

Reproducibility in human cognitive neuroimaging: a community-driven data sharing framework
for provenance information integration and interoperability

Boyce Nolan Nichols III

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

James F. Brinkley, Chair

Thomas J. Grabowski

Nicholas R. Anderson

Program Authorized to Offer Degree:
Biomedical and Health Informatics, School of Medicine

University of Washington

Abstract

Reproducibility in human cognitive neuroimaging: a community-driven data sharing framework
for provenance information integration and interoperability

Boyce Nolan Nichols III

Chair of the Supervisory Committee:

Professor James F. Brinkley

Biological Structure

Access to primary data and the provenance of derived data are increasingly recognized as an essential aspect of reproducibility in biomedical research. While productive data sharing has become the norm in some biomedical communities, human brain imaging has lagged in open data and descriptions of provenance. The overarching goal of my dissertation was to identify barriers to neuroimaging data sharing and to develop a fundamentally new, granular data exchange standard that incorporates provenance as a primitive to document cognitive neuroimaging workflow.

For my dissertation research, I led the development of the Neuroimaging Data Model (NIDM), an extension to the W3C PROV standard for the domain of human brain imaging. NIDM provides a language to communicate provenance by representing primary data, computational workflow, and derived data as bundles of linked Agents, Activities, and Entities. Similar to the way a sentence conveys a standalone thought, a bundle contains provenance statements that parsimoniously express the way a given piece of data was produced. To demonstrate a system that implements NIDM, I developed a modern, semantic Web application platform that provides neuroimaging workflow as a service and captures provenance statements as NIDM bundles. The course of this work necessitated interaction with an international community, which adopted

and extended central elements of this work into prevailing brain imaging software. My dissertation contributes neuroinformatics standards to advance the current state of computational infrastructure available to the cognitive neuroimaging community.

Table of Contents

LIST OF FIGURES	VII
LIST OF TABLES	VIII
ABBREVIATIONS	IX
ACKNOWLEDGMENTS	XI
CHAPTER 1: INTRODUCTION	1
1.1 OVERVIEW	1
1.2 MOTIVATION FOR RESEARCH	2
1.2.1 DIGITAL BRAIN ATLAS TERMINOLOGIES ARE NOT HARMONIZED	4
1.2.2 NEUROIMAGING DATA IS STORED IN HETEROGENEOUS SOURCES	5
1.2.3 DATA ACQUISITION AND ANALYSIS METHODS EVOLVE RAPIDLY	6
1.3 STATEMENT OF THE PROBLEM	7
1.4 SCOPE OF THE STUDY	8
1.5 CONTRIBUTIONS	8
1.5.1 CONTRIBUTIONS TO COGNITIVE NEUROIMAGING	9
1.5.2 CONTRIBUTIONS TO NEUROINFORMATICS	9
1.5.3 CONTRIBUTIONS TO BIOMEDICAL INFORMATICS	10
1.6 SUMMARY	10
CHAPTER 2: BACKGROUND	12
2.1 OVERVIEW	12
2.2 HUMAN COGNITIVE NEUROIMAGING	12
2.2.1 MAGNETIC RESONANCE IMAGING	13
2.2.2 BRAIN ATLAS-BASED LABELING	14
2.3 APPLICATION OF NEUROINFORMATICS	18
2.3.1 ADVANCES FROM LARGE-SCALE INITIATIVES	18
2.3.2 FACILITATING OPEN COLLABORATION AND COORDINATION	21
2.4 BIOMEDICAL INFORMATICS METHODS	23
2.4.1 DATA MANAGEMENT AND INTEGRATION	24
2.4.2 SEMANTIC WEB TECHNOLOGIES	28
2.4.3 DATA PROVENANCE INFORMATION	30
2.5 SUMMARY	31
CHAPTER 3: STUDY DESIGN	32
3.1 OVERVIEW	32
3.2 GENERAL RESEARCH APPROACH	32
3.2 SPECIFIC AIM 1: RESEARCH AND DESIGN A FRAMEWORK TO REPRESENT, ACCESS, AND QUERY NEUROIMAGING PROVENANCE INFORMATION	37
3.2.1 DESIGN AND METHODS	38
3.4 SPECIFIC AIM 2: DEVELOP AN INFORMATION SYSTEM OF WEB SERVICES TO COMPUTE AND DISCOVER DATA PROVENANCE FROM BRAIN IMAGING WORKFLOW	39
3.4.1 DESIGN AND METHODS	40
3.5 SUMMARY	41
CHAPTER 4: NEUROIMAGING DATA MODEL FRAMEWORK	42
4.1 OVERVIEW	42

4.2 A NEUROIMAGING DATA SHARING SCENARIO	42
4.3 FRAMEWORK RESEARCH AND DESIGN STUDY	45
4.3.1 RESEARCH APPROACH	46
4.3.2 OUTCOMES AND DESIGN	56
4.4 CONCLUSIONS	75
CHAPTER 5: NEUROIMAGING APPLICATION FRAMEWORK	77
5.1 OVERVIEW	77
5.2 PROTOTYPE INFORMATION SYSTEM STUDY	77
5.2.1 RESEARCH APPROACH	78
5.2.2 OUTCOMES AND INFORMATION SYSTEM	85
5.3 CONCLUSIONS	95
CHAPTER 6: CONCLUSIONS AND SUMMARY	97
6.1 DISSERTATION SUMMARY	97
6.2 IMPLICATIONS FOR COGNITIVE NEUROIMAGING	99
6.2 IMPLICATIONS FOR BIOMEDICAL INFORMATICS	99
6.4 LIMITATIONS	100
6.4.1 DATA MODEL LIMITATIONS	100
6.4.2 APPLICATION FRAMEWORK LIMITATIONS	102
6.5 FUTURE DIRECTIONS	103
6.5.1 NIDM OBJECT MODEL EXTENSIONS	103
6.5.2 NiQUERY SOFTWARE DEVELOPMENT	104
6.6 FINAL CONCLUSION	106
REFERENCES	107

List of Figures

Figure Number	Page	
2.1	Fundamental Theorem of Biomedical Informatics	24
2.2	PROV Data Model core constructs	31
3.1	Biomedical Informatics Perspective	34
3.2	Research Phases and Events Timeline	35
4.1	FMA Brain Atlas Mapping	48
4.2	Image Annotation Service	53
4.3	XCEDE API	55
4.4	NIDM Layer Cake	57
4.5	NIDM Information Domains	58
4.6	OpenfMRI and Study Forrest Mapping to NIDM	59
4.7	Database Descriptor Element	61
4.8	Details of NIDM Dataset Descriptor	61
4.9	Project Descriptor Element	62
4.10	Release Descriptor Element	63
4.11	Aggregate Descriptor and Linkset Descriptor Elements	64
4.12	Component Descriptor Element	65
4.13	OpenfMRI Data Organization	66
4.14	NIDM Experiment Modeling Pattern	67
4.15	NIDM Experiment Elements	68
4.16	NIDM Project Element	69
4.17	Questionnaire Study Element	71
4.18	MRI Study Element	71
4.19	Acquisition Element	72
4.20	NIDM Modeling Process	75
5.1	SNI System Architecture	81
5.2	NiQuery Prototype	83
5.3	NiQuery Time series Demo	84
5.4	NiQuery System Architecture	87
5.5	Query Metadata for FreeSurfer	90
5.6	FreeSurfer Workflow Interface Query	90
5.7	Check cache for previously computed results	91
5.8	Nipype Entity, Activity, and SoftwareAgent	92
5.9	NIDM Results representation of FreeSurfer statistics files	94
5.10	Prototype App to work with NIDM and NiQuery	95

List of Tables

Table Number
2.1

5-Star Linked Open Data rating system

Page
29

Abbreviations

AAL	Automated Anatomical Labeling
ADHD	Attention Deficit Hyperactivity Disorder
AIM	Annotation Image Markup
API	Application Programming Interface
CUMBO	Common Upper Mammalian Brain Ontology
DICOM	Digital Imaging and Communications in Medicine
DUA	Data Use Agreement
FMA	Foundational Model of Anatomy
fMRI	functional Magnetic Resonance Imaging
HBP	Human Brain Project
IAS	Image Annotation Service
INCF	International Neuroinformatics Coordinating Facility
LOD	Linked Open Data
MRI	Magnetic Resonance Imaging
NAKFI	National Academies Keck Futures Initiative
NIF	Neuroscience Information Framework
NIFSTD	Neuroscience Information Framework Standard Ontology
NIFTI	Neuroimaging Informatics Technology Initiative
NIMS	Neurobiological Imaging Management System
OBO	Open Biomedical Ontologies
OWL	Web Ontology Language
PI	Principle Investigator
QI	Query Integrator
RO	Relations Ontology
ROI	Region of Interest
TD	Talairach Daemon

UW University of Washington
VoID Vocabulary of Interlinked Datasets
XCEDE XML-based Clinical and Experimental Data Exchange
XNAT eXtensible Neuroimaging Archive Toolkit

Acknowledgments

I wish to express gratitude to my wife, Lindsay, and to my parents, Boyce and Debbie Nichols, for encouraging and supporting my pursuit of a doctoral degree in Biomedical and Health Informatics at the University of Washington. I would also like to thank my mentors and committee members, Drs. James Brinkley, Thomas Grabowski, Nicholas Anderson, and Susan Coldwell for many thoughtful discussions that led me down this research path. Additionally, I am appreciative of my collaborators at the Integrated Brain Imaging Center and Structural Informatics Group to whom I am indebted for much of the knowledge I acquired during my studies.

A special thanks to the International Informatics Coordinating Facility (INCF) for providing me with funding to travel and collaborate with the Neuroimaging Data Sharing (NIDASH) task force, and for the leadership of David Kennedy (University of Massachusetts, Worcester, USA) and Jean-Baptiste Poline (Commissariat à l'Energie Atomique, Gif-sur-Yvette, France). I am also deeply grateful to the other task force members who were influential in my data sharing efforts: Satrajit Ghosh (Massachusetts Institute of Technology, Cambridge, USA), Rich Stoner (University of California, San Diego, USA), Arno Klein (Columbia University, New York, USA), David Keator (University of California, Irvine, USA), Jessica Turner (The Mind Research Network, Albuquerque, NM, USA), Cameron Craddock (Child Mind Institute, USA), Guillaume Flandin (University College London, UK), Chris Gorgolewski (Stanford University, Stanford, USA), Yaroslav Halchenko (Dartmouth College, Hanover, NH, USA), Michael Hanke, Institut für Psychologie (Otto-von-Guericke-Universität, Magdeburg, Germany), Christian Haselgrove (University of Massachusetts, Worcester, USA), Daniel Marcus (Neuroinformatics Research Group, Washington University in St. Louis, MO, USA), Camille Maumet (University of Warwick, Coventry, UK), Thomas Nichols (University of Warwick, Coventry, UK), Russell Poldrack (Stanford University, Stanford, USA).

Dedication

To my grandparents, Lewis and Dolores Long, who always nurtured my creative pursuits, and to my wife, Lindsay, whose warmth, support, and love brightens my every day.

Chapter 1: Introduction

"We have to overthrow the idea that it's a diversion from 'real' work when scientists conduct high-quality research in the open. Publicly funded science should be open science. Improving the way that science is done means speeding us along in curing cancer, solving the problem of climate change and launching humanity permanently into space."

- Michael Nielsen, Reinventing Discovery: the new era of networked science

1.1 Overview

Access to primary data is increasingly recognized as an essential aspect of reproducibility in biomedical research (1). While widespread sharing is being adopted in some biomedical communities (2-4), it largely remains an exception in the field of human brain imaging, except for several progressive examples (5-10). The discussion of neuroimaging data sharing is not new (11-14), but the adoption of values that support a cultural shift embracing data sharing has yet to take place. Of particular concern is reproducibility in neuroimaging research, where openly available data can help to validate findings and eliminate confounds due to, for example, specific preprocessing or analysis methods, which are nearly as prevalent as the number of studies themselves (15). For the neuroimaging community to fully embrace open data sharing, tools and standards are needed to make the process and liabilities related to sharing as predictable as possible. Furthermore, the community itself needs a set of qualified individuals to take on the role of developing and maintaining these tools and standards by incorporating them into the everyday workflow of neuroimaging researchers.

This dissertation advances the current state of reproducibility in neuroimaging by contributing informatics approaches that can be applied to facilitate the broad adoption of open data exchange standards that capture provenance information for use in the brain imaging community. To accomplish this task, I proposed two specific aims to realize a community-driven framework for collaboratively improving the current state of neuroimaging data sharing and reproducibility, which are to 1) research and design a framework to represent, access, and query neuroimaging provenance information, 2) develop an information system of Web services to compute and discover data provenance from brain imaging workflow.

1.2 Motivation for research

Human cognitive neuroimaging is a scientific discipline that investigates the relationship between brain structure and function in normal and neuropsychiatric conditions using a variety of medical imaging technologies. Additionally, most neuroimaging studies also collect cognitive and neuropsychological data that evaluate performance along cognitive dimensions, such as attention and memory. This broad spectrum of data leads to complex datasets that are difficult to manage and analyze, and, at nearly \$1,000 per hour to operate medical imaging equipment (e.g., Magnetic Resonance Imaging, MRI), neuroimaging studies are expensive to conduct. While limited resources for funding make competition for grants in this domain hypercompetitive, there is great potential for researchers to provide insight into clinically relevant questions that may eventually improve human health. Those talented, and lucky enough, to be funded are afforded the opportunity to acquire rich phenotypic information from study participants to answer a diverse set of hypotheses. However, with so many open questions that can be tested with a given dataset, it is unlikely that any given lab will conduct every possible and valid analysis. To maximize the return on investment in neuroimaging research projects, it is clear that more can be done to ensure that these data are used to their greatest potential.

In order to maximize the amount of knowledge generated from neuroimaging datasets, it is vital to ensure that the data collected, from raw to post-processed to fully analyzed, along with the accompanying metadata, are made openly available to as many scientists as possible. This is particularly relevant for data aggregation, where large datasets can be used to reproduce results with an appropriate level of statistical power. Across the scientific biomedical research enterprise, there is concern about the reproducibility of discoveries (16), and, in neuroscience research, a lack of statistical power has been identified as a key issue in meta-analyses attempting to identify reproducible findings (17). Additionally meta-research on neuroimaging studies has indicated a reporting bias in studies of brain volume abnormalities (18) and functional MRI (fMRI) activation foci (19). To help address statistical power issues, massive and openly available datasets are needed, and may be aggregated, through data sharing initiatives that target multi-site research consortia, as well as individual investigations.

While large-scale data sharing sounds like a logical step forward, there are a number of socio-technical barriers that prevent this paradigm from becoming routine. Part of the solution is to address limitations in informatics infrastructure, but these barriers cannot be addressed by information technology alone. At the root of the issue is a lack of resolve in the neuroimaging community to prioritize transparency and reproducibility. Instead, low powered studies exist in siloes that obscure the data and methods used to conduct research. Consensus for communicating scientific results is starting to emerge for some types of analyses (e.g., task-based fMRI, (20)), but scientific communication norms are not necessarily compatible across analyses or software packages (e.g., due to fundamentally different statistical approaches or brain atlases). Research silos were partially fueled by competition between labs where rewards focused on scientific innovation and establishing local expertise rather than interoperability and collaboration. This hypercompetitive, "survival of the fittest," culture may have been necessary to build up the neuroimaging infrastructure that now provides robust analysis tools and data management applications; however, as brain imaging research has expanded in scope to include multi-site research consortia, the need for collaboration and application of a broad range of analysis methods has come to the fore. With this distributed and collaborative space in mind, three examples are provided below that highlight barriers to reproducibility caused by heterogeneous brain atlases, data management systems, and analysis methods.

Before I examine these examples in detail and their relationship to reproducibility, I will present my definition of reproducibility - a term with a more broad scope than is readily recognized. The spectrum of reproducibility can be divided into four subcategories that are defined below:

- Repeatable: same researchers in the same lab can obtain consistent results using the same methodology
- Replicable: different researchers from a different lab can obtain consistent results using the same methodology and data

- Reproducible: different researchers from a different lab can obtain consistent results using a different methodology and data
- Reusable: different researchers from a different lab can confidently apply a methodology and/or access shared data from different researchers in a different lab

The union of these four subcategories constitutes the reproducibility spectrum (21). The motivations addressed below influence each of these stages, but my primary focus is on reusability. Reusable research must provide a description of the analysis methodology, as well as transparent access to the entire experimental environment (i.e., data and software). Ideally, the entire computational environment (e.g., a virtual machine image) would be deployable where a lab in a physically distinct location could reuse the same methodology and original data to make any findings repeatable. With the goal of achieving reusability, the next section continues with a discussion of specific research motivations.

1.2.1 Digital brain atlas terminologies are not harmonized

The diverse set of human brain structural and functional analysis methods represents a difficult challenge for reconciling multiple views of (functional) neuroanatomical organization. While different views of anatomical organization are expected and valid, no widely adopted approach exists to harmonize different views of neuroanatomical organization. In human brain imaging research, brain atlases are constructed using cortical parcellation protocols and terminologies that serve as spatial frameworks for representing complex neuroimaging datasets. However, each method used to create a brain atlas applies a specific set of terms that are based on different parcellation schemes (10,22-25). As a result, the anatomical entities from different parcellation schemes do not have a one-to-one mapping and are difficult to compare. Anatomical structure provides a natural organizing framework that can be used to correlate different terminologies, and thus a basic requirement for facilitating the integration of basic and clinical neuroscience data from diverse sources is a well-structured ontology that can incorporate, organize, and associate annotated neuroanatomical data (26).

In research conducted by Dr. Brinkley's Structural Informatics Group (SIG), the Foundational Model of Anatomy (FMA) ontology was created to provide a computable representation of human anatomy (27). The structural framework provided by the FMA was extended to capture the variation across neuroanatomical labeling schemes and protocols, thus creating a mechanism to correlate labels from different atlases. As brain atlas labels are used to annotate the results of neuroimaging studies, the FMA can be used to enhance information retrieval of such annotated datasets or publications using these labels, as demonstrated by Turner et al (28). While the FMA was developed using Protégé Frames (29), efforts have been made to translate it into the Web Ontology Language (OWL, (30)), and these efforts have provided versions of the FMA using the OWL standard (31,32). With the FMA in OWL, I experimented with query-based data integration technologies (33) that enabled me to understand how the FMA could be used in a practical way to facilitate performing intelligent queries over neuroimaging datasets, thus providing additional features to data shared with annotations from the FMA. The intelligent query features provided by the FMA inspired me to pursue data sharing technologies and approaches that would readily facilitate the use of OWL and other semantic Web technologies, particularly for the integration of data from clinical data management systems.

1.2.2 Neuroimaging data is stored in heterogeneous sources

While individual clinical data management systems provide mechanisms to query and download information within a given software application (34-37), there is no simple way to integrate information across applications. Users seeking to aggregate data must access one system at a time, generally by visiting a web application where they can browse for datasets that match their research questions. After learning the unique user interface of each system, they then download relevant datasets and quickly discover heterogeneous file naming conventions and directory structures. Supporting information (e.g., data dictionaries that describe imaging parameters, demographics, neuropsychological tests etc.) from each clinical data management system must then be acquired to provide the context needed for analysis, which may or may

not be readily available. It is then up to the user to map the context from one system to another, which is a dauntingly time consuming and error prone process that is commonly avoided altogether. Only after working through these details can they start to evaluate data quality and determine which datasets are useable for their analysis.

After coming to understand this scenario, I wanted to study the socio-technical barriers that were limiting data integration efforts to aggregate and index metadata about the neuroimaging datasets that were coming available to the community. Following from my experience with the FMA and semantic Web technologies, it was clear that an approach and process was needed to capture data dictionaries and term definitions in a reusable and computable representation. Such a contribution would greatly improve the process of data integration and facilitate interoperability between clinical data management systems used for neuroimaging research; however, technology is only one piece of the puzzle. I learned that for data sharing to work it takes a community to drive the development of data exchange standards to encourage broad adoption, a finding also noted in the Synthetic Biology community (38). While community-driven data exchange standards meets the requirements for reusable data, it neglects the need to capture provenance information about the computational environment (e.g., software versions, parameters, etc.) that are constantly changing and necessary to track to make a given methodology transparent.

1.2.3 Data acquisition and analysis methods evolve rapidly

The past two decades of human brain imaging have witnessed a surge of innovations in Magnetic Resonance Imaging (MRI) acquisition and analysis technology. These advances have afforded cognitive neuroscientists with a much richer description of both normal and pathological brain structure and function. However, the rapid generation of new knowledge afforded by these advances has also increased the computational complexity by which knowledge is generated, thus undermining the scientific community's ability to replicate and reproduce results. This is an issue in many areas of science that face the problem of understanding and interacting with imaging data whose scale and scope is relentlessly

expanding. The challenge is particularly acute in brain imaging and cognitive neuroscience because the interplay between rapidly evolving instrumentation and analysis methods have outpaced the infrastructure for integrating and validating biologically meaningful interpretations of neuroimaging results. For example, the FreeSurfer data analysis package is used broadly to reconstruct the cerebral cortical surface and make measurements (e.g., brain region volume or cortical thickness) of segmented or parcellated structures (39); however, there are significant differences in these measures across software versions and compute platforms (i.e., operating systems) making it difficult to rely on quantified results (40). This variability in results influenced researchers to recommend guidelines for software evaluation using publicly available datasets that help to determine the reproducibility of computational algorithms (41).

As the neuroimaging field evolves, there is a growing need for neuroinformatics to provide technical solutions that facilitate replication studies and reusable computational workflow. Replication studies are a central tenet of the scientific method that values findings validated by an independent investigator using an independent sample; however, much of the neuroimaging literature is based on data that is difficult to collect due to time, finances, or specific participant populations - generally limiting the feasibility of replication studies. A data sharing framework may be developed to enhance reproducibility by capturing the computational environment, software tools, workflow, and data used to generate neuroimaging findings. With semantically annotated results using terms from a curated terminology source (e.g., the FMA) and data exchange standards that support annotations (e.g., semantic Web technologies), a platform could be developed capable of making all the components necessary for portable and reproducible research.

1.3 Statement of the problem

The human brain mapping field is producing a tremendous amount of medical imaging, clinical, and neuropsychological data through a broad range of research activities; however, there is little coordination across these research activities to develop standards and best practices that facilitate data integration and interoperability. The scarcity of coordination has led to overly

burdensome data management practices and a lack of provenance information that exacerbates data sharing barriers and limits the reproducibility of neuroimaging findings. In response to these issues, I took a leadership role in the neuroinformatics community to design a framework for data sharing that addresses the shortcomings of previous approaches.

1.4 Scope of the study

The overall scope of my dissertation is constrained to open access data sharing in the domain of human brain imaging. In the course of this work I demonstrate a neuroimaging data exchange framework that is capable of representing a full chain of provenance information, from dataset descriptors of acquired data to workflow that operates on the data and the derived results. My focus is to create a process that facilitates community-driven efforts to harmonize metadata descriptions across openly available neuroimaging data repositories and derived results. I provide software tools to access and query these metadata descriptors, as well as demonstrate a distributed computing system architecture that uses these metadata descriptors to perform analyses. The system is an informatics research project that serves as a prototype for future work. As such, I limit the scope of this project as follows:

1. using publicly available human neuroimaging datasets
2. using anatomical MRI and resting-state fMRI modalities
3. an open and unsecured network architecture
4. a prototype implementation of the system architecture
5. the current state of an evolving data model

1.5 Contributions

Since starting my dissertation, I've had the unique opportunity to actively participate in a community effort to improve the state of data sharing in human brain imaging. I engaged this community, led by an organization called the International Neuroinformatics Coordinating Facility (INCF), with a project called NiQuery that aimed to provide a distributed computing framework for publicly accessible neuroimaging data. My effort with NiQuery was closely related to an ongoing activity within INCF called the Neuroimaging Data Sharing (NIDASH) task

force, where I was invited to contribute to a broad range of issues that became incorporated into my dissertation. As I worked on NiQuery, it became clear that the vision surrounding the project would require a community to bring to fruition. To that end, I took a leadership role in NIDASH by attending and presenting my work at data sharing satellite meetings and major conferences, as well as contributing to weekly videoconference calls - all focused on advancing the mission of inciting a cultural shift around neuroimaging data sharing. In the following sections I detail my specific contributions to the three fields that my dissertation work impacted - cognitive neuroimaging, neuroinformatics, and biomedical informatics.

1.5.1 Contributions to cognitive neuroimaging

As an informatician, my contributions to cognitive neuroimaging have had an impact on the availability and use of neuroscience information. I was a principal data architect of the community-driven Neuroimaging Data Model (NIDM) framework for developing metadata standards that intrinsically tracks data provenance. By designing this framework, I enabled interoperable specifications to model data exchange across neuroimaging databases and analysis methods. Additionally, I provided the community with the framework to build a repository of neuroimaging datasets described using NIDM, which encapsulate recommendations for best practices in neuroimaging data sharing.

1.5.2 Contributions to neuroinformatics

In the neuroinformatics community, particularly neuroimaging informatics, I worked closely with international experts in neuroimaging database development, workflow environments, and reporting of statistical results to guide the development of specifications that conform to the principles of Linked Open Data, and harmonize with existing data sharing initiatives beyond neuroimaging. I took a leadership role in developing the NIDM framework, which not only provides several example model representations, but also provides a process for community involvement in designing standard object models. Additionally, I created a prototype software library, called NiQuery, which provides developers and users with a set of queries to access and

validate NIDM components and to execute workflows that consume and produce NIDM Linked Open Data.

1.5.3 Contributions to biomedical informatics

Within the broader Biomedical Informatics community, my contributions can be viewed as a generalizable set of processes for engaging a community for developing interoperable data exchange standards. I also introduced the concept of using data provenance as a central component in biomedical data exchange standards. This approach enables transformations of data to be captured intrinsically within the data model, where a single document can describe a primary observation alongside computational workflow details, as well as statistical results reported in a manuscript. Within this context, I have described and demonstrated a general-purpose system architecture for reproducible workflow environments that utilizes concepts from distributed and cloud computing. Finally, I developed an approach to deploying the prototype system architecture that is fully automated and overcomes many of the limitations of other informatics systems that are overly complicated to configure and deploy.

1.6 Summary

In this chapter, I introduced my dissertation by reviewing the motivation for research, the problem addressed, its scope, and contributions to basic and applied Biomedical Informatics. The motivation for my research broadly targets neuroimaging data sharing from the perspective of data and software interoperability. I described how interoperability issues manifest in datasets annotated with anatomical labels from diverse brain atlases, and in databases that use inconsistent vocabulary and heterogeneous data structures, as well as analysis methods that rapidly evolve and make analyses difficult to reproduce over time. I then provided a concise description of the problem I addressed and the scope of my research, which follows several constraints that I targeted to work within to provide structure to the breadth of topics I have covered. My contributions demonstrate the interdisciplinary approach that I have taken to not only develop a technology in isolation, but to work directly with the neuroimaging

and neuroinformatics communities to solve problems in neuroimaging data sharing in a real world setting.

This dissertation addresses the need for a cultural shift towards widespread neuroimaging data sharing that covers a broad range of topics. In Chapter 2, I provide background information on previous work that has shaped the current environment of neuroinformatics tools and identify their shortcomings. This assessment provides the foundation upon which innovative information systems can be developed that will accelerate the rate of scientific discoveries that advance our understanding of how brain structure and function explain cognition and neuropsychiatric disease. Towards this end, I outline my overall study design in Chapter 3 and then describe my approach and outcomes to Aim 1 in Chapter 4, providing several studies that address issues related to standardizing neuroimaging data exchange and representation. In Chapter 5, I discuss my approach and outcomes for Aim 2, which applies the standards I designed to develop an information system that executes computational workflow while tracking data provenance. Finally, I discuss my conclusions and contributions in Chapter 6, which advance the current state of neuroimaging data sharing and reproducibly. I start this discussion by providing a background on the biomedical and informatics disciplines that my dissertation has impacted.

Chapter 2: Background

"A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

- Herbert A. Simon, Designing Organizations for an Information-Rich World

2.1 Overview

The previous chapter provided an executive summary of the motivations, problems addressed, and contributions of this dissertation that builds upon a broad and interdisciplinary body of work. This chapter is designed to provide the reader with a background on the topics covered in this dissertation. It begins with a discussion on a specific biomedical domain, the field of human cognitive neuroimaging, where the interdisciplinary nature of this field and the wide adoption of Magnetic Resonance Imaging (MRI) and brain atlases for reporting results are presented. Then, the discussion turns towards the influence that large-scale brain imaging initiatives, such as the Human Brain Project in the 1990's (42), have had on how everyday research is conducted, which has changed dramatically with the introduction of neuroinformatics software. After reviewing the current state of neuroinformatics, and neuroimaging informatics in particular, I discuss biomedical informatics (BMI) more generally by introducing the core methods and technologies that I applied and extended upon in my dissertation. The organization of this content follows the general framework provided by BMI, where a biomedical domain of interest (human cognitive neuroimaging) motivates applied informatics research and practice (neuroinformatics) that is then used to research and apply general BMI methods, techniques, and theories.

2.2 Human cognitive neuroimaging

Human cognitive neuroimaging is an interdisciplinary branch of neuroscience where researchers rely on multi-modal imaging technologies and analytical tools to study normal and diseased brain structure and function. The breadth of imaging devices available enables researchers to collect data that probe diverse aspects of neurobiology that include neuroanatomical structure, physiology, and chemical composition. Among the primary medical

devices used in neuroimaging is Magnetic Resonance Imaging (MRI), which has proven to be one of the most versatile of the imaging tools available to neuroscientists, particularly for creating brain maps based on anatomical structure and function. Additionally, the collected demographic and phenotypic data can be correlated with imaging-based measures. For example, standardized neuropsychological test batteries are administered to measure performance on tasks that range from facets of memory and attention to reaction time and impulsivity. These data can then be analyzed to reveal group differences in brain structure or function along a given behavioral or cognitive dimension. In the following sections I provide a brief overview of cognitive neuroimaging research that uses MRI and related acquisition modalities and then discuss the use of brain atlases for annotating analysis results with neuroanatomical labels.

