries. For instance, a query to find all data derived using the "fugue" tool in the pipeline could be executed or a query to find the pipeline called "FIPS_MBTS_preprocs" (name stored in *nc_analysis* table for the PreProc pipeline, **Figure 4**).

There were many analyses done using the data collected prospectively by the FBIRN consortium. Details about the data processing pipelines and results can be found in publications Friedman and Glover (2006), Magnotta and Friedman (2006), Friedman et al. (2008), Ford et al. (2009), Potkin and Ford (2009), Potkin et al. (2009a,b) and Wible et al. (2009). The StructMorph and PreProc workflows were chosen to illustrate two different use cases for the derived data constructs presented here. Design of the derived data system was focused on the capability to represent the derived data generated as part of the publications listed above. Currently the derived datasets are being loaded into the data management system using the components discussed in this manuscript.

The most challenging aspect has been obtaining sufficient provenance from the applications used for processing. Convincing the tool developers to output detailed provenance records is time

consuming and difficult even when there is a good relationship between the developer and the users. Extracting provenance information from software log files is very demanding, error-prone, incomplete, and brittle. What has worked best for the FBIRN test bed, but far from satisfying, is a combination of working with developers (where possible) and scripting/automating analysis pipelines such that provenance is automatically documented during script execution. Processing pipelines are effectively wrapped with code to populate XCEDE formatted XML files with proper provenance detailing the analysis. There are no guarantees of provenance accuracy when wrapping pipelines. One could easily change a parameter and it not be reflected in the provenance output. FBIRN has found that regardless of the provenance capturing system and analysis automation method used, a human curator is invaluable for maintaining high quality data within the federation.

DISCUSSION

Storing and documenting derived results in data management systems along with important provenance information about the original input data and the pipeline itself in the context of a federated system is a challenging yet critically important endeavor. In large multi-site consortia where many geographically distributed investigators process the original data in different ways, providing a mechanism for them to contribute their work back to the federation and inform collaborators is desirable.

In the "The First Provenance Challenge" by Moreau et al. (2008), a challenge pipeline is presented along with a set of criteria to categorize and compare provenance systems (Moreau et al., 2008). **Table 1** describes the derived data system presented here in terms of the Moreau et al. (2008) categorization criteria. The derived data system is capable of storing the provenance challenge workflow described in Moreau et al. (2008) and addressing all of the core provenance queries. Core queries Q5, Q8, and Q9 in Moreau et al. (2008) filter on specific key-value pairs extracted from derived intermediate outputs or command line parameters of processing stage execution. Our system provides a very flexible method of allowing the researchers to specify which metadata and/or key-value pairs from the pipeline execution should be made query-able in the database graphical user interface (through the XCEDE XML representation, Section "Derived Data Exchange Schema").

The derived data management system introduced here is a joint effort by many collaborators across the BIRN consortium and the authors believe have promise in facilitating knowledge discovery through collaborative, distributed, data collection and analysis. The design and implementation is still being tested on many derived datasets produced and published by consortium members. Further testing of the query capabilities and automatic creation of derived data query forms is needed. Ultimately the goal is to create a dynamic federated system where collaborators can download original data (or derived data), perform novel analyses, and contribute that information back to the federation in a consistent and well documented way with minimal programmer input. The generic structures presented here are a good start and have been useful to the FBIRN consortium.

Table 1 | Characteristics of the derived data system with respect to the categorization presented in Moreau et al. (2008), "The First Provenance Challenge".

1.	**Characteristics of provenance systems**	
1.1	Execution environment	Web
1.2	Challenge execution environment	Not applicable
1.3	Provenance representation	XML and RDBMS
1.4	Query language	SQL
1.5	Research emphasis	R/S/Q
1.6	Challenge implementation	Not applicable
2.	**Properties of provenance representation**	
2.1	Includes workflow representation	Yes
2.2	Data derivation vs. causal flow events	D/E
2.3	Arbitrary annotations in scope/ implemented	+AS
2.4	Time supported/required	(+TS/+TR)
2.5	Naming required	URIs
2.6	Tracked data and granularity	File collections or process
2.7	Abstraction mechanisms	Layered provenance model

ACKNOWLEDGMENTS

This research was supported by U24-RR021992 to the Function Biomedical Informatics Research Network, U24-RR021382 to the Morphometry Biomedical Informatics Research Network, and U24-RR019701 to the Biomedical Informatics Research Network Coordinating Center (BIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH). The authors thank Randy Notestine at the University of California, San Diego for contributing the fMRI PreProc use case, Steven Potkin at the University of California, Irvine for his support of the work presented here, and the FBIRN consortium members for the data collection and support of the scientists and engineers involved in the project.

REFERENCES

Arzberger, P., and Finholt, T. A. (2002). Data and collaboratories in the biomedical community. In Report of a Panel of Experts Meeting, September 16–18, 2002, Ballston, VA.

Fissell, K., Tseytlin, E., Cunningham, D., Iyer, K., Carter, C. S., Schneider, W., and Cohen, J. D. (2003). Fiswidgets: a graphical computing environment for neuroimaging analysis. *Neuroinformatics* 1, 111–125.

Ford, J. M., Roach, B. J., Jorgensen, K. W., Turner, J. A., Brown, G. G., Notestine, R., Bischoff-Grethe, A., Greve, D., Wible, C., Lauriello, J., Belger, A., Mueller, B. A., Calhoun, V., Preda, A., Keator, D., O'Leary, D. S., Lim, K. O., Glover, G., Potkin, S. G., and Mathalon, D. H. (2009). Tuning in to the voices: a multisite FMRI study of auditory hallucinations. *Schizophr. Bull.* 35, 58–66.

Foster, I., Vockler, J., Wilde, M., and Zhao, Y. (2003). The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration. Asilomar, CA, Proceedings of the Conference on Innovative Data Systems Research.

Freire, J., Santos, D., and Silva, E. (2008). Provenance for computational tasks: a survey. *Comput. Sci. Eng.* 10, 11–21.

Friedman, L., and Glover G. H. (2006). Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage* 33, 471–481.

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., Gollub, R. L., Lauriello, J., Lim, K. O., Cannon, T., Greve, D. N., Bockholt, H. J., Belger, A., Mueller, B., Doty, M. J., He, J., Wells, W., Smyth, P., Pieper, S., Kim, S., Kubicki, M., Vangel, M., and Potkin, S. G. (2008). Test–retest and between-site reliability in a multicenter fMRI study. *Hum. Brain Mapp.* 29, 958–972.

Grethe, J. S., Baru, C., Gupta, A., James, M., Ludascher, B., Martone, M.,

Papadopoulos, P. M., Peltier, S. T., Rajasekar, A., Santini, S., Zaslavsky, I. N., and Ellisman, M. H. (2005). Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. From grid to healthgrid: proceedings of healthgrid 2005. Amsterdam, IOS Press.

Keator, D. B. (2009). Management of information in distributed biomedical collaboratories. *Methods Mol. Biol.* 569, 1–23.

Keator, D., Gadde, S., Grethe, J., Taylor, D., Potkin, S., and FBIRN. (2006). A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels. *Neuroinformatics* 4, 199–211.

Keator, D., Grethe, J., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., and Bockholt, J. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172.

MacKenzie-Graham, A. J., Payan, A., Dinov, I. D., Van Horn, J. D., and Toga, A. W. (2008). Neuroimaging data provenance using the LONI pipeline workflow environment.

Magnotta, V. A., and Friedman, L. (2006). Measurement of signal-to-noise and contrast-to-noise in the fBIRN multicenter imaging study. *J. Digit. Imaging* 19, 140–147.

Moreau, L., Ludascher, B., Altintas, I., Barga, R. S., Bowers, S., Callahan, S., Chin, Jr, G., Clifford, B., Cohen, S., Cohen-Boulakia, S., and Others. (2008). Special issue: the first provenance challenge. *Concurrency Comput. Pract. Exp.* 20, 409–418.

Olsen, J. S., Ellisman, M., James, M., Grethe, J. S., and Puetz, M. (2008). The biomedical informatics research network. In Scientific Collaboration on the Internet, G. M. Olson, A. Zimmerman and N. Bos, eds (Cambridge, MA, MIT Press).

Ozyurt, B., Wei, D., Keator, D., Gadde, S., Bockholt, J., Pease, K., and Grethe, J. (2006). A Complete Scientific Data Management Environment. Atlanta, GA, Society for Neuroscience.

Ozyurt, B., Wei, D., Keator, D., Potkin, S., Brown, G., and Grethe, J. (2004a). A General and Extensible Database System for the Storage, Retrieval and Maintenance of Human Brain Imaging and Clinical Data. Budapest, Organization of Human Brain Mapping.

Ozyurt, B., Wei, D., Keator, D., Potkin, S., Brown, G., and Grethe, J. (2004b). Web-Accessible Clinical Data Management within an Extensible Neuroimaging Database. San Diego, CA, Society for Neuroscience.

Paton, N. (2008). Managing and sharing experimental data: standards, tools and pitfalls. *Biochem. Soc. Trans.* 36(Pt 1), 33–36.

Potkin, S. G., and Ford, J. M. (2009). Widespread cortical dysfunction in schizophrenia: the FBIRN imaging consortium. *Schizophr. Bull.* 35, 15–18.

Potkin, S. G., Turner, J. A., Brown, G. G., McCarthy, G., Greve, D. N., Glover, G. H., Manoach, D. S., Belger, A., Diaz, M., Wible, C. G., Ford, J. M., Mathalon, D. H., Gollub, R., Lauriello, J., O'Leary, D., van Erp, T. G., Toga, A. W., Preda, A., and Lim, K. O. (2009a). Working memory and DLPFC inefficiency in schizophrenia: the FBIRN study. *Schizophr. Bull.* 35, 19–31.

Potkin, S. G., Turner, J. A., Guffanti, G., Lakatos, A., Fallon, J. H., Nguyen, D. D., Mathalon, D., Ford, J., Lauriello, J., and Macciardi, F. (2009b). A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophr. Bull.* 35, 96–108.

Rajasekar, A., Wan, M., Moore, R., Schroeder, W., Kremenek, G., Jagatheesan, A., Cowart, C., Zhu, B., Chen, S., and Olschanowsky, R. (2003). Storage resource broker – managing distributed data in a grid. *Comput. Soc. India J.* 33, 42–54 (special issue on SAN).

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *Sigmod Rec.* 34, 31.

Strother, S. C. (2006). Evaluating fMRI preprocessing pipelines. *IEEE Eng. Med. Biol. Mag.* 25, 27–41.

Wible, C. G., Lee, K., Molina, I., Hashimoto, R., Preus, A. P., Roach, B. J., Ford, J. M., Mathalon, D. H., McCarthey, G., Turner, J. A., Potkin, S. G., O'Leary, D., Belger, A., Diaz, M., Voyvodic, J., Brown, G. G., Notestine, R., Greve, D., and Lauriello, J. (2009). fMRI activity correlated with auditory hallucinations during performance of a working memory task: data from the FBIRN consortium study. *Schizophr. Bull.* 35, 47–57.

Zhao, Y., Wilde, M., and Foster, I. (2006). Applying the virtual data provenance model. *Lect. Notes Comput. Sci.* 4145, 148.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2009; paper pending published: 07 July 2009; accepted: 16 August 2009; published online: 07 September 2009.

Citation: Keator DB, Wei D, Gadde S, Bockholt J, Grethe JS, Marcus D, Aucoin N and Ozyurt IB (2009) Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. Front. Neuroinform. ***3****:30. doi: 10.3389/neuro.11.030.2009*

frontiers in
NEUROINFORMATICS

METHODS ARTICLE
published: 28 July 2009
doi: 10.3389/neuro.11.026.2009

Visualizing data mining results with the Brede tools

Finn Årup Nielsen*[1,2,3]

[1] *Center for Integrated Molecular Brain Imaging, Copenhagen, Denmark*
[2] *DTU Informatics, Technical University of Denmark, Lyngby, Denmark*
[3] *Neurobiology Research Unit, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark*

Edited by:
John Van Horn, University of California, Los Angeles, CA, USA

Reviewed by:
Kate Fissell, University of Pittsburgh Medical Center, Pittsburgh, PA, USA
John Van Horn, University of California, Los Angeles, CA, USA

***Correspondence:**
Finn Årup Nielsen, DTU Informatics, Richard Petersens Plads, DTU, 2800 Kongens Lyngby, Denmark.
e-mail: fn@imm.dtu.dk

A few neuroinformatics databases now exist that record results from neuroimaging studies in the form of brain coordinates in stereotaxic space. The Brede Toolbox was originally developed to extract, analyze and visualize data from one of them – the BrainMap database. Since then the Brede Toolbox has expanded and now includes its own database with coordinates along with ontologies for brain regions and functions: The Brede Database. With Brede Toolbox and Database combined, we setup automated workflows for extraction of data, mass meta-analytic data mining and visualizations. Most of the Web presence of the Brede Database is established by a single script executing a workflow involving these steps together with a final generation of Web pages with embedded visualizations and links to interactive three-dimensional models in the Virtual Reality Modeling Language. Apart from the Brede tools I briefly review alternate visualization tools and methods for Internet-based visualization and information visualization as well as portals for visualization tools.

Keywords: neuroimaging, visualization, software, meta-analysis, database, text mining, Web service, Brede

INTRODUCTION

In a narrow sense, neuroimaging workflows involve neuroimaging image processing and analysis. In a more broader sense, the workflow in a neuroimaging study involves a number of other processes: gathering information, designing the experiment, brain scanning, interpretation of the study, relating it to other studies and communicating the study. Data mining in neuroimaging may not only be applied as the standard neuroimaging analysis but also set to work on other components in workflow, and visualization of the data mining results may help the individual researcher in understanding his or her data as well as in communication with other researchers.

A number of tools exists for visualizing neuroimaging data mining results when the result is a volumetric neuroimage. There are, however, also visualization tools for other aspects of the neuroimaging process, and one example is our Brede Toolbox (Nielsen and Hansen, 2000a). Starting out as a program for handling and visualization of data from the BrainMap database of Fox et al. (1994) the Brede Toolbox now includes its own database of results from neuroimaging – the Brede Database (Nielsen, 2003) – as well as analysis and visualization functions for a range of tasks. We have setup an automated workflow involving a few non-interactive batch scripts that construct practically the entire Web presence of the Brede Database with static Web pages and visualizations. Furthermore, automated workflows using the ontologies of the Brede Database can perform mass meta-analysis across brain functions or brain regions (Nielsen, 2005; Nielsen et al., 2006a).

PORTALS FOR VISUALIZATION TOOLS

The abundance of tools for visualization as well as for other aspects of the neuroimaging process has spawned an interest in generating overviews for these tools, and now there exist several Web-based directories: Neuroscience Database Gateway (NDG) (Gardner and Shepherd, 2004), Neuroscience Information Framework (NIF) (Gardner et al., 2008), Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) (Buccigrossi et al., 2008), *I Do Imaging* and Internet Analysis Tools Registry (IATR), see also (Dinov et al., 2008). Some of these have an API so that HTML or XML for a specific tool can be requested. The NIF resource may be downloaded as an XML file. NITRC, IATR and I Do Imaging have Web 2.0 components with user-provided tool ratings and NITRC has an associated wiki. Since 2001, I have updated the *Bibliography on Neuroinformatics* which also lists numerous tools. Recently I began the Brede Wiki with structured information about neuroscience including neuroimaging visualization tools. Anyone can 'micro-publish' relevant information, and the structured content allows for off-wiki database queries (Nielsen, 2009).

META-ANALYTIC VISUALIZATION

Many meta-analyses use so-called *forest plots* and *funnel plots*, where scatter plots with whiskers display effect sizes and estimators of their variations in two dimensions (Lewis and Clark, 2001), see **Figure 1**. These meta-analyses typically investigate a single variable – continuous or dichotomous – and its relation to another variable, e.g., a personality trait and its association with a genetic polymorphism. In neuroimaging meta-analysis, we have a quite different situation: The neuroimage result contains not just one variable but many variables, i.e., voxels. One would need thousands of standard meta-analysis plots to capture the result across studies. Another much more fundamental problem stems from the fact that neuroimaging researchers typically only report the positive results, e.g., areas with activation to a given task, – not signal changes for brain regions that did not survive the statistical threshold selected. Meta-analysts usually regard the discarding of negative results as a heresy, referring to it as the *file drawer problem* or with the term *publication bias*. All the standard statistical meta-analysis technique require that also negative results are reported, – at least to some extent (Hedges and Olkin, 1985). So

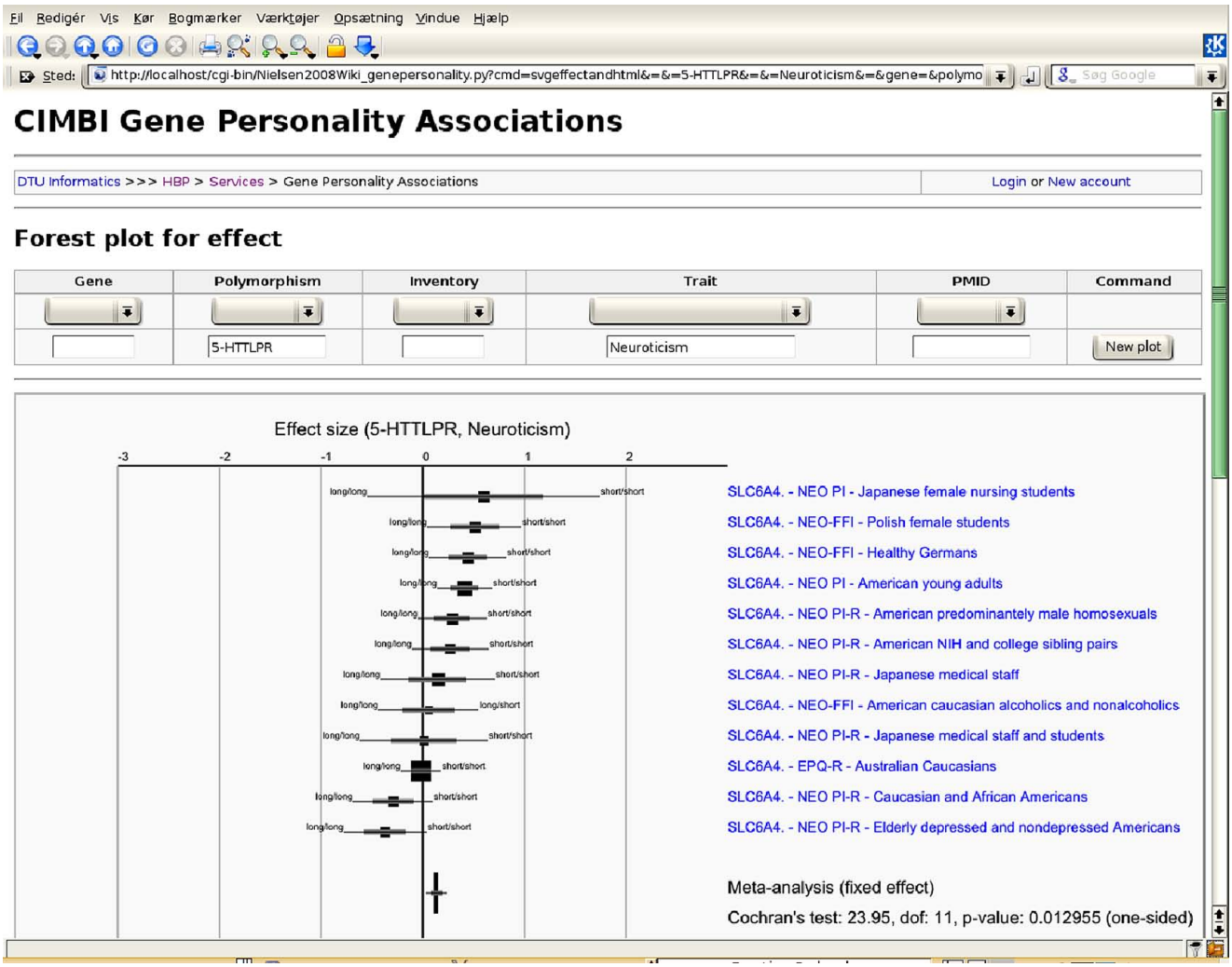


FIGURE 1 | Meta-analytic forest plot as a Web service with studies on personality genetics. Components in the Scalable Vector Graphics image file are hyperlinked and the content may be controlled interactively through a HTML form. Recent work with image-based meta-analysis has shown the possibility of constructing sensible forest and funnel plots for functional neuroimaging data (Salimi-Khorshidi et al., 2009b).

we may ask if it is at all possible to make appropriate analyses and visualizations across studies in neuroimaging?

One simple visualization simply plots the positive results – the reported coordinates – in stereotaxic space. The program associated with the original BrainMap database displayed coordinates in 2D tri-planar plot (Fox et al., 1994). This type of visualization is maintained in a newer version of the database with the program Sleuth (Laird et al., 2005). WebCaret may display coordinates in 3D as colored spheres together with an inflated cortical surface (Van Essen and Dierker, 2007), see **Figure 2**. The Brede Toolbox can generate 3D visualizations in the corner cube style of Rehm et al. (1998). Plotting points in 3D is not straightforward, – simple 'zero' dimensional graphics do not give an important perception of depth, therefore we use 3D glyphs of different color and shape. To help the viewer in spatial localizing the coordinates we can add components in a configurable workflow such as AC/PC axes, stalks for the glyphs, glyph shadows on the tri-planar walls, contour and cerebral cortex outlines from the atlas of Talairach and Tournoux (1988). **Figure 3** shows two visualizations of this kind with **Figure 3A** displaying all coordinates in the Brede Database from papers authored by Edward T. Bullmore and **Figure 3B** displaying cingulate coordinates colored according results from a text mining of the associated abstracts (Nielsen et al., 2005, 2006a). The batch script setup for the Brede Database will automatically generate a plot like **Figure 3A** for each author mentioned in the author ontology. Sometimes these simple plots reveal interesting features: The Bullmore coordinates appear somewhat limited to the middle of the inferior-superior axis perhaps reflecting a restricted field of view selected for some of the studies. The elaborate and automated workflow for generating a plot like **Figure 3B** involves:

1. Select a brain region and from the Brede Database brain region ontology get all naming variation of the brain region and its subareas. With these names extract coordinates from papers recorded

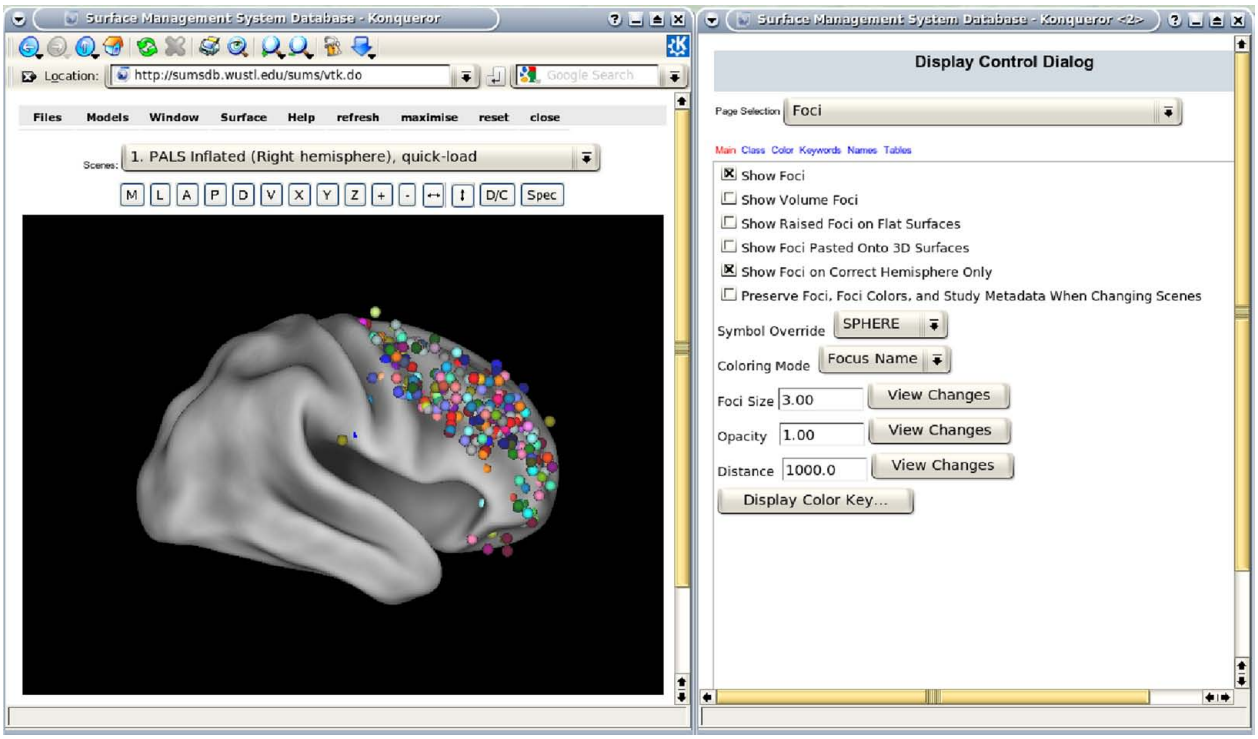


FIGURE 2 | WebCaret server-side display of returned coordinates from the Surface Management System Database (SumsDB) with a query on 'middle frontal gyrus'. The right window offers some control over the rendering and the buttons in the left window may rotate the cortical surface. SumsDB allows the query on a neuroanatomical label to be invoked from another program or Web site by simple Web linking, and the Brede Wiki automatically constructs such links.

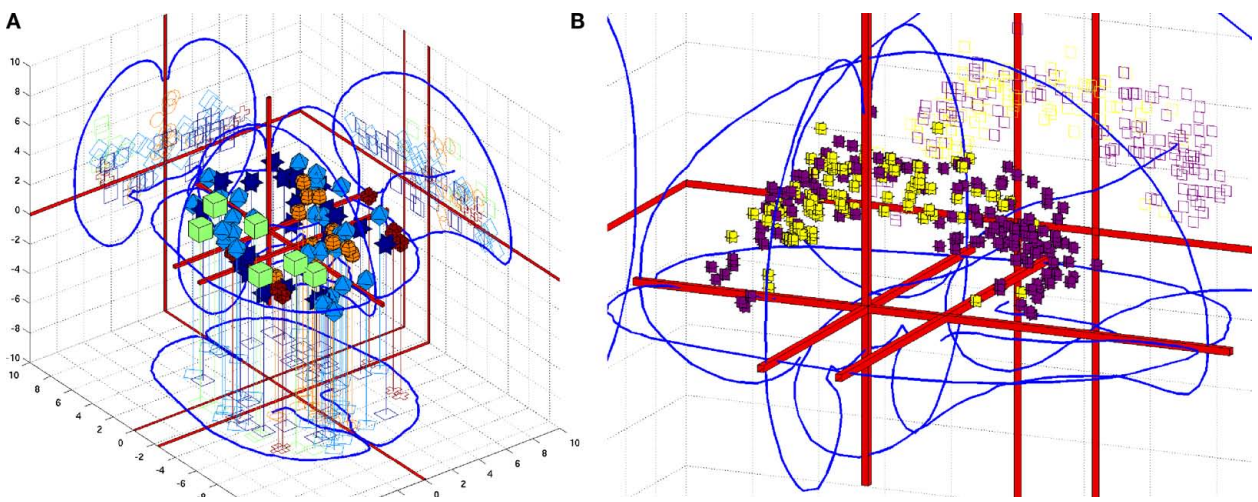


FIGURE 3 | Two examples of coordinates in a 3D corner cube visualization. **(A)** Coordinates from the five studies in the Brede Database authored by Edward T. Bullmore. The 3D glyphs have type and color according to paper: Dark blue (Phillips et al., 1997), light blue (Phillips et al., 1998), light green boxes (Bullmore et al., 1996), orange spheres (Hunkin et al., 2002), red (Calvert et al., 1999). **(B)** Cingulate coordinates colored according to the clustering results after a text mining of abstracts in the Brede Database. Dark magenta glyphs are from the 'memory' cluster while the light yellow are from the 'pain' cluster. From Nielsen et al. (2006a).

in the database, model their spatial distribution and include extra non-matched coordinates that lies within the region.

2. Get abstracts from the Brede Database that – for the brain region in question – have one or more coordinates and perform text mining, which results in clusters of themes, such as 'pain' and 'memory' and documents belonging to these clusters.
3. Perform statistical tests on the spatial distribution of the coordinates grouped according to the text mining clusters to

determine if the text mining has discovered functions that are segregated in the region.

The procedure is done for all brain regions in the Brede Database brain region ontology and **Figure 3B** shows one of the regions that listed high after sorting brain regions according to statistical significance in the spatial distribution test.

Data mining directly with the coordinates has been termed coordinate-based meta-analysis (CBMA) and several methods exists (Wager et al., 2009), see also Laird et al. (2009), this issue. For the most part they involve a form of estimation of a conditional probability density $p(\mathbf{v}|c)$ in stereotaxic space $\mathbf{v}$. The conditioning, c, may be, e.g., for a specific brain function or a specific anatomical label. Once the probability density is estimated it can be converted to a volume by sampling the probability density in voxels and visualized in the same way as standard neuroimages, or the density can be used to color-code the cortical surfaces in a 3D visualization, see Wager et al. (2009).

Fox et al. (1997) introduced the method to model the probability density: a single confined area – the primary motor area for the mouth – were examined so only a model with mean and standard deviation was devised, i.e., a simple Gaussian model. As more complex brain functions are distributed in brain space, more flexible models are needed. Our first effort in modeling the probability density was by Gaussian mixture models (Nielsen and Hansen, 1999):

$$p(\mathbf{v} \mid c) = \sum_{k}^{K_c} p(\mathbf{v} \mid k) P(k \mid c), \quad (1)$$

where each $p(\mathbf{v}|k)$ estimates a 3D Gaussian probability density. **Figure 4A** shows the isosurfaces in a model of this type where the parameters have been fitted to data from the BrainMap database. Here, each ellipsoids corresponds to a single Gaussian $p(\mathbf{v}|k)$ and c corresponds to three different labels of 'behavioral domain' from the BrainMap database that are associated with each coordinate. Although the Gaussian mixture model may generalize, the ellipsoids do not look neuroanatomical plausible and call for yet more flexible models. **Figure 4B** is generated with kernel density estimation using a Gaussian kernel (Nielsen and Hansen, 2000b). Such models seems to generate probabilities that are somewhat more neuroanatomical plausible than the Gaussian mixture model.

The isosurfaces in the probability densities in both subplots of **Figure 4** has been set for display purpose. More statistically grounded values can be obtained with the methods by Turkeltaub et al. (2002); Nielsen (2005); Costafreda et al. (2009). The methods for probability density estimation of coordinates are not limited to activations but may be applied to any kind of coordinates in stereotaxic space from 'deactivations', cortical stimulations, lesions or structural changes, e.g., obtained with voxel-based morphometry.

When a probability density estimate is constructed for a set of coordinates and it is converted to a voxel-volume, then the volumes across multiple sets of coordinates may be aggregated into a single data matrix $\mathbf{X}$(sets × voxels). This data matrix may then be decomposed with multivariate analysis in a number of ways, e.g., with singular value decomposition for principal component analysis, $\mathbf{ULV} = \mathbf{X}$, where the left factorization matrix $\mathbf{U}$(sets × components) contains loading over sets of coordinates for each principal component and the right factorization matrix $\mathbf{V}$(vowel × components) contains loadings over voxels. Other types of decomposition for this matrix are independent component analysis ($\mathbf{MS} = \mathbf{X}$, with $\mathbf{M}$ the mixing matrix and $\mathbf{S}$ the source matrix), non-negative matrix factorization ($\mathbf{WH} = \mathbf{X}$) and K-means clustering ($\mathbf{CA} = \mathbf{X}$, with $\mathbf{C}$ a centroid matrix and $\mathbf{A}$ an assignment matrix). The right decomposition matrices, $\mathbf{V}$, $\mathbf{S}$, $\mathbf{H}$ and $\mathbf{A}$ all contain vectors that each represents a volume. As part of the workflow for presenting the information in the Brede Database on the Web the decompositions work on data matrices formed from sets of papers and sets of experiments, and corner cube visualizations are automatically constructed with

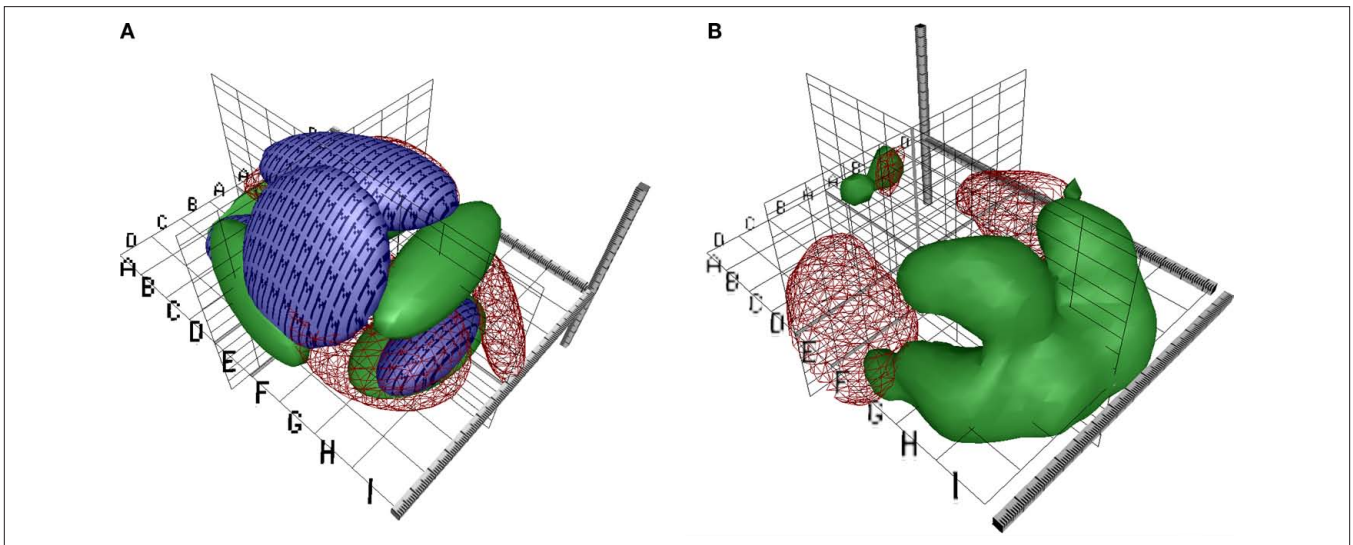


FIGURE 4 | VRML visualizations with coordinate-based meta-analysis of data from BrainMap with isosurfaces in conditional probability densities. **(A)** Gaussian mixture model of the three main behavioral domains in BrainMap: Perception (red wireframe), cognition (green surface) and motion ('M'-textured surface). From Nielsen and Hansen (1999). **(B)** Kernel density modeling of auditory (red wireframe) and vision (green) studies. From Nielsen and Hansen (2000b).

isosurfaces in the volumes contained in the right decomposition matrices. **Figure 5** shows such a visualization for a component from non-negative matrix factorization, i.e., a row in the **H** matrix. Such visualizations may be useful for navigating among the studies in the database, and to a certain extent they reveal spatial distributions of the 'cognitive components' of the brain. Together with the visualization on the Web page are listed the experiments that have high association with the component, i.e., experiments associated with large elements in a column of the left matrix **W**. For the component in **Figure 5** they are experiments described as, e.g., 'Visual object decision', 'Buildings visual objects', 'Color perception during free viewing' and 'Passively viewed scenes'.

Before putting too much trust in visualizations and analysis across studies one needs to remember that the study results may have arisen in quite different ways. In standard meta-analysis the only variations between studies that are usually modeled is the number of subjects and the standard deviation of the data in the individual studies. In neuroimaging meta-analysis and visualization these variables are not usually modeled, for exceptions see Fox et al. (1997). Besides there are several other variables that neither are considered: The varying thresholds applied, e.g., corrected and uncorrected *P*-values (Nielsen et al., 2006b), the difference in field of view between studies, the reporting style of coordinates (e.g., 'extent threshold', 'number of maxima per cluster') as well as the variation from the different preprocessing and analysis choices that have been made. Furthermore, the different CBMA models may produce different results on the same material. Salimi-Khorshidi et al. (2009a) compared different CBMA models and their application of a threshold makes a 'blob' appear and disappear depending on the type of CBMA.

INTERNET-BASED VISUALIZATION

Quite a few tools exist for interactive neuroimaging visualization across the Internet. Often these tools are based on a client–server model with the client implementing the visualization and graphical user interface in Java. Among these tools are JIV that renders multiple volume data by orthogonal slice views implemented as a Java applet (Cocosco and Evans, 2001). iiV implements a similar functionality (Lee et al., 2008), and MindSeer can also render in 3D remotely (Moore et al., 2007). NeuroTerrain implements 3D visualization and has demonstrated its use in connection with a Mouse atlas (Gustafson et al., 2007). The Talairach Applet renders a digital representation of the Talairach Atlas and combines it with neuroanatomical labeling of coordinates via the Talairach Daemon described by Lancaster et al. (2000). Also in connection with the BrainMap database the Java client-program Sleuth plots 3D points in orthogonal 2D slices based on user query to the BrainMap server (Laird et al., 2005).

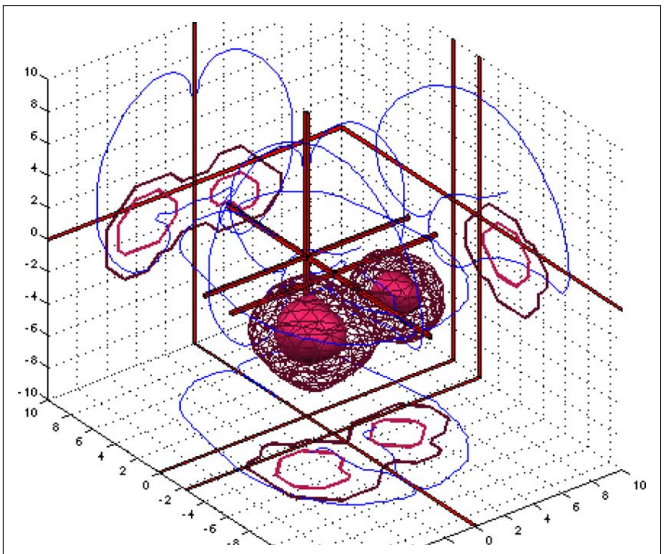


FIGURE 5 | Corner cube visualization on the Web page of the Brede Database with results from a non-negative matrix factorization of experiments in the database.

The *Internet Brain Volume Database* (IBVD) records published values for brain region volumes across variables such as gender and diagnosis (Kennedy et al., 2003). Since the neuroimaging data analysis arrives at one single value – the brain volume in cubic centimeters – the visualization of the data is relatively simple compared to other neuroinformatics visualizations: From Web-based user queries IBVD generates on-the-fly PNG image-files with the brain volumes from the different studies plotted as a function of age with color-coding and the variability indicated. Interactive visualization systems for neuroimages with server-side 3D rendering have been described by Poliakov et al. (2005) and a public system is available with the WebCaret Web service, see **Figure 2**.

With the Brede Toolbox we construct 3D visualizations browsable on the Web by using the Virtual Reality Modeling Language (VRML) (ISO/IEC, 1997; Nielsen and Hansen, 2000a), see the VRML examples in **Figure 4**. When defined in the middle of 1990s VRML held great promise to get wide-spread use for 3D interactive and hyperlinked visualizations, but since then it has had limited growth: VRML lacks good browser implementations and there has been erratic adoption of a scripting language. Nevertheless, it is one of the few means for Web distribution of 3D content in free standardized form. An alternative format is the *Universal 3D File Format* (U3D) that can be embedded in newer versions of the PDF format. Apart from the Brede Toolbox *ImageSurfer* described by Feng et al. (2007) implements VRML export.

For the Web presentation of the Brede Database, we generate 3D corner cube visualizations of the coordinates in the database with an offline Matlab batch script, – both as image files embedded on the Web page as well as VRML files, see **Figure 6**. Matlab is not well suited to work as a Web script, and for the interactive Web scripts associated with the Brede Database, there are presently no visualization implemented. The INC *Interactive Talairach Atlas* renders 2D orthogonal slices from the Talairach and the MNI single subject atlases. This Web service can merge a user-given coordinate with the visualization, and as such we use it for visualization of individual coordinates from the Brede Database and the Brede Wiki, see **Figure 7** for an example.

Besides Java, VRML and standard image files such as PNG the *Scalable Vector Graphics* (SVG) format may prove useful for Internet-based visualizations, see **Figure 1** for an example. These files may contain hyperlinks and JavaScript. However, Web browsers do not yet consistently implement the standard.

INFORMATION VISUALIZATION

Data mining results from neuroimaging analysis are not the only type of information for visualization. Information about the background,

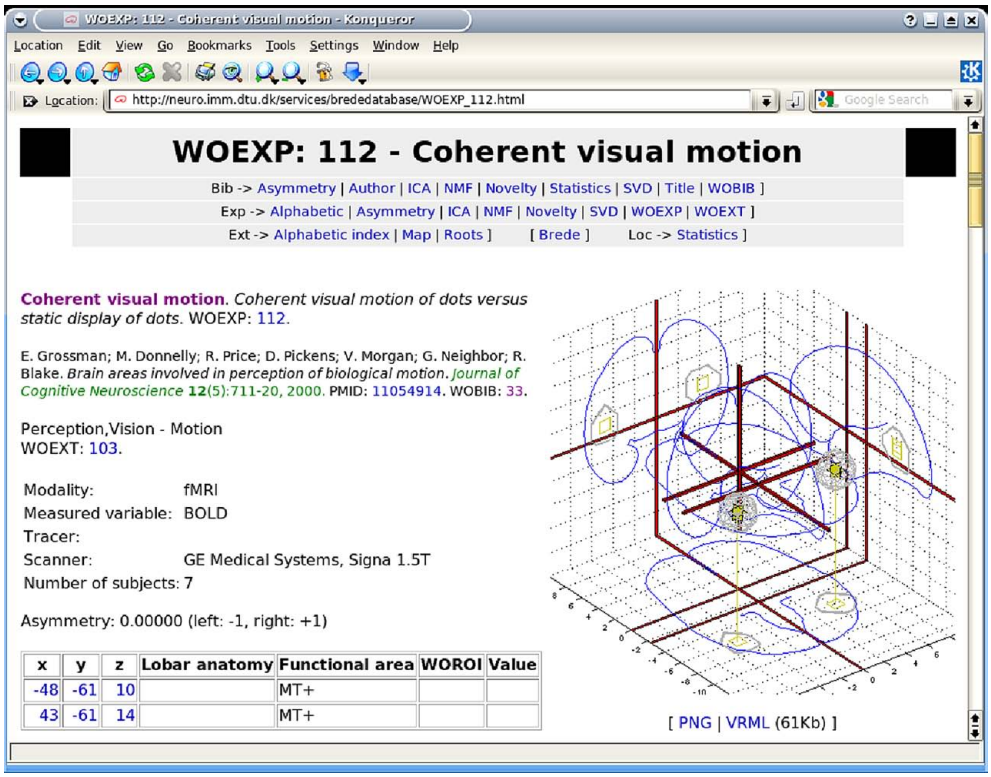


FIGURE 6 | Screenshot of the Web page for an experiment in Brede Database with a corner cube visualization of the coordinates in a experiment together with a wireframe indicating an isosurface of the kernel density estimate with the coordinates. An interactive rendering is provided with the link to a generated corner cube visualization in a VRML file.

design, scanning, analysis procedure, and interpretation surrounds the data mining results of a typical neuroimaging study. In scientific articles, the body text mostly carries this 'context' information, though sometimes authors also use tables to describe, e.g., subject information. Authors rarely apply visualizations for this kind of information except in situations with explanation of the experimental design and scanning. The experimental design has a natural temporal evolution and as such the visualization often displays the design as a function of time. Users of the behavioral experiment software from Psychology Software Tools is familiar with the graphical programming environment of *E-Prime* which has this kind of visualization as an integral part of the development of the experiment. Other parts of the neuroimaging study may be visualized with what is usually referred to as *information visualization*.

In a demonstration visualization, we employed a torus topology for an entire neuroimaging study process constructing 3D icons for 'funding', the experimental design, authors, experimental subjects, etc. (Nielsen and Hansen, 1997), see also **Figure 8**. The usefulness of such a visualization depends on how effective it conveys information compared to standard text, and if the visualization format requires specialized and limited distributed programs for rendering and interaction the impact may be small. Manual creation of these visualizations is infeasible, – the visualization should be constructed automatically from description of the study, e.g., the so-called 'provenance' (Fissell, 2007). In related visualizations, some workflow management systems display the processing flow graphically (Dinov et al., 2008).

When neuroimaging studies get reported in articles the relationships between the articles can be turned in to visualizations. Many types of visualizations exist and many relationships may be revealed: Between terms, concepts, citations to and from articles as well as between authors, cited authors and cited journals. The visualizations are of course not limited to articles only in neuroimaging, see, e.g., Card et al. (1999); Chen (1999). For an example in neuroscience Naud et al. (2007) use a spherical embedding algorithm to display a bipartite graph in 3D space with two spheres. One of their illustrations visualized the relationship between poster sessions in the Society for Neuroscience 2006 meeting together with words from the abstracts in the sessions. Another example of text mining result visualization is what we termed a 'cluster bush', that describe the clusters in a hierarchical multivariate analysis (Nielsen et al., 2005): Clusters are indicated with dots and thick lines indicate a large similarity between two clusters. Given a set of abstracts the automated workflow for generating a plot like **Figure 9** involves the conversion of the texts to a bag-of-words matrix, the exclusion of a large number of words (stop words), hierarchical non-negative matrix factorization and lastly the 'cluster bush' visualization all implemented with the functions of the Brede Toolbox.

Coordinate-based meta-analysis and text mining can be combined to form visualizations, see **Figure 10** and Nielsen et al.

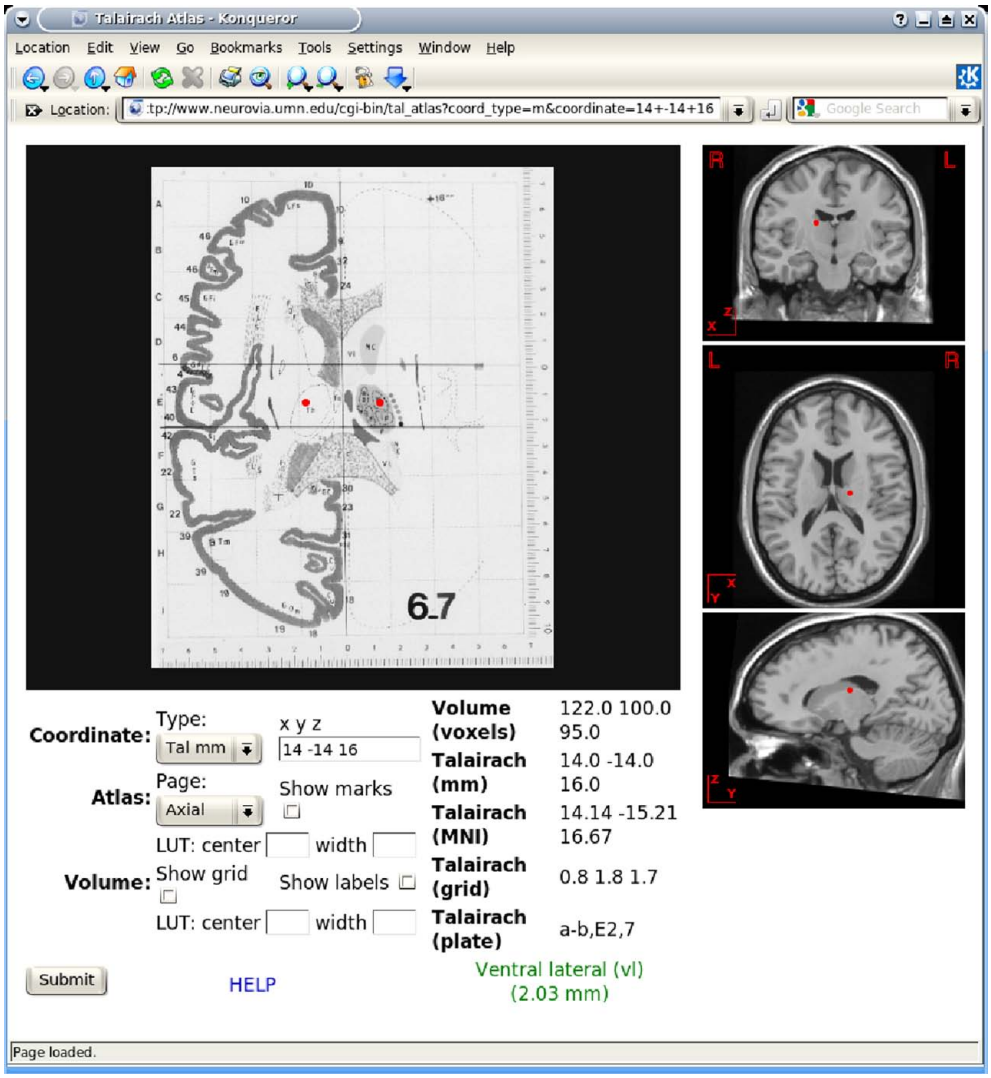


FIGURE 7 | The Web-based INC Interactive Talairach Atlas queried with a coordinate from the Brede Wiki.

(2004). The workflow for constructing the visualization in the figure involves the setup of a matrix describing the words in the abstract of papers and the construction of another matrix from kernel density estimation with the coordinates in each paper. After non-negative matrix factorization each individual factor may be rendered in 3D and associated with words from the abstract, e.g., the blue area in **Figure 10A** in the occipital lobe is associated with words such as 'visual' and 'eye'.

Based on a corpus of articles published between 1997 and 2000 in the journal *NeuroImage* we could plot cited authors and cited journals in 2D. The data mining with visualization would for example reveal a dichotomy between PET and fMRI (Nielsen, 2002), see **Figure 11**. Here, the workflow involves specialized algorithms that extract citations and the use of matrix computations, particularly singular value decomposition, for multidimensional scaling-like projection of the data onto 2D. For the Brede Database, we automatically construct what we have termed 'bullseye plots' to display the network of coauthors for each recorded author. **Figure 12** shows a larger bullseye plot on coauthors in the *NeuroImage* corpus. Authors near the center, such as Friston and Dolan, have high network degrees, which here corresponds to the number of authored articles (Nielsen, 2002).

The well-tested and widely used GraphViz package provides spatial graph layout for a given network (Gansner and North, 2000). At one point the *PubGene* Web service used GraphViz in a large-scale application for displaying relations between genes based on literature in PubMed (Jenssen et al., 2001). GraphViz layouts graphs for the Web presentation of the Brede Database. These graphs display the brain function and brain region ontologies, e.g., indicating that 'vision' has 'perception' as taxonomic parent or that the cingulate area is a parent for the posterior cingulate, see **Figure 13**. Our workflow with the Brede Toolbox involves extraction of the ontology from Brede Database XML files, construction of a file with the graph that GraphViz reads, invoking GraphViz for generation of an image file, and then finally construction of the Web page with

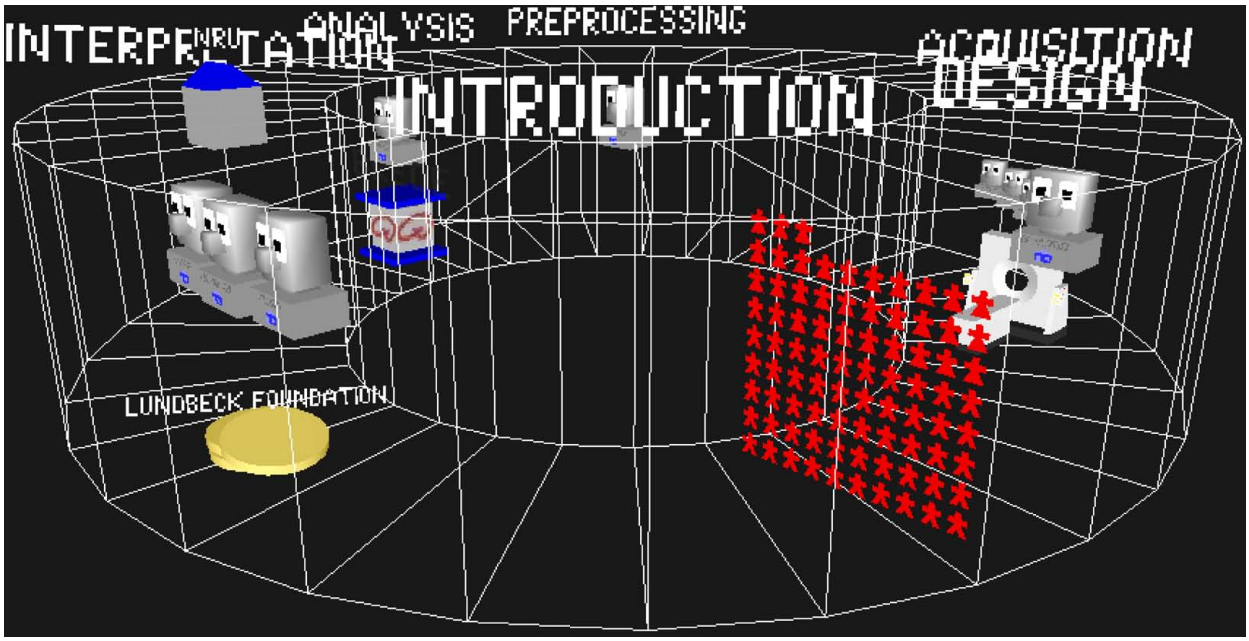


FIGURE 8 | Screenshot of a VRML rendering seeking to convey parts of the information surrounding a neuroimaging study: 3D icons for funding, research organization, researchers, software, subjects, and scanner placed in a torus.

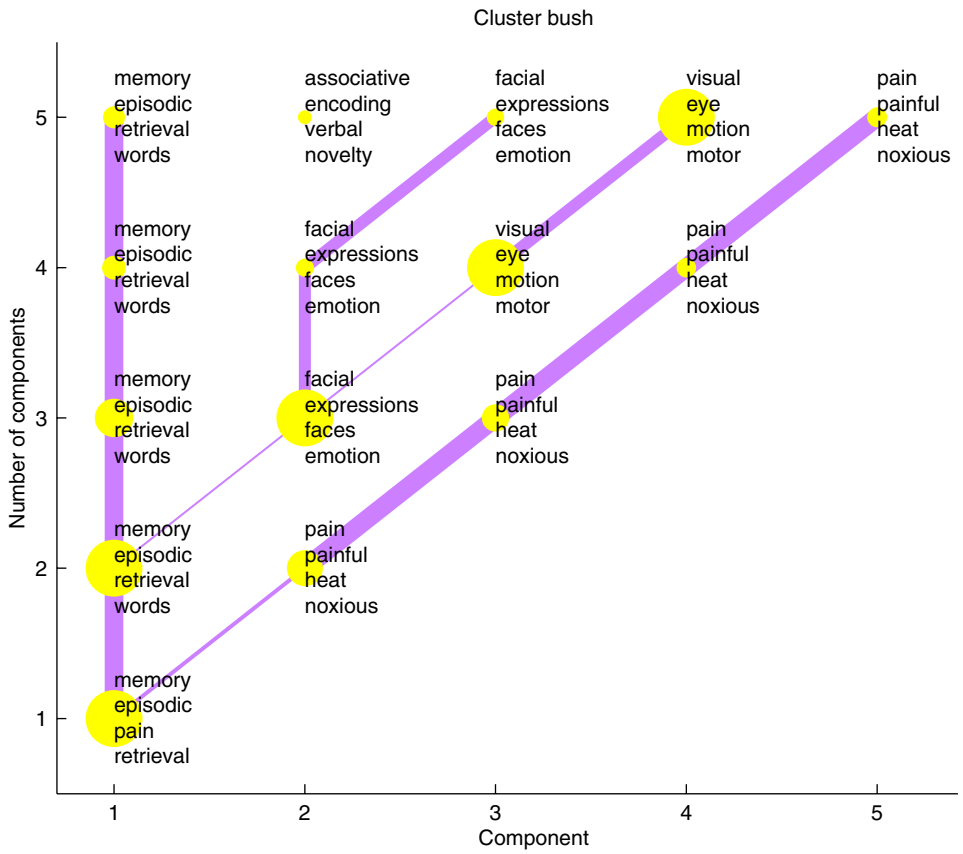


FIGURE 9 | A so-called 'cluster bush' visualization of the text mining results of the abstracts in the Brede Database. Each yellow dot is a cluster of articles and words in the article. The four words with highest load on each cluster are listed.

