

MAPPING: MANAGEMENT AND PROCESSING OF IMAGES FOR POPULATION IMAGING

EDITED BY : Michel Dojat, Wiro Niessen and David N. Kennedy
PUBLISHED IN: Frontiers in ICT and Frontiers in Neuroinformatics



frontiers

Frontiers Copyright Statement

© Copyright 2007-2017 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714

ISBN 978-2-88945-260-6

DOI 10.3389/978-2-88945-260-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

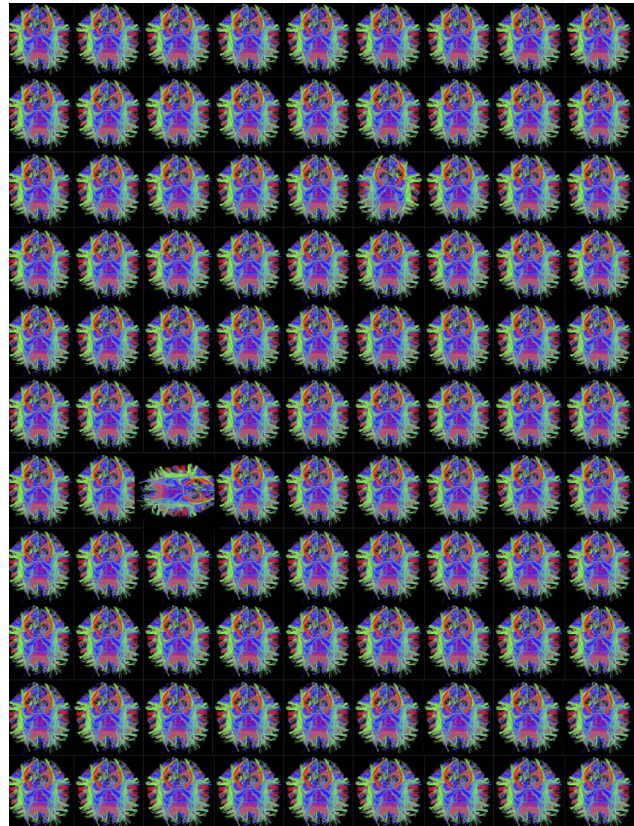
MAPPING: MANAGEMENT AND PROCESSING OF IMAGES FOR POPULATION IMAGING

Topic Editors:

Michel Dojat, INSERM, Université Grenoble Alpes, France

Wiro Niessen, Erasmus University Rotterdam, Netherlands

David N. Kennedy, University of Massachusetts Medical School, United States



Fiber tracking using Diffusion Tensor Imaging on a Large Data Set.

Copyright M. Dojat, INSERM

Several recent papers underline methodological points that limit the validity of published results in imaging studies in the life sciences and especially the neurosciences (Carp, 2012; Ingre, 2012; Button et al., 2013; Ioannidis, 2014). At least three main points are identified that lead to biased conclusions in research findings: endemic low statistical power and, selective outcome and

selective analysis reporting. Because of this, and in view of the lack of replication studies, false discoveries or solutions persist. To overcome the poor reliability of research findings, several actions should be promoted including conducting large cohort studies, data sharing and data reanalysis.

The construction of large-scale online databases should be facilitated, as they may contribute to the definition of a “collective mind” (Fox et al., 2014) facilitating open collaborative work or “crowd science” (Franzoni and Sauermann, 2014). Although technology alone cannot change scientists’ practices (Wichert et al., 2011; Wallis et al., 2013, Poldrack and Gorgolewski 2014; Roche et al. 2014), technical solutions should be identified which support a more “open science” approach. Also, the analysis of the data plays an important role. For the analysis of large datasets, image processing pipelines should be constructed based on the best algorithms available and their performance should be objectively compared to diffuse the more relevant solutions. Also, provenance of processed data should be ensured (MacKenzie-Graham et al., 2008). In population imaging this would mean providing effective tools for data sharing and analysis without increasing the burden on researchers.

This subject is the main objective of this research topic (RT), cross-listed between the specialty section “Computer Image Analysis” of *Frontiers in ICT* and *Frontiers in Neuroinformatics*. Firstly, it gathers works on innovative solutions for the management of large imaging datasets possibly distributed in various centers. The paper of Danso et al. describes their experience with the integration of neuroimaging data coming from several stroke imaging research projects. They detail how the initial NeuroGrid core metadata schema was gradually extended for capturing all information required for future metaanalysis while ensuring semantic interoperability for future integration with other biomedical ontologies. With a similar preoccupation of interoperability, Shanoir relies on the OntoNeuroLog ontology (Temal et al., 2008; Gibaud et al., 2011; Batrancourt et al., 2015), a semantic model that formally described entities and relations in medical imaging, neuropsychological and behavioral assessment domains. The mechanism of “Study Card” allows to seamlessly populate metadata aligned with the ontology, avoiding fastidious manual entrance and the automatic control of the conformity of imported data with a predefined study protocol. The ambitious objective with the BIOMIST platform is to provide an environment managing the entire cycle of neuroimaging data from acquisition to analysis ensuring full provenance information of any derived data. Interestingly, it is conceived based on the product lifecycle management approach used in industry for managing products (here neuroimaging data) from inception to manufacturing. Shanoir and BIOMIST share in part the same OntoNeuroLog ontology facilitating their interoperability. ArchiMed is a data management system locally integrated for 5 years in a clinical environment. Not restricted to Neuroimaging, ArchiMed deals with multi-modal and multi-organs imaging data with specific considerations for data long-term conservation and confidentiality in accordance with the French legislation. Shanoir and ArchiMed are integrated into FLI-IAM¹, the national French IT infrastructure for in vivo imaging.

Secondly, dedicated software and hardware infrastructures are proposed for the sharing and execution of image processing workflows making easier the replication and comparison of data analysis procedures. The contribution of Das et al. presents the functionalities added to the LORIS-CBRAIN software ecosystem to fulfill the technical challenges raised by supporting an Open Science approach. Specific mechanisms have been introduced for ensuring privacy and security of the stored data, quality control checking and heterogeneous tools integration. Fastr is a workflow engine dedicated to the automation of complex medical imaging processing pipelines. It allows the composition of different software elements to design pipelines, checks datatype compatibility of linked outputs and inputs, ensures data provenance and finally creates a list of jobs for execution. In the same vein, OpenMOLE is designed to optimize execution of workflows on distributed computing architectures. Although no specific application domain is targeted by OpenMOLE, case studies are reported to illustrate its suitability to neuroimaging data processing. How to document data provenance to facilitate processed data sharing and reuse is the question explored by Pauli et al. from datasets processed using the most common

software package used in Neuroimaging. They provide a set of results as a benchmark for testing automated provenance software.

Finally, two papers are more concerned with the usage of such platforms. Serag et al. propose SEGMA, a supervised solution for brain tissue and structure segmentation combining sparse training data selection, linear registration and random forest classification for processing large MR datasets with a reduced computational time. Brain atlases are often used by automated workflows for imaging population studies. The paper by Dickie et al. reviews the brain MRI atlases currently available, which appear of modest size, based on limited image sequences and where some populations are under-represented. The next challenge is then to develop nonparametric brain atlases including a wide number of parameters extracted from different imaging sequences from a large set of individuals, representative of more different classes of population.

To conclude, this RT demonstrates that, since the pioneer experiments of neuroimaging data sharing with the fMRIDC project (Van Horn and Gazzaniga, 2013) or the BIRN initiative (Keator et al., 2008), many technical efforts have been performed or are currently underway to facilitate data and tools sharing. Solutions now exist that are mature enough to help us make substantial changes to how we conduct health research (Chan et al., 2014), improving reproducibility and quality of published research findings.

(1) <https://project.inria.fr/fli/en/>

References

- Carp J (2012) The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63(1):289-300.
- Ingre M (2013) Why small low-powered studies are worse than large high-powered studies and how to protect against “trivial” findings in research: comment on Friston (2012). *NeuroImage* 81:496-498.
- Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, & Munafo MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14(5):365-376.
- Ioannidis JP (2005) Why most published research findings are false. *PLoS Med* 2(8): e124.
- Mackenzie-Graham AJ, Van Horn JD, Woods RP, Crawford KL, & Toga AW (2008) Provenance in neuroimaging. *NeuroImage* 42(1):178-195.
- Fox PT, Lancaster JL, Laird AR, & Eickhoff SB (2014) Meta-analysis in human neuroimaging: computational modeling of large-scale databases. *Annu Rev Neurosci* 37:409-434.
- Franzoni C & Sauermann H (2014) Crowd science: The organization of scientific research in open collaborative projects. *Research Policy* 43(1):1-20.
- Wicherts JM, Bakker M, & Molenaar D (2011) Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One* 6(11): e26828.
- Wallis JC, Rolando E, & Borgman CL (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One* 8(7): e67332.
- Poldrack RA & Gorgolewski KJ (2014) Making big data open: data sharing in neuroimaging. *Nat Neurosci* 17(11):1510-1517.
- Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, & Kruuk LE (2014) Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* 12(1):e1001779.
- Temal L, Dojat M, Kassel G, & Gibaud B (2008) Towards an ontology for sharing medical images and regions of interest in neuroimaging. *J Biomed Inform* 41:766-778.
- Gibaud B, Kassel G, Dojat M, Batrancourt B, Michel F, Gaignard A, Montagnat J. 2011. NeuroLOG: sharing neuroimaging data using an ontology-based federated approach. *AMIA Annu Symp Proc*. 2011:472-480.
- Batrancourt B, Dojat M, Gibaud B, & Kassel G (2015) A multilayer ontology of instruments for neurological, behavioral and cognitive assessments. *Neuroinformatics* 13(1): 93-110.
- Danso SO, Job DE, Gonzalez DR, Dickie DA, Palmer J, Ure J, Bath PM, Sandercock PAG, Wardlaw JM (2016) Developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular Disease: Our Experience from the Stroke Image Bank Project. *Frontiers in ICT*. 3(32).
- Das S et al. (2017) Cyberinfrastructure for Open Science at the Montreal Neurological Institute. *Frontiers in Neuroinformatics*. 10(53).
- Pauli R, Bowring A, Reynolds R, Chen G, Nichols TE, Maumet C (2016) Exploring fMRI Results Space: 31 Variants of an fMRI Analysis in AFNI, FSL, and SPM. *Frontiers in Neuroinformatics*. 10(24).

- Serag A, Wilkinson AG, Telford EJ, Pataky R, Sparrow SA, Anblagan D, Macnaught G, Semple SI, Boardman JP (2017) SEGMA: An Automatic SEGmentation Approach for Human Brain MRI Using Sliding Window and Random Forests. *Frontiers in Neuroinformatics*. 11(2).
- Dickie DA, Shenkin SD, Anblagan D, Lee J, Blesa Cabez M, Rodriguez D, Boardman JP, Waldman A, Job DE, Wardlaw JM (2017) Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging. *Frontiers in Neuroinformatics*. 11(1).
- Van Horn JD & Gazzaniga MS (2013) Why share data? Lessons learned from the fMRIDC. *NeuroImage* 82:677-682.
- Keator DB, Grethe JS, Marcus D, Ozyurt B, Gadde S, Murphy S, Pieper S, Greve D, Notestine R, Bockholt HJ, Papadopoulos P, Function B, Morphometry B, & Coordinating B (2008) A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans Inf Technol Biomed* 12(2):162-172.
- Chan A-W, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, Krumholz HM, Ghersi D, & van der Worp HB (2014) Increasing value and reducing waste: addressing inaccessible research. *The Lancet* 383(9913):257-266.

Citation: Dojat, M., Niessen, W., Kennedy, D. N., eds. (2017). Mapping: Management and Processing of Images for Population Imaging. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-260-6

Table of Contents

- 08 Editorial: MAPPING: MANagement and Processing of Images for Population ImagiNG**
Michel Dojat, David N. Kennedy and Wiro Niessen
- 11 Developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular Disease: Our Experience from the Stroke Image Bank Project**
Samuel O. Danso, Dominic E. Job, David Rodriguez Gonzalez, David Alexander Dickie, Jeb Palmer, Jenny Ure, Philip M. Bath, Peter A. G. Sandercock and Joanna M. Wardlaw
- 23 Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories**
Christian Barillot, Elise Bannier, Olivier Commowick, Isabelle Corouge, Anthony Baire, Ines Fakhfakh, Justine Guillaumont, Yao Yao and Michael Kain
- 38 BIOMIST: A Platform for Biomedical Data Lifecycle Management of Neuroimaging Cohorts**
Marianne Allanic, Pierre-Yves Hervé, Cong-Cuong Pham, Myriam Lekkal, Alexandre Durupt, Thierry Brial, Arthur Grioche, Nada Matta, Philippe Boutinaud, Benoit Eynard and Marc Joliot
- 57 ArchiMed: A Data Management System for Clinical Research in Imaging**
Emilien Micard, Damien Husson, CIC-IT Team and Jacques Felblinger
- 68 Cyberinfrastructure for Open Science at the Montreal Neurological Institute**
Samir Das, Tristan Glatard, Christine Rogers, John Saigle, Santiago Paiva, Leigh MacIntyre, Mouna Safi-Harab, Marc-Etienne Rousseau, Jordan Stirling, Najmeh Khalili-Mahani, David MacFarlane, Penelope Kostopoulos, Pierre Rioux, Cecile Madjar, Xavier Lecours-Boucher, Sandeep Vanamala, Reza Adalat, Zia Mohaddes, Vladimir S. Fonov, Sylvain Milot, Ilana Leppert, Clotilde Degroot, Thomas M. Durcan, Tara Campbell, Jeremy Moreau, Alain Dagher, D. Louis Collins, Jason Karamchandani, Amit Bar-Or, Edward A. Fon, Rick Hoge, Sylvain Baillet, Guy Rouleau and Alan C. Evans
- 81 Fastr: A Workflow Engine for Advanced Data Flows in Medical Image Analysis**
Hakim C. Achterberg, Marcel Koek and Wiro J. Niessen
- 92 Reproducible Large-Scale Neuroimaging Studies with the OpenMOLE Workflow Management System**
Jonathan Passerat-Palmbach, Romain Reuillon, Mathieu Leclaire, Antonios Makropoulos, Emma C. Robinson, Sarah Parisot and Daniel Rueckert
- 108 Exploring fMRI Results Space: 31 Variants of an fMRI Analysis in AFNI, FSL, and SPM**
Ruth Pauli, Alexander Bowring, Richard Reynolds, Gang Chen, Thomas E. Nichols and Camille Maumet

114 ***SEGMA: An Automatic SEGmentation Approach for Human Brain MRI Using Sliding Window and Random Forests***

Ahmed Serag, Alastair G. Wilkinson, Emma J. Telford, Rozalia Pataky, Sarah A. Sparrow, Devasuda Anblagan, Gillian Macnaught, Scott I. Semple and James P. Boardman

125 ***Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging***

David Alexander Dickie, Susan D. Shenkin, Devasuda Anblagan, Juyoung Lee, Manuel Blesa Cabez, David Rodriguez, James P. Boardman, Adam Waldman, Dominic E. Job and Joanna M. Wardlaw



Editorial: MAPPING: MAnagement and Processing of Images for Population ImagiNG

Michel Dojat^{1*}, David N. Kennedy² and Wiro Niessen³

¹U1216-GIN, INSERM, Site Santé, La Tronche, France, ²Medical School, University of Massachusetts, Worcester, MA, United States, ³Erasmus University, Rotterdam, Netherlands

Keywords: data sharing, neuroimaging, brain, magnetic resonance imaging, image processing, computer-assisted

Editorial on the Research Topic

MAPPING: MAnagement and Processing of Images for Population ImagiNG

Several recent papers underline methodological points that limit the validity of published results in imaging studies in the life sciences and especially the neurosciences (Ioannidis, 2005; Carp, 2012; Button et al., 2013; Ingre, 2013). At least three main points are identified that lead to biased conclusions in research findings: endemic low statistical power, selective outcome, and selective analysis reporting. Because of this, and in view of the lack of replication studies, false discoveries or solutions persist. To overcome the poor reliability of research findings, several actions should be promoted including conducting large cohort studies, data sharing, and data reanalysis. The construction of large-scale online databases should be facilitated, as they may contribute to the definition of a “collective mind” (Fox et al., 2014) facilitating open collaborative work or “crowd science” (Franzoni and Sauermann, 2014). Although technology alone cannot change scientists’ practices (Wicherts et al., 2011; Wallis et al., 2013; Poldrack and Gorgolewski, 2014; Roche et al., 2014), technical solutions should be identified, which support a more “open science” approach. Also, the analysis of the data plays an important role. For the analysis of large datasets, image processing pipelines should be constructed based on the best algorithms available and their performance should be objectively compared to diffuse the more relevant solutions. Also, provenance of processed data should be ensured (MacKenzie-Graham et al., 2008). In population imaging, this would mean providing effective tools for data sharing and analysis without increasing the burden on researchers. This subject is the main objective of this research topic (RT), cross-listed between the specialty section “Computer Image Analysis” of Frontiers in ICT and Frontiers in Neuroinformatics. First, it gathers works on innovative solutions for the management of large imaging datasets possibly distributed in various centers. The paper of Danso et al. describes their experience with the integration of neuroimaging data coming from several stroke imaging research projects. They detail how the initial NeuroGrid core metadata schema was gradually extended for capturing all information required for future meta-analysis while ensuring semantic interoperability for future integration with other biomedical ontologies. With a similar preoccupation of interoperability, Shanoir relies on the OntoNeuroLog ontology (Temal et al., 2008; Gibaud et al., 2011; Batrancourt et al., 2015), a semantic model that formally described entities and relations in medical imaging, neuropsychological, and behavioral assessment domains. The mechanism of “Study Card” allows to seamlessly populate metadata aligned with the ontology, avoiding fastidious manual entrance and the automatic control of the conformity of imported data with a predefined study protocol. The ambitious objective with the BIOMIST platform is to provide an environment managing the entire cycle of neuroimaging data from acquisition to analysis ensuring full provenance information of any derived data. Interestingly, it is conceived based on the product lifecycle management approach used in industry for managing products (here

OPEN ACCESS

Edited and Reviewed by:

Kaleem Siddiqi,
McGill University, Canada

*Correspondence:

Michel Dojat
michel.dojat@univ-grenoble-alpes.fr

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 28 April 2017

Accepted: 29 June 2017

Published: 17 July 2017

Citation:

Dojat M, Kennedy DN and Niessen W
(2017) Editorial: MAPPING:
MAnagement and Processing of
Images for Population ImagiNG.
Front. ICT 4:18.
doi: 10.3389/fict.2017.00018

neuroimaging data) from inception to manufacturing. Shanoir and BIOMIST share in part the same OntoNeuroLog ontology facilitating their interoperability. ArchiMed is a data management system locally integrated for 5 years in a clinical environment. Not restricted to Neuroimaging, ArchiMed deals with multimodal and multi-organs imaging data with specific considerations for data long-term conservation and confidentiality in accordance with the French legislation. Shanoir and ArchiMed are integrated into FLI-IAM,¹ the national French IT infrastructure for *in vivo* imaging.

