NeuroImage xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Sharing data in the global alzheimer's association interactive network

Scott C. Neu, Karen L. Crawford, Arthur W. Toga *

Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Accepted 27 May 2015
Available online xxxx

Keywords:
GAAIN
Data sharing
Neuroimaging
Data repository
Federation

ABSTRACT

The Global Alzheimer's Association Interactive Network (GAAIN) aims to be a shared network of research data, analysis tools, and computational resources for studying the causes of Alzheimer's disease. Central to its design are policies that honor data ownership, prevent unauthorized data distribution, and respect the boundaries of contributing institutions. The results of data queries are displayed in graphs and summary tables, which protects data ownership while providing sufficient information to view trends in aggregated data and discover new data sets. In this article we report on our progress in sharing data through the integration of geographically-separated and independently-operated Alzheimer's disease research studies around the world.

Introduction

At present there are many geographically-separated and independent studies of Alzheimer's disease and aging around the world (Erten-Lyons et al., 2012; Stanziano et al., 2010). A primary focus shared among these groups is to identify quality- and length-of-life predictors that can be used to develop strategies for reducing the burdens of chronic illness due to aging and disease (Stanziano et al., 2010). For example, a better understanding of how social and behavioral factors influence the effectiveness of interventions may lead to improved lives and reduced healthcare costs. Unifying these research efforts has the potential to reveal more insights into the causes of Alzheimer's disease, improve treatments, and design preventative measures that delay the onset of physical symptoms.

The statistical significance of research findings is dependent upon the amount of data available to study. Therefore aggregating data into larger pools is essential for effective data analysis. This not only increases the precision of measured results but also reveals trends and correlations that are not apparent from the smaller data sets themselves (Ferguson et al., 2014). Additionally, pooled data can be reused in new studies. This reduces the costs of those studies because the data does not need to be recollected and the different naming conventions and terminologies of the smaller data sets have already been harmonized (Poldrack and Gorgolewski, 2014).

There is currently a great deal of interest in promoting data sharing in neuroimaging (Ferguson et al., 2014; Poldrack and Gorgolewski, 2014). For our purposes, it is important to distinguish between two tiers of neuroimaging data sharing because the participants that share data in each tier have different needs and concerns. In the first tier is the individual research scientist who collects subject data, studies a particular cohort, and wishes to share the data with other researchers (Ferguson et al., 2014; Poline et al., 2012,). These scientists commonly lack resources to share data and tend to be focused upon the completeness and correctness of their publications (Ferguson et al., 2014). A "single bucket" system (all data resides in a single remote location) is often sufficient for them to share their data with other scientists. The system provides storage and retrieval services, user registration and authentication, and user access controls. Well known examples include the LONI IDA (Neu et al., 2012), PING Data Portal (Bartsch et al., 2014), LORIS (Das et al., 2011), COINS (Scott et al., 2011), XNAT (Marcus et al., 2007), and FITBIR.[1]

The second tier of sharing neuroimaging data consists of organizations that manage their own data repositories (Erten-Lyons et al., 2012; Stanziano et al., 2010) and have computing infrastructure, personnel resources, and software for data distribution and often make their data available to collaborators. These data repositories can include the single bucket systems in the first tier. Sharing data across repositories is complex because data repositories are inherently designed to manage and distribute data, not to interact with other repositories. Also, each data repository may be subject to local policies, ethical considerations, and legal obligations; often an Institutional Review Board places significant limits on the way that data can be shared and even stronger limits on its re-distribution (Hall et al., 2012). Federated approaches have been implemented [e.g., BIRN Human Imaging Database (HID) (Ozyurt et al., 2010), NeuroBase (Barillot et al., 2006), and NeuroLOG (Gibaud et al., 2011)] that use a server to distribute queries

* Corresponding author at: Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, 2001 North Soto Street, Room 102, Los Angeles, CA 90032, USA.
E-mail address: toga@loni.usc.edu (A.W. Toga).

[1] https://fitbir.nih.gov/jsp/about/index.jsp.

http://dx.doi.org/10.1016/j.neuroimage.2015.05.082
1053-8119/

to each data repository. The queries are accordingly reformulated at each data repository and the query results are returned and combined into a single result set. The National Database for Autism Research (NDAR) (Hall et al., 2012) is a noteworthy example of a platform that manages data sharing in both tiers. It not only receives and archives data from individual researchers of autism but is also federated with four other private databases.

The primary objective of the Global Alzheimer's Association Interactive Network[2] (GAAIN) is to establish a virtual community for sharing Alzheimer's-related data stored in independently-operated repositories around the world. Neuroimaging, demographic, genetic, and biologic data are integrated together while respecting the boundaries of existing repositories and protecting the ownership of shared data. In the next sections we describe the architecture and discuss how our implementation choices address the practical concerns of GAAIN's data partners.