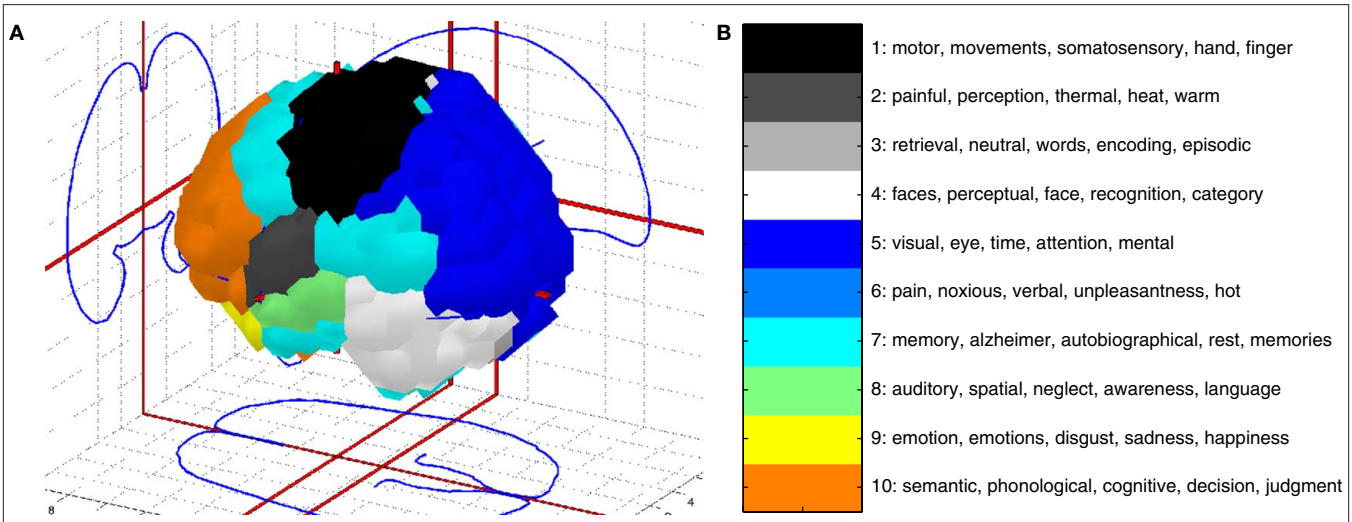


FIGURE 10 | A 'functional' atlas generated from a combined data mining of text and brain coordinates from the Brede Database. (A) Corner cube visualization with labeled brain areas. **(B)** Automatically generated legend with words from the text mining of abstracts.

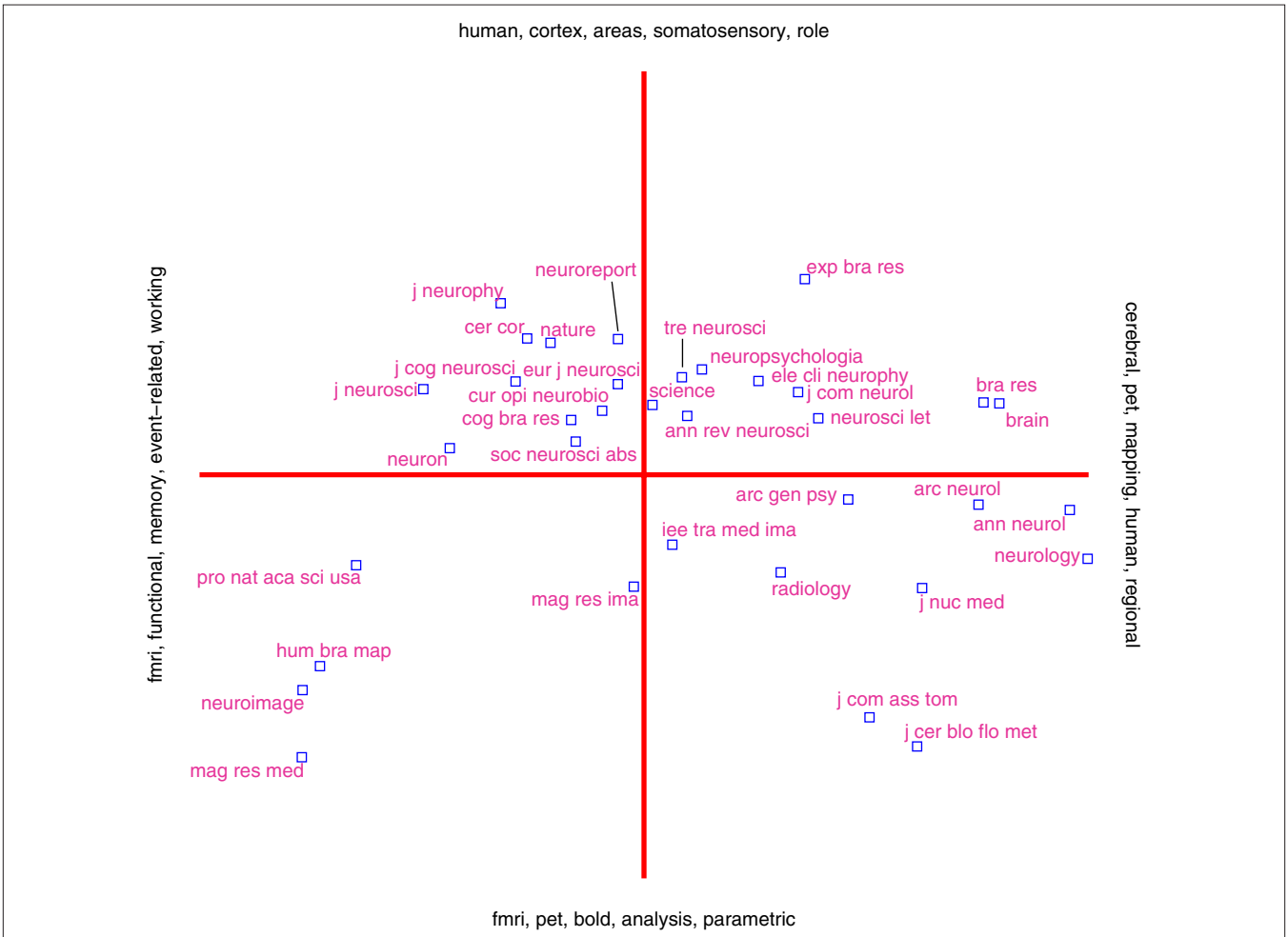


FIGURE 11 | Visualization of data mining result of journal co-citation analysis with singular value decomposition on citation data from *NeuroImage*.

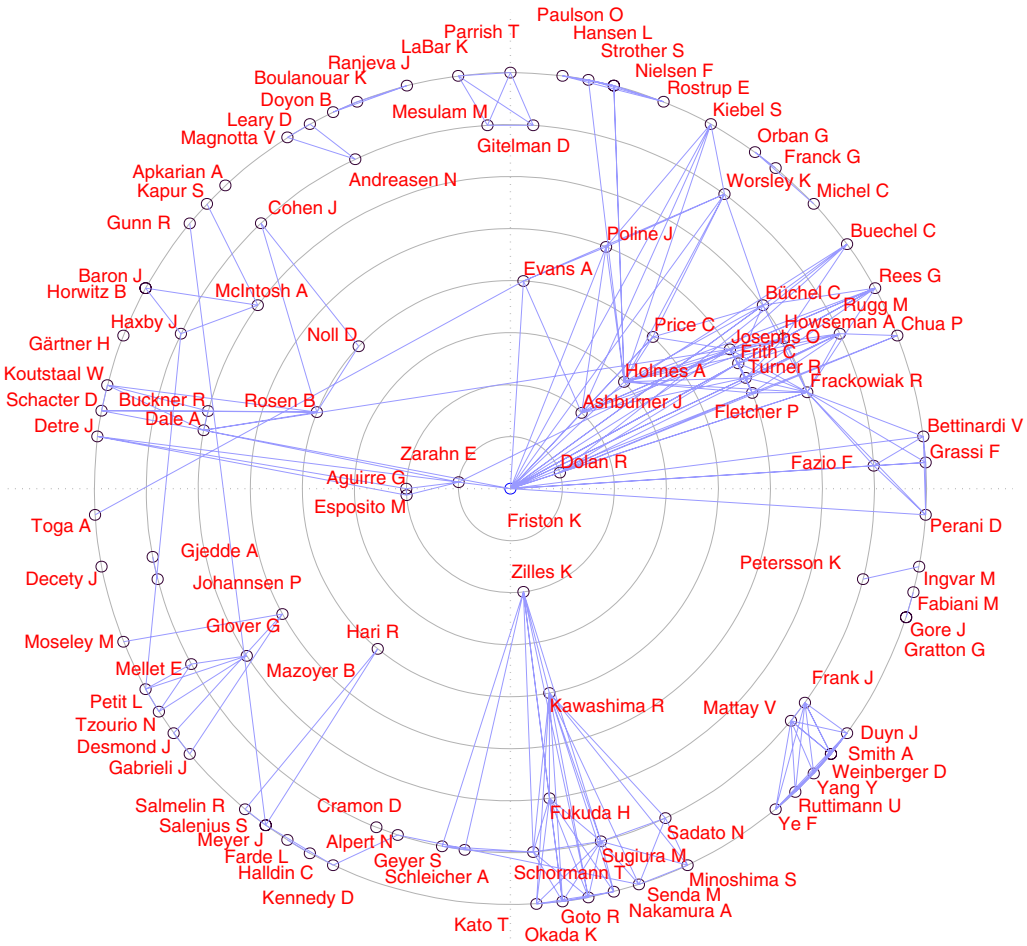


FIGURE 12 | Coauthor bullseye plot (target diagram) with data from *NeuroImage* 1997–2000. A line between two authors indicates that they co-wrote a paper. The concentric circles indicate the number of articles written by the author in the corpus. The Brede Toolbox automatically constructs similar, albeit smaller, bullseye visualizations for each author represented in the Brede Database author ontology. These are available on the Web.

the image file embedded. GraphViz can construct HTML image maps so the nodes in the graph image are associated with clickable hyperlinks. On the final Web page a reader may navigate the brain region and brain function ontologies by clicking on the nodes in the graph. The Brede Toolbox can also use GraphViz for layout of other types of data that can be described as a network, e.g., from structural equation modeling of regional neuroimaging data. A number of journal Web sites use plots called *Citation map* in the style of GraphViz for visualizing in- and out-going citations of each article, see, e.g., *BMJ* and *The Journal of Neuroscience* Web sites.

Another type of graph visualization within neuroimaging is the interactive graph visualization with a hyperbolic browser that features in tools from the Laboratory of Neuro Imaging (LONI): LOVE and iTools (Dinov et al., 2006, 2008). *ISI Web of Knowledge* provides a Java applet to render their citation information with a similar topology.

CONCLUSION AND FUTURE WORK

With the Brede Toolbox we are able to build a workflow with extraction of data from the Brede Database, automated data mining and visualizations. The automated procedures generate publicly accessible Web pages with interactive visualizations. An advantage of the automated procedure is that little human intervention is required to update the visualizations as new data is added to the database. The visualizations can display not only spatial neuroimages, but for example also results from text mining, and visualization can take place across the Internet with data originating on one server and displayed on another.

The Brede Database represents just a small fragment of the results from the published literature (Derrfuss and Mar, 2009). Databases such as NeuroNames, BrainMap and SumsDB are much larger. However, no universal database exist for coordinates from functional neuroimaging. To gain a higher degree of coverage future work may attempt to aggregate data from different databases for combined visualizations. Since typical meta-analytic data is anonymous and small (compared to a typical neuroimaging study), it is easier to share such data and we may see collaborative Internet-based analyses and visualizations. Our wiki for personality genetics (**Figure 1**) is such a collaborative system. Building a collaborative system for neuroimaging data requires

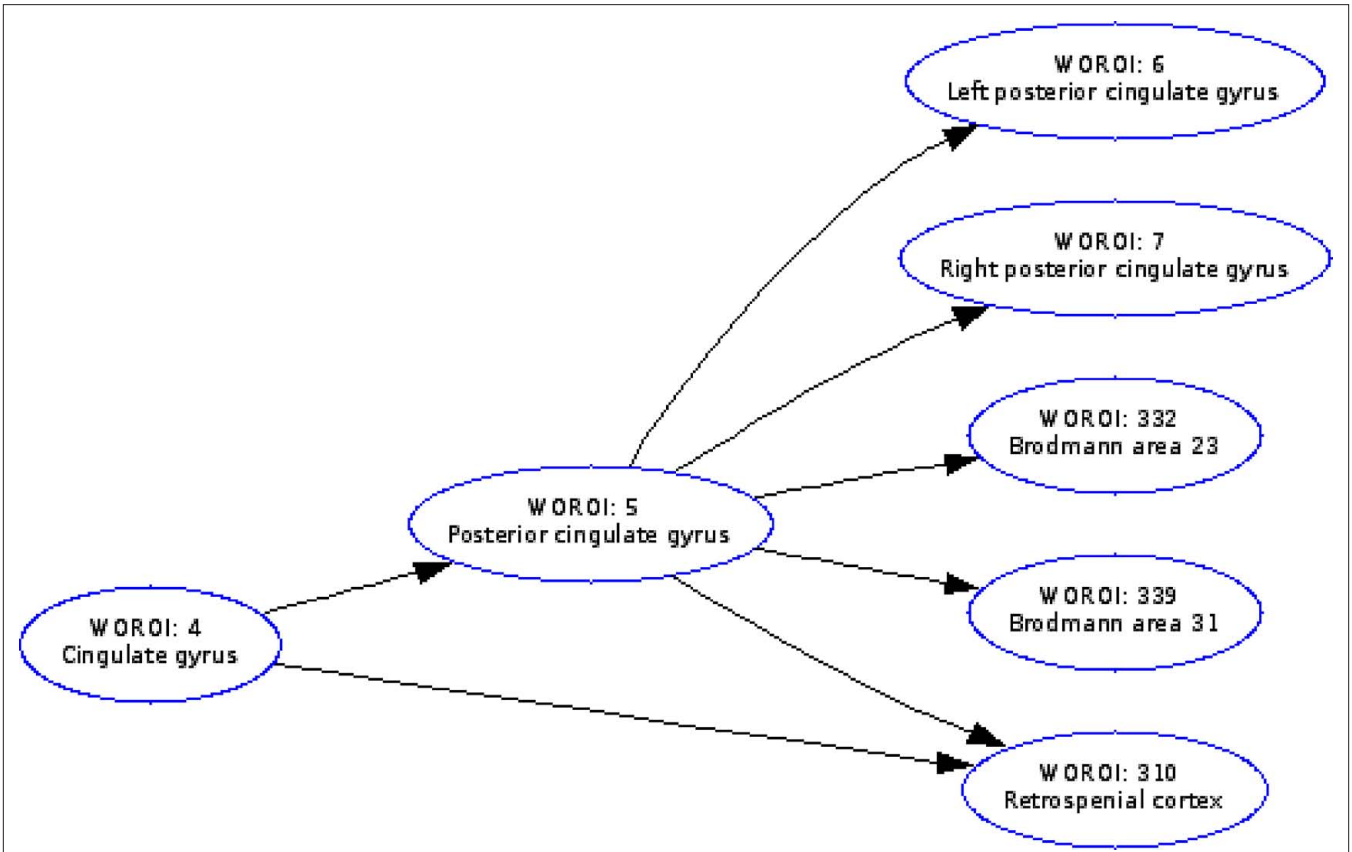


FIGURE 13 | For presenting the Brede Database brain region ontology on the Web the workflow with the Brede Toolbox invokes the GraphViz program which generates hyperlinked plots of the brain region hierarchy, here for the 'posterior cingulate gyrus'.

more effort, and in the Brede Wiki only simple visualizations are presently available. A target for future development should be towards 'Science 2.0' where data, analyses and visualizations can be shared in Web-based collaborative and user-friendly environments.

ACKNOWLEDGMENTS

Many thanks to Kristoffer Hougaard Madsen, Daniela Balslev and Lars Kai Hansen for comments on an earlier version of the manuscript. This work was supported by the Lundbeck Foundation through the Center for Integrated Molecular Brain Imaging.

REFERENCES

Buccigrossi, R., Ellisman, M., Grethe, J., Haselgrove, C., Kennedy, D., Martone, M., Preuss, N., Sullivan, M., and Wagner, K. (2008). The neuroimaging informatics tools and resources clearinghouse (NITRC). In 14th Annual Meeting of the Organization for Human Brain Mapping, (Organization for Human Brain Mapping), pp. 319 T–AM.

Bullmore, E. T., Rabe-Hesketh, S., Morris, R. G., Williams, S. C. R., and Gregory, L. (1996). Functional magnetic resonance image analysis of a large-scale neurocognitive network. *Neuroimage* 4, 16–33.

Calvert, G. A., Brammer, M. J., Bullmore, E. T., Campell, R., Iversen, S. D., and David, A. S. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport* 10, 2619–2623.

Card, S. K., MacKinlay, J. D., and Shneiderman, B. (eds) (1999). Readings in Information Visualization. Using Vision to think. The Morgan Kaufmann Series in Interactive Technologies. San Francisco, CA, Morgan Kaufmann Publishers.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Inf. Process. Manage.* 35, 401–420.

Cocosco, C. A., and Evans, A. C. (2001). Java internet viewer: a WWW tool for remote 3D medical image data visualization and comparison. In Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention, Volume 2208 of Lecture Notes in Computer Science (London, Springer), pp. 1415–1416.

Costafreda, S., David, A. S., and Brammer, M. J. (2009). A parametric approach to voxel-based meta-analysis. *Neuroimage* 46, 115–122.

Derrfuss, J., and Mar, R. A. (2009). Lost in localization: the need for a universal coordinate database. *Neuroimage*. In press.

Dinov, I. D., Rubin, D., Lorensen, W., Dugan, J., Ma, J., Murphy, S., Kirschner, B., Bug, W., Sherman, M., Floratos, A., Kennedy, D., Jagadish, H. V., Schmidt, J., Athey, B., Califano, A., Musen, M., Altman, R., Kikinis, R., Kohane, I., Delp, S., Parker, D. S., and Toga, A. W. (2008). *iTools*: a framework for classification, categorization and integration of computational biology resources. *PLoS ONE* 3, e2265.

Dinov, I. D., Valentino, D., Shin, B. C., Konstantinidis, F., Hu, G., MacKenzie-Graham, A., Lee, E.-F., Shattuck, D., Ma, J., Schwartz, C., and Toga, A. W. (2006). LONI visualization environment. *J. Digit. Imaging* 19, 148–158.

Feng, D., Marshburn, D., Jen, D., Weinberg, R. J., Taylor, R. M. II, and Burette, A. (2007). Stepping into the third dimension. *J. Neurosci.* 27, 12757–12760.

Fissell, K. (2007). Workflow-based approaches to neuroimaging analysis. *Methods Mol. Biol.* 401, 235–266.

Fox, P. T., Lancaster, J. L., Parsons, L. M., Xiong, J.-H., and Zamarripa, F. (1997). Functional volumes modeling: theory and preliminary assessment. *Hum. Brain Mapp.* 5, 306–311.

Fox, P. T., Mikiten, S., Davis, G., and Lancaster, J. L. (1994). BrainMap: a database of human function brain mapping. In Functional Neuroimaging: Technical Foundations, Chap. 9, R. W. Thatcher, M. Hallett, T. Zeffiro, E. R. John, and M. Huerta, eds (San Diego, CA, Academic Press), pp. 95–105.

Gansner, E. R., and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Softw. Pract. Exp.* 30, 1203–1234.

Gardner, D., Akil, H., Ascoli, G., Bowden, D. M., Bug, W., Donohoe, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., Halavi, M., Kennedy, D. N., Marenco, L., Martone, M. E., Miller, P. L., Müller, H.-M., Robert, A., Shepherd, G. M., Sternberg, P. W., Van Essen, D. C., and Williams, R. W. (2008). The Neuroscience Information Framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 6, 149–160.

Gardner, D., and Shepherd, G. M. (2004). A gateway to the future of neuroinformatics. *Neuroinformatics* 2, 271–274.

Gustafson, C., Bug, W. J., and Nissanov, J. (2007). NeuroTerrain – a client–server system for browsing 3D biomedical image data. *BMC Bioinformatics* 8, 40.

Hedges, L. V., and Olkin, I. (1985). Statistical Methods for Meta-Analysis. Orlando, FL, Academic Press.

Hunkin, N. M., Mayes, A. R., Gregory, L. J., Nicholas, A. K., Nunn, J. A., Brammer, M. J., Bullmore, E. T., and Williams, S. C. R. (2002). Novelty-related activation within the medial temporal lobes. *Neuropsychologia* 40, 1456–1464.

ISO/IEC (1997). Information Technology – Computer Graphics And Image Processing – The Virtual Reality Modeling Language (VRML) – Part 1: Functional Specification and UTF-8 Encoding. International Organization for Standardization/International Electrotechnical Commission. International Standard ISO/IEC 14772–1.

Jenssen, T.-K., Læreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28.

Kennedy, D. N., Haselgrove, C., and McInerney, S. (2003). MRI-based morphometric analysis of typical and atypical brain development. *Ment. Retard. Dev. Disabil. Res. Rev.* 9, 155–160.

Laird, A. R., Eickhoff, S. B., Kurth, F., Fox, P. M., Uecker, A. M., Turner, J. A., Robinson, J. L., Lancaster, J. L., and Fox, P. T. (2009). ALE meta-analysis workflows via the BrainMap database: progress towards a probabilistic functional brain atlas. *Front. Neuroinform.* 3:23. doi: 10.3389/neuro.11.023.2009.

Laird, A. R., Lancaster, J. L., and Fox, P. T. (2005). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 3, 65–78.

Lancaster, J. L., Woldorff, M. G., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., and Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Hum. Brain Mapp.* 10, 120–131.

Lee, J. T., Munch, K. R., Carlis, J. V., and Pardo, J. V. (2008). Internet image viewer (iiV). *BMC Med. Imaging* 8, 10.

Lewis, S., and Clark, M. (2001). Forest plots: trying to see the wood and the trees. *BMJ* 322, 1479–1480.

Moore, E. B., Poliakov, A. V., Lincoln, P., and Brinkley, J. F. (2007). MindSeer: a portable and extensible tool for visualization of structural and functional neuroimaging data. *BMC Bioinformatics* 8, 389.

Naud, A., Usui, S., Ueda, N., and Taniguchi, T. (2007). Visualization of documents and concepts in neuroinformatics with the 3D-SE viewer. *Front. Neuroinformatics* 1, 7.

Nielsen, F. Å. (2002). Neuroinformatics in Functional Neuroimaging. PhD Thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby (IMM-PHD-2001-87).

Nielsen, F. Å. (2003). The Brede database: a small database for functional neuroimaging. *Neuroimage* 19, 2. Presented at the 9th International Conference on Functional Mapping of the Human Brain, June 19–22, 2003, New York, NY. Available on CD-Rom.

Nielsen, F. Å. (2005). Mass meta-analysis in Talairach space. In Advances in Neural Information Processing Systems 17, L. K. Saul, Y. Weiss, and L. Bottou, eds (Cambridge, MA, MIT Press), pp. 985–992.

Nielsen, F. Å. (2009). Brede Wiki: neuroscience data structured in a wiki. In Proceedings of the Fourth Workshop on Semantic Wikis – The Semantic Wiki Web, Volume 464 of CEUR Workshop Proceedings, C. Lange, S. Schaffert, H. Skaf-Molli, and M. Völkel, eds (Aachen, RWTH Aachen University), pp. 129–133.

Nielsen, F. Å., Balslev, D., and Hansen, L. K. (2005). Mining the posterior cingulate: segregation between memory and pain component. *Neuroimage* 27, 520–532.

Nielsen, F. Å., Balslev, D., and Hansen, L. K. (2006a). Data mining a functional neuroimaging database for functional segregation in brain regions. In Den 15. Danske Konference i Mønstergenkendelse og Billedanalyse, S. I. Olsen, ed. (Copenhagen, The Department of Computer Science, University of Copenhagen).

Nielsen, F. Å., Christensen, M. S., Madsen, K. H., Lund, T. E., and Hansen, L. K. (2006b). fMRI neuroinformatics. *IEEE Eng. Med. Biol. Mag.* 25, 112–119.

Nielsen, F. Å., and Hansen, L. K. (1997). Interactive information visualization in neuroimaging. In Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation, D. S. Ebert and C. K. Nicholas, eds (New York, NY, ACM), pp. 62–65.

Nielsen, F. Å., and Hansen, L. K. (1999). Modeling of BrainMap Data. Available at http://isp.imm.dtu.dk/publications/1999/nielsen.nips99.ps.gz.

Nielsen, F. Å., and Hansen, L. K. (2000a). Experiences with Matlab and VRML in functional neuroimaging visualizations. In VDE2000 – Visualization Development Environments, Workshop Proceedings, Princeton, New Jersey, USA, April 27–28, 2000, S. Klasky and S. Thorpe, eds (Princeton, NJ, Princeton Plasma Physics Laboratory), pp. 76–81.

Nielsen, F. Å., and Hansen, L. K. (2000b). Functional Volumes Modeling Using Kernel Density Estimation. Available at http://www.imm.dtu.dk/pubdb/views/edoc_download.php/4688/pdf/imm4688.pdf.

Nielsen, F. Å., Hansen, L. K., and Balslev, D. (2004). Mining for associations between text and brain activation in a functional neuroimaging database. *Neuroinformatics* 2, 369–380.

Phillips, M. L., Bullmore, E. T., Howard, R., Woodruff, P. W., Wright, I. C., Williams, S. C., Simmons, A., Andrew, C., Brammer, M., and David, A. S. (1998). Investigation of facial recognition memory and happy and sad facial expression perception: an fMRI study. *Psychiatry Res.* 83, 127–138.

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C., Gray, J. A., and David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature* 389, 495–498.

Poliakov, A. V., Albright, E., Hinshaw, K. P., Corina, D. P., Ojemann, G., Martin, R. F., and Brinkley, J. F. (2005). Server-based approach to web visualization of integrated three-dimensional brain imaging data. *J. Am. Med. Inform. Assoc.* 12, 140–151.

Rehm, K., Lakshminarayan, K., Frutiger, S. A., Schaper, K. A., Sumners, D. L., Strother, S. C., Anderson, J. R., and Rottenberg, D. A. (1998). A symbolic environment for visualizing activated foci in functional neuroimaging datasets. *Med. Image Anal.* 2, 215–226.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E. (2009a). Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *Neuroimage* 45, 810–823.

Salimi-Khorshidi, G., Smith, S. M., and Nichols, T. E. (2009b). Bias and heterogeneity in neuroimaging meta-analysis. In 15th Annual Meeting of the Organization for Human Brain Mapping Abstracts Online. 406 SA-PM.

Talairach, J., and Tournoux, P. (1988). Co-planar Stereotaxic Atlas of the Human Brain. New York, Thieme Medical Publisher Inc.

Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16(Pt 1), 765–780.

Van Essen, D. C., and Dierker, D. L. (2007). Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron* 56, 209–225.

Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., and Snellenberg, J. X. V. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* 1(Suppl. 1), S210–S221.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2009; paper pending published: 01 May 2009; accepted: 10 July 2009; published online: 28 July 2009.
*Citation: Nielsen FA (2009) Visualizing data mining results with the Brede tools. Front. Neuroinform. (2009) **3**:26. doi: 10.3389/neuro.11.026.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 09 July 2009
doi: 10.3389/neuro.11.023.2009

ALE meta-analysis workflows via the BrainMap database: progress towards a probabilistic functional brain atlas

Angela R. Laird[1]*, Simon B. Eickhoff[2,3,4], Florian Kurth[2], Peter M. Fox[1], Angela M. Uecker[1], Jessica A. Turner[5], Jennifer L. Robinson[6], Jack L. Lancaster[1] and Peter T. Fox[1]

[1] *Research Imaging Center, University of Texas Health Science Center, San Antonio, TX, USA*
[2] *Institute for Neuroscience and Medicine (INM – 2), Research Center Jülich, Jülich, Germany*
[3] *Department of Psychiatry and Psychotherapy, RWTH Aachen University, Aachen, Germany*
[4] *Jülich Aachen Research Alliance–Translational Brain Medicine, Germany*
[5] *Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA*
[6] *Neuroscience Institute, Scott and White Memorial Hospital, Temple, TX, USA*

Edited by:
John Van Horn, University of California, USA

Reviewed by:
Ivo Dinov, University of California, USA
John Van Horn, University of California, USA

***Correspondence:**
Angela R. Laird, Research Imaging Center, University of Texas Health Science Center San Antonio, 7703 Floyd Curl Drive, San Antonio, TX 78229-3900, USA.
e-mail: lairda@uthscsa.edu

With the ever-increasing number of studies in human functional brain mapping, an abundance of data has been generated that is ready to be synthesized and modeled on a large scale. The BrainMap database archives peak coordinates from published neuroimaging studies, along with the corresponding metadata that summarize the experimental design. BrainMap was designed to facilitate quantitative meta-analysis of neuroimaging results reported in the literature and supports the use of the activation likelihood estimation (ALE) method. In this paper, we present a discussion of the potential analyses that are possible using the BrainMap database and coordinate-based ALE meta-analyses, along with some examples of how these tools can be applied to create a probabilistic atlas and ontological system of describing function–structure correspondences.

Keywords: BrainMap, meta-analysis, activation likelihood estimation, ontology, functional atlas

INTRODUCTION

Over the last three decades, neuroimaging research has produced an enormous amount of data localizing the neural effects of specific mental operations in both healthy and diseased populations. Community-wide standards of spatial normalization and the reporting of peak activation locations in stereotactic coordinates allow researchers to compare results across studies when the primary data are unavailable or difficult to obtain. Due to the nearly universal adherence to these standards, the BrainMap project was designed to create tools for large-scale data mining and meta-analysis of the brain mapping literature (Fox and Lancaster, 2002; Laird et al., 2005a).

BrainMap is a community accessible database[1] that allows a user to relate behavioral functions to specific brain locations through retrieval and visualization of peak coordinates and their associated metadata. These metadata allow each coordinate to be linked with how the observed activation was experimentally derived, a formulation that lends itself to very rich data mining. BrainMap was originally conceived by Peter Fox in 1987 and received its original funding from the James S. McDonnell Foundation (1988–1990). Continued BrainMap development was funded by the Office of Naval Research (1991–1992), the EJLB Foundation (1992–1996), and the National Library of Medicine (2000–2003). BrainMap is currently funded by the Human Brain Project of the National Institute of Mental Health.

BrainMap SOFTWARE

There are three desktop applications (*Scribe*, *Sleuth*, and *GingerALE*) and one web application (*BrainMapWeb*) that allow interaction with the BrainMap database, all of which are coded in Java. The desktop applications run in the Java Runtime Environment on Macintosh, Windows, Linux, and Unix operating systems, while the web application uses Java server-side technologies. *Scribe*[2] is used to code papers for entry into BrainMap. Peer-reviewed publications can be submitted to the database by the original authors (uncommon) or by investigators performing a meta-analysis (very common). Most data fields have candidate responses presented in scrollable lists. Coordinate tables of peak locations can be imported from a tab-delimited file or entered by hand. Upon insertion into the database, each *x*,*y*,*z*, coordinate is assigned an anatomical location using the Talairach Daemon[3] (Lancaster et al., 2000). All entries are reviewed for quality control by BrainMap staff and faculty before being entered into the database to ensure the accuracy and consistency of coding. The *Sleuth* application[4] allows a user to search the BrainMap database and retrieve data, which can then be filtered and visualized on a standard Talairach atlas image. Specific locations may be searched for according to user-defined, coordinate-based regions of interest (defined by Talairach or MNI coordinates) or anatomical labels from the Talairach Daemon nomenclature. BrainMap queries may also be implemented via an internet browser using *BrainMapWeb*[5], which includes query functions that are similar to those of *Sleuth*, but lack 3D brain visualizations. Data view and manipulation capabilities are much more restricted than

[1]http://brainmap.org

[2]http://brainmap.org/scribe
[3]http://talairach.org/
[4]http://brainmap.org/sleuth
[5]http://brainmap.org/bmapWeb

in *Sleuth*, which can output data in publication-ready graphics, text files of Talairach or MNI coordinates, or workspaces that specify search rules and filters for meta-analyses. By archiving coordinates of activation locations rather than raw image data, BrainMap focuses on discoveries derived from coordinate-based meta-analyses of functional neuroimaging data. The last BrainMap application, *GingerALE*[6], is used for performing activation likelihood estimation (ALE) meta-analyses on sets of coordinates extracted from the database in Talairach or MNI space. *Scribe*, *Sleuth*, and *GingerALE* are closely integrated to transition seamlessly from database submission to search refinement to meta-analysis results (**Figure 1**).

BrainMap CODING SCHEME

To summarize the experimental design and results of a published study for inclusion in the database, BrainMap utilizes a rigorous taxonomy that is composed chiefly of structured keywords. BrainMap entries include descriptions on the scanned subjects and experimental conditions, including the presented stimuli, instructions, and responses. A complete metadata listing can be seen in **Figure 2**. To facilitate meta-analysis, several hierarchically structured keywords have been developed that categorize the nature of each experimental contrast to allow rapid, comprehensive retrieval of results. "Context" broadly categorizes the purpose of the work; for example, normal mapping (the comparison of different experimental conditions in a group of healthy subjects), age effects, disease effects, or drug effects. "Behavioral Domain" classifies the research in terms of the neural systems studied according to six main categories and their related subcategories: cognition, action, perception, emotion, interoception, or pharmacology (**Figure 3**). "Paradigm Class" categorizes the challenge presented, preferably in the jargon of the field, such as anti-saccades, Stroop, delayed match to sample, or mental rotation tasks (**Figure 4**).

While a given a given paradigm class (e.g., n-back) is often immediately associated with a given behavioral domain (e.g., working memory), this is not always the case. Each experiment must be evaluated based on the conditions contrasted since many comparisons are designed to elicit processing in domains not directly linked to the paradigm class employed. For example, n-back and working memory are appropriate choices when comparing an n-back condition to a control condition, but if two n-back conditions are contrasted that differ only in the modality of stimulus presentation (e.g., visual or auditory), then additional perceptual domains should be coded for that experiment. We have found that the context, behavioral domain, and paradigm class represent the three most critical components of a functional neuroimaging study; their orthogonality fully defines and gives contextual meaning to the coordinates archived in the database.

Typically, investigators use citation indexing services such as PubMed to search the literature according to user-defined keywords. This results in the identification of a subset of desired studies that explicitly match the search criteria. BrainMap's strategy has been to design paradigm class entries in order to pool similar studies, rather than segregate them according to domain-specific keywords. Categories and sub-categories are created and refined only as needed, based on the demands of the literature and the continued development of functional brain imaging. For example, the first study entered into BrainMap that utilized the Wisconsin Card Sorting Test was initially coded under the paradigm class of "Deductive Reasoning". This was done to classify the study in the same set as other similar papers in the database (i.e., "best fit"), a practice that the classes are rich enough to be useful. Once four papers of the same sub-category are entered into the database, a new class is created and defined (e.g., "Wisconsin Card Sorting Test"). Then all entries matching this new class are manually searched for and updated to reflect the new designation. This procedure requires continuous and labor-intensive maintenance of the database, yet yields a high-quality database full of rich metadata categories and provides an evolving and flexible tool.

Given the sheer volume of neuroimaging data that is currently being produced, it is rapidly becoming overwhelming for an investigator to reconcile new results to those previously published, particularly when studies pertain to different areas of research. Derrfuss and Mar (2009) estimated that BrainMap contains approximately one-fifth

[6]http://brainmap.org/ale

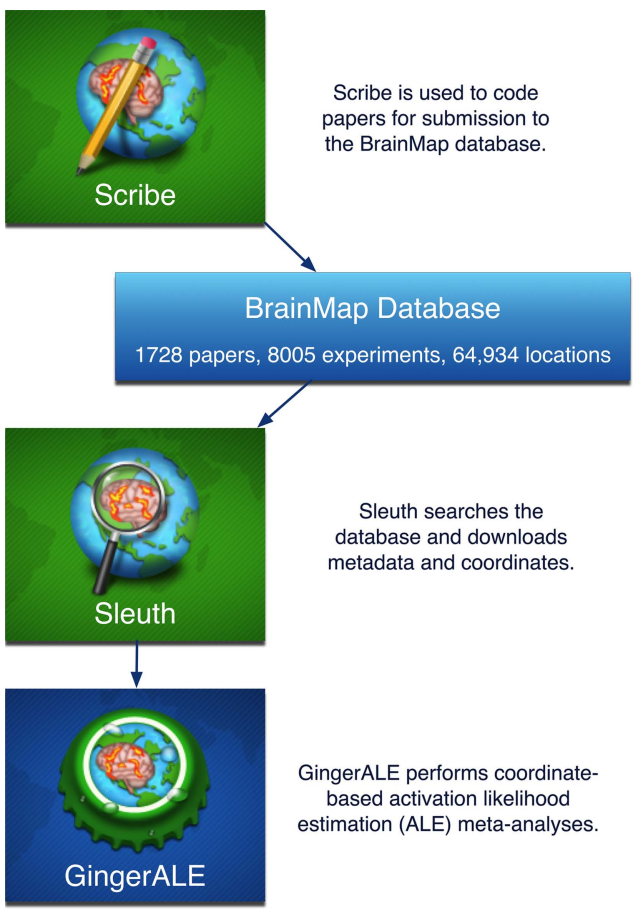


FIGURE 1 | BrainMap software applications. There are three desktop applications that allow interaction with the BrainMap database: *Scribe* (data entry), *Sleuth* (search and retrieval), and *GingerALE* (meta-analysis). These Java applications are freely downloadable from the BrainMap website (http://brainmap.org). Database statistics are current as of 04/01/09.

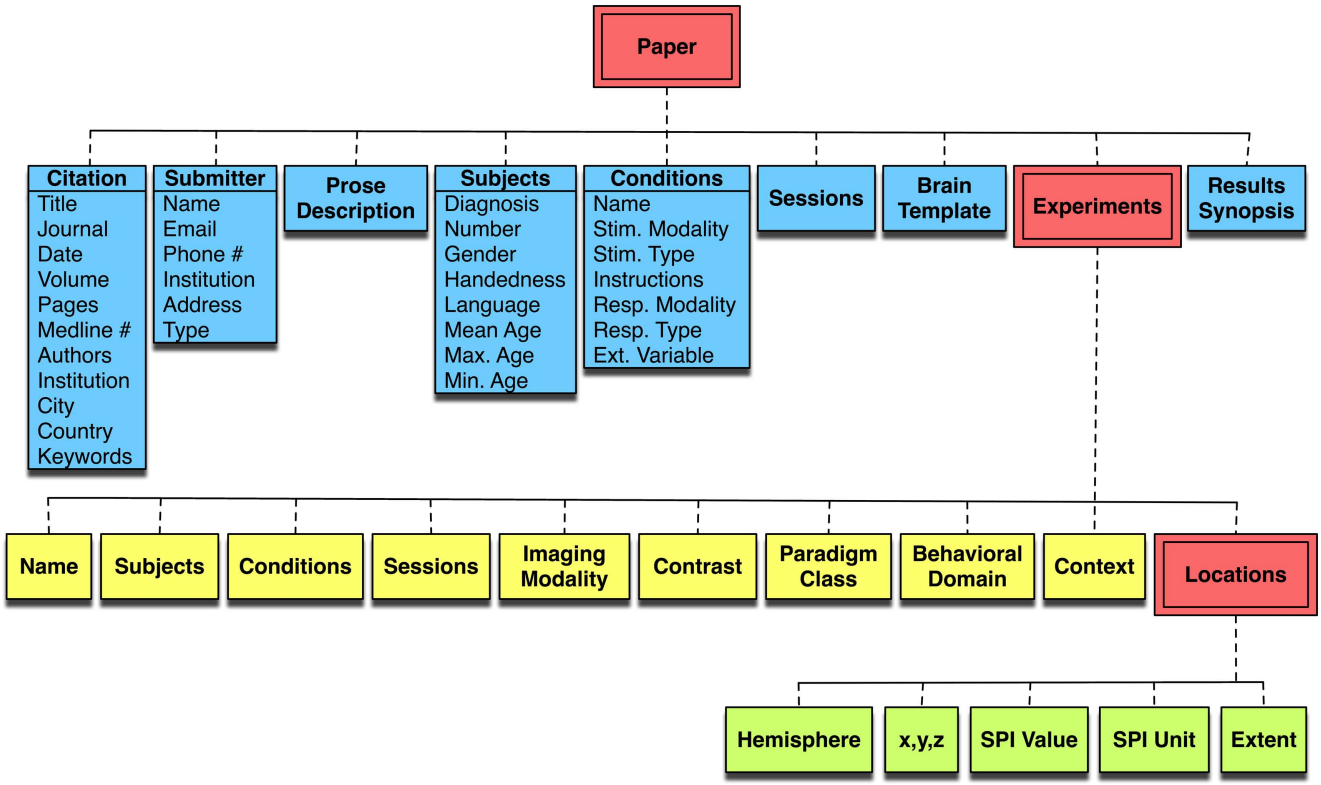


FIGURE 2 |The BrainMap coding scheme. The BrainMap metadata coding scheme follows a hierarchy naturally occurring within the brain mapping literature. Every *Paper* reports experimental results drawn from one or more subject groups, the members of which have been functionally imaged during one or more behavioral *Conditions*. In BrainMap, an *Experiment* is defined by the production of a statistical parametric image. From each *Experiment* one or more functional activations (*Locations*) are extracted in the form of peak coordinates.

of the relevant published studies, making it the largest coordinate-based database in functional neuroimaging to date. In BrainMap, an ROI of 1 cm^3 currently contains an average of 23 experiments, and includes results from 15 paradigm classes. Databases designed to simply retrieve studies reporting activations in proximate locations result in subsequent manual filtering and interpretation. While BrainMap's data entry procedure is labor-intensive (Laird et al., 2005a), the depth of the current coding strategy is what provides diverse data mining opportunities and establishes the overall value of the database. BrainMap was structured with the goal of not only retrieving studies returned by regional searches without domain-specific biases, but also allowing the results to be synthesized. Specifically, BrainMap development of neuroinformatics tools focuses on knowledge discovery that is made possible by coordinate-based function–location meta-analysis (Fox et al., 1998).

ACTIVATION LIKELIHOOD ESTIMATION META-ANALYSIS

Activation likelihood estimation (ALE) is a method of coordinate-based meta-analysis that is supported within the BrainMap software environment. It is a useful tool for integrating the neuroimaging literature wherein consistent regions of activation are identified across a collection of studies. In particular, peak coordinates are collected from studies that share a similar feature of interest, which can be a specific task (e.g., go/no-go) or a more generalized cognitive process (e.g., inhibition). In ALE, coordinates are then modeled with a Gaussian function to accommodate the spatial uncertainty associated with a reported coordinate and are analyzed for where they converge.

Since its introduction (Chein et al., 2002; Turkeltaub et al., 2002), ALE has been applied in many aspects of normal brain function (Costafreda et al., 2008; Decety and Lamm, 2007; Eickhoff et al., 2006a; Grosbras et al., 2005; Soros et al., 2009; Spreng et al., 2009), as well as in studies of neuropsychiatric and neurological disorders, such as schizophrenia (Glahn et al., 2005; Minzenberg et al., 2009; Ragland et al., 2009), obsessive-compulsive disorder (Menzies et al., 2008), depression (Fitzgerald et al., 2008), and developmental stuttering (Brown et al., 2005). Recently, ALE has been extended to voxel-based morphometry (Ellison-Wright et al., 2008; Glahn et al., 2008; Schroeter et al., 2007) and diffusion tensor imaging studies (Ellison-Wright and Bullmore, 2009).

The most interesting ALE applications do not merely merge previous results in a retrospective fashion, but instead generate or test a new hypothesis (Eickhoff et al., 2009a; Price et al., 2005), identify a previously unspecified region (Derrfuss et al., 2005), resolve conflicting views (Laird et al., 2005b; Petacchi et al., 2005), or validate a new paradigm (McMillan et al., 2007). Several studies have used ALE as a preliminary step, followed by an analysis of network co-occurrences (Lancaster et al., 2005; Neumann et al., 2005, 2008; Toro et al., 2008) or structural equation modeling

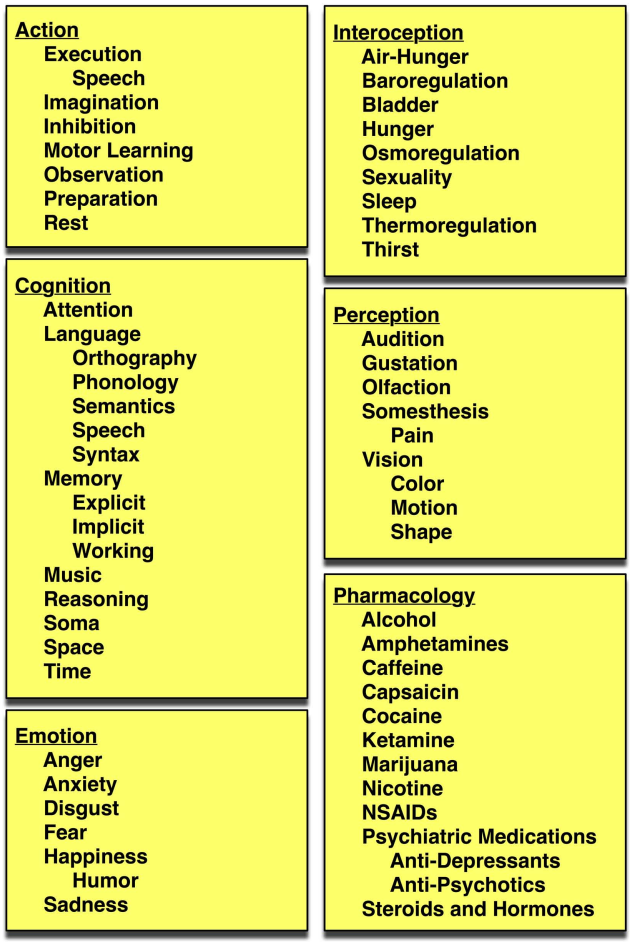


FIGURE 3 | BrainMap behavioral domains. In the BrainMap coding scheme, the behavioral domain is recorded for each statistical contrast, which describes the research in terms of the neural systems studied according to 6 main categories: cognition, action, perception, emotion, interoception, or pharmacology. The yellow background represents that this field lies at the *Experiment* level.

FIGURE 4 | BrainMap paradigm classes. The paradigm class is recorded for each statistical contrast in BrainMap. This field categorizes the task or challenge presented, preferably in the jargon of the field. BrainMap currently contains 80 paradigm classes. The yellow background represents that this field lies at the *Experiment* level.

(Laird et al., 2008). Each of these novel meta-analytic applications was carried out using typical ALE analysis parameters, and many of them involved the comparison of multiple meta-analyses. For example, Price et al. (2005) examined the results of a picture naming meta-analysis and found that use of a high-level baseline condition to control for speech production and perceptual processing resulted in increased sensitivity to activation in areas associated with semantic processing, visual-speech integration, and response selection. A prospective fMRI study was performed to test this hypothesis, which subsequently allowed the picture naming system to be decomposed into its perceptual, semantic, and phonological components. In a different application of the ALE method, a meta-analysis was performed on studies in which transcranial magnetic stimulation was applied to left motor cortex (Laird et al., 2008). The results of this meta-analysis were used to determine the location of regions of interest in a prospective study of TMS/PET data that examined the effective connectivity of the motor system using structural equation modeling. These examples of how meta-analysis results have been applied to guide analyses in newly acquired experimental data demonstrate the power of the ALE method and provide evidence of its efficacy beyond that of a purely retrospective tool.

MODIFICATIONS TO THE ALE ALGORITHM

The ALE method was originally developed and validated by Turkeltaub et al. (2002) in a meta-analysis of single word reading. BrainMap developers obtained the algorithm from the Georgetown University CSL group, ported the code into Java, and created a graphical user interface (*GingerALE*). A cluster analysis script was added that identifies ALE clusters (areas of high activation likelihood) and returns the cluster extent above a user-specified threshold, *x-y-z* coordinates of the weighted center-of-mass and peak locations, and an anatomical label assigned by the Talairach Daemon (Lancaster et al., 2000). A coordinate conversion utility was also included to convert MNI coordinates to Talairach space (Lancaster et al., 2007). Two extensions of the original ALE method, a correction for multiple comparisons and a method for computing statistical contrasts of pairs of ALE images, were also added (Laird et al., 2005c).

In the original implementation of the ALE method, several limitations were known to exist: (1) the size of the modeled Gaussian distribution was user-specified and therefore subjective, (2) the permutation test for significance was not anatomically constrained, leading to some modeled activation in white matter, and (3) the analysis tested for above-chance clustering of individual coordinates (a fixed-effects analysis), preventing the generalization of results that is possible in a random-effects analysis (Wager et al., 2007, 2009). Recent advances in the ALE technique have overcome these limitations to provide a more valid and statistically reliable meta-analysis framework (Eickhoff et al., 2009b). Rather than relying on user-dependent Gaussian distributions, quantitative estimates of the between-subject and between-template variability were empirically determined in order to more explicitly model the spatial uncertainty associated with each coordinate (a correction that also includes a weighting of each study by the number of included subjects). In addition, the permutation test was limited to regions of gray matter and modified to test for the above-chance clustering between experiments, resulting in a transition from a fixed-effects to a random-effects method of statistical inference. By progressing from an analysis based on the clustering across coordinates to the clustering across experiments, ALE results no longer may potentially be driven by a single study. The new ALE formulation was validated against the classical algorithm and experimental data (Grefkes et al., 2008) and found to increase the specificity of results without losing the sensitivity of the original approach. These improvements have been implemented in the most recent version of *GingerALE*, which is currently available for beta testing on the BrainMap website.

FUNCTION-BASED META-ANALYSES

The ALE meta-analysis method can be applied in a variety of ways to answer specific research questions. Most frequently, ALE is applied to sets of neuroimaging studies that share some similar aspect of evoked brain function. These function-based meta-analyses usually involve pooling studies with similar experimental designs, and these studies may also be segregated into different collections to evaluate the functional specificity of the task or process being investigated. In the BrainMap coding scheme, experimental conditions are described according to the presented stimuli, instructions given, and response requested. Each of these fields has candidate entries to choose from, which collectively capture the essence of the scanned conditions (**Figure 5**). These three condition axes (stimulus, response, and instructions) provide a structure in which differential activation patterns can be systematically probed for variations of a given task.

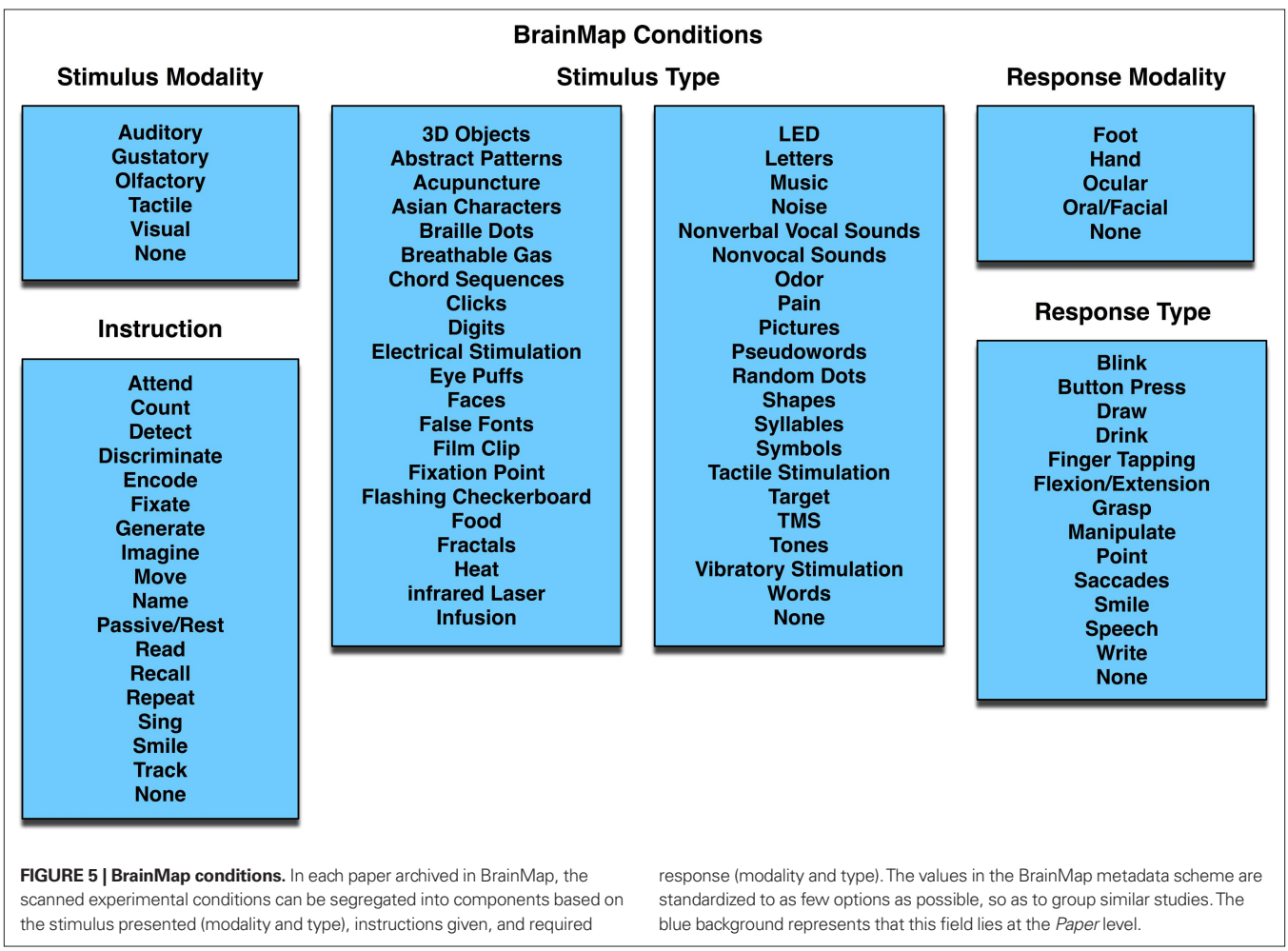


FIGURE 5 | BrainMap conditions. In each paper archived in BrainMap, the scanned experimental conditions can be segregated into components based on the stimulus presented (modality and type), instructions given, and required response (modality and type). The values in the BrainMap metadata scheme are standardized to as few options as possible, so as to group similar studies. The blue background represents that this field lies at the *Paper* level.

For example, we performed an ALE meta-analysis of all studies in BrainMap that were coded with a paradigm class of "word generation" (66 papers, 197 experiments, 1552 coordinates), a widely used test of neuropsychological function. The meta-analytic results revealed extensive convergence in large portions of the left inferior frontal gyrus, centering on Brodmann area 44/45 (Broca's area), and the left dorsolateral prefrontal cortex (DLPFC, BA 46/9), regions commonly known to be associated with word retrieval and executive function. ALE clusters were also observed in the bilateral insula, anterior cingulate cortex (ACC, BA 32), supplementary motor area (SMA, BA6), precuneus (BA 7), posterior cingulate cortex (PCC, BA 31), left posterior temporal cortex (Wernicke's area, BA 22), left inferior parietal cortex (BA 40), right posterior cerebellum, and left thalamus. These areas are generally understood to be involved with the production of language and executive processing that is characteristic of verbal fluency tasks (Heim et al., 2008; Petersen et al., 1988; Warburton et al., 1996; Wise et al., 1991).