Second, dedicated software and hardware infrastructures are proposed for the sharing and execution of image-processing workflows making easier the replication and comparison of data analysis procedures. The contribution of Das et al. presents the functionalities added to the LORIS-CBRAIN software ecosystem to fulfill the technical challenges raised by supporting an Open Science approach. Specific mechanisms have been introduced for ensuring privacy and security of the stored data, quality control checking, and heterogeneous tools integration. Fastr is a workflow engine dedicated to the automation of complex medical imaging processing pipelines. It allows the composition of different software elements to design pipelines, checks datatype compatibility of linked outputs and inputs, ensures data provenance, and finally creates a list of jobs for execution. In the same vein, OpenMOLE is designed to optimize execution of workflows on distributed computing architectures. Although no specific application domain is targeted by OpenMOLE, case studies are reported to illustrate its suitability to neuroimaging data processing. How to document data provenance to facilitate processed data sharing and reuse

is the question explored by Pauli et al. from datasets processed using the most common software package used in Neuroimaging. They provide a set of results as a benchmark for testing automated provenance software.

Finally, two papers are more concerned with the usage of such platforms. Serag et al. propose SEGMA, a supervised solution for brain tissue and structure segmentation combining sparse training data selection, linear registration, and random forest classifier for processing large MR datasets with a reduced computational time. Brain atlases are often used by automated workflows for imaging population studies. The paper by Dickie et al. reviews the brain MRI atlases currently available, which appear of modest size, based on limited image sequences and where some populations are underrepresented. The next challenge is then to develop non-parametric brain atlases including a wide number of parameters extracted from different imaging sequences from a large set of individuals, representative of more different classes of population.

To conclude, this RT demonstrates that, since the pioneer experiments of neuroimaging data sharing with the fMRIDC project (Van Horn and Gazzaniga, 2013) or the BIRN initiative (Keator et al., 2008), many technical efforts have been performed or are currently underway to facilitate data and tools sharing. Solutions now exist that are mature enough to help us make substantial changes to how we conduct health research (Chan et al., 2014), improving reproducibility, and quality of published research findings.

AUTHOR CONTRIBUTIONS

The authors contributed equally to this editorial.

REFERENCES

- Batrancourt, B., Dojat, M., Gibaud, B., and Kassel, G. (2015). A multilayer ontology of instruments for neurological, behavioral and cognitive assessments. *Neuroinformatics* 13, 93–110. doi:10.1007/s12021-014-9244-3
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi:10.1038/nrn3475
- Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63, 289–300. doi:10.1016/j.neuroimage.2012.07.004
- Chan, A.-W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Götzsche, P. C., et al. (2014). Increasing value and reducing waste: addressing inaccessible research. *Lancet* 383, 257–266. doi:10.1016/S0140-6736(13)62296-5
- Fox, P. T., Lancaster, J. L., Laird, A. R., and Eickhoff, S. B. (2014). Meta-analysis in human neuroimaging: computational modeling of large-scale databases. *Annu. Rev. Neurosci.* 37, 409–434. doi:10.1146/annurev-neuro-062012-170320
- Franzoni, C., and Sauermann, H. (2014). Crowd science: the organization of scientific research in open collaborative projects. *Res. Policy* 43, 1–20. doi:10.1016/j.respol.2013.07.005
- Gibaud, B., Kassel, G., Dojat, M., Batrancourt, B., Michel, F., Gagnard, A., et al. (2011). “NeuroLOG: sharing neuroimaging data using an ontology-based federated approach,” in *AMIA 2011 Annual Symposium*, ed. R. Scott Evans.
- Ingre, M. (2013). Why small low-powered studies are worse than large high-powered studies and how to protect against “trivial” findings in research: comment on Friston (2012). *Neuroimage* 81, 496–498. doi:10.1016/j.neuroimage.2013.03.030
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi:10.1371/journal.pmed.0020124
- Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., et al. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172. doi:10.1109/TITB.2008.917893
- MacKenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage* 42, 178–195. doi:10.1016/j.neuroimage.2008.04.186
- Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi:10.1038/nn.3818
- Roche, D. G., Lanfear, R., Binning, S. A., Haff, T. M., Schwanz, L. E., Cain, K. E., et al. (2014). Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol.* 12:e1001779. doi:10.1371/journal.pbio.1001779
- Temal, L., Dojat, M., Kassel, G., and Gibaud, B. (2008). Towards an ontology for sharing medical images and regions of interest in neuroimaging. *J. Biomed. Inform.* 41, 766–778. doi:10.1016/j.jbi.2008.03.002
- Van Horn, J. D., and Gazzaniga, M. S. (2013). Why share data? Lessons learned from the fMRIDC. *Neuroimage* 82, 677–682. doi:10.1016/j.neuroimage.2012.11.010
- Wallis, J. C., Rolando, E., and Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS ONE* 8:e67332. doi:10.1371/journal.pone.0067332
- Wicherts, J. M., Bakker, M., and Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE* 6:e26828. doi:10.1371/journal.pone.0026828

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Dojat, Kennedy and Niessen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the

original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Developing an Integrated Image Bank and Metadata for Large-scale Research in Cerebrovascular Disease: Our Experience from the Stroke Image Bank Project

Samuel O. Danso¹, Dominic E. Job¹, David Rodriguez Gonzalez¹, David Alexander Dickie¹, Jeb Palmer¹, Jenny Ure², Philip M. Bath³, Peter A. G. Sandercock¹ and Joanna M. Wardlaw^{1*}

¹ Brain Research Imaging Centre (BRIC), Centre for Clinical Brain Sciences (CCBS), University of Edinburgh Medical School, Edinburgh, UK, ² MRC Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK, ³ Stroke Trials Unit (STU), Division of Clinical Neuroscience (DCN), University of Nottingham, Nottingham, UK

OPEN ACCESS

Edited by:

David N. Kennedy,
University of Massachusetts
Medical School, USA

Reviewed by:

Paul Kwan,
University of Tsukuba, Japan
K. C. Santosh,
University of South Dakota, USA

*Correspondence:

Joanna M. Wardlaw
joanna.wardlaw@ed.ac.uk

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 26 August 2016

Accepted: 08 December 2016

Published: 26 December 2016

Citation:

Danso SO, Job DE, Gonzalez DR,
Dickie DA, Palmer J, Ure J, Bath PM,
Sandercock PAG and Wardlaw JM
(2016) Developing an Integrated
Image Bank and Metadata for
Large-scale Research in
Cerebrovascular Disease:
Our Experience from the Stroke
Image Bank Project.
Front. ICT 3:32.
doi: 10.3389/fict.2016.00032

A framework for building an infrastructure that semantically integrates, archives, and reuses data for various research purposes in human brain imaging remains critical. In particular, problems of aligning technical, clinical, and professional systems in order to facilitate data sharing are a recurring issue in brain imaging. However, large samples of well-characterized images with detailed metadata are increasingly needed. This paper outlines the experience of the NeuroGrid Stroke Exemplar and further work in the Brain Research Imaging Centre and Stroke Trials Unit in developing an infrastructure that facilitates the linkage, archiving, and reuse of imaging data from stroke patients for large-scale clinical and epidemiological studies. We examined data from 12 past stroke projects carried out over the past two decades in our center and two large trials with 329 centers. We assessed previously published schemas and those developed specifically for large multicentre ischemic and hemorrhagic stroke treatment trials. We then developed our own harmonized and integrated schema and database with a web-based interface system, Longitudinal Online Research and Imaging System (LORIS), aiming to be flexible and adaptable to future trials and observational studies. We then linked image and metadata from 3,079 patients acquired in stroke research in one center in a 14-year period (1996–2010) with prospective central hospital health statistics to obtain long-term follow-up. Our integrated database includes 3,079 subjects and over 550 federated and searchable data items including imaging details, medical history, and examination, stroke, and laboratory details, which map to large multicentre stroke trials with imaging data from over 10,000 patients from 30 countries. The central linkage identified 879 of 3,079 patients had died, 525 had recurrent strokes, and 291 developed dementia during up to a 19-year period (range = 0–19; median = 9.04; IQR = 12.17) of follow-up, demonstrating its utility. The core metadata schema has benefited from extensive development in large clinical trials. Further trials' data can now be added. It provides an opportunity to crosslink and reuse data for a range of large-scale stroke

brain imaging clinical and research purposes including developing data analytics models for research into common brain diseases and their consequences.

Keywords: multicenter imaging, heterogeneous data, metadata schema, ischemic and hemorrhagic, image bank, neuroimaging, data sharing, stroke

INTRODUCTION

There is a global drive to develop strategies and frameworks to facilitate archiving, sharing, and reuse of data obtained from original research projects in order to maximize the value of the data (Pilat and Fukasaku, 2007; Walport and Brest, 2011; Mennes et al., 2013; Ferguson et al., 2014; Poldrack and Gorgolewski, 2014). This involves developing the required infrastructure that aligns technical, clinical, and biomedical systems and semantically integrates data from multiple sources, archiving, and making it available to be reused. Such integration is particularly important when creating large datasets from smaller individual studies for use in large-scale image analysis projects, especially for stratified medicine and machine learning which require very large amounts of individualized subject-specific data. In spite of the significant progress made in several neuroimaging domains such as the Biomedical Informatics Research Network (Keator et al., 2008); LORIS (Das et al., 2012), XNAT Central (Marcus et al., 2007); the Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2008); the Human Connectome Project (Van Essen et al., 2013); and the BRAINS project (Job et al., 2016), the problem remains partially solved particularly for neurological diseases such as stroke (Warach et al., 2016).

Stroke researchers have access to imaging and associated data from multiple sources, in many different formats and at different levels of granularity. However, despite stroke being one of the most advanced fields among common neurological diseases in terms of (a) having a standard outcome measure for trials [the modified Rankin Scale (Lees et al., 2012)] and (b) effective treatments and prevention (Lindley et al., 2015), in general, the data collection protocols lack widely used standards, vary considerably, without clearly published provenance information between and within studies, which has significantly impeded the utility of the data (Ferguson et al., 2014; Nichols et al., 2016). Meanwhile, there would be numerous benefits that can be derived from semantically integrated data for various endeavors. Specifically, trials of new treatments for stroke require imaging data as part of the patient assessment (Wintermark et al., 2013), but the sample size needs to be large enough to obtain reliable results, particularly where treatment effects are likely to be modest (Lindley et al., 2015): the ability to combine image as well as clinical data facilitates meta-analyses (Laird et al., 2011). Furthermore, a semantically integrated patient database could be an efficient and cost-effective way to obtain data from many different centers and many different countries in order to obtain the sample size required to be able to observe a statistically significant difference between the subtypes of stroke and other key clinical variables or treatment effects in observational studies or clinical trials (Poldrack and Gorgolewski, 2014). Additionally, an integrated image bank offers the potential for building data analytics models, which

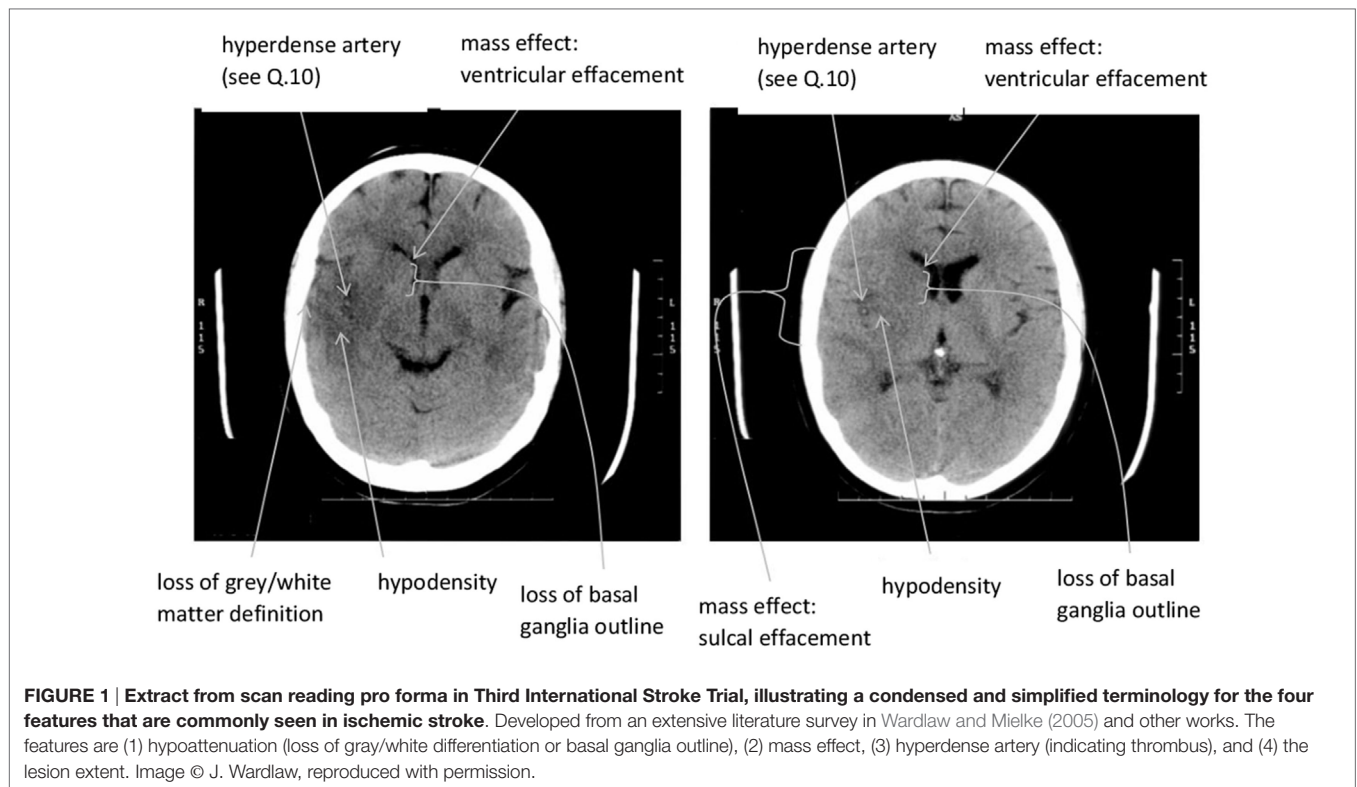
will offer researchers the opportunity to develop new insight and understanding (Gomez-Cabrero et al., 2014).

The paper details our experience on the NeuroGrid Stroke Exemplar (Wardlaw et al., 2007) and further work that was carried out at the Brain Research Imaging Centre (BRIC), University of Edinburgh in collaboration with Stroke Trials Unit, University of Nottingham. The aim of the project was to develop an infrastructure to facilitate linkage, archiving, and reuse of neuroimaging data from stroke patients for large-scale clinical trials, focused observational, mechanistic, and epidemiological studies. We outline the recurring challenges associated with integrating neuroimaging data from multiple sources. We then describe the approach employed to develop an integrated metadata and schema for ischemic and hemorrhagic stroke, as the first step toward integrating neuroimaging data that combines clinical, demographic, and treatment data from patients. We further describe how we developed an integrated schema and database with a web-based interface system, with the aim of being flexible and adaptable to future trials and observational studies. We finally demonstrate the utility of the schema by linking the images and data to prospective central hospital health statistics.

Recurring Issues in Integrating Neuroimaging Data from Multiple Sources

Integrating and sharing imaging and associated data across multiple studies requires shared understanding of the datasets within the domain. Data from patients with common neurological disorders such as stroke are collected increasingly from a growing range of imaging modalities, especially computerized tomography (CT) and magnetic resonance (MR) imaging, and both produce multiple types of images. Images from different sites reflect differences in the scanner manufacturer and models used, and calibrations employed (Warach et al., 2016), even when similar MR sequences are deployed, although frequently MR in stroke still omits key sequences such as T2* weighted or T1 weighted. **Figure 1** shows an example of four early ischemic signs commonly seen in stroke patients imaged soon after stroke, distilled from a large literature survey to represent common features and terminologies (Wardlaw and Mielke, 2005) and which can then be captured efficiently by expert scan readers, e.g., in multicenter clinical trials, providing a simplified shared naming convention for ischemic lesions that allows translation between research and clinical practice.

However, even in an apparently simple process such as plain CT brain scanning (the commonest method used in stroke), there is variability in image and associated clinical data acquisition, transfer and storage that reflects the complexity, and variability in clinical practice as well as those that exist in the structural



representation of the heterogeneous brain data (Keator et al., 2008). These issues have major integration challenges for machines (less so for humans), which can be addressed by metadata schema harmonization to achieve a simplified shared naming convention required in order to be accessible for machines (Keator et al., 2009). “Metadata” are facts about a given dataset that provides additional information regarding the parameters in which the dataset was acquired and the assumptions made about the experiment or analyses that helps one understand and use the data. For example, in the context of medical imaging data, metadata will allow machine-based reference models to be built and embedded into software for rapid determination of the validity of imaging data at the point of image acquisition. This is applicable to all data acquisition where imaging has a key role.

Progress toward Integrating Neuroimaging Data for Stroke Image Bank

Attempts are being made toward developing infrastructures to facilitate sharing and reuse of neuroimaging data from heterogeneous sources. To the best of our knowledge, **Table 1** shows all image banks specifically developed for stroke. We examine each briefly to determine their relevance and scope for stroke clinical trials.

The descriptions provided in **Table 1** demonstrate the scope and limitations of the existing stroke image banks, with respect to facilitating clinical trials of new treatments for stroke, which was the focus of the NeuroGrid project (Geddes et al., 2005; Wardlaw et al., 2007). NeuroGrid focused on two exemplar large multicenter clinical stroke trials that were ongoing at the time,

the Third International Stroke Trial (IST-3) (Sandercock et al., 2012) and the Efficacy of Nitric Oxide in Stroke (ENOS) trial (The ENOS Trial Investigators, 2015). In order to create an integrated searchable database that could ultimately house the image data of both trials for future meta-analyses and data sharing to which other trials could be added, we had to design purpose-specific stroke imaging metadata and a related schema to accommodate different data structures and purposes, including, in addition to the actual images, collection of data on initial clinical assessments across several domains, long-term outcomes, treatments, and radiological interpretations of the images, which would be sufficiently flexible and adaptable for use in any future clinical trial or observational study in ischemic or hemorrhagic stroke (Wardlaw et al., 2007).

MATERIALS AND METHODS

The concepts and methods described here arose from NeuroGrid, followed by our work in developing an image bank of normal subjects across the lifespan in the BRAINS project¹ and also described in Job et al. (2016). The BRAINS project was carried out in parallel with adapting the stroke data schema to accommodate all data acquired in a series of 12 observational mechanistic and diagnostic studies in patients with various subtypes of stroke acquired in one center between 1996 and 2013 (but to which subsequent studies are being added).

¹<http://www.brainsimagebank.ac.uk>.

TABLE 1 | Stroke image banks.