2.2.1 Magnetic resonance imaging

MRI scanners are one of the most flexible data acquisition instruments available for cognitive neuroimaging research. This flexibility is afforded by the programmability of MRI using pulse sequences that detect magnetic properties of brain (or other) tissue that are specific to the given pulse sequence. For a review of different pulse sequences and the tissue properties they measure see (43). This enables an MRI to detect a variety of magnetic properties that are spatially localized and encoded as voxels (i.e., volumetric picture elements). The signal in each spatially encoded voxel is used to construct a contrast between different types of tissue. T1 and T2 weighted MRI scans construct contrasts that reveal purely anatomical features related to water and fat content, where one slice is acquired at a time along the sagittal, axial, or coronal axis; however, more recent pulse sequences can acquire a whole brain volume. Regardless of slice-by-slice images or whole brain acquisitions, each voxel in an anatomical MRI can be used to differentiate between tissue types, such as gray matter, white matter, and cerebrospinal fluid. As MRI is reconfigurable, additional pulse sequences are used to construct contrasts using other magnetic properties of tissues that can then be registered to anatomical scans.

For example, functional MRI (fMRI) uses a property of hemoglobin, referred to as the blood oxygenation level dependent (BOLD) signal, in which a contrast is constructed between

the differing magnetic properties of oxygenated and deoxygenated blood (44). Neural activity is linked to metabolic and blood flow demand, resulting in changes to hemoglobin oxygenation that are detectable with MRI. Cognitive neuroscientists and psychologists have used fMRI to examine a variety of cognitive processes ranging from understanding basic brain functions, such as language, attention, and memory, to neuropsychiatric disease, such as Alzheimer's disease, Attention Deficit Hyperactivity Disorder (ADHD), and Autism. Other imaging modalities have also emerged, such as Diffusion Weighted Imaging (DWI, (45)), which provides a measure of how water diffuses within the myelin encased axons of the brain (i.e., white matter) that can be used to construct measures of white matter integrity, as well as tractography streamlines that represent an abstraction of white matter bundles connecting brain regions and the broader central nervous system. While both fMRI and DWI provide complementary information about brain function and structural connectivity, respectively, the resolution they are acquired at (e.g., 2-4 mm³) is typically much more coarse than anatomical T1 acquisitions (e.g., 1 mm³ isotropic). Therefore, anatomical images are used to provide the neuroanatomical and structural framework for interpreting and reporting the results of fMRI and DWI acquisitions, which is accomplished through a process called registration (46).

High resolution anatomical scans provide the basis for grounding more coarse grained imaging modalities to anatomical structure through the use of brain atlases and analysis techniques. The next section discusses issues related to brain atlas labels and how the diversity of brain atlas labels complicate data integration tasks, such as those approached in this dissertation.

2.2.2 Brain atlas-based labeling

One important component of cognitive neuroimaging research involves the development of digital brain atlases, which serve as the spatial framework for representing complex neuroimaging datasets. Brain atlases consist of a brain template and a labeling protocol. A brain template is an image in a standard coordinate system that is typically generated by registering the brains from one or more participants into an average brain. The average brain is then

delineated according to a protocol that specifies the boundaries for anatomical structures, which are each given a label (e.g., left hippocampus). Generally, anatomical imaging data from individual participants are aligned to brain templates in a standard spatial coordinate system that provides a set of anatomical labels used to interpret the anatomical location of an observation and annotate the results (e.g., a t-statistic for an activation foci from fMRI). Brain atlases can be based on manual or automatically labeled brain parcellations that are derived using volume-based or surface-based methods. Each method used to create a brain atlas applies a specific set of terms that are based on different parcellation schemes (10,22-25). As a result, the anatomical entities defined in different parcellation schemes do not directly correspond with a one-to-one mapping; therefore, data annotated with terms from disparate atlases are difficult to compare and integrate.

Brain atlas terminologies lack the multi-faceted hierarchical organization provided by an ontological framework, which is needed to explicitly declare the semantics of the terms used, both by providing definitions to terms and relationships to other structures. Neuroimaging data and information encoded by brain atlas terms alone cannot be accurately interpreted, compared, correlated and applied across different studies. This becomes particularly relevant during information retrieval tasks that require integrating datasets labeled with terms from different atlases. Similarly, white matter connectivity atlases are being developed (47,48), but only limited "ground truth" information is available about *in vivo* human brain connectivity. Different parcellation methods are represented in a growing number of white matter atlases designed to better segment and identify structures in clinical MRI research studies. To better grasp how each atlas or labeling scheme is structured, a summary of brain parcellation protocols is provided that each capture anatomical knowledge and spatial relationships including the Talairach Daemon, Desikan-Killiany (i.e., FreeSurfer), Anatomical Automatic Labeling (AAL), and NeuroLex.

The Talairach Daemon (TD) is an information system that provides a mapping between 3D coordinates (i.e., Talairach coordinates) and specific brain structure labels (49). It is a digital representation of the original Talairach atlas (23) that is hierarchically organized into five

hierarchical levels: 1) Hemisphere, 2) Lobe, 3) Gyrus, 4) Tissue type, and 5) Cell type. For example, the label "Right Cerebrum.Temporal Lobe.Inferior Temporal Gyrus.Gray Matter.Brodmann area 20" represents a number of 3D coordinates in the Brodmann area 20 cell type-level, the gray matter tissue-type level and so on. While this approach has been broadly applied in human brain mapping, there are limitations when normalizing patient MRI scans due to natural morphological differences between individuals. Furthermore, the Talairach atlas brain was not analyzed microscopically to determine regional cell types, thus the accuracy of Brodmann area boundaries in this atlas are unknown.

The Desikan-Killiany (DK, (50)) atlas is a gyral-based parcellation scheme for labeling anatomical MRI scans. The parcellation scheme protocol was manually applied to 40 MRI scans to build an atlas with 34 cortical regions of interest (ROI) per hemisphere. This atlas is incorporated into the FreeSurfer MRI data analysis package (39) that provides researchers with access to a variety of image processing tools that includes labeling anatomical ROIs with a predefined set of terms.

The Automated Anatomical Labeling (AAL) brain atlas provides an anatomical parcellation of a single participant using magnetic resonance imaging (MRI) (24). The AAL provides the location and labels for 90 anatomical structures (45 per hemisphere) that were manually identified in a high-resolution MRI. The AAL Toolbox for the Statistical Parametric Mapping MatLab package (51,52) provides researchers with a method for labeling their data using the AAL protocol and vocabulary.

The Neuroscience Information Framework (NIF) standard ontologies (NIFSTD) are developed to provide a consistent source of terminology for neuroscience concepts (53). NIFSTD is neither an atlas, nor is it tied to a particular spatial arrangement of brain regions, but is a collection of brain region labels and inter-relationships. It represents a general mammalian hierarchy of brain parts, as opposed to the other primate-centered ontologies like the Foundational Model of Anatomy ((27), FMA). NIFSTD is a formal ontology constructed through the import of community ontologies with specific extensions for neuroscience, covering the major domains of neuroscience (53,54). For community contributions, NIF maintains the

NeuroLex lexicon. An important feature of this project was to explicitly define all of the terms that are used to describe data (e.g., anatomical terms, techniques, organism names). The NIF gross anatomy module was largely based on the NeuroNames hierarchy (55-57) and recoded in the Web Ontology Language (OWL), but has been extensively modified through contributions to NeuroLex. NeuroLex serves as a community platform where those with minimal knowledge of building ontologies can still contribute their expertise.

Clearly, the diversity in the approaches discussed above is warranted, and even expected, in a field that advances as rapidly as neuroimaging. However, the need to interoperate with data labeled using these different approaches needs further investigation through both computational and ontological means. With the expectation that more such atlases will be constructed, it is also expected that new and existing software will implement the analytical framework(s) necessary to make use of new atlases. Aside from NIFSTD/NeuroLex, the aforementioned atlases are used as part of the every day research activities performed by neuroscientists, generally through the use of neuroimaging data analysis software. This software enables researchers to perform complex analyses with freely available open source software that makes it possible to reuse analysis methodologies that would otherwise need to be implemented by each lab conducting a given analysis. However, the results from these analyses are labeled with a heterogeneous set of neuroanatomical labels that makes correlating terms for information retrieval tasks challenging. Neuroinformatics can help mediate the correlation of terms across atlases, as well as address other information-related challenges. The next section provides an overview of neuroinformatics and issues related to the ever-evolving landscape of open source neuroimaging software and challenges with which researchers are faced.

2.3 Application of neuroinformatics

The goal of neuroinformatics is to coordinate efforts to standardize data formats, meta-data representations, and to recommend best practices for data acquisition, storage, analysis, and sharing across all scales of neuroscience – from genes to behavior. This area of applied informatics stems from a two-decade lineage of neuroinformatics projects rooted in the efforts of the Human Brain Project (HBP) (42,58). The landscape carved out by the HBP set the stage for neuroscience advances at all levels of analysis by putting open source analytical tools into the hands of individual investigators. Neuroinformatics tools enabled the neuroscientist without a background in software development to perform analyses that would not otherwise be feasible. The informatics of human cognitive neuroimaging, and specifically issues related to data management, integration, and sharing, has great potential to facilitate collaborative and reproducible research. A number of large scale projects painted the current neuroinformatics landscape and provided solutions to many data analysis and data management issues, as well as demonstrated successes and failures in data integration and sharing. What emerged from these initiatives is a neuroinformatics community interested in conducting informatics research and developing software to help accelerate the rate of discovery in neuroscience. By providing scientists with tools that ensure more robust research practices, neuroinformaticians are also paving the way for reproducibility in neuroimaging. The following sections discuss how large-scale initiatives shaped the current landscape of advanced neuroimaging informatics software and data management applications before turning towards community-based efforts to collaboratively coordinate the evolutions of the neuroscience cyberinfrastructure.

2.3.1 Advances from large-scale initiatives

Neuroscience research benefitted greatly from large-scale initiatives, such as the Human Brain Project (42), which helped to develop a vast assortment of informatics tools and cyberinfrastructure for the research community. By providing the resources to explore how information technology can facilitate neuroscience research, critical advances were made during the HBP that transformed our understanding of the brain and opened up a vast collection of new questions to explore. To answer questions about, for example, neuropsychiatric disorders

or cognitive functions using MRI technologies, it was necessary to enable a diverse set of researchers, that may or may not have a strong computational background (e.g., psychologists, clinicians), with neuroinformatics software that implements a set of analysis methodologies. For example, the FreeSurfer package is used widely to process anatomical scans by segmenting subcortical structures and parcellating the cortex into a number of regions using brain atlases like those discussed above (39). By providing an automated pipeline that provides a variety of statistics, such as anatomical structure volumes and cortical thickness, researchers can focus on biologically oriented science, rather than attempting to implement software themselves. There are many models of how to develop analysis tools for biomedical communities, such as human cognitive neuroimaging, that range from heavily funded and large-scale consortia to small grassroots efforts. Neuroinformaticians play an important role along the spectrum of projects and makes significant contributions to advancing the tools available to neuroscientists. Neuroinformatics software that is open-source and freely available further enhances adoption by users and provides a transparent view of how the code was implemented.

The development of tools can take time before they are ready to be released to the community, but when they are released the tools should be open source and freely available, particularly for publicly funded research. For example, CBRAIN (59) is a Canadian government initiative that provides access to high-performance computing facilities across Canada and the World. CBRAIN is also used to drive an online collaborative Web platform from which users control brain imaging data, compute, and visualize results using 2D and 3D Web browser applications. Currently, only pieces of the CBRAIN framework are open source (e.g., BrainBrowser is now open source, but was initially private), while the remaining platform is only available to collaborators. Similarly, The Laboratory Of NeuroImaging (LONI) offers a large selection of well-engineered data management (60), analysis (61), and visualization (62) tools that are free to the public, but not open source. The LONI infrastructure is also used to host the Alzheimer's Disease Neuroimaging Initiative ((7), ADNI), which is arguably the most successful data sharing effort in brain imaging. However, there is debate over whether or not the ADNI Data Use Agreement (DUA), which requires researchers to list the ADNI Consortium as the

senior author on any manuscripts, is the proper way to appropriate credit for data sharing (63). Both CBRAIN and LONI provide benefits to the community and many investigators rely on the tools they have created, but a lack of transparency may limit thorough evaluation and broad adoption.

The Allen Institute for Brain Science is another large-scale initiative that started out by working primarily on the mouse brain (64) and has recently made contributions to clinical research by developing an atlas of the human brain transcriptome (65). The Allen Institute demonstrated how wet lab science could be scaled up through the use of informatics tools that streamline workflow and add findings to a web accessible database. This institute is unique to neuroscience in that they make all of their data, including images, public through an Application Programming Interface (API) that supports a variety of queries and brain atlas functions. The human brain data available include microarray, anatomical MRI, and diffusion-weighted MRI datasets. The underlying infrastructure implementation details, however, are not publicly available, but the resources necessary to produce this research are unprecedented. While the internal operations of the Allen Institute are not fully transparent (i.e., source code is unavailable), other large-scale initiatives have provided more complete transparency in their internal operations.

At the large-scale and open source level is the Human Connectome Project (HCP) (66). The HCP is a major endeavor to acquire and analyze functional and structural connectivity data plus other neuroimaging, behavioral, and genetic data from 1,200 healthy adults. It is already a key resource for the neuroscience research community with several data releases already available, enabling discoveries about structural and functional connectivity. The HCP consortium developed an informatics platform that handles: 1) primary and processed data storage, 2) systematic data processing and analysis, 3) open-access data sharing, and 4) mining and exploration of the data. This informatics platform includes ConnectomeDB and the Connectome Workbench. ConnectomeDB is based on the eXtensible Neuroimaging Archive Toolkit (XNAT), which provides a database for storing and distributing the data, and executing data analysis pipelines. The Connectome Workbench provides visualization and exploration capabilities (67).

The projects discussed above provide only a small sample of efforts that have shaped the current landscape of neuroimaging cyberinfrastructure. With such a variety of implementation details in each of these projects, it is not hard to imagine why interoperability remains an important issue to overcome. While it is important for neuroinformatics developers and scientists to harness creativity and build novel solutions to complex problems, each of the solutions provided only offers a glimpse into the goal of understanding the human brain. If we are to put the pieces back together and form a holistic model of human brain structure and function, there needs to be a community that complements the efforts of scientists to integrate and interoperate across data and software.

2.3.2 Facilitating open collaboration and coordination

Advocates of open neuroscience who value methodological transparency and maximizing efficient access and exchange of information through collaboration, are actively contributing to the vision of Neuroinformatics, including many from the groups listed above. One organization that seeks to facilitate communication across borders and break down barriers to collaboration is the International Neuroinformatics Coordinating Facility (INCF). The role of INCF is to bring scientists, software engineers, and informaticians together to work on cyberinfrastructure projects that catalyze the integration and interoperability of neuroscience data across spatial scales. The INCF defines several programs that tackle crosscutting issues in neuroscience by targeting specific task force teams on a domain specific given topic. For example, the Program on Standards for Data Sharing includes a task force for Electrophysiology and Neuroimaging. By drawing on a growing number of member countries, currently sixteen, the INCF engages experts to volunteer in each task force by organizing satellite events at international congresses, as well as at the Neuroinformatics Congress each year. In this way, INCF interacts with many of the initiatives discussed in the previous section to help mediate communication over specific issues. With regards to reproducibility in neuroimaging and the motivations listed above, there are three relevant INCF programs:

1. Program on Ontologies of Neural Structures (PONS)

2. Digital brain atlasing

3. Standards for data sharing

The PONS task force facilitates the development of ontologies and vocabularies of neuroanatomy terms. PONS task force members collaborate to define properties and hierarchical relationships between neurons and brain structures, which are then captured in either the Neuron Registry (68), NeuroLex (69), or the Common Upper Mammalian Brain Ontology (CUMBO) (70). Much of this work is directly related to the brain atlas labeling section above; however, their mission is to focus on the lexical, rather than spatial, organization of structures and does not address the mapping of terms from one brain atlas to another.

The digital brain atlasing group addresses the spatial concordance issue related to brain normalization (71), as well as to integrate brain atlas resources. Neuroinformatics tools increasingly rely on brain atlases and standard coordinate systems in order to compare across subjects and investigate normal and abnormal brain structure and function. In particular, this INCF program developed a distributed query and image integration system to access several mouse brain resources, called the Digital Atlasing Infrastructure ((72), DAI). This program primarily focuses on web services for the mouse brain by providing a middleware layer to access a series of mouse brain atlases. The mouse brain atlases integrated into the DAI are accessible from an API that handles queries about the given atlas. The DAI provides a way to map each database API into a common object model, and register a set of services. An atlas with web services that conform to the INCF Digital Atlasing specifications is called a "hub", and currently there are four hubs in the DAI network. The Web service infrastructure proposed by the DAI provides one model for integrating atlas resources that may eventually be applicable to human neuroimaging data sharing.

The program on standards for data sharing is divided into the electrophysiology and neuroimaging task forces. The overall goal for this program is to capture metadata that represents the information needed to facilitate data integration and interoperability for easy data sharing. Of particular interest is the Neuroimaging Data Sharing (NIDASH) task force that works in the applied informatics space to identify barriers to data sharing and to propose

technical solutions that overcome these barriers. In an evaluation of the current state of neuroimaging data sharing, the NIDASH task force identified four projects targeted at easing the burden of data sharing, including 1) a "One-Click Share Tool" that allows researchers to make their data publicly available, 2) a common data model and API to facilitate data exchange and database interoperability, 3) a mechanism to capture data within a single data container, and 4) a means to automatically store metadata and processing stream results to a database (13). The overall NIDASH vision is to make data sharing a straightforward process that investigators would have no choice but to share; however, these ambitious projects do not address the many socio-technical issues that arise during deployment, such as community involvement and competing efforts.

This overview of INCF, particularly the NIDASH task force, provides a description of how applied informatics in neuroscience is conducted in a community-driven fashion. The methods employed by NIDASH and other INCF programs are not all specific to neuroscience, rather they draw heavily from core technologies and methods developed by other disciplines. For example, the ontology work on the Common Upper Mammalian Brain Ontology (CUMBO) also drew on efforts from more general human anatomy ontologies, such as the Foundational Model of Anatomy (27), thus highlighting the need for interaction with a broader community outside of neuroscience with more general solutions. The following section discusses the crosscutting methods from biomedical informatics that neuroinformatics draws upon and contributes to by examining the areas of data integration and interoperability, the semantic Web, and data provenance.

2.4 Biomedical informatics methods

Biomedical informatics (BMI) provides an interdisciplinary framework for integrating methods from a variety of disciplines and then specializing those methods for problems in a given biomedical domain of interest, such as neuroscience. The American Medical Informatics Association (AMIA) defines BMI as "the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem

solving, and decision making, motivated by efforts to improve human health" (73), which eloquently captures a domain agnostic version of many goals articulated by INCF and the discipline of Neuroinformatics. As its defining characteristic, Charles Friedman put forth a "fundamental theorem" of BMI (Figure 2.1), which can be interpreted as "a person working in partnership with an information resource is 'better' than that same person unassisted" (74). Along these lines, neuroscientists use a variety of information resources to augment their interaction with data using much of the cyberinfrastructure discussed in the sections above, thus working within the context of the BMI fundamental theorem. Whether it is the use of brain atlases, data management and integration technologies, or data analysis software packages, the methods employed by BMI work to enhance researchers' ability to communicate research findings, streamline research processes, and extract information and knowledge from biomedical data. The sections below discuss the BMI methods and component disciplines that I build upon in this dissertation, with the goal of providing the reader with a technical background in several relevant areas.



Figure 2.1 Fundamental Theorem of Biomedical Informatics. An individual performing a task in concert with an information resource, depicted as a world with people circling it, will perform a better job than the same individual acting alone.

2.4.1 Data management and integration

Biomedical data management evolution follows a trajectory of scale that starts with the use of readily available applications (e.g., spreadsheets) and may then progress to produce one-off solutions to meet their functional needs. While homegrown systems are perceived to be

beneficial, many of the researchers who create their own method for organizing information readily recognize the shortcomings for certain tasks, such as data retrieval or exchange between labs, and are not well equipped with technical expertise or finances to adopt more robust data management practices (75). Neuroimaging data management and integration solutions evolve out of a need for more streamlined approaches to doing high-quality scientific research.

Streamlining workflow helps to address quality control and scalability issues, where labs will generally start off using Microsoft Excel and eventually need to implement special macros to, for example, validate data entry before moving on to something more robust, such as Microsoft Access. However, the resources that labs spend on scaling up operations organically will eventually become unmanageable, particularly in the context of multi-site research consortia that require data to be integrated from each site. A more practical approach would be to use an open source data management system from the start and eliminate scalability issues up front. For example, in a review of clinical data management systems by Franklin et al., the REDCap system (37) was found to be preferred by users among several other systems with similar features (76), which would suggest to knowledgeable researchers to utilize a proven data management solution rather than expend resources developing a new system. Similarly, managing medical imaging data can hit scalability barriers; however, implementing a broadly used research Picture Archiving and Communications System (PACS), such as the extensible Neuroimaging Archive Toolkit ((34), XNAT) may alleviate these issues. Consequently, the overhead for such systems may only be warranted if a lab is expecting to grow rapidly or if the institution where the research is being conducted administers the system. Although adopting such systems may alleviate scalability issues, no open source clinical or imaging data management system exists that provides a combination of support for Digital Imaging and Communications in Medicine (DICOM), a medical imaging standard, and easy-to-use data entry forms, nor do any two systems interoperate seamlessly.

Achieving a seamless level of interoperability between information systems necessitates a data exchange format, particularly when data are shared across labs. A common misconception

is that information residing in a database with a defined schema and a set of terms is unambiguous. Without a data catalog or dictionary of each term and data structure, it becomes impossible for users to ascertain the precise meaning of information exported from a data management system, thus limiting the data's overall utility and ease of reuse and integration. It is also common for the same term to have multiple definitions that depend on context. For example, handedness can refer to a simple "right/left/ambidextrous" checkbox or a structured assessment that provides a handedness score, such as the Edinburgh Handedness Scale (77). To make data sharing meaningful the distinction between such concepts needs to be made explicit by providing precise definitions for terms, ideally accessible at a URL on the Web. Biomedical data integration approaches commonly apply standard vocabularies or ontologies and data models to harmonize disparate databases (78,79), including examples from brain imaging (28,80,81). However, these proposed solutions are not as widely adopted in cognitive neuroimaging as they are in fields like genomics, possibly because the data sharing culture in human neuroscience is lacking. Fortunately data sharing is starting to take hold by the virtue of early advocates demonstrating the utility of sharing and large scale data aggregation (6,82). With more brain images available and growing interest in the research community, the time seems ripe to push initial data integration efforts forward and fill in the missing gaps with tools and services researchers are now keen on using.

With access to more robust tooling, secondary use of research data can be made more efficient through the use of standards or recommendations. Many examples exist that demonstrate how standards have enabled neuroimaging researchers to conduct their work in more streamlined fashion. One successful example from neuroimaging is the Neuroimaging Informatics Technology Initiative (NIFTI) standard for representing binary imaging data along with minimal metadata. The NIFTI developers implemented this standard in several data analysis packages, which enabled researchers to load the same files into multiple software packages for interoperable analysis and visualization of medical images. From a users perspective, interoperability is a key component that enables flexibility in the choice of software, where some software outperforms others for a given task. In a field with rapidly

evolving methods, such as neuroimaging, it is crucial for investigators to move seamlessly between software applications and choose the optimal tool for the task at hand. An analogy can be seen in the use of workflow/pipeline tools (e.g., Nipype (83), PSOM (84), LONI Pipeline (85)) that provide a common interface between analysis software, some of which even converts between formats at runtime to enable interoperability. Similarly, agreed upon standards for representing metadata about datasets and derived data can enable interoperable data management systems.

To date, the data models used in data management systems have focused on a hierarchical syntax that maps well to the relational database world; however, the heterogeneous schemas utilized by leading neuroimaging databases have required significant data integration efforts. For example, Ashish et al. developed a database mediator to integrate XNAT and the Human Imaging Database ((86), HID) and provide a harmonized view of the data (81). Their system used a series of complex rules to rewrite database specific queries that were shipped out to each database. The query results were returned as XML documents consistent with the XML-based Clinical Experiment Data Exchange ((87), XCEDE) schema. XCEDE served as a data model standard/specification for the exchange of scientific data between databases, and it provided a structured metadata hierarchy for storing information relevant to various aspects of an experiment (e.g., project, subject, etc.). While the Ashish et al. approach demonstrated the utility of XCEDE, the extensive configuration details made it impractical for wide spread adoption, thus not scalable. However, the XCEDE schema demonstrated several key components for effective data exchange in neuroimaging.

XCEDE provides an extensible core structure for describing a hierarchy relevant to clinical experiments. By using these features, XCEDE enables the representation of metadata that spans the experimental process from project titles and subject demographics to data acquisition, image header information, and data provenance. The XCEDE Experiment Hierarchy represents a set of information common to many brain imaging studies, which is also reflected in other neuroimaging data management systems, such as those mentioned above. In particular, information about Projects, Subjects, Studies, and Acquisitions are prevalent pieces of

information that are captured across neuroimaging databases and ripe for metadata harmonization efforts. Although XCEDE was never as broadly implemented as the NIfTI image format, it demonstrated the capabilities and shortcomings of information modeling with XML Schema. While effective at describing the syntax and structure of data, XML Schema was not designed to capture the semantics, or meaning, of data elements in a format that readily facilitates data integration. Semantic Web technologies were designed to extend XML in order to address these shortcomings and now mature tools are now available that make their use practical in a research setting.

2.4.2 Semantic Web technologies




The Semantic Web is a vision originally articulated by the inventor of the World Wide Web, Tim Berners-Lee. The semantic Web provides a framework for a Web of both machine-readable data and knowledge, as well as in addition to human readable documents (88). The technology stack to support this vision builds upon XML with a standard data model (Resource Description Framework, RDF, <http://www.w3.org/RDF/>), query language (SPARQL Protocol And RDF Query Language, SPARQL (89) and ontology language (Web Ontology Language, OWL (30)) that are designed for loosely coupled, distributed systems that intrinsically support data sharing. However, these tools are designed to be flexible and general, thus they do not provide the solution to any specific problem alone, rather they provide a toolkit on which to build modern and scalable systems. By adopting Semantic Web standards as a base framework and working from the conceptual foundation provided by XCEDE, a next generation specification suite can be defined for neuroimaging that captures elements from the experimental process in a fashion that facilitates data exchange and models provenance in a reproducible fashion. From the RDF specification (<http://www.w3.org/TR/2002/WD-rdf-concepts-20020829>), the following quote provides a provocative statement about the potential that this technology stack enables:

"The real value of RDF comes not so much from any single application, but from the possibilities for sharing data between applications. The value of information thus increases as it becomes accessible to more and more applications across the entire Internet."

Tracking data provenance and capturing structured metadata only enhances the utility of shared neuroimaging data if it can be indexed and discovered. Linked Open Data (LOD) refers to a set of rules that describe how data behaves when the Semantic Web is "done right," thus allowing data to be traversed in much the same way as one following links while browsing the Web (90). The W3C Design Issues page itemizes the four rules for LOD (<http://www.w3.org/DesignIssues/LinkedData.html>) as follows:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Each rule adds an additional layer of accessibility and interconnectedness that is designed to broadly link structured data, which can then go on to be used in unexpected ways. Additionally, the W3C developed a 5-Star LOD rating scale that describes how well a given dataset conforms to the above rules (Table 2.1), as well as other necessary requirements (e.g., an open license and non-proprietary formats). In the context of provenance information, the LOD principles are foundational to representing scientific information in a transparent way that facilitates reproducibility, which can be extended by a data model for describing provenance on the Web.

Number of Stars	Expected Behavior
	Available on the web (whatever format) <i>but with an open license, to be Open Data</i>
	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
	As (2) plus non-proprietary format (e.g. CSV instead of excel)



	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
	All the above, plus: Link your data to other people's data to provide context

Table 2.1. 5-Star Linked Open Data rating system (<http://www.w3.org/DesignIssues/LinkedData.html>).

2.4.3 Data provenance information

Data provenance provides an explicit record of events or activities that lead to the generation of a new piece of data, information, or knowledge. With workflow tools, a provenance record can include details about software versions, computational environment, and specific parameters used during an analysis. Analysis tools take time to mature, and it is not uncommon for errors or bugs to be discovered in software that impact the measures scientists use to publish findings (40,91). When these kinds of errors are identified, a structured provenance record provides a computational way of identifying impacted research. The need to capture structured data provenance arises from complications in reproducing research findings, particularly as the complexity of analyses has increased in recent years. The W3C designed the PROV specifications to capture provenance information on the Web and interoperate with semantic Web technologies discussed in the sections above. Among the PROV specifications is the PROV Data Model (PROV-DM, (92)) that defines a core set of high-level structures for capturing provenance information that bolsters trust in how a given piece of information was generated.

