

Neuroimaging Data Provenance Using the LONI Pipeline Workflow Environment

Allan J. MacKenzie-Graham¹, Arash Payan¹, Ivo Dinov¹,
John D. Van Horn¹, and Arthur W. Toga^{1*}

¹Laboratory of Neuro Imaging (LONI), Department of Neurology,
University of California Los Angeles School of Medicine, 635 Charles E.
Young Drive South, Suite 225, Los Angeles, CA 90095-7334
{amg, apayan, ivo.dinov, jvanhorn, toga}@loni.ucla.edu

* Corresponding Author

Abstract. Provenance, the description of the history of a set of data, has become important in the neurosciences with the proliferation of research consortia-related neuroimaging efforts. Knowledge about the origin, preprocessing, analysis and post hoc processing of neuroimaging volumes is essential for establishing data and results quality, the reproducibility of findings, and their scientific interpretation. Neuroimaging provenance also includes the specifics of the software routines, algorithmic parameters, and operating system settings that were employed in the analysis protocol. The LONI Pipeline (<http://pipeline.loni.ucla.edu>) is a Java-based workflow environment for the construction and execution of data processing streams. We have developed a provenance framework for describing the current and retrospective data state integrated with the LONI Pipeline workflow environment. Collection of provenance information under this framework alleviates much of the burden of documentation from the user while still providing a rich description of an image's characteristics, as well as the description of the programs that interacted with that data. This combination of ease of use and highly descriptive meta-data will greatly facilitate the collection of provenance information from brain imaging workflows, encourage subsequent data and meta-data sharing, enhance peer-reviewed publication, and support multi-center collaboration.

Keywords: Provenance; Workflow; Neuroimaging; Grid

1. Introduction

One of the fundamental challenges in neuroimaging, and in fact all biological sciences, involves devising ways to manage the enormous amounts of data currently being gathered. This challenge is compounded not only by the proliferation of collaborative efforts and the necessity of sharing data across multiple sites, but also of making that data openly available and useful to the scientific community at large. The scientific community has recognized the need for solutions that facilitate the process of tool and data exchange and numerous efforts are underway to achieve this goal [1]. Yet to be truly meaningful, the data obtained and the analytic tools employed must be adequately described and documented. The meta-data detailing the origin and subsequent processing of biological images is referred to as "provenance" [2].

Recently, leading computer scientists have recognized the unique issues associated with neuroimaging datasets that often exceed several tens of gigabytes in a full set of raw data. Simon Miles, Luc Moreau, Mike Wilde, Ian

Foster, and others proposed a “provenance challenge” to determine the state of available provenance systems [3]. The challenge consisted of collecting provenance information from a simple neuroimaging workflow [4] and documenting each system’s ability to respond to a set of predefined queries. Some of these existing provenance systems have previously been proposed as mechanisms for capturing provenance in neuroimaging, though they have not been widely adopted [5]. The main difficulty appears to be the need of a system to capture provenance information accurately, completely, but with minimal user intervention. Minimizing an individual’s burden for providing the details on provenance, as well as facilitating a comprehensive data and process tracking system, will dramatically improve compliance, thereby freeing the user to focus on performing neuroimaging research rather than exhaustively documenting provenance.

In the biological sciences, a description of how data was obtained is crucial for assessing its quality and usefulness, as well as enabling analysis in an appropriate context. It is therefore imperative that the provenance of biological images be easily captured and readily accessible. In multiple sclerosis research, for example, increasingly complex analysis workflows are being developed to extract information from large cross-sectional or longitudinal studies [6]. This is also true of Alzheimer’s disease [7-9], autism [10], depression [11], schizophrenia [12], and even studies of normal populations [13]. The implementation of the increasingly complex processing workflows associated with these investigations requires the institution of quality-control practices to ensure the precision, reproducibility, and reusability of the results. In effect, provenance.