Reference	Stroke image bank project	Scope
Hanser et al. (2007)	neurlST	Focuses on very specific terminologies for describing vascular abnormality, clinical features, treatments and outcomes for subarachnoid hemorrhage
Colombo et al. (2010)	NeuroWeb	Focuses on genetics means that there is less priority given to recording image data in the detail required for many acute stroke treatment trials or other types of stroke research where highly specialized phenotyping including detailed imaging is required
Gibaud et al. (2011)	NeuroLOG	Focuses on the neuropsychological aspects of stroke and computational image analysis and does not provide for documenting more clinically relevant acute treatment and outcomes
Wang et al. (2011)	Medical Image Management System	This is particularly useful for managing imaging data in clinical trials but neither relevant to stroke specifically nor to observational studies with heterogeneous data
Ali et al. (2012)	Virtual International Stroke Trials Archive	Focuses on clinical stroke research for prevention, rehabilitation, imaging, and intracerebral hemorrhage. However, data are limited to demographic and clinical data from baseline and follow-up visits (2 h–90 days)
Wintermark et al. (2013)	Stroke Imaging Repository	Focuses on terminology and standardization for acute ischemic stroke trials but not metadata schema required for integrating heterogeneous imaging data (initiated with early terminology from NeuroGrid Stroke exemplar, an early version of the Stroke Schema in the present paper)
Kim et al. (2014)	CRCS-5	Focuses on ischemic stroke monitoring and management in hospitals. Also, although data are collected from multiple centers, it does not require metadata schema for integration as it uses a single data management with web-based interface system
Seghier et al. (2016)	PLORAS	Data are not heterogeneous and also focuses on only speech and language abilities-related outcomes of stroke

Our Approach

Image bank development begins with data integration. Data integration approaches could be broadly grouped into two. The “centralized approach” is where data sources are accessed through a single access point based on a predefined common metadata schema (Keator et al., 2009). The alternative is the “federation-based approach,” which requires a framework in order to present a unified view of the data from multiple sources (Wiederhold, 1992). Our framework is, of necessity, federation-based, based on semantic rules derived from expert knowledge underpinned by many years of professional experience in stroke research including in clinical trials. **Figure 2** shows the schematic diagram of the framework, which we subsequently describe in detail.

Step1: Examination of Datasets from Past Projects and NeuroGrid Stroke Example Metadata

As a first step toward developing an integrated schema, we started with the NeuroGrid schema based on the two large multicentre international trials, ENOS and IST-3, and examined data from 12 past stroke imaging research projects with various different objectives including different stroke subtypes and types of imaging, carried out over the past two decades in our center. These projects varied in research objectives and data collection protocols. This is demonstrated with two examples.

First, the Salvageable Tissue study (Wardlaw et al., 2013) was a multicenter study carried out in three acute stroke centers in Scotland (Aberdeen, Glasgow, and Edinburgh) between 2008 and 2010. The objective was to assess the practicalities of performing acute stroke imaging with CT and MR including perfusion

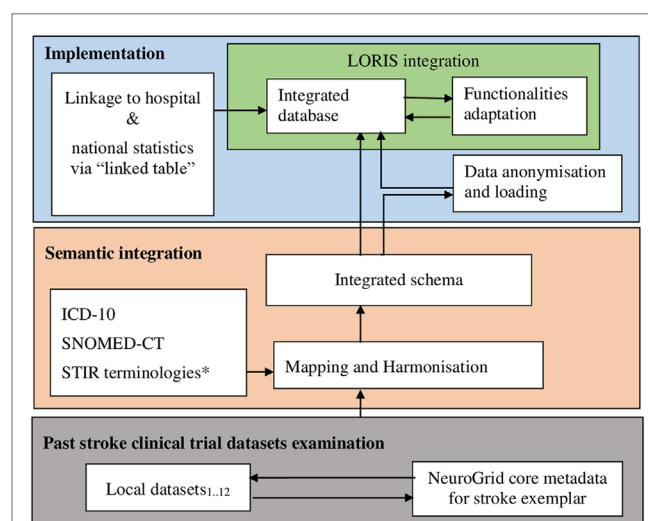


FIGURE 2 | Schematic diagram of the framework for the stroke image bank. LORIS, Longitudinal Online Research and Imaging System; ICD-10, the World Health Organization’s International Classification of Diseases coding version 10; SNOMED-CT, a systematized nomenclature of medicine—clinical terms; STIR, Stroke Imaging Repository coding standards. *Initiated with terminology from NeuroGrid stroke exemplar, i.e., an early version of the present schema.

imaging, to assess the proportion of patients with perfusion-evidence of salvageable tissue [perfusion-diffusion mismatch on MRI or reduced flow on CT perfusion (CTP)], and markers of

subsequent lesion growth on follow-up imaging to provide sample size estimates for future treatment trials. This involved recruiting patients with moderate to severe cortical ischemic stroke in three centers, performing imaging [diffusion weighted imaging (DWI), perfusion-weighted imaging, fluid attenuation inversion recovery (FLAIR), gradient echo (GRE/T2*), MR angiography (MRA); or with CT, CTP, and CT angiography (CTA)] within 6 h of stroke, repeated at 2–5 days (mostly MR) and 1 month (MR T2, GRE, DWI, and MRA). A final clinical follow-up was performed at 3 months.

The second is the Mild Stroke Study (Wardlaw et al., 2009) performed between 2005 and 2009. The aim was to investigate causes of lacunar stroke and associations with retinal vascular appearances (as a surrogate for cerebral small vessels). This was to test the theory that lacunar stroke and small vessel disease arise through blood–brain barrier damage. It recruited patients with lacunar or minor cortical ischemic stroke, all of whom had diagnostic MR imaging with DWI, FLAIR, T2-weighted, GRE, T1-weighted, and (in a subset) blood–brain barrier permeability imaging. A subset was followed up clinically and had follow-up imaging at 3 years after stroke.

The stroke exemplar metadata designed originally in the NeuroGrid project was an extension to the NeuroGrid core metadata and was designed to be scalable and modifiable to suit other stroke studies using imaging. The NeuroGrid core metadata was constructed to accommodate studies in stroke, dementia, and psychosis and was in response to one of the key infrastructure objectives of NeuroGrid—to develop management systems to allow large “living archives” of images linked to key metadata for diseases that require long-term study to understand their true natural history and the effects of treatment (Wardlaw et al., 2007). This involved developing a simple repository browser to perform *ad hoc* searches against the core metadata and display user-readable, navigable listings of search results including the images for administration and quality control. An example of a search could be to generate a list of all patients in trial X who were scanned at location Y and had a clinical feature Z and an imaging feature A.

In the stroke exemplar, the NeuroGrid core metadata schema was extended significantly based on the two large multicentre randomized stroke trials, IST-3 and ENOS. IST-3 was a 3035-patient multicenter randomized controlled trial of alteplase given up to 6 h after onset of acute ischemic stroke (Sandercock et al., 2008, 2012). IST-3 sought to determine whether a wider range of patients might benefit from intravenous recombinant tissue plasminogen activator (rt-PA). ENOS (The ENOS Trial Investigators, 2006, 2015) was a 4011-patient multicentre randomized controlled trial in patients with acute (<48 h of onset) ischemic or hemorrhagic stroke. ENOS tested the safety and efficacy of transdermal GTN, and of continuing or stopping temporarily prior antihypertensive medication. Both the trials required a CT brain scan at randomization (minimum requirement plain non-contrast CT brain), but MRI could be used instead (minimum sequences T2-weighted, FLAIR, DWI, and GRE). Advanced imaging, such as CTA, MRA, or perfusion imaging, was also collected where performed. Both the trials involved multiple centers ($n = 329$), and therefore,

inevitably the images came from a very large variety of scanners (Wardlaw et al., 2007).

The extension of the core metadata schema was governed by issues relating to where, when, and how datasets are collected, published to the database, or required by clinicians. Thus, the resulting extended NeuroGrid core metadata for stroke allowed a search across a wide range of patient baseline characteristics (including history factors: vascular risk factors, prior treatments, past medical history), stroke clinical characteristics (severity, clinical subtype, neurological examination details), type and timing of imaging, appearance of the stroke lesion on imaging (including site and size), laboratory test results, details of trial treatment administration, details of any non-trial treatments, subacute and late clinical functional measures (symptomatic intracranial hemorrhage or brain swelling, modified Rankin Scale, death), cognitive and imaging outcomes, and adverse events.

We then compared our 12 study datasets from our center with the NeuroGrid stroke exemplar metadata. We noted the differences and overlaps that existed and iterated modifications to address items that were not covered in the original NeuroGrid exemplar or that were present but required more granularity and fed this into the subsequent developments of the data schema. We demonstrate this with some examples of the differences that were observed in data collection protocols between the Salvageable Tissue and Mild Stroke Studies described earlier. For example, the NeuroGrid exemplar schema required information about stroke severity using the National Institute of Health Stroke Scale (NIHSS) (Goldstein et al., 1989). While the Salvageable Tissue protocol required a detailed data to be recorded for each symptom (e.g., “Bast gaze,” which is one of the items on the NIHSS is recorded as either “forced deviation” or “Normal” or “Partial gaze palsy”), the Mild Stroke Study protocol, on the other hand, required summary data, which is the total score assigned to each NIHSS symptom to be recorded. The reverse of this was observed in another instance. The NeuroGrid exemplar schema required data on classification of stroke based on the Oxford Community Stroke Project classification—OCSP (Bamford et al., 1987). In this instance, The Salvageable Tissue protocol required a summary of the data by recording either “present” or “not present” for each of the classifications [e.g., Partial Anterior Circulation Syndrome (PACS) is to be recorded as either “present” or “not present”] based on the assessment and knowledge of the clinician. On the other hand, the Mild Stroke Study protocol did not rely on the knowledge of the clinician to classify but only required data to be collected on symptoms such as weakness/sensory deficit in arm, leg, and face. The differences in data as result of differences in collection protocols demand some amount of adaptation from data integration and image bank perspective, which is subsequently described in step 2. The guiding principles adopted in this work were that the approach must be pragmatic; the metadata and schema should be relevant to clinical practice, as well as scalable to other researches where details might need to be added or switched off in particular domains, without requiring major redesign.

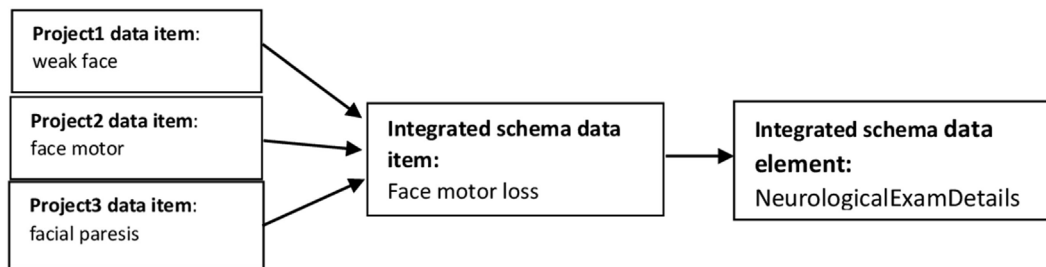


FIGURE 3 | Mapping “face motor loss” variables as expressed in various datasets.

Step 2: Semantic Integration

“Semantic integration” is the process of ensuring that all semantically related data elements and items are grouped together based on expert knowledge of domains and other resources. This was achieved through a series of steps described below.

Mapping and Harmonization

Mapping ensures that data items that have different names, but that are considered to be semantically the same or very similar, are captured as a single schema data item. This involved mapping the IST-3 and ENOS trials metadata and schema developed in NeuroGrid, then refining, and extending the schema based on the process described in step 1 above. Examination of the 12 local prior stroke research projects showed a high degree of variability in the datasets (from the machine point of view though not the human point of view), which is noted to be a common issue associated with data from multiple sources (Gomez-Cabrero et al., 2014), or in this case, even from a series of studies of one disease in one center that basically collected the same clinical variables even though each study might collect some other information. **Figure 3** illustrates an example of the variabilities and how these are handled.

For example, **Figure 3** shows three different variables (“weak face,” “face motor,” and “facial paresis”) in three different projects being mapped to a single search item “face motor loss,” which is part of the integrated schema data element, “NeurologicalExamDetails.” On the other hand, harmonization is a process that ensures uniformity in how schema search items are encoded and represented. For example, “lesion age” in one dataset is encoded in categories (1 = “less than 6 h”; 2 = “6–12 h”; 3 = “greater than 12 h”), whereas in another dataset, different encoding scheme (e.g., raw values) are employed. Specifically, with regards to the examples of the problems between the Salvageable Tissue study and Mild stroke dataset described in step 1 above, the data on the individual symptoms were mapped to the corresponding numeric values for each symptom based on the NIHSS documentation (Goldstein et al., 1989). This enabled us to transform the responses into a total score representing the severity of stroke for each patient as required by our new metadata schema. Again, to be able to harmonize the OSCP data, rules were developed to transform the symptoms collected by the Mild stroke study based on the OSCP classification rules. So for example, if a patient had weakness and/or sensory problems in

the face, arm, or leg and also has dysphasia, the stroke is classified as PACS being “present,” otherwise “not present.” Thus, reasonable encoding and representation were achieved through harmonization. This strategy was applied to all issues that were identified and documented as part of the provenance, which is also made available to potential users of the image bank. This process was automated using the Python programming language (version 3.2, see Python Software Foundation²).

Use of Coding Standards

In order to further enhance the interoperability and reusability of the integrated schema and image bank to facilitate future integration with other biomedical ontologies, we cross compared our terms with other data coding standards and medical taxonomies. This included standard terminologies that were originally derived from the NeuroGrid work with additional modification for use in the Stroke Imaging Repository of acute treatment and secondary prevention stroke trials (Wintermark et al., 2013), which also aligns with the National Institute of Neurological Disorders and Stroke Common Data Elements.³ The World Health Organization’s International Classification of Diseases coding version 10⁴ and the systematized nomenclature of medicine—clinical terms (SNOMED-CT) (Cote and Robboy, 1980) provide a familiar and useful common vocabulary in clinical practice where other relevant data may be cross-referenced. ICD-10 and SNOMED-CT, in particular, are implemented as standards by health services in many countries hosting multi-site trials and has the additional benefit that allows integration with national health information systems and electronic health records (Westra et al., 2015).

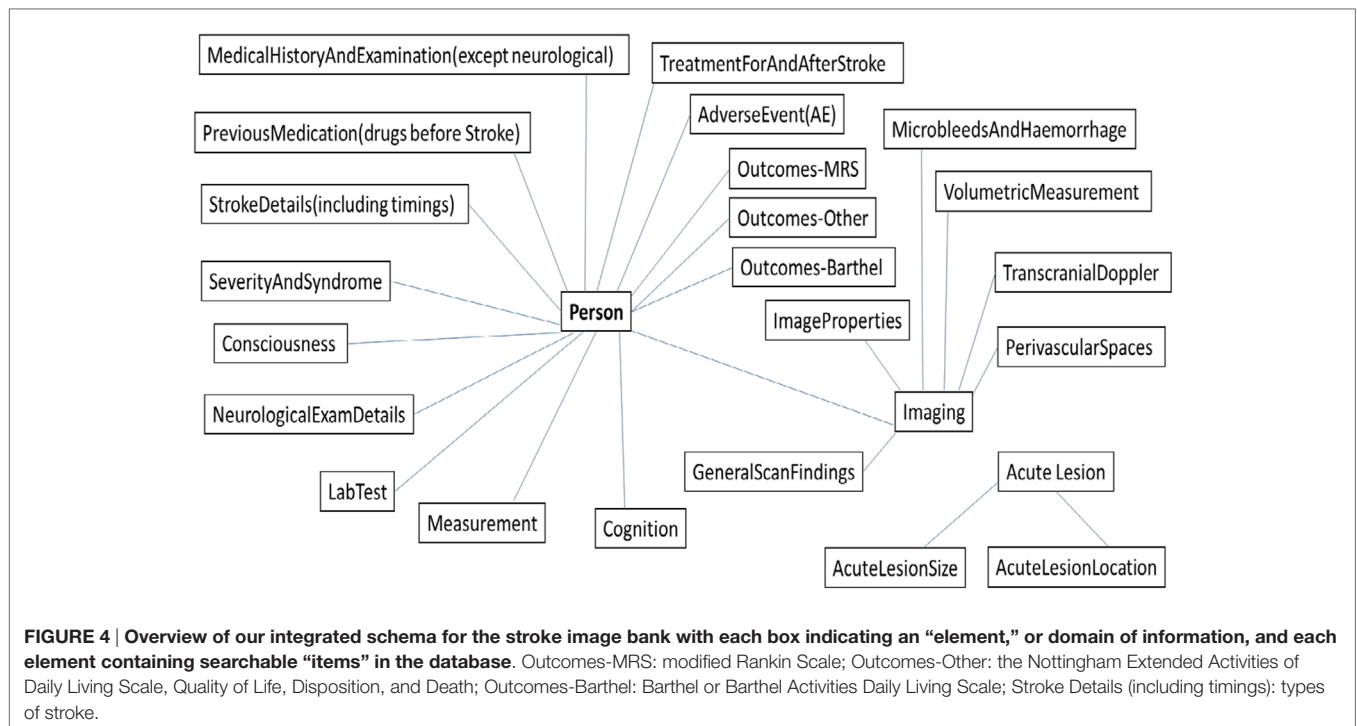
Figure 4 shows schematic diagram of the integrated metadata schema with its data elements, which have over 550 integrated searchable data items contained within them.

As demonstrated in **Figure 4**, the resulting integrated schema will allow searches across a wide range of patient baseline and outcome characteristics described as part of the stroke exemplar and additional searchable data elements and items including read-by-an-expert, visual scores, and computationally measured imaging features. This includes categorization of the acute stroke

²<https://www.python.org/>.

³http://www.ninds.nih.gov/research/clinical_research/toolkit/common_data_elements.htm.

⁴<http://apps.who.int/classifications/icd10/browse/2016/en>.



lesion (infarct or hemorrhage, extent, background brain changes); volumetric measurements (e.g., intracranial volume, brain volume, infarct volume, white matter hyperintensity volume); other visual scores as relevant to, for example, small vessel stroke (e.g., perivascular spaces, lacunes, microbleeds by brain region); and lesion-specific anatomical locations (e.g., thalamus, gray white matter, deep white matter) where relevant.

Step 3: Implementation

Our implementation took advantage of available open source technologies as described below.

Longitudinal Online Research and Imaging System (LORIS) Integration

We integrated our integrated schema with the Longitudinal Online Research and Imaging System (LORIS) database in order to take advantage of its capabilities. LORIS is an open-source data management system, well engineered for managing imaging and associated behavioral longitudinal data, and implemented using MySQL and NoSQL (CouchDB)⁵ for back-end web interface and Hypertext Preprocessor (PHP) programming language⁶ for front-end web interface (Das et al., 2012), which we deployed in Linux Ubuntu 14.04 box.

