Harmonization of Methods to Facilitate Reproducibility in Medical Data Processing: Applications to Diffusion Tensor Magnetic Resonance Imaging

Jeffrey Jenkins[1,2], Lin-Ching Chang[2], Elizabeth Hutchinson[1], M. Okan Irfanoglu[1], Carlo Pierpaoli[1]

[1]*Dept. of Electrical Engineering and Computer Science, the Catholic University of America, Washington D.C.*
[2]*Section for Quantitative Imaging and Tissue Sciences, NICHD, National Institutes of Health, Bethesda MD*
54jenkins@cua.edu, changl@cua.edu, elizabeth.hutchinson@nih.gov, irfanogl@gmail.com, cp1a@nih.gov

Abstract — **Data and methodology sharing is essential for progression of scientific research. Several research groups have built tools for medical big data (MBD) processing applicable to Diffusion Tensor MRI (DTI) processing pipelines. In this paper, we propose a framework enabling methodology sharing (i.e. harmonization) to facilitate the reproducibility in DTI processing.**

Keywords: medical big data, method sharing, diffusion tensor MRI, reproducibility, workflow automation

I. Introduction

Data and methodology sharing is important for science, as it helps researchers to more easily reproduce the results of others. The National Institutes of Health (NIH) has described data sharing as ‘essential for expedited translation of research results into knowledge, products, and procedures to improve human health’, and encourages the sharing of final research data. Methodology sharing would be the next focus to streamline research.

In 2014, the International Data Corporation revealed an estimate that 0.15% of the world’s population was considered expert software developers [1]. One can assume that the majority of scientific software users are not in this small percentile. Research projects often possess a constantly growing collection of data, where integrating even a single new data point may require reprocessing of the entire dataset which may be challenging if the processing workflow has not been saved. Aside from management of the data, the next biggest hurdle for the medical big data (MBD) community lies in improving research efficiency through workflow automation. Numerous applications have been developed which combine common operations into a user interface to simplify the data exploration and workflow creation process [2]. Other tools are available which require programming knowledge and are often very complex to learn, so users typically import/export data through a string of multiple applications.

II. Background

The workflow concept has been mathematically formalized, extended to virtually all academic disciplines, and utilized by manufacturing and other industrial entities worldwide [3]. Figure 1 shows the ecosystem diagram of our proposed framework. Over the past decade, many tools for processing and visualizing medical images have been released. For examples, tools for processing multi-dimensional data, filtering and enhancing 2D histological stain imagery, and registration modules etc.

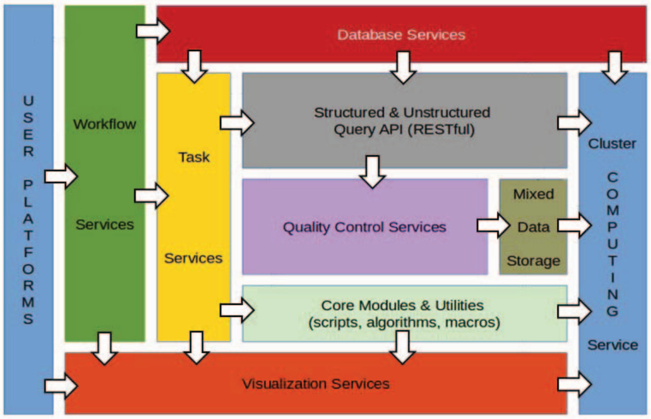

Figure 1 - Ecosystem diagram for the proposed layered workflow and visualization framework.

The DTI acquisition and processing pipeline, generates an enormous amount of information regarding the microscopic properties of biological tissue. Hidden microstructural features are revealed via quantitative maps derived from *in-vivo* DTI data. These maps often resemble histological stains of microstructural components within *ex-vivo* tissue obtained via optical imaging [4].

Performing a typical DTI workflow is an extremely tedious and time consuming process that is prone to errors from both the users as well as algorithms used. Several image processing pipelines have been envisioned by the TORTOISE lab and their collaborators and implemented as scripts which one can run on the command line using a set of input parameters [5]. Some of these pipelines have already been implemented as preset workflows using the proposed framework and exposed in the web repository, found on GitHub [6]. Additional, modules to perform evaluation of different algorithms performing similar operations can be easily compared, which is useful for determining their accuracy and ultimately, their medical utility for diagnosis [7].