Overview

The system architecture of GAAIN contains a central server that communicates with multiple client applications (Data Partner Clients or DPC's) that are installed at the data partner sites. As shown in Fig. 1, data is locally exported into CSV (comma-separated values) files and loaded into the DPC's. When GAAIN investigators query the network through its web interfaces, search requests are sent to the central server. The central server in turn sends requests to the DPC's that are "online" (accepting requests). The database in each online DPC is queried and the results are sent back to the central server where they are aggregated into the response passed on to the web interfaces.

The DPC is a Java jar file that contains both a light-weight web server[3] and database.[4] While the jar file is running, the DPC can be configured using its administrative web pages and its online/offline status can be changed. It uses a single directory for file storage and two user-configurable ports for administration and secure communications to the GAAIN central server. The data stored in the DPC may be updated at the convenience of the data partner, and the only institutional requirement is to change local firewall configurations to allow HTTPS traffic from the central server into the data partner's network.

Fig. 2 lists some of the data partners currently sharing or in ongoing discussions to share data through GAAIN. Data partner recruitment is in its early stages and is ongoing. Every data partner manages a significant data repository of Alzheimer's disease data housed in North America or Europe. There is a DPC running at each data partner site, with the exception of LAADC (at their request we manage their DPC). Fig. 3 summarizes the data that is available for sharing from each data partner. Searchable attributes include demographic data (e.g., age, gender, race), cognitive measurements [e.g., Mini-Mental State Examination (MMSE) (Folstein et al., 1975) and Global Clinical Dementia Rating (CDR) (Hughes et al., 1982) scores], historical and genetic information [e.g., parent history of Alzheimer's disease and Apolipoprotein E (APOE) genotype], biological measurements (e.g., CSF level of phosphorylated tau protein and glucose metabolism in the right hippocampus), and data segmented from neuroimaging scans (e.g., volume of the brain and hippocampi). GAAIN provides the resources to map the nomenclature and conventions used by data partners into the global schema used within GAAIN. These mappings are currently constructed in Java code and added to the DPC, but in the future we plan on developing a mapping tool to expedite the creation of new mappings as well as to update existing mappings.

Prospective data partners can apply to join GAAIN from the GAAIN website.[5] We ask that each data partner agrees to and signs a Memorandum of Understanding (MOU) that explicitly states that the data shared by the data partner will be de-identified and that the data partner will receive recognition on the GAAIN website and in all GAAIN-related presentations. Investigators can join GAAIN if they have a valid email address and they agree to acknowledge GAAIN and its data partners in all related publications.

Philosophy

The prominence of a data sharing network depends upon providing search functionality to those looking for data while addressing the concerns of those sharing data. As such, GAAIN aims to help scientists find Alzheimer's data for their research while protecting the data ownership rights of each of its data partners. GAAIN search results are displayed using graphs so that scientists can intuitively interact with the results and visualize trends in the data without having direct access to the shared data sets. As an added benefit, GAAIN search interfaces essentially advertise the data shared by each data partner and increase the public visibility of each partner data repository. Participation in GAAIN can also help its partners comply with data sharing requirements of their funding agencies.

GAAIN has specifically designed its architecture to address the practical concerns of its data partners. Design and policy choices that motivate membership are critical because participation in GAAIN is voluntary. Most Alzheimer's disease researchers have invested considerable time and resources in building their data repositories and therefore will be receptive to joining a data sharing network only if it requires little investment of their resources and only if their data repositories continue to manage their data. GAAIN recognizes and addresses these concerns:

Control. GAAIN data partners retain complete control over their data. GAAIN investigators are directed to the data use application pages of its data partners where they follow existing application processes. GAAIN does not grant access to partner data nor has access to the user authentication methods and access controls of its data partners. Every DPC has an "on/off switch" which provides the freedom to immediately disconnect data from the network ("go offline") at any time for any reason.

Light footprint. The GAAIN DPC at each data partner site does not interfere with or consume resources of the local production system. The DPC is typically installed on a computer system separate from the production system. This is possible because data is imported into the DPC from a CSV file that is created by exporting data from the production database. Since it does not have direct access to the production database, the DPC cannot disrupt the normal operations of the production system.

No copy policy. At no time will GAAIN store data from any data partner on any GAAIN central server computer disk, unless requested by the data partner. GAAIN will manage a DPC on its computers if a data partner does not wish to do so. GAAIN central servers do not save and manage copies of the data. However, data fulfilling investigator searches may be cached in server memory to optimize search performance but are never copied or written to disk. Whenever a partner goes offline, all data cached from the partner is erased.

Security. All communications between the DPC's and the GAAIN central server are performed securely using HTTPS. During the client registration process, privately-signed security certificates are exchanged and used to establish secure identities. When data partners are required to conduct security reviews of externally-developed software, GAAIN makes the DPC source code available for inspection.

Not all architectures that have been used for sharing neuroimaging data have taken these concerns into consideration. The BIRN HID

[2] http://www.gaain.org.
[3] http://www.eclipse.org/jetty/.
[4] http://www.h2database.com/html/main.html.
[5] http://www.gaain.org.

Please cite this article as: Neu, S.C., et al., Sharing data in the global alzheimer's association interactive network, NeuroImage (2015), http://dx.doi.org/10.1016/j.neuroimage.2015.05.082