The PROV specification details three core objects that are used to describe provenance, which are Entities, Activities, and Agents (Figure 2.2). These core objects are related to each other with a set of defined relations. Entities are used to capture information that tends to persist over time, for example, a dataset or spreadsheet file, which can be modified or derived from other entities. For an Entity to be created (e.g., a file), an Activity is needed to describe how the Entity came into existence, and an Agent is needed to describe who (e.g., a person) or what (e.g., an organization or software) is responsible for generating an Entity. The binding of these three objects provides the structure needed to trust how and who created the file, dataset, or analysis. As a standard model for tracking provenance, PROV is more constrained

than RDF, yet too generic to represent the specific types of information common to neuroimaging in a meaningful way. By augmenting the basic types provided by PROV with domain-specific brain imaging terms, the PROV standard can be extended to meet the needs of the brain imaging community for representing provenance.

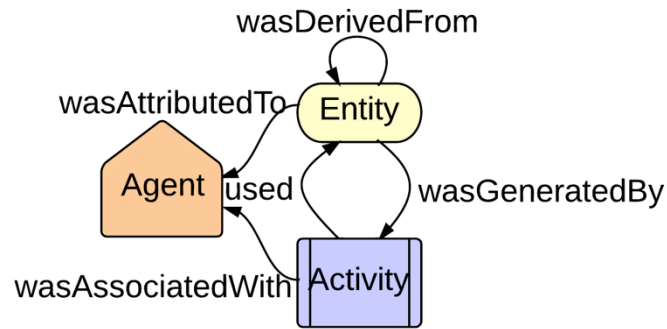


Figure 2.2. PROV Data Model core constructs. The three core PROV structures are Entity, Activity, and Agent, as well as a set of directional relationships.

2.5 Summary

To summarize, the computational landscape within cognitive neuroimaging is rapidly changing. As the neuroimaging community advances knowledge about the brain, it also identifies methodological flaws in how data is analyzed and software implementation errors. With no signs of slowing down, this rapidly changing scientific environment calls for agile approaches to data analysis that need to be constantly refactored, which greatly impacts the efficiency in which research can be done. Neuroinformatics offers relief to some of these challenges, while introducing the burden of learning new software tools. The area of data sharing shows great promise as one area that neuroinformatics can offer more relief than burden by providing standards that can help software developers make design decisions that facilitate interoperability. Biomedical informatics provides the organizing framework around which neuroimaging can interact with the broader biomedical community and incorporate general informatics methods. The next chapter describes the study design for building a data sharing framework that enables a community to collaborate in the development of a data exchange standard, and then use that standard as part of a scientific workflow.

Chapter 3: Study design

"I think IT projects are about supporting social systems-about communications between people and machines. They tend to fail due to cultural issues. "

- Tim Berners-Lee

3.1 Overview

In Chapter 1, the challenges of heterogeneous brain atlas terminologies, data management schemas, and computational analysis software were introduced as motivating factors behind this investigation into reproducible research. Chapter 2 provided background information on these challenges by examining specific cases in a biomedical domain of interest (i.e., cognitive neuroimaging), as well as in applied and basic informatics disciplines. Chapter 2 also introduced the "Fundamental Theorem" of biomedical informatics that states "a person working in partnership with an information resource is 'better' than that same person unassisted," which forms a basis for defining informatics research goals and evaluation (74). In this chapter, the discussion turns toward research goals and evaluation by introducing the general research approach and study design of two specific aims 1) to research and design a framework to represent, access, and query neuroimaging provenance information and 2) to develop an information system of Web services to compute and discover data provenance from brain imaging workflow. To achieve these aims, three communities of stakeholders participated in a research process that took place over two phases through a series of workshops and weekly meetings. The work completed during the first phase provided the preliminary results and insight necessary to address the specific aims in the second. The general research approach used to organize these two research phases is presented in the following section, followed by the study design details for each specific aim. I was a major instigator in all of the following activities, and my specific contributions are described in Chapter 6.

3.2 General research approach

To investigate the design and implementation of a framework for reproducible research, a mixed approach was chosen that engages stakeholders from three interdisciplinary scientific

communities: cognitive neuroimaging, neuroinformatics, and biomedical informatics. Each stakeholder community has a unique set of skills, information needs, and available resources that needed to be identified and incorporated into the development lifecycle of this framework. Without direct interaction with each stakeholder community there is an increased risk that previously identified solutions would overlook or that use cases would not reflect the needs of system users. Conceptually, Figure 1 captures the general contributions and dependencies between each community-stakeholder pair, which is modeled after a guide developed by leaders in the biomedical informatics community to clarify synergistic interactions among stakeholders in biomedical informatics research (73). With this model in mind, partnerships were established with stakeholders in representative labs and organizations in order to iteratively identify requirements for the framework design and implementation over two phases. In phase one, partnerships were formed *a priori* to address the need for image-specialized database tools through a collaboration between basic informatics in the Structural Informatics Group (SIG, <http://sig.biostr.washington.edu>) at the University of Washington (UW) and cognitive neuroimaging experts located in the Integrated Brain Imaging Center (IBIC, <https://www.ibic.washington.edu>), also at UW, and the Center for Cognitive and Neurobiological Imaging (CNI, <https://cni.stanford.edu>) at Stanford University. The phase two partnerships were formed *ad hoc* with applied informatics experts at the International Neuroinformatics Coordinating Facility (INCF), which shared the common vision of a reproducible research framework. Given a climate of constrained resources, there was a pragmatic need to work collaboratively towards a mutual goal of a standards-driven framework of interoperable tools for data sharing and reproducibility.

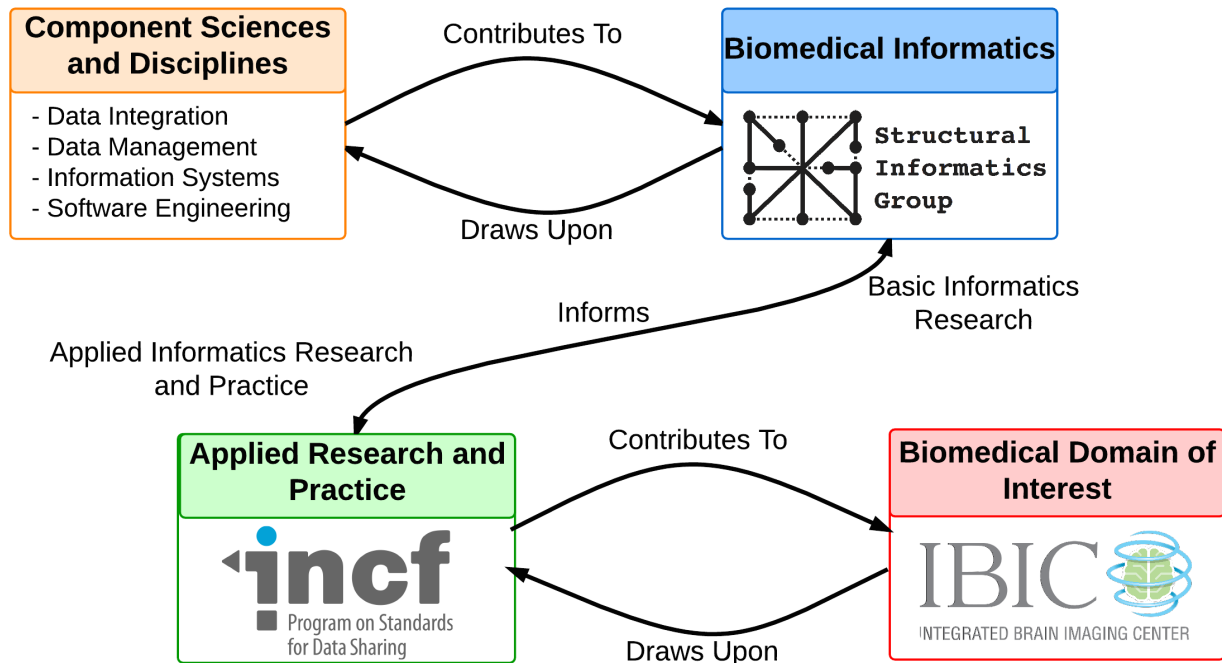


Figure 3.1. Biomedical Informatics Perspective. A conceptual framework to organize the interaction between cognitive neuroimaging, neuroinformatics, and biomedical informatics biomedical informatics stakeholders from the, respective, biomedical domain of interest (IBIC), applied informatics research and practice (INCF), and biomedical informatics (SIG) organizations and labs.

The partnerships with stakeholder communities took place over three years and are divided into two overlapping phases: 1) Scalable Neuroimaging Initiative (SNI) and 2) INCF Neuroimaging Data Sharing (NIDASH). During each phase, a series of weekly meetings and workshops were completed to facilitate the framework development lifecycle. The key events are summarized as a timeline in Figure 3.2. The SNI research phase was an initial effort to integrate the expertise of stakeholders in two representative human cognitive neuroimaging research centers with a biomedical informatics lab to assess the need for image-specialized database tools and a develop prototype application. A National Academies Keck Futures Initiative (NAKFI) conference on Imaging Science in 2010, where Dr. Thomas Grabowski worked with an interdisciplinary team challenged to “develop image-specialized database tools for data stewardship and system design in large-scale applications,” catalyzed this phase. The team identified Neuroimaging and Cognitive Neuroscience as areas ripe for innovation, where evolving instrumentation, evolving conceptualization of brain systems, and differentiation of modalities and analysis approaches promised to be disruptive factors for the foreseeable

future. While the team concluded that basic informatics concepts are available, they indicated a lack of interdisciplinary interaction between those with specialized scientific knowledge and those with specialized informatics knowledge. Pilot funding for SNI was provided by NAKFI to design and develop a system architecture prototype that addressed the expertise gap, which was then evaluated and redesigned as phase 2.

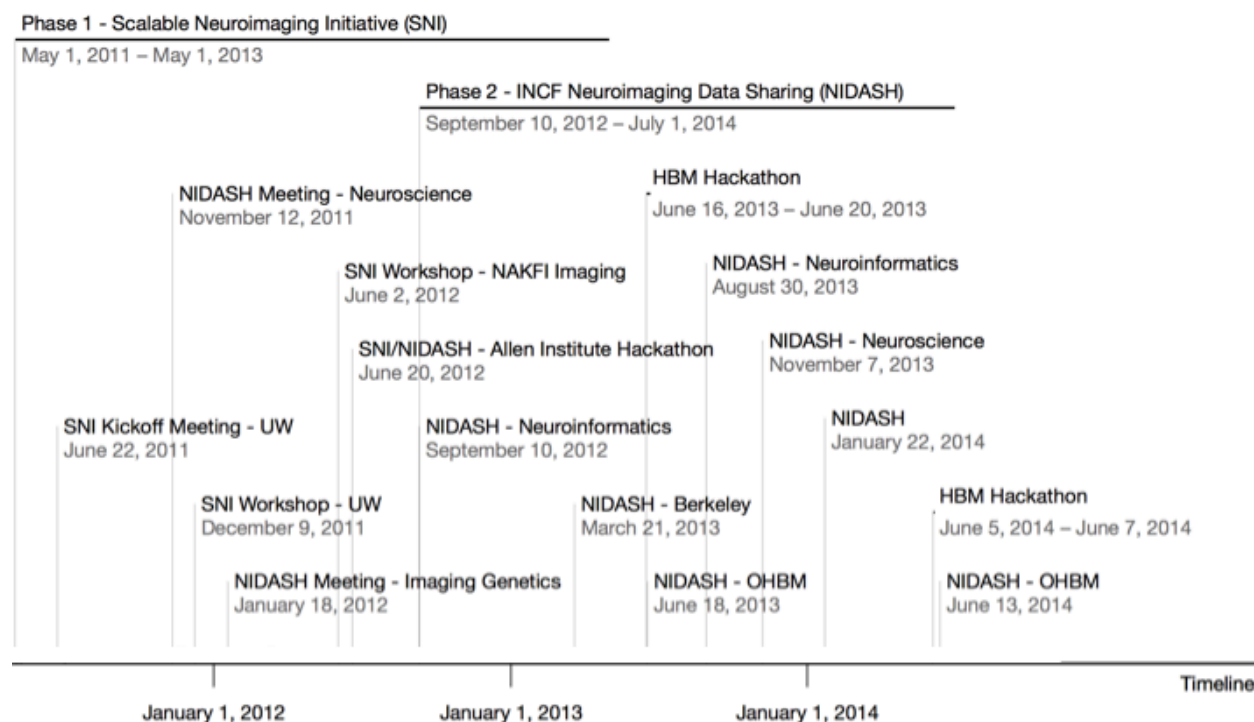


Figure 3.2. Research Phases and Events Timeline. Highlights the key workshops and meetings for the SNI and NIDASH research phases.

The NIDASH phase was motivated by feedback on the prototype software developed during the SNI phase, which was presented at the 2012 Neuroinformatics Congress (93) and garnered interest from the INCF Data Sharing Program. The SNI team decided to engage the broader Neuroinformatics community and leverage the limited resources afforded by the grant from NAKFI with those of INCF. By working with fellow Neuroinformatics researchers and software developers in a grassroots fashion, the SNI vision and prototype software were refactored and harmonized with community-based efforts to accomplish similar goals with a broader set of expertise. Further, INCF made resources available to participants working on standards related efforts to present work at conferences and attend workshops. These

contributions made it possible to bring community members together to extend and refine the work completed during the SNI phase, and make a real world impact by encouraging adoption of SNI concepts in the field of cognitive neuroimaging. The specific aims are tailored to a participatory design and development process that builds on the preliminary results established by SNI by encouraging iterative and agile methods that favor a motto of "fail early, fail fast, fail often."

The specific aims in this dissertation are driven by the needs of stakeholders in the human cognitive neuroimaging community, which were initially established during the SNI phase. There was also a need for participation from stakeholders in the neuroinformatics community with expertise in addressing reproducibility in neuroimaging research. To address this broad issue, research questions were identified to climb the so called "tower of achievement in biomedical informatics", which progresses from "Model Formulation" and "System Development" to "System Installation" and "Study of Effects" (94). The first aim addresses the question, "What are the requirements and design of an information resource that improves the ability of a cognitive neuroimaging scientist to conduct reproducible research?" By investigating this question and identifying the specific "reproducibility needs" of this research community, an information resource could be designed to meet those needs. Aim 1 also addresses the design question, "What processes, models, and/or systems can be designed to meet the needs identified." Aim 2 turns towards the system development and installation levels of the "tower of achievement" to create an information resource artifact that answers the question of "what open source software and services are available and appropriate to implement prototypes of the framework design?"

These aims follow an approach modeled after agile software development (95), where incremental progress is used to provide feedback that may change the course and direction of previous aims. The intention of this design is to facilitate vertical integration across each "tower of achievement" level and incorporate best practices that have emerged from the bioinformatics community's use of agile approaches in other contexts (96,97). Each specific aim also incorporates an evaluation sub-aim that addresses the "Study of Effects" level and is designed

to address questions such as, "Does the framework design meet the specified requirements and will it function as expected?" and "What is the overall impact that the framework has had on being able to conduct reproducible neuroimaging research?" This general approach is designed to enrich shared brain imaging datasets with structured provenance information that will lower barriers to new explorations of neuroimaging data on the Web.

To advance the progress being made in the neuroimaging data sharing space, working examples needed to be developed and distributed in an open and community -driven way. The scope was structured to provide prototype standards and applications using open access datasets, databases, knowledge bases, and computational services to demonstrate novel data sharing standards and resources. The following specific aims sections provide further details about these aims and their sub-aims.

3.2 Specific Aim 1: Research and design a framework to represent, access, and query neuroimaging provenance information

The goal of this aim was to form a bridge between scientific and informatics disciplines to lower data reuse barriers. A specifications suite was designed to provide the data structures and terminologies necessary to aggregate and enrich public brain imaging metadata with provenance information. Standardized provenance information is not supported by present-day neuroimaging databases. Such provenance information is an essential component of reproducibility that ties together a range of information from different research stages such as raw experimental data, analysis workflow, and statistical results. By identifying the terms used in each research stage, a model can be constructed that captures relevant information as an extension of the PROV Data Model discussed in Chapter 2. The model for each research stage can then be linked together as a single flow of information, where apps can be designed specifically for each research stage. If the apps use this same provenance model and vocabulary, they would then be interoperable and able to exchange common sets of information.

To enable the generation, aggregation, and query of provenance information from neuroimaging data management systems and computational workflows, four sub-aims were

proposed that led to the design and demonstration of specifications for 1) a vocabulary of terms for cognitive neuroimaging metadata, 2) extensions to the PROV Data Model that correspond to the vocabulary of terms, 3) the design of an API to access and query a provenance information repository, and 4) evaluation of each specification with use-case driven demonstrations. These specifications provide the design blueprint for developing an information system of Web services to compute and discover provenance information from brain imaging workflow, as well as a process for community feedback and contribution.

3.2.1 Design and methods

To successfully design the specifications in Aim 1, an interdisciplinary working group of individuals with necessary technical expertise and experience was assembled. Oversight of the working group was provided by INCF, which was comprised of members from SNI, INCF NIDASH Task Force, and the biomedical informatics Research Network (BIRN) Derived Data Group. Working group members during eleven in-person workshops and forty remote weekly meetings, as well as informal discussions between team members, designed the specifications for each sub-Aim. A model and process were defined to enable distributed collaboration and to track contributions from team members as the data modeling and specifications were developed. The outcomes for each sub-Aim indicate specific deliverables and expected results.

Outcomes for the four sub-Aims were split into two types of deliverables. The first type of deliverable consisted of draft specification documents to capture cognitive neuroimaging metadata terms, related PROV extensions, and an API for access and query. The working group priorities and resources drove which terms were identified, vetted, and incorporated into specification documents. The second type of deliverable was an evaluation of the specifications based on specific use-cases that were identified by the working group to fulfill sub-aim 4. The evaluation includes descriptive statistics on the number of metadata terms, classes, and properties defined for each use case, as well as demonstrations of how the terms are used with PROV extensions. The API specification was evaluated with a prototype web application developed in specific Aim 2.

This specific Aim had a number of limitations. Modeling neuroimaging metadata required a deep understanding of the problem domain and priorities needed to follow use-cases defined by domain experts. The decentralized, grass-roots nature of the working group explicitly allowed for flexibility in pursuing new metadata terms, access and query requirements, and brain atlas vocabularies as necessary, making it difficult to predict modeling priorities. The PROV Data Model was chosen because it is general, flexible, and extensible, and thus allowed for an ongoing and iterative development approach amenable to changing priorities. This includes an expectation that the working group may change use-cases in the event of new contributors with additional resources, which is an unavoidable consequence of any grassroots effort.

3.4 Specific Aim 2: Develop an information system of Web services to compute and discover data provenance from brain imaging workflow

The second aim of this dissertation was to build software tools that implemented the specifications designed in Aim 1 and to use these components to develop a data management system architecture to compute and discover data provenance from brain imaging workflow. While the products of Aim 1 included specifications for terms, data model extensions, and an API, this aim turned to the system architecture level to develop an information resource that can be evaluated using feedback from the neuroimaging and neuroinformatics communities. The preliminary feedback received on the framework components (e.g., design process, data model, system architecture, etc.) from the SNI Phase were used to prioritize and evolve system components to meet the needs of stakeholders, including INCF.

To refine the initial framework components from the SNI Phase into an information system for executing and tracking the provenance of neuroimaging workflow four sub-aims were proposed: 1) Create a reference implementation of the vocabulary, data model, and API specifications; 2) Develop an updated data management architecture of the NiQuery system; 3) Implement a prototype Web application to demonstrate system features for query, workflow execution, and provenance discovery; and 4) Evaluate the overall positive and negative

outcomes of the framework implementation. The next section describes the design and methods used to implement and evaluate the framework components.

3.4.1 Design and methods

To fulfill the above sub-aims, an evaluation was conducted of a prototype data management system developed during the SNI Phase. The system was evaluated from the experimental research perspective of biomedical informatics, where an experiment is the development and demonstration of a biomedical information resource that is assessed to identify positive and negative outcomes of deploying the system (98,99). The outcomes are then used in a subsequent redesign of the resource, which then becomes another experiment, thus resulting in continued improvement. The evaluation was conducted based on requirements distilled from meetings and workshops with members from the neuroimaging and neuroinformatics communities, in parallel with those discussed above in Aim 1. The requirements were assessed in the context of emerging technologies that were used to redesign the original prototype system architecture, called NiQuery (93), and fulfill the four sub-aims.

Outcomes for the four sub-aims were split into three categories including features, applications, and an evaluation of the implemented system. The feature outcomes were defined as a demonstration of the vocabulary, data model, and API specifications for a neuroimaging data sharing scenario discussed in Chapter 4. The application outcomes were defined as an updated system architecture diagram and a working prototype for example use-cases. Members of the working group collaboratively defined use-cases to demonstrate system features for query, workflow execution, and provenance discovery. To evaluate the overall positive and negative outcomes of the framework implementation, the system is compared to the set of requirements necessary to fulfill the use-cases.

The components of this aim are implementation heavy and have several limitations. Redesign and implementation of a project using open-source and novel technologies takes significant technical expertise to execute in a timely fashion. To address the amount of effort needed to complete the overall framework, a minimal set of functionality may need to be

identified such that it would yield a working system. By specifying a minimally sufficient set of features, the full set of specification can be pending even as a working system with demonstrable features that can be evaluated. Additionally, it is reasonable to expect contributions from working group members and stakeholders seeking to leverage or influence the direction of the framework and system architecture.

3.5 Summary

This chapter presented the general approach and specific aims of the dissertation project. The general approach outlined a framework for biomedical informatics designed to facilitate collaboration among key stakeholders, which took place in weekly meetings and face-to-face workshops summarized in Figure 2. The goals, design, and methods for two specific aims were then discussed, including a brief description of the sub-aims that will be elaborated in the respective research approach sections in Chapter 4. The focus of Aim 1 was to produce specifications for cognitive neuroimaging terms, term mappings to the PROV standard for representing data provenance, and an API for access and query of each data model component. The focus of Aim 2 was to evaluate a system prototype and re-design the system to incorporate the specifications from Aim 1 and implement a working prototype. Successful realization of these aims will lead to tools and standards that will accelerate the meaningful reuse of brain imaging data by providing methods to capture, access, and query previously unavailable provenance information, as well as preserving the computational details and environment necessary for reproducible research in cognitive neuroimaging. The next chapter describes the studies and research outcomes for each of these aims and their corresponding sub-aims.

Chapter 4: Neuroimaging data model framework

"Just as bioinformatics has profoundly impacted molecular biology and related fields, neuroinformatics has the potential for a transformative impact on neuroscience research."

- David van Essen, Washington University

4.1 Overview

In Chapter 3, the general research approach for this dissertation was discussed in the context of a system for collaborative Biomedical Informatics research. The study design was introduced as a collaborative endeavor that engaged stakeholders from cognitive neuroimaging, neuroinformatics, and biomedical informatics through a series of workshops and meetings to address two specific aims during two research phases. The first research phase provided the preliminary results that phase two expanded upon, in order to produce specifications for neuroimaging data sharing and a prototype system that implements these specifications. In this chapter, a study is presented that builds upon the phase one results, reported in the research approach section, to improve the reproducibility of cognitive neuroimaging research through adoption of shared data models. A neuroimaging data sharing scenario is presented to describe the vision for how a scientist may interact with neuroinformatics resources that are enabled by the overall data sharing framework. This data sharing scenario is used to provide context when reporting outcomes and to demonstrate progress on key components of the vision.

4.2 A neuroimaging data sharing scenario

A vision for reproducible human cognitive neuroimaging research is emerging through a growing number of neuroinformatics resources designed to streamline scientific workflow. A data sharing scenario is presented that embodies this vision to enable neuroimaging researchers with neuroinformatics applications using a new data interchange standard called the Neuroimaging Data Model (NIDM) and a new computing platform, NiQuery. In this scenario, a second-year graduate student in cognitive neuroscience is rotating through a lab that investigates the functional neuroanatomy of auditory systems in the human brain. After

meeting with the lab's principal investigator (PI), the student identifies a project that will introduce them to neuroimaging data analysis methods starting with anatomical MRI and later moving on to functional MRI. In the first stage, the student is tasked by the PI to conduct a review of brain imaging literature, identify an open access dataset with anatomical and functional scans, process the anatomical data using FreeSurfer (100), and present the resulting derived data in the weekly lab meeting.

The student first visits PubMed to conduct a search on auditory system studies that returns over a thousand manuscripts. Even after filtering for studies based on specific interests, the student recognizes the need to further consolidate the literature. Luckily, a postdoctoral fellow in the lab recommends that the student explore Neurosynth (101), a neuroinformatics application that extracts activation foci from task-based fMRI manuscripts and performs an automated meta-analysis for different types of studies (e.g., auditory studies). The student then follows a link from the Neurosynth application to Brainspell (<http://brainspell.org>) where they are able to manually curate and annotate the studies in their literature review with terms from the Cognitive Atlas (102). With literature review complete, the student moves on to identify open access datasets and follows a link from Brainspell to NeuroVault (<http://neurovault.org>), an application for hosting unthresholded statistical maps that result from fMRI analysis. NeuroVault enables the student to explore the results from individual studies in their literature review and to identify studies that provide access to raw data they will need to run FreeSurfer. Of particular interest to the student are studies with an accompanying "data paper" that provides a detailed account of the methods used for data acquisition, which go beyond the level of detail typically covered in the methods section of a manuscript (103,104).

The student became familiar with the concept of data papers from their cognitive neuroscience training, which included a course on reproducibility that emphasized the importance of publishing data papers and using open access data. One relevant study on NeuroVault provided the student with a link to a Nature Publishing Group's open source publication journal called "Scientific Data" (<http://www.nature.com/sdata>) where the primary article type is a "Dataset Descriptor" that allows public description of valuable scientific

datasets and is intended maximize the utility such datasets. The Dataset Descriptor manuscript the student found was carefully curated with metadata represented in the ISA-Tab format (105), which is organized into files that represent three levels: Investigation, Study, and Assay. While ISA-Tab provides a standardized, machine-readable format for describing the layout and organization of a study, it does not provide many of the details expected to accurately curate neuroimaging research, such as an fMRI study (20). The fMRI Dataset Descriptor manuscript found by the student was published in Scientific Data and called the "Study Forrest Project" (8), which includes an auditory paradigm. The manuscript provided human-readable text that fulfilled the majority of the guidelines proposed by Poldrack et al. (20); however, the machine-readable ISA-Tab representation was not designed to capture the necessary details for neuroimaging.

To gain access to these additional details, the student follows a link to the Study Forrest dataset hosted on the OpenfMRI database (106). In addition to an ISA-Tab description, OpenfMRI provides a neuroimaging specific format called NIDM that captures detailed metadata the student will need for processing, particularly downstream fMRI analysis. Unbeknown to the student to this point, each of the links between Neurosynth, Brainspell, NeuroVault, and OpenfMRI were all created using NIDM in the background to seamlessly exchange information between the systems. Given the student's limited experience with computing environments such as Linux and command line tools, the same clever postdoctoral fellow recommends the student to use NiQuery, a neuroinformatics computing platform that can use NIDM to automate many data management and computational tasks. By accessing NiQuery, the student is able to chose the Study Forrest project on OpenfMRI and select the option to run FreeSurfer. Without the student needing to think about how the data is accessed or where the computation will take place, NiQuery reads the NIDM description hosted on OpenfMRI and automatically identifies the correct anatomical scans needed as input and triggers a FreeSurfer task for each scan. The student is informed that the task will take approximately one day to complete, based upon the time the task has taken to run previously, and then subscribes to a notification system that will send them an email when the task completes. When the student checks their email the next

morning, there is a message waiting from NiQuery indicating that their task is complete. After clicking on a link, they are presented with a webpage view of a NIDM Results document with several buttons to open NIDM-compliant data visualization environments, for example an IPython Notebook (<http://ipython.org/notebook.html>).

The student clicks the IPython Notebook button, which opens the application and loads the NIDM document. The student is free to explore interactive plots and 3D rendering of the FreeSurfer derived data. For example, the NiQuery FreeSurfer workflow generates statistics for brain regions annotated with terms from the Foundational Model of Anatomy (FMA) ontology (107). The semantic annotations provided by the FMA enables the student to perform intelligent queries of the data (28). The student explores several interesting queries and generates a few figures using the tools provided in the visualization environment. During the next weekly lab meeting, the student is able to present findings and get suggestions on the next steps to take when moving from anatomical to functional MRI data analysis with the Study Forrest Project. The NIDM data interchange standard and NiQuery platform are two essential components to enable this vision of workflow for students and researchers alike. The studies presented here and in Chapter 5 explore the requirements and implementation details of a framework that realizes this vision.

4.3 Framework research and design study

In this study, the goals of Specific Aim 1 were addressed to research and design a framework to represent, access, and query neuroimaging provenance information. The four sub-aims proposed in Chapter 3 were pursued and led to the design and demonstration of specifications for 1) a vocabulary of terms for cognitive neuroimaging metadata, 2) extensions to the PROV Data Model that correspond to the vocabulary of terms, 3) the design of an API to access and query a provenance information repository, and 4) evaluate each specification with use-case driven demonstrations. The work supporting these sub-aims was completed during the Scalable Neuroimaging Initiative (SNI) Phase, which is described in the following section on research approach. The outcomes of these preliminary studies informed the development of the

Neuroimaging Data Model (NIDM) data interchange framework that was alluded to in the above data sharing scenario. In the outcomes and design section, the details of the NIDM framework are presented with examples that follow from the data sharing scenario with a use-case driven demonstration of modeling OpenfMRI and the Study Forrest project. This description of the NIDM framework provides an overview of the standards used for data interchange in the second study to develop a prototype information system called NiQuery.