Our clinical trial datasets also have longitudinal characteristics as projects required subjects to be followed up after the initial visit, sometimes over many years. Therefore, it was prudent to take advantage of the functionalities available in LORIS in order to avoid duplication of effort. MySQL, NoSQL and PHP are both

open source and widely used relational database management systems and frameworks (Bakken et al., 1997; Bretthauer, 2002). Both MySQL and NoSQL as employed in LORIS offered us the following database design capabilities: (a) performance, which was to ensure speed processing of queries and a quick access to the data; (b) integrity, which was to ensure accurate storage of the data as obtained from the original sources; (c) comprehensibility, which was concerned with ensuring coherence in the structure of the database as presented to users; and (d) extensibility, which was to ensure the database can be extended without the need to redesign. The functionalities adaptation process involved integrating our Python-based scripts with the PHP-based script functionalities used in LORIS. The integration process was achieved through collaboration and support from the LORIS software development team.⁷

Data Anonymization and Loading

All images had already been anonymized of metadata by passing through DICOM Confidential (González et al., 2010), a freely available data anonymization tool for imaging.⁸ It is a Java-based de-identification toolkit that enforces confidentiality policies as defined by the Medical Research Council.⁹ It is also specifically designed to support batch processing for multicentre clinical trials. Additionally, all identifiable information contained in the columns of the associated clinical data was also removed to ensure complete anonymity. After the data anonymization process, we

⁷<http://loris.ca/>.

⁸<https://sourceforge.net/projects/privacyguard/>.

⁹<https://www.mrc.ac.uk/documents/pdf/personal-information-in-medical-research>.

⁵<http://couchdb.apache.org/>.

⁶<http://php.net/manual/en/intro-what-is.php>.

Stage	Status	Date
Screening		
Visit	Pass	2010-05-19
Approval	Pass	

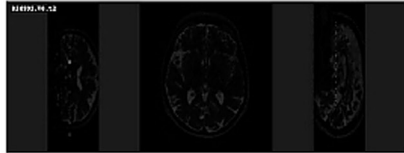
Details of Imaging Performed

Date of Administration	2010May19
Candidate Age (Months)	740
Window Difference (+/- Days)	N/A
Examiner	
Date MRI was done	2010-05-20
MRI Diffusion weighted imaging	yes
MRI T2 imaging	yes
MRI T1 imaging	yes
MRI FLAIR imaging	yes
MRI Susceptibility weighted imaging	yes
MRI Gradient Echo imaging	no
MRI Spectroscopy imaging	no
MRI Spectroscopy imaging type	not_answered
MRI Chemical shift imaging	no
Other MRI imaging	not_answered
Date CT was done	not_answered
CT Plain	not_answered
CT Perfusion imaging	not_answered
Date Transcranial Doppler Ultrasound (TCD) was done	not_answered

QC Status	PSCID	DCCID	Visit Label	Site	QC Pending	DOB	Gender	Output Type	Scanner	Subproject
	MSSB009	836993	V0	Edinburgh Imaging (BRIC)		1935-03-17	Female	native	GE MEDICAL SYSTEMS Signa HDxt 00000000200MRS03	MSS2

2 file(s) displayed.

☐ loris_836993_V0_I2_001.mnc



QC Status (★ New)

Selected

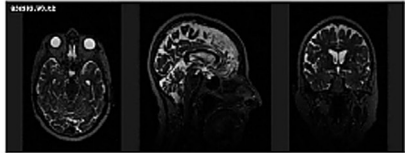
Caveat

False

QC Comments

Download MINC

☐ loris_836993_V0_I2_002.mnc



QC Status (★ New)

Selected

Caveat

False

QC Comments

Download MINC

FIGURE 5 | A LORIS-based web interface of the stroke image bank showing details of an anonymized patient's imaging and clinical data as contained in the integrated database at baseline visit (visit 0).

then loaded the data by populating the integrated database with data from the clinical trial datasets described in step 1 above. The loading process also accounts for the mapping and harmonization

process that was carried out to ensure that the correct data items were populated to conform to our new integrated schema. This process was also automated using Python-based scripts.

Linkage to Hospital and National Statistics

We made provision for linking the integrated imaging database to hospital and national statistics to obtain long-term outcomes such as recurrent stroke, dementia, other vascular events, and death. We first obtained regulatory approvals from the relevant institutions. This include Caldicott Guardian and Community Health Index Advisory Board, NHS Lothian (reference: CG/DF/1559); NHS Lothian Research & Development (reference: 2015/0296); Information Services Division (ISD) and Scottish Stroke Care Audit (reference eDRIS-1516-0337); and West of Scotland Research Ethics Service (reference: 15/WS/0157). This allowed us to create a database of identifiable details of subjects scanned at our center in Edinburgh for the purpose of central matching with routinely collected health data by the Information Services Division of NHS Scotland.¹⁰ In order to achieve the linkage between our integrated database and hospital and national statistics database, a “linked table” was created which holds the patients’ hospital primary IDs and randomly generated IDs assigned to subjects in the integrated database by LORIS-based ID generation algorithm. Access to the linked table is restricted and only accessible to key approved members of research team covered by the data access agreements. The data anonymization and loading step described above also populated the integrated database with the individual “key” stored in the linked table.

Quality Control

In order to ensure data accuracy and consistency, an end-to-end quality control procedure was performed on samples of the data. This involved randomly selecting sample records from the web interface and checking data values against the source as well as data provenance.

RESULTS

Our integrated schema contains over 550 searchable data variables. Additionally, the integrated schema maps to IST3¹¹ and ENOS,¹² which are the two original NeuroGrid exemplar large multicentre stroke trials with over 7,000 patients from 30 countries between them. This demonstrates its utility within the context of ensuring data standards to facilitate seamless integration of heterogeneous multicentre neuroimaging data for ischemic and hemorrhagic stroke as well as stroke subtypes such as small vessel lacunar stroke. Moreover, our integrated database contains over 3,079 unique subjects from our 12 research studies, who were scanned in our local BRIC, Edinburgh, with neuroimaging data for ischemic and hemorrhagic stroke and small vessel disease studies. **Figure 5** shows the LORIS-based interface of our integrated database.

We submitted records on 3,245 patients from the combined dataset of 12 stroke studies in our 1 center for central linkage with routinely collected health records achieving an overall linkage success rate of 95% with the National Health Service (NHS)

Hospital Information System and Stroke Audit databases of Scotland. A detailed breakdown showed that up to 19 years since inclusion in the research project and scanning (median = 9.04; IQR = 12.17, range 0–19 years) of follow-up, 879/3079 patients had died, 525 had had one or more recurrent stroke, and 291 had developed dementia, which further demonstrates the utility of our integrated database. The metadata schema for the integrated database and provenance information including data dictionary are available online under Apache 2.0 and CC-YB 4.0 licenses, respectively.¹³

DISCUSSION

Our neuroimaging data acquisition and management for stroke research has evolved from large pragmatic clinical stroke trials of acute stroke treatments with fairly basic imaging in NeuroGrid in the mid-2000s to include much more detailed bespoke observational mechanistic studies with much more complex imaging and longer follow-up linked with more detailed outcomes. This evolution demanded new approaches and also presents new opportunities. With the advent of “big data” science for medical and clinical research (Wang and Krishnan, 2014) and also for neuroimaging (Van Horn and Toga, 2014), our image bank will provide stroke researchers with new opportunities to explore big data science for stroke. An image bank with special focus on ischemic and hemorrhagic stroke and subtypes such as small vessel disease adds substantially to the dynamic range of capabilities of secondary research with cerebrovascular diseases data, thereby contributing to the volume and veracity of stroke data which characterize big data (Laney, 2001). Furthermore, employing international data standards facilitates the creation of Linked Data (Heath and Bizer, 2011), thus expanding the data space useful for new data management and technological initiatives for stroke. Also, the provision made in our integrated database to allow data from hospital information systems and national statistics to be linked provides opportunities to investigate a range of clinically highly relevant issues in stroke and to make use of centrally housed routinely collected image data in National Picture Archiving and Communication Systems PACS, such as the many thousands of brain scans collected in the first 8 years of the Scottish National PACS, now stored at the Farr Institute, Edinburgh.¹⁴ To demonstrate this potential, for example, we are currently using imaging data from our 12 stroke studies linked to data from NHS Scotland’s Information System and Stroke Audit databases to investigate imaging predictors of neurodegeneration measured at presentation with suspected stroke and subsequent adverse outcomes of recurrent stroke, dementia, or death.

From image analysis perspective, well-characterized images with detailed metadata are increasingly needed for studies that typically need larger samples or more variety of cases than are available in individual studies—these include studies to develop machine learning methods for image analysis, in stratified medicine, and large studies of genetics, e.g., genome wide association

¹⁰<http://www.isdscotland.org/>.

¹¹<http://www.dcn.ed.ac.uk/ist3/>.

¹²<http://www.strokecenter.org/trials/clinicalstudies/the-efficacy-of-nitric-oxide-in-stroke-enos-trial>.

¹³<https://sourceforge.net/projects/cvd-db.brainsimagebank.p/>.

¹⁴<http://www.farrinstitute.org/>.

studies where typically many thousands of cases are needed (Hernández et al., 2013; Caligiuri et al., 2015). The availability of large amount of data could help develop models that can be generalizable based on the patterns the underlying algorithms are able to “learn” from the data. Large amounts of data can also provide enough statistical power for valid conclusions to be drawn (Cooper et al., 2011). This could be achieved by having access to selected cases with particular characteristics that are pulled from multiple studies for testing these algorithms and hypothesis. For example, Maillard et al. (2008) demonstrated the usefulness of image bank when they pulled over 1,100 of elderly subjects (with similar characteristics) from two large MRI studies to evaluate the performance of an automated method for detection, quantification, localization, and statistical mapping of white matter hyperintensities in T2-weighted images. An integrated image bank such as this will afford researchers the opportunity to carry out similar studies.

The framework that we employed offers an alternative to other frameworks proposed in the literature. The ontology-based federation is the most common approach within the neuroimaging domain (Hanser et al., 2007; Colombo et al., 2010; Gibaud et al., 2011). These approaches tend to rely on some specialized ontology to serve as a mediation layer between databases to integrate heterogeneous neuroimaging datasets (Wiederhold, 1992) and require that all potential submitters of data to the database stick religiously to the described schema terminology, which in reality is difficult across multiple sites. Within the context of stroke, the neurIST Project employed description logic-based ontology to represent concepts that are associated with cerebral aneurysms and subarachnoid bleedings (Hanser et al., 2007). Similarly, an ontology-based approach was also employed in the NeuroLOG (Gibaud et al., 2011) as well as NeuroWeb (Colombo et al., 2010) projects. A hybrid approach has also been proposed by Keator et al. (2013), where an ontology-based resource, NeuroLex (Larson and Martone, 2013), is combined with information obtained from other resources such as the Human Imaging Database¹⁵ and XNAT.¹⁶ None of these were suitable for stroke, thereby suggesting that lack of ontology for a given specialized domain raises significant neuroimaging data integration challenges (Smith et al., 2015). Furthermore, it has been noted that ontology-based approaches result in tensions between logical (research) and clinical representations of a domain, which make it difficult to create shared models resulting in tensions between ontological consistency and clinical usability (Bodenreider, 2004; Bodenreider and Stevens, 2006; Rector and Rogers, 2006). Thus, our approach is an important advance that overcomes the lack of a specialized ontology for ischemic and hemorrhagic stroke.

Moreover, there is an implicit expectation that medical concepts of disease, based on signs and symptoms, can be transposed as formally defined classes and relations, which are often much more complex to model in practice and resistant to simplification. Thus, the pragmatic and simplified approach adopted here makes our framework and data integration approach easy to implement. However, it is important to note that this is heavily dependent for

its development on domain knowledge. In our case, the domain experts lead the project and were motivated to combine their datasets from individual studies, thus providing the required domain and semantic knowledge. Such exercises are not achievable without the close working of experts in the disease of interest (and in this case its imaging) with experts in the technological infrastructure required to host complex interrelated medical and imaging data, the former having the motivation and the content knowledge and the latter the essential knowledge to manage the data efficiently.

The mapping and harmonization process described as part of our framework involved data provenance documentation of the integrated schema.¹⁷ This provides a detailed account of processes carried out on the datasets from the point of acquisition, descriptions of the imaging hardware and parameters used in the acquisition of the data, as well as mapping and harmonization (including transformations) as previously described (MacKenzie-Graham et al., 2008). The importance of this information has been emphasized (Keator et al., 2013) and documented as one of the guiding principles of data sharing best practices (Nichols et al., 2016).

CONCLUSION

This paper summarizes our experience in developing an integrated image bank and schema suitable for hosting data from multiple individual stroke imaging research projects and enabling large-scale research in cerebrovascular diseases, with a particular focus on ischemic and hemorrhagic stroke and small vessel diseases. This will facilitate research into new treatments for stroke by enabling large meta-analysis as well as testing computationally based image analysis methods (e.g., machine learning) for building predictive models specifically for stroke and other related conditions. In addition to adding more research data, we open the door to adding new data such as that routinely collected in health services, for example, by using Natural Language Processing (Chapman et al., 2011). Additionally, the past decade has seen unprecedented attempts to develop frameworks and infrastructure that can facilitate integration, archiving, and reuse of neuroimaging from multiple sources. We believe that the experience and framework described in this manuscript could be applied to neuroimaging data from other domains where resources such as ontologies do not currently exist.

AUTHOR CONTRIBUTIONS

JW, DJ, JP, JU, PB, and PS designed and carried out the NeuroGrid Stroke Exemplar project. JW, SD, DJ, DG, and DD created the integrated stroke metadata schema and image bank. All the authors contributed to the drafting of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank a wide range of colleagues in NeuroGrid and other HealthGrid projects who have

¹⁵<http://www.nitrc.org/projects/hid/>.

¹⁶<http://www.xnat.org/>.

¹⁷<https://sourceforge.net/projects/cvd-db.brainsimagebank.p/>.

contributed to the work described in this paper. Special thanks go to Andrew Duffy, the Farr Institute, and the NHS Scotland for extracting data for the image data bank. Finally, the authors are also thankful to Christine Rogers and the LORIS team at the McGill Centre for Integrative Neuroscience for their support in integrating LORIS with our federated database.

FUNDING

The Medical Research Council (Grant Ref no: GO600623 ID number 77729) for the NeuroGrid project. The SINAPSE Collaboration (Scottish Imaging Network, A Platform for Scientific Excellence, www.sinapse.ac.uk) through the Scottish Funding Council part funded JW. PB is Stroke Association Professor of Stroke Medicine and a NIHR Senior Investigator. This work was further supported by INNOVATE-UK (reference 102167) for the vascular linkage project. Also, as part of this work, SD received additional funding from Scottish Funding Council through the SINAPSE Postdoctoral and Early Career Researcher

REFERENCES

- Ali, M., Bath, P., Brady, M., Davis, S., Diener, H. C., Donnan, G., et al. (2012). Development, expansion, and use of a stroke clinical trials resource for novel exploratory analyses. *Int. J. Stroke* 7, 133–138. doi:10.1111/j.1747-4949.2011.00735.x
- Bakken, S. S., Aulbach, A., Schmid, E., Winstead, J., Wilson, L. T., Lerdorf, R., et al. (1997). *PHP Manual*. Zend Technologies, Ltd. Available at: [ftp://ftp.nymh.edu/doc/php/manual_m-x.pdf](http://ftp.nymh.edu/doc/php/manual_m-x.pdf)
- Bamford, J., Sandercock, P., Jones, L., and Warlow, C. (1987). The natural history of lacunar infarction: the Oxfordshire Community Stroke Project. *Stroke* 18, 545–551. doi:10.1161/01.STR.18.3.545
- Bodenreider, O. (2004). “The ontology-epistemology divide: a case study in medical terminology,” in *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)* (Torino: IOS Press).
- Bodenreider, O., and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Brief. Bioinformatics* 7, 256–274. doi:10.1093/bib/bbl027
- Bretthauer, D. (2002). Open source software: a history. *Inform. Technol. Libr.* 21, 3–10.
- Caligiuri, M. E., Perrotta, P., Augimeri, A., Rocca, F., Quattrone, A., and Cherubini, A. (2015). Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: a review. *Neuroinformatics* 13, 261–276. doi:10.1007/s12021-015-9260-y
- Chapman, W. W., Nadkarni, P. M., Hirschman, L., D’Avolio, L. W., Savova, G. K., and Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J. Am. Med. Inform. Assoc.* 18, 540–543. doi:10.1136/amiajnl-2011-000465
- Colombo, G., Merico, D., Boncoraglio, G., De Paoli, F., Ellul, J., Frisoni, G., et al. (2010). An ontological modeling approach to cerebrovascular disease studies: the NEUROWEB case. *J. Biomed. Inform.* 43, 469–484. doi:10.1016/j.jbi.2009.12.005
- Cooper, R., Hardy, R., Sayer, A. A., Ben-Shlomo, Y., Birnie, K., Cooper, C., et al. (2011). Age and gender differences in physical capability levels from mid-life onwards: the harmonisation and meta-analysis of data from eight UK cohort studies. *PLoS ONE* 6:e27899. doi:10.1371/journal.pone.0027899
- Cote, R. A., and Robboy, S. (1980). Progress in medical information management: systematized nomenclature of medicine (SNOMED). *Jama* 243, 756–762. doi:10.1001/jama.1980.03300340032015
- Das, S., Zijdenbos, A. P., Vins, D., Harlap, J., and Evans, A. C. (2012). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinformatics* 5:37. doi:10.3389/fninf.2011.00037
- Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: data-sharing in the ‘long tail’ of neuroscience. *Nat. Neurosci.* 17, 1442–1447. doi:10.1038/nn.3838
- Exchange programme, which enabled him to visit Harvard Medical School, USA. ENOS was funded by the Bupa Foundation and Medical Research Council. The IST-3 trial was funded by the following agencies: the Australian Heart Foundation (Australian, grant number G 04S 1638); Australian NHMRC (grant number 457343); Danube University (Austria); the Dalhousie University Internal Medicine Research Fund (Canada); Norwegian Research Council (Norway); Polish Ministry of Science and Education (Poland, grant number 2PO5B10928); AFA Insurances (Sweden), the Swedish Heart Lung Fund (Sweden), Karolinska Institutet (Sweden), Stockholm County Council and Karolinska Institute Joint ALF-project grants (Sweden); Swiss National Science Foundation (Switzerland); Swiss Heart Foundation (Switzerland); The Foundation of Marianne and Marcus Wallenberg (Sweden); Foundation for health and cardio-/neurovascular research (Switzerland); the Medical Research Council (UK, grant numbers G0400069 and EME 09-800-15), The Health Foundation (UK), The Stroke Association (UK); DeSACC (UK); The University of Edinburgh (UK); The Lothian Health Board (UK); The Assessorato alla Sanita (Italy).
- Geddes, J., Lloyd, S., Simpson, A., Rossor, M., Fox, N., Hill, D., et al. (2005). “NeuroGrid: collaborative neuroscience via grid computing,” in *Proceedings of All Hands Meeting*. Available at: https://www.researchgate.net/profile/Stephen_Lawrie/publication/268202484_NeuroGrid_Collaborative_Neuroscience_via_Grid_Computing/links/5469cf06cf20dedafid10822.pdf
- Gibaud, B., Kassel, G., Dojat, M., Batrancourt, B., Michel, F., Gaignard, A., et al. (2011). “NeuroLOG: sharing neuroimaging data using an ontology-based federated approach,” in *Proceedings of American Medical Informatics Association, October 2011* (Washington, DC). Available at: <https://hal.archives-ouvertes.fr/hal-00683087>
- Goldstein, L. B., Bertels, C., and Davis, J. N. (1989). Interrater reliability of the NIH stroke scale. *Arch. Neurol.* 46, 660–662. doi:10.1001/archneur.1989.00520420080026
- Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merckenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8:11. doi:10.1186/1752-0509-8-S2-11
- González, D., Carpenter, T., van Hemert, J. I., and Wardlaw, J. (2010). An open source toolkit for medical imaging de-identification. *Eur. Radiol.* 20, 1896–1904. doi:10.1007/s00330-010-1745-3
- Hanser, S., Boeker, M., Kumpf, K., and Schulz, S. (2007). “Design of an ontology on cerebral aneurysms: representing the conceptual space of the @neurIST project. Medinfo 2007,” in *Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems* (Amsterdam: IOS Press).
- Heath, T., and Bizer, C. (2011). “Linked data: evolving the web into a global data space,” in *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 1, 1–136. Available at: http://seco.cs.aalto.fi/u/jwtuomin/svn/secoweb/public_html/publications/2012/hyvonon-ch-book-2012.pdf
- Hernández, M., Piper, R. J., Wang, X., Deary, I. J., and Wardlaw, J. M. (2013). Towards the automatic computational assessment of enlarged perivascular spaces on brain magnetic resonance images: a systematic review. *J. Magn. Reson. Imag.* 38, 774–785. doi:10.1002/jmri.24047
- Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imag.* 27, 685–691. doi:10.1002/jmri.21049
- Job, D. E., Dickie, D. A., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2016). A brain imaging repository of normal structural MRI across the life course: brain images of normal subjects (BRAINS). *Neuroimage*. doi:10.1016/j.neuroimage.2016.01.027
- Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., et al. (2008). A National Human Neuroimaging Collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172. doi:10.1109/TITB.2008.917893

- Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi:10.1016/j.neuroimage.2013.05.094
- Keator, D. B., Wei, D., Gadde, S., Bockholt, H. J., Grethe, J. S., Marcus, D., et al. (2009). Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. *Front. Neuroinformatics* 3:30. doi:10.3389/neuro.11.030.2009
- Kim, B. J., Han, M.-K., Park, T. H., Park, S.-S., Lee, K. B., Lee, B.-C., et al. (2014). Current status of acute stroke management in Korea: a report on a multicenter, comprehensive acute stroke registry. *Int. J. Stroke* 9, 514–518. doi:10.1111/ij.12199
- Laird, A. R., Eickhoff, S. B., Fox, P. M., Uecker, A. M., Ray, K. L., Saenz, J. J., et al. (2011). The BrainMap strategy for standardization, sharing, and meta-analysis of neuroimaging data. *BMC Res. Notes* 4:349. doi:10.1186/1756-0500-4-349
- Laney, D. (2001). “3D data management: controlling data volume, velocity and variety,” in *META Group Research Note*, Vol. 6, 70. Available at: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- Larson, S. D., and Martone, M. (2013). NeuroLex.org: an online framework for neuroscience knowledge. *Front. Neuroinformatics* 7:18. doi:10.3389/fninf.2013.00018
- Lees, K. R., Bath, P. M. W., Schellinger, P. D., Kerr, D. M., Fulton, R., Hacke, W. D., et al. (2012). Contemporary outcome measures in acute stroke research: choice of primary outcome measure. *Stroke* 43, 1163–1170. doi:10.1161/STROKEAHA.111.641423
- Lindley, R. I., Wardlaw, J. M., Whiteley, W. N., Cohen, G., Blackwell, L., Murray, G. D., et al. (2015). Alteplase for acute ischemic stroke: outcomes by clinically important subgroups in the Third International Stroke Trial. *Stroke* 46, 746–756. doi:10.1161/STROKEAHA.114.006573
- MacKenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage* 42, 178–195. doi:10.1016/j.neuroimage.2008.04.186
- Maillard, P., Delcroix, N., Crivello, F., Dufouil, C., Gicquel, S., Joliot, M., et al. (2008). An automated procedure for the assessment of white matter hyperintensities by multispectral (T1, T2, PD) MRI and an evaluation of its between-centre reproducibility based on two large community databases. *Neuroradiology* 50, 31–42. doi:10.1007/s00234-007-0312-3
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit. *Neuroinformatics* 5, 11–33. doi:10.1385/NI:5:1:11
- Mennes, M., Biswal, B. B., Castellanos, F. X., and Milham, M. P. (2013). Making data sharing work: the FCP/INDI experience. *Neuroimage* 82, 683–691. doi:10.1016/j.neuroimage.2012.10.064
- Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2016). Best practices in data analysis and sharing in neuroimaging using MRI. doi:10.1101/054262
- Pilat, D., and Fukasaku, Y. (2007). OECD principles and guidelines for access to research data from public funding. *Data Sci. J.* 6, 4–11. doi:10.2481/dsj.6.OD4
- Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi:10.1038/nn.3818
- Rector, A., and Rogers, J. (2006). *Ontological and Practical Issues in Using a Description Logic to Represent Medical Concept Systems: Experience from GALEN. Reasoning Web 2nd International Summer School* (Lisbon: Springer), 197–231.
- Sandercock, P., Lindley, R., Wardlaw, J., Dennis, M., Lewis, S., Venables, G., et al. (2008). The third international stroke trial (IST-3) of thrombolysis for acute ischaemic stroke. *Trials* 9, 1–17. doi:10.1186/1745-6215-9-37
- Sandercock, P., Wardlaw, J. M., Lindley, R. I., Dennis, M., Cohen, G., Murray, G., et al. (2012). The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial. *Lancet* 379, 2352–2363. doi:10.1016/S0140-6736(12)60768-5
- Seghier, M. L., Patel, E., Prejawa, S., Ramsden, S., Selmer, A., Lim, L., et al. (2016). The PLORAS database: a data repository for predicting language outcome and recovery after stroke. *Neuroimage* 124, 1208–1212. doi:10.1016/j.neuroimage.2015.03.083
- Smith, B., Arabandi, S., Brochhausen, M., Calhoun, M., Ciccarese, S., Doyle, B., et al. (2015). Biomedical imaging ontologies: a survey and proposal for future work. *J. Pathol. Inform.* 6, 37. doi:10.4103/2153-3539.159214
- The ENOS Trial Investigators. (2006). Glyceryl trinitrate vs. control, and continuing vs. stopping temporarily prior antihypertensive therapy, in acute stroke: rationale and design of the efficacy of nitric oxide in stroke (ENOS) trial (ISRCTN99414122). *Int. J. Stroke* 1, 245–249. doi:10.1111/j.1747-4949.2006.00059.x
- The ENOS Trial Investigators. (2015). Efficacy of nitric oxide, with or without continuing antihypertensive treatment, for management of high blood pressure in acute stroke (ENOS): a partial-factorial randomised controlled trial. *Lancet* 385, 617–628. doi:10.1016/S0140-6736(14)61121-1
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: an overview. *Neuroimage* 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041
- Van Horn, J. D., and Toga, A. W. (2014). Human neuroimaging as a “Big Data” science. *Brain Imaging Behav.* 8, 323–331. doi:10.1016/j.neuroimage.2013.05.041
- Walport, M., and Brest, P. (2011). Sharing research data to improve public health. *Lancet* 377, 537–539. doi:10.1016/S0140-6736(10)62234-9
- Wang, F., Lee, R., Zhang, X., and Saltz, J. (2011). “Towards building high performance medical image management system for clinical trials,” in *Proceedings of SPIE 7967, Medical Imaging 2011: Advanced PACS-based Imaging Informatics and Therapeutic Applications*. Lake Buena Vista, FL.
- Wang, W., and Krishnan, E. (2014). Big data and clinicians: a review on the state of the science. *JMIR Med Inform* 2, e1. doi:10.2196/medinform.2913
- Warach, S. J., Luby, M., Albers, G. W., Bammer, R., Bivard, A., Campbell, B. C. V., et al. (2016). Acute stroke imaging research roadmap III: imaging selection and outcomes in acute stroke reperfusion clinical trials: consensus recommendations and further research priorities. *Stroke* 47, 1389–1398. doi:10.1161/STROKEAHA.115.012364
- Wardlaw, J., Bath, P., Sandercock, P., Perry, D., Palmer, J., Watson, G., et al. (2007). The NeuroGrid stroke exemplar clinical trial protocol. *Int. J. Stroke* 2, 63–69. doi:10.1111/j.1747-4949.2007.00092.x
- Wardlaw, J. M., Doubal, F., Armitage, P., Chappell, F., Carpenter, T., Muñoz Maniega, S., et al. (2009). Lacunar stroke is associated with diffuse blood-brain barrier dysfunction. *Ann. Neurol.* 65, 194–202. doi:10.1002/ana.21549
- Wardlaw, J. M., and Mielke, O. (2005). Early signs of brain infarction at CT: observer reliability and outcome after thrombolytic treatment – systematic review. *Radiology* 235, 444–453. doi:10.1148/radiol.2352040262
- Wardlaw, J. M., Muir, K. W., Macleod, M.-J., Weir, C., McVerry, F., Carpenter, T., et al. (2013). Clinical relevance and practical implications of trials of perfusion and angiographic imaging in patients with acute ischaemic stroke: a multi-centre cohort imaging study. *J. Neurol. Neurosurg. Psychiatry* 84, 1001–1007. doi:10.1136/jnnp-2012-304807
- Westra, B. L., Latimer, G. E., Matney, S. A., Park, J. I., Sensmeier, J., Simpson, R. L., et al. (2015). A national action plan for sharable and comparable nursing data to support practice and translational research for transforming health care. *J. Am. Med. Inform. Assoc.* 22, 600–607. doi:10.1093/jamia/ocu011
- Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer* 25, 38–49. doi:10.1109/2.121508
- Wintermark, M., Albers, G. W., Broderick, J. P., Demchuk, A. M., Fiebach, J. B., Fiehler, J., et al. (2013). Acute stroke imaging research roadmap II. *Stroke* 44, 2628–2639. doi:10.1161/STROKEAHA.113.002015

Conflict of Interest Statement: The authors are aware of no conflict of interest that might bias the work presented here. Our funding sources had no involvement in this work.

Copyright © 2016 Danso, Job, Gonzalez, Dickie, Palmer, Ure, Bath, Sandercock and Wardlaw. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories

Christian Barillot^{1,2,3*}, Elise Bannier^{1,2,3,4}, Olivier Commowick^{1,2,3}, Isabelle Corouge^{1,2,3}, Anthony Baire¹, Ines Fakhfakh^{1,2,3}, Justine Guillaumont^{1,2,3}, Yao Yao^{1,2,3} and Michael Kain^{1,2,3}

¹Inria, Rennes, France, ²Visages U746, INSERM, Rennes, France, ³UMR 6074, CNRS, IRISA, University of Rennes I, Rennes, France, ⁴Department of Neuroradiology, CHU Rennes, Rennes, France

OPEN ACCESS

Edited by:

Marleen De Bruijne,
Erasmus MC, Netherlands

Reviewed by:

Suyash P. Awate,
Indian Institute of Technology
Bombay, India
Henri Vrooman,
Sophia Children's Hospital,
Netherlands

*Correspondence:

Christian Barillot
christian.barillot@irisa.fr

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 08 April 2016

Accepted: 20 October 2016

Published: 01 December 2016

Citation:

Barillot C, Bannier E, Commowick O,
Corouge I, Baire A, Fakhfakh I,
Guillaumont J, Yao Y and Kain M
(2016) Shanoir: Applying the
Software as a Service Distribution
Model to Manage Brain Imaging
Research Repositories.
Front. ICT 3:25.
doi: 10.3389/fict.2016.00025

Two of the major concerns of researchers and clinicians performing neuroimaging experiments are managing the huge quantity and diversity of data and the ability to compare their experiments and the programs they develop with those of their peers. In this context, we introduce Shanoir, which uses a type of cloud computing known as software as a service to manage neuroimaging data used in the clinical neurosciences. Thanks to a formal model of medical imaging data (an ontology), Shanoir provides an open source neuroinformatics environment designed to structure, manage, archive, visualize, and share neuroimaging data with an emphasis on managing multi-institutional, collaborative research projects. This article covers how images are accessed through the Shanoir Data Management System and describes the data repositories that are hosted and managed by the Shanoir environment in different contexts.

Keywords: neuroimaging, database, data sharing, neuroinformatics, software as a service, cloud computing, web application, web services

INTRODUCTION

Context

Two of the major concerns for researchers and clinicians performing neuroimaging experiments are managing the huge quantity and diversity of data and the ability to compare their experiments and the programs they develop with those of their peers. In practice, researchers and clinicians in the neuroimaging field are encouraged to set up large-scale experiments, but the inability to recruit sufficient local subjects who meet specific criteria results in the need for cooperation to gather the relevant imaging data. Pooling experimental results *via* the Internet and cooperative efforts by centers provide larger and more specific subject populations that expand the scope and value of scientific research.

Abbreviations: DICOM, Digital Imaging and Communications in Medicine; DOLCE, Descriptive Ontology for Linguistic and Cognitive Engineering; IRC, imaging resource center; J2EE, Java platform enterprise edition; JAX-WS, Java API for XML web services; JWS, Java web start; NIfTI, neuroimaging informatics technology initiative; OFSEP, Observatoire Français de la Sclérose en Plaques (French multiple sclerosis observatory); OWL, web ontology language; PACS, picture archiving and communication system; PI, principal investigator; SaaS, software as a service; Shanoir, sharing neuroimaging resources; SOAP, simple object access protocol; WSDL, Web Service Description Language.

Searches on distributed neuroimaging databases for similar results and images containing singularities (quirk, peculiarities, etc.) or the use of data mining techniques may highlight possible similarities. Such efforts also broaden the possible panel of people involved in neuroimaging studies while maintaining the quality of the work. Indeed, the explosion of data generated by the neurosciences community in the early 1990s has resulted in the need for innovative techniques for data and knowledge sharing and reuse (Roland and Zilles, 1994; Mazziotta et al., 1995; Shepherd et al., 1998). This has led to the emergence of large-scale projects on the human brain. A recent objective added to these initial issues is the application of data analysis and data processing software to various data repository systems for knowledge discovery and data mining, including its more recent extension to merging imaging and genetic data (Hibar et al., 2015). In parallel, the development of web applications has stimulated the interest of researchers and clinicians in distributed databases and information sharing.

Background

It is now commonly accepted in the neuroimaging community that sharing data and image processing services will play a crucial role in translational research (Barillot et al., 2003; Walport and Brest, 2011; Poline et al., 2012; Keator et al., 2013; Van Horn and Gazzaniga, 2013; Poldrack and Gorgolewski, 2014). Research funding agencies now clearly identify the sharing of scientific resources (data processing) as a top priority. International organizations such as the International Neuroinformatics Coordinating Facility (INCF)¹ are now dedicated to promoting the field of neuroinformatics (Book et al., 2013; Kennedy et al., 2015). Sharing data and image processing services for translational research are needed for:

- (1) the integration of large data sets for population-wide studies and *construction of imaging cohorts* (Shepherd et al., 1998; Van Horn et al., 2001; Barillot et al., 2006; Evans and Brain Development Cooperative Group, 2006; Jack et al., 2008; Hall et al., 2012; Weiner et al., 2012; Marcus et al., 2013; Van Essen et al., 2013),
- (2) the validation of image processing tools on reference datasets for *validation and quality control of image processing procedures* (Styner et al., 2008; Menze et al., 2015),
- (3) the reuse of image processing pipeline on different sets of data and different peers for *sharing processing tools* (Keator et al., 2009, 2013; Ooi et al., 2009; Dinov et al., 2010; Gorgolewski et al., 2011; Bellec et al., 2012; Glatard et al., 2014), and
- (4) the validation of research results based on proofed control statistical analysis of images for *validation and quality control of experimental research* (Carp, 2012; Button et al., 2013; Ioannidis, 2014; Ioannidis et al., 2014).

This is particularly significant in the field of neuroimaging as several large recent multicenter initiatives have shown. These include Evans and Brain Development Cooperative Group (2006), which performed a study using magnetic resonance

imaging (MRI) of normal brain maturation from birth to adulthood in approximately 500 children with behavior disorders, and the Alzheimer's disease neuroimaging initiative (ADNI), which has assembled a very large variety of images for its work (Weiner et al., 2012). The Human Connectome Project (HCP), which worked with 1,200 healthy volunteers to investigate brain connectivity in the normal brain (Marcus et al., 2013; Van Essen et al., 2013), is another well-known example illustrating the importance of aggregate imaging data and relating data warehouses to image processing resources.

To provide archiving solutions for large or various multicenter projects, several architectures have already been proposed. The Biomedical Informatics Research Network (BIRN) has been a pioneer in launching brain imaging solutions (Gupta et al., 2003; Keator et al., 2008, 2009; Ashish et al., 2010). Another early initiative, the FMRIDC project sought to share task-based fMRI imaging data (Van Horn and Gazzaniga, 2013). @NeurIST set up a dedicated solution (funded by an Integrated European Project) to support research and treatment of cerebral aneurysms using heterogeneous data, computing, and complex processing services (Benkner et al., 2010). The LORIS/CBRAIN project is an initiative to develop a pan-Canadian platform for distributed processing, analysis, exchange, and visualization of brain imaging data (Das et al., 2011; Sherif et al., 2014). Finally, other generic data management systems have been proposed to offer shared solutions for managing multicenter studies. These include the Extensible Neuroimaging Archive Toolkit (XNAT) (Marcus et al., 2007), which has been successful due to its integration in the management of large projects (Marcus et al., 2013) and ability to communicate with data management servers *via* dedicated REST web services, and the Collaborative Informatics and Neuroimaging Suite (COINS), which provides a web-based neuroimaging and neuropsychology software suite (Scott et al., 2011). Although the extensibility of these platforms is part of the motivations, none of them are built on top of a formal semantic model or ontology that can guarantee the sustainability of any evolution of the original data scheme.

Significance

In this context, the Sharing Neuroimaging Resources (Shanoir) environment enables sharing between distributed sources of neuroimaging information over the Internet, whether the sources are located in various centers of experimentation, clinical departments of neurology, or research centers in cognitive neurosciences or image processing. A large variety of users can thus share, exchange, and have controlled access to neuroimaging information using the software as a service (SaaS) type of cloud computing (Rimal et al., 2009) almost as easily as if the data were stored locally.