III. Framework Overview

In the traditional software development process, translating requirements into a working distributed system is both time-consuming and difficult, requiring several stages of manual development and deployment. This complex, error-prone task can be effectively streamlined using a higher-level, component-based framework.

The components in a framework are tools which utilize web services to strictly import and export information. Encapsulation of the low-level data transformation methodologies allows for these tools to be easily translated into

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

distributed, high-level services that are easier to develop, manipulate, and debug. Tools can be easily composited into implementation-level data flows without the user or developer having to track complex middleware concepts. Furthermore, the implementation-level services can run on any machine across the network by virtue of the built-in dynamic deployment [8]. Our software design required an approach for providing a loose coupling between disparate, heterogeneous data services as well as legacy software components. For that reason, we followed approaches for taking the packaged application through the Service Oriented Architecture maturity model, and the resulting system diagram is shown in figure 2. This figure is an end-to-end, high level snapshot of how services within our software are sequentially utilized for both processing and visualization of group templates with annotations.

A holistic data management strategy capable of handling multi-modal representations of image data sets is paramount for maximizing the utility of large medical image data sets. This is especially true for voxel-wise comparison of high-resolution ground truth histological tissue features directly with quantitative DTI metrics derived from the same organism. This type of comparison is very difficult and visually driven. For example, if one is investigating a region of interest (ROI) in DTI, and trying to locate the same ROI in histology images, defining the boundary of an ROI cannot be easily performed due to the mismatched image resolution.

IV. Application

We will now describe how we utilize the proposed architecture to construct a workflow able to perform multi-modal, multi-scale image co-localization, analysis and visualization. The different resolutions of images generated by different imaging devices and countless possible experiment workflows can preclude a general registration strategy. Microscale features must be preserved over all scales in order to provide reliable interrogation of co-localized data. Interactive data operations like zooming into heterogeneous datasets requires visualization and quantitative analysis services. Direct comparison of DTI to histology requires co-registration pre-processing of both, DTI images and histological sections. Histology images are often very high resolution 2D RGB slices while DTI images are relatively low resolution 3D volumes. In figure 3 (top row), zooming into an ultra-high resolution coronal mouse tissue section, captured with fluorescence microscopy is shown with co-localized DTI data (bottom row). This ROI reveals neuronal fiber tracts within the tissue micro-architecture.

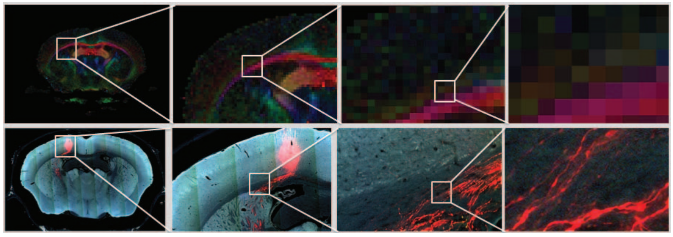

Figure 3 – Level of Detail available depends on imaging modality. Direction encoded color map of mouse (top) demonstrating resolution challenges for *in-vivo* DTI imaging; (bottom) Fluorescent axonal staining shown for histology of the same mouse and tissue slice as in top.

Histology data is immensely useful when desiring to investigate tissue microstructure. However, DTI data is useful for inspecting more macroscopic, whole-brain structures. To co-register histology and MRI data, the registration procedure must reverse histological and MRI deformations on a slice-by-slice basis and then find a global transformation that maps the 2D histological slice onto the correct position in the DTI 4D volume.