4.3.1 Research approach

This section elaborates on the study design introduced Chapter 3, Aim 1. It is organized by sub-Aim, according to metadata terms, PROV extensions, and the API. The section on metadata terms describes 1) a method to represent and harmonize different views of neuroanatomy using the Foundational Model of Anatomy (FMA) (27) and 2) a method to capture MRI acquisition parameters using the controlled vocabulary of the DICOM standard (108). The section on PROV extensions describes initial work to harmonize the XML-based Clinical and Experimental Data Exchange (XCEDE) (87) with the PROV XML Schema for representing data provenance. This endeavor provided the key insight that all experimental data could be modeled as a provenance information flow in NIDM. Finally, the section on the API presents initial experiments with Web service API's for voxel-level access to brain imaging data that are annotated with terms from the FMA. The outcomes reported in each section provide a summary of the lessons learned and requirements determined to be important for the NIDM framework that is subsequently described in the outcomes and design section.

4.3.1.1 Cognitive neuroimaging metadata terms

A basic requirement for interoperability between biomedical information systems is a controlled vocabulary of terms that provides a common meaning for data. Two categories of terms in neuroimaging research that need to be captured unambiguously are MRI acquisition metadata and neuroanatomical labels. Acquisition metadata for MRI studies is captured using the DICOM standard, which has been implemented by all major medical imaging vendors.

However, perhaps surprisingly, there is no machine-readable resource providing a dictionary of all DICOM terms that includes definitions and datatype information. Similarly, the proliferation of brain atlas terminologies poses a considerable challenge for reconciling datasets annotated with terms from different atlases, as discussed in Chapter 2, and there is no online resource that provides an anatomical framework with lexical mappings between atlas terms. To fill these terminology gaps for cognitive neuroimaging research, two preliminary studies were completed during the SNI phase to deliver 1) machine-readable MRI acquisition terms and 2) brain atlas mappings for use in the NIDM framework.

To fill the DICOM terminology gap, a study was conducted to develop a resource of unambiguously defined terms accessible on the Web. This resource would enable MRI acquisition metadata to be represented independently of the DICOM header (109). Once an MRI dataset is acquired in DICOM format it is typically converted to NifTI format for data analysis and the original data acquisition metadata from the DICOM header is lost. This conversion is lossy and omits metadata that is essential for secondary data use, thus undermining the potential of shared NifTI data. To develop a DICOM terminology, the INCF NIDASH task force extracted DICOM terms from natural language specification documents and imported them into a shared spreadsheet where terms and definitions were reviewed during weekly meetings. The outcomes of this study provides 1,757 DICOM terms that are now available through NeuroLex (69), where each term is assigned a unique identifier and URI that can be accessed on the NeuroLex DICOM terms webpage (http://uri.neuinfo.org/nif/nifstd/nlx_149624).

To address the need for harmonized brain atlas terms, a study was conducted to augment the FMA with the spatio-structural properties needed to represent different brain labeling schemes within a coherent framework (26). By accommodating different views of neuroanatomical organization within the same framework, the Structural Informatics Group enhanced the FMA properties to correlate disparate brain labeling schemes and to annotate brain imaging datasets, such as those produced by FreeSurfer in the data sharing scenario that introduced this chapter. Figure 4.1 demonstrates how the FMA provides a formalized structure to lexically map the disparate terminologies discussed in Chapter 2. As reported in Turner et

al., the FMA was used to annotate a large dataset of task-based fMRI signal activations both in participants with schizophrenia and healthy controls. This was then used to demonstrate how the enhanced FMA properties could be used to answer novel questions about the data with "intelligent" queries, such as "What are all the brain regions connected by the dorsal segment of the right superior longitudinal fasciculus?"

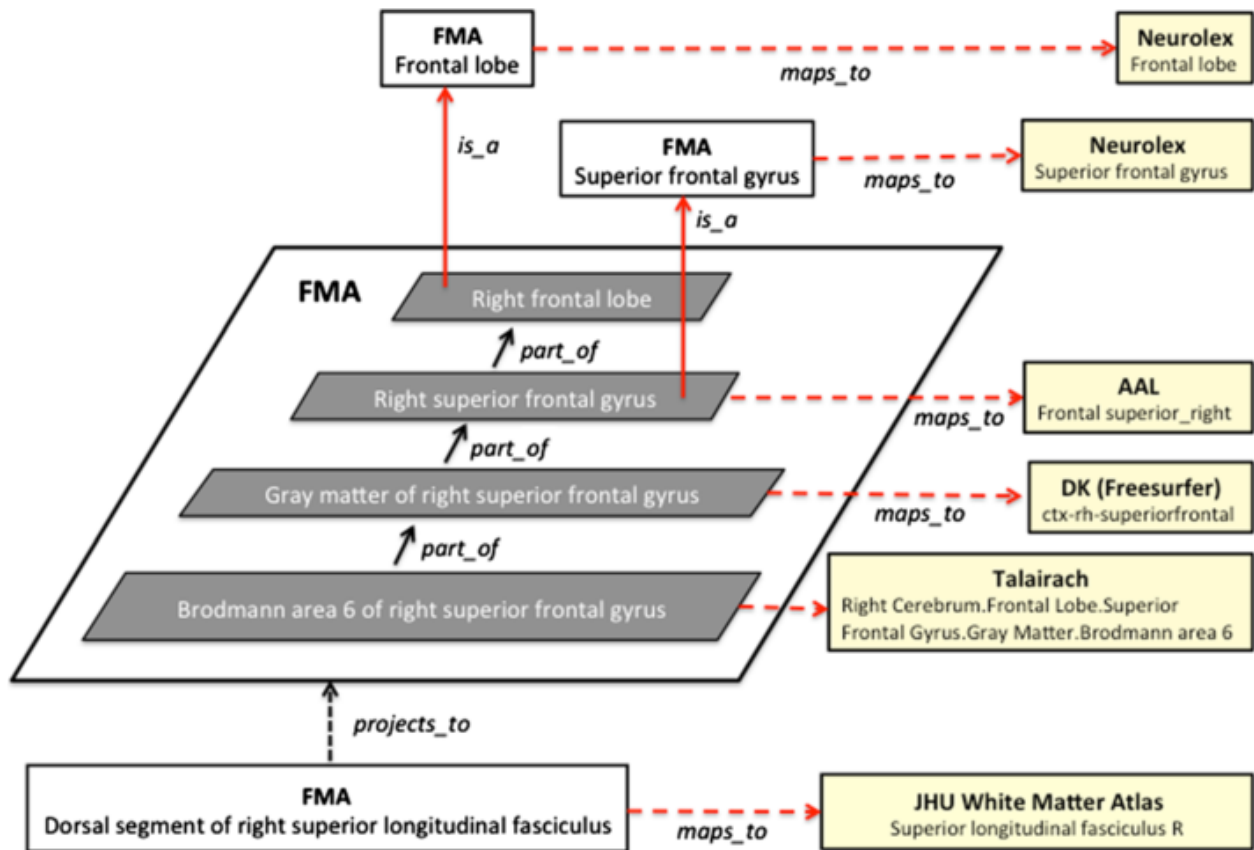


Figure 4.1. FMA Brain Atlas Mapping. An example of how terms from brain atlases and vocabularies can be correlated by mapping to the corresponding class in the FMA hierarchy.

The outcomes from the SNI phase for these two preliminary studies provided terms to annotate MRI acquisition metadata and neuroanatomical assays (e.g., FreeSurfer anatomical parcellation and segmentation statistics). Additionally, annotations from the FMA were shown to enhance information retrieval tasks by providing a computable set of relations for linking levels of anatomical granularity, connectivity, and correlating across brain atlas terminologies. By incorporating these terminologies, as well as others, into the NIDM framework, extensions to the PROV standards can be further linked into a broad set of online information resources.

The two requirements that emerged from these studies are: 1) to develop domain specific reference terminologies or ontologies and 2) to reuse terms by importing them in specific applications. By providing a common vocabulary of terms, these requirements enable applications using NIDM to fulfill the data exchange and interoperability envisioned in the neuroimaging data sharing scenario that introduced this chapter.

4.3.1.2 PROV data model extensions

A basic requirement for reproducing the experimental conditions, under which data were acquired and processed, is a record of provenance. Provenance information is used to capture detailed metadata about the original state of a piece of data, as well as the computational environment used to transform the data *in any way*. In biomedical and neuroimaging research, provenance information is best recorded in context with links to project information that include metadata about the study participants, principal investigator, experimental design, etc. The XCEDE XML Schema, introduced in Chapter 2, provides an Experimental Hierarchy used to model information at the levels of Project, Subject, Visit, Study, Episode, and Acquisition information, as well as data provenance. Thus, XCEDE can capture data provenance in the context of project information; however, PROV (110), a W3C standard, has emerged as the gold standard for representing provenance on the Web. To facilitate data exchange between neuroimaging systems and to capture provenance information with the terms discussed above, two preliminary studies were conducted during the SNI Phase to 1) identify the advantages and disadvantages of using XCEDE standard for data exchange and 2) harmonize the XCEDE and PROV XML Schemas to enable provenance tracking in the context of an experimental project.

To evaluate the XCEDE format for standardized data exchange, a study was conducted to integrate data from two neuroimaging data management systems using XCEDE as a common schema (93). Data was extracted from XNAT (34) and the Neurobiological Imaging Management System (NIMS, <https://github.com/cni/nims>) and was transformed into XCEDE before being loaded into a Web accessible location. XCEDE provided the necessary data structures to conceptually map information from both XNAT and NIMS into a common XML format; however,

both systems contained additional information that would not validate with the XCEDE XML Schema without additional modeling effort. The rigid structure of XML Schema did not fit well with the rapidly evolving set of terminologies and concepts found in neuroimaging. In contrast, the Open World assumptions provided by semantic Web standards, such as RDF and OWL, were found to provide the additional flexibility necessary for the dynamic landscape of metadata in neuroimaging, albeit at the cost of the robust validation provided by XML Schema. The outcomes of this study indicate the need for a flexible modeling environment that assumes incomplete and complementary information from external terminologies found on the Web.

To harmonize the XCEDE and PROV XML Schemas, a study was conducted to extend XCEDE with the formalized notion of provenance provided by the W3C PROV specifications (111). XCEDE specifies an XML Element for provenance that can express an ordered sequence of actions performed on data including command line executable programs, parameters, and associated metadata. While this matches with traditional notions of provenance, it lacks the PROV notion of "views" that provide different provenance perspectives including a "data flow view," "process flow view," and "responsibility flow view." The PROV Entity, Activity, and Agent concepts, detailed in Chapter 2, endow each of these views with specific properties. In XCEDE, all XML elements are considered to be containers of information that roughly map to the PROV concept of Entity, which neglects the richness provided by the additional PROV concepts. To harmonize these two schemas, the PROV XML Schema was imported and mapped into an XCEDE XML Schema document. Instances of the schema were manually generated to capture the richness of PROV within the experimental project context of XCEDE. While it was possible to map, for example, an XCEDE Project element to a PROV Entity and an XCEDE Subject Element to a PROV Agent, it became clear that doing so resulted in redundant information and the additional complexity did not provide significant benefit. However, the outcomes of this study generated a key insight into the generic nature of the PROV model, namely that it could be extended to capture the same information expressed by XCEDE, but entirely within the framework of provenance information.

4.3.1.3 Access and query application programming interface

A basic requirement for a decentralized, Web-based application framework is an application programming interface (API) that provides access and query functionality. In the data sharing scenario section, several neuroinformatics applications were described that operate on overlapping, yet distinct sets of information. For example, the OpenfMRI database contains the type of information modeled by the XCEDE Experiment Hierarchy, while the NiQuery application requires additional provenance information to record the FreeSurfer workflow environment. The information requirements for a given application can be considered a "view" over a common data model, where the data model contains more information than any single application may require. To explore the types of data model "views" that would provide applications with useful information, two preliminary studies were conducted during the SNI Phase to 1) evaluate the Annotation Image Markup (AIM) standard (112) for semantic annotation of brain images, and 2) design an initial API specification based on the XCEDE Experiment Hierarchy to access information common to neuroimaging data management systems.

To evaluate the Annotation Image Markup (AIM) standard for the semantic annotation of brain images, a study was conducted to apply the brain atlas mappings provided by the FMA to generate an AIM document that represents voxel-wise activation in a statistical fMRI dataset (113). AIM is an information model for describing regions of interest in images by labeling voxels with semantic information, such as anatomical labels or clinical findings. The AIM standard was designed to work with the Digital Imaging and Communications in Medicine (DICOM) standard, which is ubiquitous in medical imaging systems, but does not provide a mechanism for annotation. Brain atlases are used to identify brain regions in images by using numerical labels and a lookup table that maps the number to a given anatomical structure. In a statistical fMRI map, the value at a given coordinate corresponds to an activation level, but the coordinate lacks an anatomical label. By registering the statistical map into the same coordinate system as a brain atlas, a lookup table can be used to label a given statistical value with a brain region label. AIM provides a standard mechanism to represent the coordinates, statistical value,

and anatomical region in an XML document. An AIM XML document can then be transferred over the Web and used to exchange images that are semantically annotated with terms from an ontology, such as the FMA. To assess the feasibility of using AIM as a data exchange format for fMRI datasets annotated with brain atlas labels from the FMA, an Image Annotation Service (IAS) was developed to generate AIM XML documents that captured annotated brain activation measures.

The IAS system architecture is outlined in Figure 4.2, which provides an API to annotate fMRI datasets on a voxel-by-voxel basis with FMA identifiers. As an initial experiment for defining an API, the system revealed several limitations of using the AIM standard for brain imaging. First, much of the shared brain imaging data is stored in NIfTI format, but AIM requires the use of DICOM. While a mock DICOM header can be generated, it is considered bad practice to generate fake DICOM identifiers. Second, the AIM standard is much more verbose than the compact binary representation provided by NIfTI. For example, a compressed NIfTI file may be 2.3 megabytes, whereas the same information modeled in AIM for the whole brain is over 500 megabytes and is not scalable. The outcomes of this study indicate that a voxel-by-voxel XML representation is unnecessarily verbose and that a lighter weight solution consisting of summary information and a link to the original file would be preferable, such as the approach taken by NeuroVault in the data sharing scenario.

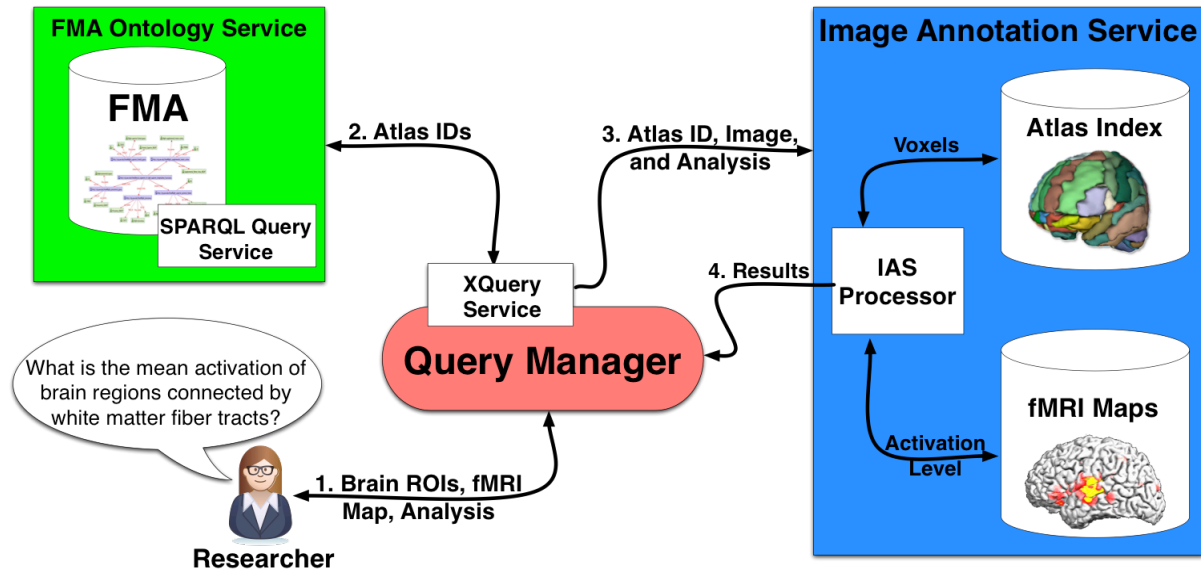


Figure 4.2. Image Annotation Service. A researcher can 1) use a query management application to 2) execute a query over the FMA to retrieve a mapping of brain atlas identifiers that can be 3) submitted to the IAS where 4) an AIM representation is generated for all of the brain regions identified in the initial FMA query and returned to the user. The code is open source and available at: <https://github.com/nicholsn/pyAIM>

To design an initial API specification based on the XCEDE Experiment Hierarchy, a study was conducted to specify a data exchange layer between neuroimaging databases with a common interface to access and query shared brain imaging data (114). A large number of databases are developed to store and manage human neuroimaging data. While individual neuroscience databases provide mechanisms to query and download information within a given framework (34-36,86,115) there is no standardized way to programmatically access related information stored in these heterogeneous systems. Creating a data exchange layer with a common interface to access and query shared brain imaging data would enable the development of interoperable client applications capable of consuming resources available across disparate brain imaging data management systems.

The initial API design focused on providing uniform access to neuroimaging databases. Conceptually, the API was defined as a service for accessing common entity types (e.g., '**nidm:Project**', '**nidm:Participant**') and their relationships (e.g., ':*subject*' '*prov:wasAssociatedWith*' ':*project*') specified by the metadata terms and XCEDE-based PROV extensions discussed in the previous two sections. Neuroimaging databases conforming to the

API would implement a mapping of their local resources to XCEDE entities and provide a mechanism to request resources. The API is not tied to a specific language or technology, but for web-accessible databases the Representational State Transfer (REST) pattern (116) fit the system requirements. The initial REST API would respond to an Hypertext Transfer Protocol (HTTP) request for a Subject URI (e.g., http://www.example.com/xcede_query/subject?uri=<uri:string>) by listing the relationship and Uniform Resource Identifier (URI) of related entities (e.g., *'subject' 'prov:wasAssociatedWith' 'project_id'*). The specified HTTP resources are listed in Figure 4.3. The response to an API request at a given URI would return XCEDE XML or another serialization (e.g., Turtle) conforming to the XCEDE data model. The API design was discussed with the developers of many neuroimaging databases to identify limitations of this approach. The outcomes of this study provided a prototype common API specification to use for further refinement in Phase 2 of this dissertation.

Resources - Elements

Resource	Methods	Description
Entity	GET	This resource contains <i>prov:entity</i> metadata and properties
Activity	GET	This resource contains <i>prov:activity</i> metadata and properties
Agent	GET	This resource contains <i>prov:agent</i> metadata and properties
Project	GET	This resource contains <i>prov:activity</i> metadata and properties for <i>prov:type="xcede:project"</i>
Visit	GET	This resource contains <i>prov:activity</i> metadata and properties for <i>prov:type="xcede:visit"</i>
Study	GET	This resource contains <i>prov:activity</i> metadata and properties for <i>prov:type="xcede:study"</i>
Episode	GET	This resource contains <i>prov:activity</i> metadata and properties for <i>prov:type="xcede:episode"</i>
Participant	GET	This resource contains <i>prov:agent</i> metadata and properties for <i>prov:type="xcede:study"</i>
Acquisition	GET	This resource contains <i>prov:agent</i> metadata and properties for <i>prov:type="xcede:acquisition"</i>

Figure 4.3. XCEDE API. The initial common API specification provides resources to access the base PROV types, as well as extensions based on XCEDE.

These two SNI phase preliminary studies from the SNI Phase provided an assessment of the AIM standard for representing voxel-level semantic annotations and a specification for the initial set of API resources for accessing XCEDE-based extension to PROV. While AIM was capable of modeling fMRI statistical images and annotating values with terms from the FMA, the requirement to use DICOM and the verbosity of AIM were not a good fit for brain imaging use cases. The voxel-by-voxel annotation method was found to be unnecessary, as access to full image files was sufficient. The IAS system also provided an initial example of a computational Web service, which is explored further in the prototype system study. The initial API specification was used to gather feedback from neuroimaging database developers, who

were generally supportive; however, these developers had limited resources to implement the API as part of their system and were unable to add this feature request to their software development roadmap. It was suggested that a separate server be used to extract, transform, and load information from each neuroimaging system and expose an API as a separate server. The results of these two studies identified key limitations of AIM and provided a starting point for further development of a common API.

The research approach discussed in this section outlined the preliminary studies guiding the development of a neuroimaging data sharing framework. These initial studies demonstrated vocabularies useful for annotating MRI data with DICOM terms and brain atlas labels, which can be used in conjunction with a data model that intrinsically captures data provenance.

4.3.2 Outcomes and design

This approach captures provenance not as an afterthought but as explicitly modeled relationships between Entities, Activities and Agents that support modeling research as a process. This section reports the outcomes and design of a community-driven framework that emerged from these initial studies by following the data sharing scenario as a use-case to introduce the Neuroimaging Data Model (NIDM). First, a conceptual overview of NIDM is presented to familiarize the reader with the framework components before tying these components back to the use-case of a student accessing data from the OpenfMRI database. Each of the subsections continue with this use-case by examining two key NIDM components for data management, the Dataset Descriptor and Experiment Components, using detailed examples from OpenfMRI and the Study Forrest project (8). The section concludes with a description of the data modeling process used to develop these data sharing standards.

At its core, NIDM builds upon semantic Web technologies to extend existing vocabulary and ontology standards that support reproducibility and data reuse in the domain of cognitive neuroimaging. By extending the generic concept of a provenance data model, specialized components were defined to model information from each stage of the research process

starting with project-level information about participants and acquired data, to information about computational workflow, as well as derived statistical results. By incorporating each of these elements within the same conceptual framework (i.e., provenance information), a rich description of research processes can be captured in a machine-readable form and investigated computationally. Figure 4.4 provides a set of four conceptual layers and five NIDM Components that are introduced here and detailed in the sections below.

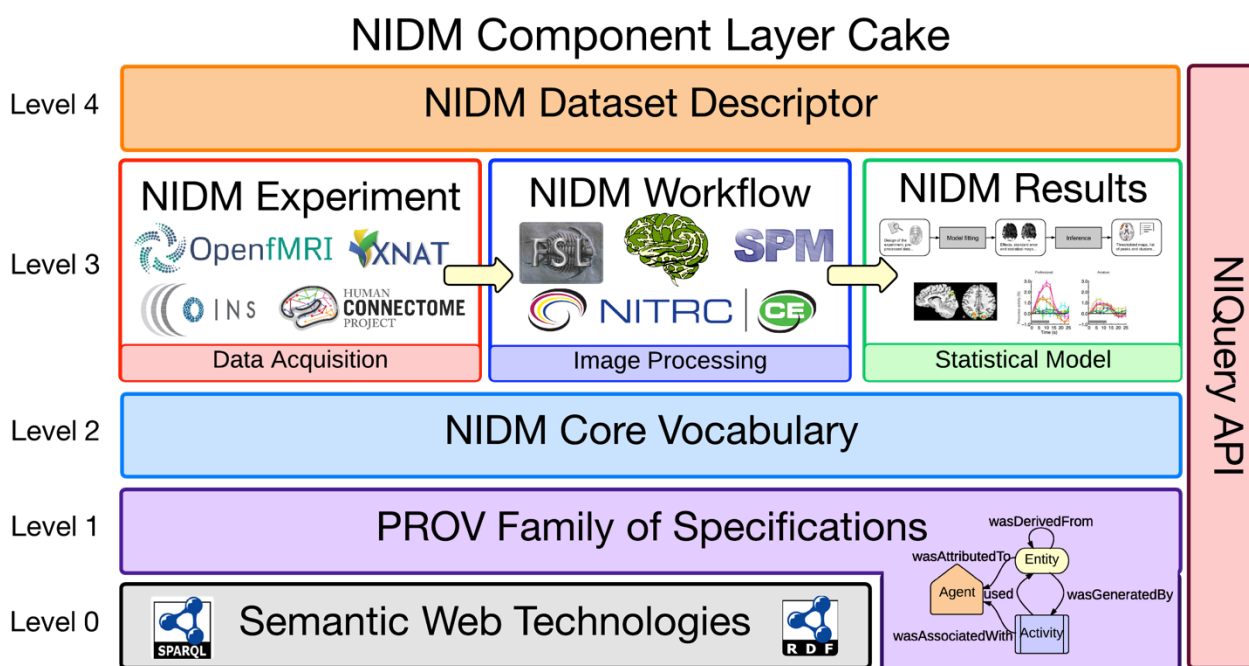


Figure 4.4. NIDM Layer Cake. NIDM is organized into a set of layers that each build upon the layers beneath. The NIDM Components include 1) 'NIDM Dataset Descriptor' information to provide an overall description of available documents, 2) 'NIDM Experiment' information to model the XCEDE Experiment Hierarchy, 3) 'NIDM Workflow' information implemented by processing tools that transform data, 4) 'NIDM Results' information to model essential details of derived data, and 5) 'NiQuery API' to access and query views of information modeled in other components.

Each NIDM Component is defined by a specification that includes terms with definition and an information scope. The three components in level 3 (i.e., NIDM Experiment, NIDM Workflow, and NIDM Results) relate directly to information that is generated during a scientific investigation, while level 4 (i.e., NIDM Dataset Descriptor) is a meta-layer to describe additional information about how a given dataset, or NIDM bundle, is distributed. The NiQuery API Component specifies how information can be retrieved from the Object Models that are defined by each Component, where an Object Model is a particular "view" of the underlying PROV data

model designed for a specific application. To transition into how NIDM relates to the data sharing scenario, Figure 4.5 provides an example of how the Entities, Activities, and Agents defined by PROV are extended to represent Object Models for an experimental project information (e.g., OpenfMRI and Study Forrest), workflow information (e.g., FreeSurfer), and derived data information (e.g., FreeSurfer statistics). Each of the following sections provides a detailed account of each NIDM Component, starting with the Dataset Descriptor that the student in the data sharing scenario would pull into their workspace to identify scans with which to run FreeSurfer.

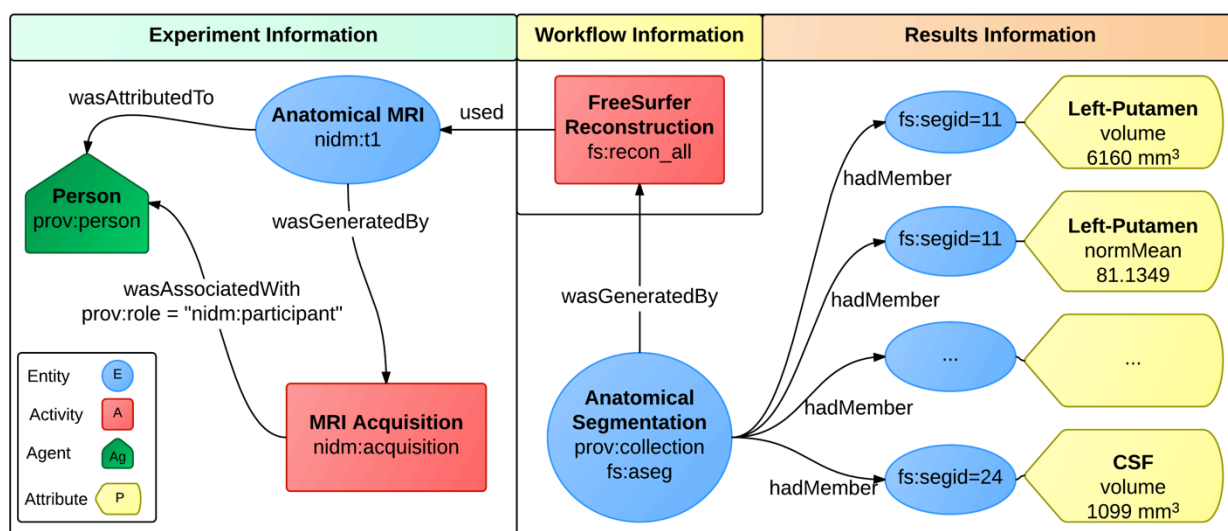


Figure 4.5. NIDM Information Domains. Provenance data model extensions showing how the relationships between experiment, workflow, and results information can be described in the context of provenance using the core Agent, Activity, and Entity elements of the W3C PROV data model.

4.3.2.1 NIDM dataset descriptor component

The data sharing scenario describes how a student identified the Study Forrest data paper during a literature review and found that the data for this study was available on the OpenfMRI database. One goal of publishing a data paper is to encourage the reuse and reanalysis of a given biomedical dataset, which NIDM can facilitate. As more datasets become available from different biomedical domains it will be essential to aggregate heterogeneous data in meaningful ways. Interdisciplinary collaboration is necessary to develop a common, high-level dataset descriptor that provides a generic representation of a dataset's attributes. The NIDM Dataset Descriptor Component (NDDC) is modeled after an effort to produce such a harmonized

dataset descriptor, which was developed by the Health Care and Life Sciences (HCLS, <http://www.w3.org/blog/hcls>) Interest Group at the W3C. HCLS created a generic dataset descriptor specification that captures a general set of metadata that is applicable across biomedical domains. Particular focus is given to representing information at three distinct levels; Summary, Version, and Distribution, which represent static, version specific, or format specific information about a dataset, respectively. By adopting the HCLS recommendations and extending them, NDDC can remain harmonized with HCLS, and interoperable with dataset descriptors from other domains, while providing the additional structure and vocabulary necessary for neuroimaging. To provide an intuitive introduction to the HCLS Dataset Descriptors and to demonstrate how NDDC extends this effort, information from the OpenfMRI website is used to create a dataset descriptor of the Study Forrest project (Figure 4.6).

The figure consists of two side-by-side screenshots. The left screenshot shows the OpenfMRI website's 'Data Sets' page. A table lists datasets with columns for 'Study', 'Investigators', and 'Accession Number'. The right screenshot shows the 'Study Forrest' dataset page, which includes a title, submission date, description, and additional resources. Red boxes and arrows labeled A through F point to specific elements on both pages that are used for data extraction.