In a broad sense, provenance can be divided into two subtypes: *data provenance* and *processing provenance*. Data provenance is the metadata that describes the subject being imaged, how an image of that subject was collected, who acquired the image, what instrument was used, what settings or parameters were used, and how the sample was prepared. However, most scientific image data is not obtained directly from such measurements, but rather derived from other data by the application of computational processes. Processing provenance is the metadata that defines what processing an image has undergone; for example, how the image was skull-stripped, what form of image inhomogeneity correction was employed, how the volume was spatially aligned to a standard atlas space, etc. Even data that is presented as “raw” often has been subjected to reconstruction software or converted from the scanner’s native image format (k-space) to a more commonly used and easily shared file format [14]. A complete data provenance model would capture all this information, making the history of a complete set of data transparent, thus enabling seamless sharing across the neuroimaging community.

Certain neuroimaging data provenance may be gathered at the site where the data is collected, in the headers of image files or in databases that record image acquisition [15, 16]. A highly abbreviated form of this kind of provenance is often reported in method descriptions or even in the image files themselves [17]. However, this data is seldom propagated along with the images themselves, since it is commonly removed or ignored in the course of file conversion – not being critical for further data processing.

Processing provenance can be obtained concerning any resource in the data processing system and may include multiple levels of detail. Two major models for collecting processing provenance have been described: a *process-oriented* model [18] and a *data-oriented* model [2]. The process-oriented model collects lineage information from the deriving processes and provenance is inferred from

that processing and through an inspection of the input and output data. This mechanism is well suited for situations where individual data products are tracked within comprehensive frameworks and where the deriving processes can easily be reapplied to the original data to reproduce the data product. In the data-oriented model, lineage information is explicitly gathered about the set of data. This method may be better suited for situations where data sharing occurs across heterogeneous environments and intermediate data products may not be available for reproduction. This would be the case, for example, when neuroimaging data sets are shared between two or more collaborating laboratories.

The analysis of raw data in neuroimaging has become a computationally rich process with many intricate steps run on increasingly larger datasets [6]. Many commonly available software packages exist that provide either complete analyses or specific steps in neuroimaging data analysis. These packages often have diverse input and output requirements, utilize different file formats, run under particular computer environments, and may have limited abilities for certain types of data. The combination of these packages to achieve more sensitive and accurate results has become a widespread strategy in brain mapping studies, though requires much work to ensure valid interoperability between programs.

Simplicity of use cannot be overstated when developing software tools for the scientific community at large. Many outstanding software tools are not adopted due to difficult learning curves or because their use places too great a burden on the end user. One of the main requirements of a successful provenance system would be the simplicity and unobtrusiveness.

The LONI Pipeline was developed to facilitate ease of workflow construction, validation, and execution [19] freeing the user to focus on image analysis. In this article we describe a simple yet comprehensive provenance system that has been incorporated into the LONI Pipeline Processing Environment, placing little or no burden on the end user for documentation of processing provenance.

2. The LONI Pipeline Workflow Environment

The LONI Pipeline (<http://pipeline.loni.ucla.edu>) is a simple, efficient, and distributed computing environment, enabling software inclusion from different laboratories in different environments. It provides a visual programming interface for the design, execution, and dissemination of neuroimaging analyses. Individual executables are represented as “modules” that can be included, deleted, and substituted for other modules within a user-friendly graphical user interface. Connections between the modules that establish an analysis methodology are represented as “workflows”. The environment handles bookkeeping, controls the details of the computation, and information transfer between modules and within the workflow. It permits files, intermediate results, and other information to be accurately passed between individually connected modules. The DRMAA API (www.drmaa.net), backed by the Sun Grid Engine (<http://gridengine.sunsource.net>), acts as an interface to grid environments. Modules and workflows can be saved to disk at any stage of development and recalled at a later time for modification, use, or distribution.

2.1 Goals of the LONI Pipeline environment

The overarching goals of the LONI Pipeline are to:

2.1.1 *Graphical User Interface:* Create a robust environment for scientific software tool interoperability, Grid integration and low-cost interactive user interface. For maximum portability, scalability and efficiency, this environment is built in Java and utilizes XML for storing and communication of meta-data, and descriptors for tools and services.

2.1.2 *New Tool Discovery:* Enable expert researchers to quickly design, test and validate novel experimental designs and to rapidly examine new data analysis protocols. This is achieved via dynamic, responsive and extensible graphical user interface.