In this paper, we introduce the Shanoir software environment for managing neuroimaging data production in the context of clinical neurosciences and show how the images are accessible through the Shanoir Data Management System. Shanoir is an open source neuroinformatics environment designed to structure, manage, archive, visualize, and share neuroimaging data with an emphasis on multi-institutional, collaborative research projects. The software offers features commonly found in neuroimaging

¹<http://www.incf.org/>.

data management systems along with research-oriented data organization capabilities and enhanced accessibility. It also provides user-friendly secure web access and an intuitive workflow that facilitates the collection and retrieval of neuroimaging data from multiple sources.

In Section “Shanoir Software Environment,” we provide a brief overview of the software environment including its core (web portal, Study Card, and quality control) and extensions for loading, querying, and processing data. Section “Data Repositories” describes the data repositories, while Section “Conclusion and Perspectives” covers the use of these repositories and potential evolution.

SHANOIR SOFTWARE ENVIRONMENT

General Description of the Software Environment

Shanoir is an open source software environment with QPL licensing designed to archive, structure, manage, visualize, and share neuroimaging data with an emphasis on managing collaborative research projects. It includes the common features of neuroimaging data management systems along with research-oriented data organization and enhanced accessibility. Shanoir is based on a secure J2EE application running on a JBoss server that is accessed *via* graphical interfaces in a browser or by third-party programs *via* web services using simple object access protocol (SOAP). It behaves like a repository of neuroimaging files coupled with a relational database containing metadata (**Figure 1**).

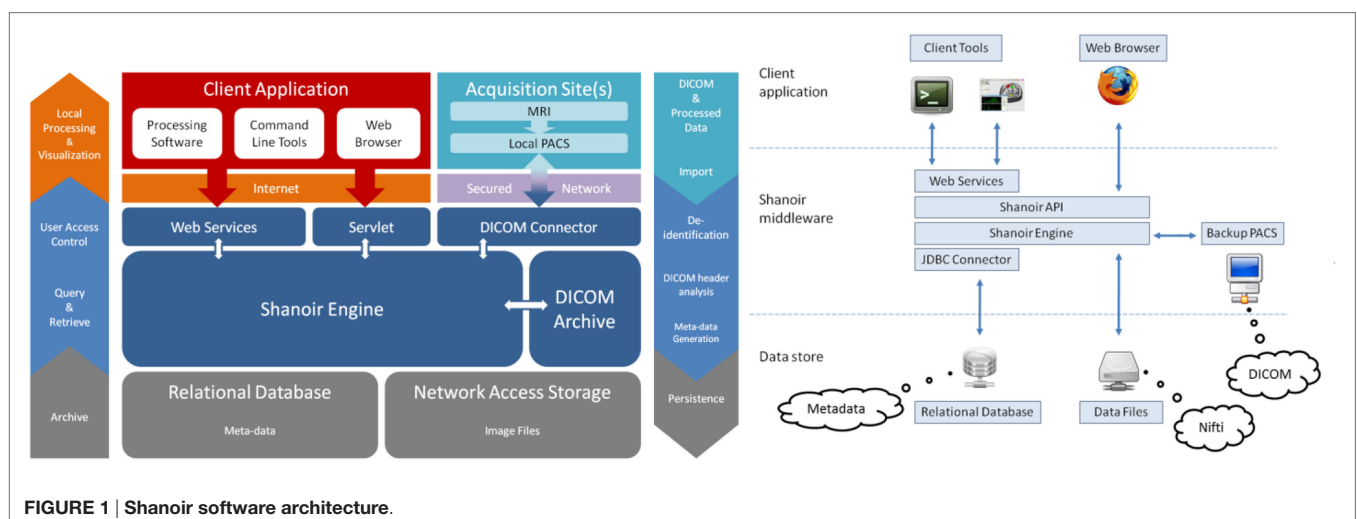
Shanoir uses semantics for structuring the concepts as defined by the OntoNeuroLOG² ontology (Temal et al., 2008; Michel et al., 2010). OntoNeuroLOG reuses and extends the OntoNeuroBase ontology defined earlier (Barillot et al., 2006) (see **Figure 2**). Both were designed using the methodological framework (Temal et al., 2008) of the foundational Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Masolo et al., 2003) and

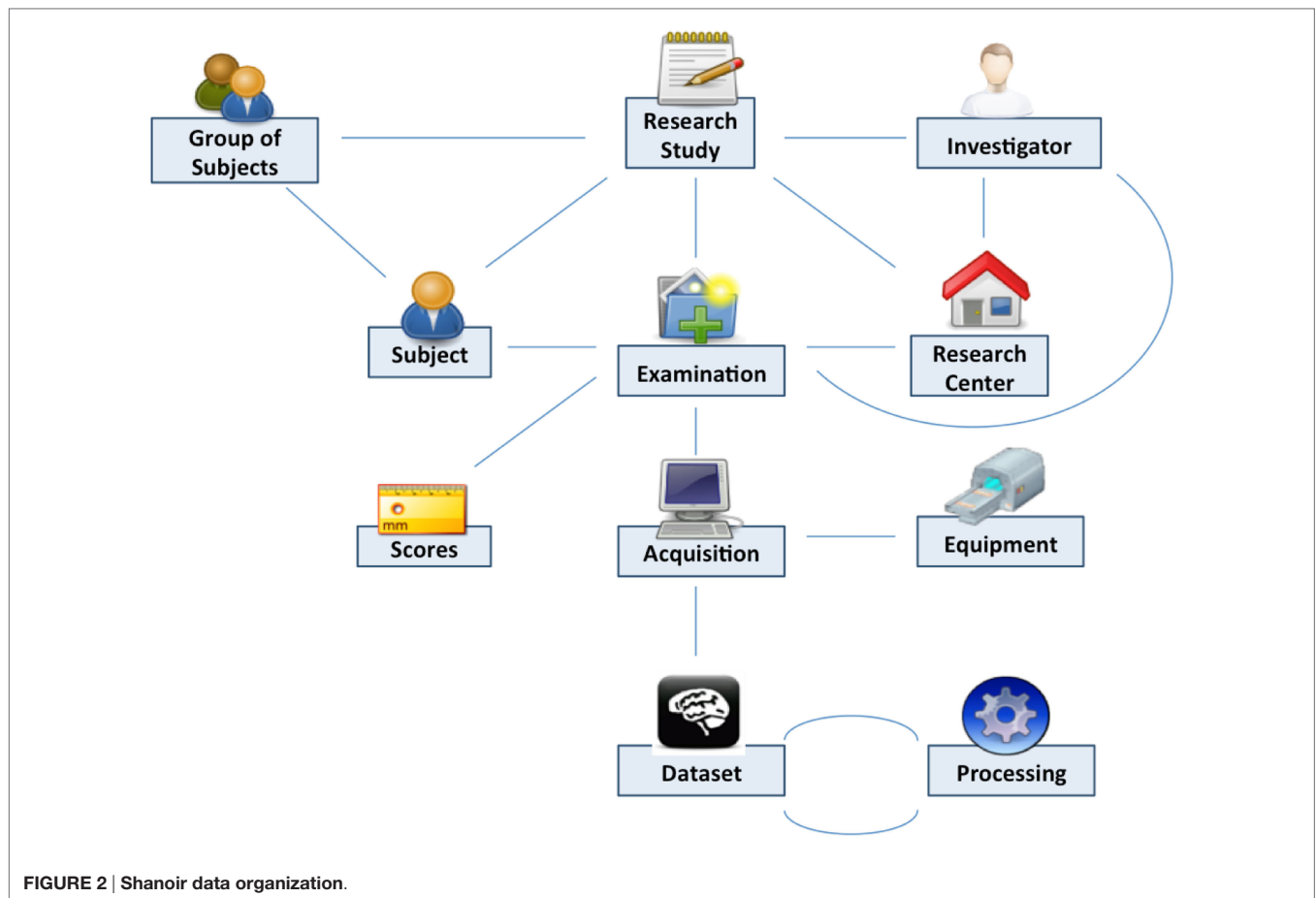
a number of core ontologies that provide generic, basic, and minimal concepts and relationships in specific fields such as artifacts, participant roles, information, and discourse acts. In Shanoir, the OWL-Lite implementation was manually derived from the OntoNeuroLOG initial expressive representation to Java classes. The data model based on this ontology is dedicated to the neuroimaging field and is structured around research studies in which patients are examined to produce image acquisitions or clinical scores. Each image acquisition is composed of datasets represented by acquisition parameters and image files. For security and legal reasons, all data on the system are anonymous by default, this can be customized with specific algorithm (e.g., defacing is not currently implemented but can easily be embedded in a specific anonymizer that Shanoir will call).

Raw as well as derived (i.e., post-processed) image files can also be imported into the system using medical imaging technology [e.g., media based on the Digital Imaging and Communications in Medicine (DICOM) standard, picture archiving and communication system (PACS), or image files in the Neuroimaging Informatics Technology Initiative (NifTI)/Analyze-style data format] using online wizards, which complete related metadata, command line tools, or SOAP web services. Once identity information has been removed from raw data during the importation process, the DICOM header content is automatically extracted, enriched, and inserted into the database with the customizable “Study Card” feature. Shanoir can also record any execution process for retrieval of workflows applied to a particular dataset along with the derived data.

Clinical scores from instrument assessments (e.g., neuropsychological tests) can be recorded and easily retrieved and exported in different formats (Excel, CSV, and XML). The instrument database is scalable and new measures can be added in order to meet specific project needs (**Figure 3**). Scores, image acquisitions, and post-processed images are bound together, so that relationships can be analyzed. Using cross-data navigation and advanced search criteria, the user can quickly indicate a subset of data for download. Client-side applications have also been developed to locally access and exploit data through web services. The security

²OntoNeuroLOG: http://neurolog.i3s.unice.fr/public_namespace/ontology.





features of the system require authentication with user rights set for each study. A study manager can define the users allowed to see, download, or import data into his/her study or simply make it public.

In practice, Shanoir serves neuroimaging researchers by efficiently organizing their studies while cooperating with other laboratories. By managing patient privacy, Shanoir offers the possibility of using clinical data in a research context. Finally, it is a handy solution for publishing and sharing data with a broader community.

Study Card and Quality Control Concepts

Images can be imported in Shanoir from various sources: DICOM media, PACS (with DICOM Query and Retrieve), and 3D/4D image files (in NIfTI/Analyze format). Users are guided step-by-step through online forms to perform imports. In addition to archiving DICOM files, NIfTI copies are automatically generated and saved. This is convenient since the NIfTI format is better suited to perform image processing (such as registration, segmentation, and statistical analysis) than the DICOM format.

The Study Card

During archiving, the DICOM files are processed in two phases. The first phase de-identifies the images. The second phase

populates the database with the new metadata items generated from the DICOM header and enriched with the Study Card, which enables online metadata wrapping between the local data to be imported (center, acquisition equipment, etc.) and the semantic concepts of the research study to which the data will be assigned. The actual DICOM metadata can thus be aligned with the ontology and also provides additional allocation of concepts to the stored images that are more closely related to the research study protocol (e.g., functional MRI, perfusion imaging, contrast agent, diffusion imaging, etc.). The mechanism behind this feature is based on a set of rules that the user predefines to associate specific acquisition equipment and a specific data production site to the desired research study. Each rule determines the specific value of a metadata item according to the value(s) of one or more specific DICOM tag(s) (e.g., Series Description, see **Figure 4**). This greatly facilitates the consistent recording and alignment to the ontology of metadata for all data in a research study without the need for tedious workflow during the online import of images. Due to the simplicity of the process, no specific skills are required to perform data import, and it only takes a few minutes over the Internet. The “Study Card” concept makes possible an automatic quality control of the imported data using their metadata. For instance, a conformal statement can be attached to the imported data according to a match score to the Study Card rules.



FIGURE 3 | Shanoir “instrument” database can be used for attaching clinical scores to images (e.g., EDSS score in MS). An instrument can be any record where an alphanumerical value can be attached.

Quality Assessment

Shanoir’s next major functionality concerns the quality check of the images for conformity of the imported data with the pre-defined study protocol and ensures the integrity of the archived data. We have identified three levels of control:

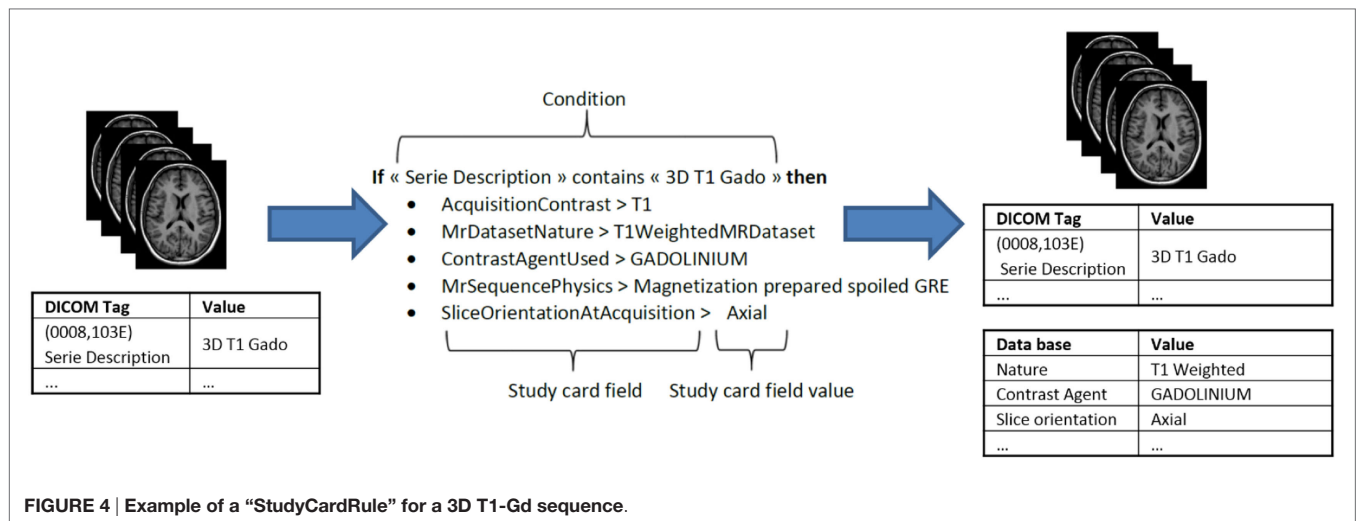
- study protocol: controls the time interval between examinations (expected visits) as defined by the principal investigator (PI) of the study;
- acquisition protocol: controls the presence of all the sequences of the imaging protocol as defined by the PI of the study; and
- raw data:
 - the software automatically checks the range of parameters for a given protocol, experimental center, or acquisition

scanner as defined in the Study Card by the PI’s technical representative;

- visual inspection of the image quality and integrity can be reported and assigned to the imported data; however, the mechanism to detect the visual quality is not yet integrated in the Shanoir environment.

In the next release of Shanoir, quality assessment will be present as flags (flawless, acceptable, or inadmissible) in the database.

This QA capability does not address the control of image formation as, for instance, to check image artifacts (bias, motions, ghosting, etc.). This category of QA can be implemented in a dedicated image visualization and processing tools that interoperate with Shanoir through the dedicated web services.



Web Portal

Shanoir provides user-friendly secure web access and offers an intuitive workflow to facilitate the collection and retrieval of neuroimaging data from multiple sources (Figure 5). On the home page, the user has direct access to the most frequent functionalities: Find and Download Datasets, Explore the Research Studies, Find Clinical Scores, and Import Data (Figure 6). On the top of all pages, the user always has a very complete navigation menu that leads to all services.

Interoperability

Interoperability is a very important concern for the Shanoir environment. Shanoir offers web services interface that is open to a large variety of clients. We already offer several dedicated interface that are already in used by different external applications. Hereafter, we described four of these external services that are currently available and run independently to each other: ShanoirUploader, QtShanoir, medInria, and iShanoir developed either in C++, Java, or Objective-C environments.

SOAP for Integration of Services

The Shanoir web services interface is based on the SOAP. Messages between clients and the server are exchanged using Extensible Markup Language (XML) with well-defined elements. The Hypertext Transfer Protocol (HTTP) is used with Transport Layer Security (TLS). Elements and services are described with the Web Service Description Language (WSDL). Based on this description, client stubs can be automatically generated to simplify the connection of new clients. The web service layer is implemented with the Java API for XML web services (JAX-WS). Shanoir offers numerous dedicated web services:

- “EntityCreator”: creates new entities, such as creating a new subject in the database
- “CredentialTester”: validates if username and password are correct
- “Downloader”: downloading files/datasets on base of dataset IDs

- “CenterFinder”: find center(s) based on different search criteria, i.e., study or investigator
- “DatasetAcquisitionFinder”: find acquisition(s) based on IDs or examinations
- “DatasetFinder”: find dataset(s) based on multiple search criteria/filters
- “DatasetProcessingFinder”: find dataset processing(s) based on IDs
- “ExaminationFinder”: find examination(s) based on multiple search criteria/filters
- “ExperimentalGroupOfSubjectsFinder”: find group of subjects based on multiple criteria
- “InvestigatorFinder”: find investigator(s) based on IDs or centers
- “MrDatasetFinder”: find MR dataset(s) based on multiple search criteria/filters
- “StudyFinder”: find study/-ies based on multiple search criteria/filters
- “SubjectFinder”: find subject(s) based on IDs with multiple filters
- “DatasetImporter”: import dataset files to already existing entities in the database
- “ReferenceLister”: shows list of reference strings stored in the database
- “FileUploader”: upload files in local archive for later import, used by ShanoirUploader

ShanoirUploader for Seamless Integration of Data

“ShanoirUploader” is a Java desktop application that transfers data securely between a PACS and a Shanoir server instance (e.g., within a hospital). It offers both a direct DICOM query/retrieve connection to search and download images from a local PACS and a DICOM CD upload facility. After retrieval, the DICOM files are locally anonymized and then uploaded to the Shanoir server (the anonymization algorithm can be customized according to specific operational/regulation constraints). The primary goals of the application are to enable mass data transfers between different remote server instances and reduce

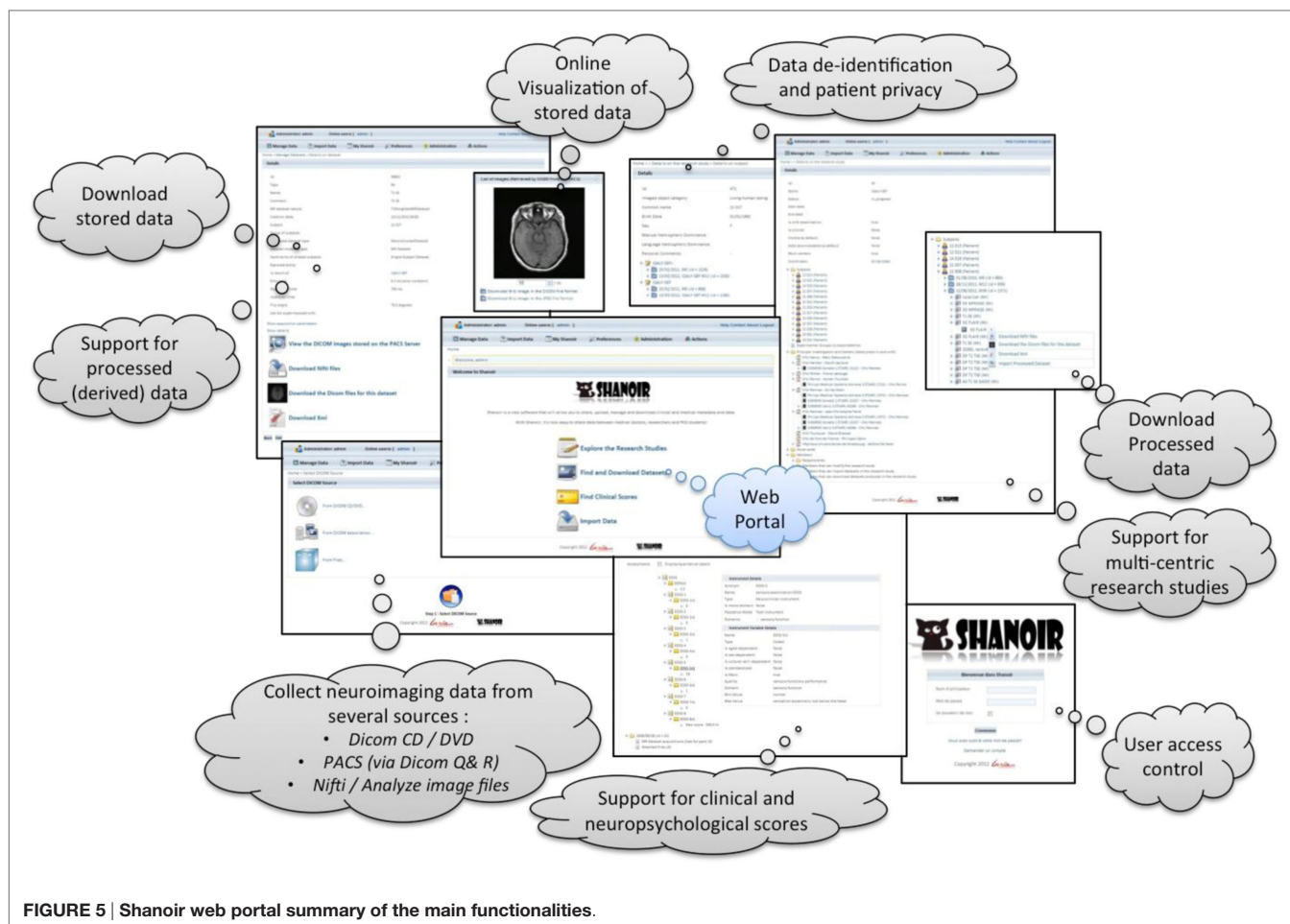


FIGURE 5 | Shanoir web portal summary of the main functionalities.