We then examined BrainMap metadata for these studies and found that the imaged tasks varied according to the modality of responses (covert or overt), the modality of stimulus presentation (visual or auditory), or the stimulus type (words or letters). Supplemental ALE meta-analyses were then performed according to each of these task variations, yielding differential patterns of activation likelihood (**Figure 6**). Covert word generation yielded more extensive engagement of lateral and medial prefrontal areas (similar to Basho et al., 2007), precuneus, and posterior cingulate cortex, while analysis of overt tasks revealed distinct concordance in the right cerebellum. ALE results of studies using visual stimuli were observed in visual cortex (BA 17/18) and the fusiform gyrus (BA 37), while auditory stimuli were localized to left auditory cortex (BA 41), Wernicke's area (BA 22), and precuneus. Generation of words in response to a presented word (visual or auditory) was associated with the middle frontal gyrus (BA 9). Semantic verbal fluency was also associated with the ventral portion of the left inferior

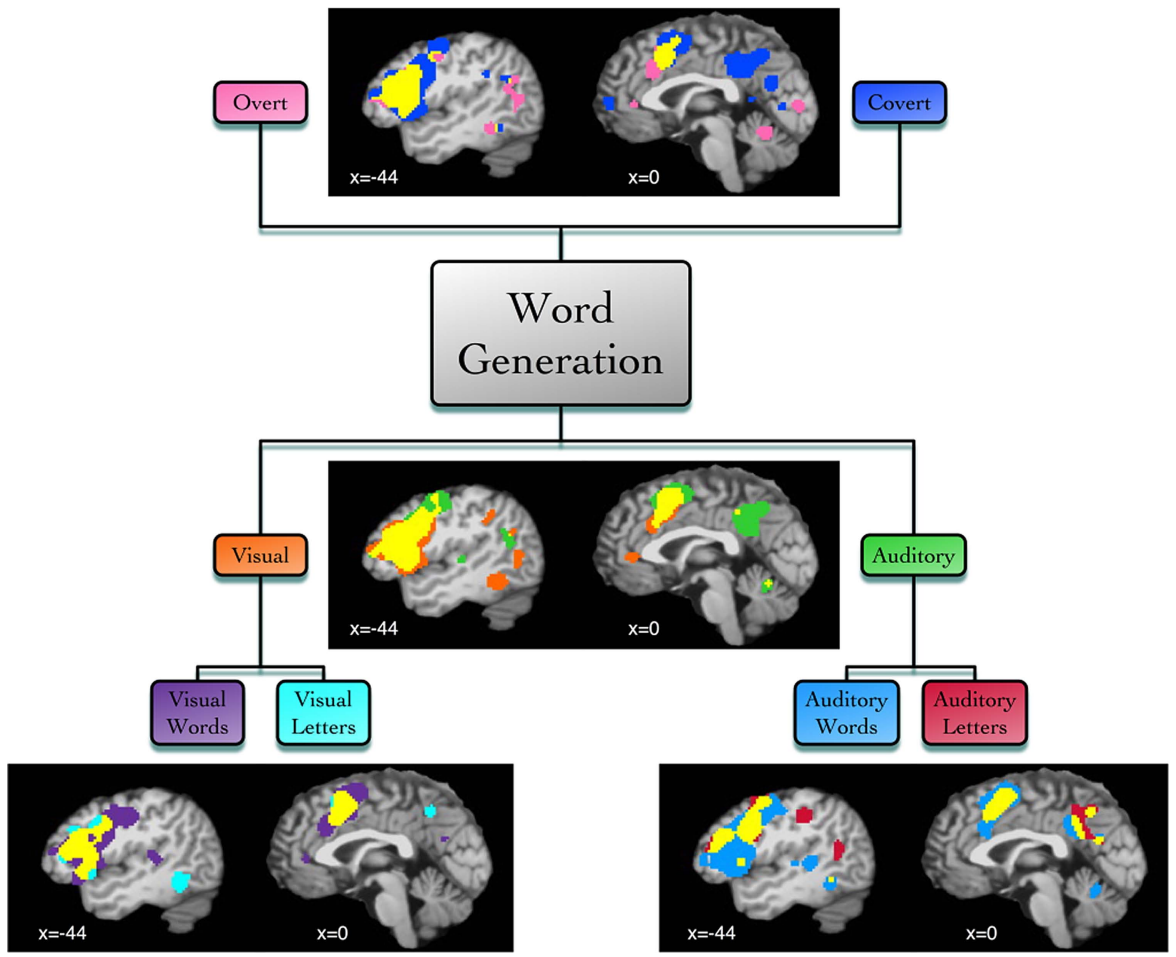


FIGURE 6 | Function-based meta-analysis of word generation. Studies utilizing a word generation paradigm were downloaded from the BrainMap database. Papers were segregated according to stimulus and response features. Separate ALE meta-analyses were performed for each set of coordinates: overt speech responses (30 papers, 78 experiments, 604 coordinates), covert speech responses (35 papers, 92 experiments, 729 coordinates), visual stimuli (39 papers, 115 experiments, 906 coordinates), auditory stimuli (32 papers, 86 experiments, 621 coordinates), visual words (25 papers, 88 experiments, 638 coordinates), visual letters (14 papers, 38 experiments, 373 coordinates), auditory words (39 papers, 124 experiments, 1001 coordinates), and auditory letters (14 papers, 35 experiments, 279 coordinates). Results are displayed at $P < 0.05$, corrected. Left lateral ($x = -44$) and medial ($x = 0$) sagittal slices display the comparative meta-analytic results. In all images, the overlap between meta-analysis results is shown in yellow.

frontal gyrus, in agreement with previous meta-analysis results (Costafreda et al., 2006), as well as BA 45 (Amunts et al., 2004). The observation that Wernicke's area was preferentially involved auditory stimulation, particularly with words, likely reflects this region's role in auditory processing and the comprehension of spoken words. This word generation meta-analysis highlights the rich data mining that is possible using the BrainMap coding scheme.

STRUCTURE-BASED META-ANALYSES

While function-based meta-analyses tend to dominate the literature, structure-based meta-analyses offer an alternative meta-analytic strategy. Instead of pooling studies that share a common experimental design, structure-based meta-analyses focus on a specific anatomical region and look for global coactivation patterns across a diverse range of tasks. The theory behind this type of meta-analysis is that groups of coordinates that coactivate across experiments can be pooled to identify functionally connected networks in the brain. Like other methods of analyzing functional connectivity (Cordes et al., 2000; Rogers et al., 2007; Xiong et al., 1999), structure-based meta-analyses are based on the co-occurrence of spatially separate neurophysiological events. Koski and Paus (2000) used this technique to study the meta-analytic connectivity of the anterior cingulate cortex, although their analysis was limited to the frontal lobe. In their study, the authors manually collected and filtered data from 107 studies tasks to examine regional co-occurrences and found evidence for functional heterogeneity within the ACC. A similar meta-analysis was performed on basal ganglia activation using 126 published studies to determine the functional connectivity between cortex and striatum (Postuma and Dagher, 2006).

Increasing both the size and diversity of structure-based meta-analyses adds to the generalizability of the results. When used in conjunction with the BrainMap database, the procedure is more automated, resulting in larger meta-analyses that include decades of neuroimaging data from a diverse range of paradigms and behavioral domains. Recently, a large-scale meta-analysis of the functional connectivity of the amygdala was carried out in which the ROIs for left and right amygdala were defined according the Harvard-Oxford structural probability atlas distributed with FSL (Smith et al., 2004) and seeded in BrainMap (Robinson et al., 2009). This anatomically defined meta-analysis of 240 papers (326 experiments with 3842 coordinates) revealed that the amygdala plays a integrative role in both emotion and cognition, and was validated according to the non-human primate database, CoCoMac (Stephan et al., 2001).

To illustrate the use of structure-based meta-analyses, we note that the function-based meta-analysis of verbal fluency studies (**Figure 6**) was characterized by extensive activation of the left inferior frontal gyrus, centered on Broca's area in BA 44 and 45. To determine what connectivity differences exist between these two cytoarchitectonic regions, a location query was performed within BrainMap for studies activating these regions, using ROIs of the cytoarchitectonically defined areas (Amunts et al., 1999) as distributed with the SPM Anatomy Toolbox (Eickhoff et al., 2005, 2006b). The returned studies were then analyzed in a structure-based meta-analysis of both left BA 44 and BA 45 (**Figure 7**). For both regions, extensive bilateral connectivity was observed across the inferior frontal gyrus, precentral gyrus, inferior parietal lobule, fusiform gyrus, and insula, as well as medial ACC and SMA.

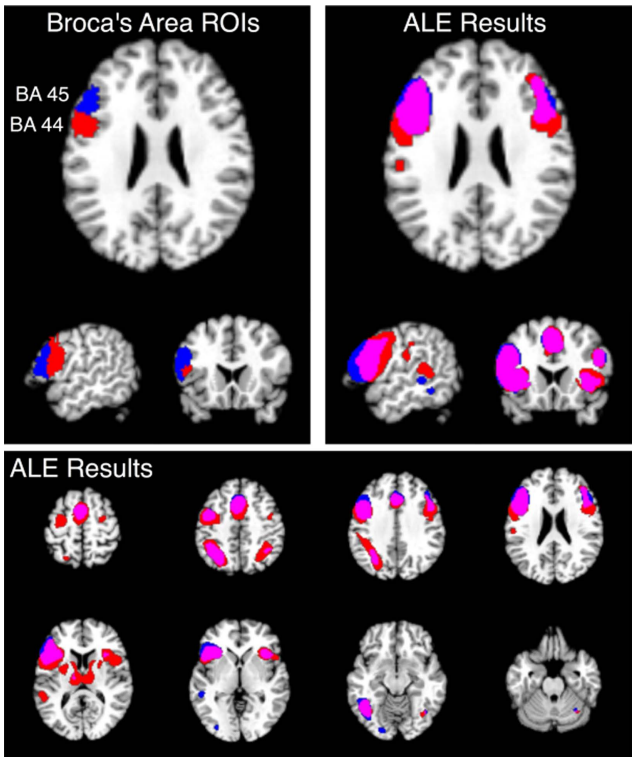


FIGURE 7 | Structure-based meta-analysis of Brodmann areas 44 and 45. ALE meta-analyses were performed for all experiments in BrainMap that reported activation in either BA 44 or BA 45 to determine the meta-analytic functional connectivity of these cytoarchitectonically-defined regions. Results are displayed at $P < 0.05$, corrected. The top slices (left and right) are centered at $x = -54$, $y = 18$, $z = 24$. The upper left slices display the ROIs used to search BrainMap for BA 44 (red) and BA 45 (blue), which were obtained from the SPM Anatomy Toolbox (Eickhoff et al., 2005; Eickhoff et al., 2006b). ALE results for these regions are shown on the top right in axial, sagittal, and coronal slices, as well as on the bottom panel in axial slices from $z = 58$ to $z = -26$. The overlap between meta-analyses for BA 44 and 45 is shown in purple.

Joint connectivity was also observed in the left thalamus and right cerebellum. In contrast to the word generation meta-analysis, no connectivity was observed in the precuneus or posterior cingulate cortex, likely reflecting a memory retrieval component (Cabeza and Nyberg, 2000) of verbal fluency processing. Much overlap was observed between the two images; however, comparison of the maps for BA 44 and BA 45 revealed strong dissociation in subcortical regions. BA 44 exhibited extensive connectivity with bilateral thalamus, caudate, and putamen in agreement with Eickhoff et al. (2009a), while only the left thalamus was observed for BA 45. In addition, only the BA 44 map returned connectivity in Wernicke's area, reflecting preferential engagement of this region in comparison to BA 45. Structure-based meta-analyses can therefore find distinct differences in functionally connected networks even for regions that lie very close to each other, such as BA 44 and 45.

Toro et al. (2008) expanded upon the idea of structure-based meta-analyses and developed an algorithm to test the likelihood of a functional connection between regions, yielding a 3D

meta-coactivation map for every voxel in the brain. In their analysis of 825 papers (3402 experiments with 27,909 coordinates), Toro et al. observed distinct and recognizable functional networks that are commonly associated with processes such as attention, motor function, and the resting state. Given that meaningful networks were extracted from the coordinates contained in BrainMap via a coactivation analysis, a recent study proposed that known functional networks of the brain during explicit activation could be derived using independent component analysis (ICA) of BrainMap data (Smith et al., 2009). These networks, when compared to resting state networks (RSNs) obtained by ICA of resting state fMRI data (Damoiseaux et al., 2006) were virtually identical. That is, the set of major covarying activation networks identified from a massive-scale meta-analysis (1687 papers, 7342 experiments, 58,620 coordinates) matched the set of networks that are present in the resting brain. These results provide strong evidence that RSNs reflect functional neural networks, and that these dynamic networks are engaged even at rest (Fox and Raichle, 2007). Given the independent nature of these two analyses on fundamentally different types of data, as well as the heterogeneity of data contained in BrainMap due to differences in subjects, scanners, analyses, and paradigms, it is remarkable that such strong correspondence was observed between the resting state and meta-analytic results. In sum, this study supports the validity of using BrainMap and coordinate-based meta-analyses to identify functional neural networks on a large scale.

MAPPING FUNCTION–STRUCTURE RELATIONSHIPS IN THE BRAIN

One of the broad goals of functional neuroimaging research is to determine function–structure relationships in the brain. A concrete deliverable of this aim is a probabilistic functional atlas, in which specific mental operations are mapped to discrete networks of brain regions. Price and Friston (2005) point out that the relationship between a brain region and a mental function is not a one-to-one mapping. Instead this relationship is a many-to-many mapping, as a single region can be involved in many cognitive processes, and a single process usually activates multiple regions. Evaluating these mappings will require collating the immense amount of neuroimaging data that has been acquired thus far and continuing the development of advanced meta-analytic techniques in order to efficiently and effectively synthesize all of these data.

As the development of comprehensive neuroinformatics tools progresses, the need for comprehensive data ontologies increases. An ontology is a machine-interpretable description of concepts and their relationships with the purpose of sharing of ideas and information in a manner facilitated by semantic interoperability (Stevens et al., 2000). Until a foundational ontology for neuroimaging is established and adopted, the communication within and between databases will be limited, hindering the creation of a functional brain atlas. In the field of neuroimaging, ontology development is proceeding rapidly in the domains of representing neuroanatomical findings [e.g., NeuroNames, Bowden and Dubach, 2003; Bowden and Martin, 1995 and the Foundational Model of Anatomy (FMA), Rosse and Mejino, 2003], describing imaging acquisition strategies (e.g., RadLex, Langlotz, 2006; Rubin, 2008), and identifying clinical assessments [e.g., the Systematized Nomenclature of Medicine (SNOMED), Coté and Robboy, 1980]. In addition, the Neuroscience Information Framework Standardized (NIFSTD) Ontology, developed by BIRN and the NIF, is a collection of these and other neuroscience ontologies (Bug et al., 2008).

There is currently no accepted ontology for describing the range of mental operations performed by the human brain, although need for such an ontology is increasingly being discussed (Binder et al., 2009; Poldrack, 2006; Price and Friston, 2005; Toga, 2002). A research question such as "find the data examining the relationship between hippocampal volume in Alzheimer's disease" requires knowledge about clinical diagnosis, neuroanatomy, and MR scan acquisition. As mentioned previously, there are ongoing ontological efforts designed to curate knowledge in these domains. However, a query such as "find the data examining neural activations observed during sustained attention" involves knowledge about cognitive processing, for which no ontology exists to date. Creating a realistic functional brain atlas will require a systematic description of mental operations that are reported in the literature so that informative and standardized labels can be applied to different brain networks.

Investigators frequently utilize alternate and sometimes competitive terminologies when referring to the cognitive processes elicited by specific tasks. When different words are employed to represent the same concepts, the grouping of related ideas across different resources is impeded. For example, a semantically interoperable ontology will allow the linking of data designated with terms such as "declarative" and "explicit" memory types and be capable of relating "working memory" to either "memory" or "executive processing". Cognition represents the most difficult domain to explicate as it contains the most intricate of all concepts, such as language, attention, and memory. But while much imaging research focuses on cognitive processing, important results are also being published in the areas of perception, action, emotion, and autonomic functions. To comprehensively describe the many-to-many mappings of structure and function being investigated in neuroimaging research, a complete mental ontology must be developed.

Poldrack (2008) suggests that the many-to-many mapping dilemma may be complicated by our current understanding of what cognitive processes exist and how they are defined across functional neuroimaging experiments. Optimally, an appropriate and useful cognitive ontology will not merely be a catalogue of various mental operations parsed down to very fine detail in accord with current theories of cognitive psychology. While the consideration of competing theories often results in new knowledge discovery, the development of such a top-down ontology would be so continuously and vehemently debated that it could never reach a sufficient degree of consensus in order to be considered adoptable by the neuroimaging community. In contrast, a biologically based ontology that is driven by the way we observe the brain to operate in imaging experiments may reveal a cognitive architecture that has not previously been considered.

BrainMap's FUTURE ROLE IN ATLAS AND ONTOLOGY DEVELOPMENT

The synergy that exists between the BrainMap database and the ALE meta-analysis method was designed to facilitate the creation

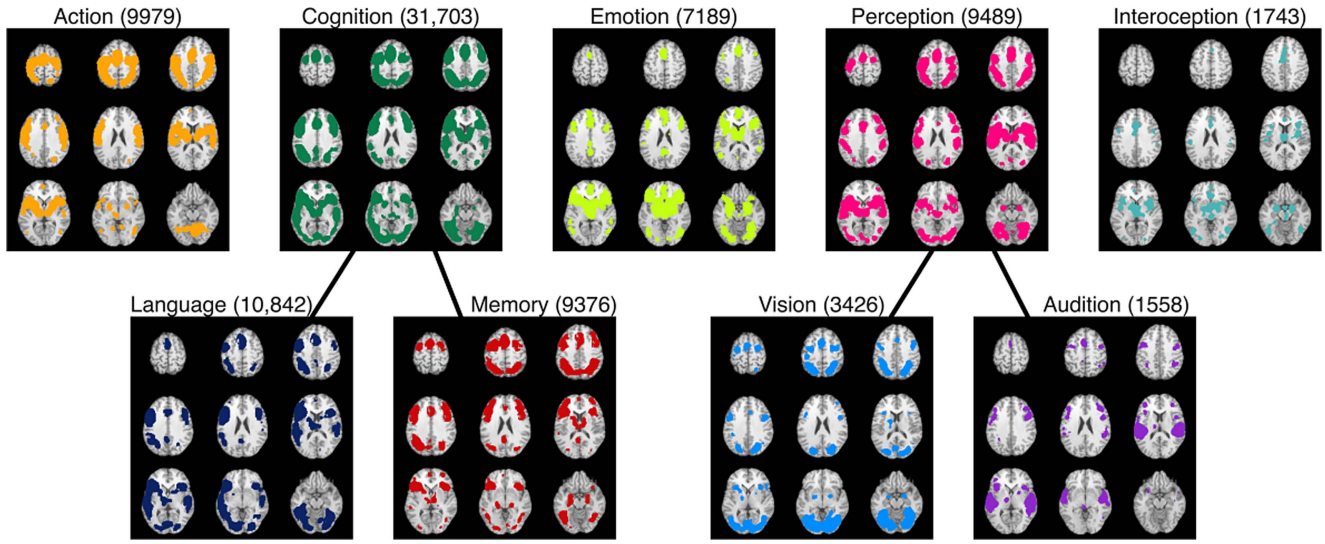


FIGURE 8 | BrainMap's functional atlas development strategy. ALE meta-analyses were performed on nine arbitrarily selected behavioral domain categories and sub-categories: action, cognition, emotion, perception, interoception, language, memory, vision, and audition. These results illustrate the mapping of function-to-regions correspondences that are currently capable using the BrainMap database. Refinement of these maps can be accomplished through high-level filtering and mining of the database. Results are displayed at $P < 0.05$, corrected. The total number of coordinates in each meta-analysis is listed after the domain heading for each map in parentheses.

of a functional brain atlas. BrainMap's search capabilities can support different types of queries, such as "for a given function, what regions are typically engaged?", "for a given region, what tasks elicit activation?", or "for a given region, what other regions are coactivated?" These questions highlight the value of meta-analytic results in comparison to results from individual studies. We believe that these correspondences (function-to-regions, region-to-tasks, or region-to-network) must be constructed according to a bottom-up strategy, using knowledge gleaned from data-driven analyses. Current probabilistic structural atlases (e.g., the Harvard-Oxford structural probability atlas, Smith et al., 2004, or the Jülich cytoarchitectonic atlas, Eickhoff et al., 2005, 2006a) have proven to be useful in testing region-to-tasks or region-to-network associations, as shown in **Figure 7**. Theoretically, any method of determining regions of interest can be used to query the BrainMap database, whether structurally or functionally defined.

BrainMap is also capable of generating function-to-regions associations, albeit at a coarse resolution. Whole brain meta-analytic maps can be created for each behavioral domain category in BrainMap, which can then be decomposed into sub-networks based on different levels of the domain hierarchy. To illustrate, ALE meta-analyses were performed on nine different behavioral domain categories and sub-categories: action, cognition, emotion, perception, interoception, language, memory, vision, and audition (**Figure 8**). Each ALE image provides a unique mapping of the neural network associated with the relevant domain. Many regions are observed in multiple domain maps, and some maps are very similar, but none are identical. However, a more detailed domain structure is needed to fully characterize the range of human cognition. We propose that applying high-level filters from the entire BrainMap coding scheme to these meta-maps, in the way condition-based filters were applied in the word generation meta-analysis (**Figure 6**), can be an effective strategy for refining the spatial specificity of these images. Thus, while paradigm class and behavioral domain are important metadata fields in the BrainMap coding scheme, *all* fields have the potential to assist in unraveling the brain's systems and their interactions.

As the BrainMap database increases in size, these results will evolve and grow more powerful, perhaps leading to a multi-layered, multi-modal probabilistic functional brain atlas derived from many different large-scale coordinate-based meta-analyses. This approach would likely be enhanced by the development of a standardized mental ontology. Differences in competitive terminology must be resolved to allow for the union of experimentally similar data sets. Perhaps the best strategy would be to combine all of BrainMap's data-driven methods in establishing function–structure relationships with other ontology initiatives, such at the Cognitive Atlas[7] (Bilder et al., 2009), the NIFSTD ontology (Bug et al., 2008), or the Neural ElectroMagnetic Ontologies (NEMO) (Frishkoff et al., 2009). Given the complex nature of human brain function, it is reasonable to suggest that no single approach will be powerful enough to solve the fundamental challenges associated with mapping the mind, but rather a joint effort will be required.

ACKNOWLEDGEMENTS

This work was supported by the NIMH (R01-MH074457-03; PI = Peter Fox and R01-MH084812-01; PI = Angela Laird and Jessica Turner), NINDS (T35-NS051166-04; PI = Peter Fox), NIDCD (F32-DC009116-02; PI = Matthew Cykowski), NCRR (U24-RR021992; PI = Steven Potkin), and the Helmholz Initiative on Systems-Biology (SBE).

[7]http://cognitiveatlas.org

REFERENCES

Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H. B., and Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *J. Comp. Neurol.* 412, 319–341.

Amunts, K., Weiss, P. H., Mohlberg, H., Pieperhoff, P., Eickhoff, S., Gurd, J. M., Marshall, J. C., Shah, N. J., Fink, G. R., and Zilles, K. (2004). Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotactic space – the roles of Brodmann areas 44 and 45. *Neuroimage* 22, 42–56.

Basho, S., Palmer, E. D., Rubio, M. A., Wulfeck, B., and Muller, R. A. (2007). Effects of generation mode in fMRI adaptations of semantic fluency: paced production and overt speech. *Neuropsychologia* 45, 1697–1706.

Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., Freimer, N. B., and Poldrack, R. A. (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cogn. Neuropsychiatry* (in press).

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* (in press).

Bowden, D. M., and Dubach, M. F. (2003). NeuroNames 2002. *Neuroinformatics* 1, 43–59.

Bowden, D. M., and Martin, R. F. (1995). NeuroNames brain hierarchy. *Neuroimage* 2, 63–83.

Brown, S., Laird, A. R., Ingham, R. J., Ingham, J. C., and Fox, P. T. (2005). Stuttered and fluent speech production: an ALE meta-analysis of functional neuroimaging studies. *Hum. Brain Mapp.* 25, 105–117.

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., and Martone, M. E. (2008). The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics* 6, 175–194.

Cabeza, R., and Nyberg, L. (2000). Imaging cognition II: an empirical review of 275 PET and fMRI studies. *J. Cogn. Neurosci.* 12, 1–47.

Chein, J. M., Fissell, K., Jacobs, S., and Fiez, J. A. (2002). Functional heterogeneity within Broca's area during verbal working memory. *Physiol. Behav.* 77, 635–639.

Cordes, D., Haughton, V. M., Arfanakis, K., Wendt, G. J., Turski, P. A., Moritz, C. H., Quigley, M. A., and Meyerand, M. E. (2000). Mapping functionally related regions of brain with functional connectivity MR imaging. *Am. J. Neuroradiol.* 21, 1636–1644.

Costafreda, S. G., Brammer, M. J., David, A. S., and Fu, C. H. (2008). Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and fMRI studies. *Brain Res. Rev.* 58, 57–70.

Costafreda, S. G., Fu, C. H. Y., Lee, L., Everitt, B., Brammer, M. J., and David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: role of the left inferior frontal gyrus. *Hum. Brain Mapp.* 27, 799–810.

Coté, R. A., and Robboy, S. (1980). Progress in medical information management. Systemized nomenclature of medicine (SNOMED). *JAMA* 243, 756–762.

Damoiseaux, J. S., Rombouts, S. A., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., and Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13848–13853.

Decety, J., and Lamm, C. (2007). The role of the right tempoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist* 13, 580–593.

Derrfuss, J., Brass, M., Neumann, J., and Yves von Cramon, D. (2005). Involvement of the inferior frontal junction in cognitive control: meta-analyses of switching and Stroop studies. *Hum. Brain Mapp.* 25, 22–34.

Derrfuss, J., and Mar, R. A. (2009). Lost in localization: the need for a universal coordinate database. *Neuroimage* (in press).

Eickhoff, S., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., and Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* 25, 1325–1335.

Eickhoff, S. B., Amunts, K., Mohlberg, H., and Zilles, K. (2006a). The human parietal operculum. II. Stereotaxic maps and correlation with functional imaging results. *Cereb. Cortex* 16, 268–279.

Eickhoff, S. B., Heim, S., Zilles, K., and Amunts, K. (2006b). Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage* 32, 570–582.

Eickhoff, S. B., Heim, S., Zilles, K., and Amunts, K. (2009a). A systems perspective on the effective connectivity of overt speech production. *Philos. Transact. A Math. Phys. Eng. Sci.* 367, 2399–2421.

Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., and Fox, P. T. (2009b). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* (in press).

Ellison-Wright, I., and Bullmore, E. (2009). Meta-analysis of diffusion tensor imaging studies in schizophrenia. *Schizophr. Res.* 108, 3–10.

Ellison-Wright, I., Glahn, D. C., Laird, A. R., Thelen, S. M., and Bullmore, E. T. (2008). The anatomy of first-episode and chronic schizophrenia: an anatomical likelihood estimation meta-analysis. *Am. J. Psychiatry* 165, 1015–1023.

Fitzgerald, P. B., Laird, A. R., Maller, J., and Daskalakis, Z. J. (2008). A meta-analytic study of changes in brain activation in depression. *Hum. Brain Mapp.* 29, 683–695.

Fox, M. D., and Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700–711.

Fox, P. T., and Lancaster, J. L. (2002). Mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* 3, 319–321.

Fox, P. T., Parsons, L. M., and Lancaster, J. L. (1998). Beyond the single study: function–location meta-analysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* 8, 178–187.

Frishkoff, G. A., Dou, D., Frank, R., LePendu, P., and Liu, H. (2009). Development of Neural Electromagnetic Ontologies (NEMO): representation and integration of event-related brain potentials. *Proc. Int. Conf. Biomed. Ontologies* (in press).

Glahn, D. C., Laird, A. R., Ellison-Wright, I., Thelen, S. M., Robinson, J. L., Lancaster, J. L., Bullmore, E., and Fox, P. T. (2008). Meta-analysis of gray matter anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. *Biol. Psychiatry* 64, 774–781.

Glahn, D. C., Ragland, J. D., Abramoff, A., Barrett, J., Laird, A. R., Bearden, C. E., and Velligan, D. I. (2005). Beyond hypofrontality: a quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. *Hum. Brain Mapp.* 25, 60–69.

Grefkes, C., Eickhoff, S. B., Nowak, D. A., Dafotakis, M., and Fink, G. R. (2008). Dynamic intra- and interhemispheric interactions during unilateral and bilateral hand movements assessed with fMRI and DCM. *Neuroimage* 41, 1382–1394.

Grosbras, M.-H., Laird, A. R., and Paus, T. (2005). Cortical regions involved in eye movements, shifts of attention, and gaze perception. *Hum. Brain Mapp.* 25, 140–154.

Heim, S., Eickhoff, S. B., and Amunts, K. (2008). Specialisation in Broca's region for semantic, phonological, and syntactic fluency? *Neuroimage* 40, 1362–1368.

Koski, L., and Paus, T. (2000). Functional connectivity of the anterior cingulate cortex with the human frontal lobe: a brain-mapping meta-analysis. *Exp. Brain Res.* 133, 55–65.

Laird, A. R., Lancaster, J. L., and Fox, P. T. (2005a). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics* 3, 65–78.

Laird, A. R., McMillan, K. M., Lancaster, J. L., Kochunov, P., Turkeltaub, P. E., Pardo, J. V., and Fox, P. T. (2005b). A comparison of label-based review and activation likelihood estimation in the Stroop task. *Hum. Brain Mapp.* 25, 6–21.

Laird, A. R., Fox, M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., Turkeltaub, P. E., Kochunov, P., and Fox, P. T. (2005c). ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25, 155–164.

Laird, A. R., Robbins, J. M., Li, K., Price, L. R., Cykowski, M. D., Narayana, S., Laird, R. W., Franklin, C., and Fox, P. T. (2008). Modeling motor connectivity using TMS/PET and structural equation modeling. *Neuroimage* 41, 424–436.

Lancaster, J. L., Laird, A. R., Fox, M., Glahn, D. E., and Fox, P. T. (2005). Automated analysis of meta-analysis networks. *Hum. Brain Mapp.* 25, 174–184.

Lancaster, J. L., Tordesillas-Gutierrez, D., Martinez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J. C., and Fox, P. T. (2007). Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* 28, 1194–1205.

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., and Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Hum. Brain Mapp.* 10, 120–131.

Langlotz, C. P. (2006). RadLex: a new method for indexing online educational materials. *Radiographics* 26, 1595–1597.

McMillan, K. M., Laird, A. R., Witt, S. T., and Meyerand, M. E. (2007). Self-paced working memory: validation of

verbal variations of the n-back paradigm. *Brain Res.* 1139, 133–142.

Menzies, L. A. C., Chamberlain, S. R., Laird, A. R., Thelen, S. M., Sahakian, B. J., and Bullmore, E. T. (2008). Integrating evidence from neuroimaging and neuropsychological studies of obsessive compulsive disorder: the orbitofronto-striatal model revisited. *Neurosci. Biobehav. Rev.* 32, 525–549.

Minzenberg, M. J., Laird, A. R., Thelen, S. M., Carter, C. S., and Glahn, D. C. (2009). Meta-analysis of 41 functional neuroimaging studies of executive cognition reveals dysfunction in a general-purpose cognitive control system in schizophrenia. *Arch. Gen. Psychiatry* (in press).

Neumann, J., Lohmann, G., Derrfuss, J., and Yves von Cramon, D. (2005). The meta-analysis of functional imaging data using replicator dynamics. *Hum. Brain Mapp.* 25, 165–173.

Neumann, J., Yves von Cramon, D., and Lohmann, G. (2008). Model-based clustering of meta-analytic functional imaging data. *Hum. Brain Mapp.* 29, 177–192.

Petacchi, A., Laird, A. R., Fox, P. T., and Bower, J. M. (2005). Cerebellum and auditory function: an ALE meta-analysis of functional neuroimaging studies. *Hum. Brain Mapp.* 25, 118–128.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature* 331, 585–589.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63.

Poldrack, R. A. (2008). The role of fMRI in cognitive neuroscience: where do we stand? *Curr. Opin. Neurobiol.* 18, 223–227.

Postuma, R. B., and Dagher, A. (2006). Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. *Cereb. Cortex* 16, 1508–1521.

Price, C. J., Devlin, J. T., Moore, C. J., Morton, C., and Laird, A. R. (2005). Meta-analyses of object naming: effect of baseline. *Hum. Brain Mapp.* 25, 70–82.

Price, C. J., and Friston, K. J. (2005). Functional ontologies for cognition: the systematic definition of structure and function. *Cogn. Neuropsychol.* 22, 262–275.

Ragland, J. D., Laird, A. R., Ranganath, C. S., Blumenfeld, R. S., Gonzales, S. M., and Glahn, D. C. (2009). Prefrontal activation deficits during episodic memory in schizophrenia. *Am. J. Psychiatry* (in press).

Robinson, J. L., Laird, A. R., Glahn, D. C., Lovallo, W. R., and Fox, P. T. (2009). Meta-analytic connectivity modeling: Delineating the functional connectivity of the human amygdala. *Hum. Brain. Mapp.* (in press).

Rogers, B. P., Morgan, V. L., Newton, A. T., and Gore, J. C. (2007). Assessing functional connectivity in the human brain by fMRI. *Magn. Reson. Imaging* 25, 1347–1357.

Rosse, C., and Mejino, J. L. V. (2003). A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *J. Biomed. Inform.* 36, 478–500.

Rubin, D. L. (2008). Creating and curating a terminology for radiology: ontology modeling and analysis. *J. Digit. Imaging* 21, 355–362.

Schroeter, M. L., Raczka, K., Neumann, J., and Yves von Cramon, D. (2007). Towards a nosology for frontotemporal lobar degenerations – a meta-analysis involving 267 subjects. *Neuroimage* 36, 497–510.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., and Beckmann, C. F. (2009). The brain's functional architecture: Correspondence between rest and activation. *Proc. Natl. Acad. Sci. USA* (in press).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., and Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23(Suppl. 1), S208–S219.

Soros, P., Inamoto, Y., and Martin, R. E. (2009). Functional brain imaging of swallowing: an activation likelihood estimation meta-analysis. *Hum. Brain Mapp.* (in press).

Spreng, R. N., Mar, R. A., and Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind and the default mode: a quantitative meta-analysis. *J. Cogn. Neurosci.* 21, 489–510.

Stephan, K. E., Kamper, L., Bozkurt, A., Burns, G. A., Young, M. P., and Kötter, R. (2001). Advanced database methodology for the Collation of Connectivity data on the Macaque brain (CoCoMac). *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* 356, 1159–1186.

Stevens, R., Goble, C. A., and Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.* 1, 398–414.

Toga, A. W. (2002). Neuroimage databases: the good, the bad, and the ugly. *Nat. Rev. Neurosci.* 3, 302–309.

Toro, R., Fox, P. T., and Paus, T. (2008). Functional coactivation map of the human brain. *Cereb. Cortex* 18, 2553–2559.

Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16, 765–780.

Wager, T. D., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* 2, 150–158.

Wager, T. D., Lindquist, M., Nichols, T. E., Kober, H., and Van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* 45, S210–S221.

Warburton, E., Wise, R. J., Price, C. J., Weiller, C., Hadar, U., Ramsay, S., and Frackowiak, R. S. (1996). Noun and verb retrieval by normal subjects. Studies with PET. *Brain* 119, 159–179.

Wise, R., Chollet, F., Hadar, U., Friston, K., Hoffner, E., and Frackowiak, R. (1991). Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain* 114, 1803–1817.

Xiong, J., Parsons, L. M., Gao, J. H., and Fox, P. T. (1999). Interregional connectivity to primary motor cortex revealed using MRI resting state images. *Hum. Brain Mapp.* 6, 151–156.

Conflict of Interest Statement: The authors report no competing interests.

Received: 01 April 2009; paper pending published: 12 May 2009; accepted: 26 June 2009; published online: 09 July 2009.

*Citation: Laird AR, Eickhoff SB, Kurth F, Fox PM, Uecker AM, Turner JA, Robinson JL, Lancaster JL and Fox PT (2009) ALE meta-analysis workflows via the BrainMap database: progress towards a probabilistic functional brain atlas. Front. Neuroinform. (2009) **3**:23. doi: 10.3389/neuro.11.023.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 21 April 2010
doi: 10.3389/neuro.11.036.2009

Mining the Mind Research Network: a novel framework for exploring large scale, heterogeneous translational neuroscience research data sources

Henry J. Bockholt[1]*, Mark Scully[1,2], William Courtney[1,2], Srinivas Rachakonda[3], Adam Scott[1], Arvind Caprihan[3], Jill Fries[3], Ravi Kalyanam[3], Judith M. Segall[3], Raul de la Garza[1], Susan Lane[1] and Vince D. Calhoun[1,2,3,4]

[1] *Neuroinformatics, Mind Research Network, Albuquerque, NM, USA*
[2] *Computer Science, The University of New Mexico, Albuquerque, NM, USA*
[3] *Medical Image Analysis, Mind Research Network, Albuquerque, NM, USA*
[4] *Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM, USA*

Edited by:
John Van Horn, University of California at Los Angeles, USA

Reviewed by:
Shantanu Joshi, University of California at Los Angeles, USA
John Van Horn, University of California at Los Angeles, USA

****Correspondence:***
Henry J. Bockholt, The Mind Research Network, 1101 Yale Blvd NE, Albuquerque, NM 87131, USA.
e-mail: jbockholt@mrn.org

A neuroinformatics (NI) system is critical to brain imaging research in order to shorten the time between study conception and results. Such a NI system is required to scale well when large numbers of subjects are studied. Further, when multiple sites participate in research projects organizational issues become increasingly difficult. Optimized NI applications mitigate these problems. Additionally, NI software enables coordination across multiple studies, leveraging advantages potentially leading to exponential research discoveries. The web-based, Mind Research Network (MRN), database system has been designed and improved through our experience with 200 research studies and 250 researchers from seven different institutions. The MRN tools permit the collection, management, reporting and efficient use of large scale, heterogeneous data sources, e.g., multiple institutions, multiple principal investigators, multiple research programs and studies, and multimodal acquisitions. We have collected and analyzed data sets on thousands of research participants and have set up a framework to automatically analyze the data, thereby making efficient, practical data mining of this vast resource possible. This paper presents a comprehensive framework for capturing and analyzing heterogeneous neuroscience research data sources that has been fully optimized for end-users to perform novel data mining.

Keywords: data mining, magnetic resonance imaging, XML, XCEDE, Mind Clinical Imaging Consortium

INTRODUCTION

Modern science is marked by an accumulation of massive amounts of data and neuroscience is no exception. The different neuroimaging modalities, such as diffusion tensor imaging (DTI), functional magnetic resonance imaging (fMRI), structural MRI (sMRI), electroencephalography (EEG), positron emission tomography, or magnetoencephalography, each produce a huge amount of data that when combined with genetic information, psychological assessment results, and socio-demographics makes it impossible for researchers to draw conclusions without sophisticated storage, recall, and inference methods. As research has moved to multi-site collaborations, the difficulties of working with large datasets have only increased, underlining the need for comprehensive tools to address these problems (Amari et al., 2002).

Neuroinformatics (NI) aims to solve these problems and increase the effectiveness of researchers through intelligent use of data storage, data analysis, and data presentations. NI makes storage and retrieval of data easy and transparent to researchers, but also assists them by supplying only the data that is relevant to their needs (Toga, 2002). Combining these services with data repositories enables easy sharing and reduces the difficulties of scanning enough subjects to draw meaningful conclusions.

The importance of a NI framework and system cannot be overstated. A NI system is critical in order to shorten the time between study conception and results. Second, a scalable system is required when large numbers of participants are studied. Further, when multiple sites participate in research projects, organizational issues become difficult. Optimized NI applications mitigate these problems. Finally, NI software enables coordination across multiple studies, leveraging the advantages of each to potentially lead to exponentially greater research discoveries. The web-based Mind Research Network (MRN) system has been designed and improved through our experience with several multi-site translational neuroscience research studies and feedback from researchers from seven different institutions. The MRN tools permit the collection, management, reporting and efficient use of large scale, heterogeneous data sources, e.g., multiple institutions, multiple principal investigators, multiple research programs and studies, and multimodal acquisitions (Carneiro and Vasconcelos, 2005; Bockholt et al., 2007).

Applications typically contain complex features often found to be non-intuitive by end-users, especially when they are first starting to use them. Our framework has been shaped by the requirements of several years of experience in providing NI tools to a full-spectrum of investigators and researchers conducting data acquisition, storage,

management, analysis, and retrieval. The MRN approach and tools have proven to be effective and scalable. When researchers have access to an existing, well-designed, well-documented turnkey solution, that is already specialized to their domain of research, they can use the tools for their own projects, providing a distinct advantage to the group in both startup time and in minimizing future data integrity problems. However, the ultimate goal of data mining is to effectively use data sources to their full potential. The framework presented herein strives to achieve this end for the scientists that access the vast MRN data sources by providing intuitive access to fully annotated, anonymous data sources for novel exploration.

MATERIALS AND METHODS

The MRN Clinical Imaging Consortium (MCIC) is one example of a multi-institutional program for which the described framework was initially developed, built, and deployed (Demirci et al., 2008; Kim et al., 2009; Segall et al., 2009; Sui et al., 2009). The MCIC project needed sophisticated tools to analyze and support the multi-site heterogeneous data sources that were collected by the consortium of investigators (Carneiro and Vasconcelos, 2005). The tools within the framework needed to provide security, querying, reporting, analyzing, summarizing, exporting, and archiving capabilities (see **Figure 1**). The MCIC project is composed of sources from more than 400 human research volunteers that have had comprehensive baseline and longitudinal neuroimaging (sMRI, fMRI, DTI), genetic, clinical, socio-demographic, and neuropsychological assessments. This NI capability has been actively used by several investigators and researchers distributed across The University of New Mexico, The University of Iowa, The University of Minnesota, and Massachusetts General Hospital at Harvard University. In addition to the MCIC project, the framework presented has benefited enormously through years of collaboration with the Biomedical Informatics Research Network (BIRN[1]), the National Alliance for Medical Imaging Computing (NA-MIC[2]) and collectively, all of the investigators and researchers within the scope of the MRN.

A large volume of data is collected, managed, and made available for exploration in any type of neuroscience research project. In **Figure 2** is an overview of the applications that commonly access and use the MRN clinical research tools. The framework focuses on real-time neuropsychological assessment acquisition via a tablet-PC platform, real-time annotation via web services, collaborative web portals for data management and reporting, automated neuroimaging analyses, web application tools for monitoring and staging data analyses, quality assurance (QA) methods, and data mining capabilities. The full implementation details for this framework will be made available on The Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC)[3].

DATABASE

A system for storing, archiving, accessing, and integrating the various sources of data is clearly needed. One tier of the system is a relational database management system (RDBMS) (Farn and Hu, 1995). The advantage of using a RDBMS over other types of databases is that the RDBMS technology is mature, stable, portable, scalable, and easy to integrate (Brinkley and Rosse, 2002; Bly et al., 2004; Bota and Arbib, 2004; Bota et al., 2005). In the MRN data-mining framework, we have determined that the following items in **Table 1** should be supported within the RDBMS schema.

COLLABORATIVE WEB PORTAL APPLICATION

The public face of the MRN framework is a collaborative portal that provides secure access to data sources for the participating researchers and investigators. This web-tier application manages requests between a user's desktop browser and the RDBMS tier. To accompany the RDBMS, we have identified functional requirements and designed and implemented a comprehensive web-based system to support the translational neuroscience research needs within the MRN organization. These requirements have been summarized in **Table 2**.

The specific requirements for an end-user's ability to create, modify, query, or export a given item of research data depends on the site and role of the user requesting the data manipulation event (Prasad et al., 1987; Brinkley and Rosse, 2002; Bota and Arbib, 2004; Costa, 2004; Bota et al., 2005; Jovicich et al., 2005). We have developed tools for attaching roles to portal users, such as principal investigator, co investigator, study coordinator, rater, etc. The features that a given user has access to depends upon the assigned role that user has in the study. The MRN framework provides a mechanism for indicating who the principal investigator is on a given study and a means for managing the users and their role on each study.

WEB-BASED DATA-ENTRY

The socio-demographic, clinical, and neuropsychological assessments collected in the MCIC protocol, along with many other types of multi-site consortium studies, generate a large amount of data that must be made electronic so that it can be integrated with data

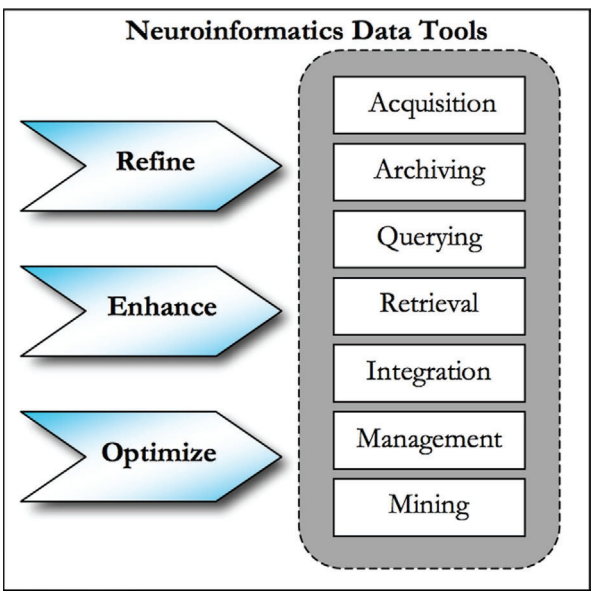


FIGURE 1 | Neuroinformatics data tools.

[1]http://www.nbirn.net

[2]http://www.na-mic.org

[3]http://www.nitrc.org/projects/mindknowdb/

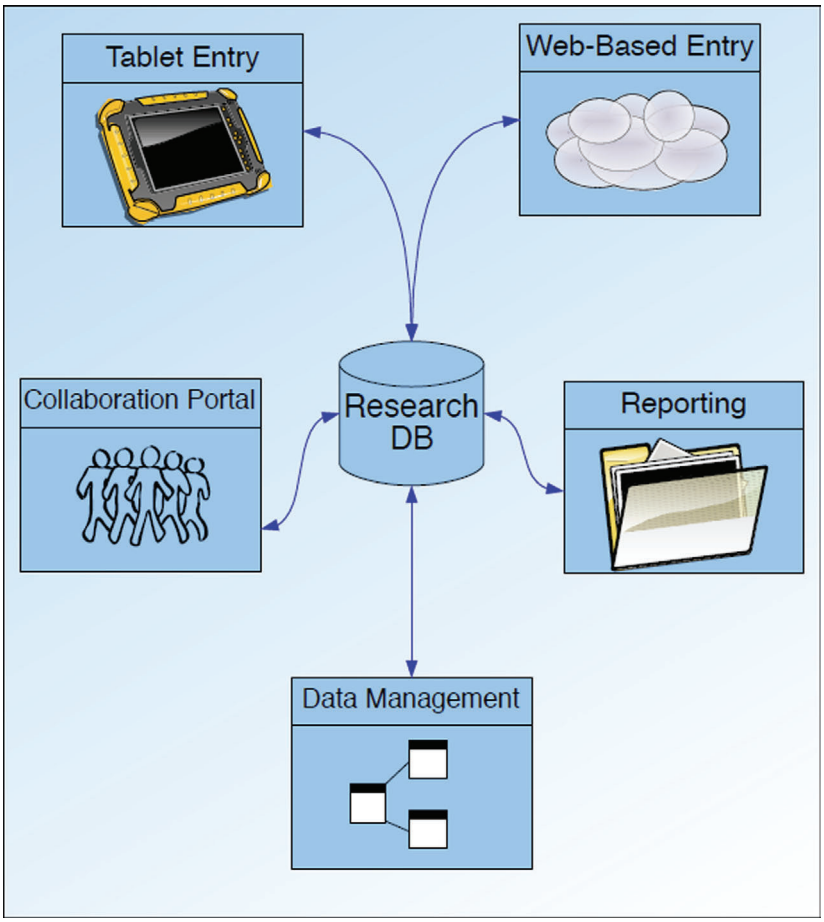


FIGURE 2 | Overview of MRN Neuroinformatics System.

Table 1 | Items that should be supported within the RDMBS schema.

One or more data collection sites
One or more research participants at one or more sites
One or more studies across one or more sites
One or more subjects that can be assessed at one or more sites
One or more assessments conducted by one or more raters
One or more visits by a given subject participating in one or more studies
One or more neuroimaging modalities across one to many sessions
Support for multiple image analysis pipelines
Support for multiple image analysis results from one or more pipelines
Support for genetic polymorphisms (SNP results)
Support for real-time annotation of all data sources

Table 2 | Functional requirements for the web-based system supporting the translational neuroscience research needs within the MRN organization.

Protocol and consents to participate in research	Document library
Timeline of required events for each cohort	Weekly progress reports
PDF documents of all required assessments	Presentations and publications
Meeting information, agendas, minutes	Investigator initiated reports
Simple summary of collection by site demographics	Clinical raters
Data requests	Roster of participants
Metadata summary	Calibration information
Real-time annotation tool	Training information
Summary of requests by other investigators	*Ad hoc* queries
Archive of delivered data requests	

collected in other research domains. After completing a training program, raters, the individuals that conduct the assessment events, are trained to document the interview results on standard paper-based forms. When a complete set of assessments has been collected and documented for a given subject, the stack of assessments for that subject is shipped to a centralized data-entry. This data acquisition process generates specific requirements for an application to manage the multitude of paper-based assessments.

Data-entry of clinical, socio-demographic, neuropsychological and other types of assessments performed is necessary since most of the time these data sources are collected as pen-and-paper-based assessments. The following web application requirements for an

assessment data-entry system have been determined: perform first entry of assessment by data-entry operator; perform second entry of assessment by alternative data-entry operator; perform conflict screening and logical checks of doubly-entered assessments by clinical program manager; summarize data acquisition by assessment, subject, site, and other custom report generating features as needed (Andreasen et al., 1995; Vessey et al., 2003).

The pen-and-paper forms must be data-entered in a secure and fault tolerant manner. The web-based data entry application, accessible via the intranet, facilitates the first and second entry of assessment data by two different data-entry operators. A clinical program manager then utilizes the application to perform conflict screening and logical checks on the double-entered assessments. Data acquisition summaries may then be generated by assessment, subject, and site along with other custom report generating features as needed.

TABLET-BASED DATA-ENTRY

The purpose of the tablet PC entry capability is to provide end-users with a capability for the real-time collection and quality validation of clinical neuroscience research assessment data. This tool (written for use on any tablet hardware running Windows XP Tablet Edition with SDK 1.7, operating system patched to SP2, Microsoft.Net 2.0 framework 1.7) provides our researchers a means to capture assessment data electronically in settings where a network connection may not be possible or permitted. Electronic acquisition of assessment events also permits a more efficient research process since data-entry of paper-based assessments is not required. Additionally, quality control can be conducted in real-time, since the tablet PC can provide feedback to the rater during the data acquisition process. Finally, tablet PC based data collection was found to be preferred by raters (Pace and Staton, 2005; Cole et al., 2006).

The Tablet Assessment software validates data as it is entered, including: required fields; data type for the response (e.g., numeric, character string or date); bounds checking information (e.g., systolic blood pressure is a number between 0 and 300); question dependencies (e.g., question 2 "How many cigarettes do you smoke a day?" does not need to be answered if the answer to question 1 "Do you smoke?" is no).

During an interview, the rater is notified immediately when a required field is skipped or data entered does not meet quality criteria, but the software does not constrain the rater to fix the data immediately. This allows the rater to complete the interview smoothly and fix data issues at a later time if necessary. Assessments that do not pass data quality validation may be stored on the rater's tablet and edited at any time, but they may not be submitted to the database until all issues are resolved. The tablet-based product stores and maintains the data that it manages in XML and is capable of exporting data via a SOAP webservice[4] using XCEDE[5] or other XML schema.

SCAN ANNOTATION

In providing NI tools for the MCIC project, we have developed a utility for having integrated data sources and real-time documentation of what, when, and where items (such as neuroimaging events) succeed or fail. This documentation permits timely, efficient processing and maximizes data-usability. In **Figure 3**, we present a screenshot of a real-time, web-based, image annotation tool. During a neuroimaging session, this annotation tool allows the end-user to track and document each imaging series. The web application is connected to both a custom DCM4CHE-based DICOM receiver[6] and the MRN RDBMS database described above. The order of events, whether or not the event was completed, whether or not the end-user thinks that the imaging data is usable for analysis can be annotated by using this tool. Furthermore, the end-user may attach additional detailed documentation such as why an image may not be usable. Finally, auxiliary files, such as behavioral data, may be attached and submitted in real-time.

AUTO-ANALYSIS DESCRIPTION

We have standards in place at The MRN for researchers to follow for scanning and naming data. When research subjects are scanned, information is input into a database form on the scanner

[4]http://www.w3.org/TR/soap/

[5]http://www.xcede.org/

[6]http://dicom.offis.de/dcmtk.php.en

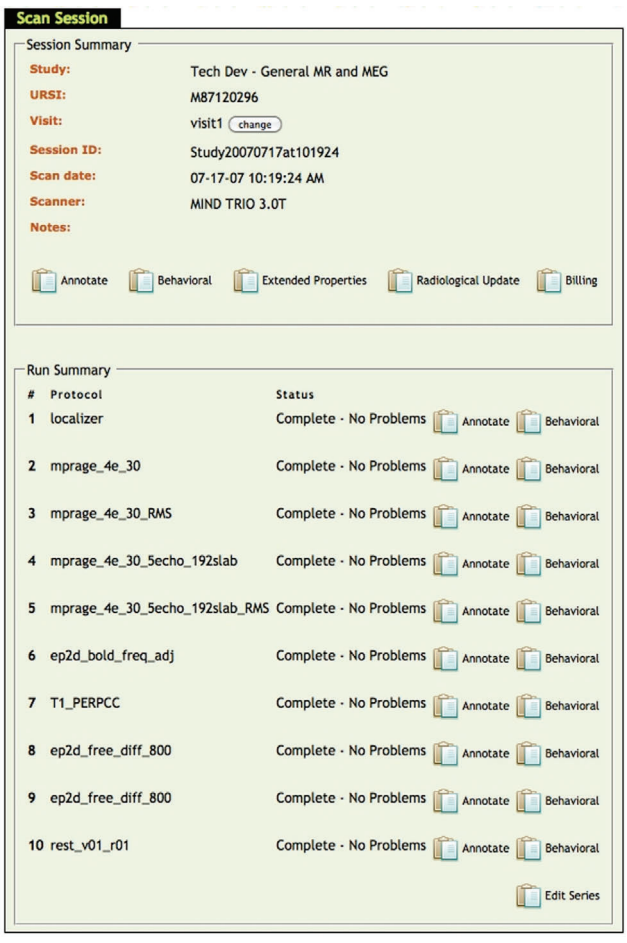


FIGURE 3 | Real-time annotation tool.

console that feeds the information directly into the NI database about the scan session. The scanned data is then archived onto a backup storage space. The subject data is then transferred by auto-analysis scripts to a local analysis area. Here the data is reformatted, so that all scans are in an analyzable format. Automation performs the first level analysis for fMRI, the two modalities of sMRI, and DTI. Preprocessing is done on both the fMRI and sMRI data that allows researchers to work on the statistical analysis, instead of having to pre-process their data first. For the structural scans, FreeSurfer provides cortical and subcortical results for individual subjects. VBM provides volume and density results for grey matter and white matter tissues. In the case of FreeSurfer, which is processed on a computing cluster, auto-analysis has saved investigators lots of time and local computing resources, for it is a computationally intensive software package. DTI gives us water diffusion results for white matter tracts. We are currently in the process of automating magnetic resonance spectroscopy. This auto-analysis pipeline benefits the PIs, boosting their effectiveness, and it also magnifies the value of the information, allowing it to be pooled from smaller datasets to larger datasets, yielding large Ns to analyze effects, such as gender, that are not seen in smaller datasets.