2.1.3 *Compatibility:* Provide the necessary means for integration of LONI Pipeline XML workflow descriptions with other established graphical environments for scientific Grid computing. This functionality facilitates the translation of existent analysis paradigms from other environments to the LONI Pipeline and vice-versa.

The LONI Pipeline differs from many similar workflow environments. For instance, the LONI Pipeline does not require the use of an application programming interface (API) – it considers all resources as well-described external applications that may be invoked with standard remote execution protocols. The LONI Pipeline XML description protocol allows any command-line driven process, web-service or data-server to be encapsulated into the environment by reference. There is no need to reprogram, revise or recompile external resources to make them usable within the LONI Pipeline. This is a deliberate design we have imposed to reduce the integration/utilization costs of including new resources within the LONI Pipeline environment. This approach provides the benefit of quick and easy management of large and disparately located resources and data. In addition, this choice significantly minimizes the hardware requirements for user-client machine (e.g., memory, storage, CPU). Finally, while the LONI Pipeline is primarily used in the context of neuroimaging, we wish to stress the important point that the Pipeline is agnostic to any particular scientific domain and can be used to manage workflows under any other scientific domain.

3. LONI Pipeline Provenance Architecture

To begin the discussion of how the LONI Pipeline manages provenance, we have defined some terms in order to prevent any ambiguity when software is discussed. To facilitate the current discussion we define the following terms; a binary is a pre-compiled program that is ready to run under a given operating system, a script is a simple program written in a utility language that is interpreted at runtime, and an executable is either a binary or script.

3.1 *Data Provenance:* As mentioned above, an important aspect of provenance is the description of the subject. Subject provenance includes birth and death dates (for post-mortem studies), in addition to the age of the subject at the time of the data collection (or death). Sex and species are captured, further qualified by strain and genetic manipulation in the case of non-human subjects. Treatments, such as disease induction in experimental models, drug treatment, and combinations of treatments can be documented in the schema. Subject name has explicitly been excluded in order to protect patient privacy (<http://www.hhs.gov/ocr>), SubjectID standing in as a unique identifier for a given subject. These elements are extensible, allowing for multiple treatments

or clinical evaluations. Subject provenance has been described in a simple, yet flexible format in order to make it easily accessible to the community with a minimum of work to adapt it for specialized use.

The description of how a set of data was acquired is of critical importance for data provenance. Different information is required from the user based on the kind of data acquired. For example, when collecting acquisition provenance about an MRI image, information about the acquisition type (2D vs. 3D), weighting (proton density, T1, T2, etc.), pulse sequence, flip angle, echo time (TE), repetition time (TR), inversion time (TI), matrix dimensions, step sizes, magnet field strength, coil used, equipment manufacturer and model are explicitly captured in the XSD. These elements are far from exhaustive, but are easily expanded and/or extended to accommodate other imaging modalities from diffusion tensor imaging (DTI) to positron emission tomography (PET).

An XML schema document (XSD) describing the neuroimaging data provenance presented here is available to the public for use and discussion (<http://provenance.loni.ucla.edu>).

3.2 Processing Provenance: In our model, binary provenance describes how a piece of software was compiled. It comprises two parts, a description of the environment and a description of the binary itself. The environment description includes the operating system, environment variables, compiler used, and libraries installed. The binary description includes configuration flags and/or modifications made to configuration files or *makefiles*. Our goal is to provide the user with the ability to reproduce the binary exactly (Table 1).

A fundamental difference between executables is the hardware platform on which they were compiled. Differences in floating-point implementation across different architectures can have a profound impact on outcome of a calculation and have been widely publicized in the popular media [20]. The LONI Pipeline executable provenance description captures not only architecture, but also the specific processor and the flags that are enabled on it.

Capturing important details about the operating system is complicated, particularly for Linux and open-source Unix distributions, since each distribution contains many individually updated components. Essential information must be captured such as the operating system name, version, distribution, kernel name, and kernel version. For example, an application running on Ubuntu Dapper Drake (<http://www.ubuntu.com>) must have the following operating system metadata: Linux, 6.06, Ubuntu Desktop, #1 PREEMPT, 2.6.15-27-386; whereas an application built on the current Mac OS X Leopard platform must have the following operating system metadata: Mac OS, 10.5.2, n/a, Darwin, 9.2.0.