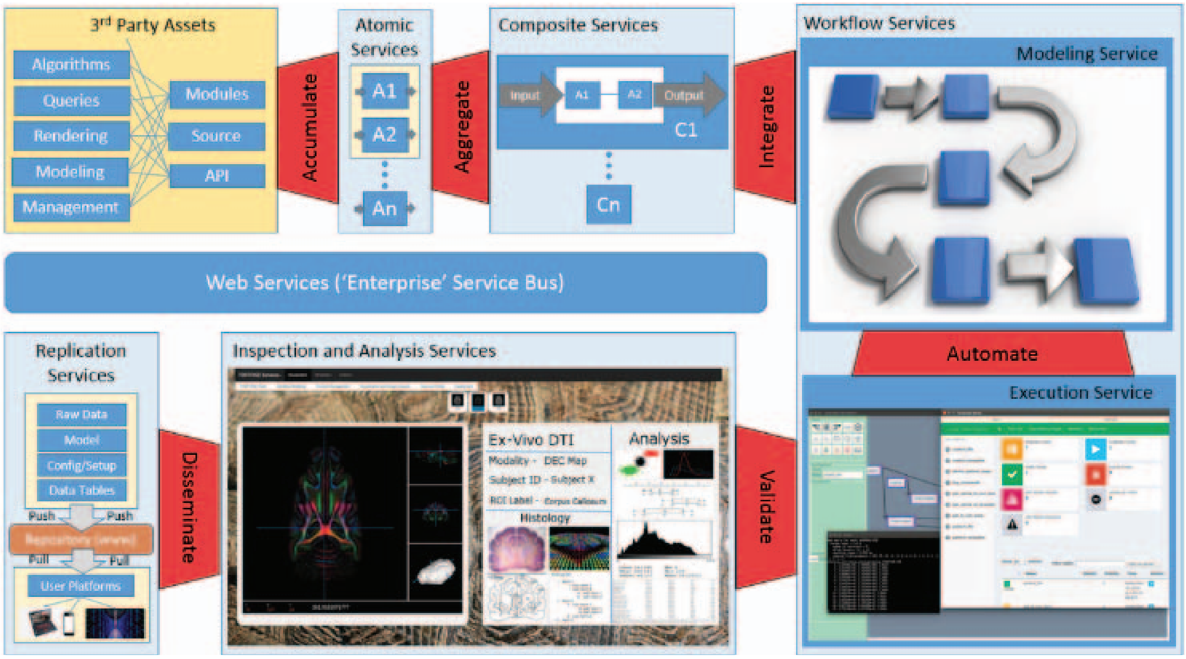

Figure 2. DTI Service Component Architecture for Reproducibility as a Service delivered via a web browser.

There are several challenges when dealing with this medical big data that spans the micro and macroscopic image modalities. For examples, artifacts that often appear in raw DTI data can be corrected automatically, but in some cases, determining whether data is an artifact or real is difficult [9]. Tractographic maps generated from quantitative DT maps have the potential to help surgeons gain a spatial understanding of tissue for surgical planning, but again, artifacts in the data can cause spurious tracts to be generated. The flexible nature of our proposed architecture encourages pre-existing and new modules to be seamlessly integrated as 'nodes' and added to the workflow to form a directed acyclic graph. Development of specific nodes for longitudinal analysis of DTI data and combined analysis of histological and DTI data have been added to the atomic services.

In whole-slice histological sections, powerful multi-resolution file types are able to encode many levels of detail, but one must manually explore pixel ROIs in order to find meaningful features. The result of execution of these services in the same histology and DTI co-localization workflow is shown in figure 4 (bottom row) for coronal whole-slice Rhesus monkey data. The heat-map overlay represents a histological stain of cell bodies by using the Nissl method. Zooming into multi-scale, multi-modal results like the species-invariant workflow here, allows users to observe features of histology that are co-located with features within DTI maps. This particular workflow is an extremely interesting approach for producing combined DTI and histological atlases based on spatial correlates.

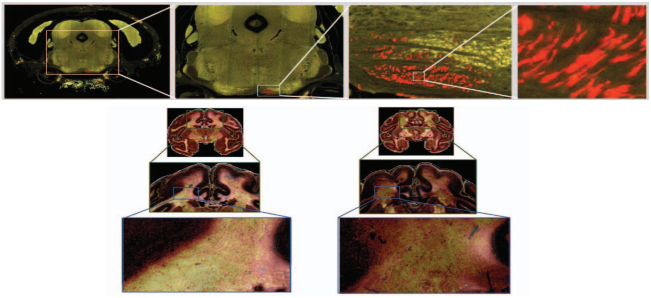

Figure 4 –Workflow harmonization: (top) Zooming into mouse Fluorescence Microscopy coronal slice reveals biomarkers via neuronal fibers. (bottom) Rhesus monkey coronal slice DTI and Histology co-localization; left is control, right lesioned.