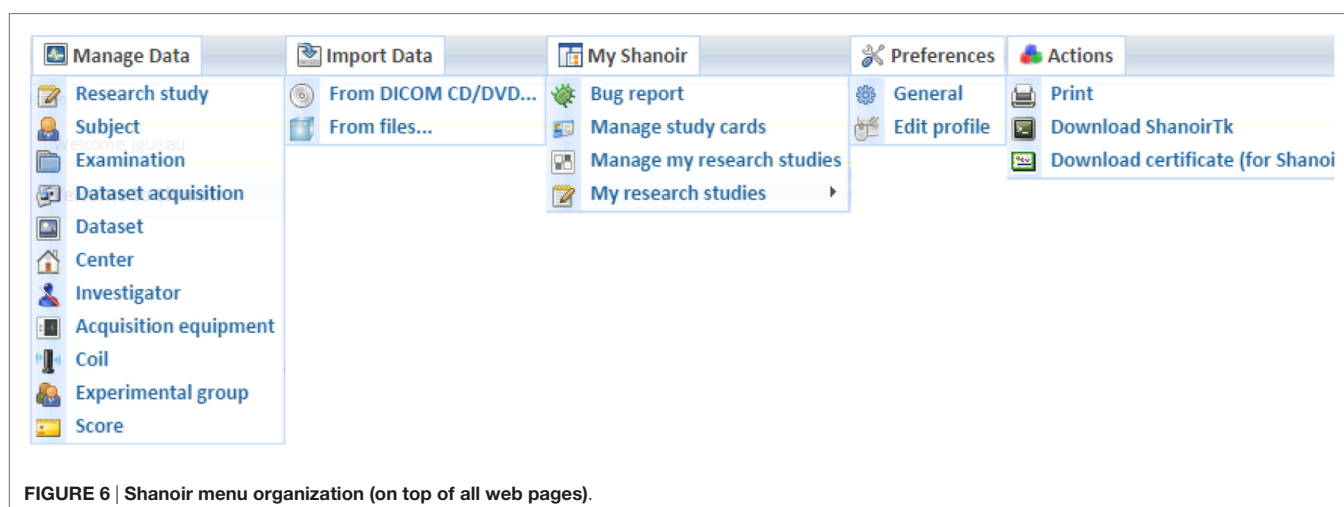


FIGURE 6 | Shanoir menu organization (on top of all web pages).

user waiting time when importing data into Shanoir. Most of the import time thus involves data transfer.

“ShanoirUploader” requires a local Java installation. For a simpler distribution and installation of the software, Java

Web Start (JWS) can be used. The application can be installed with a simple web link that is opened in a web browser. Java takes care of the installation and, later, of automatic updates. Internal components are based on Java Swing for the graphical

user interface (**Figure 7**), *dcm4chee*³ libraries to connect with a PACS and Java, and WebServices (JAX-WS) to transfer data to a Shanoir server.

Apache Solr for Metadata Querying

Shanoir integrates the open source enterprise search platform Apache Solr,⁴ which provides users with a vast array of advanced features such as near real-time indexing and queries, full-text searching, faceted navigation, autosuggestion, and autocomplete.

One of the most important features of the Solr search is the faceted navigation. Facets correspond to properties of the Solr information elements and are derived by analyzing pre-existing metadata that are related to the ontology model used by Shanoir.

Shanoir users can access all metadata with a simple Solr search bar. After entering at least one character, a user will be automatically guided to complete his search. Data are sorted by categories and dynamically displayed once a facet is chosen. By clicking on Solr data results, users access all the additional information available in Shanoir corresponding to their search, and then use these queries for local downloading (**Figure 8**).

All metadata are indexed in a JBoss server that hosts the Solr servlets. A custom security post-filter has also been developed and implemented in Shanoir to control user access. This filter retrieves user identification and access rights in Shanoir and interacts with the Solr server to show relevant results that the user is allowed to access.

iShanoir for Mobile Data Access

An iOS application, iShanoir, has been developed for iPhones and iPads. It opens a secure connection with a Shanoir server

and enables the user to access data stored on a Shanoir server. With iShanoir, the user can navigate within the Shanoir data tree structure on the server. After data are selected from the mobile app, the images can be downloaded to the local device, displayed, and analyzed with any local DICOM viewer or through cloud services (i.e., Dropbox, iCloud, Google, or OneDrive).

The iShanoir application has been developed with Xcode and implemented in Objective-C. For the graphical user interface, two storyboards have been developed to fit the different display sizes between iPhones and iPads (**Figure 9**). It uses the following iOS frameworks: Foundation, CoreFoundation, UIKit, and CFNetwork. For implementation of the SOAP web services client, the WSDL2ObjC utility has been used as it offers a client stub code generation based on the server WSDL document.

QtShanoir for Image Processing

Shanoir web services may also be queried from standalone C++/Qt applications through the QtShanoir library,⁵ which uses SOAP web services provided by a Shanoir server to access and display studies, patients, and data with their associated metadata. In QtShanoir, a set of Qt widgets are defined that can be embedded in any Qt application. The library was used to implement a Shanoir query plugin inside the medInria visualization and processing software⁶ for interrogation and downloading of image data from Shanoir for processing within medInria, for example, using the available processing tools and then upload the processing results back to the Shanoir server with the correct metadata values (**Figure 10**).

Distribution of Shanoir

The Shanoir server can be freely downloaded on request. It is currently deployed using Docker containers running on a Linux

³<http://www.dcm4che.org>.

⁴<http://lucene.apache.org/solr/>.

⁵QtShanoir library: <http://qtshanoir.gforge.inria.fr>.

⁶medInria: <http://med.inria.fr>.

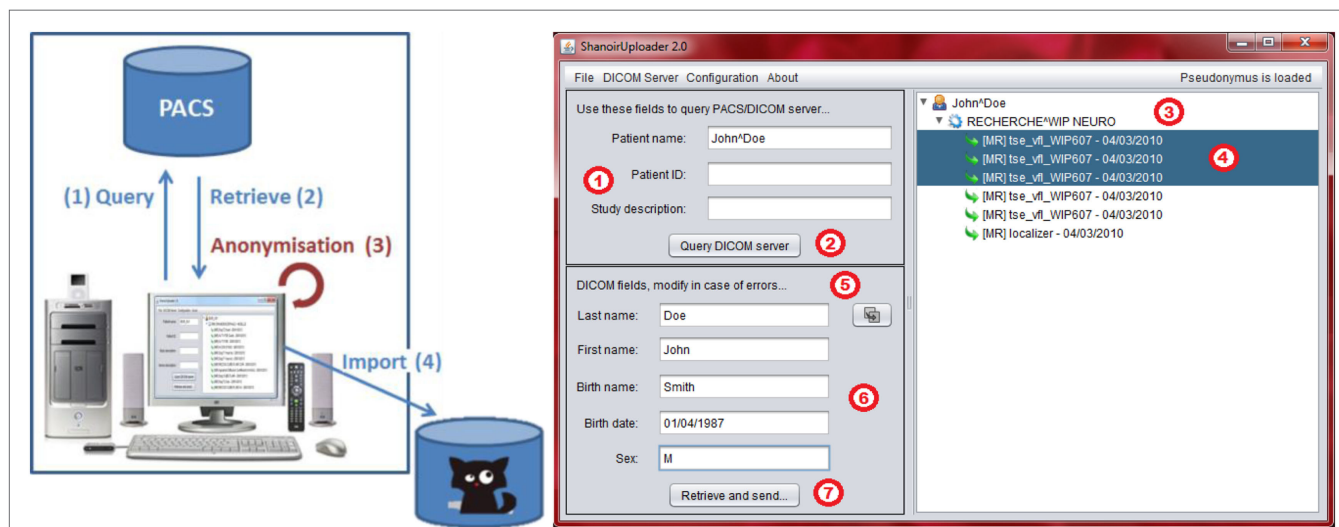


FIGURE 7 | Shanoir Uploader architecture for secure transfer of local PACS data to a Shanoir Server (left) and user interface (right).

Administrator: admin Online users: [admin] Solr Search [] OK Help Contact About Logout

Manage Data Import Data My Shanoir Preferences Administration Actions

Current Selection

- > remove all
- > (x) study_name:USPIO-6
- > (x) subject_name:01002

Study Name

USPIO-6

Subject Name

01002

Dataset Type

2D FLASH 2D FLASH MT 3D FLAIR 3D MPRAGE 3D MPRAGE GADO
DIFF 3D DIR DP T2 TSE FIELD MAP RELAXO T1 30Å* RELAXO T2
RELAXO T2 USPIO RELAXO T2STAR
RELAXO T2STAR USPIO T1 SE GADO T2 SE_MC 7 echos uspio
T2 SE_MC TRA 7 echos T2STAR GRE3D TRA T2STAR GRE3D uspio localizer

Dataset Creation Date

August 2009

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

< 1 2 3 ... 47 48 > displaying 1 to 10 of 478

Dataset Name: T2STAR GRE3D uspio 1
Dataset ID: 1200
Dataset Creation Date: 2009-08-10T22:00:00Z
study_name: USPIO-6
subject_name: 01002
dataset_comment: T2STAR GRE3D uspio

Dataset Name: T2STAR GRE3D uspio 2
Dataset ID: 1201
Dataset Creation Date: 2009-08-10T22:00:00Z
study_name: USPIO-6
subject_name: 01002
dataset_comment: T2STAR GRE3D uspio

Dataset Name: T2STAR GRE3D uspio 3
Dataset ID: 1202
Dataset Creation Date: 2009-08-10T22:00:00Z
study_name: USPIO-6
subject_name: 01002
dataset_comment: T2STAR GRE3D uspio

Dataset Name: T2STAR GRE3D uspio 4
Dataset ID: 1203
Dataset Creation Date: 2009-08-10T22:00:00Z
study_name: USPIO-6
subject_name: 01002
dataset_comment: T2STAR GRE3D uspio

Dataset Name: T2STAR GRE3D uspio 5
Dataset ID: 1204
Dataset Creation Date: 2009-08-10T22:00:00Z
study_name: USPIO-6
subject_name: 01002
dataset_comment: T2STAR GRE3D uspio

FIGURE 8 | Example of Apache Solr search of the Shanoir server.

kernel. Linux containers are implemented using namespaces for locating each type of resource. Dockers are tools for managing lightweight method of virtualization (named containers) on Linux that are lighter than traditional virtual machines. The host and guest systems share the same kernel. The kernel is responsible for host ↔ guest and guest ↔ guest isolation (the result of system calls depends on the container in which the calling process is running). As described in **Figure 11**, a minimal Shanoir deployment consists of four servers running in at least four separate containers:

- “*shanoir_container*”: the actual Shanoir server. It relies on a *mysql* container (for the database) and on the PACS container (for archiving DICOM data),

- “*pacs_container*”: the DICOM PACS server, currently managed by *dcm4chee*,⁷
- “*mysql_container*”: the database server that is hosting two databases: *shanoirdb* and *pacsdb*,
- “*nginx_container*”: the web frontend server based on a *nginx*⁸ HTTP server configured as a reverse-proxy for reaching the Shanoir server. It is the only server that is publicly reachable. It provides TLS encryption and security filtering, and
- “*smtpsink_container*”: an optional SMTP server for outgoing e-mails.

⁷<http://www.dcm4che.org/>.

⁸<http://wiki.nginx.org/>.



FIGURE 9 | Example of storyboard interfaces under the iShanoir iOS mobile application connected to a Shanoir server.

DATA REPOSITORIES

Each Shanoir repository has an administrator that manages the access rights of the repository. Each user requests an account through a web-based form and specifies which study he/she wants to access, contact, role in the study, required level of expertise/access (guest, user, expert, and admin), etc. According to the information provided, the Shanoir administrator of the repository determines whether the user can access the system. Access to a specific study is granted by the person responsible for the study (i.e., the PI of the research study or the official representative). Depending on these settings, the new user will be able to see, download, and import datasets or even modify the study parameters. The corresponding rights are set for a limited time and must be renewed regularly. If requested, the user can receive a report by e-mail each time data are imported into the study.

The Shanoir@Neurinfo Repository

Started in 2009, the Neurinfo MRI research facility⁹ promotes translational clinical research and supports the development

⁹<http://www.neurinfo.org/>.

of clinical research, technological activity, and methodological activity. It offers resources for *in vivo* human imaging acquisition, image data analysis, and image data management. A large community of users, both clinicians and scientists, uses the resources as part of local, national, and international imaging-based research projects.

All data produced at Neurinfo for academic or clinical research purposes are managed through a dedicated Shanoir@Neurinfo repository (Figure 12) administered by the facility's staff. The Shanoir@Neurinfo server also hosts data from imaging studies at multiple sites. In total, around 2To of data from 42 centers and 50 MR scanners are archived at this repository. The amount of data increases by 30 GB per month (see table in Figure 12).

In daily practice, DICOM data are imported by a technician from either a local PACS, a CD/DVD, or a disk drive containing the DICOMDIR in its root directory and the DICOM files.

The clinical studies stored on the Shanoir@Neurinfo server concern the whole body (brain, spine, heart, lung, pelvis, vasculature, etc.) with a major focus on brain anatomy and function in normal control and pathological populations.

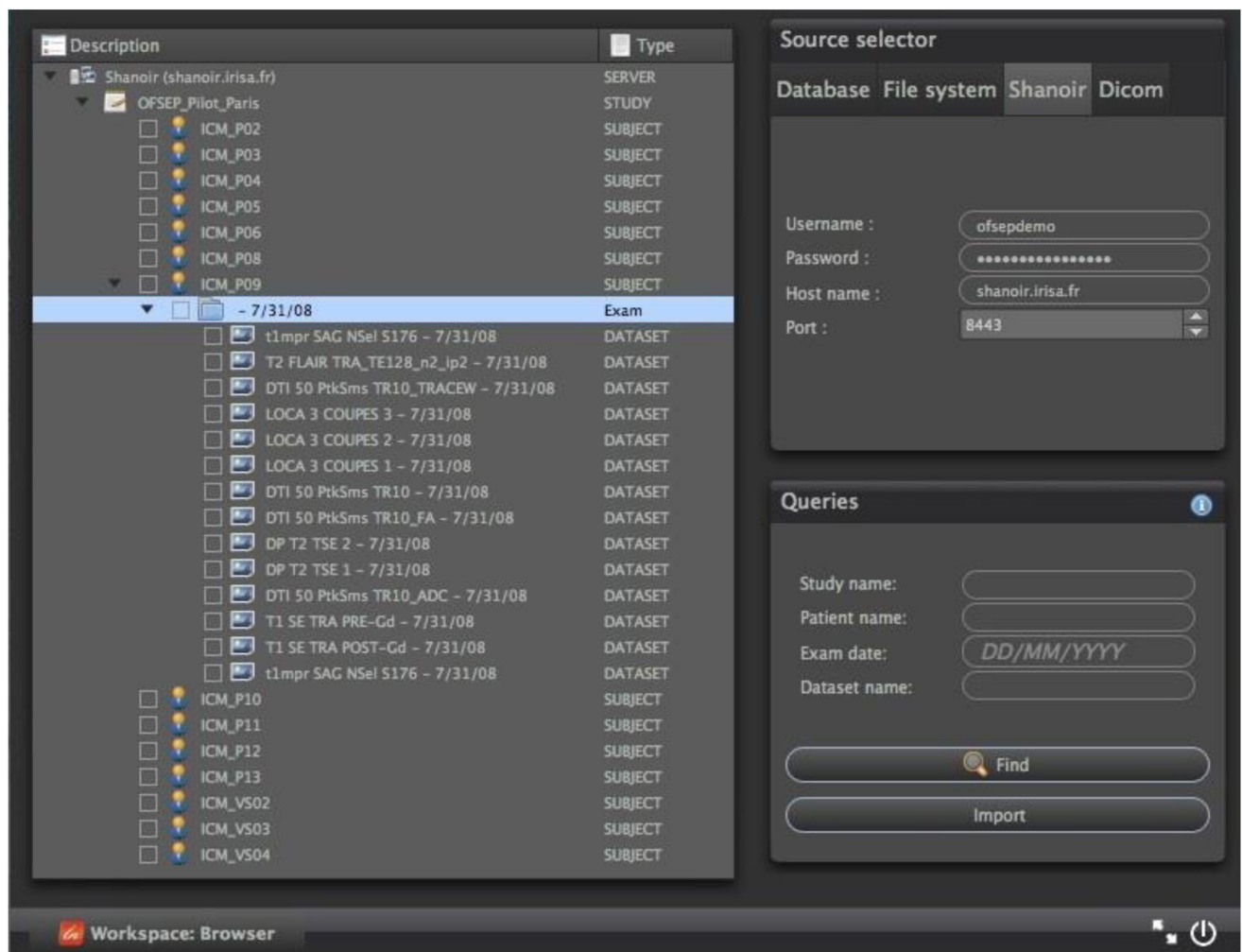


FIGURE 10 | Example of a Shanoir query service within the medInria environment by using QtShanoir web services.