The compiler used and libraries linked during compilation are a crucial aspect of the environment. In addition to compiler name and version, a list of which updates have been applied is also captured. This section of the provenance metadata also records which flags were used when the compiler was invoked, architecture and optimization flags being of special interest. Libraries used for compilation are described similarly to the binary itself and are recursive.

Binaries also can be configured prior to compilation. Some packages are distributed in a format for use with the GNU build system or Autotools [21]. Modification of the configure script or the *makefile* can yield substantially different results after compilation. The LONI Pipeline executable provenance description captures flags to the configure script, modifications to configure scripts and makefiles.

Executable provenance need only be collected once, when a binary is

compiled or when a script is written. This data is then included in the LONI Pipeline module description of the executable and thus is propagated with both the module and any workflows created with those modules.

Processing provenance describes the actual invocation of an individual executable or the invocation of an executable in the context of a series of steps or workflow. Recording the command-line that was used to invoke it captures arguments to the executable. The processing environment is described similarly to the environment for compilation, but also includes environmental variables that may modify the behavior of the executable.

Often image processing is complex and non-linear and cannot be represented in a simple script or directed acyclic graph. Data may converge along several lines of processing only to diverge again after a common step. These complex workflows are difficult to document, either for publication or later re-use. Capturing the provenance for these workflows is equally complex, not only requiring the execution order of the individual steps, but how those steps are related to one another, especially in the case of multiple lines of data being processed simultaneously. In order to address this issue we have used the LONI Pipeline Processing Environment (<http://pipeline.loni.ucla.edu>) [19] to capture not only executable provenance and description, but also the relationships between the executables.

Using the LONI Pipeline as an example of workflow software, we have designed the provenance framework to take advantage of context information that can only be kept while using workflow software. Specifically, the use of conditionals between executables, and loops can all be represented in a higher workflow language and associated with a series of executable events in the provenance. More generally, we want to be able to track how data is derived with sufficient precision that one can create or recreate it from this knowledge.

Continuing discussion and development of the LONI Pipeline Provenance Architecture can be found at <http://provenance.loni.ucla.edu>.

- Workflow provenance
 - Pipeline workflow
 - Executable provenance
 - Environment
 - Options
 - Input files
 - Output files
 - Binary provenance
 - Binary configuration
 - Configuration options
 - System configuration
 - Architecture
 - Operating system
 - Compiler
 - Libraries
 - Script provenance
 - Shell
 - Script
 - Binary provenance

Table 1. Workflow Provenance. Outline describing the major elements of workflow provenance contained in a LONI Pipeline workflow file.