V. Discussion and Conclusion

Reproducibility as a Service will be a focal point for the next generation of scientific software, but we must define a set of standards of 'data quality', as well as standards for software which transforms data.

Thus, the data science community could aggregate a robust library of tools and test data, then define metrics for assessing the integrity of modality-specific performance quality scores for a general task. This judging system might provide an unbiased aid to non-expert users. In the near future, we intend to host a DTI pipeline 'challenge' using our proposed framework. This challenge seeks to systematically evaluate the performance of these pipelines, and provide results to the community. One could construct a 'social media of science' platform for users to build, execute, share, and tender workflows. Data mining techniques can then be applied to the embedded knowledge locked in the raw data and extracted by user defined transformations.

Software efficiency is observed when the total input costs are minimized yet the output quality is maximized. Input costs for developing a workflow can become exponential due to a combination of the 'learning curve' for tools and 'processing time', which is incurred from data processing and quality control. The goal of the proposed framework is to simultaneously increase efficiency and quality of service for MBD analytics, through minimization of error-prone resources [10, 11].

Dealing head on with the complexities of Big Data efforts is often the catalyst for creation of value-added project specific capabilities. Empowering researchers who are masters of their domain to master data efficiency will add value to the community of science in more ways than ever before. Although the DTI workflow was used as an application to showcase the design concept, the proposed framework can be easily adapted to other applications such as Biology big data and Astronomy big data.

Acknowledgment

This work was supported in part by CDMRP grant #W81XWH-13-2-0019, and award #3251 from The Henry M. Jackson Foundation. We thank Cibu Thomas, Aaron Renberg, and Alexander Cummins for tissue section data sets and Anne Uyeda for editing the document and useful suggestions.

References

1. International Data Corporation, (2013). 2014 *Worldwide Software Developer and ICT-Skilled Worker Estimates*, IDC #244709
2. Rex, D. E., Ma, J. Q., and Toga, AW. (2003) *The LONI Pipeline Processing Environment*. Neuroimage, 19(3): 1033-48.
3. Hilton, B. C., (2005). *A History of Production Planning and Control, 1750-2000,* p. 64
4. Basser, P. J., Pierpaoli, C. (2011), *Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI*, Journal of Magnetic Resonance, Volume 213, Issue 2, Pages 560-570.
5. Pierpaoli, C., Walker, L., Irfanoglu, M. O., Barnett, A., Basser, P., Chang, L-C, Koay, C., Pajevic, S., Rohde, G., Sarlls, J. and Wu, M. (2010). *TORTOISE: an integrated software package for processing of diffusion MRI data*. in *ISMRM*.
6. Jenkins, J., (2016) Tortoise Flow Builder; software available at https://github.com/jjenki11/tortoise-flow-builder.
7. Irfanoglu, M. O., Nayak, A., Jenkins, J., Hutchinson, E. B., Sadeghi, N., Thomas, C. P., Pierpaoli, C., (2016). *DR-TAMAS: Diffeomorphic Registration for Tensor Accurate Alignment of Anatomical Structures*, NeuroImage, Volume 132, Pages 439-454.
8. Webster, S., Miller, G., Mayott, G., (2012) Software as a service approach to sensor simulation software deployment. Proc. SPIE 8403.
9. Chang, L.-C., Jones, D. K. and Pierpaoli, C. (2005), RESTORE: Robust estimation of tensors by outlier rejection. Magn. Reson. Med., 53: 1088–1095.
10. Monahan, 2013. https://www.cebglobal.com/ blogs/big-data-calls-for-big-judgment-in-finance-and-beyond.
11. Ziemann M., Eren Y., El-Osta A., (2016). Gene name errors are widespread in the scientific literature. Genome Biol. 17(1):177.