DATA QUERYING

We have learned that users of our MRN tools wish to perform customized queries within individual research studies and across studies, where permitted. To that end, we have developed a prototype application to handle queries within and across research studies for all data domains stored in the MRN database. For custom queries, users typically wish to first be able to select the study or studies they wish to query, and then perform some high-level filtering of the major data domains in order to set the criteria for the subjects they wish to analyze.

In **Figure 4**, we demonstrate the applications filtering capability. The user is able to select assessment criteria, such as the instrument, the visit type, the field, and operator and a value. In the example, the user wished to query all MCIC subjects where neuropyschological batteries were conducted at a baseline visit and where the total reading score was assessed at greater than a value of 50. The result is a filtered list of subjects for which the user is then asked what they wish to report from that filtered list of subjects. The example continues where the user is able to select and report all of the data sources available on that filtered list of subjects. Finally, we demonstrate how the user may export the data in a format that suits their needs. The application currently permits a customizable field delimiter, line terminator, and selectable data orientation. This functional prototype permits extensive customized querying, and given that it may be used across all data sources from all studies stored in the MRN database, it will prove to be an invaluable tool, forming the foundation for planned data mining activities.

QUALITY ASSURANCE AND QUALITY CONTROL

We now present two QA protocols: one for morphological data and one for behavioral data. We have used individuals control charts on the morphological data because of an automated segmentation algorithm that allows us to inspect every brain. When multiple structural scans are taken, the variation within session is too small to be identifiable (Spiring, 2007). Each segment is normalized to total brain volume due to differences across gender, age, and scanner differences (Tofts, 2004). Our control limits are set by the data, but as our database size continues to increase, the variability will decrease. With an increasingly large dataset that has multiple subject types, a

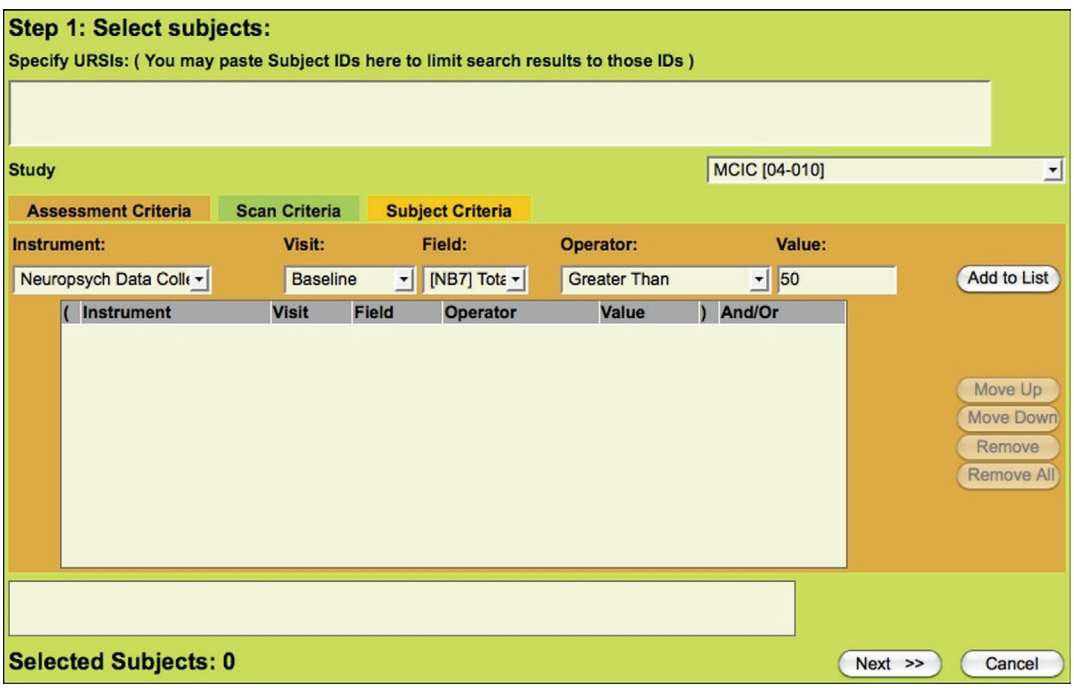


FIGURE 4 | Data querying application tool.

regression control chart (Aroian and Levene, 1950) can be used to detect outliers by subject type. The morphological findings show that within psychometrically normal subjects, neuromorphometric outliers are detected. These outliers will begin to lead us into researching more of the predictive potential of neuromorphometric data. **Figure 5** illustrates an example of a control chart. In the example, a single subject is found to be an outlier on the left thalamic proper label (that has been normalized by total brain volume). When this particular subject is flagged as a statistical outlier on this measure, the end-user is prompted to review the entire neuromorphometric results for that subject and make a decision on whether or not to use that subject in their particular analysis.

In tandem with the QA of the neuromorphometric data is the QA of the assessment data. The three aims of using QA of assessment data are: make certain that collected data falls within acceptable boundaries, use subject type to determine quality of data, and, integrate neuroimaging data and clinical assessment data to create multivariate control charts.

Quality assurance of the processed data is always a concern. We have designed measures to ensure that both the automation process is working correctly on all of the data analysis methods and that the quality is consistent. The NI database has a field to include notes about issues that come up during the individual scans. Incoming scans and processed scans are monitored daily by a team of people involved in automation. A weekly report of disk space and total number of scans is generated to make sure the process is operating properly. QA measures are being built in to the automation stream that notifies us when data fails to meet QA standards. These issues can then be resolved and the corrections implemented into future analyses to prevent concerns.

RESULTS

Within the MRN system across five sites and 280 system users, the framework encompasses access to 8502 subjects with 10,410 MRI scan sessions, 1200 EEG session, 752 unique instruments have been developed for 140,692 assessment events with a total of 2,533,868 questions available for use in mining nearly 150 TB of raw and analyzed data. We are now actively sequencing one million SNP arrays on prospective subjects, as well as continuing to collect vast amounts of baseline and longitudinal clinical, neuropsychological, behavioral, and treatment assessment results. While the preliminary research studies that drove the

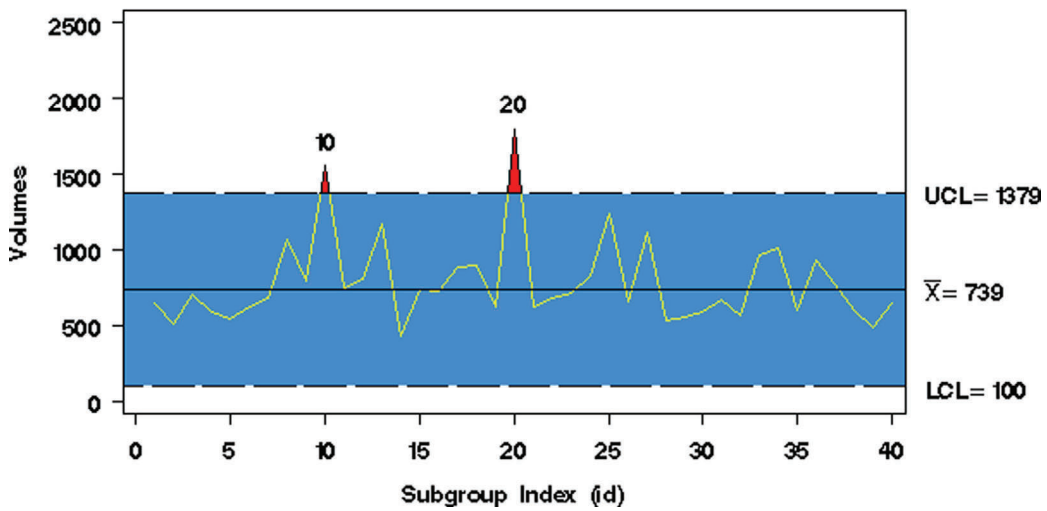


FIGURE 5 | Quality control chart example.

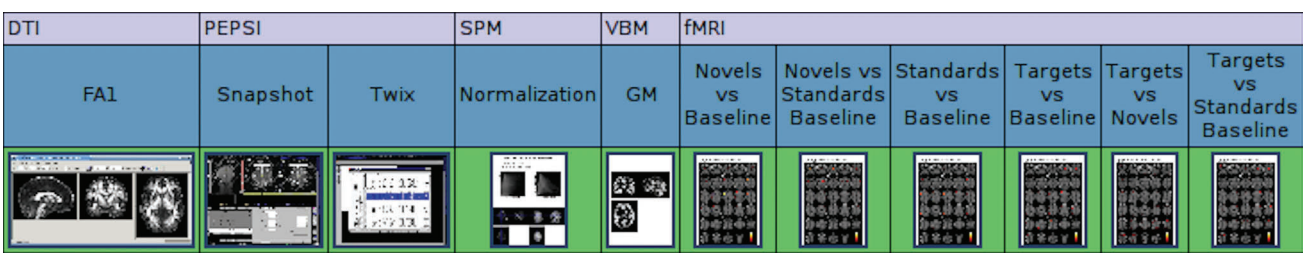


FIGURE 6 | First level analysis planning.

initial development of MRN tools were primarily based on schizophrenia, MRN studies managed in MRN tools now involve a wide range of psychiatric, psychological, and neurological disorders including post-traumatic stress disorder, psychopathy, addiction, traumatic brain injury, lupus, vascular dementia, stroke, mild cognitive impairment, Alzheimers, as well as studies of creativity and accelerated learning.

The following three case studies provide examples of results of data-mining using the MRN framework.

CASE STUDY ONE: NOVEL ANALYSIS PLANNING

As part of a study's configuration in the NI system, a protocol must be devised regarding the necessary assessments, tasks, and automated analysis pipelines that are to be performed on each subject entered into the study. These protocols are specialized to a specific subject type under which each subject may be registered. From this information, along with the data results and metadata contained in the database, the system can determine which subject data has been completed, which data is not yet scheduled to be completed, and which data is delinquent.

The system is flexible enough to accommodate multiple types of protocols to enable growth that may come with future analysis techniques. For example, it currently supports the management of assessment data and analysis snapshots, but will soon be used to drive automated quality control systems that will rely on professional, human confirmation. Furthermore, this unified protocol schema enables a commonized system of viewing the data collected from the subjects.

Persons with the necessary privileges may view the summarized results from their study's web portal. The user must choose a type of analysis to be summarized and may optionally filter their results by subject type. The results are displayed in a color-coded grid and can be sorted by the links at the tops of the columns. Data results that have been completed can be viewed by clicking on the appropriate link in their representative box. In the case of analysis pipelines, various images are displayed for fast reviewing purposes in thumbnail form (see **Figure 6**). These can be expanded, along with metadata concerning their entry, to be viewed in full size for a more detailed qualitative review. Assessment questions and their responses can be viewed in a similar way.

This tool provides investigators with a tool to summarize the results of the analysis done on their subjects' data and bring to light the results that are tardy in their completion. Much time can be spared from the waste of manually sifting through filesystem-based data storage to view results. As an added boon, problems with analysis pipelines may also now be found more easily.

CASE STUDY TWO: CONDUCTING ANALYSIS

An image processing module in the database can be used as a quick diagnostic tool to compare groups such as controls vs patients. A number of tests are supported including one sample *t*-tests, two sample *t*-tests, Class mean, and K-means clustering.

The input data for these algorithms are the contrast images obtained from the first level analyses using Statistical Parametric Mapping. **Figure 7** shows an example where six healthy and six schizophrenics contrast images are used to generate a one sample

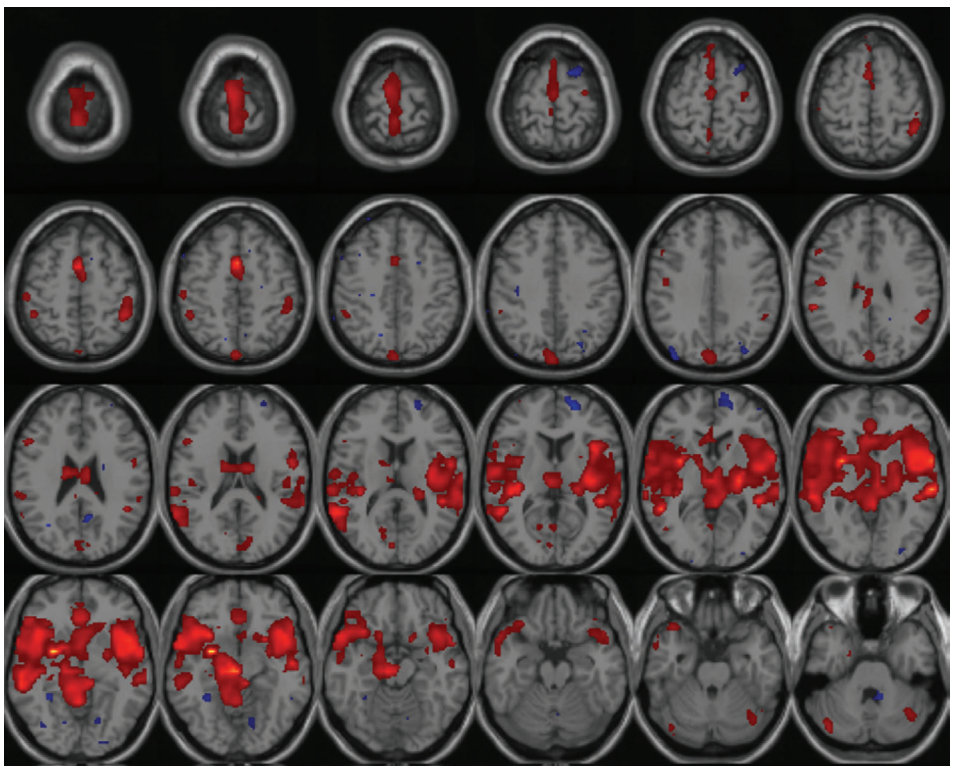


FIGURE 7 | One sample *t*-test calculated on 12 images (six healthy and six schizophrenics) with a T-threshold of 1.5 applied to the *t*-map.

t-test map. In addition, we plan to provide data mining tools which work with the preprocessed spatiotemporal fMRI data for example. In this case, the processing must be done in an offline manner as the wait time will be considerably longer.

CASE STUDY THREE: CLASSIFICATION

In this case study, the entire system can be tested, including determining selection criteria for each group, ensuring that first and second level analyses are performed and available, after which, a further classification analysis is performed. In this example, a class mean is used to classify a given image by computing the distance from the mean of each of the input groups.

To illustrate the method, we use five subjects from a healthy group and five subjects from schizophrenics group and two subjects from an unknown group that needs to be classified. Based on the Euclidean distance measure, both the unknown subjects belonged to the first group (**Figure 8**).

DISCUSSION

This novel workflow utilizes a custom web application client that communicates with a database back-end along with a custom DICOM receiver that permits the end-user to conduct real-time annotation of neuroimaging data during acquisition. The user has the ability to annotate each imaging series with metadata such as the order of events, whether or not the event was completed, whether or not the end-user thinks that the imaging data is usable for analysis, and any other notes or relevant information. In addition, the web application allows the end-user to upload auxiliary data, such as stimulus response time files, supporting video or other files that may be needed for full analysis of functional neuroimaging datasets.

For automated image analysis to be feasible in the NI framework presented here, the protocol metadata (what type, kind, and condition an imaging run belongs to) and the usability of an imaging session, are needed to perform analyses. As soon as the end-user has set the status of an imaging session to usable, an automated process evaluates the constraints of the protocol, metadata, and usability status in order to execute the appropriate image analysis pipeline. In functional imaging runs where behavioral data is needed to process activation maps, when the end-user attaches the behavioral data, it triggers the functional imaging pipeline processing. The other strength of managing research protocols is the ability to monitor and enforce compliance as well as provide a platform for QA.

Since we have integrated this annotation tool with a DICOM receiver, and a comprehensive RDBMS, we are able to provide end-users with rich metadata associated with each neuroimaging session and run. This integration of annotation along with comprehensive NI tools that combine clinical, socio-demographic, and neuropsychological data sources collected in a study greatly enhances the usability of data and establishes the foundation for efficient, semantic-based retrieval of complex images via a secure web application. When combined with fully automated image analyses, this annotation tool can serve as a powerful quality

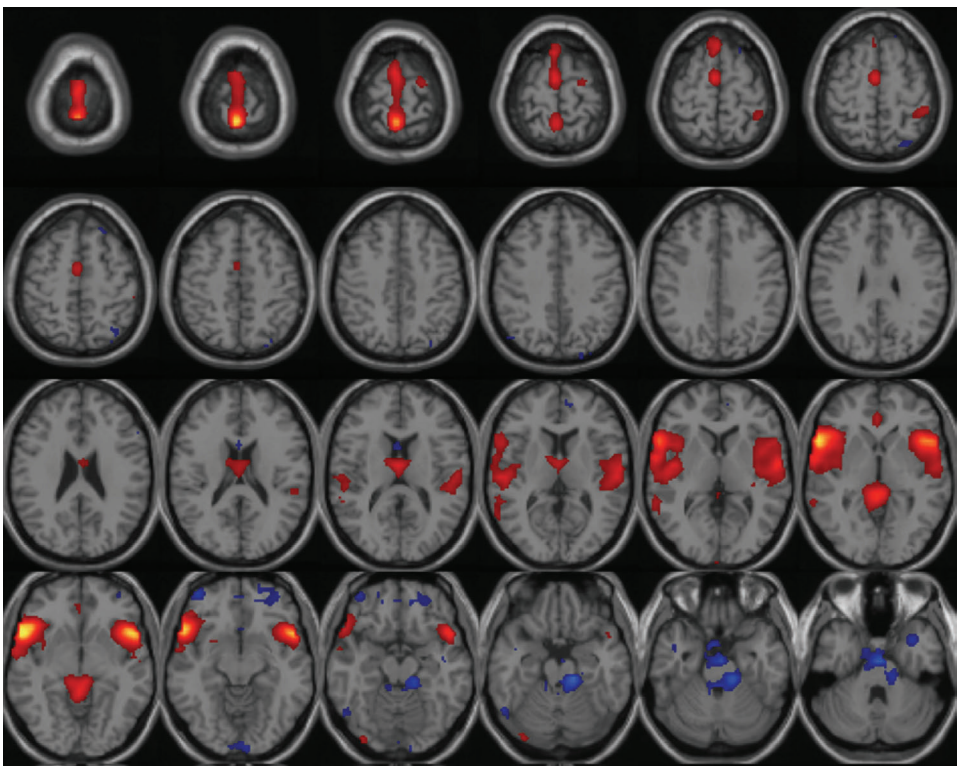


FIGURE 8 | Class mean run on five healthy, five schizophrenics and two images unknowns. Image shown is the mean image of group 1 thresholded at 1.0.

control mechanism for the end-user to flag problematic cases or guide automated procedures and subsequent users of the data with pertinent information that may otherwise be lost, such as protocol deviations, subject noncompliance, poor data acquisition, etc.

We have developed a capability to handle queries within and across research studies for all data domains stored in the MRN database. The user is able to select assessment criteria, such as the instrument, the visit type, the field, an operator and a value. The result is a filtered list of subjects for which the user is then prompted for what data they want reported. Following that step, the user is able to select and report all of the data sources available on that filtered list of subjects. Finally, the user may export the data in a format that suits their needs (CSV, SAS, SPSS, XCEDE, or custom format). This capability permits extensive customized querying, and given that it may be used across all data sources from all studies stored in the MRN database, it will prove to be an invaluable tool, forming the foundation for planned data mining activities. Once queries take place they can be saved and run again to reflect new subjects being available, and finally the queries can be leveraged to plan and execute meta-analyses across all subjects and research studies permitting analyses of individual data sources that were not envisioned by the investigator that may have collected the original data.

Within an active study, the NI system provides investigators with a tool to manage and evaluate the quality of their data through the structured protocol schema and its associated display. The user may evaluate their study both on a subject-by-subject basis and by viewing a summary of the study as a whole. In the near future, this tool will further lead to the implementation of automated quality control mechanisms that can flag suspicious data for review by a human expert and collect the results of their assessment. The image processing module is especially advantageous because it helps people to perform a quick group comparison and classification within the database and thus avoids the need to use analysis packages for doing these diagnostic tests. Furthermore, the database-driven analysis will grant a more detailed on-the-fly analysis of the quality of the existing data to provide insight into the progress of a given study as well as supporting the likelihood of a hypothesis proposed for future studies.

We believe this novel framework represents an enormous step toward the efficient mining of large scale heterogeneous translational neuroscience research. Data mining of such large NI repositories can lead to the creation of classifiers with the ability to perform diagnosis, predict treatment outcomes, and identify novel targets for pharmaceuticals. We provided a data mining example of classification, but current users are also using the NI system to perform clustering, regression, and associative rule learning. Ultimately this type of mining should hasten translation neuroscience discoveries to 1 day lead to better treatments, cures, and more complete understanding of the basic neurosciences.

ACKNOWLEDGMENTS

This research was supported by DOE DE-FG02-99ER6274, NIH 1R01EB006841, NIH U24RR021992, and NIH U54EB005149. The authors thank the Mind Clinical Imaging Consortium members for data collection and support of the scientists and engineers involved in the project.

REFERENCES

Amari, S.-I., Beltrame, F., Bjaalie, J. G., Dalkara, T., De Schutter, E., Egan, G. F., Goddard, N. H., Gonzalez, C., Grillner, S., Herz, A., Hoffmann, K. P., Jaaskelainen, I., Koslow, S. H., Lee, S. Y., Matthiessen, L., Miller, P. L., Da Silva, F. M., Novak, M., Ravindranath, V., Ritz, R., Ruotsalainen, U., Sebestra, V., Subramaniam, S., Tang, Y., Toga, A. W., Usui, S., Van Pelt, J., Verschure, P., Willshaw, D., and Wrobel, A. (2002). Neuroinformatics: the integration of shared databases and tools towards integrative neuroscience. *J. Integr. Neurosci.* 1, 117–128.

Andreasen, N. C., Arndt, S., Alliger, R., Miller, D., and Flaum, M. (1995). Symptoms of schizophrenia: methods, meanings, and mechanisms. *Arch. Gen. Psychiatry* 52, 341–351.

Aroian, L. A., and Levene, H. (1950). The effectiveness of quality control charts. *J. Am. Stat. Assoc.* 45, 520–529. Available at: http://www.jstor.org/stable/2280720.

Bly, B. M., Rebbechi, D., Hanson, S. J., and Grasso, G. (2004). The rumba software: tools for neuroimaging data analysis. *Neuroinformatics* 2, 71–100.

Bockholt, H. J., Ling, J., Scully, M., Magnotta, V. A., Gollub, R. L., White, T., Schulz, S. C., Lauriello, J., and Andreasen, N. C. (2007). *MIND Clinical Imaging Consortium as a Case Study of Novel Multi-Center Neuroinformatics Software*. Colorado Springs: International Congress on Schizophrenia Research.

Bota, M., and Arbib, M. A. (2004). Integrating databases and expert systems for the analysis of brain structures: connections, similarities, and homologies. *Neuroinformatics* 2, 19–58.

Bota, M., Dong, H.-W., and Swanson, L. W. (2005). Brain architecture management system. *Neuroinformatics* 3, 15–48.

Brinkley, J. F., and Rosse, C. (2002). Imaging and the human brain project: a review. *Methods Inf. Med.* 41, 245–260.

Carneiro, G., and Vasconcelos, N. (2005). Formulating semantic image annotation as a supervised learning problem. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2, 163–168.

Cole, E., Pisano, E. D., Clary, G. J., Zeng, D., Koomen, M., Kuzmiak, C. M., Seo, B. K., Lee, Y., and Pavic, D. (2006). A comparative study of mobile electronic data entry systems for clinical trials data collection. *Int. J. Med. Inform.* 75, 722–729.

Costa, L. da F. (2004). Bioinformatics: perspectives for the future. *Genet. Mol. Res.* 3, 564–574.

Demirci, O., Clark, V. P., Magnotta, V., Andreasen, N. C., Lauriello, J., Kiehl, K. A., Pearlson, G. D., and Calhoun, V. D. (2008). A review of challenges in the use of fMRI for disease classification/characterization and a projection pursuit application from multi-site fMRI schizophrenia study. *Brain Imaging and Behav.* 2, 207–226.

Farn, K. J., and Hu, S. L. (1995). *Practical Issues for RDBMS Application Development*. Proceedings of the 11th International Conference on Data Engineering, Taipei, 353.

Jovicich, J., Beg, M. F., Pieper, S., Priebe, C., Miller, M. M., Buckner, R., Rosen, B., and Birn, B. M.(2005). *Biomedical Informatics Research Network: Integrating Multi-Site Neuroimaging Data Acquisition, Data Sharing and Brain Morphometric Processing.* The 18th IEEE International Symposium on Computer-Based Medical Systems, Dublin, 288–293.

Kim, D., Manoach, D. S., Mathalon, D., Turner, J., Brown, G., Ford, J. M., Gollub, R. L., White, T., Wible, C. G., Belger, A., Bockholt, H. J., Clark, V. P., Lauriello, J., O'Leary, D., McCarthy, G., Mueller, B., Lim, K., Andreasen, N. C., Potkin, S., and Calhoun, V. D. (2009). Dysregulation of working memory and default-mode networks in schizophrenia during a Sternberg item recognition paradigm: an independent component analysis of the multisite Mind and fBIRN studies. *Hum. Brain Mapp.* 30, 3795.

Pace, W. D., and Staton, E. W. (2005). Electronic data collection options for practice-based research networks. *Ann. Fam. Med.* 3(Suppl. 1), S21–S29.

Prasad, B. E., Gupta, A., Toong, H. M. D., and Madnick, S. E. (1987). A microcomputer-based image database management system. *IEEE Trans. Ind. Electron.* 34, 83–88.

Segall, J. M., Turner, J. T., Van Erp, T., White, T., Bockholt, H. J., Gollub, R. L., Ho, B. C., Magnotta, V., Jung, R., McCarley, R., Schulz, S. C., Lauriello, J., Clark, V. P., Voyvodic, J., Diaz, M. T., and Calhoun, V. D. (2009). Voxel-based morphometric multi-site

collaborative study on schizophrenia. *Schizophr. Bull.* 35, 82–95.

Spiring, F., (2007). Introduction to statistical quality control (5th Edn.), by Douglas C. Montgomery. *Technometrics* 49 (1), 108–109. Available at: http://www.ingentaconnect.com/content/asa/tech/2007/00000049/00000001/art00026.

Sui, J., Adali, T., Pearlson, G., and Calhoun, V. D. (2009). An ICA-based method for the identification of optimal FMRI features and components using combined group-discriminative techniques. *Neuroimage* 46, 73–86.

Tofts, P., ed. (2004). *Quantitative MRI of the Brain: Measuring Changes Caused by Disease*. Chichester: John Wiley and Sons.

Toga, A. W. (2002). Neuroimage databases: the good, the bad and the ugly. *Nat. Rev. Neurosci.* 3, 302–309.

Vessey, J. A., Broome, M. E., and Carlson, K. (2003). Conduct of multisite clinical studies by professional organizations. *J. Spec. Pediatr. Nurs.* 8, 13–21.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 April 2009; paper pending published: 28 May 2009; accepted: 19 September 2009; published online: 21 April 2010.

Citation: Bockholt HJ, Scully M, Courtney W, Rachakonda S, Scott A, Caprihan A, Fries J, Kalyanam R, Segall JM, de la Garza R, Lane S and Calhoun VD (2010) Mining the Mind Research Network: a novel framework for exploring large scale, heterogeneous translational neuroscience research data sources. Front. Neuroinform. ***3****:36. doi: 10.3389/neuro.11.036.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 06 November 2009
doi: 10.3389/neuro.11.038.2009

Interactive exploration of neuroanatomical meta-spaces

Shantanu H. Joshi*, John Darrell Van Horn and Arthur W. Toga

Laboratory of Neuro Imaging, Department of Neurology, University of California, Los Angeles, CA, USA

Edited by:
Jan G. Bjaalie, International Neuroinformatics Coordination Facility, Sweden; University of Oslo, Norway

Reviewed by:
Shiro Usui, RIKEN Brain Science Institute, Japan
Jeffrey S. Grethe, University of California, USA
Jan G. Bjaalie, International Neuroinformatics Coordination Facility, Sweden; University of Oslo, Norway

****Correspondence:***
Shantanu H. Joshi, Laboratory of Neuro Imaging, Department of Neurology, University of California, Los Angeles, CA 90095, USA.
e-mail: sjoshi@loni.ucla.edu

Large-archives of neuroimaging data present many opportunities for re-analysis and mining that can lead to new findings of use in basic research or in the characterization of clinical syndromes. However, interaction with such archives tends to be driven textually, based on subject or image volume meta-data, not the actual neuroanatomical morphology itself, for which the imaging was performed to measure. What is needed is a content-driven approach for examining not only the image content itself but to explore brains that are anatomically similar, and identifying patterns embedded within entire sets of neuroimaging data. With the aim of visual navigation of large-scale neurodatabases, we introduce the concept of brain meta-spaces. The meta-space encodes pair-wise dissimilarities between all individuals in a population and shows the relationships between brains as a navigable framework for exploration. We employ multidimensional scaling (MDS) to implement meta-space processing for a new coordinate system that distributes all data points (brain surfaces) in a common frame-of-reference, with anatomically similar brain data located near each other. To navigate within this derived meta-space, we have developed a fully interactive 3D visualization environment that allows users to examine hundreds of brains simultaneously, visualize clusters of brains with similar characteristics, zoom in on particular instances, and examine the surface topology of an individual brain's surface in detail. The visualization environment not only displays the dissimilarities between brains, but also renders complete surface representations of individual brain structures, allowing an instant 3D view of the anatomies, as well as their differences. The data processing is implemented in a grid-based setting using the LONI Pipeline workflow environment. Additionally users can specify a range of baseline brain atlas spaces as the underlying scale for comparative analyses. The novelty in our approach lies in the user ability to simultaneously view and interact with many brains at once but doing so in a vast meta-space that encodes (dis) similarity in morphometry. We believe that the concept of brain meta-spaces has important implications for the future of how users interact with large-scale archives of primary neuroimaging data.

Keywords: meta-analysis, neuroanatomical data mining, visual data mining, 3D visualization

INTRODUCTION

The past decade has seen an explosive rise in the volume of brain image scans for clinical, diagnostic as well as research purposes. Fortunately, the neuroimaging research community recognized early on that facilitating data sharing among collaborative research centers is the key to boosting neuroscientific knowledge and discovery. Drawing a parallel with genomics research which has immensely benefitted with such data sharing strategies, a position paper (Eckersley et al., 2003) even goes far to suggest the use of public domain licensing policies, not unlike the GNU public license, for neuroscience data. The consensus on the archiving and sharing of primary neuroimaging data has fostered several large-scale initiatives: The Biomedical Informatics Resource Network (BIRN), the Morphometry and Function BIRN testbed projects (Grethe et al., 2005); The NIH MRI Study of Normal Brain Development (Pediatric MRI Study) and resulting Pediatric MRI Data Repository (Evans, 2006); and The fMRI Data Center (fMRIDC) (Van Horn et al., 2001; Van Horn and Gazzaniga, 2002). Much recently, the Neuroscience Information Framework (NIF) (Hurd, 2005) has initiated the development of a comprehensive experimental, clinical and translational databases, knowledge bases, atlases etc for processing, analysis, or simulation of brain data. Additionally the Clinical Data Interchange Standards Consortium (CDISC) (Souza et al., 2007) strives to improve data exchange across multiple domains and platforms for medical research as well as health care initiatives.

Notably, the LONI Image Data Archive (IDA) contains neuroanatomical data from nearly 30 research projects and serves as the primary repository for large studies such as the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data sharing has also indirectly benefitted and affected computational neuroscientific tool development. As algorithms get tested on more and more diverse datasets, they evolve to become more general and robust. Data sharing has been the first step in neuroinformatics research efforts and has largely been the focus of the past decade, and will continue to be so. The neuroscientific community is now getting ready to prepare for the next logical step – database integration (Forsberg and Roland, 2008). Most data storage facilities like the ones above, have implemented centralized repositories in proprietary formats. The challenge that the informatics community faces in the near future is the unification of existing large, heterogeneous neurodatabases in a user-transparent manner. This goes above and beyond

data sharing, where the user can not only access a single database, but can sift through multiple repositories at once without having the database to be localized in a central place. This is very much like the WWW, where there is an interconnection of data processing and storage nodes in a decentralized network. An important step towards this goal will be designing and standardizing robust database exchange protocols, while maintaining compatibility with privacy regulations and laws.

However, an alternate parallel goal complementing these efforts is the ability to *graphically* navigate, browse and query such aggregations of repositories. With the increasing progress in computational processing, and visualization, textual queries and interactions continue to be a severe drawback in future database access, especially with the enormity of the data involved. Recently, Herskovits et al. (Herskovits and Chen, 2008) have developed an open source implementation for a database system with data mining capabilities for managing, querying, analyzing and visualizing brain-MR images. We anticipate a compelling need for similar tools in the neuroscience community that facilitate informatics-driven approaches for users to better examine databases and explore the inter-relatedness of subjects in the population. Our goal then is to facilitate the large-scale informatics, mining, and visualization of the contents of existing neuroimaging data repositories by developing streamlined data processing workflows to decompose the contents of an archive, compare each image volume against all others in the archive, and visually display the results in a user friendly client application. We claim that the neuroimaging data itself can form the basis for such mining, that visualization of how brains relate to one another carries essential information, and that well-designed tools can permit data from outside the archive to be used as the basis for similarity-based searching.

This paper is organized as follows: the Section "Introduction" makes an argument for visual explorative interfaces for large-scale neuroimaging databases. The Section "Materials and Methods" outlines the main idea of this paper. It proposes neuroimaging workflows (see Introduction) focusing towards discriminative analysis for visualization. The Section "Materials and Methods" introduces the concept of a neuroanatomical meta-space built on top of the dissimilarity measures generated by the workflows. A meta-space is constructed in a case study (see Introduction) on a sample dataset of 400 subjects from the ADNI dataset. Finally Section "Discussion" proposes a 3D visualization environment for interactively navigating through this meta-space followed by a discussion.

NEED FOR VISUAL MINING OF NEURODATABASES

There is a growing interest in content-based searches for neuroimaging because of the limitations inherent in meta-data-based systems (Nielsen et al., 2006), as well as the large range of possible uses for efficient image retrieval. Without the ability to examine image content, searches currently rely on meta-data such as captions or keywords, which may be laborious or expensive to produce manually. While textual information about images can be easily searched using existing technology, it requires humans to personally tag and annotate every image in the database. This can be impractical for very large databases. Similarly, there are added benefits for manipulating the search criteria and results visually. A visual interface will present an opportunity to cluster, classify, and graphically represent data in ways not possible based on textual meta-data alone. We identify the following scenarios where such a graphical navigation system can be applied.

Visualizing anatomical differences and relatedness simultaneously

A single brain image scan may give rise to a variety of anatomies. Pertaining to a specific neuroscientific study, researchers may choose to directly work with MRI images, or work with suitable anatomical representations deconstructed from an MRI image. For e.g. the boundary of the volume, the cerebral cortex can be represented by a topographic two-dimensional geometrical structure (Thompson et al., 2001; Hinds et al., 2008). This structure can be further differentiated by the anatomical folds also known as the sulci and the gyri. One can further descend beneath the cortex to delineate various other structures such as the limbic system, thalamus, hypothalamus, corpus callosum (Narr et al., 2005) etc. Existing neuroimaging analysis and visualization tools restrict users to a single, individual brain image or surface for anatomical studies. While this is useful for structural analysis or evaluation of pertinent anatomies, neuroimaging studies often consist of large population of subjects and resulting brain images. Especially for large-scale statistical or discriminative analyses focusing on disease, genetic, or heritable effects and changes according to neuromorphology, it would be useful to simultaneously visualize the morphology in an appropriate metric space resulting from the analysis. Currently most neurodatabases are accessible solely by textual queries. Furthermore there is no existing application or workflow that enables the neuroscientist to manipulate neuroinformatics search criteria, and the resulting queries and outputs in a visual manner.

Educational resource or a training environment for neuroscientists

Developments in the area of content representation, interaction, and search has been employed for graphical data with the notable example of Microsoft's Photosynth that been used to mine the Flickr[1] photo sharing site to then graphically depict a collection of images from a spatial reconstruction of their taken vantage point. Likewise, Google Earth[2] displays satellite imagery, mapping, and geographic data, permitting interactive search, annotation, and other functions. In the astronomy community, the recent launch of the World wide Telescope® [for historical context, see (Szalay and Gray, 2001)] has revolutionized the exploration and search capabilities for astral, galactic, and planetary data obtained from multiple imaging sources. These applications continue to enhance educational instruction, both for the general public and the specialists alike. Similar tools do not yet exist in the neuroimaging community where there is a tremendous potential for computer-simulated training for neuroscientists.

Visual cataloging of neurodatabases

Visual data mining (VDM) is useful in exploratory analysis, where one has limited views and information of the data. With the recent advances in computing and storage, VDM has been

[1]http://www.flickr.com/

[2]http://earth.google.com/

used for diverse applications such as exploring geospatial data (Keim, 2002; Keim et al., 2004), internet web resource databases (Chen et al., 2007), and analyzing business intelligence patterns (Hao et al., 2000). Such an effort currently does not exist in the field of neuroimaging. The development of visual catalogs of neuroimaging data would enable and enhance large-scale scientific interaction among users. Though some basic image viewing tools exist, we believe a different approach is needed altogether. A content-based solution is beneficial for researchers to more easily examine (dis)similarity between brains and to dynamically visualize patterns that may be indicative of the demographic and clinical attributes of the data themselves. By navigating through a virtual environment via an easy-to-use, web driven application, users will be able to examine large collections of brain data using only their computer mouse.

Visualization of atlas spaces

Individual brain anatomies, have their own local coordinate systems that measure local distortions of features such as curvatures, intensities, and surface areas. For large populations of such anatomies, most approaches construct an atlas template (Mazziotta et al., 2001) and transform all individuals to the atlas. This yields a single anatomical object that is then analyzed or visualized as a representative of the population. This approach also transforms the individual local variation to the atlas thus providing the researcher with an at-a-glance view of the variation across population. The drawback of atlas visualization is that the atlas depicts a single view of the population, and it is difficult to get an overview of the underlying dissimilarity patterns between individual subjects in the study. Often, these atlases are probabilistic in nature and thus only provide a statistical interpretation of the relationship between the template and the individual. Thus one has to continually go back and forth between the template and the individual to relate to, and observe the changes in the native brain space. Instead, a visualization scheme that simultaneously displays the atlas and the data used for its construction, in a meta-space is highly desirable. Moreover, one could technically extend this idea to multiple atlases grouped together with their respective populations.

MATERIALS AND METHODS

This section outlines the concept of a meta-space that follows from large-scale discriminative analyses on a brain population. Essential to the construction of the meta-space is the data processing framework that enables complete workflows leading from the original data in the form of images to the various metrics that attempt to classify, cluster and separate individuals in the population. Due to the enormity of the data, as well as the types of processing involved, we employ a grid-based execution environment. While large-scale distributed processing is an essential component in scientific computing, it has only recently (Rex et al., 2003; Callahan et al., 2009) evolved to adapt itself to biomedical or neuroimaging workflows. The main hindrance for adapting such technologies is the specialized knowledge required to maintain, develop, and execute applications for common neurocomputing tasks. However with latest advances in interfaces and visualization, much of these tasks have become oblivious to the end-user. For e.g. the LONI Pipeline (Rex et al., 2003)[3] is a highly flexible, distributed computing environment that enables parallelized execution of application software especially dealing with brain mapping protocols. It offers an efficient GUI interface to the user, where one can quickly build complete applications using pluggable components called pipeline modules. The pipeline user can further extend the functionality of the pipeline by developing modules in addition to the ones existing in the library. The pipeline communicates back and forth with the grid scheduler to queue up user tasks in an efficient manner. It also provides a feedback mechanism to the user where he can monitor program execution real time from the pipeline interface. Moreover the pipeline also allows the data as well as programs required for analysis, to reside on the user's local machine that launches the pipeline thereby integrating both local and remote resources in a seamless manner. Throughout this discussion, the LONI Pipeline will serve as a convenient execution environment for our architecture. We would also like to stress that the user is not restricted to the LONI computing infrastructure to take advantage of the LONI pipeline. The LONI pipeline is independent of the underlying grid computing environment and can be tailored and adapted to other suitable execution infrastructures[4].

DATA MINING WORKFLOW FOR DISCRIMINATIVE ANALYSIS

The data is typically stored as MRI images either in the Analyze or the NIFTI format. For the purpose of this discussion, we focus on neuroanatomical volumes, though in the future functional imaging could be incorporated. Depending upon the experimental setup, the data is usually corrected to minimize geometric distortions or non-linearities, and any non-uniform intensities resulting due to magnetic properties of the RF coils. The data can further be sharpened using histogram techniques that can further lead to a reduction of intensity non-homogeneities. The corrected MR images are then stripped of skulls, unwanted tissue, and other extra unneeded anatomical features such as the cerebellum or the brain stem. We used the Brain Extraction Tool (Smith, 2002) tool for skull stripping MRI images in our workflow, although any such similar tool can be used. All image volumes in the database are registered (Woods et al., 1998) to a standard Montreal Neurological Institute (MNI) atlas image. The resulting gray/white matter image is then processed in parallel to i) extract the cortical (gray/CSF boundary) surface (Shattuck and Leahy, 2002), and ii) extract about 56 sub-cortical features (Tu et al., 2008) such as the major gyri, hippocampus, the putamen, etc. This process exclusively gives rise to a geometrical representation that is stored in the form of a triangular mesh using a suitable file format. Henceforth in this paper, brain anatomies will be taken to mean the cortical surface as well as surface parameterizations of the individual sub structures beneath the cortex. The top portion of **Figure 1** shows the pre-processing and feature extraction steps. These steps are implemented as completely automated LONI pipeline modules. **Figure 2** shows an example of a parcellated volume colored according to different anatomies. The original input data

[3]http://pipeline.loni.ucla.edu

[4]For more information about the pipeline, the reader is referred to the article "Efficient, Distributed and Interactive Neuroimaging Data Analysis using the LONI Pipeline" by Dinov et al., 2009.

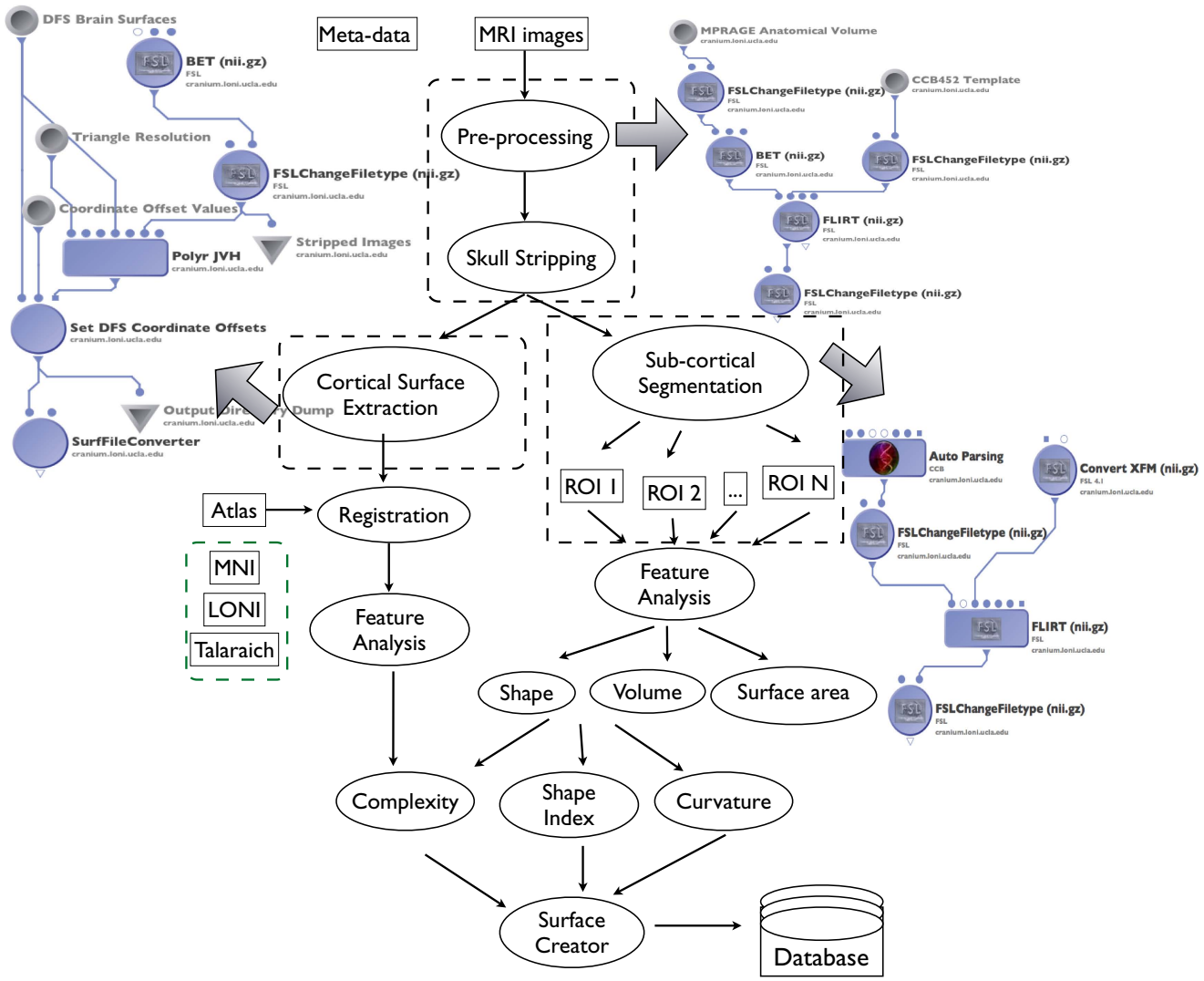


FIGURE1 | A schematic of the data mining workflows exposed through the LONI pipeline (Rex et al., 2003). The workflow is divided into three parts, i) Processing, ii) Feature Extraction – extracting anatomical features such as cortical surfaces, sub-cortical structures etc. and having 3D mesh representations for each feature, and iii) Feature Analysis – calculating the local curvature, shape index, cortical complexity, and encoding each surface mesh with these attributes. Each stage is implemented via pipeline without user intervention.

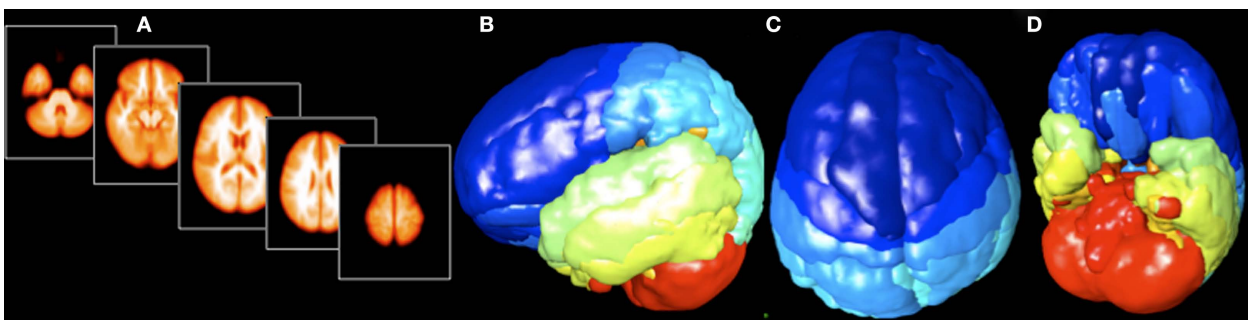


FIGURE 2 | Surface rendering of segmented sub-cortical structures labeled according to regions. (A) Examples of image slices along the axial view. **(B–D)** Parcellated cortical and sub-cortical regions along three views.

is usually accompanied by appropriate meta-tags using predefined XML schemas. The above pre-processed data is then stored hierarchically for a streamlined access in a database.

A NEUROANATOMICAL META-SPACE

The central idea behind atlas meta-spaces is the modeling of the dissimilarities between individuals in a population. The population can be analyzed all at once, or the individual subjects can be grouped according to some well-defined categories. Before performing any type of discriminatory analysis, one needs to consider an appropriate metric space of objects and define a notion of distance associated with it. Metric spaces are mathematically easier to define in case of simple tractable objects such as two-dimensional, or three-dimensional points, multidimensional vector measurements, or objects that lend themselves to functional representations in a well-defined space. In case of neuroanatomical structural data, such as cortical surfaces and sub-cortical structures, it is difficult to have a rigorous definition of a metric space of such entities directly. There are ongoing research efforts to model the geometry of the cortex, or define the shape of three dimensional closed surfaces corresponding to the sub-cortical structures. The goal is to have a mathematical representation of the shape geometry, independent of indeterminacies such as the scale, position, orientation etc. Various researchers have used harmonic functions on a sphere to represent closed genus zero surfaces. In this case the shape distance is measured by a $\mathbb{L}^2$ distance between the coefficients in the space of harmonic functions. Others have used level set representations for shapes of surfaces, again using the $\mathbb{L}^2$ metric between two signed distance representations for surfaces. Yet another approach by researchers uses global measurements such as the volume, average curvature, or the surface area of cortex, or the sub-cortical structures. Although, this may interpreted as a gross simplification of brain geometry, numerous studies have shown the effectiveness of such simple metrics in capturing a global underlying pattern of the data. A study based on simple volume analysis of the Hippocampus (Chupin et al., 2008) was able to correctly classify 82% of Alzheimer's disease (AD) patients with respect to the elderly controls. Another study (Gosche et al., 2002) has also shown that hippocampal volume can be used as an indicator for Alzheimer neuropathology. A recent study (Dubois et al., 2007) also supports that quantitative volumetric analysis on the hippocampus was able to distinguish AD across young and old ages. We will follow a similar approach and utilize metrics that are simple to compute, and lend themselves to an easy interpretation. A few of the metrics considered in this paper are cortical complexity, shape indices, volume and surface area of the segmented structures. We then represent these quantities in the space of real numbers and adopt the standard Euclidean metric. As a specific example, for a database with N subjects and L delineated sub-volumes, given by $\{V_k^i\}, i = 1,\ldots,N,\ k = 1,\ldots,L$ we first normalize all the volumes to have a unit scale. We then calculate a N-by-N distance matrix given by,

$$D(S_i, S_j) = \frac{\sum_{k=1}^{L} \left(V_k^i - V_k^j\right)^2}{L} \quad (i, j = 1\ldots, N, i \neq j) \tag{1}$$

Likewise, one can in practice utilize different metrics to generate different distance matrices. In order to establish a frame-of-reference, we construct a template atlas for the population and also include it in the distance calculation. This process yields a distance space of neuroanatomical structures with the atlas conveniently treated as the "origin". One can even define distance units in this space and define a centralized coordinate system with the atlas at the origin. This resulting distance space is extremely high dimensional and not straightforward to visualize. In the analysis stage, we will pre-store multiple such distance matrices based on different metrics for the above features in the database. As new data enters the LONI IDA, the workflow engine will automatically detect new entries and subject them to regional extraction, surface modeling, and regional measurement, etc. Random spot checking to ensure accuracy will help to reduce improper data from entering into and possibly biasing the comparison of image data-sets. Each new data set's relative distance from each of the brain volumes already in the overall distance matrix will be performed and this new information will take its place in the matrix. Upon updating of the distance matrices, the multidimensional scaling (MDS) will then be recomputed and the positions of each brain surface in the space adjusted accordingly. We expect that once fully deployed the continuous processing of new entries into the IDA and the updating of the geometric similarities will not require extensive computational loads or interfere with other jobs being processed on the LONI grid. Lastly, we will post the automated processing meta-algorithm via the community web forum so that others may download the workflow and use the LONI Pipeline on their own systems to validate results. Using the online web forum as well as client-side user interface tool, users will be able to post their reviews of the validity of meta-algorithm, note outlier subjects, or annotate interesting cases. These publicly given annotations will form additional meta-data information to be made available to other users of these tools.

In order to explore the dissimilarities between brain volumes, we need to project the dissimilarity matrix into an appropriate 2D or 3D space. There are numerous techniques to project high dimensional data into lower dimensional spaces for analysis or visualization. As discussed above, one could calculate principal coefficients, or principal factors explaining the maximum observed population variability in terms of a few determining factors. For visualization purposes, only the first 3 eigen projections can be used to display objects in a 3D space. Sophisticated visualization tools (Swayne et al., 2003) exist for performing such high dimensional data visualizations, as well as plotting multivariate statistics of the data. However, these tools usually represent objects by points in 3D space and thus limit the interaction with the original objects themselves. Moreover, since the construction of our meta-space relies on dissimilarities among neuro-structures, we will use the multidimensional scaling (Kruskal and Wish, 1978) approach for projecting the dissimilarity matrix into a 3D space. Multidimensional scaling is an optimization technique that projects a high dimensional dissimilarity matrix into a low dimensional space that most accurately represents the pair-wise distances between the objects. This is achieved by minimizing a cost function that minimizes a Euclidean cost between the original dissimilarity matrix and a set of low dimensional (3D in our case) vectors. Additionally, since most studies come equipped with meta-data tags along with the images, one can easily perform comparative analyses of individual brain locations with the mean brain locations for each categorical meta-data type. For example, in case of an Alzheimer's study, this implies that a brain whose standardized distance from the mean

AD patient location is smaller than the normal subject will increase the likelihood that the brain belongs to an AD patient. **Figure 3** shows an illustrative visualization of the meta-space after the MDS projection of pair-wise distances between a group of brains with respect to an atlas.

