NeuroImage xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Review

Q12 Q16 Towards structured sharing of raw and derived neuroimaging data across existing resources

Q1 D.B. Keator [a,b,*], K. Helmer [c,1], J. Steffener [d,2], J.A. Turner [e,3], T.G.M. Van Erp [a,4], S. Gadde [f,5], N. Ashish [g,6], G.A. Burns [h,7], B.N. Nichols [i,j,8]

Q2 [a] *Department of Psychiatry and Human Behavior, University of California, Irvine, USA*
[b] *Department of Computer Science, University of California, Irvine, USA*
[c] *Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Department of Radiology, Harvard Medical School, Boston, MA, USA*
[d] *Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University, USA*
[e] *The Mind Research Network, Albuquerque, NM, USA*
[f] *Brain Imaging and Analysis Center, Duke University, USA*
[g] *California Institute of Telecommunications and Information Technology (CalIT2), University of California, Irvine, USA*
[h] *Information Sciences Institute, University of Southern California, USA*
[i] *Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, USA*
[j] *Integrated Brain Imaging Center, Department of Radiology, University of Washington, Seattle, USA*

A R T I C L E I N F O

Article history:
Accepted 18 May 2013
Available online xxxx

Keywords:
XCEDE
Provenance
Data model
Neuroimaging
Database
Web services

A B S T R A C T

Data sharing efforts increasingly contribute to the acceleration of scientific discovery. Neuroimaging data is accumulating in distributed domain-specific databases and there is currently no integrated access mechanism nor an accepted format for the critically important meta-data that is necessary for making use of the combined, available neuroimaging data. In this manuscript, we present work from the Derived Data Working Group, an open-access group sponsored by the Biomedical Informatics Research Network (BIRN) and the International Neuroimaging Coordinating Facility (INCF) focused on practical tools for distributed access to neuroimaging data. The working group develops models and tools facilitating the structured interchange of neuroimaging meta-data and is making progress towards a unified set of tools for such data and meta-data exchange. We report on the key components required for integrated access to raw and derived neuroimaging data as well as associated meta-data and provenance across neuroimaging resources. The components include (1) a structured terminology that provides semantic context to data, (2) a formal data model for neuroimaging with robust tracking of data provenance, (3) a web service-based application programming interface (API) that provides a consistent mechanism to access and query the data model, and (4) a provenance library that can be used for the extraction of provenance data by image analysts and imaging software developers. We believe that the framework and set of tools outlined in this manuscript have great potential for solving many of the issues the neuroimaging community faces when sharing raw and derived neuroimaging data across the various existing database systems for the purpose of accelerating scientific discovery.

Contents

Q3 * Corresponding author at: University of California, Irvine, Brain Imaging Center, Irvine Hall rm. 163-Zot 3960, Irvine, CA 92617, USA.
E-mail address: dbkeator@uci.edu (D.B. Keator).
Q4 [1] Present address: Massachusetts General Hospital, Athinoula A Martinos Center for Biomedical Imaging, 149-13th St Room 2301, Charlestown, MA 02129, USA.
Q5 [2] Present address: The Taub Institute, Columbia University College of Physicians and Surgeons, 630 West 168th St., P&S 16, New York, NY 10032, USA.
Q6 [3] Present address: Mind Research Network, 1101 Yale Blvd. NE, Albuquerque, NM 87106, USA.
Q7 [4] Present address: University of California, Irvine, Department of Psychiatry and Human Behavior, 5251 California Ave., suite 240, Irvine, CA 92617, USA.
Q8 [5] Present address: Brain Imaging and Analysis Center, Duke University Medical Center, 2424 Erwin Road, Suite 501, Durham, NC 27705, USA.
Q9 [6] Present address: University of California, Irvine, 4308 Calit2 Bldg, Irvine, CA 92617, USA.
Q10 [7] Present address: 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA.
Q11 [8] Present address: Integrated Brain Imaging Center, University of Washington, Box 357240, 1959 NE Pacific Street, Seattle, WA 98195, USA.

1053-8119/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.neuroimage.2013.05.094