Study	Investigators	Accession Number
Balloon Analog Risk-taking Task	Tom Schonberg	ds000001
A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie	Michael Hanke	ds000113

Annotations:

- A:** Points to the OpenfMRI logo and 'Home View Data Sets' link.
- B:** Points to the investigator name 'Michael Hanke' in the table.
- C:** Points to the accession number 'ds000113' in the table.
- D:** Points to the dataset title 'A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie'.
- E:** Points to the 'Additional resources' section.
- F:** Points to the 'Investigator Info' link in the 'Additional resources' section.

Figure 4.6. OpenfMRI and Study Forrest Mapping to NIDM. A Screenshot of the OpenfMRI datasets page (left) and the Study Forrest dataset page (right). Each red box highlights information that is extracted and converted to attributes in the NIDM Dataset Descriptor Component. A) The database title and logo are mapped to '*dct:title*' and '*schemaorg:logo*', B) the investigator is mapped to '*dct:creator*', C) the accession

number is mapped to '*dct:hasPart*', D) the title of the Study Forrest paper is mapped to '*dct:title*', E) the abstract is mapped to '*dct:description*', and F) additional attributes about the dataset could also be mapped to NIDM, but are not currently.

NDDC extends HCLS Dataset Descriptors by introducing two Summary-level descriptions, which are split into Database and Project parts. This split allows a database, such as OpenfMRI, to include static details about the overall database and then provide links to its specific parts (i.e., projects). This style of organization is ubiquitous with data sharing websites, where there is generally more than one project available. The information denoted on these websites contains features that are consistently represented across databases, although each database may use their own vocabulary for representing them. For example, OpenfMRI refers to each collection as a "dataset," while an XNAT database refers to each collection as a "Project" (34) - in NIDM the latter is adopted. In Figure 4.6, we highlight a few pieces of information that can be captured and converted into a Summary-level dataset descriptor with two levels - the Database Level (left) and the Project Level (right), which are NDDC extensions to HCLS.

The two Summary-level dataset descriptors in Figure 4.6 can be extracted and converted into a harmonized Linked Data representation that simplifies aggregating datasets. Figure 4.7 provides a detailed example of a NDDC Database Summary-level document as RDF, using Turtle syntax (<http://www.w3.org/TR/rdf-sparql-query/>), which captures the minimal information necessary to be compliant with HCLS, as well as NDDC specific extensions. For example, NIDM includes two additional types that declare OpenfMRI to be a '**prov:Collection**' and a '**nidm:Database**', and requires the '*dct:hasPart*' attribute. An important distinction between the OpenfMRI webpage and RDF representations is that RDF uses standard terms with definitions to represent the underlying data, rather than the purely textual representation provided by HTML. In this way, the data and presentation layers are distinct, enabling the RDF data to be portable and reusable, while the HTML alone does not readily facilitate reuse. Within the Database Summary-level is the '*dct:hasPart*' attribute, which provides an entry point to the Project Summary-level where project specific details are available for each of the OpenfMRI datasets, such as '*:ds000113*' (i.e., the Study Forrest project), as well as the Version level, and Distribution levels of NDDC (Figure 4.8).

```

@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dctypes: <http://purl.org/dc/dcmitype/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix pav: <http://purl.org/pav/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix schemaorg: <http://schema.org/> .
@prefix void: <http://www.w3.org/TR/void/> .
@prefix nidm: <http://www.incf.org/ns/nidash/nidm#> .
@prefix : <http://openfMRI.s3.amazonaws.com/nidm.ttl#> .

# Database Descriptor of OpenfMRI.
:openfMRI
  a dctypes:Dataset, prov:Collection, nidm:Database ;
  dct:title "OpenfMRI"@en ;
  dct:description ""OpenfMRI.org is a project dedicated to the free and
    open sharing of functional magnetic resonance imaging
    (fMRI) datasets, including raw data.""@en ;
  dcat:accessURL <https://openfMRI.org> ;
  dct:license <http://www.opendatacommons.org/licenses/pddl/1.0/> ;
  dct:publisher <https://openfMRI.org> ;
  schemaorg:logo <https://openfMRI.org/sites/all/themes/openfMRI/logo.png> ;
  dcat:theme <http://dbpedia.org/page/Neuroimaging> ;
  dct:hasPart :ds000001, :ds000005, :ds000113 .

```

Figure 4.7. Database Descriptor Element. Provides a structured RDF representation of the OpenfMRI Database that includes corresponding namespaces, as well as links its dct:parts, such as the :ds000113 - the Study Forrest Project.

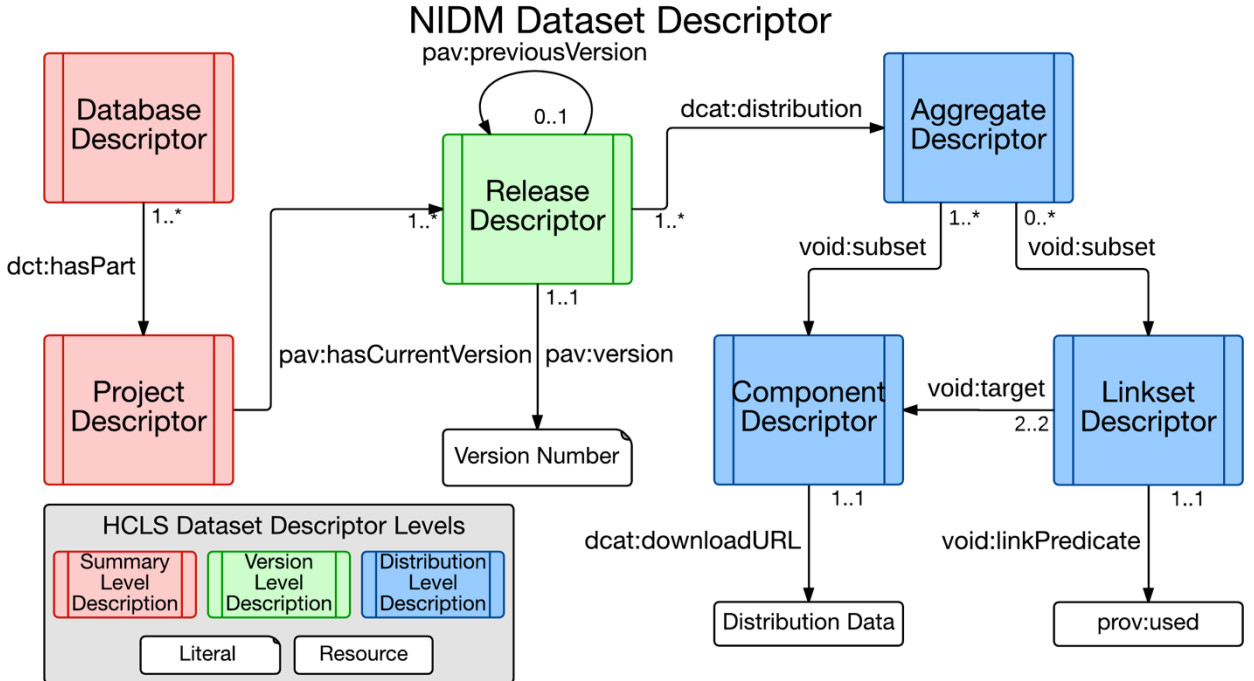


Figure 4.8. Details of NIDM Dataset Descriptor shown in Figure 4.4. Colors correspond to HCLS Dataset Descriptor levels. The Database Descriptor and Project Descriptor elements provide static details about one or more datasets. The Release Descriptor element provides version-specific information. The HCLS Distribution level Descriptor is represented by an Aggregate Descriptor, which bundles together the

Component Descriptor that describes how to access data, and the optional Linkset Descriptor describe how Components are linked together.

The Project Descriptor captures static details about the Study Forrest project (Figure 4.9). The structure of this descriptor is similar to that of the Database Descriptor (Figure 4.7), which shares several common attributes; however, it is differentiated by additional type information (i.e., '**prov:Entity**' and '**nidm:Project**'), a link to the current HCLS Version level description of the project's data (i.e., '*nidm:hasCurrentVersion*'), and a link to the OpenfMRI database (i.e., '*prov:specializationOf*'). Figure 4.9 features a basic set of required attributes, but additional elements can be included as long as they are not likely to change over time (e.g., fMRI paradigm(s) used or scanner type). The Version-level descriptor affords those attributes that fluctuate over time.

```
# Project Descriptor of Study Forrest.
:ds000113
  a dctypes:Dataset, prov:Entity, nidm:Project ;
  dct:title          """"A high-resolution 7-Tesla fMRI dataset from complex
                    natural stimulation with an audio movie""""@en ;
  dct:description    """"This is a high-resolution functional magnetic
                    resonance (fMRI) dataset – 20 participants recorded
                    at high field strength (7 Tesla)...""""@en ;
  dcat:accessURL     <https://openfmri.org/dataset/ds000113> ;
  dct:license        <http://www.opendatacommons.org/licenses/pddl/1.0/> ;
  dct:publisher      <https://openfmri.org> ;
  pav:hasCurrentVersion :ds000113-1.0.0 ;
  prov:specializationOf :openfmri .
```

Figure 4.9. Project Descriptor Element. Contains the core set of static attributes for a project that are used in the Version and Distribution levels.

At the HCLS Version-level, as represented by the Release Descriptor in Figure 4.8, NIDM only extends the HCLS recommendations to include additional types and attributes (Figure 4.10). As the Version-level is meant to capture information that changes from one data release to the next, an additional NIDM type called '**nidm:Release**', which is a '**prov:Entity**', is also included. Also included is the '*prov:specializationOf*' attribute, which keeps an upstream provenance link to the Project Descriptor, as opposed to '*prov:wasDerivedFrom*' attribute that links to the HTML version of the project information. While not listed here explicitly, the Release Descriptor (Figure 4.8) has several repeated pieces of information from the Project

Descriptor (i.e., `'dct:title'`, `'dcat:accessURL'`, etc.) that allow this resource to be listed in a separate document from the Project Descriptor and remain interpretable. The Distribution Descriptor figures also omit these repeated attributes that exist for the same purpose. The `'dcat:distribution'` link points to the Distribution-level dataset descriptor, which delivers a specific data format that represents a given data release - in the case of NIDM it is recommended to be RDF, but could also include a database, CSV file, etc.

```
# Release Descriptor of Study Forrest
:ds00013-1.0.0
  a dctype:Dataset, prov:Entity, nidm:Release ;
  # ... repeated attributes
  dct:creator      <http://mih.voxindeserto.de/> ;
  pav:version      "1.0.0" ;
  dct:isVersionOf  :ds00013 ;
  prov:wasDerivedFrom <https://openfmri.org/dataset/ds00013> ;
  prov:specializationOf :ds00013 ;
  dcat:distribution :ds00013-1.0.0-rdf .
```

Figure 4.10. Release Descriptor Element. Describes a data release that includes a version number and a link to a format-specific distribution of the data.

The elements of the HCLS Distribution-level (blue boxes in Figure 4.8) are intended to inform a user or application where NIDM Components can be accessed on the Web. A distribution can come in one or more formats (e.g., relational database, tarball, or RDF) - with NIDM, the Turtle serialization of RDF is recommended. For RDF datasets, HCLS recommends using the VoID vocabulary (117) that focuses on describing RDF graphs and includes features to, for example, split large graphs into more manageable chunks and describe how they are interconnected. NIDM documents are generally represented as RDF, thus NIDM uses a hybrid of VoID and HCLS to specify additional structure and vocabulary extensions amenable to neuroimaging.

As shown in Figure 4.8, NIDM extends the HCLS notion of a Distribution-level dataset descriptor by providing three descriptor elements - Aggregate, Component, and Linkset (optional). The Aggregate Descriptor is intended to capture the overall organization of a data release. An Aggregate Descriptor provides links to one or more RDF datasets using the `'void:subset'` relationship (Figure 4.11), where each `'void:subset'` conforms to a model defined in a

NIDM specification. The primary *void:subset* links belong to the Component Descriptor (Figure 4.12); however, it can also include an optional Linkset. NIDM uses Void Linksets as a virtual collection of relationships between components, where the object for a given triple in one dataset links to the subject in another dataset. This allows the data publisher to attach additional metadata about, for example, how many links there are between datasets and what predicate to use when querying across them (e.g., *prov:used*). Finally, the Component Descriptor defines the type of component a given dataset conforms to (e.g., **nidm:Experiment**), as well as the format and location from where that RDF file can be downloaded.

```
# Aggregate Descriptor of Study Forrest
:ds000113-1.0.0-rdf
  a void:Dataset, dcat:Distribution, prov:Entity, nidm:Aggregate ;
  # ... repeated attributes
  dct:format <http://www.w3.org/ns/formats/Turtle>, "text/turtle" ;
  prov:specializationOf :ds000113-1.0.0 ;
  void:subset :ds000113-1.0.0-rdf-experiment,
              :ds000113-1.0.0-rdf-workflow,
              :ds000113-1.0.0-rdf-results ;
# Optional Linkset Descriptor of Study Forrest
void:subset [
  a void:LinkSet ;
  void:subjectTarget :ds000113-1.0.0-rdf-workflow ;
  void:linkPredicate prov:used ;
  void:objectTarget :ds000113-1.0.0-rdf-experiment ;
], [
  a void:LinkSet ;
  void:subjectTarget :ds000113-1.0.0-rdf-results ;
  void:linkPredicate prov:used ;
  void:objectTarget :ds000113-1.0.0-rdf-workflow ;
] .
```

Figure 4.11. Aggregate Descriptor and Linkset Descriptor Elements. Provides a link to where NIDM RDF datasets can be downloaded and includes one or more Linksets that capture how NIDM components are interlinked.

```

# Component Descriptor of Study Forrest
:ds000113-1.0.0-rdf-experiment
  a void:Dataset, dcat:Distribution, prov:Entity nidm:Experiment;
  # ... repeated attributes
  void:dataDump <http://openfmri.s3.amazonaws.com/ds000113/experiment.ttl> ;
  dcat:format <http://www.w3.org/ns/formats/Turtle>, "text/turtle" .

:ds000113-1.0.0-rdf-workflow
  a void:Dataset, dcat:Distribution, prov:Entity, nidm:Workflow ;
  # ... repeated attributes
  void:dataDump <http://openfmri.s3.amazonaws.com/ds000113/provenance.ttl> ;
  dcat:format <http://www.w3.org/ns/formats/Turtle>, "text/turtle" .

:ds000113-1.0.0-rdf-results
  a void:Dataset, dcat:Distribution, prov:Entity, nidm:Results ;
  # ... repeated attributes
  void:dataDump <http://openfmri.s3.amazonaws.com/ds000113/results.ttl> ;
  dcat:format <http://www.w3.org/ns/formats/Turtle>, "text/turtle" .

```

Figure 4.12. Component Descriptor Element. Provides a description and data format details, as well as a download link for each of the metadata files that conform to NIDM.

The NIDM Dataset Descriptor Component (level 4 of the NIDM Layer Cake shown in Figure 4.4) provides a high-level overview of the database contents and acts as the entry point into a set of NIDM documents. As such, it is agnostic to a given biomedical domain, which facilitates data aggregation and interoperability. However, this level of abstraction does not represent the details of an investigation in any meaningful way. In the next section the NIDM Experiment Component (left box in level 3 of Figure 4.4) is discussed, which is used to capture fine-grained information about the actual data acquired during a study, such as the anatomical MRI scans that the student in the data sharing scenario will need to import into the NiQuery application. A description of the remaining NIDM Workflow and Results Components from level 3 of Figure 4.4 is postponed until Chapter 6, where they are introduced in an example of the NiQuery system for tracking provenance and reporting results.

4.3.2.2 NIDM experiment component

Before collecting data for a research project, an important task is to decide how the acquired data will be organized on a file system or database and prepared for analysis. The best practices for how this preparation is accomplished vary greatly from one lab to the next, where some labs use a data management system (34-36,86) and others simply use a directory

structure on a file system. While specific layouts and naming conventions are rarely standardized across labs, there are common types of information that are monitored. Most labs will have one or more projects with associated files and data collated within a given directory, which in turn contain directories for data that is collected from each of the study participants. The goal of the NIDM Experiment Component (NEC) is to deliver a flexible representation of project metadata and data provenance that can be layered on top of any organizational scheme. The NEC accomplishes this by modeling acquired data as part of an ongoing research process expressed using PROV. The OpenfMRI standard directory structure was used as an example to construct a NEC Object Model that is detailed using the Study Forrest Project as an example. Figure 4.13 presents a simplified overview of the OpenfMRI directory structure for the Study Forrest project, where the project directory contains participant directories and summary files for the whole project. Each participant directory contains folders for a certain modality of data acquisition, and, if that modality has more than one acquisition, a folder is included for each run; otherwise the directory contains raw data and metadata files.

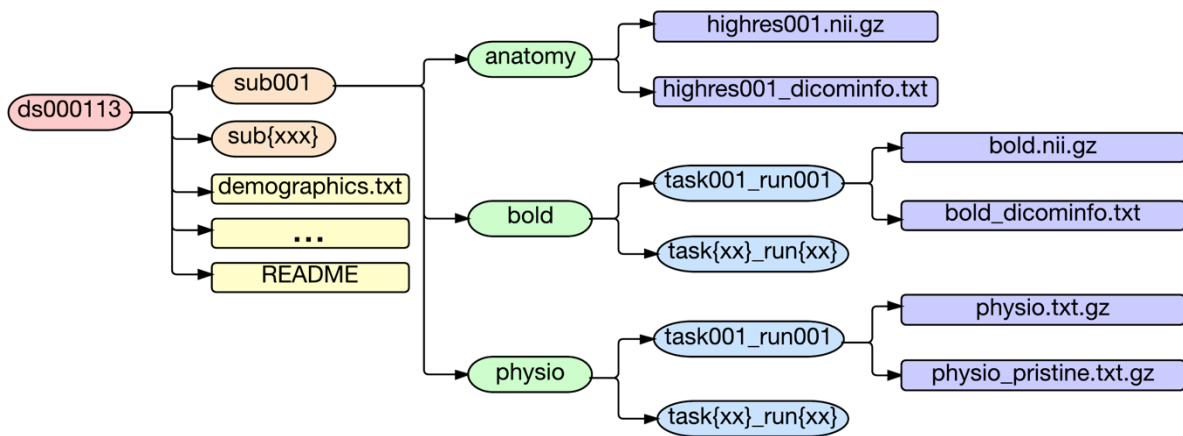


Figure 4.13. OpenfMRI Data Organization. Project Directory (red), Participant Directories (orange), Project Summary Files (yellow), Data Acquisition Modality Directories (green) Data Acquisition Protocol Directories (blue), Raw Data and Metadata files (purple).

To extract a NIDM representation of the directory structure, each directory and file was defined as an extension of the core PROV model described in Figure 2.2. A modeling pattern was defined to capture information within the scope of Project Information from 4.5. This

modeling pattern mirrors PROV with the '**nidm:ExperimentObject**', '**nidm:ExperimentProcess**', and '**nidm:ExperimentAgent**' class depicted in Figure 4.14. A '**nidm:ExperimentObject**' provides a mechanism to capture metadata and links to files in the OpenfMRI directory, while an '**nidm:ExperimentProcess**' explicitly models an investigation as an activity with a starting and ending point. A '**nidm:ExperimentAgent**' defines the person, organization, or software responsible for creating a file or acquiring data, for example. These three classes were further refined into NIDM Experiment Elements that map directly to XCEDE Experiment Hierarchy (87).

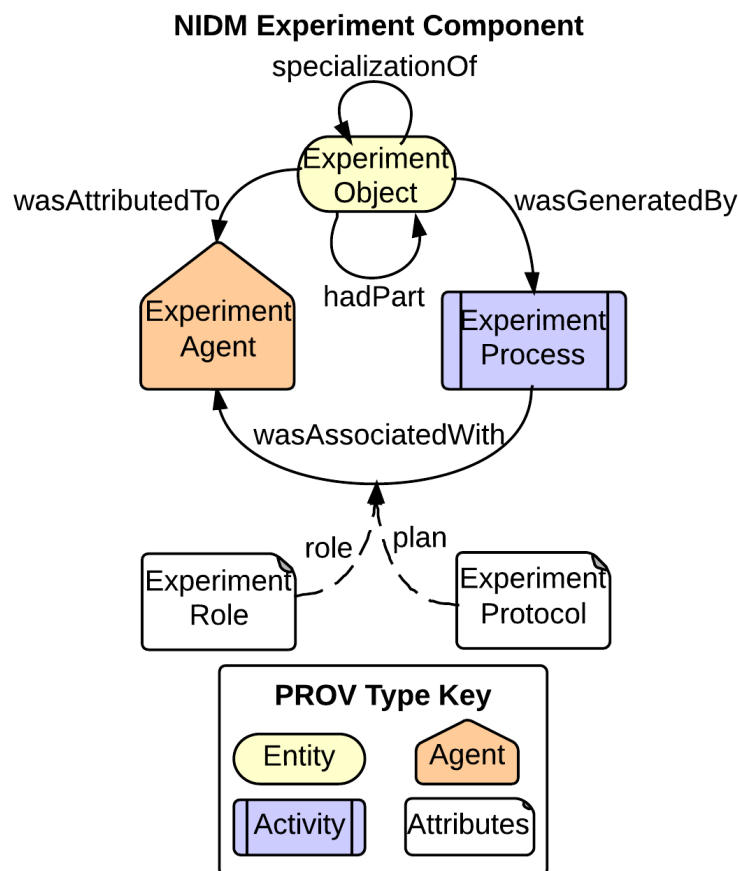


Figure 4.14. NIDM Experiment Modeling Pattern. The basic set of resources and relations that are used to capture the provenance of data acquired during a scientific investigation and to assign responsibility to individuals or software with specific roles and plans.

These extensions are grouped into the Project, Study, and Acquisition elements depicted in Figure 4.15 and follow the modeling pattern above. The element levels are linked using the

'*dct:hasPart*', and, inversely, '*prov:specializationOf*' relations to form a part and specialization hierarchy, respectively. Each element represents a level from the XCEDE Experiment Hierarchy, where additional XCEDE concepts, such as Visit, Subject, and Episode, are captured by other PROV constructs, such as '*prov:role*' and '*prov:plan*.' The type of project information modeled across each element is distinct and follows several rules to limit the scope of what is expressed. Examples from the Study Forrest project are modeled for each element, starting with the Project element.

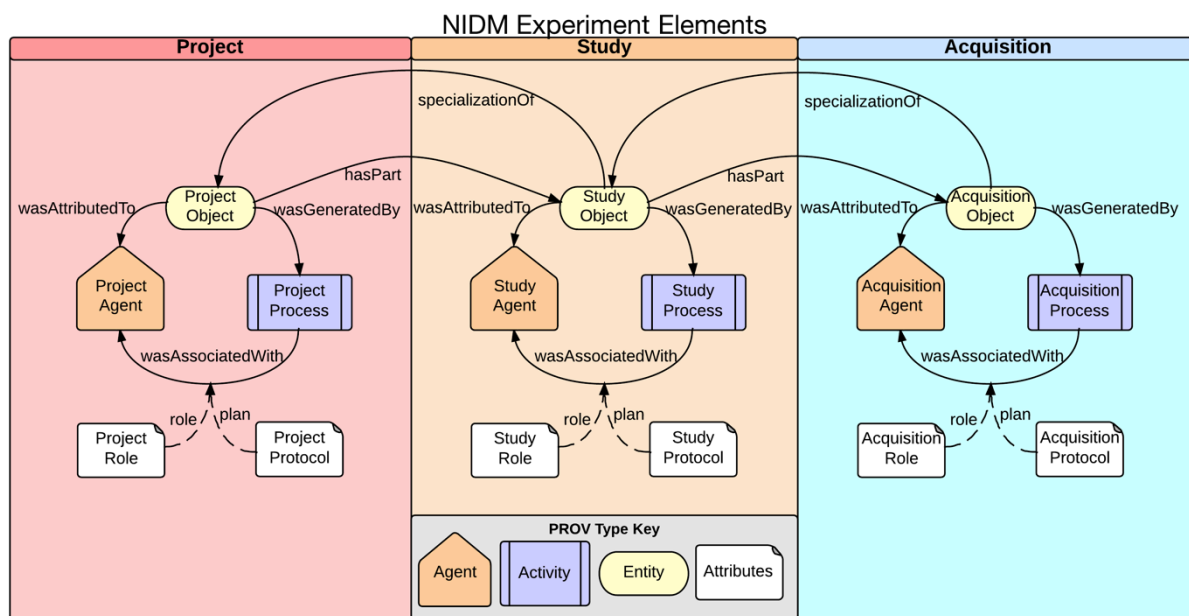


Figure 4.15. NIDM Experiment Elements. Project Element (red), Study Element (orange), Acquisition Element (blue).

The NEC Project Element is intended to capture high-level details about a project and to provide an entry point to the provenance of acquired data. On the surface, the Project element has similar features to the Project Summary-level dataset descriptor (Figure 4.9); however, it includes additional provenance structures that break from the HCLS recommendations to provide finer granularity about a project and provenance. By using the modeling pattern, a '**nidm:ProjectAgent**' can be assigned responsibility for initiating a '**nidm:ProjectProcess**' (i.e., the implied activity of doing research) and creating a '**nidm:ProjectObject**' (i.e., details about the project) (Figure 4.16). These three resources are then bound together by attributing the

'nidm:ProjectAgent' with a specific role during the project (e.g., a principle investigator) and associating a protocol to which the Project conforms.

```
@prefix : <http://openfmri.s3.amazonaws.com/ds000113/experiment.ttl#> .

# Project Element of Study Forrest.
:ds000113-project-object
  a void:Dataset, prov:Entity, nidm:Project ;
  dct:title      """"A high-resolution 7-Tesla fMRI dataset from complex
                 natural stimulation with an audio movie""""@en ;
  dct:description """"This is a high-resolution functional magnetic
                 resonance (fMRI) dataset – 20 participants recorded
                 at high field strength (7 Tesla)...""""@en ;
  dcat:accessURL <https://openfmri.org/dataset/ds000113> ;
  dct:license    <http://www.opendatacommons.org/licenses/pddl/1.0/> ;
  dct:publisher  <https://openfmri.org> ;
  pav:version    "1.0.0" ;
  prov:wasGeneratedBy :ds000113-project-process ;
  prov:wasAttributedTo :principle-investigator-project-agent ;
  dct:hasPart      :sub001-study-object-questionnaire,
                  :sub001-study-object-mri .

:ds000113-project-process
  a prov:Activity, nidm:ProjectProcess ;
  prov:wasAssociatedWith :principle-investigator-project-agent ;
  prov:qualifiedAssociation [
    a prov:Association ;
    prov:agent :principle-investigator-project-agent ;
    prov:hadRole nidm:PrincipleInvestigator ;
    prov:hadPlan <http://studyforrest.org/pages/resources.html> ;
    rdfs:comment """"The :principle-investigator-project-agent fulfilled the
                 nidm:PrincipleInvestigator role using a plan from the
                 Study Forrest website.""""@en ;
  ] .

:principle-investigator-project-agent
  a foaf:Person, prov:Person .

nidm:PrincipleInvestigator
  a prov:Role .
```

Figure 4.16. NIDM Project Element. Consists of three core named resources - '*ds000113-project-object*', '*ds000113-project-process*', and '*principle-investigator-project-agent*'.

In Figure 4.16, a key feature of PROV called the Qualification Pattern (118) is highlighted, which allows attributes, such as '*prov:role*' and '*prov:plan*', to be assigned to a relationship. For example, Figure 4.16 states that the '*ds000113-project-process*' '*prov:wasAssociatedWith*'

':principle-investigator-project-agent'; however, this relation alone does not provide the necessary context for a **'prov:Agent'** to be qualified with the *'prov:role'* of **'nidm:PrincipleInvestigator'**. To capture the context of both the *'prov:role'* and *'prov:plan'* used by the principle investigator to initiate the Study Forrest **'nidm:ProjectProcess'**, a *'prov:qualifiedAssociation'* link to an RDF blank node is used to bind *'prov:agent'*, *'prov:role'*, and *'prov:plan'* to the activity. This approach enables us to model people independently of the role(s) they may play throughout the research process, which may range from grant writer or research assistant to participant, data curator, or author. Once the Project Element is in place, Study Elements can be listed using *'dct:hasPart'*, which are used to represent types of data collected during the Study Forrest Project.