Case study for a subset of ADNI dataset

As a case study for our framework, we sampled the LONI IDA and identified three groups of subjects obtained from the ADNI dataset. Subjects included $N = 244$ mild cognitively impaired (MCI) subjects, $N = 56$ Alzheimer's Disease (AD) patients, and $N = 100$ normal control subjects, for an overall group of $N = 400$ neuroanatomical Magnetization-Prepared Rapid Acquisition Gradient-Echo (MPRAGE) image volumes. All images in this example were scanned using a 3T MR scanning platform, although, a mixture of data from across scanner manufacturers and field strengths would also be possible. The distance matrix computed using Eq. 1 is visualized in **Figure 4A**. This distance matrix is projected to a three dimensional space using MDS and the results is displayed in **Figure 4B**. The spheres are drawn with a radius equal to 5% of the standard deviation from the population mean. From the MDS analysis, we extracted the first three latent dimensions which accounted for more than 66% of the distance variation between subjects (50%, 10%, and 6%, respectively). No inferential statistical test thresholding (e.g. *T*-tests, *F*-tests, etc) was performed and no significance-levels were determined concerning the differences between groups as "group" variables were not specified *a priori*. Rather, all data were considered equally in terms of processing

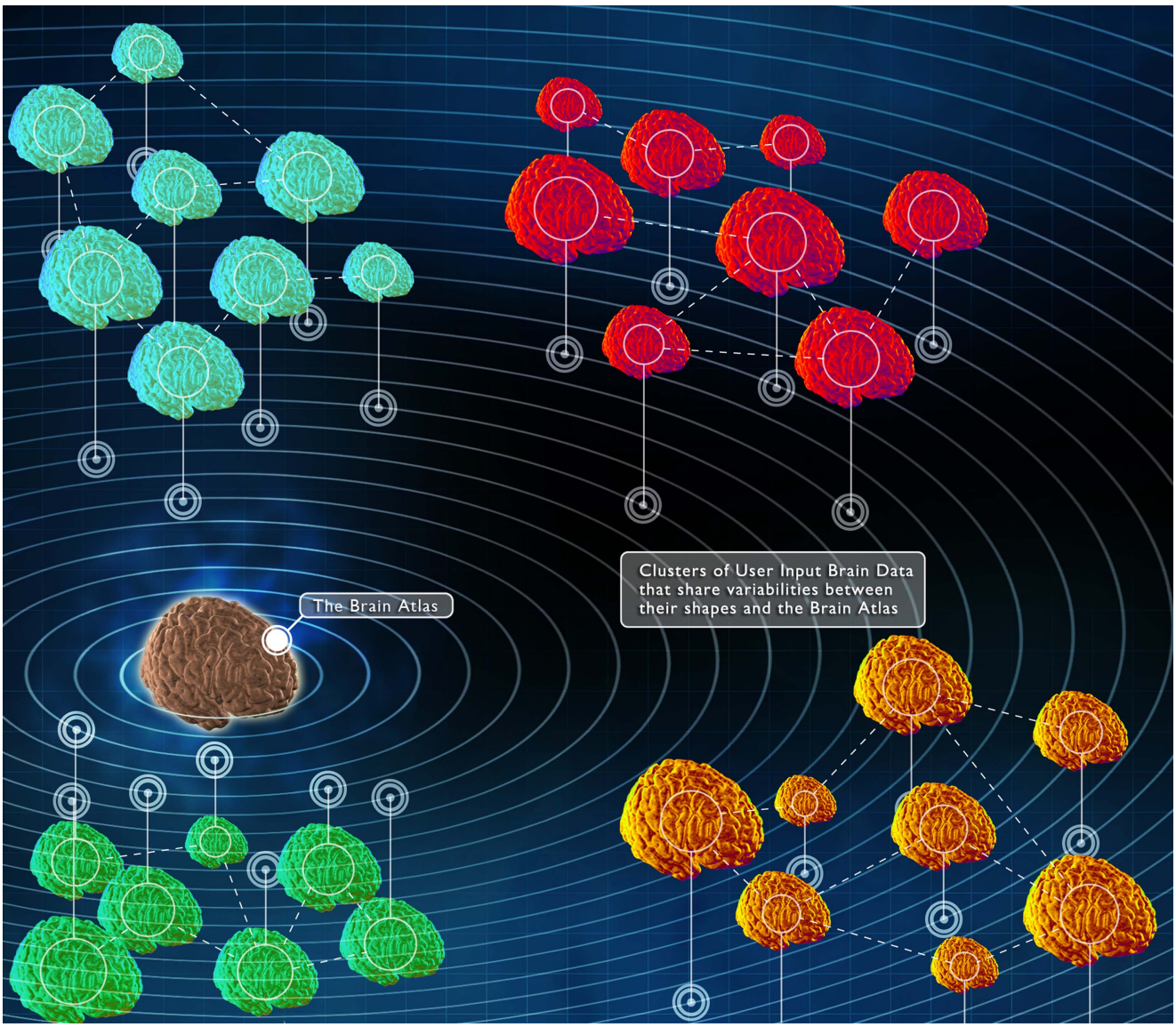


FIGURE 3 | An illustration of distributions of brain surfaces in an atlas meta-space. The atlas can be treated as the origin. The locations of the brain surfaces are derived using MDS applied to the distance matrix of discriminative features. A radial coordinate system is shown for convenience, in practice any other informative reference frame can be used.

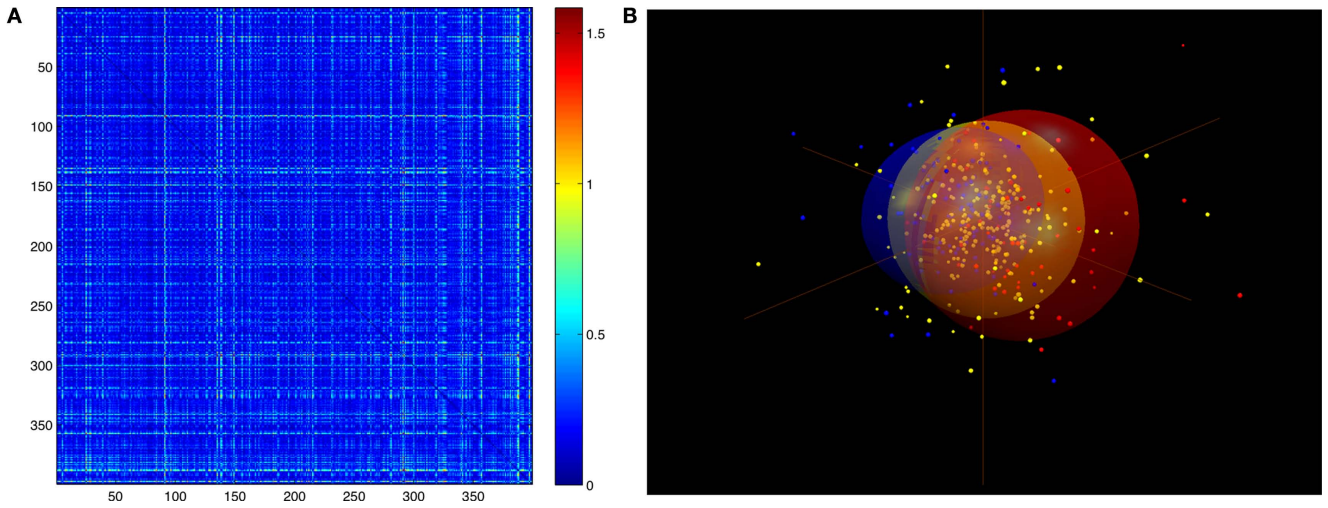


FIGURE 4 | (A) Visualization of the pair-wise distance matrix for N = 400 subjects of the ADNI dataset. **(B)** MDS projection of the dissimilarity matrix into 3D coordinates, each projection colored according to subject status, NC = blue, AD = red, MCI = yellow.

and MDS analysis. Subjects segregated maximally along the first principle axis. Along this dimension, normal subjects were clearly distinguishable from AD patients, whereas MCI patients were observed to overlap both normal and AD distributions. Each extracted brain surface was positioned in space by its MDS coordinate triad. This served to offset each brain from the origin, to position brains that are similar near one another, and those that are dissimilar far from each other. In this way a user can graphically examine similar brains to identify similar meta-data characteristics from those brains that might have bearing on etiology of disease, demographic factors, etc. The intention in processing these data in this manner was to determine and demonstrate whether the imaging data, based upon the characteristics of content-based regional brain geometry, would separate themselves in a manner that would be obvious to an end-user. Other metrics, however, besides the profile of regional volumes, might also segregate subjects equally well or even better. Different metrics of distances can lead to differing patterns of results. This scheme sets the stage for systematic evaluation of which metric discriminate between subjects best, best express individual variability, or classify subjects to heretofore unappreciated classifications based on anatomical similarity, meta-data factors, etc.

INTERACTIVE VISUALIZATION OF NEURO- META-SPACES

Finally, we provide the user with a fully interactive 3D exploration experience by allowing visual navigation of the meta-space. As seen earlier, the meta-space comprises of 3D projections of pair-wise distances among the population along with the atlas template. Intuitively, this can also be thought of as scaling, stretching and collapsing a set of 3D points (corresponding to brain anatomies) such that their pair-wise distances in the higher dimensional space are accurately approximated. Thus we have the target locations for all brain anatomies after the MDS procedure converges. We now simply scale and translate the corresponding brain cortical surfaces extracted in the data mining workflows to the appropriate location in the 3D space. The end result is a graphical rendering of a large volume of brain surfaces all at once. The visualization display is dynamic, thus enabling the user to rotate, zoom, and pan the view in real time. Additionally, the user can also navigate through the meta-space, thus discovering and verifying the brain surface geometry simultaneously in relation to it's neighbors. Each brain surface is accompanied by an XML description of its meta-data that can be quickly displayed on the screen to get more information about the individual brain.

A growing challenge to the visualization environment is the rapidly accumulating data. Both long-term storage and memory requirements for data multiply progressively with increase in the sheer data volume. Real time visualization of large data-sets presents numerous difficulties with regards to limited processing power and computer memory. For e.g. a triangular mesh parameterization of a moderate resolution cortical surface roughly includes 250 K triangles and 100 K vertices. A floating point representation for the geometry alone requires about 1.5 MBytes of storage, while attributes such as colors and normals are represented at an additional cost. For a brain volume database in excess of 500 brains, the storage requirements start becoming prohibitive for any real time manipulation of data. For this reason, it is necessary to represent the data in a multi-resolution manner. There is an ongoing research effort in the area of triangular mesh simplification for visualization or compression for storage purposes. For our visualization interface, we have implemented the quadric error mesh simplification strategy (Garland and Heckbert, 1997) that keeps on contracting edges defined by vertex-pairs until the desired number of faces are achieved. The multi-resolution representation and rendering enables faster response times, and facilitates better user interaction. Currently the surface geometry is stored as triangular meshes with faces, vertices, and colors. We also allow surfaces to be colored according to attributes for each vertex. These can represent measures such as cortical thickness, functional activity, or other statistics. **Figure 5**

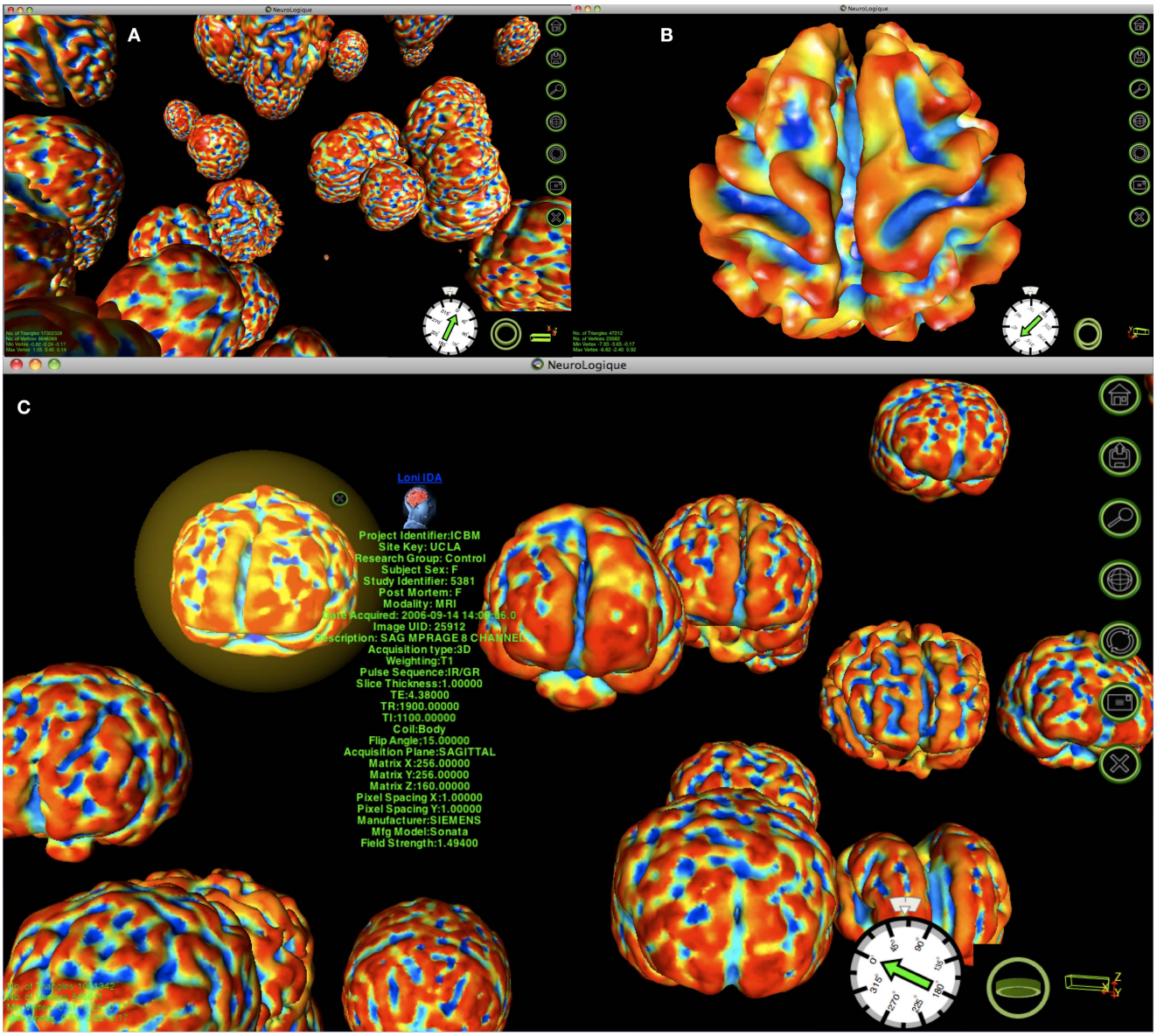


FIGURE 5 | A snapshot of the 3D Visualization environment for the neuro- meta-space displaying a group of brain surfaces from the ADNI 400 dataset. **(A)** A zoomed-in view of the 3D interface. **(B)** A close-up of an individual brain belonging to the AD category. **(C)** Alternate view of the interface with the meta-data (green text) displayed, as a result of a right-click action on one of the brains.

shows the functioning prototype of our visualization interface. The visualization environment is a desktop-based application designed in C++ and Open GL® and is available on all Windows, Mac OS X, and Unix-based platforms. The OpenGL pipeline conveniently provides a built-in framework for polygonal rendering and transformations.

DISCUSSION

We foresee the development of graphical visualization tools that enable and enhance scientific interaction with large-scale databases, as the next step in neuroimaging informatics. Though some basic image viewing tools exist, we have argued for a need for a next generation visual interaction framework. We have also demonstrated a content-based solution that can be applied to any such archive in order for researchers to more easily examine dissimilarity between brains and to dynamically visualize patterns in the degree of proximity between brains that may be indicative of the demographic and clinical attributes of the data themselves. In fact, all throughout our approach, we have made as few assumptions about the data as possible, and really let the data segregate itself based upon the characteristics of regional shape and geometry. A key component of this framework is the fully interactive, 3D visualization environment. By navigating through a virtual environment via an easy-to-use, web driven application, users will be able to examine large collections of

brain data using only their computer mouse. The underlying data distribution manifested through classification and collocated with the respective brain anatomies would be a very valuable tool for data processing, mining and interactive visualization of large-scale neuroanatomical databases. This will form a common frame-of-reference for neuroimaging informatics that is (a) familiar to most neuroimaging scientists, (b) provides a navigable space in which to position brain data, and (c) allows measurement of brain dissimilarity to be visually represented.

Our plan now is to (i) apply this meta-workflow to the thousands of MR anatomical images contained in the LONI IDA to obtain cortical surface and partition shape statistics, (ii) measure the pair-wise distances between the shapes obtained from the individual MR volumes, (iii) apply multidimensional scaling (MDS) and related decompositions of the matrix of pair-wise distances to determine which brains are most related, and (iv) broaden the concept of the standard brain atlas space to extend beyond the boundaries of the atlas to form a large space, analogous to a celestial coordinate system, wherein the atlas is centered at the origin and the individual brain surface representations are distributed in clusters with respect to it. We also plan on enhancing the user interface and scaling its performance with the increasing data.

ACKNOWLEDGMENTS

This research was partially supported by the National Institute of Health through the National Center for Research Resources (NCRR) for Center for Computational Biology (CCB), Grant U54 RR021813. Additional support was provided by Award Number RC1MH088194 from the National Institute of Mental Health. This study utilized the LONI Pipeline environment (http://pipeline.loni.ucla.edu), which was developed by the Laboratory of Neuro Imaging and partially funded by NIH grants P41 RR013642, R01 MH71940 and U54 RR021813. Additionally, we thank Mr. Vaughn Greer, Laboratory of Neuro Imaging for the graphic in **Figure 3**. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

REFERENCES

Callahan, M., Cole, M. J., Shepherd, J. F., Stinstra, J. G., and Johnson, C. R. (2009). A meshing pipeline for biomedical computing. *Eng. Comput.* 1, 115–130.

Chen, J., Zheng, T., Thorne, W., Zaiane, O. R., and Goebel, R. (2007). Visual data mining of web navigational data. In IV '07: Proceedings of the 11th International Conference Information Visualization, Washington, DC, IEEE Computer Society, pp. 649–656.

Chupin, M., Chetelat, G., Lemieux, L., Dubois, B., Garnero, L., Benali, H., Eustache, F., Lehericy, S., Desgranges, B., and Colliot, O. (2008). Fully automatic hippocampus segmentation discriminates between early Alzheimer's disease and normal aging. In th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008., pp. 97–100.

Dinov, I., Van Horn, J. D., Lozev, K. M., Magsipoc, R., Petrosyan, P., Liu, Z., MacKenzie-Graham, A., Eggert, P., Parker, D. S., and Toga, A. W. (2009). Efficient, distributed and interactive neuroimaging data analysis using the LONI pipeline. *Front. Neuroinform.* 3, doi:10.3389/neuro.11.022.2009.

Dubois, B., Feldman, H. H., Jacova, C., Dekosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J., and Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol.* 8, 734–746.

Eckersley, P., Egan, G., DeSchutter, E., Yiyuan, T., Novak, M., Sebesta, V., Matthiessen, L., Jaaskelainen, I., Ruotsalainen, U., Herz, A., Hoffmann, K.-P., Ritz, R., Ravindranath, V., Beltrame, F., Amari, S.-i., Usui, S., Lee, S.-Y., van Pelt, J., Bjaalie, J., Wrobel, A., daSilva, F., Gonzalez, C., Grillner, S., Verschure, P., Dalkara, T., Bennett, R., Willshaw, D., Koslow, S., Miller, P., Subramaniam, S., and Toga, A. (2003). Neuroscience data and tool sharing. *Neuroinformatics* 2, 149–165.

Evans, A. C. (2006). The NIH MRI study of normal brain development. *Neuroimage* 1, 184–202.

Forsberg, L. and Roland, P. (2008). 1st incf workshop on neuroimaging database integration. In Nature Precedings. available at: http://dx.doi.org/10.1038/npre.2008.1781.1/

Garland, M. and Heckbert, P. (1997). Surface simplification using quadric error metrics. In SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, New York, NY, ACM Press/Addison-Wesley Publishing Co, pp. 209–216.

Gosche, K. M., Mortimer, J. A., Smith, C. D., Markesbery, W. R., and Snowdon, D. A. (2002). Hippocampal volume as an index of Alzheimer neuropathology: Findings from the nun study. *Neurology* 10, 1476–1482.

Grethe, J. S., Baru, C., Gupta, A., James, M., Ludaescher, B., Martone, M. E., Papadopoulos, P. M., Peltier, S. T., Rajasekar, A., Santini, S., Zaslavsky, I. N., and Ellisman, M. H. (2005). Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. *Stud. Health Technol. Inform.* 112, 100–109.

Hao, M. C., Dayal, U., and Hsu, M. (2000). Visual data mining for business intelligence applications. In WAIM '00: Proceedings of the First International Conference on Web-Age Information Management, London, Springer-Verlag, pp. 3–14.

Herskovits, E. H., and Chen, R. (2008). Integrating data-mining support into a brain-image database using open-source components. *Adv. Med. Sci* 2, 172–181.

Hinds, O., Polimeni, J. R., Rajendran, N., Balasubramanian, M., Wald, L. L., Augustinack, J. C., Wiggins, G., Rosas, H. D., Fischl, B., and Schwartz, E. L. (2008). The intrinsic shape of human and macaque primary visual cortex. *Cerebral. Cortex* 11, 2586–2595.

Hurd, N. A. (2005). Neuroscience Information Framework.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph.* 1, 1–8.

Keim, D. A., Panse, C., Sips, M., and North, S. C. (2004). Visual data mining in large geospatial point sets. *IEEE Comput. Graph. Appl.* 5, 36–44.

Kruskal, J. B., and Wish, M. (1978). Multidimensional scaling. Beverly Hills, CA, Sage Publications.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., Holmes, C., Collins, L., Thompson, P., MacDonald, D., Iacoboni, M., Schormann, T., Amunts, K., Palomero-Gallagher, N., Geyer, S., Parsons, L., Narr, K., Kabani, N., LeGoualher, G., Boomsma, D., Cannon, T., Kawashima, R., and Mazoyer, B. (2001). A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 1412, 1293–1322.

Narr, K. L., Bilder, R. M., Toga, A. W., Woods, R. P., Rex, D. E., Szeszko, P. R., Robinson, D., Sevy, S., Gunduz-Bruce, H., Wang, Y., DeLuca, H., and Thompson, P. M. (2005). Mapping cortical thickness and gray matter concentration in first episode schizophrenia. *Cerebral. Cortex* 6, 708–719.

Nielsen, F. A., Balslev, D., and Hansen, L. (2006). Data mining a functional neuroimaging database for functional segregation in brain regions. In Danske Konference i Mønstergenkendelse og Billedanalyse.

Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The LONI pipeline processing environment. *Neuroimage* 3, 1033–1048.

Shattuck, D. W., and Leahy, R. M. (2002). Brainsuite: an automated cortical surface identification tool. *Med. Image Anal.* 2, 129–142.

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 3,. 143–155.

Souza, T., Kush, R., and Evans, J. P. (2007). Global clinical data interchange standards are here! *Drug Discov. Today* 3–4, 174–181.

Swayne, D. F., Buja, A., and Lang, D. T. (2003). Exploratory visual analysis of graphs in ggobi. In CompStat: Proceedings in Computational Statistics, 16th Symposium, pp. 477–488.

Szalay, A., and Gray, J. (2001). The World – Wide Telescope. *Science* 5537, 2037–2040.

Thompson, P. M., Mega, M. S., Woods, R. P., Zoumalan, C. I., Lindshield, C. J., Blanton, R. E., Moussai, J., Holmes, C. J., Cummings, J. L., and Toga, A. W. (2001). Cortical change in alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral. Cortex* 1, 1–16.

Tu, Z., Narr, K. L., Dollar, P., Dinov, I., Thompson, P. M., and Toga, A. W. (2008). Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE Trans. Med. Imaging* 4, 495–508.

Van Horn, J. D., and Gazzaniga, M. S. (2002). Opinion: Databasing fMRI studies towards a 'discovery science' of brain function. *Nat. Rev. Neurosci.* 4, 314–318.

Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., and Gazzaniga, M. S. (2001). The functional magnetic resonance imaging data center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 1412, 1323–1339.

Woods, R. P., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. (1998). Automated image registration: I. general methods and intrasubject, intramodality validation. *J. Comput. Assist. Tomogr.* 1, 139–152.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 April 2009; paper pending published: 24 June 2009; accepted: 05 October 2009; published online: 06 November 2009. Citation: Joshi SH, Van Horn JD, and Toga AW (2009) Interactive exploration of neuroanatomical meta-spaces. Front. Neuroinform. ***3****:38. doi: 10.3389/neuro.11.038.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 20 July 2009
doi: 10.3389/neuro.11.022.2009

Efficient, distributed and interactive neuroimaging data analysis using the LONI *Pipeline*

***Ivo D. Dinov*[1]*, John D. Van Horn[1], Kamen M. Lozev[1], Rico Magsipoc[1], Petros Petrosyan[1], Zhizhong Liu[1], Allan MacKenzie-Graham[1], Paul Eggert[2], Douglas S. Parker[1,2] and Arthur W. Toga[1]**

[1] *Laboratory of Neuro Imaging, University of California, Los Angeles, CA, USA*
[2] *Department of Computer Science, University of California, Los Angeles, CA, USA*

Edited by:
John Van Horn, University of California, Los Angeles, CA, USA

Reviewed by:
Yong Zhao, University of Chicago, Chicago, IL, USA
Stephen C. Strother, Baycrest, Toronto, ON, Canada; University of Toronto, Toronto, ON, Canada

***Correspondence:**
Ivo D. Dinov, Laboratory of Neuro Imaging, David Geffen School of Medicine at UCLA, 635 S. Charles Young Drive, Suite 225, Los Angeles, CA 90095-7334, USA.
e-mail: dinov@loni.ucla.edu

The LONI *Pipeline* is a graphical environment for construction, validation and execution of advanced neuroimaging data analysis protocols (Rex et al., 2003). It enables automated data format conversion, allows Grid utilization, facilitates data provenance, and provides a significant library of computational tools. There are two main advantages of the LONI *Pipeline* over other graphical analysis workflow architectures. It is built as a distributed Grid computing environment and permits efficient tool integration, protocol validation and broad resource distribution. To integrate existing data and computational tools within the LONI *Pipeline* environment, no modification of the resources themselves is required. The LONI *Pipeline* provides several types of process submissions based on the underlying server hardware infrastructure. Only workflow instructions and references to data, executable scripts and binary instructions are stored within the LONI *Pipeline* environment. This makes it portable, computationally efficient, distributed and independent of the individual binary processes involved in pipeline data-analysis workflows. We have expanded the LONI *Pipeline* (V.4.2) to include server-to-server (peer-to-peer) communication and a 3-tier failover infrastructure (Grid hardware, Sun Grid Engine/Distributed Resource Management Application API middleware, and the *Pipeline* server). Additionally, the LONI *Pipeline* provides three layers of background-server executions for all users/sites/systems. These new LONI *Pipeline* features facilitate resource-interoperability, decentralized computing, construction and validation of efficient and robust neuroimaging data-analysis workflows. Using brain imaging data from the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005), we demonstrate integration of disparate resources, graphical construction of complex neuroimaging analysis protocols and distributed parallel computing. The LONI *Pipeline*, its features, specifications, documentation and usage are available online (http://*Pipeline*.loni.ucla.edu).

Keywords: LONI *Pipeline*, software tools, resources, workflows, tool interoperability, data provenance, tool integration, neuroimaging

INTRODUCTION

Modern tools for image processing employ large amounts of heterogeneous data, diverse computational resources and distributed web-services (Toga and Thompson, 2007). Efficient analysis protocols combine diverse data, software tools and network infrastructure to obtain, analyze and disseminate results. Construction of such analysis protocols are significantly enhanced by a graphical workflow interface that provides high-level manipulation of the analysis sequence while hiding many of its technical details. In this manuscript, we discuss the challenges of development, maintenance and dissemination of integrated resources including data, software tools and web-services, as platform-independent, agile and scalable frameworks. We demonstrate the development and utilization of the LONI *Pipeline* environment for combining of user computational and biological expertise with disparate resources and Grid infrastructures. Version 4 of the LONI *Pipeline*, extends the previous implementation of this environment (Rex et al., 2003).

To provide an extensible framework for interoperability of diverse computational resources the LONI *Pipeline* employs a decentralized infrastructure, where tools, services and data are linked through an external resource-mediating-layer. This approach requires no modifications of existing tools to enable their interoperability with other computational counterparts. A new XML schema forms the backbone for the inter-resource-mediating-layer. Each XML resource (module) description includes important information about the resource location, the proper invocation protocol (i.e., I/O types, parameter specifications, etc.), run-time controls and data-types. Also included are auxiliary meta-data about the resource state, specifications, history, authorship and bibliography. This infrastructure[1] facilitates the integration of disparate resources and provides a natural and comprehensive data provenance (MacKenzie-Graham et al., 2008a). The LONI *Pipeline* also enables the broad dissemination of resource meta-data descriptions via web-services and the constructive utilization of multidisciplinary expertise by experts, novice users and trainees.

There are a number of efforts to develop environments for tool integration, interoperability and meta-analysis (Rex et al., 2004).

[1] http://*Pipeline*.loni.ucla.edu

There is a clear community need to establish efficient tool interoperability, which enables new types of analyses and facilitates new applications (Dinov et al., 2008). *Taverna* (Oinn et al., 2005) is an open-source, platform-independent graphical workflow environment, which enables linking tools after explicit rebuilding. It is mainly employed for bioinformatics applications via myGRID infrastructure. *Kepler* (Ludäscher et al., 2006) is another scientific workflow environment used for various applications, which also requires rebuilding each executable to link it with the core libraries. The *Triana* (Churches et al., 2006) workflow environment enables external data storage which significantly improves the efficiency and robustness of its user interface and optimizes the system requirements (e.g., low memory demands). VisTrails (Callahan et al., 2006) addresses the problem of visualization from a data management perspective where imaging data and meta-data are represented as a conjoint visualization product. Swift is another graphical workflow environment, which uses a scripting language, SwiftScript, to enable concise high-level specification of workflows based on various applications using large quantities of data. The Swift engine provides efficient execution of these workflows on sequential or parallel computers or distributed grids (Stef-Praun et al., 2007). There are a number of other graphical workflow environments that are proposed, tested and validated for specific applications, types of users, scientific areas or hardware infrastructures (Bowers et al., 2008).

Compared to other graphical workflow environments, the LONI *Pipeline* offers several advantages. It facilitates the back-end integration with distributed Grid-enabled and client-server infrastructures, and provides an efficient and robust framework for deployment of new resources to the community (new tools need not be recompiled, migrated or altered in anyway to be made functionally available to the community). The choice of a particular analysis workflow infrastructure always depends on the application domain, the type of user, types of access to resources (e.g., computational framework, human or machine resource interface, database, etc.), as well as, the desired features and functionalities (Bitter et al., 2007). There are inevitable similarities between the LONI *Pipeline* and other such environments. These include the graphical interface provided by most workflow environments, which facilitates the design of analysis protocols and improves the usability of these graphical protocols. Visual interfaces present complex analysis protocols in an intuitive manner and improve the management of technical details. Most of the graphical workflow environments provide the ability to save, load and distribute protocols through servers, SOAP/WSDL/XML or other means.

The LONI *Pipeline* addresses the specific needs of the neuroimaging and computational neuroscience community, but its general goals of providing portability, transparency, intuitiveness and abstraction from Grid mechanics make it appealing in other fields. The *Pipeline* is a dynamic resource manager, treating all resources as well-described external applications that may be invoked with standard remote execution protocols. The LONI *Pipeline* XML description protocol allows any command-line driven process, web-service or data-server to be accessed within the environment *by reference*, with dependencies validated (checked) dynamically on-demand. There is no need to reprogram, revise or recompile external resources to make them usable within the LONI *Pipeline*. One side effect of this design choice is that to all external *Pipeline* server installations require complete installations of all software tools, services and data as currently available on the LONI Grid. This however, is only required, if a remote Pipleine server must mirror all tools and resources as available on the LONI Grid. In most situations, each site has specific suits of tools that they utilize to meet their computational needs. This design reduces the integration/utilization costs of including new resources within the LONI *Pipeline* environment. This approach provides the benefit of quick and easy management of large and disparately located resources and data. In addition, this choice significantly minimizes the user/client machine hardware and software requirements (e.g., memory, storage, CPU). Finally, a key difference between the LONI *Pipeline* and some other environments is its management of distributed resources via its client-to-server infrastructure and its ability to export automated makefiles/scripts. These allow the LONI *Pipeline* to provide processing power independently of the available computational environment (e.g., SOLARIS, LINUX, Grid, mainframe, desktop, etc.). The LONI *Pipeline* servers communicate and interact with clients and facilitate secure transfer of processes, instructions, data and results via the Internet.

Version 4 of the LONI *Pipeline* introduces several important improvements and extensions of the previous version of the LONI *Pipeline* (Rex et al., 2003). These include a 3-tier failover mechanism for Grid hardware, Sun Grid Engine (SGE)/Distributed Resource Management Application API (DRMAA) middleware, and the *Pipeline* server, as well as client-server communication, makefile/script export and data provenance model. LONI *Pipeline* v.4 also includes a new more functional and robust graphical user interface and a significantly increased library of tools. V.4 also simplifies the inclusion of external data display modules and facilitates remote database connectivity (e.g., LONI Imaging Data Archive[2], BIRN Storage Resource Broker[3], XNAT[4], etc.)

MATERIALS AND METHODS

The main goal of developing the LONI *Pipeline* was to provide a robust and extensible infrastructure for computational neuroscience enabling efficient data utilization, construction of reliable analysis workflows, and provide the means for wide dissemination and validation of research protocols and scientific findings. The LONI *Pipeline* developments are subdivided into several complementary goals:

- *Efficient Distributed Computing*: Facilitate the integration of disparate, heterogeneous and multi-platform implementations of software tools, database protocols and remote web-services. The LONI *Pipeline* client-server communication protocol allows blending of resources that are built on remote server architectures to be accessed by the pipeline clients. This greatly lowers the usability requirements for the general user. In addition, we need a flexible export of available pipeline workflows into makefiles and bash scripts that can be submitted virtually to any computational architecture.

[2]http://ida.loni.ucla.edu

[3]http://www.nbirn.net/tools/srb

[4]http://www.xnat.org

- *Design a robust 3-tier failover mechanism for the LONI Grid*: This included the three layers of the Grid submission protocol – Sun Grid Engine (SGE), Distributed Resource Management Application API[5], and the *Pipeline* job handling server. These three layers of background-server executions enable various types of users and systems to utilize the *Pipeline* environment in any of these three execution modes: single machines, or main-frames with only one queue job submission protocol; Globus Grid infrastructure, and SGE Grid Infrastructure.
- *Provenance*: LONI *Pipeline* includes a provenance manager, which enables tracking data, workflow and execution history of all processes. This functionality improves the communication, reproducibility and validation of newly proposed experimental designs, scientific analysis protocols and research findings. This includes the ability to record, track, extract, replicate and evaluate the data and analysis provenance to enable rigorous validation and comparison of classical and novel design paradigms.
- *Tool Discovery*: Enable expert researchers to quickly design, test and validate novel experimental designs and data analysis protocols. This is achieved via a dynamic, responsive and intelligent graphical user interface for tool exploration and construction of draft pipeline workflows.
- *Friendly Graphical User Interface*: Create a robust environment for tool interoperability, Grid integration and low-cost interactive user interface. For maximum portability, scalability and efficiency, this environment is built in Java and utilizes XML for storing and communication of meta-data, and descriptors for tools and services.

The LONI *Pipeline* execution environment controls the local and remote server connections, module communication, process management, data transfers and Grid mediation. The XML descriptions of individual modules, or networks of modules, may be constructed, edited and revised directly within the LONI *Pipeline* graphical user interface, as well as saved or loaded from disk or the LONI *Pipeline* server. These workflows completely describe new methodological developments and allow validation, reproducibility, provenance and tracking of data and results. The core six types of *Pipeline* specifications are summarized below.

TYPES OF TOOLS AND SERVICES THAT CAN BE INTEGRATED WITHIN THE LONI *Pipeline*

The development and utilization of the LONI *Pipeline* environment is focused on neuroimaging data and analysis protocols. However, by design, the LONI *Pipeline* software architecture is domain agnostic and has been adopted in other research and clinical fields, e.g., bioinformatics (Dinov et al., 2008). There are two major types of resources that may be integrated within the LONI *Pipeline*. The first one is *data*, in terms of databases, data services and file systems. The second type of pipelineable resources includes stand-alone *tools*, comprising local or remote binary executables and services with well-defined command line syntax. This flexibility permits efficient resource integration, tool interoperability and wide dissemination.

GENERAL LONI *Pipeline* SPECIFICATIONS INCLUDING GRID INTEGRATION

The LONI *Pipeline* routinely executes thousands of simultaneous jobs on our symmetric multiprocessing systems (SMP) and on DRMAA[6] clusters. On SMP systems, the LONI *Pipeline* can detect the number of available processing units and scale the number of simultaneous jobs accordingly to maximize system utilization and prevent system crashes. For computer clusters, a grid engine implementing DRMAA, with Java bindings, may be used to submit jobs for processing, and a shared file system is used to store inputs and outputs from individual jobs. Later, we will extend the scope of the LONI *Pipeline* server to interact and submit jobs to other Grid infrastructures, e.g., Condor, Globus, etc.

The LONI *Pipeline* environment has been integrated with UNIX authentication using Pluggable Authentication Module (PAM), to enable a username and password challenge-response authentication method using existing credentials. A dependency on the underlying security and encryption system of the LONI *Pipeline* server's host machine offers maximum versatility in light of the diverse policies governing system authentication and access control.

Using Java binding to DRMAA interface, we have integrated the LONI *Pipeline* environment with the SUN Grid Engine (SGE), a free, well-engineered distributed resource manager (DRM) that simplifies the processing and management of submitted jobs on the grid. It is important to note, however, that other DRMs such as Condor, LSF and PBS/Torque could be made compatible with the LONI *Pipeline* environment using the same interface. DRMAA's Java foundation allows jobs to be submitted from the LONI *Pipeline* to the compute grid without the use of external scripts and provides significant job control functionality internally. We accomplished several key goals with the LONI *Pipeline*-DRMAA-SGE integration:

- the parallel nature of the LONI *Pipeline* environment is enhanced by allowing for both horizontal (across compute nodes) and vertical (across CPUs on the same node) processing parallelization;
- the LONI *Pipeline's* client-server functionality can directly control a large array of computational resources with DRMAA over the network, significantly increasing its versatility and efficacy;
- facilitate the use of a heterogeneous set of neuroimaging software tools in pipelines involving large number of datasets and multiple processing tools;
- the overall usability of grid resources is improved by the intuitive graphical interface offered by the LONI *Pipeline* environment, and
- the ability to display interim results from user-specified modules, which can be used for visual inspection of the outputs of various tools (interactive outcome checking).

LONI *Pipeline* DATA PROVENANCE

In neuroimaging studies, data provenance, or the history of how the data were acquired and subsequently processed, is often discussed but seldom implemented (MacKenzie-Graham et al., 2008b).

[5]http://www.DRMAA.org

[6]http://www.drmaa.org

Recently, several groups have proposed provenance challenges in order to evaluate the status of various provenance models (Miles et al., 2006). For instance, collecting provenance information from a simple neuroimaging workflow (Zhao et al., 2007) and documenting each system's ability to respond to a set of predefined queries. Some of the existing provenance systems are designed as mechanisms for capturing provenance in neuroimaging (MacKenzie-Graham et al., 2008a; Zhao et al., 2007). It is difficult to provide systematic, accurate and comprehensive capture of provenance information with minimal user intervention. The processes of data provenance and curation are significantly automated via the LONI *Pipeline*. Each dataset has a provenance file (*.prov) that is automatically updated by the LONI *Pipeline*, based on the protocols used in the data analysis. This data processing history reflects sequentially the steps that a dataset goes through and provides a detailed record of the types of tools, versions, platforms, parameters, control and compilation flags. The data provenance can be imported and exported by the LONI *Pipeline*, which enables utilization internally by other *Pipeline* workflows or by external resources (e.g., databases, workflow environments).

Provenance can be used for determining data quality, for result interpretation, and for protocol interoperability (Simmhan et al., 2005; Zhao et al., 2007). It is imperative that the provenance of neuroimaging data be easily captured and readily accessible (MacKenzie-Graham et al., 2008b). For instance, increasingly complex analysis workflows are being developed to extract information from large cross-sectional or longitudinal studies in multiple sclerosis (Liu et al., 2005), Alzheimer's disease (Fleisher et al., 2005), autism (Langen et al., 2007), depression (Drevets, 2001), schizophrenia (Narr et al., 2007), and studies of normal populations (Gogtay et al., 2006). The implementation of the complex workflows associated with these studies requires provenance-based quality control to ensure the accuracy, reproducibility, and reusability of the data and analysis protocols.

We designed the provenance framework to take advantage of context information that can be retrieved and stored while data is being processed within the LONI *Pipeline* environment (MacKenzie-Graham et al., 2008b). Additionally, the LONI Provenance Editor is a self-contained, platform-independent application that automatically extracts provenance information from image headers (such as a DICOM images) and generates an XML data provenance file with that information. The Provenance Editor[7] allows the user to edit the meta-data prior to saving the provenance file, correcting inaccuracies or adding additional information. This provenance information is stored in.*prov* files, XML formatted files that contain the meta-data and processing provenance and follows the XSD definition[8]. Then the data provenance is expanded by the LONI *Pipeline* to include the analysis protocol, the specific binaries used for analysis, and the environment that they were run in. The LONI *Pipeline* dramatically improves compliance by minimizing the burden on the provenance curator. This frees the user to focus on performing neuroimaging research rather than on managing provenance information.

LONI *Pipeline* INTELLIGENCE

Construction of elaborate, functional and valid workflows within the LONI *Pipeline* environment requires deep understanding of the research goals, tool specifications and neuroscientific expertise. To enhance the usability of this environment, we developed an intelligent LONI *Pipeline* component. It has two complementary features – constructive and validating. The pipeline *constructive intelligent feature* uses the spectra of available module descriptors and pipeline workflows to automatically generate valid versions of new graphical protocols according to a set of user-specified keywords. This intelligence feature uses a grammar on the set of XML module and pipeline descriptions to determine the most appropriate analysis protocol, and its corresponding module inputs and outputs, according to the keywords provided by the user. Then, it exports a.*pipe* file, which contains a draft of the desired analysis protocol, **Figure 1**.

The pipeline *validating intelligence feature* offers interactive support for running or modifying existent pipeline workflows. This feature contextually monitors the consistency of the data types, parameter matches, validity of the analysis protocol, and ensures optimal job-submission (e.g., order of module execution). The LONI *Pipeline* intelligence component reduces the need to review in details of, and double check modifications of new or existing workflows. Still, users control the processes of saving workflows and module descriptions, data input and output, and the scientific design of their experiments. This functionality significantly improves usability and facilitates scientific exploration.

LONI *Pipeline* GRAPHICAL AND SCRIPTING INTERFACES

Pipeline workflows (.*pipe* files) may be constructed in many different ways (e.g., using text editors) and these protocols may be executed in a batch mode without involving the LONI *Pipeline* graphical user interface (GUI). However, the LONI *Pipeline* GUI significantly aids most users in designing and running analysis workflows. A library of available tools for usage is presented on the left hand side of the LONI *Pipeline* client window. Users may search for, drag and drop these tools onto the main canvas to create or revise a workflow. Connections between the nodes are used to represent the piping of output from one program to another. This is accomplished without requiring the user to specify file paths, server locations or command line syntax. *Pipeline* workflows may be constructed and executed with data dynamically flowing (by reference) within the workflow. This enables trivial inclusion of pipeline protocols in external scripts and integration into other applications. Currently, the LONI *Pipeline* allows exporting of any workflow from XML (*.pipe) format to a *makefile* or a *bash* script for direct or queuing execution.

FUNCTIONALITY AND USABILITY

In the past 3 years, we have gone through several cycles of design, implementation, analysis and re-design stages of the new LONI *Pipeline*. During this process a number of *usability issues* were addressed. These included the editing and usage modes of the graphical user interface, state specific menus, pop-up and information dialogs, the handling of local and global variables within the pipeline, the integration of data sources and executable module nodes, data type checking and workflow validation, client connect and disconnects, job management and client-server communications. All of these were critical in improving the usability of the LONI *Pipeline* and are necessary before the execution of any data analysis workflow.

[7]http://www.loni.ucla.edu/Software/Software_Detail.jsp?software_id = 57

[8]http://www.loni.ucla.edu/~pipelnv4/pipeline_xsd.xsd

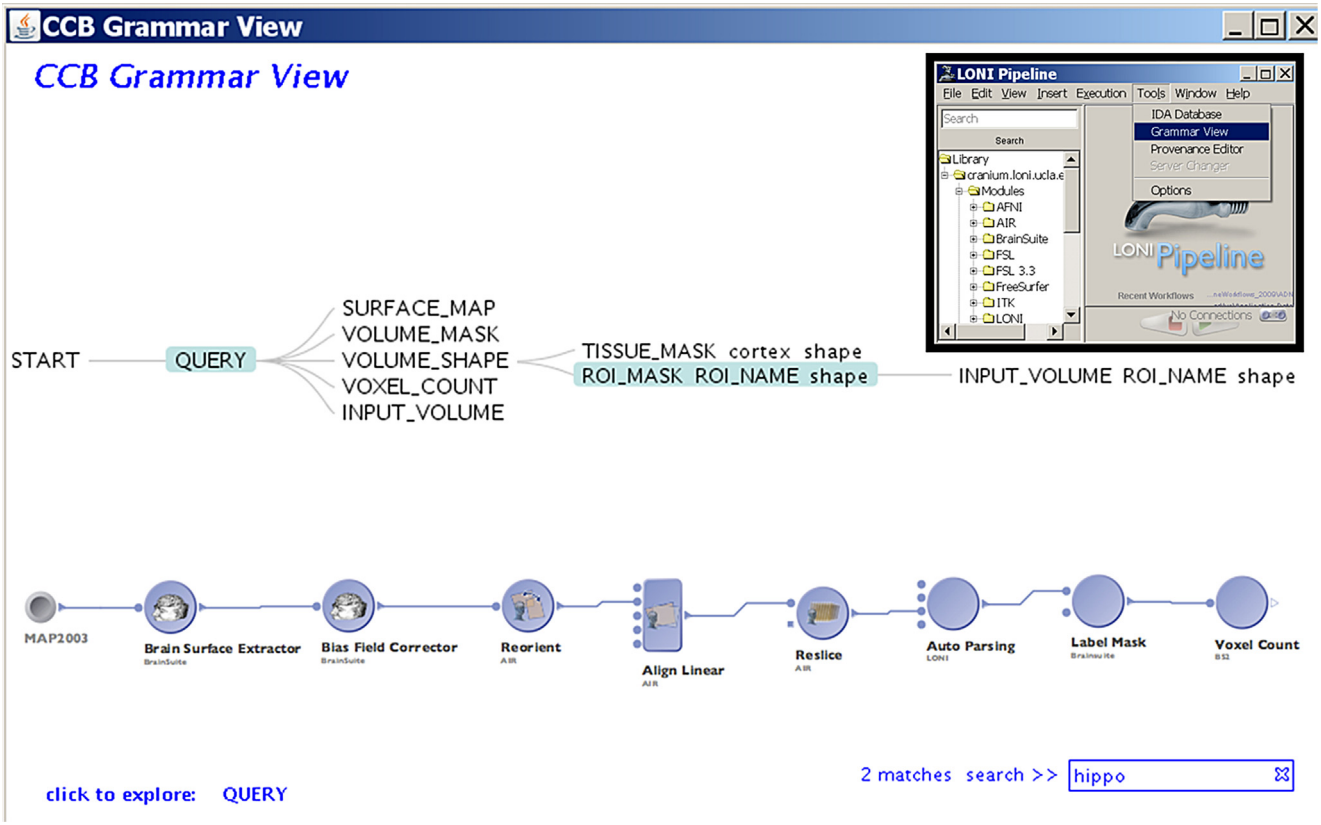


FIGURE 1 | LONI *Pipeline* intelligence component uses key-words to automatically generate a hierarchical interface and the complete analysis workflow, which represents the proposed study protocol, using semantic natural language processing of language grammar. The image insert shows the graphical user interface invoking the pipeline intelligence grammar view.

The core LONI *Pipeline* functionality is based on our prior experience (Rex et al., 2003), user feedback and information technology advancements over the past several years. The current LONI *Pipeline* functionality includes – tool discovery engine, plug-in interface for meta algorithm design, grid interface, secure user authentication, data transfers and client-server communications, graphical and batch-mode execution, encapsulation of tools, resources and workflows, and data provenance.

RESULTS

LONI *Pipeline* can be used to construct a wide variety of processing and analysis workflows. Here we demonstrate the utilization of the LONI *Pipeline* to conduct and validate new (semi)automated, robust and user-friendly protocols for (1) regional parcellation and volume extraction, (2) population-based atlas construction, and (3) the analysis of multiple population cohorts. In the following sections, we discuss the graphical *Pipeline* workflows for each of these applications:

BRAIN PARCELLATION

Regional parcellation of distinct brain regions is often needed to perform region-of-interest-based analyses between healthy as compared to diseased subjects. Manual region drawing can be labor intensive, prone to errors, and have poor reproducibility. **Figure 2** illustrates a pipeline workflow constructed to automatically extract 3D masks of 56 regions of interest using Brain Parser (Tu et al., 2008)[9]. These regions can be then be used to examine regionally specific shape characteristics among other variables of interest to the neuroimaging community. The ability to automatically obtain robust 3D masks of various brain regions is a critical step in many brain mapping studies.

BRAIN ATLAS CONSTRUCTION

Brain atlasing is a major research effort in the field and the development of efficient workflows to take large numbers of T1-weighted anatomical images, spatially warp them into a common space, and then to pool them to result in a representative atlas is often a complex process. Development of efficient workflows and utilizing a large-scale computational Grid, based at LONI, permits streamlined and rapid atlas creation in normal subjects as well as in disease, **Figure 3**. Using Automated Image Registration (Woods et al., 1998), we constructed a workflow to systematically create a whole brain atlas for use in describing the average brain anatomical structure in patients drawn from the ADNI series of Alzheimer's subject MRI data contained in the LONI Image Data Archive (IDA) (Mueller et al., 2005). Such atlases characterize "mean" population features such as shape, regional area, sulcal anatomy, etc.

[9]http://www.loni.ucla.edu/Software/BrainParser

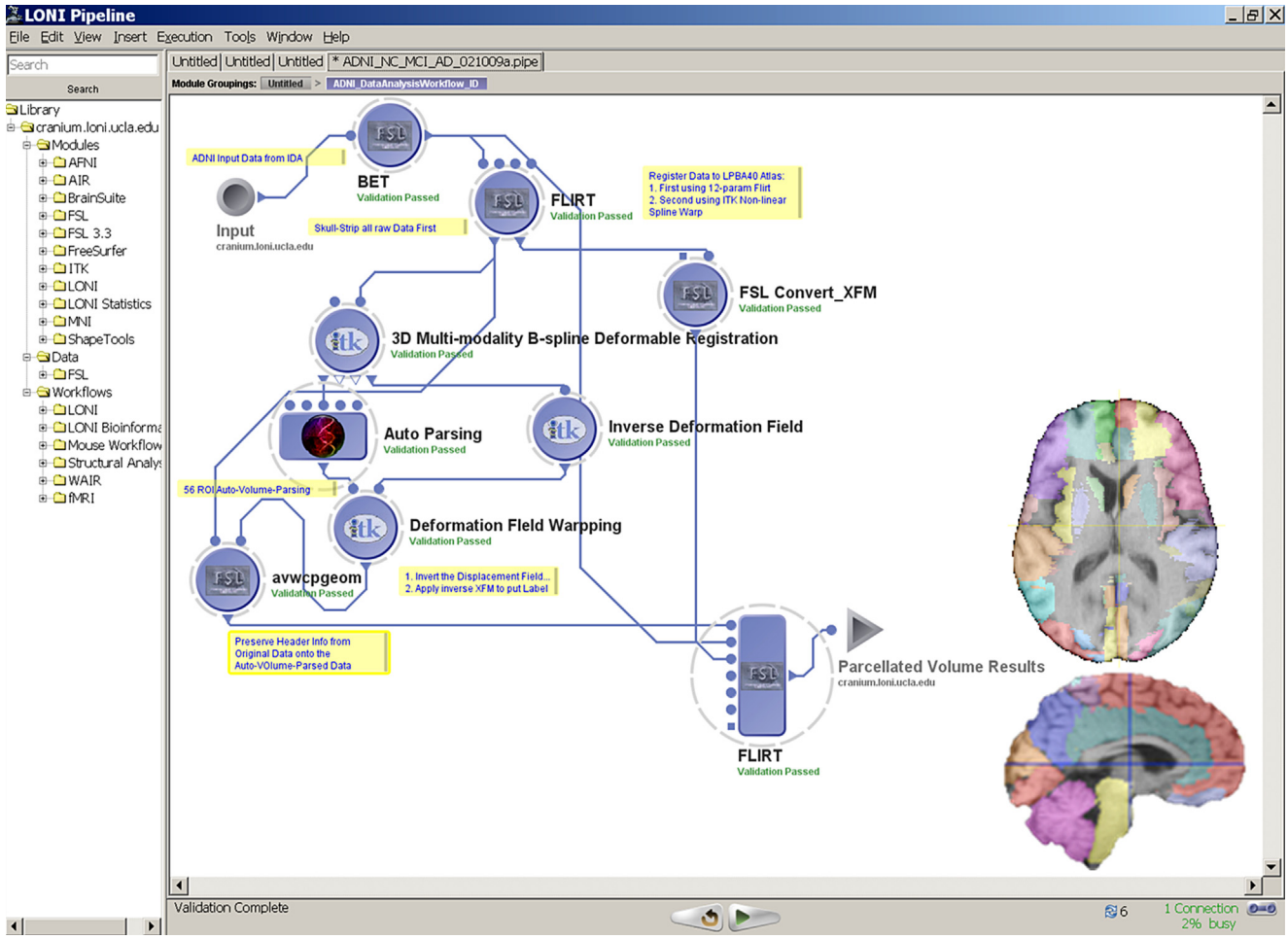


FIGURE 2 | Using a robust executable entitled Brain Parser (Tu et al., 2008), LONI *Pipeline* can be used to extract 56 predefined ROI masks from any input brain image volume (inserts).

STRUCTURAL ANALYSIS OF ALZHEIMER'S DISEASE (AD) NEUROIMAGING STUDY

We used brain imaging data from the Alzheimer's Disease Neuroimaging Initiative, ADNI (Mueller et al., 2005), to demonstrate the processes of construction, validation and execution of integrated workflow analysis protocols. This AD pipeline workflow represents a complex neuroimaging analysis protocols based on disparate tools, data and distributed parallel-computing infrastructure. **Figure 4** demonstrates this Alzheimer's disease *Pipeline* workflow. The left-panel in the *Pipeline* environment contains some predefined module definitions and complete workflows. The user may drag-and-drop these in the main workflow canvas to design new analysis protocols. The central workflow canvas shows that main six steps of the AD data analysis. These include data conversion, volumetric data pre-processing, automated extraction of regions of interest, shape processing, global shape analysis and automated cortical surface extraction. Each of these steps is itself a nested collection of groups of modules, a nested pipeline workflow, which contains a series of processing steps. The insert-figure illustrates the 3-level deep nested processing part of the Global Shape Analysis node (see the top-level tabs of the insert).

This pipeline workflow demonstrates the entire data processing and analysis protocol, from retrieval of the data from the LONI Imaging Data Archive[10], through the data manipulation, shape processing, generation of derived data (e.g., global shape measures like curvature, fractal dimension, surface area, etc.), to the final statistical analysis. In this case, the study design included three age-matched populations – asymptomatic subjects (NC), minor cognitive impairment (MCI), and Alzheimer's disease (AD) patients. There were five males and five females for each group and each subject was scanned several times longitudinally. A total of 104 brain volumes were automatically processed in about 26 h. The time of workflow completion depends on the study and workflow designs, number of subjects, and general hardware infrastructure specifications (e.g., system characteristics and user demand). The results of this completely automated

[10]http://IDA.loni.ucla.edu

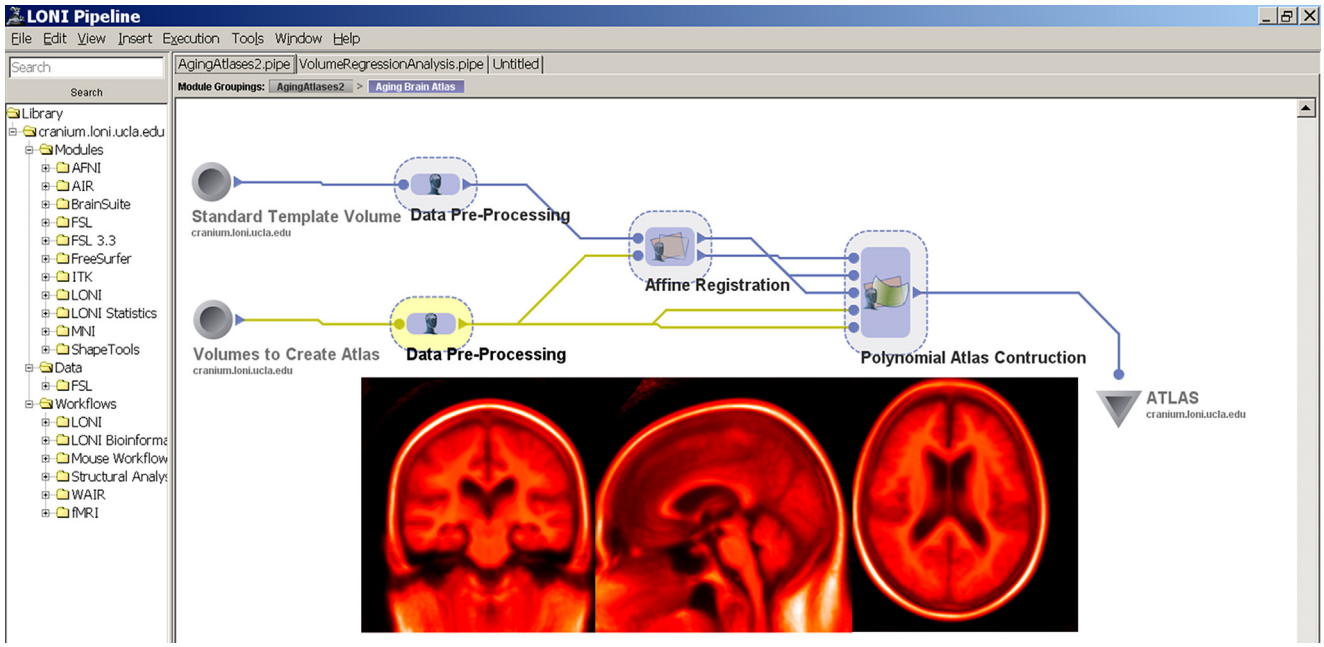


FIGURE 3 | LONI *Pipeline* workflow for constructing a population-based whole brain anatomical atlas in Alzheimer's Disease patients (insets).