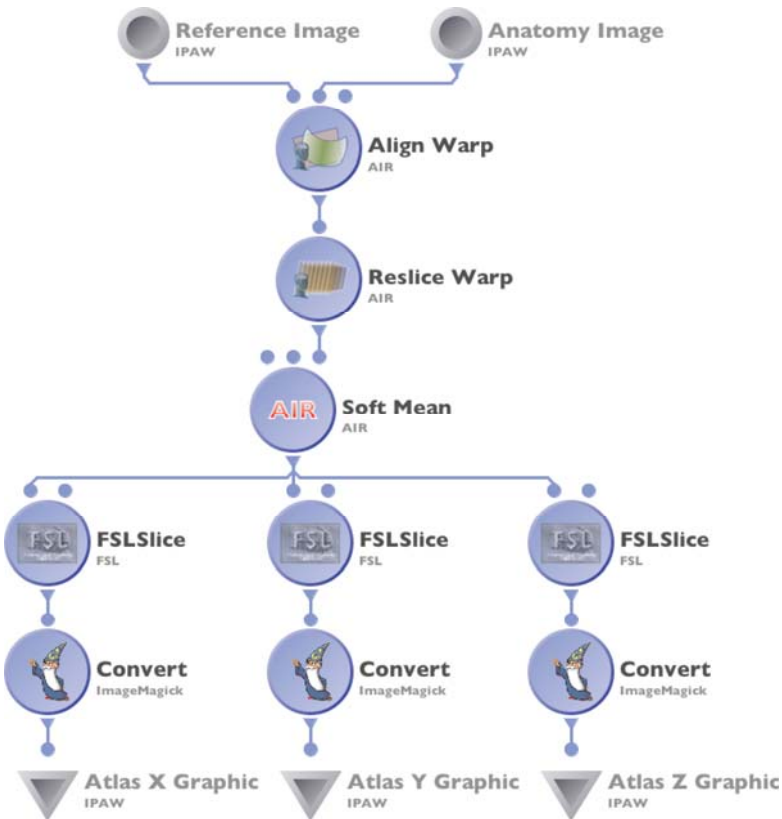


Figure 1. Simple Neuroimaging Workflow. A simple neuroimaging workflow derived from “The First Provenance Challenge” [3] in the LONI Pipeline Processing Environment.

4. Provenance Validation

In order to document the utility of this model of provenance documentation, we performed a test to demonstrate the capacity to independently recreate a workflow and its output data using only the provenance documentation.

A workflow was created in the LONI Pipeline from binaries compiled on the LONI 306-node dual processor Opteron Sun V20z grid by the LONI system administrator and modules in the pipeline library constructed by the LONI Pipeline developers. Data and processing provenance was captured and recorded in a provenance file using only the mechanism described above.

A second workflow was constructed from scratch with modules defined by the authors for a second set of executables compiled for the LONI grid, also compiled by the authors. These executables were compiled using only the provenance information captured by the mechanism described above as a guide for compilation.

The workflows used for the test were simple neuroimaging workflows (Figure 1) with both sets of executables compiled using the same options. The workflows were then run on the LONI grid using the same input data. We compared the aligned images resulting from each workflow by subtracting them from one another and verified that the difference was 0 at every voxel (data not shown).

Multiple packages can be combined and provenance information will be propagated with those workflows. Combining package elements allows the user the greatest flexibility for their analyses. For example, a workflow could correct motion artifact using tools from Freesurfer [22], perform skull stripping using the BSE [23], calculate and apply the N3 field inhomogeneity correction [24], and then align a magnetic resonance image to a standard atlas using FMRIB's Linear Image Registration Tool (FLIRT) [25, 26] from the FMRIB Software Library (FSL) [27] (Figure 2).