The NEC Study Element is used to model a set of observations using a related protocol and to reference specific Acquisition Elements. For example, a Study Element captures questionnaires, a neuropsychological test battery, or serially collected MRI series, while an Acquisition Element models the individual questionnaire items, neuropsychological test values, or MRI series modalities. Figure 4.17 details how a Study Element captures the provenance of a **'nidm:StudyQuestionnaire'** and binds it a **'prov:Person'** in the role of **'nidm:Participant'** using the same modeling approach as above. Similarly, Figure 4.18 demonstrates the representation of an MRI study with three Acquisition Elements listed under *'dct:hasPart'*, where each corresponds to a different imaging modality.

```

# Study Element for Questionnaires-
:sub001-study-object-questionnaire-
  a prov:Entity, nidm:Study, nidm:StudyQuestionnaire ;-
  dct:title      "Participant Questionnaire"@en ;-
  dct:description ""Participants' responses to a questionnaire on -
                  demographic information, musical preference and -
                  background, as well as familiarity with the 'Forrest -
                  Gump' movie.""@en ;-
  prov:wasGeneratedBy :questionnaire-study-process ;-
  prov:wasAttributedTo :sub001-study-agent ;-
  prov:specializationOf :ds000113-project-object ;-
  dct:hasPart          :questionnaire-acquisition-object .-
-
:questionnaire-study-process-
  a prov:Activity, nidm:StudyProcess ;-
  prov:wasAssociatedWith :study-agent ;-
  prov:qualifiedAssociation [
    a prov:Association ;-
    prov:agent :sub001-study-agent ;-
    prov:hadRole nidm:Participant ;-
    prov:hadPlan <http://studyforrest.org/pages/mod_annot.html> ;-
    rdfs:comment ""The sub001-study-agent performed the role of -
                  nidm:Participant during the -
                  :questionnaire-study-process using the questionnaire -
                  plan.""@en ;-
  ] .-
-
:sub001-study-agent-
  a foaf:Person, prov:Person .-
-
nidm:Participant-
  a prov:Role .-

```

Figure 4.17. Questionnaire Study Element. NIDM representation of a questionnaire at the study level.

```

# Study Element for MRI-
:sub001-study-object-mri-
  a prov:Entity, nidm:Study, nidm:StudyMagneticReasonanceImaging ;-
  dct:title      "MRI Study"@en ;-
  dct:description ""For each participant a number of different scans and -
                  auxiliary recordings have been obtained.""@en-
  prov:wasGeneratedBy :mri-study-process ;-
  prov:wasAttributedTo :sub001-study-agent ;-
  prov:specializationOf :ds000113-project-object ;-
  dct:hasPart :anatomy-acquisition-object,-
              :bold-acquisition-object,-
              :physio-acquisition-object .-

```

Figure 4.18. MRI Study Element. NIDM representation of an MRI scanning session at the study level with the 'prov:Activity,' 'prov:Person', and 'prov:Role' omitted.

To capture example questionnaire items and MRI metadata, Figure 4.19 demonstrates how demographic information, DICOM parameters, and file download locations are modeled. For clarity, the additional details related to 'prov:Activity' and 'prov:Agent' from the above examples were omitted. The 'nidm:DemographicsQuestionnaire' is a direct mapping from the CSV file listed for 'dat:downloadURL', where the column name is used as a predicate and the values for each cell are used as literals (e.g., ':gender'). While this representation is not currently harmonized with external terminologies, each column from the CSV file would ideally provide both a mapping to an external source and a range of appropriate value sets. For example, ':gender' would link to a URI for "male", rather than simply link to the string "m", thereby reducing ambiguity and adding precision. The 'nidm:MRIDAnatomicalT1' applies the DICOM terms developed in Section 4.3.1.1, which are dereferencable URIs with precise definitions that eliminate ambiguity and facilitate data integration tasks.

```
# Aquisition Element for Questionnaire
:questionnaire-acquisition-object
  a prov:Entity, nidm:Acquisition, nidm:DemographicsQuestionnaire ;
  # ... repeated attributes
  :gender      "m" ;
  :ageMin      30 ;
  :ageMax      35 ;
  :handedness  "r" ;
  # ... other items
  :listen_preference1_genre "Triphop" .
  dcat:downloadURL <http://openfmri.s3.amazonaws.com/ds113/demographics.csv>

# Aquisition Element for MRI

:anatomy-acquisition-object
  a prov:Entity, nidm:Acquisition, nidm:MRIDAnatomicalT1 ;
  # ... repeated attributes
  prov:wasGeneratedBy :mri-acquisition-process ;
  prov:wasAttributedTo :sub001-study-agent ;
  prov:specializationOf :ds000113-study-object ;
  dcm:Manufacturer      "Philips Medical Systems" ;
  dcm:ModelName         "Achieva" ;
  dcm:EchoTime           5.797 ;
  dcm:SliceThickness     0.7 ;
  # ... additional DICOM attributes
  dcm:NumberOfVolumes    1 ;
  dcat:downloadURL <http://openfmri.../ds113/sub001/anatomy/highres001.nii.gz>
```

Figure 4.19. Acquisition Element. NIDM representation of the acquisition-level for a demographics questionnaire and MRI scan.

In summary, the NIDM Experiment Component was developed to model information about an experimental investigation during the data acquisition stage, such as in the OpenfMRI directory structure. It consists of three levels designed to model Project, Study, and Acquisition Information in the context of data provenance. Where possible, standardized terms are applied to facilitate semantic interoperability and reduce ambiguity. The Qualification Pattern was introduced as a means to add contextual information to the links between activities and agents, thus enabling the use of roles and plans. By providing links to acquired demographics and MRI data, this component is the starting point from which analysis tools can query for the data necessary for processing. In the context of the data sharing scenario, the information provided here will be used as the input into the NiQuery system, which is used by the student to run the FreeSurfer reconstruction process.

4.3.2.3 NIDM workflow and results components

For this dissertation, the primary focus was on designing the overall NIDM Framework and on the NIDM Dataset Descriptor and NIDM Experiment Components. Work on the remaining NIDM Workflow and NIDM Results Components is the subject of ongoing and future work, which is being pursued through the INCF NIDASH task force. Examples of the Workflow, Results and API Components are presented in Chapter 5, where a prototype of the NiQuery application framework is discussed to provide details about the computational environment and analysis tools used for data processing.

4.3.2.4 NIDM framework design process

The NIDM Components discussed above were and are being designed using a community-driven process that engages stakeholders to participate in the identification of use-cases in need of data sharing standards. Initially, the NIDASH task force leadership identified a set of use-cases to drive development efforts and a process for collaboration was designed that made use of freely available software platforms and tools. The INCF initiated communication among

stakeholders and an in-person workshop was organized, where task force members agreed to coordinate NIDM development through a series of weekly videoconferences. A distributed version control system and the GitHub Web application were used by task force members to track progress and communicate technical aspects of the project, such as example files, code, and specifications.

The meeting minutes for weekly workshops were recorded during each in-person workshop and videoconference, which were circulated to the group for comments before being finalized and are available on the NIDASH Wiki (<http://purl.org/nidash/wiki>). The decisions made during each meeting were codified in the examples and checked into version control, which was then submitted for review to the group. Before changes were merged into the main code repository, an effort was made to form consensus through a public, online discussion forum where working group members reviewed, commented on, and verified the changes. Once any open questions were resolved, the changes were merged and the issue was closed. Working group members automatically received emails that kept them apprised of the ongoing conversation, which helped members to contribute to pieces they were interested or had expertise in.

In Figure 4.20, the modeling process used by the NIDM working group is depicted. In this process, working group members manually mock up an example instance of a NIDM document, which is used to capture how concepts are linked together using the PROV standard. The example NIDM document, a text file in Turtle format, is checked a distributed version control system (i.e., GitHub) that enables working group members to discuss the document in an online forum, called a Pull Request. The Pull Request mechanism creates an issue that allows the example(s) to be discussed online, as well as during weekly videoconferences where issues are resolved and modeling decisions are made. After a decision is made the corresponding Google spreadsheet is updated and any additional issues are submitted to GitHub for discussion during the next weekly videoconference. This process continues as a given model matures until it stabilizes enough to be captured in an OWL file that is generated using Protégé and checked into GitHub.

Currently, the OWL file is populated with simple class and property hierarchies that map to PROV classes/properties. As changes are made to the OWL files in Protégé, a similar process takes place using Github Pull Requests with online/videoconference discussions. When a NIDM release occurs, the OWL files can be extracted, transformed, and loaded into NeuroLex. Still open is the question about how modifications and discussions that take place on NeuroLex can be incorporated into the general data modeling workflow. Additionally, the group has mostly moved away from using Google Spreadsheets at the current time, and is focusing on terms and definitions using Protégé only.

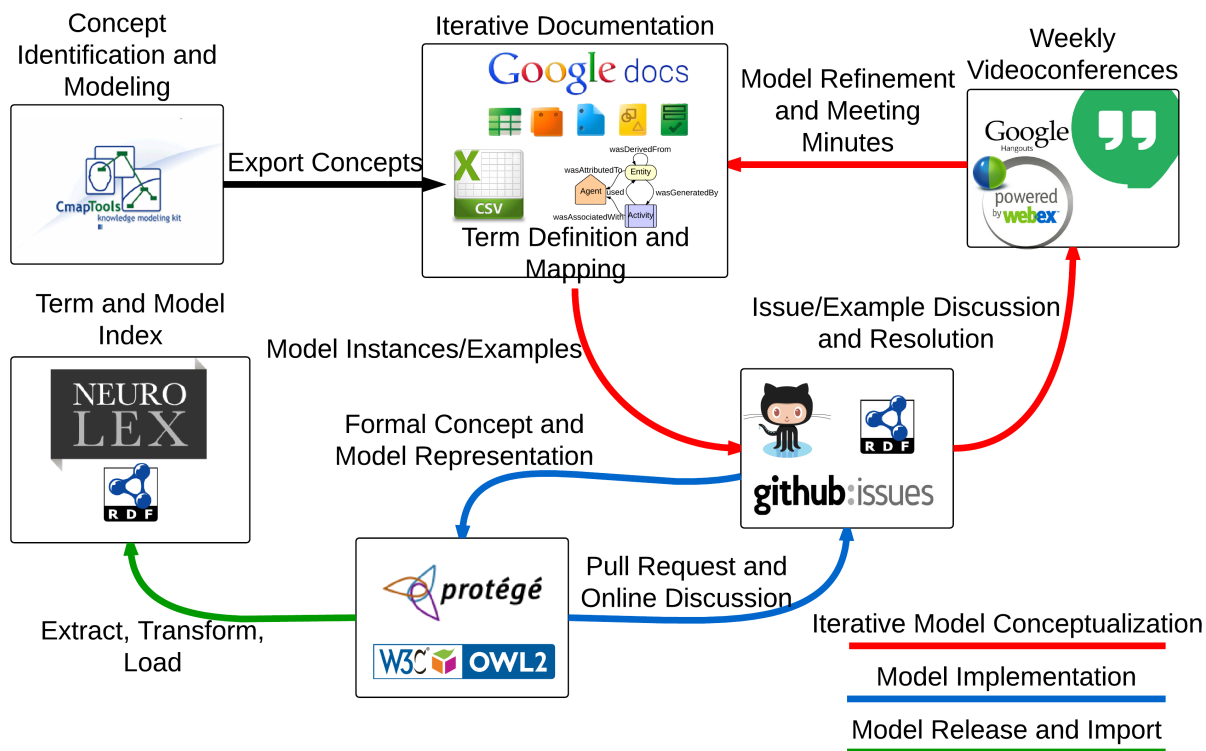


Figure 4.20. NIDM Modeling Process. The NIDM Working Group follows an iterative process to guide their data modeling efforts. The first stage is to clarify the general scope for an object model (e.g., reporting fMRI results) and then gather an initial set of concepts and relationships using the CMAP concept modeling tool. The initial CMAP diagram is then used to populate a Google spreadsheet with the concept terms and corresponding attributes, as well as an initial set of definitions and mappings to PROV ontology classes.

4.4 Conclusions

In this chapter, the preliminary studies completed during the SNI Phase were presented as the foundation for developing NIDM, a data modeling and standards development process completed during the NIDASH Phase. During the SNI Phase, the studies completed fulfilled the

first three sub-aims for Aim 1 to develop terms, data model extensions, and API specification. The NIDM framework emerged during the NIDASH Phase as a collection of conceptually distinct components linked together with provenance information. The NIDM Dataset Descriptor Component was introduced as mechanism to harmonize the NIDM Experiment, Workflow, and Results Components with the W3C Healthcare and Life Sciences working group efforts. The NIDM Experiment Component was then described with examples from the data sharing scenario that provides the use-case for developing NIDM and NiQuery. Finally, a description of the data modeling process to develop NIDM was presented. In the next chapter, the vision described in this scenario is pursued in the context of a neuroimaging Web application framework that facilitates tracking provenance and reproducible computational workflow.

Chapter 5: Neuroimaging application framework

"We're drowning in information and starving for knowledge"

- Rutherford D. Rogers, *American librarian*

5.1 Overview

In Chapter 4, a vision for how future graduate students will learn about brain imaging analysis was presented using a data sharing scenario. In the scenario, a second year PhD student needed to complete a number of tasks while rotating through a cognitive neuroimaging lab. The student was able to seamlessly transition between neuroimaging databases while conducting research because the databases and analysis tools each used the Neuroimaging Data Model (NIDM), a metadata and provenance exchange standard. Chapter 4 also reported the Aim 1 research outcomes by demonstrating how the data exchange standard described in the scenario came to fruition through research and development conducted by the INCF NIDASH task force. In this chapter, preliminary work completed during the SNI phase is reported in Section 5.2.1 before presenting a neuroimaging application framework that was developed to enable reproducible data analysis and provenance tracking. As in the previous chapter, the data sharing scenario will be the use-case to report study outcomes in the context of automated FreeSurfer analysis, where NIDM will be used as the underlying data model in a platform for neuroimaging software as a service.

5.2 Prototype information system study

In this study, the goals of Specific Aim 2 were addressed to develop an information system of Web services to compute and discover data provenance from brain imaging workflow. Four sub-aims were proposed for 1) creating a reference implementation of the vocabulary, data model, and API specifications; 2) developing an updated data management architecture of the NiQuery system; 3) implementing a prototype Web application to demonstrate system features for query, workflow execution, and provenance discovery; and 4) evaluating the overall positive and negative outcomes of the framework implementation. The work supporting these sub-aims was completed during the Scalable Neuroimaging Initiative (SNI) Phase, which is described in the

following section on research approach. The outcomes of these preliminary studies informed the system architecture design and implementation of the Web application framework referenced in the data sharing scenario in Chapter 4. In the outcomes and information system section, the details of the NiQuery framework are presented with examples that follow from the data sharing scenario. In this demonstration, a use-case follows how a user can submit a NIDM input document to a Web service that executes a FreeSurfer analysis. This description of the NiQuery framework provides an overview of the system components and their configuration for offering neuroimaging processing as a service.

5.2.1 Research approach

This section elaborates on the study design introduced in Section 3.4, Aim 2. It is organized, by sub-Aim, according to the specifications, system architecture, and prototype application. The section on specifications describes an overview and evaluation of the initial SNI protocol to enable remote, voxel-level access to neuroimaging data in a distributed context. The section on system architecture describes the technical implementation of the SNI protocol, while the section on a prototype application demonstrates the implemented system on a neuroimaging use-case for remotely interacting with neuroimaging time series data. The outcomes reported in each section provide a summary of the lessons learned and system requirements that emerged from the SNI research project, which took place over sixteen working group meetings, including two in-person workshops. The insights provided by this work guided the framework redesign and implementation reported in the outcomes and information system section.

5.2.1.1 Specification implementation

A basic requirement for any computational resource is a protocol that defines system requirements and behavior. In neuroimaging, resources are widely distributed across geographic areas, which requires researchers be provided with access to integrated imaging data with corresponding demographic and data acquisition information. Aside from textual data, the protocol must also define how to access binary images, or subsamples of images, using computational services. This adds a layer of abstraction where researchers need not visit

each database separately nor download whole datasets. To enable such distributed interaction with data, a study was conducted during the SNI phase to design a protocol that enables a scalable architecture by keeping computation close to the data and only transferring necessary data.

To define a protocol for distributed neuroimaging data access, a workshop and working group meetings were held to develop a conceptual model and identify key system requirements. Conceptually, the protocol consists of a collection of remote procedure calls (RPCs) to a neuroimaging database (e.g., XNAT). The server maps the objects specified in the RPC calls to a database wrapper layer that each individual database uses to provide access to its contents and services. A registry service provides an index of available database wrappers and methods to access additional objects. Data processing can take place at the data source or copied to a cache location (e.g., cloud) where additional computational services are supported. Only relevant results are transferred back to a client (e.g., single slice, voxel time series), which improves scalability when operating over many files. The protocol consists of three types of objects described in a draft specification and summarized below (<https://www.ibic.washington.edu/wiki/display/sni/Protocol>):

- **Session object** - wraps a neuroimaging database and exposes an API that acts as the primary interface to a given server and provides a mechanism to interrogate databases with user defined and/or predefined Web-accessible queries using a query management application (33). Results are parsed into '**DataContainer**' objects.
- **DataContainer object** - conforms to an 'image' data model as the primary interface to voxel-level data and related metadata, providing a mechanism to exercise methods remotely.
- **Workflow object** - a computational service on a '**DataContainer**' object that acts as the primary interface to computational tools that go beyond simple voxel retrieval and implements a workflow plugin type that is automatically attached to appropriate data containers.

In addition to defining these four objects, a key feature of the framework was to use Globally Unique Identifiers (GUID) for all resources available on the network. Each GUID is bound to a given object and provides a unique address to reference a given object and its provenance. For example, a '**Session**' GUID maintains a connection to a database server and, before expiring, can be used to construct '**DataContainer**' objects or execute '**Workflow**' objects, which would each be assigned a persistent GUID during initialization.

The SNI protocol provided the initial conceptual model used to pursue implementation ideas that reuse vocabularies, data exchange languages, and imaging data models. A literature review was conducted to identify open source projects that fulfilled these requirements and could aid rapid prototyping. Given the active role of the Python programming language in the neuroimaging community, preference was initially given to pursue Python libraries to implement this infrastructure. The outcomes of this study were used to design the prototype system architecture discussed in the next section.

5.2.1.2 Re-design of system architecture

To investigate the feasibility of the SNI protocol for providing distributed, voxel-level access to neuroimaging data, a prototype system architecture was designed and implemented (93). Open source software was evaluated to identify tools suitable for rapid prototyping the system requirements. This project aimed to 1) construct a '**Session**' object, 2) retrieve a listing of '**DataContainer**' objects from two or more databases, and 3) execute one '**Workflow**' RPC method to retrieve the time series from a single fMRI voxel from each '**DataContainer**' object. These core components will be used to enable the Web application described in Section 5.2.1.3.

The system architecture in Figure 5.1 depicts the additional services necessary to implement the SNI Protocol. A review of available distributed computing software identified the Python Remote Objects (Pyro4, <http://pythonhosted.org/Pyro4/>) library as a simple framework for transferring Python objects over the network, while also supporting GUIDs and a centralized registry. The Pyro4 library was used to implement the NiQuery Registry, which provides a '**Session**' object to the Pyro4-based Web application client. The '**Session**' object contains a list of

resources that represent the metadata in each common API endpoint. The Query Integrator (33) provides a number of queries that operate over this metadata to return 'DataContainer' proxy objects. These proxy objects can manipulate data via the Common API implemented by each neuroimaging database.

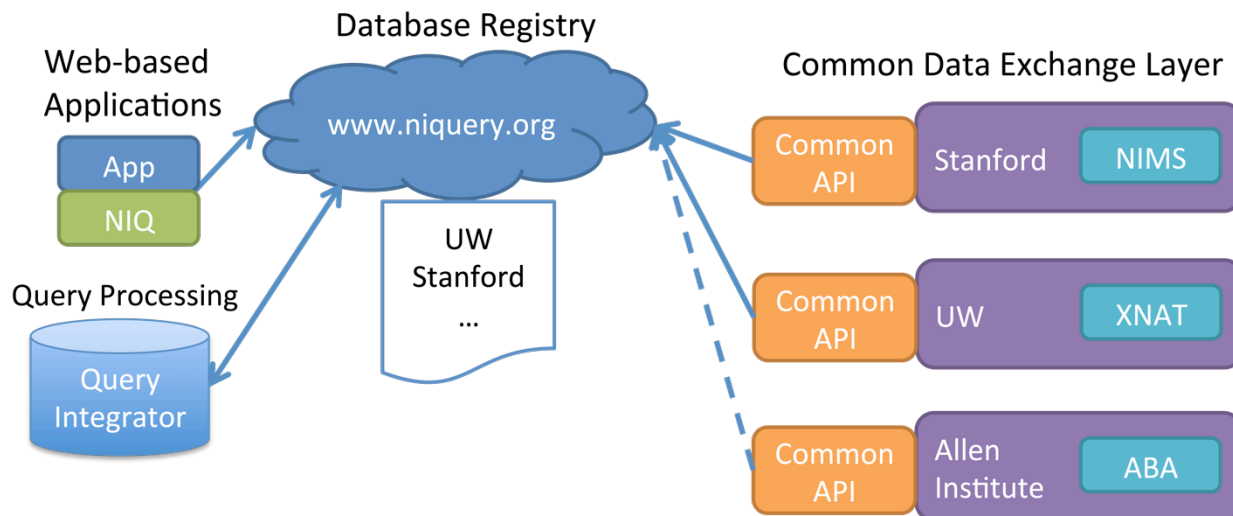


Figure 5.1. SNI System Architecture. Users can access a database registry that provides an index of neuroimaging databases exposing a common data exchange layer that can be queried using the query processing engine.

A Common Data Exchange Layer was needed to implement the Common API for the two databases in this study. XCEDE was used as a common schema to represent demographic and imaging information from both XNAT and the Neurobiological Imaging Management System (NIMS, <https://github.com/cni/nims>). The Common API wrapper was also implemented with Pyro; however, NiBabel (<http://nipy.org/nibabel/>) provided basic image manipulation of 'DataContainer' objects. Plans were made to use Nipype for 'Workflow' objects, but this was out of the scope for the initial prototype, as the requirements could be fulfilled with indexing in NiBabel. The outcomes of this study provided the necessary features to demo a prototype Web application and gather feedback for future development.

5.2.1.3 Prototype application

Prototype Web applications provide a meaningful way to assess the functionality of a protocol and system architecture. As discussed in Chapter 3, such information systems can be

considered experiments that indicate the strengths and weaknesses of a system and can be used to make incremental progress. To assess the impact and potential of the SNI Protocol and system architecture developed during the SNI Phase, two studies were conducted by designing and gathering feedback on prototype user interfaces that enable a researcher to 1) run a metadata-level query across two neuroimaging systems based on demographic information and 2) select coordinates from a brain template in standardized space and retrieve the time series of a single fMRI voxel.

To evaluate the framework for distributed metadata access, a study was conducted that attempted to answer the use-case query "Find all images from resting state fMRI studies where the participant is Female and 14 or older?" Figure 5.2. This simple query exercised the prototype NiQuery framework (Figure 5.1) by requiring the Web application to 1) create a '**Session**' object, 2) access XCEDE metadata representation of two neuroimaging databases, 3) submit a 'gender' and 'age' parameter to query processing, and 4) return the server location, sex, age, data, and viewer links. The framework successfully met the requirements to completed each of these tasks; however, there were a number of notable limitations:

1. The Pyro4 framework was difficult to configure when the Query Integrator (QI) was sufficient.
2. The query response was slower than direct QI execution due to creation of additional '**Session**' objects.
3. The XCEDE XML documents only supported a limited amount of metadata and would not validate if additional information was included from source schemas.

NIQuery Demo Application

[Home](#)

Location	Sex	Age	Data	View
niq.uw	female	68	https://s3-us-west-2.amazonaws.com/niquery/AnnArbor_sub00306/scan_rest.nii.gz	https://s3-us-west-2.amazonaws.com/niquery/AnnArbor_sub00306/scan_rest.html
niq.stanford	female	40	http://cni.stanford.edu/bobd/xcede/s002_func_rest.nii.gz	http://cni.stanford.edu/bobd/xcede/s002_func_rest.html
niq.stanford	female	59	http://cni.stanford.edu/bobd/xcede/s003_func_rest.nii.gz	http://cni.stanford.edu/bobd/xcede/s003_func_rest.html
niq.stanford	female	30	http://cni.stanford.edu/bobd/xcede/s004_func_rest.nii.gz	http://cni.stanford.edu/bobd/xcede/s004_func_rest.html

© Copyright NIQuery. <http://www.niquery.org>

Figure 5.2. NIQuery Prototype. The initial NIQuery prototype to demonstrate distributed metadata query using the SNI Protocol.

To evaluate the SNI protocol and system architecture for distributed, 'image-level' access, a study was conducted to provide researchers with a Web application (Figure 5.3). The requirements of this application were to use the NIQuery prototype system architecture to 1) create a **'Session'** object, 2) access XCEDE metadata representation of two neuroimaging databases, 3) provide a 3-plane browser to select voxel coordinates in standardized space, 5) construct a **'DataContainer'** proxy object via query processing for each fMRI scan in the XCEDE files, and 5) return and plot the time series for a single voxel in each scan (Figure 5.2). The outcomes of this study demonstrated an application that fulfilled all but one requirement. Requirement three had a technical limitation in the viewer that prevented voxels from being selected directly, requiring the use of form-based coordinate entry tools. The study also identified limitations in the overall approach as the Pyro4 framework illuminated scalability issues for computational tasks due to the slow serialization and deserialization of Python objects. These scalability issues are also noted in the use of XCEDE XML Schema at the metadata level, which was too rigid to exchange information in an academic setting with frequently changing or sparsely populated schema.

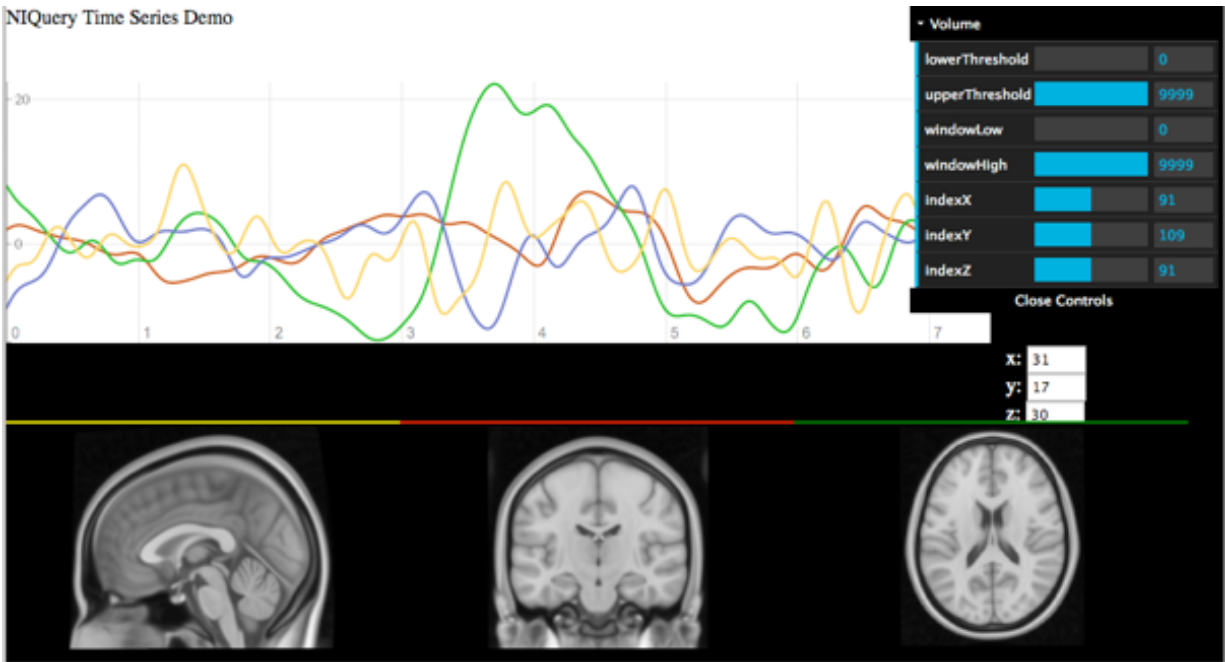


Figure 5.3. NiQuery Time series Demo. A Web application enables users to submit the coordinates of a voxel in standardized space and retrieve the time series from fMRI scans from distributed neuroimaging data sources.

The outcomes of these two preliminary studies from the SNI Phase provided an assessment of the XCEDE XML Schema as a data exchange standard and Pyro4 as a distributed computing platform for 'image-based' query. This prototype application and NiQuery's distributed architecture was demonstrated at the following meetings in order to gather feedback from stakeholders: the Allen Institute for Brain Science Hackathon 2012, Neuroinformatics 2012, Neuroscience 2012, and National Academy Keck Futures Initiative 2012. The feedback received was supportive of the vision for distributed computing in neuroimaging and SNI was encouraged to interface with other community efforts. At the metadata level, it was suggested to migrate from the XCEDE XML Schema to the INCF Neuroimaging Data Model (NIDM) detailed in Section 4.3.2, which is based on the semantic Web architecture discussed in Sections 2.4.2. It was also suggested to adopt a language agnostic computational infrastructure by using Web services and RESTful APIs, rather than the Python-centric Pyro4 framework. These lessons learned from community engagement, and the next generation of the NiQuery framework, are discussed in the following section on the NIDASH Phase outcomes.

5.2.2 Outcomes and information system

The research approach discussed in the previous section outlined the preliminary studies guiding the development of the initial NiQuery distributed computing prototype that was implemented and evaluated during the SNI Phase. The primary outcome of the SNI Phase indicated that the neuroimaging and neuroinformatics communities were interested in working collaboratively to identify a truly scalable set of technologies upon which to build. The scalability limitations raised during the initial NiQuery framework evaluation required a redesign of the vocabulary, data model, and system architecture to meet the requirements of the original SNI vision. The redesign of the vocabulary and data model was presented in Chapter 4, which described the Neuroimaging Data Model (NIDM) framework and provided examples of the NIDM Dataset Descriptor and NIDM Experiment Components. The outcomes discussed in the sections below describe the redesigned NiQuery system architecture and demonstrates an example of how NIDM is used during computational NiQuery tasks to track data provenance.

During the course of the SNI Phase, several disruptive technologies emerged as a dominant force of change for future scalable neuroimaging initiatives, including freely available applications for collaborative coding, continuous integration, automated system configuration, virtualized Linux containers, and on-demand computational infrastructure. These tools emerged as the communication between software developers and information technology operators was increasingly recognized as an essential component in any networked application infrastructure. A portmanteau of "development" and "operations," the DevOps software development method emerged in industry as an approach focusing on communication and the optimal use of tools for automation and virtualization of application components. From the perspective of biomedical informatics and academic software development, system deployment and configuration are major barriers to adoption by a given scientific community. Recognizing the importance of this emerging ecosystem of applications, with the goal of bringing Google-level scalability and availability to everyone, the NiQuery software development environment and system architecture adopted the DevOps method.