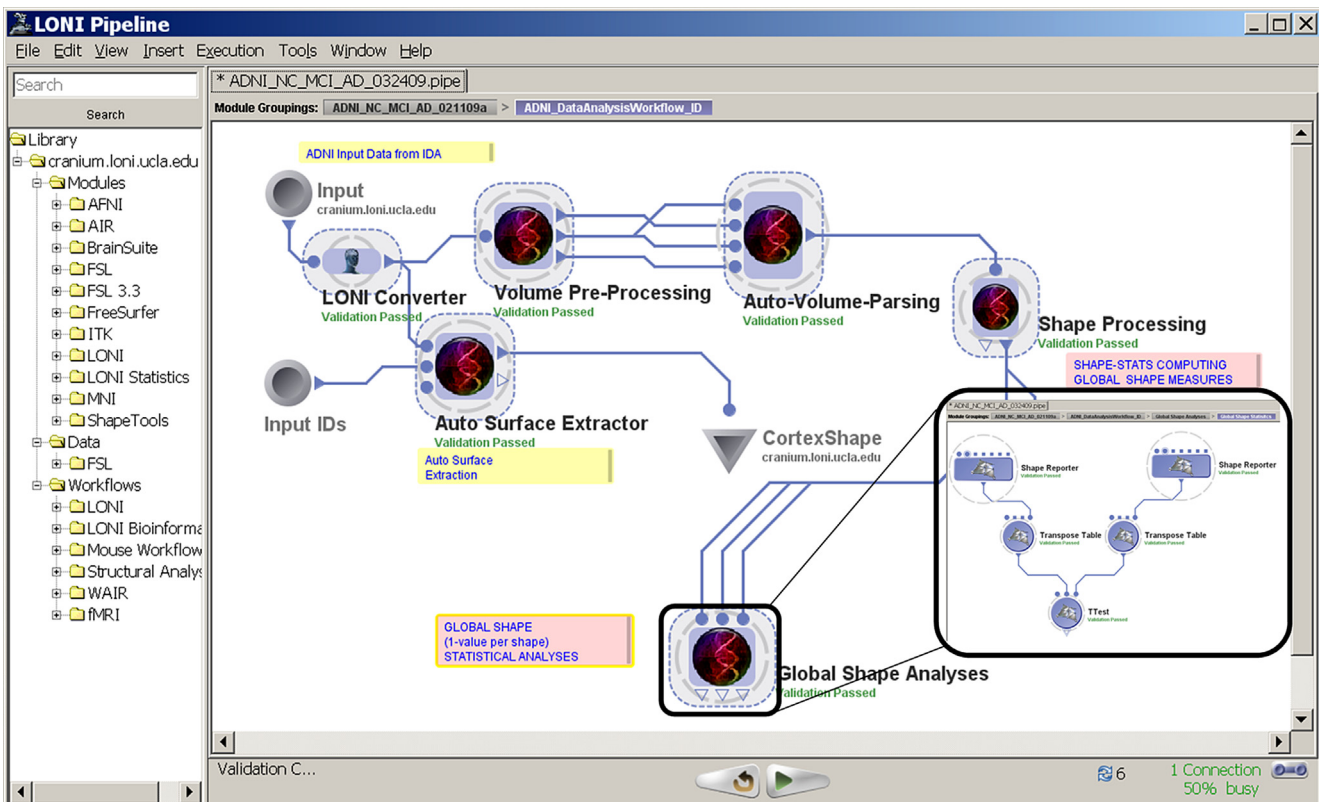


FIGURE 4 | An Alzheimer's Disease (AD) Pipeline workflow. The left-panel contains some predefined Pipeline module definitions including some complete workflows. The central-panel shows the main six steps of the data analysis – data conversion, pre-processing, automated extraction of regions of interest, shape processing, global shape analysis and automated cortical surface extraction. Each of these steps is itself a nested collection of modules, a pipeline, which contains a series of processing steps. The insert-figure illustrates the 3-level deep nested processing part of the Global Shape Analysis node (see the top-level tabs of the insert).

pipeline workflow included cortical surface representations (shapes) for each subject, parcellations of the raw MRI brain scans into 56 regions of interest (labels), surface models for each of the 56 regions for each subject and time-point, and global statistical mapping identifying the NC, MCI and AD group differences for each of the 56 regions.

Figure 5 depicts the shape-curvature measure for five regions of interest (ROI's) at two time-points – baseline (blue) and 12-month follow-up (green) for the cohort of normal subjects (NC). **Figure 6** compares the shape measures for one region (right Superior Frontal Gyrus) across all three cohorts, at baseline (time = 0). Notice the consistent decrease of shape and volume measures, for both time

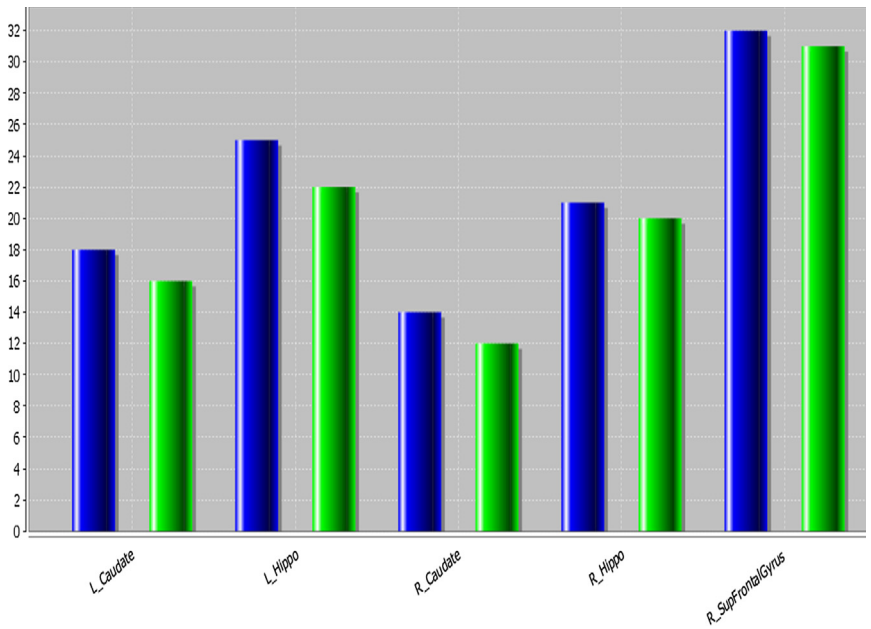


FIGURE 5 | NC *shape-curvature* measure for 5 ROI's at two times (*T1*, baseline, blue, and *T2*, 12-month follow-up, green). *L_Caudate* and *R_Caudate*, left and right caudate, *L_Hippo* and *R_Hippo*, left and right hippocampus, R_*SupFrontalGyrus*, right superior frontal gyrus.

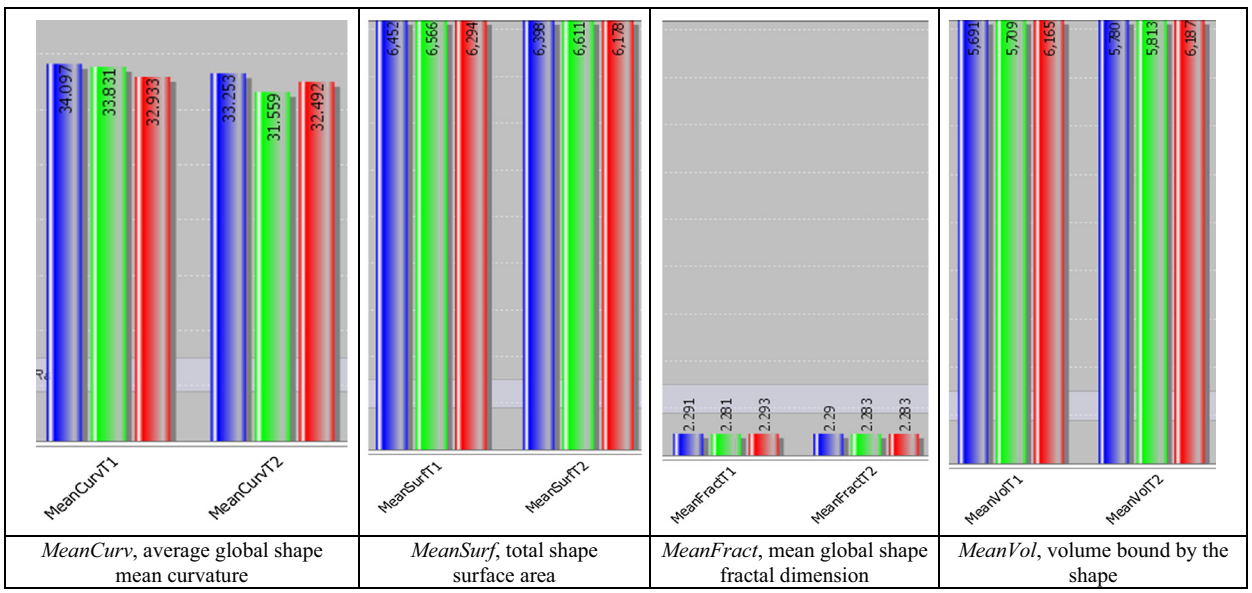


FIGURE 6 | Comparison of the four volume and shape measures for both times (baseline and 12-month follow up) across the three cohorts for one region of interest – the Right Superior Frontal Gyrus. *T1* and *T1* labels represent baseline and follow-up time scans. The statistics signature vector includes *MeanCurv*, average global shape mean curvature; *MeanSurf*, total shape surface area; *MeanFract*, mean global shape fractal dimension; and *MeanVol*, volume of the inside region of the shape. The three different cohorts, NC, MCI and AD, are colored in blue, green and red, respectively.

points, going from NC (asymptomatic) to MCI and AD (most effected individuals).

DISCUSSION

Interactive workflow environments for automated data analysis are critical in many research studies involving complex computations and large datasets (Kawas et al., 2006; Myers et al., 2006; Oinn et al., 2005; Taylor et al., 2006). There are three distinct necessities that underlie the importance of such graphical frameworks for management of novel analysis strategies – high data volume and complexity, sophisticated study protocols and demands for distributed computational resources. These three fundamental needs are evident in most modern neuroimaging, bioinformatics and multidisciplinary studies.

The LONI *Pipeline* environment aims to provide distributed access to varieties of computational resources via its graphical interface. The ability of investigators to share, integrate, collaborate and expand resources will increase the statistical power in studies involving heterogeneous datasets and complex analysis protocols. New challenges that emerge from our increased abilities to utilize computational resources and hardware infrastructure include the need to assure reliability and reproducibility of identically analyzed data, and the desire to continually lower the costs of employing and sharing data, tools and services. The LONI *Pipeline* environment attempts to provide the means to address these difficulties by providing secure integrated access to resource visualization, databases and intelligent agents.

The LONI *Pipeline* already has been used in a number of neuroimaging applications including health (Sowell et al., 2007), disease (Thompson et al., 2003), animal models (MacKenzie-Graham et al., 2006), volumetric (Luders et al., 2006), functional (Rasser et al., 2005), shape (Narr et al., 2007) and tensor-based (Chiang et al., 2007) studies. The LONI *Pipeline* infrastructure improved consistency, reduced development and execution times, and enabled new functionality and usability of the analysis protocols designed by expert investigators in all of these studies. Perhaps the most powerful feature provided by the LONI *Pipeline* environment is the ability to quickly communicate new protocols, data, tools and service resources, findings and challenges to the wider community.

The main LONI *Pipeline* page[11] provides links to the forum, support, video tutorials and usage. There are examples demonstrating how to describe individual modules and construct integrated workflows. Version information, download instructions and server/forum account information is also available on this page. There are example pipeline workflows and the XSD schema definition[12] for the.*pipe* format used for module and workflow XML description. Users may either install *Pipeline* servers on their own hardware systems, or they may use some of the available *Pipeline* servers. The primary LONI *Pipeline* server is cranium.loni.ucla.edu. It utilizes a CentOS-based compute cluster comprised of approximately 800 Core 2, 2.4GHz, 8GB RAM, AMD Opteron processors. Each dual-processor compute node has eight gigabytes of memory to accommodate memory-intensive neuroimaging applications. We selected SUN Grid Engine v6, bound by DRMAA, as the LONI *Pipeline* distributed resource manager. A highly-optimized non-blocking Cisco Gigabit network provides the connectivity infrastructure with sixteen terabytes of fault-tolerant, clustered storage from Isilon Systems acting as a cache file system for the LONI *Pipeline* environment. Users may obtain accounts on this Grid[13].

In general, some practical difficulties in validating new LONI *Pipeline* workflows may be caused by unavailability of the initial raw data, differences of hardware infrastructures or variations in compiler settings and platform configurations. Such situations require analysis workflow validation by teams of experts capable of validating the input, output and state of each module within the pipeline workflow. Further LONI *Pipeline* validation would require comparison between synergistic workflows that are implemented using different executable modules or module specifications. For example, one may be interested in comparing similar analysis workflows by choosing different sets of imaging filters, reconfiguring computation parameters or manipulating the resulting outcomes, e.g., file format, (Bitter et al., 2007). Such studies contrasting the benefits and limitations of each resource or processing workflow aid both application developers and general users in the decision of how to design and utilize module and pipeline definitions to improve resource usability.

A significant challenge in computational neuroimaging studies is the problem of reproducing findings and validating analyses described by different investigators. Frequently, methodological details described in research publications may be insufficient to accurately reconstruct the analysis protocol used to study the data. Such methodological ambiguity or incompleteness may lead to misunderstanding, misinterpretation or reduction of usability of newly proposed techniques. The LONI *Pipeline* mediates these difficulties by providing clear, functional and complete record of the methodological and technological protocols for the analysis.

Even though the LONI *Pipeline* was designed and tested to solve neuroimaging problems, its generic architecture will permit applications in other fields, where computationally intense tasks are performed and there is a need of resource interoperability. Its light-weight and platform-independent design and its low memory requirements make the LONI *Pipeline* potentially useful in many research fields relying on the integration of large and heterogeneous processing protocols. For example, the LONI *Pipeline* was recently used in conjunction with a number of bioinformatics data processing and analysis protocols (Dinov et al., 2008). We are also working on several new features of the LONI *Pipeline* including web-service-based client interface, direct integration with external resource archives (e.g., http://www.ncbcs.org/biositemaps, http://*NeuroGateway*.org, etc.) and interface enhancements using intelligent plug-in components.

ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health through the NIH Roadmap for Medical Research, Center for Computational Biology Grant U54 RR021813, NIH/NCRR 5 P41 RR013642 and NIH/NIMH 5 R01 MH71940. The authors are also indebted to members of the Laboratory of Neuro Imaging, and various collaborators and users for their patience with testing the LONI *Pipeline*. Arash Payan, Jia-Wei Tam, Celia Cheung, Cornelius Hojatkashani and Jagadeeswaran Rajendiran contributed to the implementation of *Pipeline* V.4.

[11]http://*Pipeline*.loni.ucla.edu

[12]http://www.loni.ucla.edu/~pipelnv4/pipeline_xsd.xsd

[13]http://www.loni.ucla.edu/Collaboration/Pipeline/Pipeline_Download.jsp

REFERENCES

Bitter, I., Van Uitert, R., Wolf, I., Ibanez, L. A., Kuhnigk, J. M. A., and Kuhnigk, J. M. (2007). Comparison of four freely available frameworks for image processing and visualization that use ITK. *Trans. Vis. Comput. Graph.* 13, 483–493.

Bowers, S., McPhillips, T., and Ludäscher, B. (2008). Provenance in collection-oriented scientific workflows. *Concur. Comput. Prac. Exp.* 20, 519–529.

Callahan, S., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., and Vo, H. T. (2006). VisTrails: Visualization Meets Data Management, in Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. Chicago, IL, ACM.

Chiang, M.-C., Dutton, R. A., Hayashi, K. M., Lopez, O. L., Aizenstein, H. J., Toga, A. W., Becker, J. T., and Thompson, P. M. (2007). 3D pattern of brain atrophy in HIV/AIDS visualized using tensor-based morphometry. *NeuroImage* 34, 44–60.

Churches, D., Gombas, G., Harrison, A., Maassen, J., Robinson, C., Shields, M., Taylor, I., and Wang, I. (2006). Programming scientific and distributed workflow with Triana services. *Concur. Comput. Prac. Exp.* 18, 1021–1037.

Dinov, I. D., Rubin, D., Lorensen, W., Dugan, J., Ma, J., Murphy, S., Kirschner, B., Bug, W., Sherman, M., Floratos, A., Kennedy, D., Jagadish, H. V., Schmidt, J., Athey, B., Califano, A., Musen, M., Altman, R., Kikinis, R., Kohane, I., Delp, S., Parker, D. S., and Toga, A. W. (2008). iTools: a framework for classification, categorization and integration of computational biology resources. *PLoS ONE* 3, e2265.

Drevets, W. C. (2001). Neuroimaging and neuropathological studies of depression: implications for the cognitive-emotional features of mood disorders. *Curr. Opin. Neurobiol.* 11, 240–249.

Fleisher, A., Grundman, M., Jack, C. R. Jr., Petersen, R. C., Taylor, C., Kim, H. T., Schiller, D. H. B., Bagwell, V., Sencakova, D., Weiner, M. F., DeCarli, C., DeKosky, S. T., van Dyck, C. H., and Thal, L. J. (2005). Sex, apolipoprotein E {varepsilon}4 status, and hippocampal volume in mild cognitive impairment. *Arch. Neurol.* 62, 953–957.

Gogtay, N., Nugent, T. F., Herman, D. H., Ordonez, A., Greenstein, D., Hayashi, K. M., Clasen, L., Toga, A. W., Giedd, J. M., Rapoport, J. L., and Thompson, P. M. (2006). Dynamic mapping of normal human hippocampal development. *Hippocampus* 16, 664–672.

Kawas, E., Senger, M., and Wilkinson, M. (2006). BioMoby extensions to the Taverna workflow management and enactment software. *BMC Bioinformatics* 7, 523.

Langen, M., Durston, S., Staal, W., Palmen, S., and van Engeland, H. (2007). Caudate nucleus is enlarged in high-functioning medication-naive subjects with autism. *Biol. Psychiatry* 62, 262–266.

Liu, L., Meier, D., Polgar-Turcsanyi, M., Karkocha, P., Bakshi, R., and Guttmann, C. R. G. (2005). Multiple sclerosis medical image analysis and information management. *Neuroimaging* 15(4 Suppl.), 103S–117S.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., and Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concur. Comput. Prac. Exp.* 18, 1039–1065.

Luders, E., Narr, K. L., Thompson, P. M., Rex, D. E., Woods, R. P., DeLuca, H., Jancke, L., and Toga, A. W. (2006). Gender effects on cortical thickness and the influence of scaling. *Hum. Brain Mapp.* 27, 314–324.

MacKenzie-Grahama, A., Tinsleyb, M. R., Shaha, K. P., Aguilara, C., Stricklanda, L. V., Bolinea, J., Martinc, M., Moralesd, L., Shattucka, D. W., Jacobse, R. E., Voskuhld, R. R., and Toga, A. W. (2006). Cerebellar cortical atrophy in experimental autoimmune encephalomyelitis. *Neuroimage* 32, 1016–1023.

MacKenzie-Graham, A., Payan, A., Dinov, I. D., Van Horn, J. D., and Toga, A. W. (2008). Neuroimaging data provenance using the LONI pipeline workflow environment. *LNCS* 5272, 208–220.

MacKenzie-Graham, A., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *NeuroImage* 42, 178–195.

Miles, S., Groth, P., Branco, M., and Moreau, L. (2006). The requirements of using provenance in e-science experiments. *J. Grid Comput.* 5, 1–25.

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., and Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement.* 1, 55–66.

Myers, J., Allison, T. C., Bittner, S., Didier, B., Frenklach, M., Green, W. H., Ho, Y. L., Hewson, J., Koegler, W., Lansing, C., Leahy, D., Lee, M., McCoy, R., Minkoff, M., Nijsure, S., Von Laszewski, G., Montoya, D., Oluwole, L., Pancerella, C., Pinzon, R., Pitz, W., Rahn, L. A., Ruscic, B., Schuchardt, K., Stephan, E., Wagner, A., Windus, T., and Yang, C. (2006). A collaborative informatics infrastructure for multi-scale science. *Cluster Comput.* 8, 243–253.

Narr, K. L., Bilder, R. M., Luders, E., Thompson, P. M., Woods, R. P., Robinsond, D., Szeszkod, P.R., Dimtcheva, T., Gurbani, M., and Toga, A. W. (2007). Asymmetries of cortical shape: effects of handedness, sex and schizophrenia. *NeuroImage* 34, 939–948.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2005). Taverna: lessons in creating a workflow environment for the life sciences. *Concur. Comput. Prac. Exp.* 18, 1067–1100.

Rasser, P., Johnston, P., Lagopoulos, J., Ward, P. B., Schall, U., Thienel, R., Bender, S., Toga, A. W., and Thompson, P. M. (2005). Functional MRI BOLD response to Tower of London performance of first-episode schizophrenia patients using cortical pattern matching. *NeuroImage* 26, 941–951.

Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The LONI pipeline processing environment. *Neuroimage* 19, 1033–1048.

Rex, D. E., Shattuck, D. W., Woods, R. P., Narr, K. L., Luders, E., Rehm, K., Stolzner, S. E., Rottenberg, D. A., and Toga, A. W. (2004). A meta-algorithm for brain extraction in MRI. *NeuroImage* 23, 625–637.

Simmhan, Y. L., Plale, B. and Gannon, D. (2005). A survey of data provenance in escience. *ACM SIGMOD Rec.* 34, 31–36.

Sowell, E., Peterson, B. S., Kan, E., Woods, R. P., Yoshii, J., Bansal, R., Xu, D., Zhu, H., Thompson, P. M., and Toga, A. W. (2007). Sex differences in cortical thickness mapped in 176 healthy individuals between 7 and 87 years of age. *Cereb. Cortex* 17, 1550–1560.

Stef-Praun, T., Clifford, B., Foster, I., Hasson, U., Hategan, M. S., Small, L., Wilde, M., and Zhao, Y. (2007). Accelerating medical research using the swift workflow system. In Studies in Health Technology and Informatics: From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences – Proceedings of HealthGrid 2007, H. M. Nicolas Jacq, I. Blanquer, Y. Legré, V. Breton, D. Hausser, V. Hernández, T. Solomonides, and M. Hofmann-Apitius, eds, pp. 207–216.

Taylor, I., Shields, M., Wang, I., and Harrison, A. (2006). Visual grid workflow in Triana. *J. Grid Comput.* 3, 153–169.

Thompson, P., Hayashi, K. M., de Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., Herman, D., Hong, M. S., Dittmer, S. S., Doddrell, D. M., and Toga, A. W. (2003). Dynamics of gray matter loss in Alzheimer's disease. *J. Neurosci.* 23, 994–1005.

Toga, A. W., and Thompson, P. M. (2007). What is where and why it is important. *NeuroImage* 37, 1045–1049.

Tu, Z., Narr, K. L., Dollar, P., Dinov, I., Thompson, P. M., and Toga, A. W. (2008). Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE Trans. Med. Imaging.* 27, 495–508.

Woods, R., Grafton, S. T., Holmes, C. J., Cherry, S. R., and Mazziotta, J. C. (1998). Automated Image Registration: I. General Methods and Intrasubject, Intramodality Validation. *J. Comput. Assist. Tomogr.* 22, 139–152.

Zhao, J., Goble, C., Stevens, R., and Turi, D. (2007). Mining Taverna's semantic web of provenance. *Concur. Comput. Prac. Exp.* 20, 463–472. doi: 10.1002/cpe.1231

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 April 2009; paper pending published: 18 May 2009; accepted: 26 June 2009; published online: 20 July 2009.
*Citation: Dinov ID, Van Horn JD, Lozev KM, Magsipoc R, Petrosyan P, Liu Z, MacKenzie-Graham A, Eggert P, Parker DS and Toga AW (2009) Efficient, distributed and interactive neuroimaging data analysis using the LONI Pipeline. Front. Neuroinform. (2009) **3**:22. doi: 10.3389/neuro.11.022.2009*

frontiers in
NEUROINFORMATICS

ORIGINAL RESEARCH ARTICLE
published: 09 October 2009
doi: 10.3389/neuro.11.035.2009

Bio-Swarm-Pipeline: a light-weight, extensible batch processing system for efficient biomedical data processing

Xi Cheng[1]*, Ricardo Pizarro[1], Yunxia Tong[1], Brad Zoltick[1], Qian Luo[2], Daniel R. Weinberger[1] and Venkata S. Mattay[1]

[1] *Neuroimaging Core Facility, Genes, Cognition and Psychosis Program, National Institute of Mental Health/National Institutes of Health, Bethesda, MD, USA*
[2] *Mood and Anxiety Disorders Program, National Institute of Mental Health/National Institutes of Health, Bethesda, MD, USA*

Edited by:
John Van Horn,
University of California, USA

Reviewed by:
Shiyong Lu, Wayne State University, USA
Juliana Freire, University of Utah, USA

****Correspondence:***
Xi Cheng, Neuroimaging Core Facility, Genes, Cognition and Psychosis Program, National Institute of Mental Health, Building 10, Room 3C-210, Bethesda, MD 20892, USA.
e-mail: chengx@mail.nih.gov

A streamlined scientific workflow system that can track the details of the data processing history is critical for the efficient handling of fundamental routines used in scientific research. In the scientific workflow research community, the information that describes the details of data processing history is referred to as "provenance" which plays an important role in most of the existing workflow management systems. Despite its importance, however, provenance modeling and management is still a relatively new area in the scientific workflow research community. The proper scope, representation, granularity and implementation of a provenance model can vary from domain to domain and pose a number of challenges for an efficient pipeline design. This paper provides a case study on structured provenance modeling and management problems in the neuroimaging domain by introducing the Bio-Swarm-Pipeline. This new model, which is evaluated in the paper through real world scenarios, systematically addresses the provenance scope, representation, granularity, and implementation issues related to the neuroimaging domain. Although this model stems from applications in neuroimaging, the system can potentially be adapted to a wide range of bio-medical application scenarios.

Keywords: scientific workflow, provenance, neuroimaging, neuroinformatics, swarm

INTRODUCTION

Scientific workflow systems that are capable of tracking the details of data processing history can facilitate a number of fundamental requirements in everyday scientific research, such as scheduling batch processing on multiple computers, interpreting and comparing different results, sharing and reusing existing workflow, etc. In the scientific workflow research community, the information that describes the details of data processing history is referred to as "provenance" (also "lineage" or "pedigree") (Simmhan et al., 2005). Provenance management is a critical component of scientific workflow systems and most of the existing popular scientific workflow systems have a module for management of provenance information. For e.g., the *Kepler* workflow system is able to collect the provenance information (Ludäscher, 2006), while the *Taverna* workflow system stores the provenance information for users to manage and reuse previous workflows (Oinn et al., 2004). *VisTrails* is a provenance management system (PMS) that provides infrastructure for data exploration and visualization through workflows (Callahan et al., 2006; Silva et al., 2007; Koop et al., 2008). The *Swift* workflow system builds on and includes technology previously distributed as the GriPhyN Virtual Data System to capture the provenance (Zhao et al., 2007). The *Pegasus* workflow system also uses the Virtual Data System to capture the provenance (Miles et al., 2008). The *VIEW* workflow system manages the provenance data with a provenance management module (Lin et al., 2009). The *LONI* workflow system has a provenance management framework to manage the provenance data (MacKenzie-Graham et al., 2008).

Despite its importance, however, provenance modeling and management is still a relatively new area in the scientific workflow research community (Simmhan et al., 2005) and the provenance model can vary from domain to domain (Freire et al., 2008). In particular, although the *VisTrails*, *Swift*, *VIEW* and *LONI* workflow systems have been applied to neuroimaging, a number of provenance modeling and management issues that are specific to the neuroimaging domain have to be explored further:

(Q1) The provenance model varies from domain to domain and has to be identified and appropriately customized for the neuroimaging domain. First, as the neuroimaging databases, such as *XNAT* (Marcus et al., 2007), HID (Keator et al., 2008) and NDAR (Ndar 2009), manage raw data provenance information, the neuroimaging workflow systems should be customized to work seamlessly with neuroimaging databases to minimize duplicated efforts and storage redundancy for provenance management. Second, as the neuroimaging domain involves domain specific user interaction and annotation, the provenance model should be extended to include this kind of information.

(Q2) The representation of the provenance is still not well addressed in the scientific workflow research community and needs to be adequately addressed in the neuroimaging domain. Improper representation of the provenance can result in huge redundancy. One way of minimizing the redundancy is to structure the provenance into layers of normalized components (Freire et al., 2008). However, definition of the layers and components can still vary from domain to domain. In particular, this issue needs to be appropriately addressed in the neuroimaging domain.

(Q3) The provenance granularity can vary across domains, and has not been explicitly explored for the neuroimaging domain. The provenance can be recorded at different levels of granularity, i.e., varying levels of details. Improper selection of the granularity of

provenance can produce inordinately large volume of provenance data bigger than the data it describes (Simmhan et al., 2005), which may not be useful and may be hard to manage. In the neuroimaging workflow system, theoretically, the provenance granularity can be set at voxel-level, slice-level, volume-level, session/visit-level, subject-level, or group-level. Variability in granularity can result in a big difference in performance and storage overhead. The optimal provenance granularity for the neuroimaging workflow has hitherto not been explicitly explored in the existing literature.

(Q4) The provenance model can be implemented in many different ways which vary from application to application. The different approaches can vary in the way they capture, store and retrieve the provenance information. The capturing mechanism can be at various levels, i.e., at the OS-level, processing-level and workflow-level. The storage mechanism can be either file-system based or database based. The retrieval mechanism can be a special scripting language, like SQL or a visual user interface. When a new neuroimaging provenance model is created, the corresponding implementation issues have to be properly addressed as well.

In general, proper solutions for provenance modeling and management problems need to be explored for the neuroimaging domain. In this paper, we introduce the Bio-Swarm-Pipeline (BSP), a scientific workflow management system for bio-medical research developed at the Genes, Cognition and Psychosis Program (GCAP) of NIMH/NIH. It was designed to facilitate the fundamental requirements for everyday scientific research, such as scheduling batch processing on multiple computers, interpreting and comparing different results, sharing and reusing existing workflows, etc. This system is based on a new provenance model developed to meet the needs specific to a neuroimaging workflow management system. It systematically addresses the issues involved in the provenance modeling and management in the neuroimaging domain. First, by proper extension of the provenance model, the workflow management system can work seamlessly with existing neuroimaging databases and effectively reduce unnecessary storage and developing efforts. Second, by properly structuring provenance into two layers of six independent sub-provenance components, the BSP effectively minimizes the recording redundancy of provenance; Third, by proper determination of the provenance granularity, the BSP effectively eliminates unnecessary information, makes the system more light weighted and manageable; Fourth, by providing an optimal number of user interfaces, it makes provenance management and task scheduling an efficient and effective procedure. Finally, by taking swarm as analogy, an unsophisticated user with little or no knowledge in programming can easily capture the core concepts and understand how a task is processed by the system. Although this system stems from applications in the neuroimaging domain, the system can potentially be adapted to meet the requirements for a wide range of bio-medical application scenarios.

The remainder of this paper is organized as follows: in the methods section, we describe the BSP system architecture, highlight the structured provenance model, and demonstrate how it works with real examples; in the results section, we describe the current application status and impact of the system to the work at the GCAP of NIMH; and in the discussion section, we discuss how provenance modeling and management problems were addressed in the BSP. We also discuss some additional features and future extensions.

METHODS (BSP SYSTEM ARCHITECTURE)

The BSP system architecture is made up of three layers – (I) pipeline interface layer, (II) PMS layer, and (III) data processing clients (DPCs) layer as shown in **Figure 1.** The interface layer interacts with the PMS layer to submit data processing tasks and tracks the data processing provenance information. The DPC layer interacts with the PMS layer to perform data processing and updates provenance information. In this section, we will highlight layer II, i.e., PMS, and demonstrate how it works. We will also briefly introduce layer I and layer III.

PIPELINE INTERFACE

The interface enables the user to interact with the PMS to submit data processing tasks and tracks the data processing provenance information. Each interface is presented in the next section along with the provenance data that is managed.

PROVENANCE MANAGEMENT SYSTEM

The PMS manages a structured provenance model as shown in **Figure 1**. The design and implementation is based on the MySQL relational database system. Conceptually, the model can be divided into two layers.

The first layer contains three sub-provenance components, i.e., task/job provenances, static workflow provenances and computational resource provenances. The task provenances record all the necessary information to reproduce a specific data processing result, including the run-time task provenances (such as result location, processing time, status) and user annotations as well as the references to static workflow provenances and the computational resource provenances.

In the second layer, the sub-provenance components are further decomposed. For example, the static workflow provenances are further divided into wrapper provenance, parameter provenance and data source provenance. The computational resource provenances are further divided into storage provenance and DPC provenance.

In this section, we will first introduce the static workflow and the computational resource provenances, and then introduce the task/job provenance. Later on in the discussing section, we will also discuss how this model systematically addresses the provenance modeling problems mentioned in Section "Introduction".

Static workflow provenances

Static workflow provenances are specifications about workflows which can be shared across different tasks. This includes the specification of the wrappers, processing parameters and data sources.

Wrapper provenance management. The wrapper provenance management module manages the specification of the wrapper libraries for different data processing packages. Each data processing package (e.g. SPM (Friston et al., 1995), AFNI (Cox, 1996), VBM (Ashburner and Friston, 2000), FreeSurfer (Dale et al., 1999), etc.) is encapsulated by a wrapper so that they have a uniform calling interface like *do* + *package_name* + *version* + *release*. Each wrapper is uniquely identified by a wrapper ID so that the task/job provenances component can be simplified by referring to the wrapper ID.

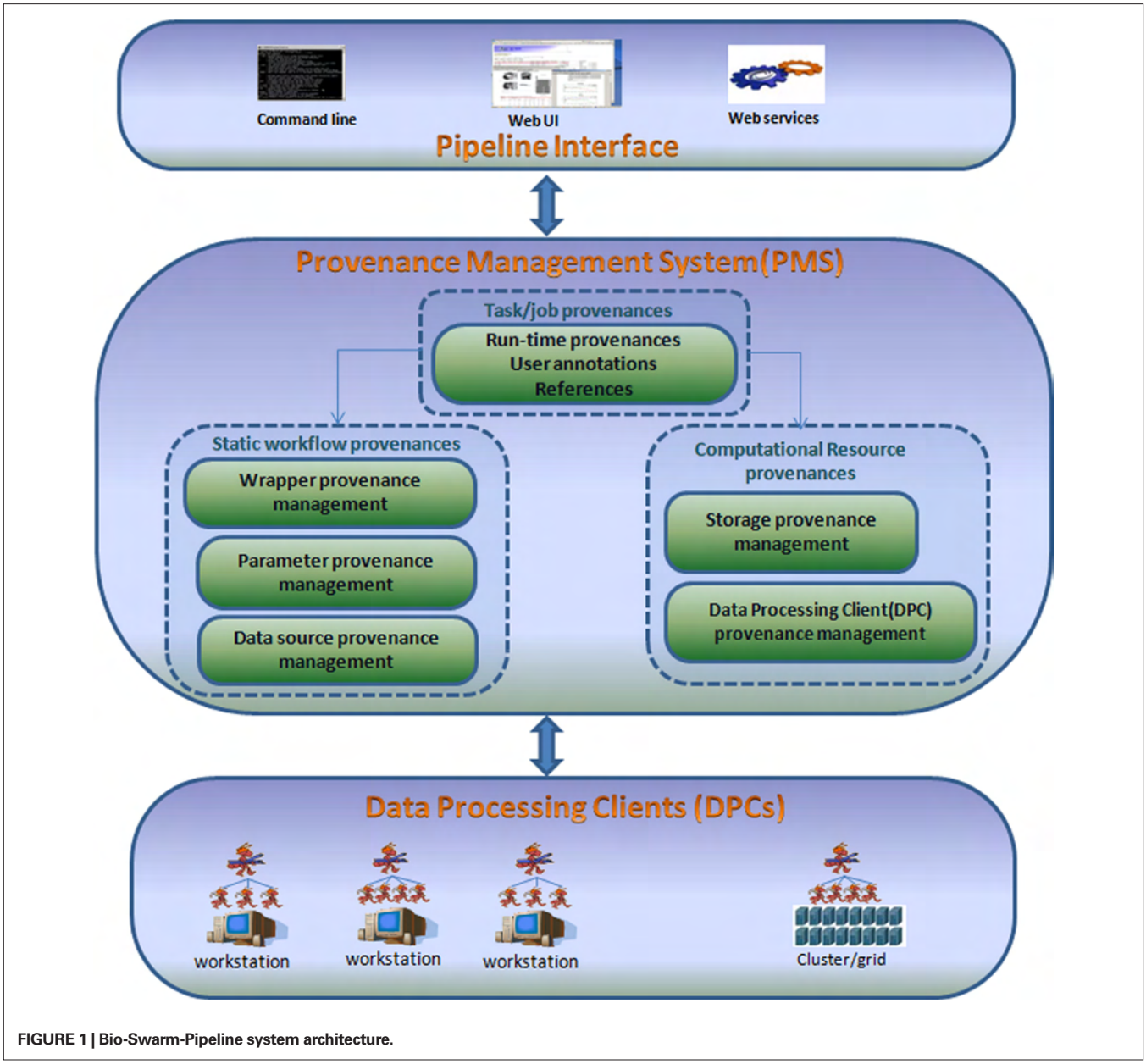


FIGURE 1 | Bio-Swarm-Pipeline system architecture.

Parameter provenance management. The parameter provenance management module manages the instantiated processing parameters for each given wrapper. It tracks everything related to the processing parameters that a user may be interested in. These include the parameter ID, the wrapper ID, the user login name who created the parameter set and the parameter body and some comments fields. Within the parameter ID, the user can easily interpret and compare the different results and check if they are generated from the same procedure.

The parameter set is managed in the parameter management interface. As an illustration, in **Figure 2** we use SPM-based first level data processing as an example to help the reader evaluate how the system works. In this illustration the processing parameters with parameter ID 11 are associated with wrapper SPM2. The parameter management interface allows the users to create new parameters or adapt from existing parameters. For the latter, users can first retrieve the parameters they want to duplicate from, and then click "borrow and create" button to make a new parameter. Then the user can modify the parameter as required. When a parameter set is first created, a unique parameter ID is automatically assigned to it. If the parameter set is derived from another parameter, users can add comments to indicate what the parent parameter ID is. This allows users to track the relationships among a family of related parameter sets.

Data source provenance management. The data source provenance management module makes it easy for the workflow management system to inter-communicate with the neuroimaging database and other heterogeneous data sources and take input data from there. By default, BSP was designed to be work

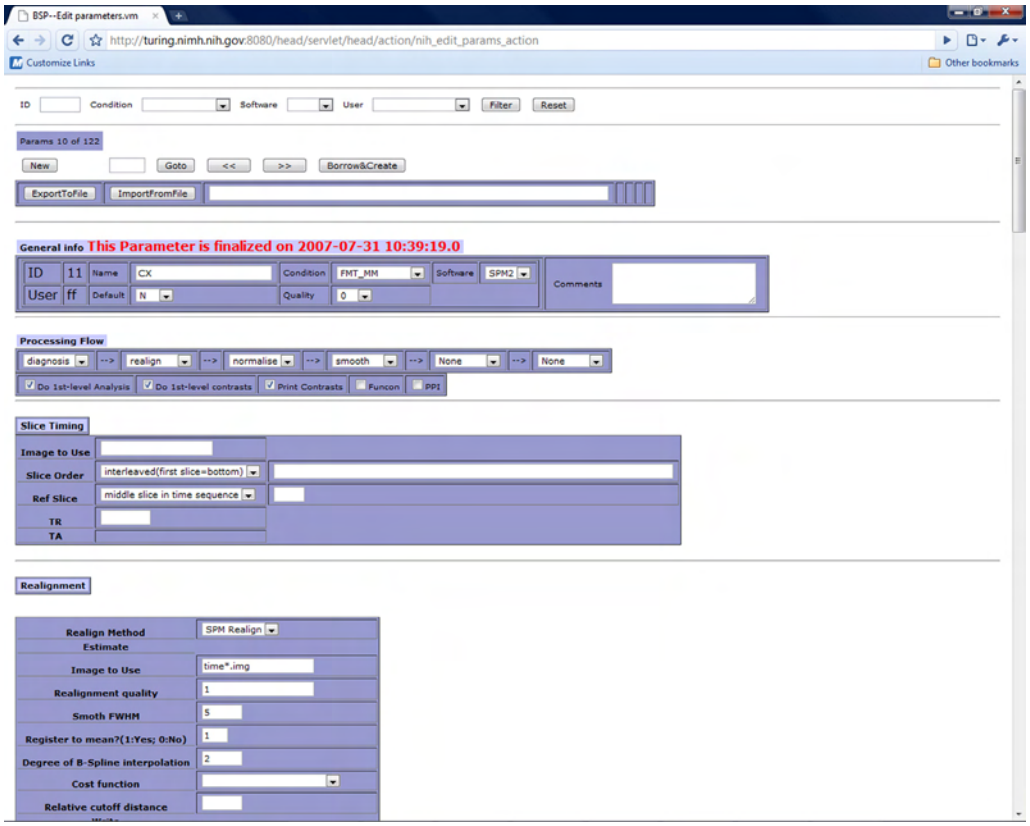


FIGURE 2 | Parameter management interface.

seamlessly with *XNAT@GCAP* neuroimaging database (Cheng et al., 2008). It assumes that the raw data provenance information, such as the raw data locations, data acquisition parameters and subject demographics information are all managed by the neuroimaging database with *XNAT* like database schema (Marcus et al., 2007). Therefore only references to the raw data provenance are kept in the system. The data source provenance component is able to retrieve raw data provenance information from the neuroimaging database when necessary.

Computational resource provenances

Computational resource provenances are specifications about computational resources that can be shared across different tasks. This includes specifications about the DPC and storage devices.

Data processing clients provenance management. The DPCs provenance management module manages the profiles of heterogeneous client workstations in the database. The profile describes the following information: hostname, system architecture, processor speed, memory capacity, operating system, network speed, available storage space, version of wrapper libraries and traffic lights, etc.

The storage provenance management. The storage provenance management module simplifies the dynamical management of the mappings between processing parameters and storage devices. The mappings are defined by the storage allocation rules in the format of (*parameter_ID, output_dir, is_active*), which specify a list of alternative output directories for each parameter IDs. When the available physical storage is below a certain threshold, the *is_active* flag will be automatically set to 0 by a daemon program so that the DPC will try to find the next available output directory with adequate space for ensuing tasks. If there is no output directory with adequate space, DPC will switch the task into "pending" status. When an output directory with adequate space is made available, DPC will automatically enable processing of the pending tasks.

Task/job provenances management

The task provenances record the most detailed information to reproduce an individual result. This includes the run-time task provenances (such as result location, processing time, status) and user annotations (such as data quality notes, etc.) as well as the references to static workflow provenances and the computational resource provenances defined above.

All the task provenance information is maintained in the task table of the database as shown in **Figure 3**. Each task is identified by a unique ID in the task table. The task table specifies the following information for task processing: such as the priority of the task, the wrapper ID and the parameter ID required to process this task, and the name of the DPC (hostname) assigned to handle the task so that different tasks can be assigned to different

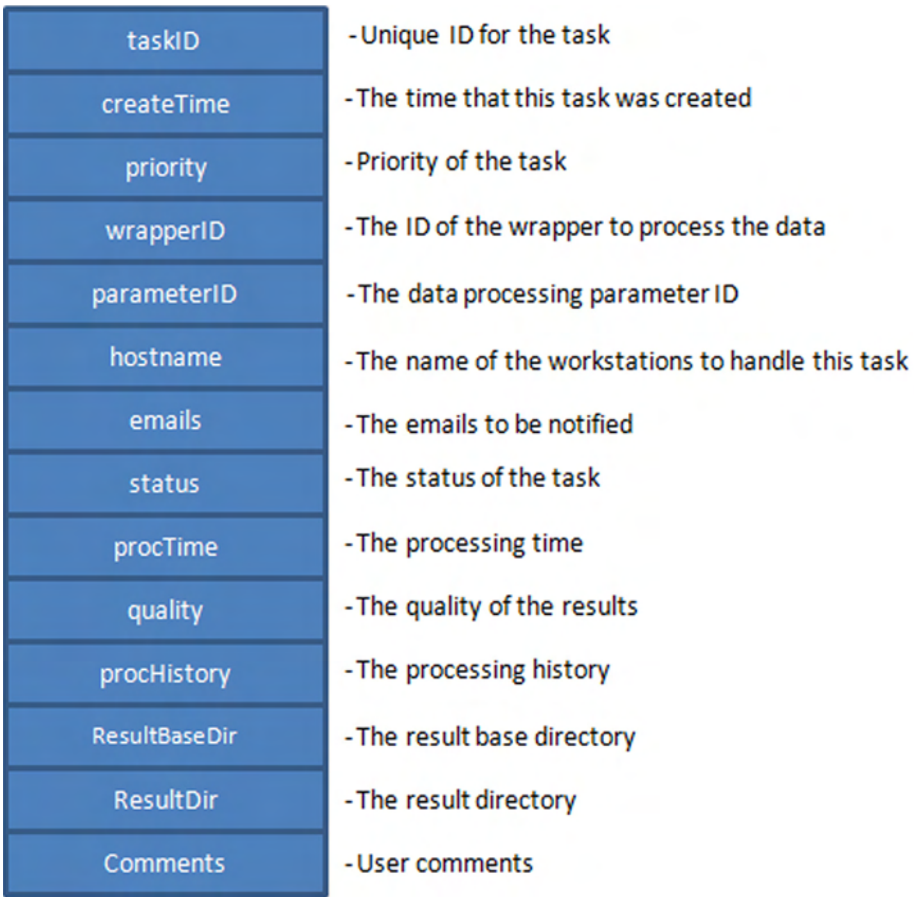


FIGURE 3 | Structure of the task table in the database.

hosts and processed in parallel, the directory where the results will be outputted, and the email addresses that should be notified upon completion of the tasks. Most of this information is collected before the task execution.

Moreover, the task status and quality annotations are customized to accommodate the requirements in neuroimaging domain. In our system, the possible status of a task can be "pending", "ready-for-processing", "failed", "ready-for-review" or "reviewed" as shown in **Figure 4**.

After a task is being created, if all the required data sources are ready and the storage space is available, the task will be set to the "ready-for-processing" status. Otherwise it will be set to a "pending" status. When processing has completed successfully, the pipeline is switched to "ready-for-review" status by the DPC. After the review is complete, it can be switched to the "reviewed" status by the user through the user interface, and the quality of the task will be marked as either + or − to indicate whether the results are usable. If the processing fails, the status will be set to "failed" by DPC, allowing the administrator to fix the error, and switch the status to "pending" or "ready-for-processing".

To make the provenance model lightweight, the optimal granularity level of the provenance must be chosen. The available choices in the neuroimaging domain are voxel-level, slice-level, volume-level, session/visit-level, subject-level, and group-level. In most occasions, researchers may only be interested in the session/visit-level provenance information. The provenance information at this level is easily manageable, so the BSP provenance

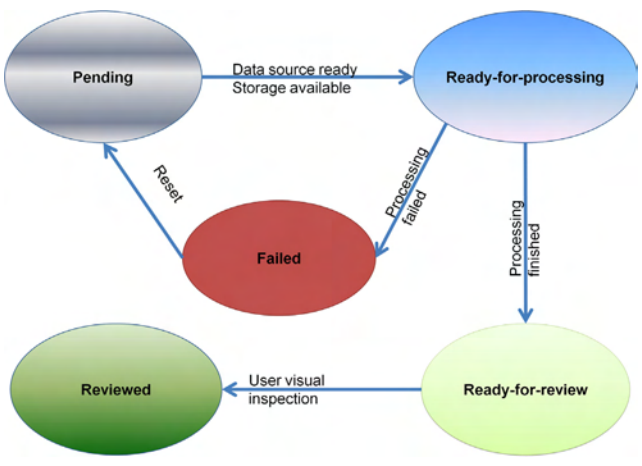


FIGURE 4 | Flow chart of task status.

model is explicitly set at this level. However, the system can be easily extended to work on a different level of granularity when necessary.

A series of web user interfaces are provided to make it easier for the user to submit tasks, retrieve tasks and review the results on-line. In the following demonstration, we will again take SPM-based batch processing as an example to help the reader further evaluate the system.

First, a user can submit tasks through the task management interface as shown in **Figure 5**. Here, the user can create a new data processing project by click the "**append" button. The user can then** specify the project name, the parameter ID, the machine list to be used and a list of e-mail addresses where notifications can be sent upon the completion of the task. Afterwards, the user can click the "**add session" button to add data and click the "place order"** button to place data processing task (we also call them orders) for them. Multiple dataset (sessions) can be added to the project at the same time. Each dataset will be assigned to an individual task. If multiple machines are provided, the tasks will be split evenly among them. The processing status of each dataset is also available in the task management screen. As can be seen in the lower part of the **Figure 5**, each processed session now has a green check mark on the left side of the row.

After the tasks have been submitted, the user can log off and wait for the process to finish. Upon the completion of each task, the user will get an email notification indicating the status of the task. An example of the email notification is shown in **Figure 6**. If a task finishes without error, the user can log into the web interface, as shown in **Figure 7**, to query the results by subject ID, task IDs or parameter ID.

When the user provides the parameter ID and clicks the "**get** results" **button (Figure 7**), the corresponding results of the tasks will be available for on-line review through the web user interface as shown in **Figure 8**. In our illustration using SPM-based first level data processing the following results are made available for inspection: preprocessed results (lower left plot in **Figure 8**), contrast map (middle plot in **Figure 8**), quality control images and

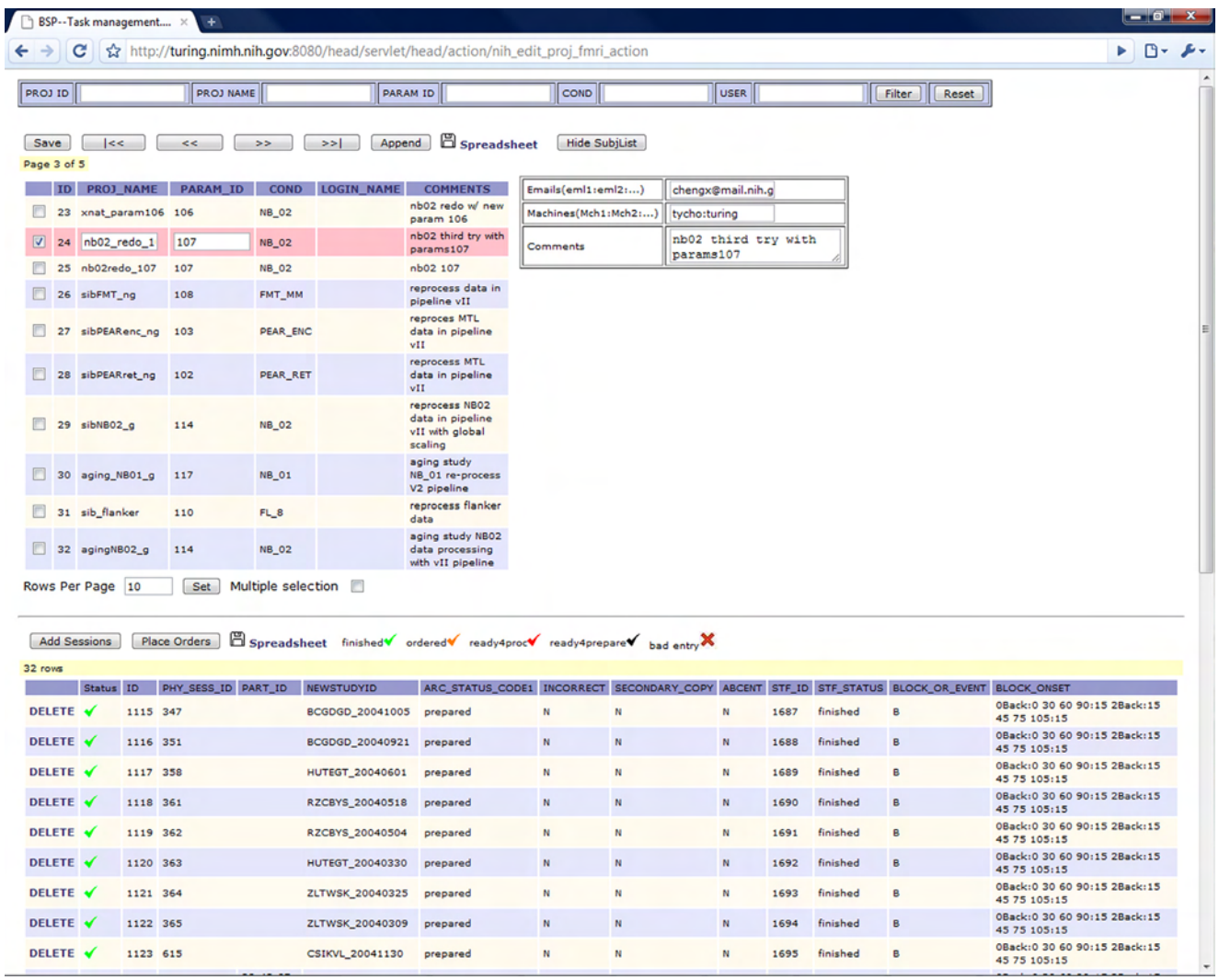


FIGURE 5 | Task management interface.