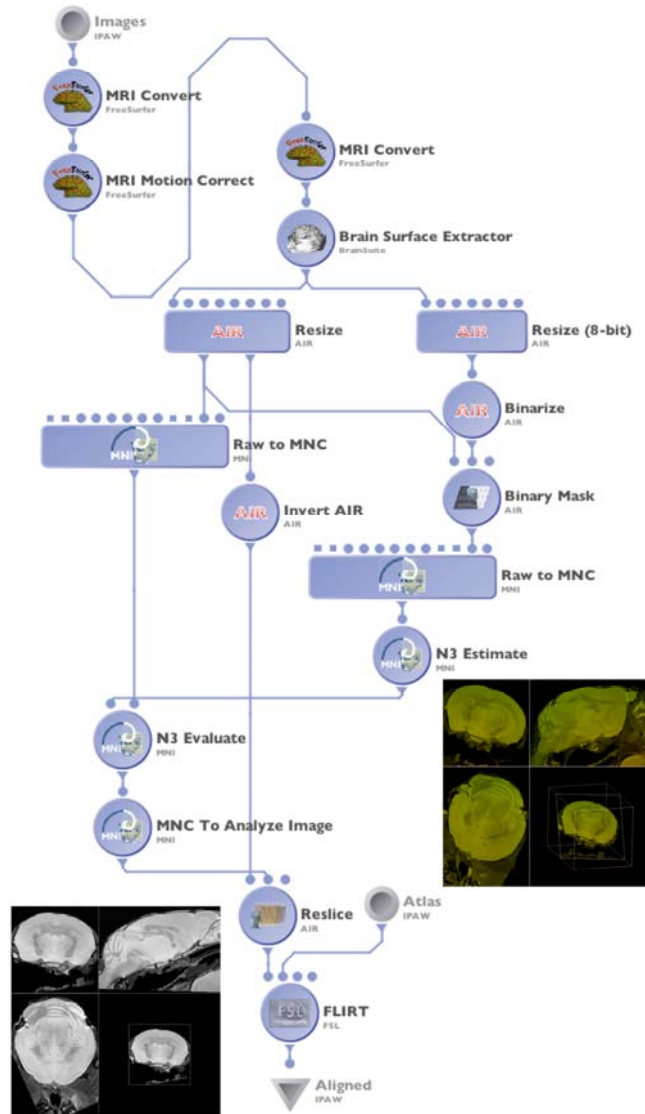


Figure 2. Complex Neuroimaging Workflow. A complex neuroimaging workflow combining multiple analysis and processing packages. Right inset: An overlay of two magnetic resonance microscopy (MRM) images, one inhomogeneity corrected (green) and one not (red). Greenish and orange areas represent field inhomogeneities. Left inset: The same MRM image, inhomogeneity corrected and aligned to an atlas.

5. Discussion

Recent interest has arisen in the field of neuroscience, and particularly in neuroimaging, in identifying or creating standards to facilitate software tool interoperability. The NIMH Neuroimaging Informatics Technology Initiative (NIFTI; <http://nifti.nimh.nih.gov>) was formed to aid in the development and enhancement of informatics tools for neuroimaging. Though best known for the Data Format Working Group (DFWG) that has defined the NIFTI image file format standard, this effort has recently turned its attentions to how provenance metadata might be standardized. The Biomedical Informatics Research Network (BIRN; <http://nbirn.net>) is another high profile effort working to develop standards among its consortia membership, including the development of study data provenance.

Descriptions of data provenance have been used successfully in other fields of endeavor. For example, the Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems (<http://dublincore.org>). This also includes meta-data related to workflow provenance.

The Collaboratory for Multi-scale Chemical Science (CMCS) project is an informatics toolkit for collaboration and data management for multi-scale chemistry [28]. CMCS collects pedigree information about individual data objects by defining input and output data and capturing “pedigree chains” describing the processing that the data has undergone (<http://cmcs.org>). The provenance data is explicitly defined in associations, placing the burden of documentation upon the user.

The Virtual Data System (VDS; formerly known as Chimera and incorporating Pegasus) [5] provides middleware for the GriPhyN project (<http://www.griphyn.org>), expressing, executing, and tracking the results of workflows. Provenance is used for the regeneration, comparison, and auditing of data derivations. Users construct workflows using a standard virtual data language (VDL) describing “transformations” (executable programs) that are executed by a VDL interpreter producing a “derivation” (the execution of a transformation). “Data objects” are entities that are consumed or produced by a derivation. In the VDS model, provenance is inferred from the processing by inverting the processing to associate the output data with the input data. This approach places very little burden on the user to document data provenance.

The myGrid project [29] provides middleware in support of computational experiments in the biological sciences, modeled as workflows in a grid environment. Users construct workflows written in XScufl language using the Taverna engine. The LogBook is a plug-in for Taverna engine that allows users to log their experiments in a MySQL database and browse, rerun, and maintain previously run workflows (<http://www.mygrid.org.uk/wiki/Mygrid/LogBook>). This provenance log contains the executables invoked, the parameters used, data used and derived, and is automatically produced when the workflow executes. This process-oriented provenance log is also inverted to infer the provenance for the intermediate and final set of data.

Within the neuroimaging community, the XCEDE (XML-based Clinical Experiment Data Exchange) schema [30] also provides for the storage of data provenance information. Provenance information manually captured includes hardware, compilation and libraries linked, operating system and software versions, and parameters used to generate and document results. XCEDE is a

data-oriented system where the provenance metadata is associated with the actual data files.

The VisTrails scientific workflow management system [31] is an excellent example of an integrated workflow and provenance collection mechanism. The VisTrails workflow system focuses on capturing provenance in exploratory workflows and saving changes that occur over time. Processes and workflows are represented as python objects prior to execution and details of the execution are stored in a relational database automatically.