Introduction

Acceleration of scientific discovery in neuroimaging and many other research areas increasingly relies on the availability of large and well-documented data sets. In fact, many of the major new discoveries in the genetics of schizophrenia and other psychiatric disorders, multiple sclerosis, diabetes, obesity, and other metabolic traits have been possible only through collaborative data sharing (Ripke et al., 2011; Sawcer et al., 2011; Speliotes et al., 2010). In the area of neuroimaging, such data sets can be obtained by a) funding large consortia to prospectively acquire large data sets (Insel et al., 2004), b) harvesting research-ready data from other sources (Kho et al., 2011; van Erp et al., 2011), and/or c) data (Biswal et al., 2010) or analysis results (Stein et al., 2012), sharing between multiple separately funded initiatives that include in-common measurements. In-common measurements, in the context of neuroimaging, refer to imaging protocols that are included in many magnetic resonance imaging (MRI) related studies such as resting state functional magnetic resonance imaging (fMRI), structural T1-weighted MRI, and diffusion tensor imaging (Nooner et al., 2012). Shared and combined use of in-common measurements is the lowest barrier in the otherwise complex and often intractable space of combining neuroimaging data collected under different initiatives; however, acquiring equivalent data sets at sites with hardware from different vendors requires careful protocol design (Jack et al., 2008, 2010; Kruggel et al., 2010). Despite efforts from consortia such as the Function and Morphometry test beds of the Biomedical Informatics Research Network (BIRN) that have published recommendations for collecting neuroimaging data with the sharing and combining of data from multiple sites in mind, the task of data sharing across scanner platforms remains difficult even though the benefits are both financially and scientifically undeniable (Glover et al., 2012; Poline et al., 2012). Sharing well-documented, often publicly funded, data sets for use by the wider research community can be cost-effective as it allows for 1) increased statistical power through mega-analyses in contrast to meta-analyses, 2) obtaining new larger data sets to answer questions not addressed by the original studies, 3) application of newly developed tools to existing data sets, and 4) replication of research findings via reanalysis of existing data by other research groups.

In the last ten years, large neuroimaging data sets have become publicly available, although, there are significant differences in the requirements for data access. These data sets are in domain-specific repositories. Some examples of completely open-access neuroimaging repositories include XNAT Central (https://central.xnat.org) which includes over 3000 subjects stored in the XNAT database (Marcus et al., 2007), the BIRN data repository (www.birncommunity.org/resources/data) which includes large cohorts of both mouse and human imaging data stored in the BIRN Human Imaging Database (Florescu et al., 1996; Ozyurt et al., 2010) and elsewhere, the 1000 Functional Connectomes project (www.nitrc.org/projects/fcon_1000/) which, at the present time, contains over 1000 subjects, and the relatively new OpenFMRI repository (www.openfmri.org) which contains imaging data from over 200 subjects. The Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC, www.nitrc.org) also hosts neuroimaging data, in addition to neuroimaging processing and analysis tools (Buccigrossi et al., 2008). Examples of neuroimaging repositories that require some form of permission to download data (e.g. prior IRB approval or simply an application to the host site), include the Alzheimer's Disease Neuroimaging Initiative (ADNI; http://adni.loni.ucla.edu/) which contains imaging data from over 800 subjects, and the National Database for Autism Research (NDAR; ndar.nih.gov) which contains data from over 6000 subjects. It is clear from this short (and by no means exhaustive) list of available neuroimaging repositories that data is accumulating in distributed domain-specific databases, rather than in a small number of central repositories. In addition, there is no integrated access mechanism, even for open-access resources, nor an accepted format for the critically important meta-data, necessary for making use of combined neuroimaging data. The Neuroscience Information Framework (NIF; www.neuinfo.org) (Gupta et al., 2008) provides integrated access to many neuroscience-related databases as well as other resources; researchers can identify imaging data sets for download from certain resources that have been mapped for the NIF interface, for example. Developing the meta-data formats and standards needed to understand the imaging data sets, or to capture the details of how the data were collected and processed, is outside the scope of NIF and other database mediators. Integrated access to existing resources, many already identified by NIF, when combined with such meta-data documentation, would provide a full-service shop for queries and download of publicly available data across projects.

A critical barrier in enabling structured sharing of raw and derived neuroimaging data across existing resources is the lack of a standard meta-data model and a set of informatics tools that enables the sharing of meta-data, including provenance, associated with neuroimaging data (Teeters et al., 2008). Meta-data are descriptive elements associated with data that provide additional clarity regarding acquisition parameters, experimental conditions, analysis procedures, and any other formation about the experiment or analyses that helps one understand and use the data. The benefits of neuroimage data sharing were introduced more than a decade ago (Van Horn and Gazzaniga, 2002; Van Horn et al., 2001), several successful data sharing projects exist (Biswal et al., 2010; Nooner et al., 2012; Weiner et al., 2012), and many of the technical, legal, and social issues of data sharing have been discussed (Mennes et al., 2012; Milham, 2012; Poline et al., 2012), but there currently is no standard format nor a set of lightweight tools that allow small laboratories or individual investigators to share imaging and *meta-data* in a structured way, nor a set of tools that allows for queries across existing databases or data sharing efforts (Poline et al., 2012). These problems are especially acute when attempting to construct large data sets from data available through online repositories, each of which uses different structures of storing meta-data.

The options for making data available are limited to putting raw data sets online, or putting raw and processed data sets online together with descriptions of the derived data in text documents (or to be gleaned

Please cite this article as: Keator, D.B., et al., Towards structured sharing of raw and derived neuroimaging data across existing resources, NeuroImage (2013), http://dx.doi.org/10.1016/j.neuroimage.2013.05.094