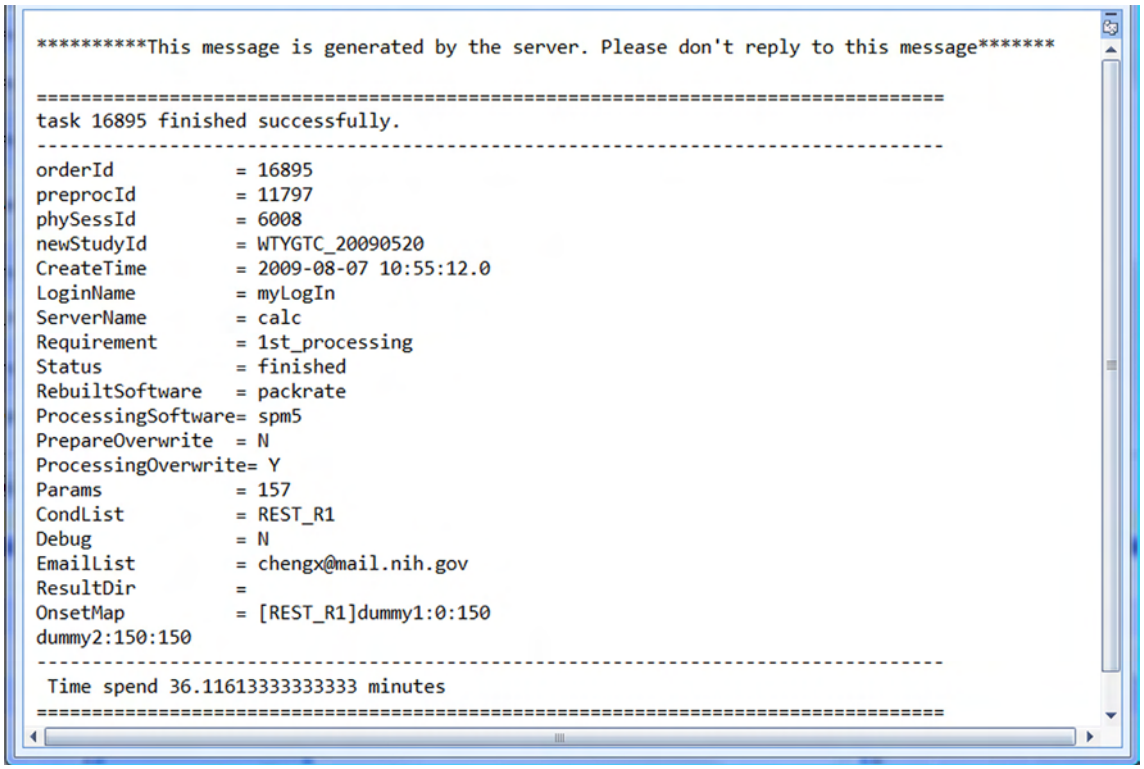


FIGURE 6 | Example of email notification.

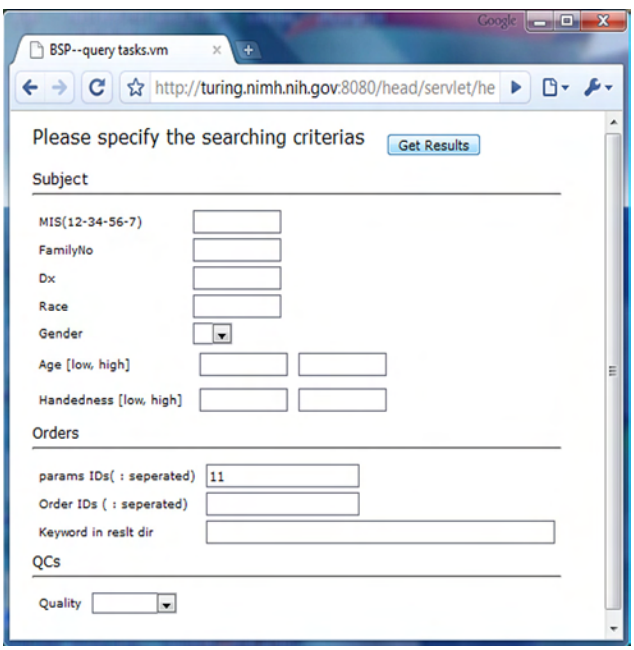


FIGURE 7 | Querying the processing results.

measures (right plot in **Figure 8**). The user can add annotations concerning the quality of the results after visual inspection. The quality information can then be used in a query to filter the datasets.

There are also a number of command line tools used for simplifying routine administrative tasks such as task scheduling, traffic control and diagnosis.

DATA PROCESSING CLIENTS

DPCs manage data processing in each client workstation by communicating with PMS. Each DPC is made up of two types of swarms, i.e., the manager-swarm (M-swarm) and the worker-swarm (W-swarm). The M-swarm, running as services on each local computer, manages (i.e., creates or kills) W-swarms in the local computer, retrieves tasks from the task table by local host name and task priority, and dispatches them to be processed by W-swarms. The W-swarm takes the task from M-swarm and processes them. The DPCs can be extended to incorporate computational resource(s) from a high-performance computing center like Beowulf cluster (Gropp et al., 2003) by installing a customized DPC.

RESULTS

The BSP has been built, maintained and supported by GCAP since 2006. To date, around 130 workflows and above 32000 data processing tasks have been completed through this scientific workflow management system, with each task taking about 10–60 min. Currently the system supports SPM (for fMRI and VBM) and Freesurfer based data processing, but other popular packages or in-house packages can be easily integrated as well. The workflow system in its current status has been playing a critical role in the day-to-day neuroimaging research within GCAP, including but not limited to the following aspects:

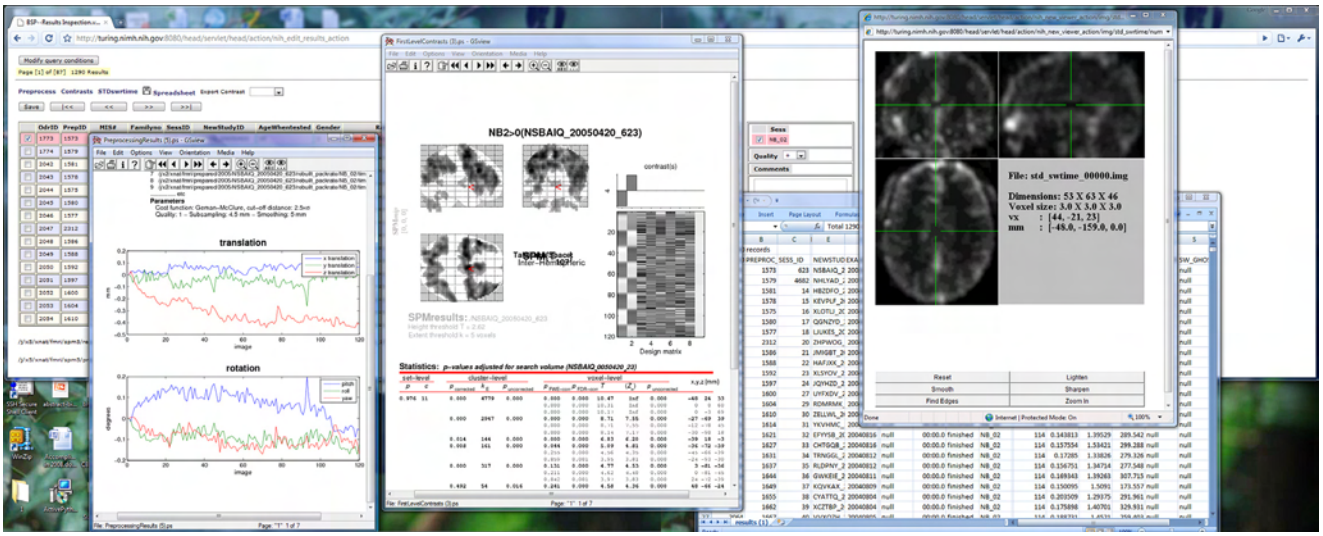


FIGURE 8 | Visual inspection interface.

- Improved productivity: due to its parallel data processing capabilities, the workflow system has refreshed data processing records in the past 2 years. The most recent accomplishment has been to successfully process almost 3000 high-resolution structural MRI preprocessing for VBM in <1 week, and nearly 5000 fMRI first level data processing for SPM in 2 weeks. This has been achieved using just one W-swarm per workstation because some of the packages were not multi-thread-able. The modern quad-core processor will be able to easily scale up to four W-swarms without compromising performance. Making the data processing packages multi-thread-able will make this a more efficient process. This kind of capability makes the BSP very efficient even when compared to the crowded Beowulf cluster.
- Improved efficiency of workflow within GCAP neuroimaging research groups. By enabling automated data processing of large datasets, it has made more time available for researchers to pursue more intellectually challenging tasks.
- Facilitates easy and efficient replication of results using identical parameters. This decreases the necessity to backup processed data and thereby decreases storage space requirements.

DISCUSSION

In this section first we discuss how the BSP provenance model addressed the provenance modeling and management problems mentioned in Section "Introduction" for the neuroimaging domain. Second we summarize the additional features of BSP along with the provenance model. Finally we list some possible future extensions.

THE BSP PROVENANCE MODEL ADDRESSES THE PROVENANCE MODELING AND MANAGEMENT PROBLEMS IN THE NEUROIMAGING DOMAIN

The BSP provenance model has systematically addressed the provenance modeling and management problems for the neuroimaging domain (Q1–Q4) as outlined in Section "Introduction:"

(P1) the BSP provenance model is extended to cover the neuroimaging domain. First, the BSP is extended to work seamlessly with the *XNAT@GCAP* (Cheng et al., 2008) neuroimaging data archiving system, i.e., the BSP data source component just keeps the references to the raw data provenance, such as the data acquisition parameters as well as the subject's demographic information, and is able to retrieve the raw data provenance information from the *XNAT@GCAP* neuroimaging database as necessary. Therefore, the duplicated efforts and storage redundancy for the maintenance of the provenance information are minimized. Second, the BSP model is extended to include information specific to neuroimaging. Particularly, the system is customized to accommodate the domain specific user interactions for reviewing the quality of the images, for example the task status field is extended to include options like **"ready-for-review", "reviewed", etc**. After a task is reviewed, the annotation and comments related to the data quality can be stored. Special user interfaces (see **Figures 2, 5, 7 and 8**) are also provided for the user to manage parameter sets, make queries, visually inspect the results, and manage the annotations. These extensions are different from existing workflow systems. For example, most workflows except *LONI* do not work with neuroimaging databases.

(P2) the BSP provenance model was structured into two layers of six independent sub-provenance components (i.e., wrapper provenance, parameter provenance, data source provenance, storage provenance, DPC provenance and task provenance) to minimize the recording of redundant information. Referring to **Figure 3**, although the task provenance component tracks all the details necessary to reproduce the results, the storage overhead are very small, as most of the common information (such as the static workflow provenance and the computational resources provenance) is stored as references. In general, the BSP provenance model structure is quite different from that of other existing provenance models. For example, in VisTrails, the provenance model is structured into three layers: the workflow evolution, the workflow instance and the execution log (Freire et al., 2008). In the LONI workflow system,

the provenance model is divided into four components: the data provenance, the binary provenance, the executable provenance, the workflow provenance, the processing provenance (MacKenzie-Graham et al., 2008). Although these systems have some features that are similar to the BSP provenance model, the overall structure is different.

(P3) in the BSP provenance model, the provenance granularity is explicitly selected to be at the session-level so that only information of interest to the user is tracked. However, there is no limitation if a user wants to extend the current model to include other levels of the provenance. Usually the provenance granularity for the neuroimaging domain is not explicitly specified in other workflows. As mentioned before, without explicitly specifying the level of granularity, the neuroimaging workflow system can potentially store too much detailed information – such as the provenance at the slice or voxel-level, which can result in a huge and unnecessary storage overhead. However, most users may not be interested in such fine-grained provenance.

(P4) the provenance model, implementation has been carefully chosen in the BSP to optimize the performance. First, most of the provenance information is collected prospectively (e.g., the wrapper provenance, parameter provenance, data source provenance, storage provenance, DPC provenance are all specified before task execution). Only little information in task provenance is collected retrospectively. Compared to the OS-level capturing mechanism, which needs to filter through all the system calls and files touched during the execution of a task's, this approach is more efficient. Second, the BSP provenance model is based on a relational database system. In comparison to file-based provenance storage system, the data storage is optimized by the database system, and the query/retrieval stage is more flexible and efficient.

Although the BSP provenance model originates in the neuroimaging domain, it can be potentially adapted to cover many other bio-medical domains as well.

ADDITIONAL FEATURES OF THE BSP

Along with the BSP provenance model, here we would like to summarize some additional features of the BSP in general.

- BSP is parallel in nature
  In contrast to workflow systems that are primarily designed to handle inter-package heterogeneities but do not facilitate parallel processing, the BSP allows optimal distribution of multiple data processing tasks across a number of computers to maximize the throughputs.
- BSP is light weighted
  This is because: (A) The swarm is conceptually simple, an unsophisticated user can capture the core concepts and understand how a task is processed by the system fairly quickly without having to read the whole manual; (B) The system boundary is properly tailored, so that duplicated work is avoided; (C) The redundancy of provenance data is minimized as the provenance model is highly normalized; (D) The granularity of the provenance is set at session level, the unnecessary provenance information is effectively ignored.
- BSP is built on top of the relational database
  This is a big advantage of the BSP over workflow systems that are not bundled with a database. With the powerful MySQL database and SQL language, routine management tasks such as wrapper management, DPCs management, task/job management, data source management and storage management can be very easy and flexible.
- BSP is reliable
  The failure of one machine will not affect the data processing on another in the network, and is therefore easy to identify and recover from failure.
- BSP is scalable
  – As there is no communication between the different processing tasks, the throughputs of the workflow system increases almost linearly with the number of workstations.
  – A work station can join or leave the workflow system at any time without affecting the overall batch processing.
- BSP is extensible
  – The workflow system can be extended to cover different data processing packages as long as the appropriate wrappers are provided.
  – The DPCs are extensible. For example, a high performance computing center like Beowulf cluster can be treated as a DPC and managed by the workflow system
  – Data sources can be extended to accommodate a wide range of different data sources as long as the appropriate data source adaptors are provided.

The BSP is flexible and has a number of other advantages. For example, when compared to the Beowulf cluster, it is: (1) capable of applying complicated and flexible data processing management; (2) free from limited license issue (e.g., the Beowulf cluster usually limits the number of Matlab licenses to 16 for each user); (3) no need to transfer data and results back and forth as is required between Beowulf and the local file systems; and (4) no waiting time (in comparison to the high performance computing center).

POSSIBLE EXTENSIONS

As some of the provenance management is currently conducted through command line, more user friendly interfaces will be provided in the new release. These include interfaces for: (1) wrapper management; (2) storage management; (3) task re-scheduling and traffic control; (4) data source management.

ACKNOWLEDGMENTS

This work was supported by GCAP, DIRP, NIMH, NIH. We would like to thank everyone in the GCAP neuroimaging core for their suggestions and feedbacks. We would also like to thank the reviewers for the informative comments and suggestions.

SUPPLEMENTARY MATERIAL

Currently we are creating extensions and developing a deployable release. A link to the downloadable version as well as the educational material will be made available at *The Neuroimaging Informatics Tools and Resources Clearinghouse* (http://www.nitrc.org/).

REFERENCES

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry – the methods. *Neuroimage* 11, 805–821.

Gropp, W., Lusk, E., Sterling, T., and Hall, J. (2003). Beowulf Cluster Computing with Linux, 2nd Edn. Cambridge, MIT Press.

Callahan, S., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., and Vo, H. T. (2006). VisTrails: visualization meets data management. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. Chicago, IL, ACM.

Lin, C., Lu, S. Y., Fei, X., Chebotko, A., Lai, Z. Q., Pai, D., Fotouhi, F., and Hua, J. (2009). A reference architecture for scientific workflow management systems and the VIEW SOA solution. *IEEE Trans. Serv. Comput.* 2, 79–92.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.

Cheng, X., Marcus, D. S., Sambataro, F., Zoltick, B., Tong, Y., Meyer-Lindenberg, A., Wienberger, D., and Mattey, V. (2008). Neuroimaging database of GCAP (XNAT@GCAP) NIMH. In Annual Meeting of the Society for Neuroscience, Washington DC, November 2008.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: segmentation and surface reconstruction. *Neuroimage* 9, 179–194.

Freire, J., Koop, D., Santos, E., and Silva, C. (2008). Provenance for computational tasks: a survey. *Comput. Sci. Eng.* 10, 11–21.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., Frackowiak, R. S. J. **(1995). Statistical parametric** maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210.

Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H. J., and Papadopoulos, P. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172.

Koop, D., Scheidegger, C. E., Callahan, S. P., Vo, H. T., Freire, J., and Silva, C. (2008). Viscomplete: automating suggestions for visualization pipeline. *IEEE Trans. Vis. Comput. Graph.* 14, 1691–1698.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J., and Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency Comput. Pract. Exp.* 18, 1039–1065.

MacKenzie-Graham, A., Payan, A., Dinov, I. D., Van Horn, J. D., and Toga, A. W. (2008). Neuroimaging data provenance using the LONI pipeline workflow environment. *LNCS* 5272, 208–220.

Marcus, D. S., Olsen, T., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit (XNAT): an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34.

NDAR (2009), National Database for Autism Research (NDAR). http://ndar.nih.gov/ndarpublicweb/.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A., and Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045–3054.

Miles, S., Groth, P., Deelman, E., Vahi, K., Mehta, G., **and Moreau, L. (2008).** Provenance: the bridge between experiments and data. *Comput. Sci. Eng.* 10, 38–46.

Silva, C., Freire, J., and Callahan, S. (2007). Provenance for visualization: reproducibility and beyond. *Comput. Sci. Eng.* 9, 82–89.

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *Sigmod Rec.* 34, 31–36.

Zhao, Y., Hategan, M., Clifford, B., Foster, I., von Laszewski, G., Nefedova, V., Raicu, I., Stef-Praun, T., and Wilde, M. (2007). Swift: fast, reliable, loosely coupled parallel computation. In IEEE International Workshop on Scientific Workflows 2007, Salt Lake City, Utah.

Conflict of interest statement: The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Received: 08 April 2009; paper pending published: 09 July 2009; accepted: 09 September 2009; published online: 09 October 2009.
*Citation: Cheng X, Pizarro R, Tong Y, Zoltick B, Luo Q, Weinberger DR and Mattay VS (2009) Bio-Swarm-Pipeline: a light-weight, extensible batch processing system for efficient biomedical data processing. Front. Neuroinform. **3**:35. doi: 10.3389/neuro.11.035.2009*

frontiers in
NEUROINFORMATICS

METHODS ARTICLE
published: 20 October 2009
doi: 10.3389/neuro.11.034.2009

Parallel workflows for data-driven structural equation modeling in functional neuroimaging

Sarah Kenny[1]*, Michael Andric[2], Steven M. Boker[3], Michael C. Neale[4], Michael Wilde[1,6] and Steven L. Small[1,2,5]

[1] *Computation Institute, The University of Chicago, Chicago, IL, USA*
[2] *Department of Psychology, The University of Chicago, Chicago, IL, USA*
[3] *Department of Psychology, University of Virginia, Charlottesville, VA, USA*
[4] *Department of Psychiatry, Virginia Commonwealth University, Richmond, VA USA*
[5] *Department of Neurology, The University of Chicago, Chicago, IL, USA*
[6] *Mathematics and Computer Science Division, Argonne National Laboratories, Argonne, IL, USA*

Edited by:
John Van Horn, University of California at Los Angeles, USA

Reviewed by:
Shantanu Joshi, University of California at Los Angeles, USA
John Van Horn, University of California at Los Angeles, USA

****Correspondence:***
Sarah Kenny, Computation Institute, University of Chicago, 5640 S Ellis Avenue, Chicago, IL 60637, USA.
e-mail: skenny@uchicago.edu

We present a computational framework suitable for a data-driven approach to structural equation modeling (SEM) and describe several workflows for modeling functional magnetic resonance imaging (fMRI) data within this framework. The Computational Neuroscience Applications Research Infrastructure (CNARI) employs a high-level scripting language called Swift, which is capable of spawning hundreds of thousands of simultaneous R processes (R Development Core Team, 2008), consisting of self-contained SEMs, on a high performance computing system (HPC). These self-contained R processing jobs are data objects generated by OpenMx, a plug-in for R, which can generate a single model object containing the matrices and algebraic information necessary to estimate parameters of the model. With such an infrastructure in place a structural modeler may begin to investigate exhaustive searches of the model space. Specific applications of the infrastructure, statistics related to model fit, and limitations are discussed in relation to exhaustive SEM. In particular, we discuss how workflow management techniques can help to solve large computational problems in neuroimaging.

Keywords: exhaustive search, OpenMx, SEM, swift, workflows

INTRODUCTION

CONCEPTS AND BACKGROUND ON CNARI AND DRIVING NEUROSCIENCE USAGE MODEL

In the past decade, there has been tremendous growth in the number and scope of functional brain imaging studies performed in the basic and applied neurosciences. These studies have been more complex than those of the past, often incorporating large numbers of participants, multiple physical sites, longitudinal follow-up, combinations of healthy groups and those with disease or injury, and/or additional types of behavioral or biological measurements. Although their numbers are increasing, the inherent complexity of data management and processing in such studies, particularly regarding anatomical and physiological data, represents a major stumbling block to their ultimate success. In studies of recovery from stroke, for example, medical data are stored in paper charts or in hospital medical information systems, behavioral and linguistic data are saved in spreadsheets on personal workstations, structural and metabolic magnetic resonance imaging (MRI) data are stored in manufacturer formats on scanners and/or with the functional MRI data in the file systems of data processing workstations. With these diverse representations of information, not even counting the possible addition of electrophysiological and other structurally unique data types, it is hard enough to perform single case studies that attempt to relate these data to each other, let alone studies that include statistically meaningful numbers of participants.

We have started building the Computational Neuroscience Applications Research Infrastructure (CNARI) to address these concerns (Stef-Praun et al., 2007; Hasson et al., 2008; Small et al., 2009). There are two basic components to CNARI: First is the use of relational database technology to represent the diverse data types of the study in a uniform representational framework that facilitates distributed data access, and permits powerful queries and data reductions to be performed significantly faster by parallelized statistical analysis procedures. Second is the use of "virtual data" grid computing, in which data and data processing are widely distributed on storage devices and computers, and where data transformation and analysis is specified in terms of abstract ("virtual") procedure descriptions. Together, these techniques enable a community of researchers to access *and share* data and perform data preparation and analysis without detailed knowledge of the internal workings of distributed computing and storage systems or of the network infrastructure that connects them.

Longitudinal functional brain imaging requires comparison of brain activation images within a single individual over time, and possibly also between single individuals and a group that represents some standard. For example, in a study of recovery from brain injury, the individual data might be compared to a normative (healthy) group. Although such comparisons can be performed using various scalar indices, we have recently begun to do this with entire activation networks. One way of modeling such networks of activation is with structural equation modeling (SEM), a method that uses known anatomy to augment the functional information with structural connectivity information, to create a model of both static and dynamic relationships (McIntosh and Gonzalez-Lima, 1994; Buchel and Friston, 1997; Horwitz et al., 1999). We have developed several such models (Solodkin et al., 2004; Skipper et al.,

2007, 2009; Walsh et al., 2008), based on a combination of primate and human data (Ban et al., 1984, 1991; Petrides and Pandya, 1984, 1988, 1999; Rosa et al., 1993; Seltzer and Pandya, 1994; Rizzolatti et al., 1997, 1998; Hackett et al., 1999; Barbas, 2000). In one of these studies, we constructed a group network model for healthy right handed individuals performing bimanual movements, and compared this normative group model to two individuals with different biological states, two healthy left handed people and one individual with stroke. The fit between a strong left hander (i.e., one who used his left hand for everything) and the model was very tight if the hemispheres were flipped in the model. The fits between either the weak left hander (i.e., someone more ambidextrous) or the person with stroke and the group model were poor. These three examples were highly informative for understanding the neurobiology of bimanual movements (Walsh et al., 2008).

Building such models can be very complex and time consuming, requiring advanced anatomical knowledge and skill. Furthermore, while these previous methods have been useful for generating a set of possible models in the absence of exhaustive techniques, they are inherently flawed since they are based on anatomical connectivity data from non-human primates. In addition, the models created depend on the hypotheses being tested, and thus there is a large number of possible models for any particular set of fMRI activation data. To address these issues, we have embarked on an extension to CNARI that aims to facilitate a more objective type of data-driven SEM via highly parallelized workflows for generating and processing large numbers of models in a manner that is easily reconfigurable and replicable. The goal for this modeling approach is to explore as much as possible of the entire space of plausible models that account for the data. In this paper, we discuss the nature of this grid-enabled SEM, and describe how it can be used and applied to various research problems in brain imaging. One of the original purposes of CNARI was to facilitate the study of stroke recovery, with particular emphasis on natural recovery and treatment for language problems (aphasia). In our presentation, we will use specific examples from language processing, though the workflows presented are generalizable to a wide variety of other SEM problems.

CONCEPTS AND BACKGROUND ON SEM: THE MOTIVATION AND DESIGN OF OPENMX

Structural equation modeling (SEM) has a long history dating back to the development of path analysis by Wright (1921). SEM is a statistical tool for estimating a set of predicted covariances between variables that may be connected with either regression (asymmetric, directional) parameters or covariance (symmetric, non-directional) parameters (see Boker and McArdle, 2005, for a review). The advent of high speed computers and high level programming languages in the 1960s, together with advances in statistical methodology led to the development of software for fitting models to observed covariance matrices by maximum likelihood (Joreskog, 1967). This procedure is now commonly known as SEM (see e.g., Bollen, 1989; Loehlin, 1992, for introductions; see McIntosh and Gonzalez-Lima, 1994 for its use in neuroimaging). SEM is widely used for fitting statistical models to epidemiological, psychological, sociological and econometric data where there are multivariate outcomes and theoretical reasons to expect that linear or non-linear systems of equations may provide explanatory power in summarizing these large data sets. For instance, in an epidemiological study of heart disease, one may wish to control for a wide variety of possible behavioral covariates while simultaneously accounting for variance due to group membership or genetic variation. For such problems, SEM models represent state-of-the-art in statistical techniques. Neuroimaging data, is a prime candidate for modeling with SEM, given overlapping sources of variance both across space and time within individual as well as sources of variance across individuals due to group membership and other covariates.

SEM models can be described as a function of two model matrices, *A, S,* a filter matrix, *F* and a residual matrix *U*, such that the expected covariance between observed variables is:

$$R = F(I - \mathrm{A})^{-1}\,\mathrm{S}((\mathrm{I} - \mathrm{A})^{-1})'\mathrm{F}' + \mathrm{U}$$

where the model matrix *A* contains the asymmetric paths (regression coefficients), *S* contains the symmetric paths (covariance coefficients), and the filter matrix, *F*, strips the latent variables from the model matrices so that the result only contains expectations for the observed covariances (McArdle and McDonald, 1984; McArdle and Boker, 1990). One implementation of SEM is the software package Mx (Neale et al., 2003). The set of built-in functions that Mx can optimize includes maximum likelihood, generalized least squares, and full information maximum likelihood analysis of covariance matrices and/or observed means. In 2007, the OpenMx development project was started in order to rewrite Mx into open source, provide a scripting interface to the R statistical language (Ihaka and Gentleman, 1996) and provide a number of extensions to the software. Among these improvements was integrating the Mx SEM optimization engine into parallel workflow management software in order to be able to estimate parameters for large numbers of SEM models simultaneously. In this way, statistical resampling techniques such as bootstrapping, simulations to verify the performance of new models, and exhaustive search routines could make use of large-scale parallel computing resources. The current article describes the first application of the OpenMx software to a real-world exhaustive search problem.

WORKFLOW MANAGEMENT

BACKGROUND AND GOALS

The ability to submit a large number of processes simultaneously to multiple grid sites is a major computational challenge and cannot be accomplished without an evolved workflow management system. In a related research project, we have been developing a workflow system called Swift (Zhao, 2007), which has been our system of choice for submission and management of large-scale workflows for neuroimaging. Using Swift, individual researchers are able to map large amounts of input and output explicitly and make calls to the cataloged executables that sit on remote grid sites. We have been investigating ways to execute and manipulate exhaustive or partially pruned, data-driven SEM workflows using Swift to operate on covariance data derived from a relational fMRI experiment database. From the standpoint of parallel computing and workflow management this poses some interesting issues and also demonstrates, quite strikingly, the convenience (to the research scientist)

of having an elegant, high-level means of expressing, reconfiguring and rerunning such workflows. Here we present several examples of such workflows and explain how they can be expressed and run using Swift, OpenMx and the computational resources of the TeraGrid (Catlett et al., 2007).

The availability of high performance computing systems (HPCs), ranging from multi-core workstations to clusters, grids, clouds, and now petascale supercomputers, creates opportunities to explore experimental datasets with SEM in ways never before possible. The availability of this computing power, however, can be difficult to harness, particularly for a neuroscientist not versed in high performance computing. For these researchers, it is undesirable to divert mental and manual effort from scientific exploration to the mechanics of large-scale parallel computing. At the same time, both the complexity and the scale of high performance environments makes it ever more challenging to assure the validity of scientific results obtained via such systems.

What scientists in general - and neuroscientists in particular - need, are ways to express the processing they want to perform in a compact, abstract, high-level notation that specifies only the logical nature of their computations, but which abstracts and automates all of the potential, varying details of implementing those computing abstractions across a wide range of computing platforms.

SWIFT AND CNARI

For the past two years our group at the University of Chicago Human Neuroscience Laboratory, in collaboration with the Computation Institute, has been developing and evaluating Swift, a parallel scripting language, for this purpose. Together with members of the OpenMx project described above, we have recently focused significant effort to create a library of Swift procedures for the flexible processing and analysis of data from fMRI and other neuroscience experiments.

We employ a programming model that "loosely couples" application programs. In this model, complete programs become our functions, and the arguments to, and results from these functions can be files, file-structured datasets, as well as database entries.

The goals of expressing data processing steps in an abstract notation are multifold: 1) to distill the computation down to the salient details and eliminate the mechanical details of file manipulation from the expression of the basic workflow steps; 2) to abstract data at a high level to relieve the programmer of concerns for the layout of the data on storage systems; 3) to enable the automatic parallelization of scripts in which independent streams of data are processed; and, 4) to enable the recording of all of the steps of a computation in an automatic, transparent manner. An overview of the scripting modules for SEM analysis, coded by the research scientist within the CNARI framework can be seen in **Figure 1**. The Swift programming language enables this model by providing the ability to represent application programs as procedures, and to define compound procedures that permit the user to create libraries of higher level processes that capture the essential protocols of an application's data preparation and analysis. The language's data model provides the ability to describe the datasets that are consumed and produced by the procedural abstractions by combining basic primitive data type definitions with a mapping mechanism of on-disk directory structures to form structures and arrays. These data objects are then automatically and transparently sent across distributed execution environments to remote and parallel Swift procedures.

The Swift language has a C-like syntax, but enforces many of the semantic aspects of a "functional" programming language. Procedures are expressed as functions, permitted to return multiple values; statements are executed in data-dependency order; variables (including array elements and structure members) are

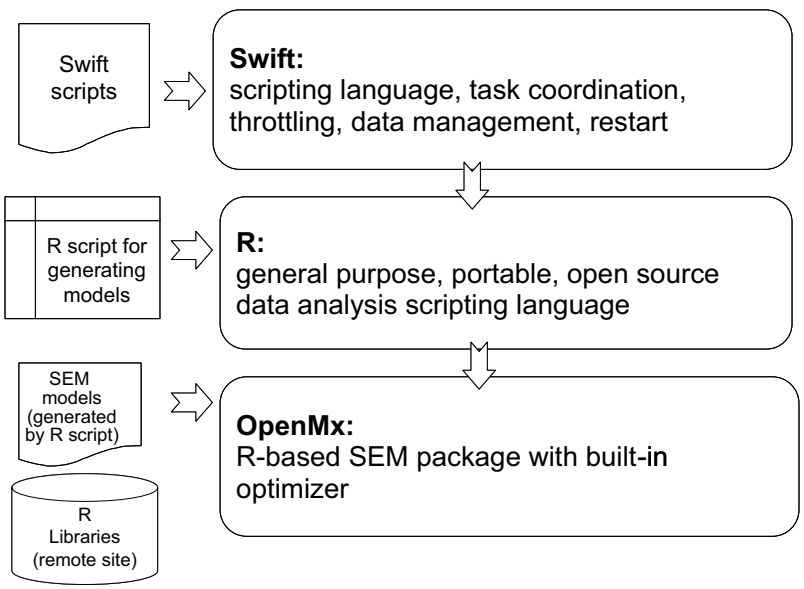


FIGURE 1 | User Interface: Overview of the CNARI scripting modules for SEM workflows.

single-assignment, making it significantly simpler to determine independent operations and threads of control, and to execute these threads in parallel; a construct called "mapping" is provided to translate between the simple, clean regular abstract data model of Swift and the potentially messy, complex model of real-world directory structures and the file naming and structuring conventions expected by real-world applications.

The notation provides a simple set of flow-of-control statements, such as *if* and *switch* (case) statements. The primary way to express a set - potentially large - of parallel operations in swift is to utilize the *foreach()* statement. This statement iterates over a collection, assigning each member of the collection to a control variable, and then evaluating the body of the *foreach()* loop once for each value of the target collection. All iterations of a *foreach()* are potentially (and conceptually) performed in parallel; the runtime system provides appropriate "throttling" and scheduling of the potentially enormous number of parallel operations that this construct can generate and submit for processing.

Atomic procedures in Swift consist of wrappers around specifications that detail the invocation of application programs. In our SEM project, this mechanism is used by Swift to invoke the individual parallel model optimizations of the many thousands of models generated in an OpenMx SEM analysis workflow. "R" is the application program of execution. The invoking (master) program that calls the individual R programs creates a (potentially very long) list of evaluations, each of which is an R expression that embodies the OpenMx engine. The master program generates a large set of model calls and marshals the model's matrix into a text character stream.

The Swift model of data abstraction was to some degree inspired and motivated by the field of fMRI data analysis. In our earliest efforts to execute fMRI preprocessing workflows on computing grids we observed that the data model of the fMRI domain had a natural tree structure in which the vast number of files stored in traditional file system directories had somewhat similar patterns. These files included data from myriad experiments, test conditions and scans, and also included various types of lower level data such as anatomical and time series data represented in the image/header file pairs of the functional data format (e.g., Analyze or AFNI formats). This suggested to us that data definition constructs could be of significant benefit for scientific workflow scripting, such that data could be described in a "typed" fashion, much like the hierarchical model of "structs" in C or "classes" in Java. To enable an organization (or even a discipline, through community curation efforts such as those managed by collaborations like BIRN)[1] to define and standardize a uniform format for describing their common data elements, Swift provides the notion of data type and "mapping" of each type to a physical representation. The logical type is simple and abstract, and reflects only the logical level of the data; the "mapping" describes how each element of a structure is mapped onto the structure's physical representation on a file system. To some extent, Swift emulates the mapped filesystem structure on the remote resources where it instantiates processing. Generic mappers with a modest degree of representational flexibility are pre-defined in the swift system; but additional mappers can be created by users for their own communities and used throughout. **Figure 2** shows the Swift modules used for execution management once a user has mapped his files, and defined processing jobs within a Swift script.

Swift is easy for users to install, and its runtime system provides the client capabilities needed to use workstation, grid and cluster computing resources. From a single client computer, e.g., a modest workstation or personal laptop, the user can launch and control scripts that send parallel work for simultaneous execution on clusters, grids and supercomputers. The user can test the correct execution of the logical script workflow, just by executing directly

[1]http://www.loni.ucla.edu/BIRN/

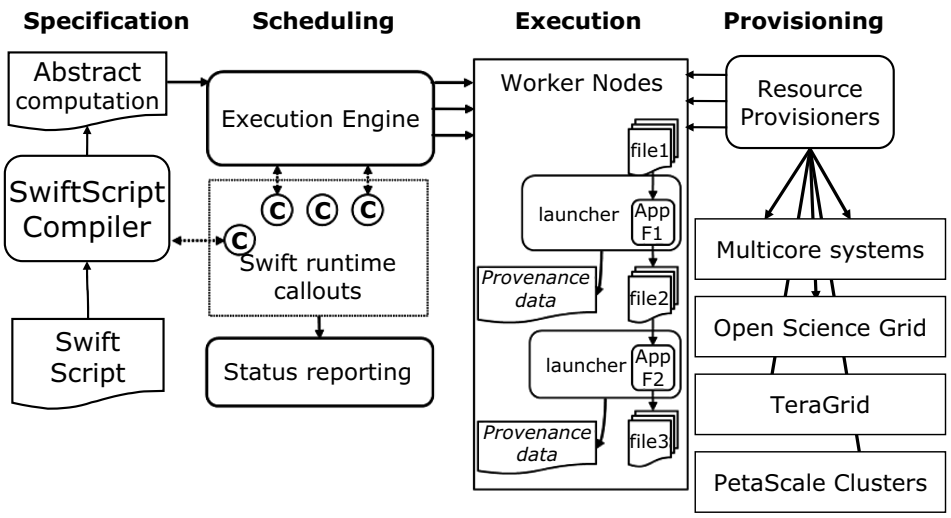


FIGURE 2 | Swift architecture: Managing workflow execution within CNARI. Specification and scheduling are implemented on the client side while execution is implemented on the remote computing resources.

on a local workstation. If the user's workstation has multiple cores, Swift can take advantage of those for modest but invaluable parallelism. And as the user's needs grow or the user is ready to scale up to increasingly large systems, Swift can readily expand to those systems with a single representation and a single client as we will show in our example workflows.

Swift scripts afford a highly productive way to produce and manage the software of neuroscience research units, whether they be local campus departments or international collaborations. In today's practice, organizations that need to process data from fMRI experiments typically develop and rely on locally produced sets of *ad hoc* scripts, usually written in a Linux "shell" language such as "c shell" (csh) or bash, or perhaps Perl, to organize the processing protocols and processes of the collaborations. In Swift, however, as all procedures are "typed" with a specific "signature" of data types for the input and output arguments, a more rigorous and less error-prone paradigm is imposed on the overall structure of the scripts. Thus Swift procedures serve as an interface-definition language for ordinary shell procedures. The overall higher-level process is then defined in a multilevel fashion, from top (highest) to bottom (lowest level) being:

- overall application (such as `multiNetworkSEM`)
- high-level scripts (such as `getCovariance()`)
- low level Swift interfaces (atomic procedures) such as `mxModelProcessor()`
- an external R wrapper script to do further argument manipulation (RInvoke.sh)
- the R tool itself (R CMD BATCH)

Special power and structure is afforded when the tool being run is not a "canned" compiled application, but rather itself a powerful data manipulation environment such as Perl or Python, or more specific to the model we describe in detail here, the R data analysis language with its vast package library of statistical and analytical procedures, including the OpenMx package used here. In this case, the actual script to be performed can be dynamically generated or selected from a template library, and sent to any computing site, which already has a suitable version of R and the OpenMx package installed.

DESCRIPTION OF THE fMRI EXPERIMENT DATA: THE EMBLEM DATABASE

We now give some concrete examples of how Swift can manipulate large datasets and enable novel analysis techniques by means of effective workflow management. The example framework we have employed our grid-enabled analysis techniques on is an fMRI investigation of the neural processing associated with emblematic gesture observation. Emblematic gestures ("emblems") are goal-directed, symbolic manual actions that, while expressed as culturally recognizable manual gestures, communicate a linguistically associable propositional meaning. Four experimental conditions were presented to participants in the MRI scanner: 1) *Emblem*, the symbolic manual gestures; 2) *Speech*, the spoken form of the linguistic propositions associated with the emblems; 3) *Emblem with Speech*, simultaneous presentations of the emblems with their verbalized linguistic associations; and 4) *Grasping*, observation of another type of goal-directed manual action, for which the neural regions associated with its processing have been well-characterized. Data were processed with AFNI (Cox, 1996) and mean normalized values of each of the hemodynamic response functions for every condition at every voxel in the brain were projected to 2-D cortical surface representations and spatially smoothed on the surfaces using SUMA (Saad et al., 2004). These surface values were then imported into MySQL database tables for relational indexing and further analyses.

SEM WORKFLOWS IN SWIFT

We have begun exploring extremely large, exhaustive SEM workflows as a means of investigating how efficient workflow tools can address computational problems that were previously considered unmanageable. Particularly, in using SEM for looking at functional connectivity many researchers are confined to hypothesis-driven approaches because they lack the tools to reliably implement data-driven methods; this situation can greatly impact mining and interpretation of datasets. In an attempt to address these issues, we are building an infrastructure that can be used by researchers to iterate over various parameters within these large sets in a reasonable amount of time and in a manner that is both dynamic and reliable. The following workflows were run on a TeraGrid HPC system known as Ranger. Ranger comprises 3,936 16-way SMP compute nodes providing 15,744 AMD Opteron™ processors for a total of 62,976 compute cores. The workflows were developed on and submitted (to Ranger) from a single-core Linux workstation running an Intel® Xeon™ 3.20 GHz CPU.

A model generator was developed for the OpenMx package and is designed explicitly to enable parallel execution of exhaustive or partially pruned sets of model objects. Given an n x n covariance matrix, it can generate the entire set of possible models with anywhere from 0 to n^2 connections; however, it can also take as input a single index from that set and it will generate and run a single model. What this means in the context of workflow design is that the generator can be controlled (and parallelized) easily by a Swift script. For example, using Swift as the interface to OpenMx we have these few lines of code:

WORKFLOW 1: 4-REGION EXHAUSTIVE SEM FOR A SINGLE EXPERIMENTAL CONDITION

```
1. app (mxModel min) mxModelProcessor(file
   covMatrix, Rscript mxModProc, int modnum,
   float initweight, string cond){
2. {
3.      RInvoke @filename(mxModProc) @
   filename(covMatrix) modnum initweight cond;
4. }
5. file covMatrix<single_file_
   mapper;file="speech.cov">;
6. Rscript mxScript<single_file_mapper;file="sin-
   glemodels.R">;
7. int totalperms[] = [1:65536];
8. float initweight =.5;
9. foreach perm in totalperms{
10.    mxModel modmin<single_file_mapper; file=@
   strcat(perm,".rdata")>;
```

```
11.    modmin = mxModelProcessor(covMatrix,
       mxScript, perm, initweight, "speech");
12. }
```

First, a covariance matrix containing activation data for 4 brain regions, over 8 time points, averaged over a group of subjects in the *Speech* condition was drawn from the experiment database and its location (in this example, on the local file system, though the file could be located anywhere) is mapped in line 5. Line 6 maps the R processing script and lines 1 through 4 define the atomic procedure for invoking R. Each iteration of the foreach loop maps its optimized model output file and calls `mxModelProcessor()` with the necessary parameters to generate and run a model. Each of these invocations of `mxModelProcessor()` is independent and is submitted for processing in parallel. Swift passes 5 variables for each invocation: (1) the covariance matrix; (2) the R script containing the call to OpenMx; (3) the permutation number, i.e., the index of the model; (4) the initialization weight for the free parameters of the given model; and (5) the experimental condition. Clearly, in this workflow all free parameters of the given model will have the same initialization weight as Swift is passing only one weight variable. When the job reaches a worker node on Ranger an R process is initialized, the generator creates the desired model by calculating where in the array that permutation of the model matrix falls. OpenMx then estimates the model parameters using a non-linear optimization algorithm called NPSOL (Gill et al., 1986), and the optimized model is returned and written out by Swift to the location specified in its mapping on line 10.

The above script completed in approximately 40 minutes. The script can then be altered to run over multiple experimental conditions by adding another outer loop:

WORKFLOW 2: 4-REGION EXHAUSTIVE SEM FOR 2 EXPERIMENTAL CONDITIONS

```
1. string conditions[] = ["emblem", "speech"];
2. int totalperms[] = [1:65536];
3. float initweight =.5;
4. foreach cond in conditions{
5.    foreach perm in totalperms{
6.     file covMatrix<single_file_mapper;file=@
       strcat(cond,".cov")>;
7.     mxModel modmin<single_file_mapper;file=@
       strcat(cond,perm,".rdata")>;
8.     modmin= mxModelProcessor(covMatrix,
       mxScript,perm, initweight, cond);
9. }
```

When the outer loop is added, the new workflow consists of 131,072 jobs since we are now running the entire set for two conditions. This workflow completed in approximately 2 hours (**Figure 3**).

WORKFLOW 3: 4-REGION EXHAUSTIVE SEM FOR MULTIPLE NETWORKS

In this workflow multiple 4-region networks are run for the *Emblem with Speech* experimental condition. The regions of interest (ROIs) designated are from FreeSurfer's[2] automatic parcellation of anatomical regions, based on the Duvernoy atlas (1991), and further manual subdivisions to delineate anterior and posterior extents of the superior temporal gyrus and sulcus, as well as superior and inferior segments of the precentral gyrus. Because *Emblem with Speech* involved subjects' perceiving simultaneously both spoken (audiovisual) and manual information, here we chose candidate regions expected to be involved in audiovisual recognition of speech and manual action: occipital pole (OP), middle occipital gyrus (MOG), anterior occipital sulcus (AOS), posterior superior temporal sulcus (STSp), posterior superior temporal gyrus (STGp), transverse temporal gyrus (TTG), and supramarginal gyrus (SMG). Covariance matrices of activation data for *Emblem with Speech* for several networks comprised of these ROIs were then queried from the database:

network 1: {OP, STGp, TTG, AOS}
network 2: {OP, MOG, AOS, STSp}
network 3: {TTG, STGp, SMG, STSp}

```
1. string conditions[] = ["emblemwithspeech"];
```

[2]http://surfer.nmr.mgh.harvard.edu/

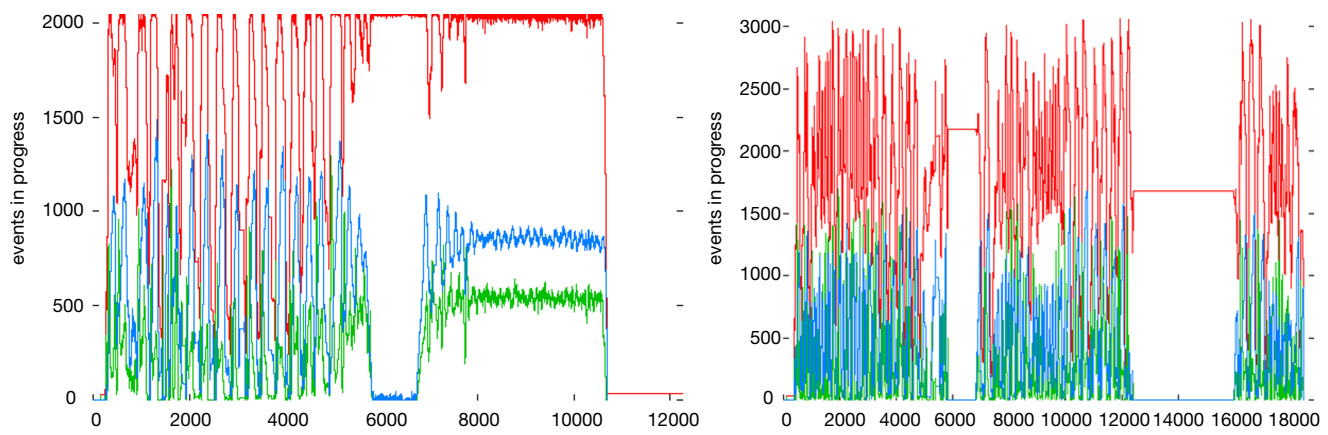


FIGURE 3 | Number of active processes during workflow execution: (left) Processing of the 4-region workflow over 2 experimental conditions**. (Right)** Processing of the 4-region workflow over multiple networks. The red line represents the execution of jobs on Ranger, while the blue and green represent the staging in and out of files respectively. Plots were generated by swift-plot-log, part of the Swift suite of tools.

```
2. int networks[] = [1:3];
3. int totalperms[] = [1:65536];
4. float initweight =.5;
5. foreach cond in conditions {
6.   foreach perm in totalperms {
7.    foreach n in networks {
8.     file covMatrix<single_file_mapper; file=@
       strcat("matrices/net",n,"_",cond,".
       cov")>;
9.     mxModel modmin<single_file_mapper; file=@
       strcat(n,"_",cond,"_",perm,".rdata")>;
10.    modmin = mxModelProcessor(covMatrix,mxScr-
       ipt,perm,initweight,
11.                              condition,@
                                strcat("net",n));
12.    }
13.  }
14. }
```

This results in a workflow containing 196,608 processing jobs (1 condition x 3 networks x 65536 models) and completed in approximately 5 hours on Ranger. For an example of how this might be used as part of a larger processing workflow see Section "Language Study Workflow in Swift" in the Appendix.

DISCUSSION AND FUTURE WORK

The workflows presented here do not result in a single "best" model representing connectivity amongst the four brain regions for the given conditions. Rather, their value lies in that they produce an exhaustive set of optimized models from which to begin searching for good-fitting models. Thus, a natural extension to this set of workflows might be a model-selection component based on a fit statistic (e.g., Bayesian information criterion, Akaike information criterion, RMSEA), an exploratory visualization component (see "Language Study Workflow in Swift" in the Appendix) or perhaps a combination of these methods. A "model-selection workflow" based on one or more fit statistics extending, for example, *workflow 1* would extract the desired fit statistic from each of the 65,536 optimized models and potentially keep or discard a given model based on whether or not it is above or below a selected threshold. It is worth noting that there is a good deal of controversy around which measures provide the most accurate model-selection (Bullmore et al., 2000) as well as some variation in how SEM software packages actually calculate those fit statistics (Clayton and Pett, 2008).

While the present workflows suggest new possibilities for exhaustive search and large-scale, parallel analysis techniques, their utility lies heavily in the ability to be easily replicated and reconfigured for use on varying datasets. Exhaustive search through a space of structural equation models is, *ab initio*, an exploratory technique. Thus, one cannot make statements concerning the probability that there are significant differences between models or that a selected parameter is significantly different from zero. The number of tested models is so great that any statistical argument concerning the likelihood of the data given a null hypothesis is overwhelmed by the number of comparisons made. In addition, one must be concerned about generalizability of results if a single data set was used—the exhaustive search may have overfit idiosyncrasies of the target data. Thus, it is imperative to cross-validate results from exhaustive search using other data sets.

On the other hand, an exhaustive search of the space of structural equation models for a particular data set does result in an empirical distribution of the fit statistics of the models. By plotting the log likelihood resulting from each fit against the number of degrees of freedom in its associated model, it is likely that clusters in the fit statistics will be observed. In this way, we may observe patterns of candidate models that are roughly equivalent given the data. Some of these models may be algebraically equivalent (vonOertzen, in press), and others may be empirically equivalent given the data. We intend the CNARI development effort to enable this type of data exploration.

Beginning with some basic pruning techniques, we can start to narrow down the space of models in the exhaustive set while leveraging Swift's ability to submit large numbers of processes, resulting in some powerful workflows. The first reduction in the exhaustive set of models is elimination of any models that are unidentified, that is, models containing negative degrees of freedom due to the presence of more unconstrained than constrained variables. The degrees of freedom can be easily calculated using the following formula:

```
(n(n+1)/2)-k
```

where n is the number of brain regions in the model and k is the number of free parameters and if the result is negative, the model is underidentified (Bollen, 1989). Additionally, a model with two-way symmetric connections is likely to fail attempts at optimization. Such a connection represents a type of cycle. In fact, most models containing cycles will be difficult to optimize as they are not usually identified in the absence of, e.g., longitudinal data (Neale and Cardon, 1992; Heath, 1993; Neale et al., 1994). The size if the acyclic set is given by

$$4^{((n*(n-1)/2)}.$$

An algorithm exists for finding cycles (Boker et al., 2002) that could potentially be used to further prune the model set. In addition to pruning cyclic and underidentified models, the set may also be pruned for models containing variables that lack residual error. The fit function cannot be evaluated under these circumstances, because the predicted covariance matrix is singular; therefore its determinant is zero, which results in the division of a negative quantity by zero in the calculation of the multivariate normal distribution probability density function, so optimization cannot be performed.

As **Table 1** shows, with a moderate degree of pruning, the set for four regions becomes trivial to run in the present infrastructure. Furthermore, the five-region set, while still a large number of processing jobs, becomes much more manageable.

Table 1 | Number of models produced for exhaustive and partially pruned workflows.

Regions	Exhaustive set	Identified	Acyclic
4	65,536	50,642	4,096
5	33,554,431	26,434,915	1,048,576
6	68,719,476,736	54,802,674,727	1,073,741,824

CNARI has been developed with the aim of managing a broad range of diverse neuroscience datasets and performing efficient, reliable parallel analysis workflows on them. Here we have demonstrated workflows that fully exercise this capability by applying this framework to large computational problems; namely, exhaustive search SEM. The need for data-driven techniques in modeling connectivity has emerged not only in our own work in studying language and aphasia but in SEM in general (Bullmore et al., 2000; Marrelec et al., 2007), though there has been little discussion of workflow management and parallel computing as means of addressing this need. Researchers, faced with seemingly insurmountable computational problems when selecting appropriate models to test, are often forced to rely on less-than-satisfactory approximations not only due to the sheer amount of processing power required but because of the daunting task of distributing those processing tasks in a cohesive manner such that the results are useful and replicable. As CNARI continues to evolve, we hope to expand these large-scale, data-driven workflows as we use them to address the complex research questions facing us.

ACKNOWLEDGMENTS

This research is supported in part by NSF grant OCI-721939, NIH grants DC08638, DA024304-02, DA-18673, 1R21DA024304—01, R21/R33 DC008638 and R01 DC07488, the U.S. Dept. of Energy under Contract DE-AC02-06CH11357, the James S. McDonnell Foundation and the TeraGrid HPC resources of the Texas Advanced Computing Center. The authors thank Ben Clifford and Mihael Hategan for creating and supporting the Swift parallel scripting system, and Michael Spiegal, Jeffrey Spies, and Tim Brick for developing and supporting the OpenMx system.

APPENDIX

LANGUAGE STUDY WORKFLOW IN SWIFT

The following is a prototype using Swift and demonstrating how the above modules can be assembled into a larger exploratory workflow. Exhaustive search is run for the *Emblem with Speech* condition on several four-region networks, and the results of the optimized models are stored in a connectivity database for visualization, further analysis, and pattern detection.