Efforts such as these examples have sought to capture data and workflow information sufficient to reproduce reported study findings and that enable cross-study comparison. Specific workflow description frameworks also exist in other fields that help to sequence data processing steps and that can be used to populate provenance descriptions. These frameworks are highly sophisticated tools that require substantial investment to learn and deploy. They do not provide a simple mechanism for the capture of provenance metadata from multiple packages, the capacity to represent complex, non-sequential analyses, nor at a sufficient level of detail to allow the reproduction of a derived set of data on a new platform. Hence the need for the development of a provenance framework that can easily be applied to complex neuroimaging analyses.

Future directions include the enrollment of LONI Pipeline workflows in a database, creating a processing and provenance database. Having a readily searchable database of commonly used (and rarely used) workflows would greatly aid investigators in recreating the conditions of a particular analysis, reproducing previous results and rerunning analyses with small modifications.

The concept of provenance can extend to knowledge of the behavior of executables, such as describing their function. The Brain Surface Extractor (BSE) [23], the Brain Extraction Tool (BET) [32], and MRI Watershed [22] are all brain extraction algorithms, however, their internal functions may not be evident to a naive user, especially since they are commonly referred to by their abbreviations. These tools can capture the expertise of algorithm developers, as well as the experience of experts at local institutions who have spent significant periods of time learning how best to apply specific tools to the analysis needs of the laboratory. The tools will inform the users of missing processing stages, suggest available and verified processing modules, and warn of incompatible data types.

6. Conclusions

We have used a combination of an executable provenance XSD incorporated into LONI Pipeline modules to capture processing provenance and description. One of the major strengths of this system is the capacity to easily recreate the processing applied to a file by viewing its provenance file, extracting the workflow, and then rerunning it in the LONI Pipeline. The LONI Pipeline can accommodate almost any form of workflow, the underlying architecture is application agnostic, not limiting the kind of science that can be examined within it. LONI Pipeline workflows can therefore serve to document workflow provenance in almost any field of endeavor.

In an era where digital information underlies much of the scientific enterprise and the manipulation of that data has become increasingly complex, the recording of data and methods provenance takes on greater importance. In this article, we describe an XML-based neuroimaging provenance description

that can be implemented in any workflow environment. We envision the LONI Pipeline as fulfilling a role for neuroimaging similar to other frameworks in chemistry or high-energy physics. We believe that data and workflow provenance form a major element of the program that promotes data processing methods description, data sharing, and study replication.

References

1. Murphy, S.N., et al., *A Web Portal that Enables Collaborative Use of Advanced Medical Image Processing and Informatics Tools through the Biomedical Informatics Research Network (BIRN)*. AMIA Annu Symp Proc, 2006: p. 579-83.
2. Simmhan, Y.L., B. Plale, and D. Gannon, *A survey of data provenance in e-science*. Sigmod Record, 2005. **34**(3): p. 31-36.
3. Moreau, L., et al., *Special Issue: The First Provenance Challenge*. Concurrency and Computation: Practice & Experience, 2007. **00**.
4. Zhao, Y., et al., *A notation and system for expressing and executing cleanly typed workflows on messy scientific data*. Sigmod Record, 2005. **34**(3): p. 37-43.
5. Zhao, Y., M. Wilde, and I. Foster, *Applying the virtual data provenance model*. Provenance and Annotation of Data, 2006. **4145**: p. 148-161.
6. Liu, L., et al., *Multiple sclerosis medical image analysis and information management*. J Neuroimaging, 2005. **15**(4 Suppl): p. 103S-117S.
7. Fleisher, A.S., et al., *Identification of Alzheimer disease risk by functional magnetic resonance imaging*. Arch Neurol, 2005. **62**(12): p. 1881-8.
8. Mueller, S.G., et al., *Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)*. Alzheimers Dement, 2005. **1**(1): p. 55-66.
9. Rusinek, H., et al., *Regional brain atrophy rate predicts future cognitive decline: 6-year longitudinal MR imaging study of normal aging*. Radiology, 2003. **229**(3): p. 691-6.
10. Langen, M., et al., *Caudate nucleus is enlarged in high-functioning medication-naïve subjects with autism*. Biol Psychiatry, 2007. **62**(3): p. 262-6.
11. Drevets, W.C., *Neuroimaging studies of mood disorders*. Biol Psychiatry, 2000. **48**(8): p. 813-29.
12. Narr, K.L., et al., *Asymmetries of cortical shape: Effects of handedness, sex and schizophrenia*. Neuroimage, 2007. **34**(3): p. 939-48.
13. Mazziotta, J.C., et al., *A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM)*. Neuroimage, 1995. **2**(2): p. 89-101.
14. Van Horn, J.D., et al., *Sharing neuroimaging studies of human cognition*. Nat Neurosci, 2004. **7**(5): p. 473-81.
15. Erberich, S.G., et al., *Globus MEDICUS - Federation of DICOM Medical Imaging Devices into Healthcare Grids*. Stud Health Technol Inform, 2007. **126**: p. 269-78.
16. Martone, M.E., et al., *The cell-centered database: a database for multiscale structural and protein localization data from light and electron microscopy*. Neuroinformatics, 2003. **1**(4): p. 379-95.
17. Bidgood, W.D., Jr., et al., *Understanding and using DICOM, the data interchange standard for biomedical imaging*. J Am Med Inform Assoc, 1997. **4**(3): p. 199-212.
18. Zhao, J., et al., *Semantically linking and browsing provenance logs for e-science*. Semantics of a Networked World: Semantics for Grid Databases, 2004. **3226**: p. 158-176.
19. Rex, D.E., J.Q. Ma, and A.W. Toga, *The LONI Pipeline Processing Environment*. Neuroimage, 2003. **19**(3): p. 1033-48.