5.2.2.1 NiQuery system architecture

In this section, the redesigned NiQuery approach to software development is presented in the context of the data sharing scenario from Section 4.2. Each component in Figure 5.4 is discussed and the design and implementation perspectives are outlined. This provides the reader with a general notion of how information flows through the system, how neuroimaging workflows are configured, and how data provenance is tracked. In the next section, a specific example is used to demonstrate how a user would interact with the API to run FreeSurfer and view the results in prototype user interface that uses the NIDM Workflow and NIDM Results Components.

The system architecture diagram in Figure 5.4 provides an overview of the system components. These components enable a fully reproducible workflow engine that supports NIDM as an information retrieval and messaging protocol. The framework incorporates the DevOps methods discussed in the previous section and provides a fully automated deployment strategy that enables the system to scale from running on a laptop to a compute cluster in the cloud seamlessly. Following each letter in Figure 5.4, the student from the data sharing scenario in Section 4.2 could interact with the system by browsing NIDM documents as Linked Open Data (LOD) (Figure 5.4.A) to identify interesting datasets using a "follow your nose" approach by clicking on links in the user interface provided by the Virtuoso Universal Server (<http://virtuoso.openlinksw.com/>).

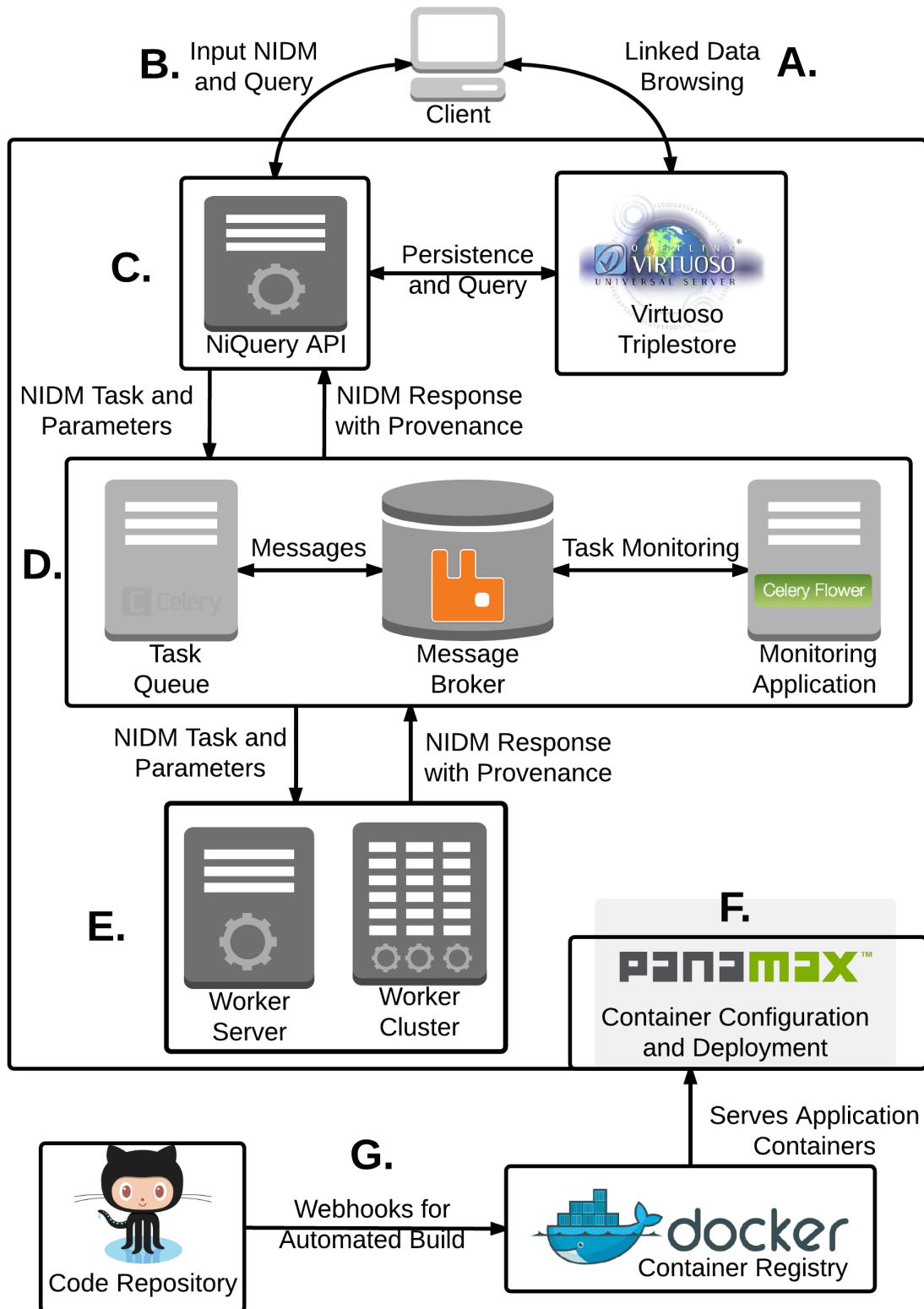


Figure 5.4. NiQuery System Architecture. High-level overview of the system components and products used to deploy NiQuery.

After identifying a relevant NIDM document, the student could open a Web application that provides access to a repository of computational workflows. By entering the Web accessible location of the NIDM document, the Web application downloads it, determines what workflows are compatible, and provides the student with a list. The student selects the FreeSurfer workflow from the list and clicks 'submit' (Figure 5.4.B). The browser app passes along the NIDM document and workflow type to NiQuery, and a SPARQL query is executed over the document to gather input files for this task (Figure 5.4.C).

The NiQuery API then submits the FreeSurfer job to a task queue that manages the distribution of incoming workflow requests and makes the results of completed or failed workflows available to NiQuery (Figure 5.4.C). The workflow messaging infrastructure is implemented with Celery (<http://www.celeryproject.org>), which uses RabbitMQ (<https://www.rabbitmq.com/>) as a message broker that keeps track of task status and provides monitoring information to Flower (<http://flower.readthedocs.org>), a graphical task monitoring utility (Figure 5.4.D).

To perform a given workflow task, a worker node or full worker cluster listens for messages posted to the message broker (Figure 5.4.E). Each worker is configured to perform specific tasks and will only consume a task from the queue that it is able to perform. The workflow tasks configured on each worker node are implemented using Nipype (83), which captures detailed information about the computational environment by producing provenance documents that conform to the NIDM Workflow Component. After a task finishes running on a worker, the NIDM document and provenance is passed back to the message broker, where NiQuery is notified the task is complete and then persists the NIDM provenance document to a Named Graph (119) in Virtuoso, where it is then made available as Linked Open Data (LOD).

The entire infrastructure described here is captured using a DevOps methodology that can be automatically tested, deployed, and scaled as necessary. The platform is implemented using Panamax (<http://panamax.io>, Figure 5.4.F), an open source application that allows developers to abstract away from physical hardware and virtual machines by enabling them to deploy services that can be composed into applications and captured as a preconfigured

template. In many ways, this enables a full application stack to be deployed as though it were a binary package.

This additional layer of abstraction is accomplished by enabling developers to build Linux containers that encapsulate individual services (Figure 5.4.G Right). Docker (<https://registry.hub.docker.com/>) popularized this approach by providing a simple file format that can import a base machine image (e.g., Ubuntu Linux) and install just the components necessary for an individual service. In this way, any Linux machine running Docker can attach a container to a running kernel and start a service within milliseconds, rather than start an entire virtual machine with the additional overhead of an operating system. Further, by checking the configuration files into a version control system (e.g., GitHub, <https://github.com>, Figure 5.4.g Left), Webhooks (<https://en.wikipedia.org/wiki/Webhook>) can be configured to trigger rebuilding Docker Containers and even entire Panamax templates. In this way, the NiQuery infrastructure provides a modern and scalable framework to develop and offer neuroimaging data analysis as a service.

5.2.2.2 Data sharing scenario requirements

In Chapter 4, NIDM was introduced as a framework for representing information about the neuroimaging research process as a flow of provenance information from data acquisition and analysis, to reporting statistical results. Chapter 4 also introduced a data sharing scenario to guide the NIDM Dataset Descriptor and NIDM Experiment examples, which provided detailed information about the Study Forrest project (8) hosted on OpenfMRI (106). In the following section, the data sharing scenario continues where it left off with a description of the NIDM Experiment Acquisition Element (Figure 4.19) to provide a specific example of interacting with the NiQuery application framework to run FreeSurfer. The requirements enabling this scenario are:

1. NIDM Acquisition object with an anatomical T1 MRI
2. SPARQL query that defines the interface to run FreeSurfer
3. NIDM Workflow object generated by Nipype

4. NIDM Results object of annotated FreeSurfer statistics
5. Prototype Web application to view the FreeSurfer results

5.2.2.3 NIDM workflow component

Each workflow interface in NiQuery is defined as a 'view' over one or more NIDM documents that returns a set of inputs and the name of a workflow task. To discover the workflow interfaces available, the NiQuery API provides a query library with metadata that is useful for query discovery. For example, Figure 5.5 shows the metadata describing the FreeSurfer workflow interface. To implement the interface to FreeSurfer, a query was written that searches for a NIDM Acquisition Object with a 'dct:downloadURL' link to an anatomical T1 MRI, as illustrated by the NIDM document in Figure 4.23. When executed against a NIDM document, the workflow query in Figure 5.6 will return list of all 'nidm:MRIDAnatomicalT1' images available, along with the NiQuery task to process the images with.

```
rq:2c5782ea-7b30-11e4-b2ce-67e96e88432a
a niq:ComputeQuery ;
dct:title      """"Compute FreeSurfer recon-all for
                every T1 image in the given rdf"""" ;
dct:description "Query a graph for T1 file URIs in NIfTI GZipped format." ;
dct:creator    <http://orcid.org/0000-0003-1099-3328> ;
dcat:format    niq:Select ;
dcat:downloadURL <http://purl.org/niquery/id/2c5782ea-7b30-11e4-b2ce-67e96e88432a> ;
dcat:keyword   "query", "compute", "freesurfer", "recon-all" ;
niq:model     niq:Experiment ;
niq:columns    "?t1_uri ?task" .
```

Figure 5.5. Query Metadata for FreeSurfer. (<http://purl.org/niquery/id/meta.ttl>)

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX nidm: <http://purl.org/nidash/nidm#>

SELECT DISTINCT ?t1_uri ?task
WHERE {
  ?uri a nidm:MRIDAnatomicalT1 ;
       dcat:downloadURL ?t1_uri ;
       dcat:format nidm:NIFTI1Gzip .
  VALUES (?task) {"tasks.fs_reconall"}
}
```

Figure 5.6. FreeSurfer Workflow Interface Query. (<http://purl.org/niquery/id/2c5782ea-7b30-11e4-b2ce-67e96e88432a>)

After the workflow interface query returns the download link and interface name, NiQuery submits a job to the task queue and a worker with FreeSurfer installed downloads the

T1 MRI file. Before starting to process the data with Nipype, a checksum is calculated and a query checks for previously computed results and only continues if there is no match (Figure 5.7). When the task completes, Nipype generates a NIDM Workflow object (Figure 5.8) with Entities, Activities, and Agents representing the computational environment and the files produced during analysis.

```
SELECT DISTINCT ?graph ?interface ?sha512
WHERE {
  GRAPH ?graph
  {
    ?file
      crypto:sha512 ?sha512 ;
      a prov:Entity .
    ?activity
      prov:used ?file ;
      nipype:interface ?interface .
  }
}
```

Figure 5.7. Check cache for previously computed results. (<http://purl.org/niquery/id/E0921842-1EDB-49F8-A4B3-BA51B85AD407>)

For example, Figure 5.8 shows the Nipype representation of such a Workflow object. The first object (*'niiri:df946b...'*) describes the T1 MRI file that was used as input as a **'prov:Entity'** that includes the shasum attribute used in Figure 5.7. The second object (*'niiri:fe69e60...'*) showing the Nipype interface to FreeSurfer is captured as a **'prov:Activity'** with a variety of attributes that detail the computational environment the interface was executed on. For example, the command line arguments (i.e., *'nipype:command'*) and version of the command line executable (i.e., *'nipype:version'*) are captured. Also included are links to the input files (i.e., *'prov:used'* *'niiri:df946b...'*) and the third object, **'prov:SoftwareAgent'** (*'niiri:0e3b89'*), is responsible for execution (i.e., *'prov:wasAssociatedWith'* *'niiri:0e3b89...'*), which includes the versions of software dependencies and platform information.

```

niiri:df946b6fa060c2c7bb693382a937e19b
  a prov:Entity ;
  crypto:sha512 "de0d52...20e0c5f1c4649dc76d4dff68" ;
  prov:atLocation <file://glia.localhighres001.nii.gz> ;
  prov:value <file://glia.localhighres001.nii.gz> .

niiri:fe69e607689d11e4ae1ab8e856385718
  a nipype:NipypeInterfacesFreesurferPreprocessReconall,
    prov:Activity ;
  rdfs:label "ReconAll" ;
  nipype:command "recon-all -i highres001.nii.gz -subjid sub001 -sd ." ;
  nipype:commandPath <file://.../Applications/freesurfer/bin/recon-all> ;
  nipype:interface "ReconAll" ;
  nipype:module "nipype.interfaces.freesurfer.preprocess" ;
  nipype:platform "Darwin-14.0.0-x86_64-i386-64bit" ;
  nipype:version "1.379.2.73" ;
  nipype:workingDirectory <file://.../Users/nolan/Downloads/openfmri> ;
  prov:endTime "2014-11-10T05:54:10.831091"^^xsd:dateTime ;
  prov:used niiri:df946b6fa060c2c7bb693382a937e19b ;
  prov:wasAssociatedWith <http://iri.nidash.org/0e3b895483...95dc68f8fe17> .

niiri:0e3b895483485b3c6c0a95dc68f8fe17>
  a prov:Agent,
    prov:SoftwareAgent ;
  rdfs:label "Nipype" ;
  nipype:networkx_version "1.9.1" ;
  nipype:nibabel_version "1.3.0" ;
  nipype:numpy_version "1.9.1" ;
  nipype:pkg_path <file://.../nipype/lib/python2.7/site-packages/nipype> ;
  nipype:scipy_version "0.14.0" ;
  nipype:sys_executable <file://.../nipype/bin/python> ;
  nipype:sys_platform "darwin" ;
  nipype:sys_version ""2.7.8 |Continuum Analytics, Inc. |
    (default, Aug 21 2014, 15:21:46)
    [GCC 4.2.1 (Apple Inc. build 5577)]"" ;
  nipype:traits_version "4.4.0" ;
  foaf:name "Nipype" .

```

Figure 5.8. Nipype Entity, Activity, and SoftwareAgent.

While an implementation of NIDM Workflow exists in Nipype, it is still an early implementation and will require additional standardization. Currently, all of the terms in the Nipype namespace are undefined and are not captured in a controlled vocabulary; rather they are artifacts of the underlying Nipype workflow engine. However, the level of detail captured in this provenance document provides computational access to the conditions under which data was processed. Even without standardization, NIDM Workflow objects can enable applications

to gain access to information that facilitates reproducibility and helps drive the creation of NIDM Results models.

5.2.2.4 NIDM results component

The NIDM Results Component captures the output of computational processes by defining object models for specific applications. An object model can be considered a 'view' over an underlying data model, in this case a NIDM. In the data sharing scenario, the driving need is for the student to generate a report for a lab meeting based on the analysis generated by FreeSurfer. The outputs from this process include a standard directory structure that contains brain surface reconstructions, parcellated and segmented brain imaging label maps, and statistical measures of anatomical structures. One approach would be to create a NIDM representation of the directory structure and extract statistical data from text files that can be displayed in a Web application.

To create a NIDM results object model that represents a directory structure, directories can be represented as a '**prov:Collection**' and files can be represented as a '**prov:Entity**'. Each of the files are then related to a directory using '*prov:hadMember*'. This pattern can be used recursively to gather up all the files in the FreeSurfer directory structure to be further annotated with attributes and type information. In the case of statistical text files, a NIDM representation can be constructed that models the files contents by anatomical structure. For example, Figure 5.9 demonstrates a representation of statistical measures, such as surface area (i.e., "*fs:SurfArea*") and gray matter volume ("*fs:GrayVol*"), that can be captured along with an anatomical annotation. As discussed in Section 4.3.1.2, annotating data with terms from an ontology (e.g., the FMA) enables developers to explore relationships and mappings between terminologies. In this case, the '*nidm:AnatomicalAnnotation*' is linked to a term from FreeSurfer, but the mappings in the FMA would allow exploration of the statistics calculated on other connected structures.