For each of the selected networks `multiNetworkSEM` is called with configuration files for the user to access the databases, information on the network to be processed, and the total number of models in the exhaustive set. First, the covariance data is pulled from the experiment database. This is seen in the `runQuery` function, which is Swift's call to a python database interface (see Small et al., 2009 for a more detailed description of this mediator component). Then for each iteration of the loop in line 34, Swift invokes `mxModelProcessor`, assigning each process a model to generate and optimize in OpenMx. The instantiation of the OpenMx model object and the call to the optimizer are encapsulated in the R script mapped on line 33, which is also passed to mxModelProcessor. Each of these processes writes out a file containing the result of the optimization, and these results can be read and inserted into the connectivity database, which is done with `insertOptMod`. It should be noted that both `insertOptMod` and `getCovariance` operate on the same principle: the user assembles a query that the python DBI will submit to the database. If the user also passes an R script (as in line 62), it will process the query result with that R script. Each result file is read, and its contents are inserted into the connectivity database where they can be further analyzed. A call to `plotLogLik` can be used to plot of the minimum values obtained by OpenMx for each model allowing for identification of patterns or clusters within the set (**Figure 4**).

```
#### MultiNetworkSEM.swift

1. type file;
2. type mxMin;
3. type Rscript;
4. type dbConnect;
5. type mxModel{
6. int modnum;
7. int dof;
8. string best;
9. }
10.# ----------- atomic procedures ----------- #
11.
12.app (file matrix) runQuery (dbConnect dbconn,
   string query, Rscript calcCov){
13.     }
14.     mysqlPythonDBI query @calcCov @dbconn;
15.    }
16.
17.app (external inserted) insertMxResult
   (dbConnect dbconn, string query, file
    datafile)
18.    {
19.     mysqlPythonDBI query @dbconn stdout=@
        filename(inserted) @datafile;
20.    }
21. app (file min) mxModelProcessor ( file
    cov, Rscript mxModProc, int modnum, float
    weight, string cond, int net)
22.      {
23.     RInvoke @mxModProc @filename(cov) modnum
        weight cond net;
24.    }
25.
26. # ------ user-defined SEM procedures ------ #
27.
28. multiNetworkSEM(string condition,dbConnect
    emblemdb, dbConnect semdb, int n, string net,
    int totalperms[])
29. {
30.  float initweight =.75;
31.  file covariance<single_file_mapper;file=@
     strcat("net",n,"/",condition,".cov")>;
32.  covariance = getCovariance(condition, n,
     net, emblemdb);
33.  Rscript mxModProc<single_file_
     mapper;file="scripts/singlemodels.R">;
34.  foreach perm in totalperms{
35.    file modmin<single_file_mapper;file=@
       strcat("net",n,"/",condition,"_",perm,".
       stat")>;
```

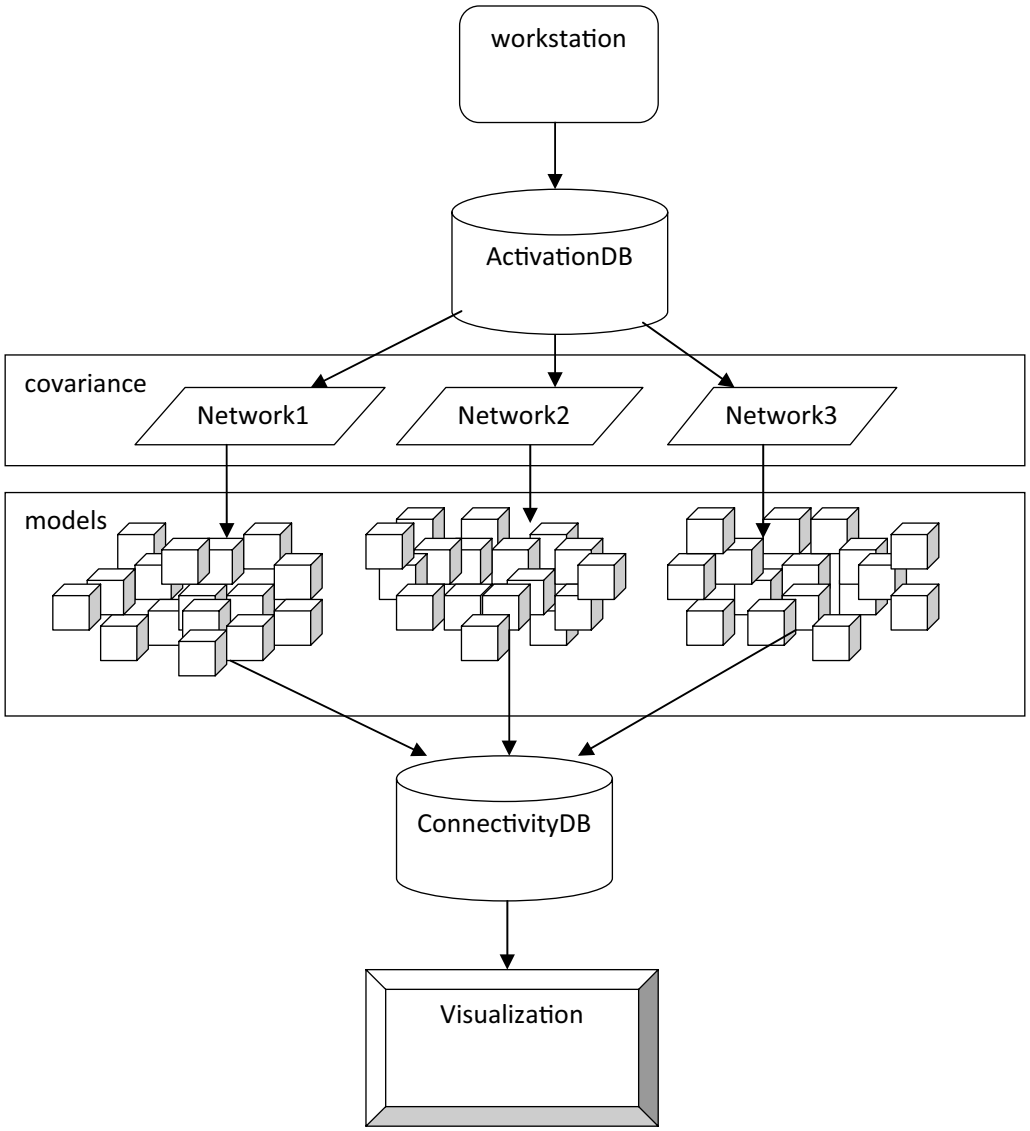


FIGURE 4 | Multinetwork Swift workflow for the *Emblem with Speech* condition.

```
36.    modmin = mxModelProcessor(covariance,mxMod
       Proc,perm,initweight,condition,n);
37.    external doneflag = insertOptMod(n, semdb,
       condition, modmin);
38. }
39.
40. (external ins) insertOptMod(int net,
    dbConnect dbconn, string cond, file modfile)
41. {
42.  string mysqlstr = @strcat("INSERT
     INTO optimized_models (network, deg_of_
     freedom, mx_minimum, modnum, cond) VALUES
     (",net,",DOF,BEST,MODNUM,",cond,");");
43.  string argList = @strcat(
44.  " --query ", mysqlstr,
45.  " --data ", @filename(modfile),
46.  " --conf ", @filename(dbconn));
47.  ins = insertMxResult(dbconn, argList,
     modfile);
48. }
49.
50. (file covariance) getCovariance (string cond,
    int net, string rois, dbConnect dbconn)
51. {
52.  string mysqlstr = @strcat("SELECT
     avg(",cond,"0B), avg(",cond,"1B),
     avg(",cond,"2B),",
53. "avg(",cond,"3B), avg(",cond,"4B),
     avg(",cond,"5B),",
54. "avg(",cond,"6B), avg(",cond,"7B),
     avg(",cond,"8B) ",
55. "FROM emblemfemlh where roi in (",rois,")
```

```
     group by roi ");
56. string argList = @strcat
57. " --conf ", "user.config",
58. " --query ", mysqlstr,
59. " --r_script ", "scripts/cov.R",
60. " --r_swift_args ", "matrices/net",net, "/",
    cond);
61. Rscript calcCov<single_file_
    mapper;file="scripts/cov.R">;
62. (covariance = runQuery(dbconn, argList,
    calcCov);
63. }
64. {
65. (file plotfile) plotLogLik(int net, string
    cond, dbConnect dbconn)
66.
67.  Rscript rplot<single_file_
     mapper;file="scripts/plotloglik.R">;
68.  string mysqlstr = @strcat("SELECT deg_of_
     freedom,mx_minimum FROM optimized_models",
69.  " where network = ",net,";");
70.  string argList = @strcat(
71.  " --conf ", @filename(dbconn),
72.  " --query ", mysqlstr,
73.  " --r_script ", "scripts/plotloglik.R",
74.  " --r_swift_args ", @filename(plotfile));
75.  plotfile = runQuery(dbconn, argList, rplot);
76.
77.  # ---------------- Main ----------------- #
78.
79. string condition = "emblemwithspeech";
80. string networks[] = ["42, 34, 33, 60", "42,
    15, 60, 80", "33, 34, 23, 80"];
81. dbConnect emblemdb <single_file_mapper;
    file="./user.config">;
82. dbConnect semdb <single_file_mapper; file="./user2.
    config">;
83. int totalperms[] = [1:65536];
84. foreach net,n in networks{
85.  multiNetworkSEM(condition,emblemdb,semdb,n,net,
     totalperms);
86.  }
```

REFERENCES

Ban, T., Naito, J., and Kawamura, K. (1984). Commissural afferents to the cortex surrounding the posterior part of the superior temporal sulcus in the monkey. *Neurosci. Lett.,* 49, 57–61.

Ban, T., Shiwa T., and Kawamura, K. (2000). Cortico-cortical projections from the prefrontal cortex to the superior temporal sulcal area (STs) in the monkey studied by means of HRP method. *Arch. Ital. Biol.*, 129, 259–272.

Barbas, H. (2000). Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. *Brain Res. Bull.* 52, 319–330.

Boker, S. M., and McArdle, J. J. (2005). Path analysis and path diagrams. In Encyclopedia of Statistics in Behavioral Science Vol. 3, B. Everitt and D. Howell, eds (New York, John Wiley & Sons), pp. 1529–1531.

Boker, S. M., McArdle, J. J., and Neale, M. (2002). An algorithm for the hierarchical organization of path diagrams and calculation of components of expected covariance. *Struct. Equ. Modeling,* 9, 174–194.

Bollen, K. A. (1989). Structural Equations with Latent Variables. New York, John Wiley & Sons.

Buchel, C., and Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cereb. Cortex* 7, 768–778.

Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S., and Sharma, T. (2000). How good is good enough in path analysis of fMRI data? *NeuroImage,* 11, 289–301.

Catlett, C. et al. (2007). TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications, HPC and Grids in Action, L. Grandinetti, ed (Amsterdam, IOS Press, Advances in Parallel Computing Series).

Clayton, M. F., and Pett, M. A. (2008). AMOS versus LISREL: One data set, two analyses. *Nursing Res.,* 57, 283–292.

Cox, R. W. (1996). AFNI: software for analysis, and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173.

Duvernoy, H. M. (1991). The Human Brain: Surface, Three-dimensional Sectional Anatomy, and MRI. New York, Springer-Verlag.

Gill, P. E., Murray, W., Saunders, M. A., and Wright, M. H. (1986). User's Guide for NPSOL (Version 4.0): A FORTRAN package for nonlinear programming. Department of Operations Research, Stanford University, Stanford.

Hackett T. A., Stepniewska I., and Kaas J. H. (1999) Callosal connections of the parabelt auditory cortex in macaque monkeys. *Eur. J. Neurosci.,* 11, 856–866.

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

Horwitz, B., Tagamets, M. A., and McIntosh, A. R. (1999). Neural modeling, functional brain imaging, and cognition. *Trends Cogn. Sci. (Regul. Ed.)* 3, 91–98.

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Stat.,* 5, 299–314.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika,* 32, 443–482.

Loehlin, J. (1992). Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis. Hillsdale, NJ: Lawrence Erlbaum Associates.

Marrelec, G., Horwitz, B., Kim, J., Pelegrini-Issac, M., Benali, H., and Doyon, J. (2007). Using partial correlation to enhance structural equation modeling of functional MRI data. *Magn. Reson. Imaging,* 25, 1181–1189.

McArdle, J. J., and Boker, S. M. (1990). Rampath. Hillsdale, NJ: Lawrence Erlbaum.

McArdle, J. J., and McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *Br. J. Math. Stat. Psychol.,* 87, 234–251.

McIntosh A. R., and Gonzalez-Lima, F. (1994) Structural equation modelling, and its application to network analysis in functional brain imaging. *Hum. Brain Mapp.*, 2, 2–22.

Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (2003). Mx: Statistical modeling, 6th edn. Richmond, VA: Department of Psychiatry.

Neale, M. C., Eaves, L. J., and Kendler, K. S. (1994). The Power of the Classical Twin Study to Resolve Variation in Threshold Traits. Vol. 24, Netherlands: Springer.

Neale, M. C., and Cardon, L. R. (1992). Methodology for Genetic Studies of Twins and Families. NATO ASI Series. Vol. 67. Dordrecht, Kluwer Academic Publishers.

Petrides, M., and Pandya D. N. (1984). Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. *J. Comp. Neurol.*, 228, 105–116.

Petrides, M., and Pandya D. N. (1988). Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *J. Comp. Neurol.* 273, 52–66.

Petrides, M., and Pandya, D. N. (1999). Dorsolateral prefrontal cortex: comparative cytoarchitectonic analysis in the human, and the macaque brain, and corticocortical connection patterns. *Eur. J. Neurosci.* 11, 1011–1036.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0, Available at: http://www.R-project.org.

Rizzolatti, G., Luppino, G., and Matelli, M. (1998). The organization of the cortical motor system: new concepts.

Electroencephalogr. Clin. Neurophysiol., 106, 283–296.

Rizzolatti, G., Fogassi, L., and Gallese, V. (1997). Parietal cortex: from sight to action. *Curr. Opin. Neurobiol.,* 7, 562–567.

Rosa, M. G., Soares, J. G., Fiorani, M. Jr., and Gattass, R. (1993). Cortical afferents of visual area MT in the Cebus monkey: possible homologies between New and Old World monkeys. *Vis. Neurosci.,* 10, 827–855.

Saad, Z. S., Reynolds, R. C., Argall, B. D., Japee, S., and Cox, R. W. (2004). SUMA: An interface for surface-based intra- and inter-subject analysis with AFNI. Arlington, VA, IEEE International Symposium on Biomedical Imaging. pp. 1510–1513.

Seltzer, B., and Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. *J. Comp. Neurol.,* 343, 445–463.

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

Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., and Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. *Curr. Biol.* 19, 661–667.

Small, S. L., Wilde, M., Kenny, S., Andric, M., and Hasson, U. (2009). Database-managed Grid-enabled analysis of neuroimaging data: The CNARI framework. *Int. J. Psychophysiol.* 73, 62–72.

Solodkin, A., Hlustik, P., Chen, E. E., and Small, S. L. (2004). Fine modulation in network activation during motor execution and motor imagery. *Cereb. Cortex,* 14, 1246–1255.

Stef-Praun, I., Foster, U., Hasson, M., Hategan, S.L., and Wilde, S. M. (2007). Accelerating medical research using the Swift Workflow System. Paper Presented at the HealthGrid 2007, Geneva.

vonOertzen, T. (in press). Power equivalence in structural equation modeling. *Br. J. Math Stat. Psychol.*

Walsh, R. R., Small, S. L., Chen, E. E., and Solodkin, A. (2008). Network activation during bimanual movements in humans. *Neuroimage,* 43, 540–553.

Wright, S. (1921). Correlation and causation. *J. Agric. Res.,* 20, 557–585.

Zhao, H., Clifford, F., von, L., Nefedova, R., and Stef-Praun, W. (2007). Swift: Fast, Reliable, Loosely Coupled Parallel Computation. IEEE Congress on Services, pp. 199–206.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 11 April 2009; paper pending published: 10 July 2009; accepted: 09 September 2009; published online: 20 October 2009.

Citation: Kenny S, Andric M, Boker SM, Neale MC, Wilde M and Small SL (2009) Parallel workflows for data-driven structural equation modeling in functional neuroimaging. Front. Neuroinform. ***3****:34. doi: 10.3389/neuro.11.034.2009*

frontiers in
NEUROINFORMATICS

REVIEW ARTICLE
published: 30 September 2009
doi: 10.3389/neuro.11.033.2009

Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies

Sergi G. Costafreda*

Biomedical Research Center Nucleus and Department of Psychiatry, Institute of Psychiatry, King's College, London, UK

Edited by:
John Van Horn, University of California, USA

Reviewed by:
Craig M. Bennet, University of California, USA
Russell A. Poldrack, University of California, USA

****Correspondence:***
Sergi G. Costafreda, Biomedical Research Center Nucleus and Department of Psychiatry, Institute of Psychiatry, King's College London, De Crespigny Park, PO 89, London SE5 8AF, UK.
e-mail: sergi.costafreda@kcl.ac.uk

The quantitative analysis of pooled data from related functional magnetic resonance imaging (fMRI) experiments has the potential to significantly accelerate progress in brain mapping. Such data-pooling can be achieved through meta-analysis (the pooled analysis of published results), mega-analysis (the pooled analysis of raw data) or multi-site studies, which can be seen as designed mega-analyses. Current limitations in function-location brain mapping and how data-pooling can be used to remediate them are reviewed, with particular attention to power aggregation and mitigation of false positive results. Some recently developed analysis tools for meta- and mega-analysis are also presented, and recommendations for the conduct of valid fMRI data pooling are formulated.

Keywords: fMRI, meta-analysis, mega-analysis, multi-center studies, power, false positive results, random effects analysis, study design

INTRODUCTION

A goal of brain mapping in healthy subjects is to associate mental functions with specific brain locations. In its clinical application, brain mapping aims at identifying the location of brain activation differences between persons suffering from a given neurological or psychiatric disorder and healthy controls during the performance of a cognitive task. Functional magnetic resonance imaging (fMRI) has become the main tool in the brain mapping field as, relative to other techniques, it is non-invasive, has increased spatial resolution, wider availability and lower cost (Pekar, 2006). Conversely, brain mapping studies represent well over half of the fMRI literature to date (Logothetis, 2008).

It has been recognized that data pooling across individual studies has the potential to significantly accelerate progress in the brain mapping field (Van Horn et al., 2004), following other successful data-sharing initiatives, such as The Human Genome Project (Collins and Mansoura, 2001). The most immediate advantage of data pooling is an increase in power due to the larger number of subjects available for analysis. Data pooling across scanning centers can also lead to a more heterogeneous and potentially representative participant sample. Finally, the study of the causes of variability across related experiments may also lead to novel scientific insights (Matthews et al., 2006; Costafreda et al., 2008).

Meta-analysis techniques based on published coordinates of activation have been used since early on to summarize research data and generate novel insights (Fox et al., 1998). Mega-analysis, defined as the pooling of the fMRI time-series, has been less successful so far in spite of its much greater potential, probably due to the difficulty in databasing and making publicly available these "raw" data, and a lack of specific analysis methods that recognize the additional heterogeneity introduced by different scanning centers. Such difficulties may be easing as the field evolves towards multi-site studies (Schumann, 2007), which can be seen as designed mega-analyses, and the necessary databasing, data-sharing and analysis tools are emerging (Keator et al., 2008, 2009; Bockholt et al., 2009).

In the following, current limitations in function-location brain mapping are examined, along with strategies for their remediation through data pooling. Following the meta/mega-analysis distinction frequently employed in the field, the advantages and shortcomings of different types of data-sharing based on the type of data used as prime matter for pooling are also discussed. Finally, the different steps for a valid data pooling exercise, from data collection to the selection of suitable analysis methods, are considered.

LIMITATIONS IN BRAIN MAPPING AND DATA-POOLING REMEDIES

ERRONEOUS RESULTS IN SINGLE-STUDY fMRI ANALYSIS

The aim of conventional group analysis of fMRI data is to detect the regions that show significant increases in BOLD signal in response to a given task. For explanatory purposes, a comparison between an active task and a baseline condition will be assumed, although the following reasoning can be easily extended to more complex designs. Localizing significant changes is often done through voxelwise hypothesis testing, where a null (H_0) and an alternative (H_1) hypothesis are compared. The null hypothesis states that there is no difference in mean signal across subjects between the active and the baseline tasks, while the alternative hypothesis states that such difference exists. The decision as to whether or not H_0 should be rejected in favor of H_1 is then made on the basis of the value of a suitable test (e.g. *t*-test). **Table 1** presents the possible decision outcomes.

False positive results

This mapping strategy is liable to false positive (FP) findings, if H_0 is rejected when it is in fact correct, that is if the area declared to be active was truly not engaged by the active task. The probability α of a

FP result can be kept acceptably low by using multiple comparisons control procedures such as the random field theory (Worsley et al., 1996). In practice, the level of FP results in the literature is likely to be higher than the conventional 5% value of α, as uncorrected results are sometimes reported and sub-optimal fixed-effects group analysis is still occasionally used.

However, under the assumption that FP appear at random brain locations, aggregating results across studies is likely to result in improved brain mapping accuracy in the sense of FP reduction, as a FP finding in a given region is unlikely to be replicated across studies (Fox et al., 1998). In other words, the more studies which have reported that a given area is recruited by a certain paradigm, the less likely it is to be a false positive result. This idea can be formalized: if an observed level of replication in a given location across studies is greater than what would be expected by chance alone, then the null hypothesis of a FP result can be rejected. Recent years have seen the development of several voxel-based meta-analysis methods (Chein et al., 2002; Turkeltaub et al., 2002; Wager et al., 2004, 2007; Laird et al., 2005a; Neumann et al., 2005; Costafreda et al., 2009a; Eickhoff et al., 2009). The initial breakthrough was provided by the Activation Likelihood Estimate (ALE) method presented by Turkeltaub et al. (2002). ALE is a kernel-based approach currently implemented in BrainMap, an online database of published studies (Laird et al., 2009). In kernel-based methods, individual studies are represented by a pattern of activation peak coordinates, which are smoothed using a spatial kernel function (Silverman, 1986). The smoothed patterns are aggregated to obtain a summary map with voxel-level scores representing the local density of activation peaks. This summary map is then thresholded using simulation (Wager et al., 2004; Laird et al., 2005a) or parametric (Costafreda et al., 2009a) approaches, and the areas that survive the threshold are declared as true positive activations. Voxel-based meta-analysis techniques have liberated the meta-analysis process from simple counting of anatomical labels reported by each study and have increased sensitivity to detect aggregate sub-regional activations. A workflow example for one of such methods, Parametric Voxel-based Meta-analysis (PVM, software available from the author; Costafreda et al., 2009a), is presented in **Figure 1**.

Table 1 | Outcomes of hypothesis testing.

	State of the world	
	H_0	H_1
Decision		
H_0	Correct acceptance	Type II error (β)
H_1	Type I error (α)	Correct rejection

False negative results

Brain mapping also suffers from False Negative (FN) reporting, when a region truly active during the task is not recognized as such. This problem is exacerbated by the low number of subjects and, hence, low power that is common in fMRI research. Using a 3T scanner, Thirion et al. (2007) estimated that at least 20 and preferably 27 or more subjects were needed to obtain reproducible results with a simple sensori-motor task under random-effects assumptions. Although specific to the particular scanner, task and analysis employed by the authors, these findings suggest that many fMRI studies may be underpowered. Additionally, Thirion et al. (2007) also found that high inter-subject variability was the key element producing low reliability of group mapping. Factors which increase inter-subject variability in BOLD response, such as the inclusion of psychiatric or neurological populations, will therefore require larger samples.

Under certain conditions, data pooling may also result in an increase of power to detect brain activations and therefore a decrease in FN results. It is this potential for increased power through the aggregation of sub-significant results that underpins meta-analysis applications in most fields (Whitehead, 2002). This

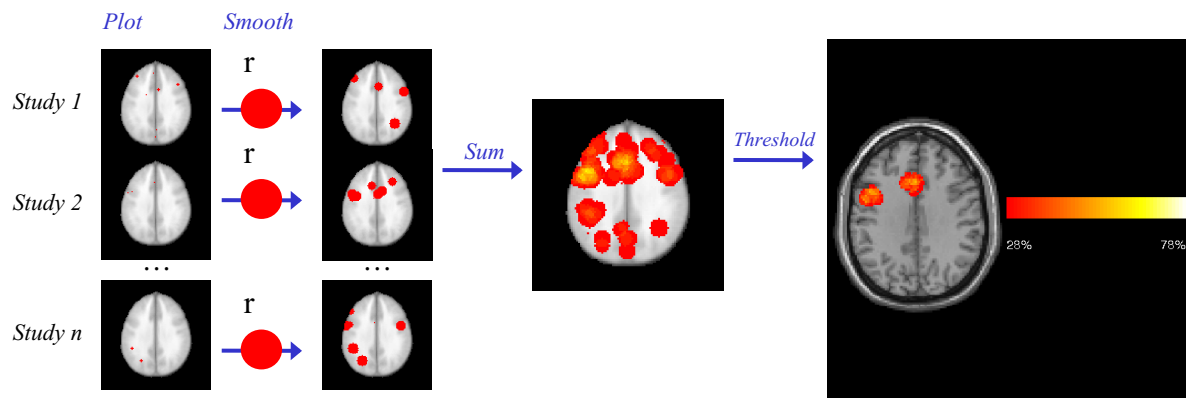


FIGURE 1 | Parametric Voxel-based Meta-analysis. Step 1: the coordinates for each study are plotted in standard space brain (MNI). Step 2: After smoothing with a uniform kernel of size *r*, each study map is transformed into an indicator map, where voxels with 1 values (red) indicate the presence of at least one activation within distance *r*. Step 3: all study-level indicator maps are summed and then divided by the number of studies *n*, to obtain a summary map reflecting the proportion of studies reporting an activation within distance *r* of each voxel. Step 4: the *p* value of the observed proportion is computed, under the null hypothesis that the activations are generated at random spatial locations. The final thresholded map reflects the areas where the proportion of studies reporting activation is too high to have been generated by such null random process alone. In this example of a meta-analysis of language production in healthy subjects, Broca's area and anterior cingulate are revealed as areas of significant activation (Costafreda et al., 2009a).

type of effect-size meta-analysis is based on study-level estimates of a given scalar effect size (e.g. difference in treatment effects across clinical trials) plus, crucially, the standard error of such estimates. Effect sizes from several studies are then statistically pooled to obtain a summary effect size, which has increased precision over any of the original studies.

An equivalent for fMRI research of this primary data would be the (group-level) effect size image or "beta map" accompanied by its corresponding standard error image. However, fMRI researchers rarely publish the complete statistical images, but instead present a highly compact and refined, but impoverished, representation of the original brain activation maps. Regions of significant brain activation, also known as "blobs", are three-dimensional structures which approximately follow grey matter distribution and its associated complicated topography. As a description of such structures, only a list of three-dimensional coordinates is available in a standard paper, usually the points of maximum activation (most statistically significant voxel) for each blob, or its centroid. Results published in this format also lack a measure of variance (i.e. standard error), which precludes the use of traditional effect-size meta-analytical techniques (Fleiss, 1993).

Kernel-based meta-analysis methods can be seen as an attempt to recover a richer representation by deeming as active not only the point of the activation coordinates, but also some neighboring area (Turkeltaub et al., 2002; Wager et al., 2007; Costafreda et al., 2009a). Non-active areas are simply represented by zero. An unavoidable consequence of this impoverished representation is that subtleties in the three-dimensional spatial distribution of the blobs are lost when studies are pooled. Another result is that because the (non-significant) measurements of non-active areas are also lost and simply coded as zero it is not possible to add non-significant findings across studies to decide whether the pooled outcome does, in fact, reach significance. In other words, meta-analysis of coordinate-based data cannot aggregate power across studies and thus cannot remediate the FN problem. Improvements in power can only be obtained through mega-analysis.

In fact, current meta-analysis techniques for brain mapping can be described, from a statistical point of view, as spatial vote-counting (Hedges and Olkin, 1980), where each study "votes" through its reported peak coordinates on whether a particular location is active or not. Vote-counting is a less than ideal technique for research synthesis in statistical terms (Hedges and Olkin, 1980). In particular for fMRI research, detection of significant activation in a given study is a factor of both activation effect size and power, mainly determined by its sample size. Given that sample size is usually limited in typical fMRI experiments, there is scope for misleading findings when aggregating vote-counting results.

VARIABILITY IN EXPERIMENTAL DESIGN, POWER AND GENERALIZATION

From the previous discussion, it can be seen that the initial appeal of pooling fMRI data is therefore a very practical one: to increase the reliability of findings and the power of the statistical analysis. However, this comes at a price: relative to a single large-scale study, a multi-site (or analogously multi-study) design of a similar scale would suffer from inflated variability in its fMRI measurements. This is because it is rare that independent fMRI experiments can be considered exact replicates of each other. For instance, Matthews et al. (2006) described how a subtle variation in the visual presentation of the cue for a simple hand-tapping task across centers in a multi-center study generated significant between-study variability in visual cortical BOLD responses. Findings such as this one suggest that minor changes in experimental conditions may result in significant differences in brain activation. Examples of experimental characteristics with empirical evidence of an effect on fMRI results include: scanner strength (Friedman et al., 2006), subject sample composition (D'Esposito et al., 2003) and analysis method (Strother et al., 2004). The resulting inflation in variability of the fMRI measurements due to these between-study or between-site factors, even when a standardised protocol across sites is enforced (Zou et al., 2005; Friedman et al., 2006) may reduce the statistical power relative to a large single-site design.

Although optimal from the point of view of maximising statistical power, recruitment and other pragmatic issues have tended to make such large-scale single site studies an exception in neuroimaging. Particularly when elusive clinical samples are necessary, recruitment difficulties may recommend a multi-site design (see for example the Alzheimer's Disease Neuroimaging Initiative, Mueller et al., 2006). Also, for many research questions, a sample of relevant studies already exists, and pooling results across this sample through meta- and mega-analysis techniques will often be a more efficient use of these data than considering the findings of each study in isolation (Salimi-Khorshidi et al., 2009).

Apart from the above practical considerations, the increased variability inherent to multi-site or multi-study design is not necessarily detrimental, and can even present advantages for certain research questions. The main potential benefit is that including participants from different sites may lead to a more representative sample of participants, an important consideration if the results of the analysis are intended to be generalized to the population at large. Additionally, activations that generalize over sites and studies are more likely to be linked to the substantive research question under consideration than to idiosyncrasies in study design. As an illustration, the discovery of the resting state brain network in an early mega-analysis was "(…) particularly compelling because these activity decreases were remarkably consistent across a wide variety of task conditions" (Raichle and Snyder, 2007). Data pooling can then be useful to quantitatively examine the generalization of a finding by pooling the results of related studies performed under different conditions. Finally, the causes of between-study variability may also be of interest in themselves. In Costafreda et al. (2008), we applied a meta-regression approach to a large sample of experiments on emotional processing to identify the study characteristics that predicted amygdala activation. Independent predictors of amygdala activation included the type of emotion depicted in the experimental stimuli (e.g. fear), along with more "methodological" variables such as modality of presentation of the stimulus or scanner strength.

REVERSE INFERENCE

Reverse inference in functional neuroimaging is the deduction of the presence of a particular cognitive process as a component of a task due to the engagement of the region (or set of regions) during the task (Poldrack, 2006). An example of reverse inference is concluding that reward may be present during a particular task on the basis

of observing activation in striatum. Although problematic from a logical point of view, used cautiously reverse inference may be useful to elucidate the component processes for a task, and it is often used by functional neuroimaging practitioners (Poldrack, 2006).

In Costafreda et al. (2008) we reported quantitative estimates of the selectivity of amygdala for different emotions relative to neutral material. For example, we found that the amygdala is four to seven times more likely to be activated by fear than by stimuli of neutral content. This probabilistic estimate may be useful in the interpretation of a particular study finding by quantifying the specificity of the link between an area (or network) and a cognitive process. This estimate also acts as an explicit reminder of the limitations in reverse inference, in that such link is not absolute, but probabilistic and necessarily relative to an alternative state (in this example, a neutral stimulus). Therefore, detecting amygdala activation in a particular experiment cannot lead to the conclusion that the task must have involved a fearful stimulus, but simply that it is more likely that the stimulus was fearful than neutral. Additionally, this single estimate cannot exclude a number of credible alternatives, such as amygdala reactivity to social stimuli *per se* or emotions other than fear.

SPATIAL RESOLUTION AND FUNCTIONAL SEGREGATION

The spatial resolution of fMRI has been estimated as a point spread function with full width at half maximum (FWHM) of 3.5 mm for 1.5 T scanners (Engel et al., 1997) and as low as 2 mm for 7 T scanners (Shmuel et al., 2007). However, inter-subject variability in cytoarchitecture is substantial (Amunts et al., 1999), which significantly reduces the resolution obtainable at group level. In addition, the analysis of fMRI data usually involves Gaussian filtering, with typical filter sizes (FWHM) being in the range of 6–15 mm, thus further limiting the effective resolution obtained in practice.

Spatial resolution is particularly relevant to the study of functional segregation. Functional segregation aims to delineate discrete cortical regions along functional lines. Very fine-grained examinations of functional segregation have been attempted by pooling results from different studies (Picard and Strick, 1996). In Costafreda et al. (2006; **Figure 2**), we developed a quantitative method to determine whether two sets of activation peaks are spatially segregated in their cortical distribution. We applied this method to the analysis of verbal fluency studies demonstrating different distributions for the activation peaks of phonological and semantic studies within Broca's area. The significant difference in mean location identified between both distributions (2–18 mm) was comparable or below the usual resolution of any single study.

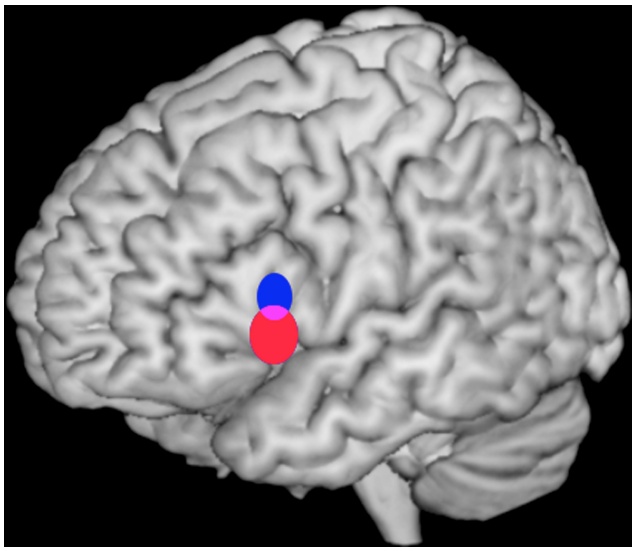


FIGURE 2 | Bootstrap 95% confidence intervals for the mean locations of peak activations in a meta-analysis of phonological and semantic verbal fluency activations in the left inferior frontal gyrus. Updated version of the analysis in Costafreda et al. (2006): the systematic literature search has been updated to September 2008 with a total of 25 studies included, and the bootstrap method has been modified to take into account the clustered nature (activations within studies) of the data. The conclusions are the same as the ones in the published paper. Left lateral view of a rendered image of the brain (MNI template). The confidence intervals (CI) for the mean location of peak BOLD responses associated with semantic verbal fluency (red) were significantly more ventral (*z*-axis) than for those for phonological verbal fluency (blue) at $\alpha = 0.05$. Areas of intersections of the CI (phonological semantic) are shown in mauve.

A TYPOLOGY OF fMRI DATA-POOLING

META-ANALYSIS

Activation coordinates as primary data

Almost all the pooling exercises to date have been meta-analysis, conducted using the coordinates of the location of activations as the primary data. Some of this popularity may be due to the availability of coordinate data, which has become a standard of neuroimaging reporting (Laird et al., 2009). As discussed earlier, its main disadvantage is the impossibility to aggregate power across studies. Therefore most meta-analyses compute estimates of between-study reliability of activations, although many other coordinate-based approaches are possible, such as the examination of between-study co-activation of brain as a proxy of functional connectivity (Toro et al., 2008).

Meta-analysis using additional descriptors

Neuroimaging publications often report both coordinates of peak or maximum activation and their associated anatomical label. Meta-analysis based on labels (Laird et al., 2005b), or a combination of labels and coordinates (Costafreda et al., 2008) is possible, and can even be more powerful than voxel-based meta-analysis when the number of studies is low (<10) as multiple testing is reduced from the number of voxels to the number of regions. However, the variability in anatomical nomenclature in published studies can be a serious limitation. Additionally, voxel-based meta-analysis may be more sensitive if the clustering of activations across studies is not well matched by the chosen anatomical label (Laird et al., 2005b).

Often, in addition to location coordinates, additional measures of the activation characteristics are reported. If the volume of the activated "blobs" was consistently reported, then it could be used for more accurate approximation of the original activations. In our experience though, volume of activation is not consistently reported.

Often the *T* or *Z* statistics of significant activations are also reported. It is possible to employ these quantities to generate effect

size meta-analyses. The difficulty with this approach, however, consists of how to handle non-significant effects, for which no effect size estimate is given: are we to assume these unknown effect sizes are zero, or just below significance, or simply exclude them from the dataset? In our view any of these alternatives leads to further difficulties in the form of potential biases of our results, while the benefit is only an apparent increase in power (apparent because the subsignificant results are unknown).

In conclusion, while acknowledging the serious limitations inherent in coordinate-based data, and short of a decided move towards full voxel-based reporting of significant and non-significant effect sizes discussed below, coordinates are currently the best available substrate for meta-analysis.

MEGA-ANALYSIS

fMRI time-series as primary data

As the raw time-series contains the record of all the measurements obtained during an fMRI experiment, it would seem the obvious prime matter for data pooling: mega-analysis can reduce both false positive and false negative results. However, three practical difficulties have severely limited the application of this approach. First, fMRI measurements from a single study typically generate gigabytes of data. Databasing such large volumes of information and making it publicly available is no trivial technical task (Van Horn et al., 2001; Bockholt et al., 2009). Secondly, fMRI data sharing initiatives have in the past sparked serious objections in the scientific community, which has often proven reticent to share data that are difficult and expensive to acquire (Koslow, 2002). Only a very small fraction of fMRI experiments are nowadays publicly available for download. Finally, there is currently a paucity of quantitative methods that are able to cope with the processing complexity that may arise in fMRI data mega-analysis. These factors create a classical egg and chicken situation: as very limited data are available for download, limited effort is put into developing mega-analysis methods, which in turn further limits the appeal of data-sharing in this format.

This situation, however, is starting to change. Empirical studies have shown low scanner-related variance relative to between-subject variability and measurement error (Costafreda et al., 2007; Suckling et al., 2008) thus encouraging multi-center designs and associated databasing technology (Keator et al., 2008). Methods of analysis are also starting to reflect the need for large-scale integration of results (Pinel et al., 2007; Costafreda et al., 2009b; Dinov et al., 2009; Salimi-Khorshidi et al., 2009), as discussed below.

Statistical maps as intermediate format

The complexity in databasing and publishing time-series data would be reduced if instead statistical brain maps were made publicly available. If effect-size brain maps were accompanied by their standard error images, then usual effect-size meta-analysis methods could be applied (Whitehead, 2002), and power could be aggregated across studies with smaller databasing overheads. Additionally, standard random-effects fMRI analysis techniques could be used validly on such summaries (Salimi-Khorshidi et al., 2009). If subject-level statistical maps, rather than group-level maps, were to be released, this would also allow the examination of the causes of between-subject variability, which has been consistently identified as the main source of heterogeneity in fMRI measurements (Zou et al., 2005; Costafreda et al., 2007; Thirion et al., 2007; Suckling et al., 2008).

In spite of its convenience, it must be stressed that such intermediate data format would also have its disadvantages. Temporal data, and therefore, connectivity information, would be lost in the translation. Relative to time-series pooling, extraneous variability would also be introduced by those statistical maps, as different labs would report maps obtained through varying pre-processing and first-level analysis approaches.

REQUIREMENTS AND ANALYSIS TOOLS FOR VALID fMRI DATA POOLING

SYSTEMATIC SEARCH STRATEGY

The validity of data pooling is crucially dependent on which studies are included. In effect-size meta-analysis, a particularly important problem is publication bias. Also known as the "file-drawer" problem, it originates from the fact that negative studies are less likely to be published, biasing the overall estimate of effect size towards higher values (Sterne and Egger, 2001). Unbiased, exhaustive and *a priori* literature-sampling strategies are necessary to ensure the inclusion of all relevant studies, or at least of a representative sample, of which only clearly flawed or inadequate studies should be excluded. It is worth insisting that these sampling considerations also apply to mega-analysis of fMRI data, as negative studies may be less likely to be represented in publicly available data repositories. In our view, databases containing the results of fMRI experiments (e.g. Brainmap, fMRIDC) (Laird et al., 2009) should be used to complement the systematic literature search bearing in mind the caveat they do not include all potential studies, and the criteria for inclusion in the database are often not explicitly stated, creating room for selection biases. By contrast, in coordinate-based meta-analysis, the focus of the analysis is usually the determination of the location of an effect, which may be less affected by the exclusion of non-significant results (Fox et al., 1998).

STUDY AS A RANDOM EFFECT

Both meta- and mega-analysis require analysis methods adapted to the specificities of pooling data across experimental designs. As discussed earlier, functional MRI experiments are highly heterogeneous in their subjective recruitment strategies, cognitive paradigms, acquisition software and hardware, and analysis methods. Even with standardized protocols and adequate data preprocessing (Zou et al., 2005; Friedman et al., 2006; Costafreda et al., 2007) two fMRI measurements coming from the same center can be expected to be more similar to each other than what would be expected by chance alone, compromising crucial independence assumptions inherent to most analysis methods. Therefore, the existence of multiple sites for data acquisition will in most cases have to be recognized during data analysis as well.

In the analysis of the efficacy of clinical interventions, meta-analysis of (scalar) data from heterogeneous trials is also the rule (Whitehead, 2002). It is often dealt with in a double strategy: (1) by employing study-level covariates that are likely to explain some of the study heterogeneity as fixed-effects in a meta-regression approach, and (2) through the inclusion of a study-level error term capturing residual inter-study variability. This second point is equivalent to treating the study factor as a random effect, in a

similar way as subjects are treated in fMRI group-level estimates (Mumford and Nichols, 2006).

Meta- and mega-analysis of functional imaging data could benefit from a similar approach. The study should therefore be recognized as a further level in the usual fMRI data hierarchy of task runs within subjects within studies (Penny et al., 2003). Most methods currently in use for fMRI meta-analysis, however, consider the foci of activation as the independent observations and ignore the clustering of coordinates in the original studies (Chein et al., 2002; Turkeltaub et al., 2002; Wager et al., 2004; Laird et al., 2005a; Neumann et al., 2005). These approaches are therefore fixed-effects meta-analysis techniques. The results of fixed-effects meta-analysis only apply to the specific sample of experiments under consideration and cannot be generalized to a population of studies if between-study heterogeneity is present. In practical terms, the main undesirable consequence of omitting study-level clustering is that statistically significant density can be obtained with fixed-effects methods simply by the report of several contiguous foci by a single paper, which may have been obtained through overly generous statistical thresholding and thus a marker of poor study quality (Wager et al., 2007). Random-effects alternatives for fMRI meta-analysis have been recently developed using simulation-based (Wager et al., 2007; Eickhoff et al., 2009) and parametric analytical approaches (Costafreda et al., 2009a), and should in our view be preferentially employed.

In particular, PVM (Costafreda et al., 2009a; **Figure 1**) is a statistical method for function-location meta-analysis that allows valid, powerful, fast and scalable detection of the areas with significance concordance between studies for maps expressed in proportions. That is, the statistic computed in this approach is, for each voxel, the proportion of studies that have reported activation within a pre-determined local neighbourhood. Proportions are "natural" random effects estimators, in the sense of taking between-study variability into account. They are also easily interpretable, even when translated into a map. Finally proportions, and ratios between proportions, can be directly used as quantitative estimates of probability, for example as guidance in reverse inference.

Regarding mega-analysis, the existence of study-level clustering effects would need to be recognized through, for example, the introduction of a study level in the analysis hierarchy (e.g. runs within tasks within subjects within group within centers/studies). If the highest-level, "top" summary map is of interest, a random-effects analysis can be obtained through the application of split-level analysis using usual software libraries, such as FSL (Salimi-Khorshidi et al., 2009). Costafreda et al. (2009b) presents a mega-analysis tool that may be useful for more complex designs, especially in the presence of clusters (families, studies) with potentially low degrees of freedom. If covariate estimation is required, then clusters with low counts may present an identifiability problem (if number of parameters ≥ items in cluster). The Bayesian all-in-one approach allows the estimates to "borrow strength" across clusters, thus stabilizing the model fitting process (Bowman et al., 2008).

STUDY DIFFERENCES AS FIXED EFFECTS

As discussed earlier, heterogeneous experimental designs are inevitable in many data pooling situations. Some of this heterogeneity may have direct consequences on the results of the experiments. Known or suspected sources of heterogeneity may be controlled at the study selection step by restriction, for example by only including studies with exclusively right-handed samples in a language meta-analysis. At the analysis step, covariates can be included as fixed effects in a meta-regression strategy (Costafreda et al., 2008). Covariate adjustment is often an attractive option, because the addition of the extraneous factor as a covariate maximizes power both by allowing the inclusion of a larger number of studies than if a restrictive approach had been used, and by removing the variability associated with the covariate factor. Whether a covariate is, in fact, influencing the summary findings can then also be determined, which may be interesting in itself.

Finally, if the covariate is associated with both the outcome under study and the predictor of primary interest, this association may result in confounding, which would lead to biased meta-analytical findings if not taken into account (Greenland et al., 1999; Lawlor et al., 2004). A hypothetical example of confounding would be created if fMRI was a more sensitive technique than PET, and experiments on negative emotions were mostly done with fMRI while those on positive emotions were conducted with PET. Thus, ignoring this potential confounding effect in the analysis would create an apparent increase in the probability of amygdala activation for negative over positive emotions. Two difficulties have to be acknowledged when dealing with confounding. First, potential confounders are not always accurately measured. For example, while functional neuroimaging publications do not always disclose enough methodological detail to ascertain whether fixed or random-effects multisubject analysis was performed, this methodological choice influences the sensitivity and generalizability of the analysis (Friston et al., 1999). Accurate and extensive meta-data collection is thus a pre-requisite for pooled data analysis, which should benefit from recent advances in automated meta-data collection (Bockholt et al., 2009). Second, the number of potential confounding factors that can be effectively introduced in the analysis depends ultimately on the size of the available dataset. A general rule-of-thumb in linear modeling is that one predictor may be included for each 10 independent observations (Harrell, 2001), although newer statistical approaches may be able to remediate this limitation (Fu et al., 2008). If these steps for heterogeneity control are not available, for example due to incomplete information, then the likely impact of potential confounding factors should be addressed when discussing the results (Costafreda et al., 2006).

Crucially, random and fixed-effects strategies are not competing alternatives to deal with between-study heterogeneity. When possible, pertinent covariates can be used in a meta-regression to explain some of the variability or to study the causes for between-study heterogeneity. Additionally, all attempts at fMRI data pooling should include a study-level error even if study factors are already included as fixed effects, because it is unlikely that the measured covariates capture all the between-study variability.

THE VALUE OF fMRI DATA POOLING

Pooling data across sites responds primarily to pragmatic necessities, such as the maximization of sample size, especially in elusive clinical populations. It can also satisfy the need to utilize already existing, but frequently underpowered, neuroimaging studies in

a more efficient way than the consideration of their individual findings. Last but not least, as fMRI research grows exponentially, quantitative synthesis of published fMRI research will remain necessary simply to allow researchers a summary of a mountain of research data. As functional neuroimaging becomes more data-rich, such computational approaches able to extract novel insights from existing large-scale datasets are likely to become increasingly valuable.

ACKNOWLEDGMENTS

Work supported by a MRC Special Training Fellowship in Neuroinformatics and a Wellcome Trust Value in People Award. The author also acknowledges support from the National Institute for Health Research (NIHR) Specialist Biomedical Research Centre for Mental Health award to the South London and Maudsley NHS Foundation Trust and the Institute of Psychiatry, King's College London.

REFERENCES

Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H. B., and Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *J. Comp. Neurol.* 412, 319–341.

Bockholt, H. J., Scully, M., Courtney, W., Rachakonda, S., Scott, A., Caprihan, A., Fries, J., Kalyanam, R., Segall, J., de la Garza, R., Lane, S., and Calhoun, V. D. (2009). Mining the Mind Research Network: A Novel framework for exploring large scale, heterogeneous translational neuroscience research data sources. *Front. Neuroinform.* 3. doi:10.3389/neuro.11.036.2009.

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *Neuroimage* 39, 146–156.

Chein, J. M., Fissell, K., Jacobs, S., and Fiez, J. A. (2002). Functional heterogeneity within Broca's area during verbal working memory. *Physiol. Behav.* 77, 635–639.

Collins, F. S., and Mansoura, M. K. (2001). The human genome project. revealing the shared inheritance of all humankind. *Cancer* 91, 221–225.

Costafreda, S. G., Brammer, M. J., David, A. S., and Fu, C. H. (2008). Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 PET and fMRI studies. *Brain Res. Rev.* 57–70.

Costafreda, S. G., Brammer, M. J., Vencio, R. Z. N., Mourao, M. L., Portela, L. A. P., de Castro, C. C., Giampietro, V. P., and Amaro, E. J. (2007). Multisite fMRI reproducibility of a motor task using identical MR systems. *J. Magn. Reson. Imaging* 26, 1122–1126.

Costafreda, S. G., David, A. S., and Brammer, M. J. (2009a). A parametric approach to voxel-based meta-analysis. *Neuroimage* 46, 115–122.

Costafreda, S. G., Fu, C. H. Y., Picchioni, M., Kane, F., McDonald, C., Prata, D. P., Kalidindi, S., Walshe, M., Curtis, V., Bramon, E., Kravariti, E., Marshall, N., Toulopoulou, T., Barker, G. J., David, A. S., Brammer, M. J., Murray, R. M., and McGuire, P. K. (2009b). Increased inferior frontal activation during word generation: a marker of genetic risk for schizophrenia but not bipolar disorder? *Hum. Brain Mapp.* 30, 3287–3298.

Costafreda, S. G., Fu, C. H. Y., Lee, L., Everitt, B., Brammer, M. J., and David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: role of the left inferior frontal gyrus. *Hum. Brain Mapp.* 27, 799–810.

D'Esposito, M., Deouell, L. Y., and Gazzaley, A. (2003). Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nat. Rev. Neurosci.* 4, 863–872.

Dinov, I., Van Horn, J., Hojatkashani, C., Magsipoc, R., Petrosyan, P., Liu, Z., Lozev, K., Mackenzie-Graham, A., Eggert, P., Stott Parker, D., and Toga, A. (2009). Effcient, distributed and interactive neuroimaging data analysis using the LONI pipeline. *Front. Neuroinform.* 3. doi: 10.3389/neuro.11.022.2009.

Eickhoff, S., Laird, A., Grefkes, C., Wang, L., Zilles, K., Fox, P., West, B., and Julich, G. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30, 2907–2926.

Engel, S. A., Glover, G. H., and Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex* 7, 181–192.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Stat. Methods Med. Res.* 2, 121–145.

Fox, P. T., Parsons, L. M., and Lancaster, J. L. (1998). Beyond the single study: function/location metanalysis in cognitive neuroimaging. *Curr. Opin. Neurobiol.* 8, 178–187.

Friedman, L., Glover, G., Krenz, D., and Magnotta, V. (2006). Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. *Neuroimage* 32, 1656–1668.

Friston, K. J., Holmes, A. P., and Worsley, K. J. (1999). How many subjects constitute a study? *Neuroimage* 10, 1–5.

Fu, C. H., Mourao-Miranda, J., Costafreda, S. G., Khanna, A., Marquand, A. F., Williams, S. C., and Brammer, M. J. (2008). Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. *Biol. Psychiatry* 63, 656–662.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* 10, 37–48.

Harrell, F. E. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, Springer-Verlag.

Hedges, L. V., and Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychol. Bull.* 88, 359–369.

Keator, D., Gadde, S., Bockholt, H. J., Marcus, D., Grethe, J., Ozyurt, B., Wei, D., Aucoin, N., BIRN, (2009). A meta-data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. *Front. Neuroinform. 3. doi: 10.3389/neuro.11.030.2009.*

Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H. J., and Papadopoulos, P. (2008). A national human neuroimaging collaboratory enabled by the biomedical informatics research network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172.

Koslow, S. H. (2002). Opinion: sharing primary data: a threat or asset to discovery? *Nat. Rev. Neurosci.* 3, 311–313.

Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., Turkeltaub, P. E., Kochunov, P., and Fox, P. T. (2005a). ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25, 155–164.

Laird, A. R., McMillan, K. M., Lancaster, J. L., Kochunov, P., Turkeltaub, P. E., Pardo, J. V., and Fox, P. T. (2005b). A comparison of label-based review and ALE meta-analysis in the stroop task. *Hum. Brain Mapp.* 25, 6–21.

Laird, A. R., Lancaster, J. L., and Fox, P. T. (2009). ALE meta-analysis workflows via the brainmap database. *Front. Neuroinform. 3. doi: 10.3389/neuro.11.023.2009.*

Lawlor, D. A., Smith, G. D., Bruckdorfer, K. R., Kundu, D., and Ebrahim, S. (2004). Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet* 363, 1724–1727.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature* 453, 869–878.

Matthews, P. M., Honey, G. D., and Bullmore, E. T. (2006). Applications of fMRI in translational medicine and clinical practice. *Nat. Rev. Neurosci.* 7, 732–744.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., Trojanowski, J. Q., Toga, A. W., and Beckett, L. (2006). The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am.* 15, 869–877.

Mumford, J. A., and Nichols, T. (2006). Modeling and inference of multisubject fMRI data. *IEEE Eng. Med. Biol. Mag.* 25, 42–51.

Neumann, J., Lohmann, G., Derfuss, J., and von Cramon, D. Y. (2005). Meta-analysis of functional imaging data using replicator dynamics. *Neuroimage* 25, 165–173.

Pekar, J. J. (2006). A brief introduction to functional MRI. *IEEE Eng. Med. Biol. Mag.* 25, 24–26.

Penny, W. D., Holmes, A. P., and Friston, K. J. (2003). Hierarchical models. In Human Brain Function II. R. S. Frackowiak, K. J. Friston, K. J. Dolan, J. Ashburner, and W. D. Penny, eds (London, Elsevier Academic), pp. 851–863.

Picard, N., and Strick, P. L. (19966). Motor areas of the medial wall: a review of their location and functional activation. *Cereb. Cortex* 6, 342–353.

Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J.-B., and Dehaene, S. (2007). Fast reproducible identification and large-scale databasing of individual functional cognitive networks. *BMC Neurosci.* 8, 91.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63.

Raichle, M. E., and Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *Neuroimage* 37, 1083–1090.

Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., and Nichols, T. E., (2009). Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. *Neuroimage* 45, 810–823.

Schumann, G. (2007). Okey lecture 2006: identifying the neurobiological mechanisms of addictive behaviour. *Addiction* 102, 1689–1695.

Shmuel, A., Yacoub, E., Chaimow, D., Logothetis, N. K., and Ugurbil, K. (2007). Spatio-temporal point-spread function of fMRI signal in human gray matter at 7 Tesla. *Neuroimage* 35, 539–552.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. London, Chapman and Hall.

Sterne, J. A., and Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J. Clin. Epidemiol.* 54, 1046–1055.

Strother, S., La Conte, S., Kai Hansen, L., Anderson, J., Zhang, J., Pulapura, S., and Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. a preliminary group analysis. *Neuroimage* 23(Suppl. 1), S196–S207.

Suckling, J., Ohlssen, D., Andrew, C., Johnson, G., Williams, S. C. R., Graves, M., Chen, C.-H., Spiegelhalter, D., and Bullmore, E. (2008). Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. *Hum. Brain Mapp.* 29, 1111–1122.

Thirion, B., Pinel, P., Meriaux, S., Roche, A., Dehaene, S., and Poline, J.-B. (2007). Analysis of a large fMRI cohort: statistical and methodological issues for group analyses. *Neuroimage* 35, 105–120.

Toro, R., Fox, P., and Paus, T. (2008). Functional coactivation map of the human brain. *Cereb. Cortex* [Epub ahead of print].

Turkeltaub, P. E., Eden, G. F., Jones, K. M., and Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *Neuroimage* 16, 765–780.

Van Horn, J. D., Grafton, S. T., Rockmore, D., and Gazzaniga, M. S. (2004). Sharing neuroimaging studies of human cognition. *Nat. Neurosci.* 7, 473–481.

Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., Rockmore, D., and Gazzaniga, M. S. (2001). The functional magnetic resonance imaging data center (fMRIDC): the challenges and rewards of large-scale databasing of neuroimaging studies. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 356, 1323–1339.

Wager, T. D., Jonides, J., and Reading, S. (2004). Neuroimaging studies of shifting attention: a meta-analysis. *Neuroimage* 22, 1679–1693.

Wager, T. D., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* 2, 150–158.

Whitehead, A. (2002). Chapter 1. In Meta-Analysis of Controlled Clinical Trials. Statistics in Practice. (Chichester, England, Wiley), pp. 1–10.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., and Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* 4, 58–73.

Zou, K. H., Greve, D. N., Wang, M., Pieper, S. D., Warfield, S. K., White, N. S., Manandhar, S., Brown, G. G., Vangel, M. G., Kikinis, R., and Wells, W. M. (2005). Reproducibility of functional MR imaging: preliminary results of prospective multi-institutional study performed by Biomedical Informatics Research Network. *Radiology* 237, 781–789.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 April 2009; paper pending published: 07 July 2009; accepted: 31 August 2009; published online: 30 September 2009.

Citation: Costafreda SG (2009) Pooling fMRI data: meta-analysis, mega-analysis and multi-center studies. Front. Neuroinform. ***3****:33. doi: 10.3389/neuro.11.033.2009*