20. Halfhill, T.R., *The Truth Behind the Pentium Bug*. Byte, 1995.
21. Vaughan, G.V., *GNU Autoconf, Automake, and Libtool*. 1st ed. 2000, Indianapolis: New Riders. xx, 390.
22. Dale, A.M., B. Fischl, and M.I. Sereno, *Cortical surface-based analysis. I. Segmentation and surface reconstruction*. Neuroimage, 1999. **9**(2): p. 179-94.
23. Shattuck, D.W. and R.M. Leahy, *BrainSuite: an automated cortical surface identification tool*. Med Image Anal, 2002. **6**(2): p. 129-42.
24. Sled, J.G., A.P. Zijdenbos, and A.C. Evans, *A nonparametric method for automatic correction of intensity nonuniformity in MRI data*. IEEE Trans Med Imaging, 1998. **17**(1): p. 87-97.
25. Jenkinson, M., et al., *Improved optimization for the robust and accurate linear registration and motion correction of brain images*. Neuroimage, 2002. **17**(2): p. 825-41.
26. Jenkinson, M. and S. Smith, *A global optimisation method for robust affine registration of brain images*. Med Image Anal, 2001. **5**(2): p. 143-56.
27. Smith, S.M., et al., *Advances in functional and structural MR image analysis and implementation as FSL*. Neuroimage, 2004. **23 Suppl 1**: p. S208-19.
28. Myers, J.D., et al., *A collaborative informatics infrastructure for multi-scale science*. Cluster Computing-the Journal of Networks Software Tools and Applications, 2005. **8**(4): p. 243-253.
29. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows*. Bioinformatics, 2004. **20**(17): p. 3045-3054.
30. Keator, D.B., et al., *A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels*. Neuroinformatics, 2006. **4**(2): p. 199-212.
31. Freire, J., et al., *Provenance for computational tasks: A survey*. Computing in Science & Engineering, 2008. **10**(3): p. 11-21.
32. Smith, S.M., *Fast robust automated brain extraction*. Hum Brain Mapp, 2002. **17**(3): p. 143-55.

Acknowledgments: This work was generously supported by a research grants from the National Institutes of Health through the NIH Roadmap for Medical Research (U54 RR021813), the National Center for Research Resources (U24 RR021760 [Mouse BIRN] and P41 RR013642), and the National Institute of Mental Health (R01 MH071940). The authors wish to acknowledge their deep appreciation to the members of the Laboratory of Neuro Imaging (LONI).