```

<http://nidm.nidash.org/iri/3abb28e0d3d011e2a755001e4fb1404c>
  a prov:Entity ;
  fs:CurvInd "6.2"^^xsd:float ;
  fs:FoldInd "63"^^xsd:integer ;
  fs:GausCurv "0.097"^^xsd:float ;
  fs:GrayVol "2078.0"^^xsd:float ;
  fs:MeanCurv "0.152"^^xsd:float ;
  fs:NumVert "1438"^^xsd:integer ;
  fs:SurfArea "715.0"^^xsd:float ;
  fs:ThickAvg "2.436"^^xsd:float ;
  fs:ThickStd "0.519"^^xsd:float ;
  nidm:AnatomicalAnnotation fs:ctx-rh-superiorfrontal .

```

Figure 5.9. NIDM Results representation of FreeSurfer statistics files.

The structured and annotated information made available by NIDM and generated by the NiQuery application framework provides developers with a rich toolset. Using an application like the prototype in Figure 5.10 (120), the student in the data sharing scenario would be able to 1) query NIDM documents for T1 MRI scans that describe the Study Forrest Project hosted by OpenfMRI, 2) remotely process the scans using FreeSurfer, and 3) analyze or review the results and report findings during the weekly lab meeting. This particular Web application was developed using different system architecture than the NiQuery application framework; however, a similar application could be developed using the NiQuery API to perform the same tasks.

For example, the "Query" tab in Figure 5.10 could access the NiQuery API and render a list of queries from the query library using the metadata illustrated in Figure 5.5. By uploading a NIDM Experiment document (e.g., Figure 4.19) the application could validate which workflows are compatible with the document by verifying that a compute query returns all of the necessary inputs (e.g., by executing the query in Figure 5.6 against the NIDM document). Then a list of compatible workflows could be displayed under the "Process" tab in Figure 5.10, which could be triggered to run the workflow and return a NIDM Workflow object (e.g., Figure 5.8). Finally, the "Analyze / Review" tab could automatically extract and display a NIDM Results object (e.g., Figure 5.9) and retrieve any surface or volumetric data generated by the workflow to display in the "DataView" tab.

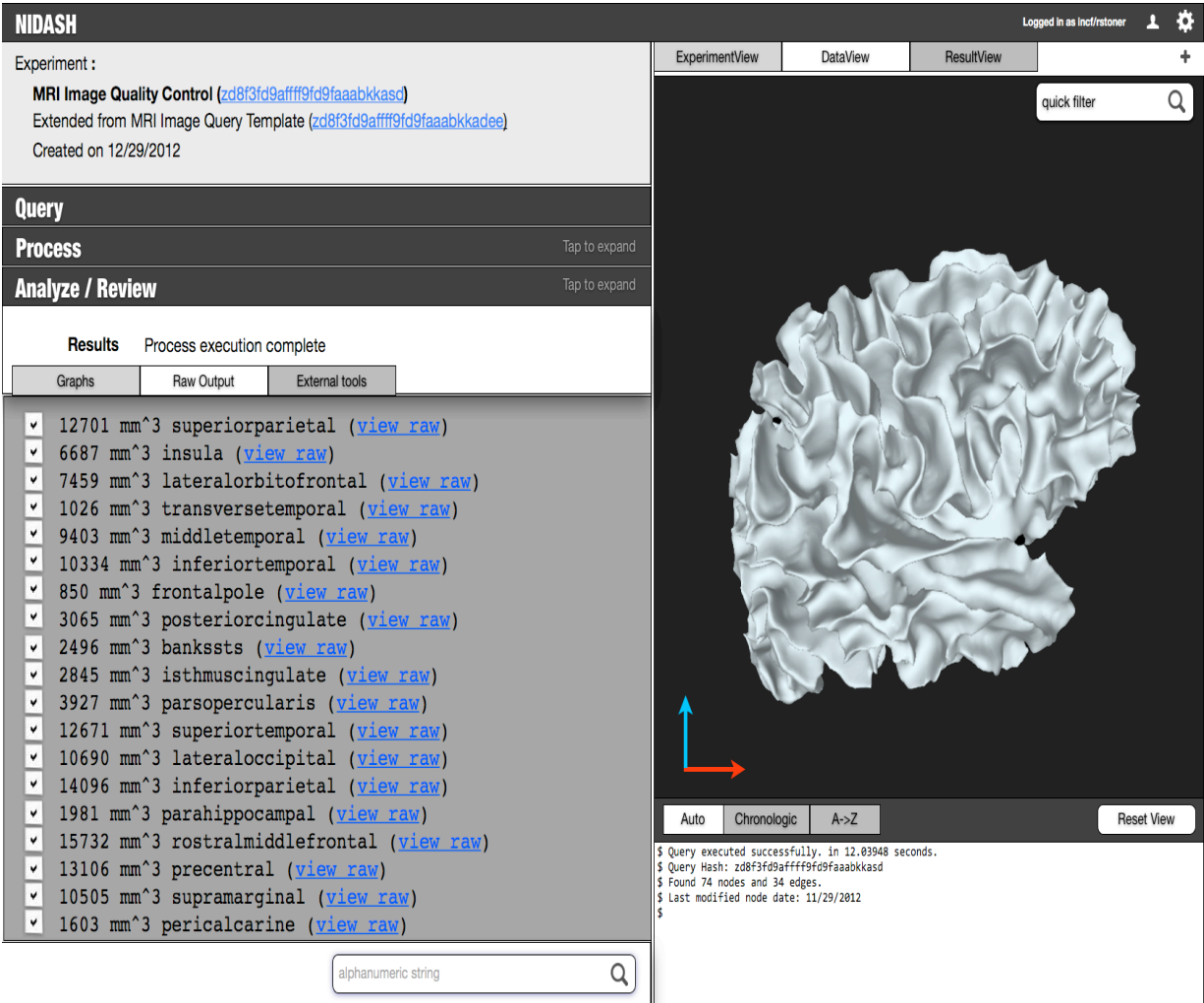


Figure 5.10. Prototype App to work with NIDM and NiQuery. (https://github.com/richstoner/incf_engine)

5.3 Conclusions

In this chapter, the preliminary studies completed during the SNI Phase were presented as the foundation for developing the NiQuery application Framework, a modern computational platform built on top of NIDM and completed during the NIDASH Phase. During the SNI Phase, the studies completed identified the limitations of the XCEDE XML Schema for modeling metadata and Pyro4 python remote objects library for distributed computing. The redesigned NiQuery application framework emerged during the NIDASH Phase due to the rapid development and availability of open source tools for Big Data management and maturing semantic Web technologies. The data sharing scenario introduced in Chapter 4 was used to demonstrate how the NiQuery application framework enabled a workflow to be configured

using NIDM as a data source, which produced provenance as NIDM representing the computational environment. Finally, a NIDM Results object model was described that captures statistical measures annotated with terms from an ontology, which was used in a prototype Web application to show summary statistics and 3D surfaces from the FreeSurfer analysis. In the next and final chapter, a summary of this dissertation is presented that focuses on my specific contributions and impact, as well as the limitations and future directions of the research I have completed.

Chapter 6: Conclusions and summary

"Making something and sending it out into the world and then people not only responding to it but adopting it for their own and making a separate thing for it, that's beautiful. It just shows you how much you can affect other people... the butterfly effect of everything you put out into the world."

- Marketa Iglova, Czech singer-songwriter, musician and actress

6.1 Dissertation summary

In this dissertation I pursued a vision for reproducible human cognitive neuroimaging research that captures the scientific process as provenance information. I presented the Neuroimaging Data Model (NIDM) and NiQuery as components that enable representing and tracking information from data acquisition to workflow and results. This framework is the start of an international, community-driven effort to address the growing concerns that brain imaging is in a reproducibility crisis.

In Chapter 2, I provide an introduction to interoperability challenges and data sharing barriers germane to cognitive neuroimaging, neuroinformatics, and biomedical informatics. In cognitive neuroimaging, scientists experience ever-evolving instrumentation and data analysis methods that make standardization and reproducibility a moving target. At the same time, neuroinformatics strives to develop software applications that adopt the crosscutting technologies and approaches provided by the many application domains of biomedical informatics. My review of these issues highlight a variety of previous approaches in the literature and casts light on emerging and maturing technologies that I apply in my dissertation to overcome barriers to data sharing.

In Chapter 3, I introduce the overarching framework of evaluation in biomedical informatics and identify three representative stakeholder communities that I work collaboratively with throughout my dissertation. In the evaluation component, I discuss how the implementation of an information system is considered part of the experimental design in informatics, where the positive and negative outcomes of deploying a system lead to a system redesign. This redesign then becomes another experiment and contributes to incremental

progress. Throughout my dissertation, I followed this pattern of design, deploy, evaluate, and redesign with both the data model and application framework I developed. This chapter also describes my interaction with the Integrated Brain Imaging Center and Structural Informatics Group at UW, as well as the International Neuroinformatics Coordinating Facility (INCF). The feedback and guidance I received from these three organizations and research groups were instrumental in the successful completion of my research.

In Chapter 4, I introduce a data sharing scenario that provides a vision of the future I would like to enable using NIDM as a data exchange mechanism. While developing NIDM, my contributions were evident in my leadership role during meetings, and I was one of the original advocates of moving away from the XCEDE XML Schema, along with Satra Ghosh. It was my research, attempting to harmonize the XCEDE and PROV XML Schemas, which provided the insight to rethink our data modeling efforts and envision PROV as a primitive from which to design NIDM. I also organized weekly NIDM Experiment calls, where I proposed and garnered support for the NIDM Dataset Descriptor and NIDM Experiment data structures reported in this dissertation.

In Chapter 5, I provide a detailed discussion about the evolution of the NiQuery application framework, which evolved from a simple system to a modern architecture designed for Web scale. During the SNI project, I held a leadership role by organizing each meeting and making implementation decisions, which ultimately led to the redesign. By presenting the initial NiQuery prototype to the INCF and neuroinformatics communities, I was able to identify experts in distributed cloud computing and educate myself in this rapidly changing field. With this experience, I was able to redesign NiQuery with a sophisticated, semantic Web-based system architecture that applies scalable components that are generally only available in an industry setting.

I lend my success in evolving both the data exchange and application frameworks to the community of volunteers who participate in the INCF Neuroimaging Data Sharing task force and openly contribute their ideas to the betterment of informatics in cognitive neuroimaging. The support of this community has led to NIDM being incorporated into the three top fMRI analysis

packages and will likely emerge as the standard for data exchange and representing data provenance in the domain of cognitive neuroimaging.

6.2 Implications for cognitive neuroimaging

The work that I have presented in this dissertation provides a foundation upon which the next generation of databases and analytical tools will be built for the field of human cognitive neuroimaging. The Neuroimaging Data Model (NIDM) I designed is a crucial innovation for reproducible brain science that not only overcomes technological barriers to data sharing, but has ignited a grassroots and community-driven effort engaging scientists to participate in the standards development process. The NiQuery application platform I developed provides a proof of concept system architecture that enabled neuroimaging data analysis as a service, where NIDM is the underlying representation, capturing provenance, to furnish a complete picture of the research process. My work also lends to a vision for the future of human brain imaging with the data sharing scenario I use to describe NIDM and NiQuery. Coupled with my leadership role that I took in developing of these products, my work has and will speed along the field of neuroimaging towards a future of interoperable neuroinformatics resources that will accelerate and streamline research objectives for discovery.

6.2 Implications for biomedical informatics

My dissertation research investigated the underrepresented role of provenance in biomedical informatics methods and demonstrated a vision for how informatics infrastructure can provide reproducible, computational services for biomedical researchers. I applied existing biomedical informatics methods in knowledge representation and standards development to incorporate the Foundational Model of Anatomy ontology into the Neuroimaging Data Model (NIDM) for annotating measures of neuroanatomy. I also took a leadership role in the INCF NIDASH task force by evaluating existing data exchange standards, researching emerging technologies, and experimenting with novel representations of data that led me to invent NIDM. I brought the biomedical informatics framework I received in my training to the neuroinformatics

community, consisting of primarily neuroscientists with computer science or engineering backgrounds, rather than formal doctoral training in biomedical informatics.

6.4 Limitations

In this dissertation, I demonstrated NIDM and NiQuery using a data sharing scenario that describes a vision for interoperable databases and reproducible neuroimaging data analysis as a service. The data sharing scenario describes a set of use cases to be completed by a student who used a combination of NIDM and NiQuery to 1) access anatomical MRI scans from OpenfMRI, 2) process the scans with FreeSurfer while tracking provenance, and 3) retrieve an annotated listing of brain region measures. These use cases only represent one imaging modality, analysis software, and set of measures, whereas in a real neuroimaging laboratory there may be many more use cases possible. Working with collaborators in the INCF NIDASH task force, I chose these use cases because of their simplicity in comparison to more advanced analyses and to highlight the role of provenance in NIDM. While the work I completed provides a proof of concept, I did not create a complete collection of Object Models to fully implement the vision of reproducible, standards-driven neuroimaging as a service. The grand vision of interoperable, standards driven data science in human brain imaging was beyond the scope of a single dissertation; however, the demonstrated adoption of NIDM by the neuroimaging community, and ongoing development by the INCF NIDASH task force, make the likelihood of additional NIDM extensions likely.

6.4.1 Data model limitations

My leadership role in developing NIDM provided me with the opportunity to engage the NIDASH task force members with use cases and development approaches. However, the grassroots and volunteer basis of task force members meant that the task force priorities and membership were subject to change according to available resources. The use cases I pursued for my dissertation were a high priority when I designed the NIDM framework, but, after presenting the work at the annual meeting of the Organization for Human Brain Mapping (OHBM) and the Neuroinformatics congress in 2013, additional contributors joined the task force to pursue the

specific use case of automating neuroimaging meta-analysis. This use case required a standardized machine-readable format for fMRI results that was captured by extending the NIDM Results Component for task-based fMRI.

The initial charter for NIDM was defined as a data interchange format to model raw experiment metadata from neuroimaging data management systems that are capable of also describing data provenance and derived data. By pursuing meta-analysis and minimum information reporting for task-based fMRI as a clear-cut use case, the priorities and time allocation of the working group shifted towards the newly available resources. Although there was a departure from the initial NIDM approach, the concept of applying the NIDM framework to meta-analysis engaged additional stakeholders and community involvement, as well as a number of unforeseen use cases where NIDM could be used as a data interchange between databases hosting neuroimaging results (e.g., activation loci), rather than the raw data and acquisition parameters. While this change in direction impacted the relative maturity of my contributions to NIDM in comparison to the fMRI results use case, it demonstrates the generalizability of my approach and of NIDM as a whole.

The technologies I used to develop NIDM also presented some limitations. Contributors working with the Git version control system (<http://www.git-scm.com>) for the first time experienced a steep learning curve. This learning curve limited the level of involvement for some individuals when working directly with NIDM documents that were checked into version control. This effect was notable for both programmers (e.g., statisticians) and non-programmers (e.g., psychologists) in the working group; however, these same individuals contributed effectively to discussions by using online comment threads (e.g., GitHub Issues and Google Spreadsheets).

A key limitation of the research I presented on NIDM is the lack of a formal evaluation that measures the impact of this data sharing standard in support of reproducibility. I took a participatory role in the international neuroinformatics and neuroimaging communities while developing NIDM, and I did not perform any observational studies or make any specific measurements that would indicate the impact of this work. However, the work I completed with

NIDM catalyzed the neuroimaging community's involvement with standards development and neuroinformatics tools can now adopt NIDM. With NIDM now in place, it will be possible to measure the impact on data exchange efforts and provenance tracking.

6.4.2 Application framework limitations

When I developed the NiQuery application platform, the evolving landscape of information technology, in combination with an agile software development approach, made implementing a robust system difficult. As I identified flaws in the system design, newly available open source technologies emerged that solved my technical requirements for NiQuery. While I found it convenient that these novel approaches were available, they were also disruptive to my collaborative approach. Each time the developers working on the system infrastructure came up to speed with a given technology, it was ready to move on to the next platform, framework, or service. For example, the three most disruptive technologies were from the DevOps (i.e., software development and IT operations) industry, which provided novel open source tools for automated system provisioning (e.g., Puppet, Chef, Ansible), virtual machine configuration (e.g., Vagrant), and Linux containers (e.g., Docker). These technologies dramatically improve how systems are developed and deployed, but come with a steep learning curve that changed the way the NiQuery platform was conceptualized. From a long-term perspective, I welcomed these disruptive technologies, but I found it difficult to develop anything beyond a proof of concept within the timeframe of my dissertation.

As with NIDM, a limitation of this work is the lack of a formal evaluation that identifies the impact of the NiQuery application framework on scientists' ability to perform reproducible research. The current technology stack that I used to deploy NiQuery is stable and ready for further development with specific use cases that could be formally studied. For example, the recent work on a NIDM Results model for task-based fMRI could be incorporated into NiQuery to produce a summary of an fMRI analysis. As use cases are identified, NiQuery can be extended in a problem-driven manner.

6.5 Future directions

My dissertation research introduced an extensible, new data exchange standard and computing platform designed for members of the human cognitive neuroimaging community. Within this community, individuals have already demonstrated that NIDM can be extended through the ongoing development of an Object Model for reporting task-based fMRI results. The vision behind this extension is to enable automated meta-analysis by encoding all of the information necessary to aggregate the statistical results from fMRI studies. There is a need to identify additional use cases for both NIDM and NiQuery with specific applications in mind, such as automated meta-analysis, to drive future development. There is also a need to evaluate similar data models, semantic Web frameworks, and computational infrastructures to ensure that NIDM and NiQuery remain harmonized with similar efforts.

6.5.1 NIDM object model extensions

The use of domain object models in NIDM provides a mechanism to continually evaluate and evolve the vocabulary used to describe the provenance of neuroimaging analyses. As the neuroimaging community moves towards more structured methods of reporting, there will be a need for new methods to be incorporated into the NIDM framework, which will require an approach to manage the provenance of NIDM terms that change. The collaborative data modeling framework I developed solves part of this problem by using the Git version control system and GitHub web application, particularly Pull Requests; however, a more detailed representation will be needed to capture versions of new models as they arise. Vocabulary and ontology versioning and migration will become increasingly necessary and although some tools do exist to address this issue (e.g., see the `simple-virtuoso-migrate` tool: <https://github.com/globocom/simple-virtuoso-migrate>), none are mature enough for use in a production environment. Thus, future research should address how communities that collaboratively develop data models and ontologies for describing their research can accurately capture the provenance of these information artifacts in a meaningful way.

Funding is needed to support collaborative modeling efforts to develop the standards needed for NIDM to work in the long term though continued efforts with organizations like

INCF. Additionally, it should be required that all NIH funded research provide a listing of standard data elements used in their research in advance of study initiation, perhaps as part of the grant application, unless the research aims to create novel measures. For the latter, researchers should be required by funding and publishing bodies to register any novel data elements in a community-vetted data element repository before publication. A registry of common data elements is necessary for any standards driven data integration efforts. This is similar to how the molecular biology community is required to register new compounds that are discovered. NIDM could provide the framework for registering such data type information in a community-driven way.

While I was developing NIDM, there were several similar efforts that emerged in other domains. In molecular biology the Investigation, Study, Assay (ISA) approach gained popularity and was adopted by Nature Publishing (105); however, ISA is not tailored for neuroimaging and lacks a deep representation of data provenance. In computer science, the concept of a "Research Object" (121) provides a collection of all the files and metadata associated with a study and shares many of the core ideas behind NIDM Components. Micropublications are another emerging approach to capturing "claim" networks (122) that model hypotheses and supporting research. Related to micropublications is the SemanticScience Integrated Ontology (SIO, (123)), which enables the capture of biomedical knowledge, as well as the integration of semantically annotated datasets and Web service discovery. As NIDM matures there is an opportunity to incorporate and harmonize the approaches taken by these similar, yet distinct projects.

6.5.2 NiQuery software development

To realize the vision for neuroimaging data analysis as a service, future software development on the NiQuery application framework is necessary. The current prototype I demonstrated is a proof-of-concept for using computational, semantic Web services in a standards-driven, scalable, and extensible information system for neuroimaging data analysis. While the core components of this framework are in place, the next iteration of software development will focus on incorporating emerging standards for describing semantic Web services, expanding to

additional workflow systems, and introducing security measures. I chose the previous use cases to explore the use of NIDM in a computational infrastructure, but, moving forward, scenarios should be identified by stakeholders with real tasks that need to be completed during their daily research. Moving development out of the informatics lab and into practice will provide more accurate ongoing feedback about the NiQuery application framework.

Semantic Web services and software tools for managing linked data are increasingly available and need to be evaluated for potential incorporation into NiQuery. The Semantic Automated Discovery and Integration (SADI) framework (124) provides an ontology and software tools that can be used to describe a computational service as an RDF triple, where the "Subject" is the input, the computational workflow is the "Predicate," and the output is the "Object." Another emerging technology is the W3C Linked Data Platform (<http://www.w3.org/TR/ldp/>), which provides an HTTP-based protocol for read/write linked data management that allows files and RDF metadata to be stored on the same server and browsed in a similar way as linked Web pages. By incorporating these approaches into NiQuery, the system will become more accessible to the research community and to application developers.

Currently, NiQuery only supports Nipype as a workflow engine. Generalizing the PROV workflow model to systems beyond Nipype will be essential for broad adoption in the research community. For example, labs that use Make to compile their workflow could include a script that converts Makefile targets into RDF (e.g., <http://www.w3.org/2000/10/swap/util/make2n3.py>) and mapped to a similar format as Nipype. By using PROV as a core model, this would make it possible for provenance-focused queries to be interoperable across workflow systems, but further work is needed to standardize the representation of workflow code.

By working with openly available data, I did not need to address security concerns within the NiQuery system design. A future version of NiQuery will need to achieve better accountability using authentication and authorization mechanisms. Using security protocols, such as HTTPPA (Accountable HTTP, (125)) will enable NiQuery to be used in applications that

include both public and private data collections. This approach will help to realize the vision of the read-write Linked Data Web (126) that requires authentication and authorization using the WebID Protocol (<http://www.w3.org/TR/webid>).

6.6 Final conclusion

My dissertation research produced informatics standards and infrastructure in order to overcome socio-technical barriers to neuroimaging data sharing and threats to reproducibility. I developed the Neuroimaging Data Model (NIDM), a fundamentally new, granular data exchange standard that provides a language to communicate provenance by representing primary data, computational workflow, and derived data. By incorporating provenance as a primitive to document cognitive neuroimaging workflow, I introduced an informatics method to make statements that parsimoniously express the way a given piece of data is produced. I conducted this research in partnership with international neuroinformatics and brain imaging communities, which adopted and extended central elements of my work into prevailing brain imaging software.

References

1. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature*. 2014 Jan 30;505(7485):612-3.
2. Field D, Sansone S-A, Collis A, Booth T, Dukes P, Gregurick SK, et al. Megascience. 'Omics data sharing. *Science*. 2009 Oct 9;326(5950):234-6.
3. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform. Elsevier Science*; 2008 Oct;41(5).
4. Piwowar HA. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*. 2011;6(7):e18657.
5. Van Horn JD, Grethe JS, Kostelec P, Woodward JB, Aslam JA, Rus D, et al. 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*. 2011 Aug 29;356(1412):1323-39.
6. Mennes M, Biswal BB, Castellanos FX, Milham MP. Making data sharing work: the FCP/INDI experience. *NeuroImage*. 2013 Nov 15;82:683-91.
7. Jack CR, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of magnetic resonance imaging: JMRI*. 2008 Apr;27(4):685-91.
8. Hanke M, Baumgartner FJ, Ibe P, Kaule FR, Pollmann S, Speck O, et al. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci Data*. Nature Publishing Group; 2014 May 27;1.
9. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*. 2007 Sep;19(9):1498-507.
10. Klein A, Tourville J. 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience*. 2012;6:171.
11. Gardner D, Toga AW, Ascoli GA, Beatty JT, Brinkley JF, Dale AM, et al. Towards effective and rewarding data sharing. *Neuroinformatics*. 2003;1(3):289-95.
12. Breeze JL, Poline JB, Kennedy DN. Data sharing and publishing in the field of neuroimaging. *Gigascience*. 2012;1(1):9.
13. Poline JB, Breeze JL, Ghosh S, Gorgolewski K, Halchenko YO, Hanke M, et al. Data sharing in neuroimaging research. *Front Neuroinform*. 2012;6:9-9.
14. Smith K, Jajodia S, Swarup V, Hoyt J, Hamilton G, Faatz D, et al. Enabling the sharing of neuroimaging data through well-defined intermediate levels of visibility. *NeuroImage*. 2004;22(4):1646-56.
15. Carp J. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage*. Elsevier; 2012;63(1):289-300.

16. Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med.* 2005;2(8):e124.
17. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience.* 2013 May;14(5):365–76.
18. Ioannidis JPA. Excess significance bias in the literature on brain volume abnormalities. *Arch Gen Psychiatry.* 2011 Aug;68(8):773–80.
19. David SP, Ware JJ, Chu IM, Loftus PD, Fusar-Poli P, Radua J, et al. Potential reporting bias in fMRI studies of the brain. *PLoS ONE.* 2013;8(7):e70104.
20. Poldrack RA, Fletcher PC, Henson RN, Worsley KJ, Brett M, Nichols TE. Guidelines for reporting an fMRI study. *NeuroImage.* 2008 Apr 1;40(2):409–14.
21. Peng RD. Reproducible Research in Computational Science. *Science.* 2011 Dec 1;334(6060):1226–7.
22. Van Essen DC, Dierker DL. Surface-based and probabilistic atlases of primate cerebral cortex. *Neuron.* 2007 Oct 25;56(2):209–25.
23. Talairach J, Tournoux P. Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging. Thieme Medical Publishers, Inc; 1988.
24. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage.* Elsevier; 2002;15(1):273–89.
25. Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM. 3D statistical neuroanatomical models from 305 MRI volumes. *IEEE;* 1993;:1813–7.
26. Nichols BN, Mejino JL Jr, Detwiler L, Nilsen TT, Martone ME, Turner JA, et al. Neuroanatomical domain of the foundational model of anatomy ontology. *J Biomedical Semantics.* 2014;5:1.
27. Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform.* 2003 Dec;36(6):478–500.
28. Turner JA, Mejino JLV, Brinkley JF, Detwiler LT, Lee HJ, Martone ME, et al. Application of neuroanatomical ontologies for neuroimaging data annotation. *Front Neuroinform.* 2010;4.
29. Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubézy M, Eriksson H, et al. The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies.* Elsevier; 2003;58(1):89–123.
30. McGuinness DL, Van Harmelen F. OWL web ontology language overview. *W3C recommendation.* 2004;10(10):2004.
31. Golbreich C, Zhang S, Bodenreider O. The foundational model of anatomy in OWL: Experience and perspectives. *Web Semant.* 2006;4(3):181–95.

32. Noy NF, Rubin DL. Translating the Foundational Model of Anatomy into OWL. *Web Semant.* 2008;6(2):133–6.
33. Brinkley JF, Detwiler LT, Structural Informatics Group. A query integrator and manager for the query web. *J Biomed Inform.* 2012 Oct;45(5):975–91.
34. Marcus D, Olsen T, Ramaratnam M, Buckner R. The extensible neuroimaging archive toolkit. *Neuroinformatics.* 2007;5(1):11–33.
35. Scott A, Courtney W, Wood D, la Garza De R, Lane S, King M, et al. COINS: An Innovative Informatics and Neuroimaging Tool Suite Built for Large Heterogeneous Datasets. *Front Neuroinform.* 2011;5:33–3.
36. Book GA, Anderson BM, Stevens MC, Glahn DC, Assaf M, Pearlson GD. Neuroinformatics Database (NiDB) - A Modular, Portable Database for the Storage, Analysis, and Sharing of Neuroimaging Data. *Neuroinformatics.* 2013;11(4):495–505.
37. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009 Apr;42(2):377–81.
38. Galdzicki M, Clancy KP, Oberortner E, Pocock M, Quinn JY, Rodriguez CA, et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nature Biotechnology.* 2014 Jun;32(6):545–50.
39. Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, et al. Automatically parcellating the human cerebral cortex. *Cereb Cortex.* 2004 Jan;14(1):11–22.
40. Gronenschild EHB, Gronenschild EHB, Habets P, Habets P, Jacobs HIL, Jacobs HIL, et al. The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE.* 2012;7(6):e38234.
41. Tustison NJ, Johnson HJ, Rohlfing T, Klein A, Ghosh SS, Ibanez L, et al. Instrumentation bias in the use and evaluation of scientific software: recommendations for reproducible practices in the computational sciences. *Frontiers in Neuroscience.* 2013;7:162.
42. Shepherd GM, Mirsky JS, Healy MD, Singer MS, Skoufos E, Hines MS, et al. The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends in Neurosciences.* 1998 Nov;21(11):460–8.
43. Bernstein MA, King KF, Zhou XJ. *Handbook of MRI Pulse Sequences.* Elsevier; 2004. 1 p.
44. Ogawa S, Lee TM, Kay AR, Tank DW. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci USA.* 1990 Dec;87(24):9868–72.
45. Le Bihan D, Mangin J-F, Poupon C, Clark CA, Pappata S, Molko N, et al. Diffusion tensor imaging: concepts and applications. *Journal of magnetic resonance imaging: JMRI.* 2001 Apr;13(4):534–46.
46. Toga AW, Thompson PM. The role of image registration in brain mapping. *Image and Vision Computing.* 2001 Jan 1;19(1-2):3–24.

47. Wakana S, Jiang H, Nagee-Poetscher LM, van Zijl PCM, Mori S. Fiber Tract-based Atlas of Human White Matter Anatomy. *Radiology*. 2004 Jan 1;230(1):77-87.
48. Oishi K, Zilles K, Amunts K, Faria A, Jiang H, Li X, et al. Human brain white matter atlas: identification and assignment of common anatomical structures in superficial white matter. *NeuroImage*. Elsevier; 2008;43(3):447-57.
49. Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, et al. Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*. 2000 Jul;10(3):120-31.
50. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*. 2006 Jul 1;31(3):968-80.
51. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*. Wiley Online Library; 1994;2(4):189-210.
52. Ashburner J, Friston KJ. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*. 1999;7(4):254-66.
53. Bug WJ, Ascoli GA, Grethe JS, Gupta A, Fennema-Notestine C, Laird AR, et al. The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. *Neuroinformatics*. 2008 Sep;6(3):175-94.
54. Imam FT, Larson SD, Bandrowski A, Grethe JS, Gupta A, Martone ME. Development and use of Ontologies Inside the Neuroscience Information Framework: A Practical Approach. *Front Genet*. 2012;3:111.
55. Bowden DM, Martin RF. NeuroNames Brain Hierarchy. *NeuroImage*. 1995 Mar;2(1):63-83.
56. Bowden DM, Dubach MF. NeuroNames 2002. *Neuroinformatics*. 2003;1(1):43-59.
57. Bowden DM, Song E, Kosheleva J, Dubach MF. NeuroNames: an ontology for the BrainInfo portal to neuroscience on the web. *Neuroinformatics*. 2012 Jan;10(1):97-114.
58. Brinkley J, Rosse C. Imaging and the Human Brain Project: a review. *Methods Inf Med*. 2002;41(4):245-60.
59. Sherif T, Rioux P, Rousseau M-E, Kassis N, Beck N, Adalat R, et al. CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front Neuroinform*. 2014;8:54.
60. Van Horn JD, Toga AW. Is it time to re-prioritize neuroimaging databases and digital repositories? *NeuroImage*. 2009 Oct 1;47(4):1720-34.
61. Dinov I, Lozev K, Petrosyan P, Liu Z, Eggert P. Neuroimaging Study Designs, Computational Analyses and Data Provenance Using the LONI Pipeline. *PLoS ONE*. 2010.
62. Shattuck DW, Leahy RM. BrainSuite: an automated cortical surface identification tool. *Medical image analysis*. Elsevier; 2002;6(2):129-42.
63. Rohlfing T, Poline J-B. Why shared data should not be acknowledged on the author

- byline. *NeuroImage*. 2012 Feb 15;59(4):4189–95.
64. Lein E, Hawrylycz M, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*. Nature Publishing Group; 2006 Dec 6;445(7124):168–76.
 65. Hawrylycz MJ, Hawrylycz MJ, Lein ES, Lein ES, Guillozet-Bongaarts AL, Guillozet-Bongaarts AL, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature*. 2012 Sep 19;489(7416):391–9.
 66. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, et al. The Human Connectome Project: a data acquisition perspective. *NeuroImage*. 2012 Oct 1;62(4):2222–31.
 67. Marcus DS, Harwell J, Olsen T, Hodge M, Glasser MF, Prior F, et al. Informatics and data mining tools and strategies for the human connectome project. *Front Neuroinform*. 2011;5:4.
 68. Hamilton DJ, Shepherd GM, Martone ME, Ascoli GA. An ontological approach to describing neurons and their relationships. *Front Neuroinform*. 2012;6:15.
 69. Larson SD, Martone ME. NeuroLex.org: an online framework for neuroscience knowledge. *Front Neuroinform*. 2013;7:18.
 70. Osumi-Sutherland D, Reeve S, Mungall CJ, Neuhaus F, Ruttenberg A, Jefferis GSXE, et al. A strategy for building neuroanatomy ontologies. *Bioinformatics*. 2012 May 1;28(9):1262–9.
 71. Bohland JW, Bokil H, Allen CB, Mitra PP. The brain atlas concordance problem: quantitative comparison of anatomical parcellations. *PLoS ONE*. 2009;4(9):e7200.
 72. Hawrylycz M, Baldock RA, Burger A, Hashikawa T, Johnson GA, Martone M, et al. Digital atlas and standardization in the mouse brain. *PLoS Comput Biol*. 2011;7(2):e1001065.
 73. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. *Journal of the American Medical Informatics Association*. 2012 Nov;19(6):931–8.
 74. Friedman CP. A “fundamental theorem” of biomedical informatics. *J Am Med Inform Assoc*. 2009 Mar;16(2):169–70.
 75. Anderson NR, Ash JS, Tarczy-Hornoch P. A qualitative study of the implementation of a bioinformatics tool in a biological research laboratory. *International Journal of Medical Informatics*. 2007 Nov;76(11-12):821–8.
 76. Franklin JD, Guidry A, Brinkley JF. A partnership approach for Electronic Data Capture in small-scale clinical trials. *J Biomed Inform*. 2011 Dec;44 Suppl 1:S103–8.
 77. Oldfield RC. The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*. 1971 Mar;9(1):97–113.
 78. Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform*. 2001

- Aug;34(4):285-98.
79. Louie B, Mork P, Martin-Sanchez F, Halevy A, Tarczy-Hornoch P. Data integration and genomic medicine. *J Biomed Inform.* 2007 Feb;40(1):5-16.
 80. Detwiler LT, Suciu D, Franklin JD, Moore EB, Poliakov AV, Lee ES, et al. Distributed XQuery-Based Integration and Visualization of Multimodality Brain Mapping Data. *Front Neuroinform.* 2009;3:2.
 81. Ashish N, Ambite JL, Muslea M, Turner JA. Neuroscience Data Integration through Mediation: An (F)BIRN Case Study. *Front Neuroinform.* 2010;4:118.
 82. Biswal BB, Mennes M, Zuo X-N, Gohel S, Kelly C, Smith SM, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences.* 2010 Mar 9;107(10):4734-9.
 83. Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform.* 2011;5:13.
 84. Bellec P, Lavoie-Courchesne S, Dickinson P, Lerch J, Zijdenbos A, Evans AC. The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Front Neuroinform.* 2012;6.
 85. Dinov I, Lozev K, Petrosyan P, Liu Z, Eggert P, Pierce J, et al. Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PLoS ONE.* 2010;5(9).
 86. Ozyurt IB, Keator DB, Wei D, Fennema-Notestine C, Pease KR, Bockholt J, et al. Federated web-accessible clinical data management within an extensible neuroimaging database. *Neuroinformatics.* 2010 Dec;8(4):231-49.
 87. Gadde S, Aucoin N, Grethe JS, Keator DB, Marcus DS, Pieper S, et al. XCEDE: an extensible schema for biomedical data. *Neuroinformatics.* 2012 Jan;10(1):19-32.
 88. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific American* [Internet]. 2001 May 1;284(May):1-4. Available from: <http://www.nature.com/scientificamerican/journal/v284/n5/pdf/scientificamerican0501-34.pdf>
 89. Prud'hommeaux E, Seaborne A, editors. SPARQL Query Language for RDF [Internet]. SPARQL Query Language for RDF. 2008 [cited 2013 Dec 13]. Available from: <http://www.w3.org/TR/rdf-sparql-query/>
 90. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *International Journal on Semantic Web and Information Systems.* 2009;5(3):1-22.
 91. Nugent AC, Luckenbaugh DA, Wood SE, Bogers W, Zarate CA, Drevets WC. Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation. *Human Brain Mapping.* Wiley Online Library; 2013;34(9):2313-29.
 92. Moreau L, Missier P. PROV-DM: The PROV Data Model [Internet]. 2012 Jul. Available from: <http://www.w3.org/TR/prov-dm/>

93. Nichols BN, Dougherty RF, Detwiler LT, Schaefer G, Frank RJ, Brinkley JF, et al. NIQuery: Neuroimaging Informatics Query Framework for Data Sharing Discovery and Analysis. Munich; 2012. p. 68. Available from: <http://www.neuroinformatics2012.org/abstracts/niquery-neuroimaging-informatics-query-framework-for-data-sharing-discovery-and-analysis>
94. Friedman CP. Where's the science in medical informatics? Journal of the American Medical Informatics Association. American Medical Informatics Association; 1995;2(1):65.
95. Martin RC. Agile Software Development: Principles, Patterns, and Practices. Prentice Hall PTR; 2003.
96. Wilson G, Aruliah DA, Brown CT, Hong NPC, Davis M, Guy RT, et al. Best Practices for Scientific Computing. PLoS Biol. Public Library of Science; 2014 Jan 7;12(1):e1001745.
97. Kane DW, Hohman MM, Cerami EG, McCormick MW, Kuhlman KF, Byrd JA. Agile methods in biomedical software development: a multi-site experience report. BMC bioinformatics. 2006;7:273.
98. Shortliffe EH. The science of biomedical computing. Inform Health Soc Care. Informa UK Ltd UK; 1984 Jan;9(3-4):185-93.
99. Clancey WJ, Shortliffe EH. Readings in medical artificial intelligence: the first decade. Addison-Wesley Longman Publishing Co., Inc; 1984.
100. Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron. 2002 Jan 31;33(3):341-55.
101. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods. 2011 Aug;8(8):665-70.
102. Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, et al. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front Neuroinform. 2011;5:17.
103. De Schutter E. Data publishing and scientific journals: the future of the scientific paper in a world of shared data. Neuroinformatics. 2010 Oct;8(3):151-3.
104. Gorgolewski KJ, Margulies DS, Milham MP. Making data sharing count: a publication-based solution. Frontiers in Neuroscience. 2013;7:9.
105. Sansone S-A, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, et al. Toward interoperable bioscience data : Nature Genetics : Nature Publishing Group. Nature genetics. 2012 Feb;44(2):121-6.
106. Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013;7:12.
107. Rosse C, Mejino JL Jr. The foundational model of anatomy ontology. Springer; 2008;;59-117.
108. Pianykh O. Digital Imaging and Communications in Medicine (DICOM): A practical

- introduction and survival guide. books.google.com. 2008.
109. Helmer KG, Ghosh SS, Nichols BN, Keator DB, Nichols TE, Turner JA. Connecting brain imaging terms to established lexicons: a precursor for data sharing and querying. Munich; 2012. p. 58.
 110. Moreau L, Groth P. Provenance: An Introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers; 2013 Sep 15;3(4):1-129.
 111. Ghosh SS, Nichols BN, Gadde S, Steffener J, Keator DB. XCEDE-DM: a neuroimaging extension to the W3C provenance data model. Munich; 2012. p. 57.
 112. Channin DS, Mongkolwat P, Kleper V, Sepukar K, Rubin DL. The caBIG annotation and image Markup project. Journal of Digital Imaging. 2010 Apr;23(2):217-25.
 113. Nichols BN, Mejino JLV, Brinkley JF. The Foundational Model of Neuroanatomy Ontology: Ontology Framework to Support Neuroanatomical Data Integration. Buffalo; 2011. p. 444.
 114. Nichols BN, Haselgrove C, Poline JB, Ghosh SS. Neuroimaging data access and query through a common application programming interface. Munich; 2012. p. 246.
 115. Das S, Zijdenbos AP, Harlap J, Vins D, Evans AC. LORIS: a web-based data management system for multi-center studies. Front Neuroinform. 2011;5:37.
 116. Fielding RT, Taylor RN. Principled design of the modern Web architecture. 2000. pp. 407-16.
 117. Alexander K, Hausenblas M. Describing linked datasets-on the design and usage of void, the 'vocabulary of interlinked datasets. In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference. Citeseer; 2009.
 118. Dodds L, Davis I. Linked data patterns. <http://patterns.dataincubator.org/book/>; 2011.
 119. Carroll JJ, Bizer C, Hayes P, Stickler P. Named graphs, provenance and trust. New York, New York, USA: ACM Press; 2005. p. 613.
 120. Nichols BN, Stoner R, Keator DB, Turner JA, Helmer KG, Grabowski TJ, et al. There's an app for that: a semantic data provenance framework for reproducible brain imaging. Seattle; 2013. p. 45.
 121. Bechhofer S, Ainsworth J, Bhagat J, Buchan I, Couch P, Cruickshank D, et al. Why Linked Data is Not Enough for Scientists. 2010. pp. 300-7.
 122. Clark T, Ciccarese PN, Goble CA. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. Journal of Biomedical Semantics [Internet]. 2014;5(1):28. Available from: <http://www.jbiomedsem.com/content/5/1/28/abstract>
 123. Dumontier M, Baker CJ, Baran J, Callahan A, Chepelev LL, Cruz-Toledo J, et al. The Semantic Science Integrated Ontology (SIO) for biomedical research and knowledge discovery. J Biomedical Semantics. 2014;5:14.

124. Wilkinson MD, Vandervalk BP, McCarthy EL. The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation. *J Biomedical Semantics*. 2011;2:8.
125. Seneviratne OW. Augmenting the web with accountability. *ACM Request Permissions*; 2012.
126. Berners-Lee T, O'Hara K. The read-write Linked Data Web. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2013 Feb 18;371(1987):20120513.