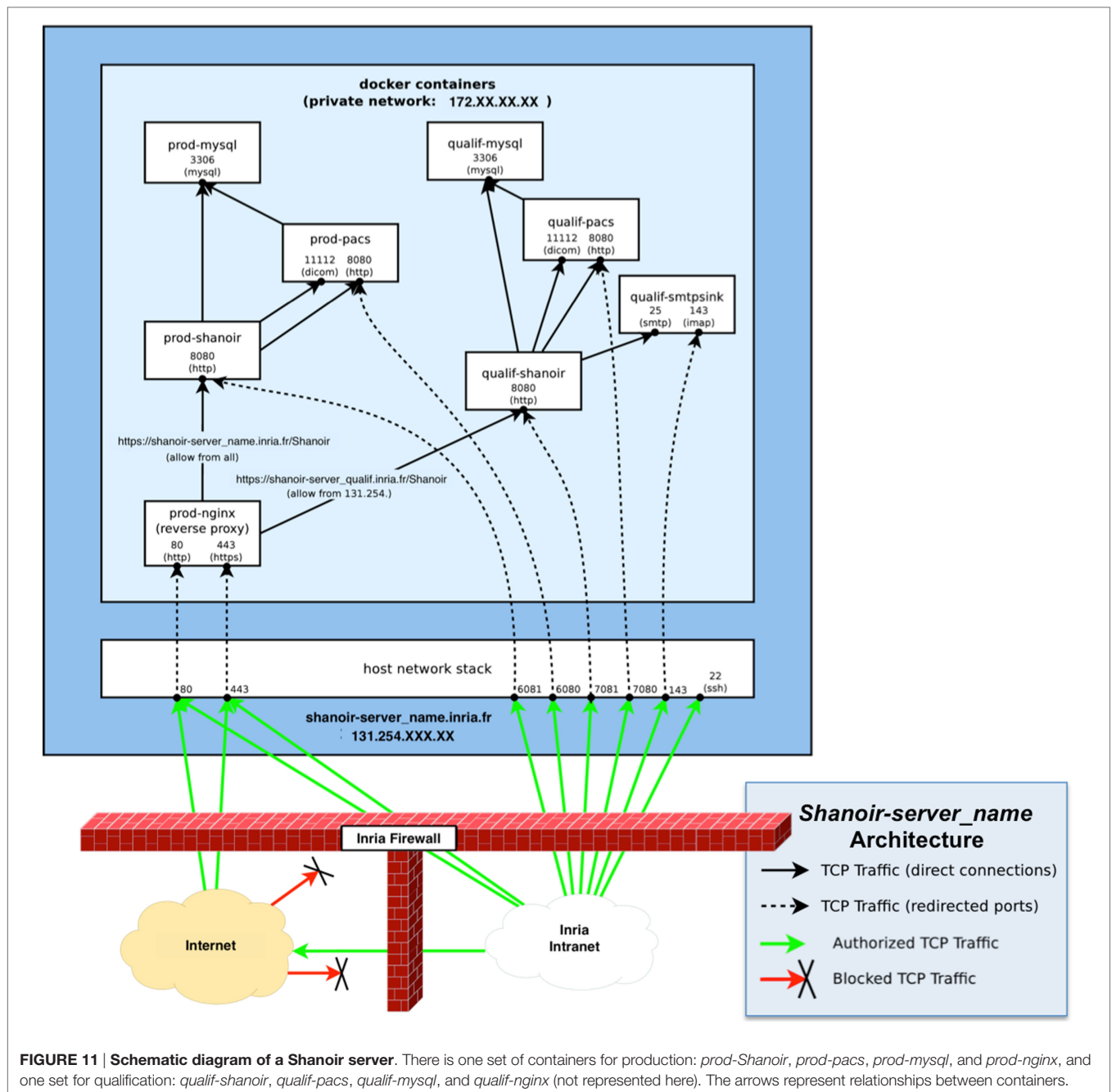
Out of the 70 or so ongoing research studies on the Neurinfo platform, 75% relates to brain imaging, 15% concern abdominal imaging, and 10% concern heart imaging. Among the neuroimaging clinical studies, multiple sclerosis, dementia, tumors, stroke, and mood disorders are the most investigated pathologies.

Depending on the specific nature of the research study, typical neuroimaging protocols include structural imaging, functional BOLD MRI, Arterial Spin Labeling perfusion imaging, diffusion imaging, relaxometry sequences, pre- and post-gadolinium T1w sequences, or vascular sequences. The following studies are examples of research carried out with the Shanoir@Neurinfo service as well as the types of data that are managed by the Shanoir@Neurinfo repository.

Along with the MRI raw data, post-processed images can be stored for each dataset. For example, in a study on functional motor activation, the motor areas were delineated by a trained radiologist and associated with each 3D T1w image. Multiple sclerosis (MS) lesion segmentation masks can also be attached

to the examination. In addition to image data, clinical scores can also be stored for each subject in the repository. Several MS clinical studies collect measurements such as the number of T2 new lesions, and number of T1w Gd enhancing lesions or clinical scores such as EDSS. These measurements are also included in the search engine and consequently easily accessible through requests. For more advanced clinical follow-ups, Shanoir can easily be interfaced with existing databases.

The general policy for the Shanoir@Neurinfo repository for dissemination of data related to a particular study is decided upon beforehand with the PI in compliance with the informed consent form approved by the ethics committee and signed by the participant. Any opening of the data to third parties is submitted to the approval of the PI prior to allowing (complete or partial) access to a third-party user. Nonetheless, to ensure dissemination and the best use of data acquired from public funding, the Neurinfo team strongly encourages investigators to share their data, which is usually done after an embargo period.



Shanoir@OFSEP Repository

The French Multiple Sclerosis Observatory (OFSEP)¹⁰, a major epidemiological tool on MS for the scientific community, was selected after a call for projects for Cohorts 2010, funded by France's Investment in the Future Program. It is a collaborative project involving over 40 MS research centers in France. The aim of the project is to build and maintain a nationwide cohort of patients with MS and enrich the clinical data with biological samples, socio-economic data, and neuro-images.

¹⁰The OFSEP MS Cohort observatory: <http://www.ofsep.org/en/>.

A dedicated imaging working group is in charge of acquiring, processing, and integrating imaging and derived imaging data into a shared imaging resource center (IRC), and ensuring that the IRC is integrated with clinical databases. The consistent assessment of MRI-based measurements on a large scale requires robust and efficient image processing pipelines. A further goal of this project is to establish an information technology infrastructure enabling audited access to imaging data, as well as a virtual laboratory environment supporting the distributed, synergistic development, validation, and deployment of specialized image analysis procedures developed by different national and international research centers. To ensure easy access to the

Last 2 years ←	09/2013	09/2014	09/2015	05/2016
Users (active)	(53)	118 (52)	178 (66)	206 (99)
Centers	27	31	42	50
Studies	53	60	70	121
Subjects	1706	2228	2833	3243
Examinations	2268	3157	4005	4531
Datasets	-	114 441	151582	176356
Raw data (Dicom)	1 126 GB (1.1 TB)	1 434 GB (1.4 TB)	1 588 GB (1.5 TB)	2 060 GB (2.1 TB)
Processed data (Nifti)	989 GB (0.9 TB)	1 331 GB (1.3 TB)	1 758 GB (1.7 TB)	1 841 GB (1.8 TB)

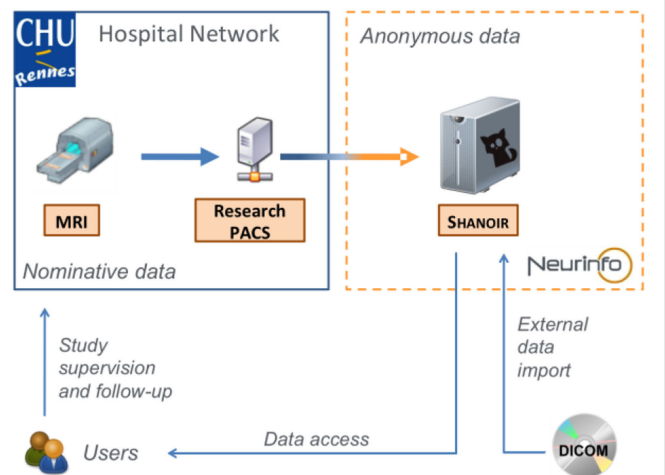


FIGURE 12 | Evolution of the Shanoir@Neurinfo repository Global Statistics (left) and Service Infrastructure (right).

imaging data and allow modifications, queries, annotations, and access control, the Shanoir environment has been selected. It will also ensure interoperability and data management related to the imaging aspect of the cohort (the clinical part is managed by the EDMUS¹¹ system). For this purpose, we have set up a specific Shanoir@OFSEP image repository that is currently in its pilot phase.

Begun in 2012, the Shanoir@OFSEP server was installed to store the imaging data of the OFSEP cohort, which will study the neuroimaging data of 40,000 MS patients over the next 10 years. A consensus has emerged concerning the acquisition protocol, which requires (at least) one brain MRI every 3 years, one spinal MRI every 6 years, i.e., 200,000 MRIs over 10 years. The Shanoir@OFSEP database will grow during this period and beyond (Cotton et al., 2015).

Since OFSEP is a nationwide project covering many patients, many IRCs, and many different kinds of MRI acquisition equipment, a national repository with nationwide access and uniform measures was therefore needed. The OFSEP imaging working group is continuously gathering new acquisition centers volunteering to take part to the cohort. In Shanoir@OFSEP, there are currently about 30 IRCs which include 31 pieces of MRI acquisition equipment representing 14 different MR scanner models from three MR manufacturers (Siemens, Philips, and GE). All the centers are importing data in one main study called the “Mother Cohort.” Each center follows the OFSEP protocol, which will be checked through the quality control module as described in Section “Quality Assessment.” If necessary, derived imaging data can then be imported back to the server in order to refer to potential post-processing information and MS-specific imaging biomarkers to make them available for authorized users.

Currently, the Shanoir@OFSEP repository is hosting five studies: the “Mother Cohort” (200,000 MRIs planned over the next

10 years) as well as four MS imaging clinical research projects. More of these “OFSEP-labeled” clinical research projects or nested cohorts will be integrated in coming years. Everyone can join the “Mother Cohort” study as long as they use the OFSEP protocol. One can also ask the OFSEP to contribute to the project through his study as soon as the PI presents his research study subject to the OFSEP scientific committee that can grant (or not) the hosting. Data hosted on Shanoir@OFSEP will remain confidential (private) throughout the duration of the study but can be made available to all researchers through a specific OFSEP application.

CONCLUSION AND PERSPECTIVES

The Shanoir SaaS manages the sharing of distributed information sources in neuroimaging over the Internet, whether these resources are located in centers of experimentation, clinical departments in neurology, or research centers in cognitive neurosciences or image processing. Through the description of two repositories that administer a Shanoir environment (Neurinfo and OFSEP), we have illustrated how a large variety of users can diffuse, share, or access neuroimaging information between peers almost as easily as if the data were stored at their local hospital, research lab, or company. Through the description of the Shanoir software environment, we have illustrated how neuroimaging data can be structured, managed, archived, visualized, and shared.

In the medium term, we plan to integrate Shanoir’s resources and services with the open community through the French National Infrastructure’s “France Life Imaging” (FLI),¹² and more specifically, the “Information Analysis and Management” (IAM) node that is dedicated to provide large scale IT infrastructure for *in vivo* imaging. For this purpose, the FLI-IAM node will

¹¹EDMUS: <http://www.edmus.org>.

¹²France Life Imaging <https://www.francelifeimaging.fr> with the IAM node (<https://project.inria.fr/fli/>) is a national infrastructure for *in vivo* imaging.

build and operate an infrastructure to store, manage, and process *in vivo* imaging data from human or preclinical procedures. The main achievements of the IAM node will consist of a versatile software platform composed of several subcomponents that will connect hardware and software facilities to build:

- an archiving and management infrastructure of *in vivo* images as well as provide solutions to process and manage the acquired data through dedicated software and hardware solutions;
- versatile image analysis and data management solutions for *in vivo* imaging to facilitate interoperability between production sites and users and provide heterogeneous and distributed storage solutions for raw and metadata indexing (e.g., through the use of semantic models).

As such, we are under integrating Shanoir as one of the data management solutions of the FLI-IAM facilities along with a collection of companion data management software platforms such as CATI-DB¹³ and ArchiMed, or a collection of processing clients or high-performance computing workflow facilities such as medInria, BrainVisa,¹⁴ and the VIP platform.¹⁵ For this purpose, within FLI-IAM, we are setting up the “glue” between these platforms that will make it possible to connect and interoperate between them. In addition, through FLI-IAM, we will provide the necessary information for additional resources to join the

FLI-IAM infrastructure by defining the basic conformal statement that will make the technology and scalability of FLI-IAM possible.

Nonetheless, as described in Section “Introduction,” there are a lot of similar initiatives going on recently in the medical imaging research field, such as Human Brain Project, ADNI, XNAT-based solutions, etc. For these initiatives, as well as for Shanoir, the goal is to share the data at a large extent. This cannot be done without a significant additional effort on standardization in the field and on interoperability between software platforms addressing similar services. This is what motivates the integration of Shanoir in the French FLI-IAM e-infrastructure initiative. This challenge of tomorrow is to continue this effort at the international level.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

This work has been supported by Inria under a “technological development program” grant for the Neurinfo platform from the Brittany regional council and the EU-Feder program and two grants provided by the French government through the Agence Nationale de la Recherche as part of the Investment in the Future Program under references ANR-10-COHO-002 (for OFSEP) and ANR-11-INBS-006 (for FLI).

REFERENCES

- Ashish, N., Ambite, J. L., Muslea, M., and Turner, J. A. (2010). Neuroscience data integration through mediation: an (F)BIRN case study. *Front. Neuroinformatics* 4:12. doi:10.3389/fninf.2010.00118
- Barillot, C., Amsaleg, L., Aubry, F., Bazin, J.-P., Benali, H., Cointepas, Y., et al. (2003). *Neurobase: Management of Distributed Knowledge and Data Bases in Neuroimaging. Human Brain Mapping*. New York, NY: Academic Press, 726.
- Barillot, C., Benali, H., Dojat, M., Gaignard, A., Gibaud, B., Kinkingnehun, S., et al. (2006). Federating distributed and heterogeneous information sources in neuroimaging: the neurobase project. *Stud. Health Technol. Inform.* 120, 3–13.
- Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P., and Evans, A. C. (2012). The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Front. Neuroinform.* 6:7. doi:10.3389/fninf.2012.00007
- Benkner, S., Arbona, A., Berti, G., Chiarini, A., Dunlop, R., Engelbrecht, G., et al. (2010). @neurIST: infrastructure for advanced disease management through integration of heterogeneous data, computing, and complex processing services. *IEEE Trans. Inf. Technol. Biomed.* 14, 1365–1377. doi:10.1109/TITB.2010.2049268
- Book, G. A., Anderson, B. M., Stevens, M. C., Glahn, D. C., Assaf, M., and Pearlson, G. D. (2013). Neuroinformatics Database (NiDB) – a modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics* 11, 495–505. doi:10.1007/s12021-013-9194-1
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Confidence and precision increase with high statistical power. *Nat. Rev. Neurosci.* 14, 585–586. doi:10.1038/nrn3475-c4
- Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* 63, 289–300. doi:10.1016/j.neuroimage.2012.07.004
- Cotton, F., Kremer, S., Hannoun, S., Vukusic, S., Dousset, V., and The Imaging Working Group of OFSEP. (2015). OFSEP, a nationwide cohort of people with multiple sclerosis: consensus minimal MRI protocol. *J. Neuroradiol.* 42, 133–140. doi:10.1016/j.neurad.2014.12.001
- Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., and Evans, A. C. (2011). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinform.* 5:37. doi:10.3389/fninf.2011.00037
- Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., et al. (2010). Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PLoS ONE* 5:e13070. doi:10.1371/journal.pone.0013070
- Evans, A. C., and Brain Development Cooperative Group. (2006). The NIH MRI study of normal brain development. *Neuroimage* 30, 184–202. doi:10.1016/j.neuroimage.2005.09.068
- Glatard, T., Rousseau, M.-E., Camarasu-Pop, S., Rioux, P., Sherif, T., Beck, N., et al. (2014). Interoperability between the CBRAIN and VIP web platforms for neuroimage analysis. *Front. Neuroinformatics*. doi:10.3389/conf.fninf.2014.18.00070
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5:13. doi:10.3389/fninf.2011.00013
- Gupta, A., Ludaescher, B., Martone, M., and Rajasekar, A. (2003). *BIRN-M: A Semantic Mediator for Solving Real-World Neuroscience Problems. ACM SIGMOD 2003*. San Diego, CA: ACM, 678.
- Hall, D., Huerta, M. F., McAuliffe, M. J., and Farber, G. K. (2012). Sharing heterogeneous data: the national database for autism research. *Neuroinformatics* 10, 331–339. doi:10.1007/s12021-012-9151-4
- Hibar, D. P., Stein, J. L., Renteria, M. E., Arias-Vasquez, A., Desrivieres, S., Jahanshad, N., et al. (2015). Common genetic variants influence human subcortical brain structures. *Nature* 520, 224–229. doi:10.1038/nature14101
- Ioannidis, J. P. (2014). How to make more published research true. *PLoS Med.* 11:e1001747. doi:10.1371/journal.pmed.1001747
- Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection,

- prevalence, and prevention. *Trends Cogn. Sci.* 18, 235–241. doi:10.1016/j.tics.2014.02.010
- Jack, C. R. Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi:10.1002/jmri.21049
- Keator, D. B., Grethe, J. S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., et al. (2008). A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Trans. Inf. Technol. Biomed.* 12, 162–172. doi:10.1109/TITB.2008.917893
- Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi:10.1016/j.neuroimage.2013.05.094
- Keator, D. B., Wei, D., Gadde, S., Bockholt, J., Grethe, J. S., Marcus, D., et al. (2009). Derived data storage and exchange workflow for large-scale neuroimaging analyses on the BIRN grid. *Front. Neuroinformatics* 3:30. doi:10.3389/neuro.11.030.2009
- Kennedy, D. N., Haselgrove, C., Riehl, J., Preuss, N., and Buccigrossi, R. (2015). The three NITRCs: a guide to neuroimaging neuroinformatics resources. *Neuroinformatics* 13, 383–386. doi:10.1007/s12021-015-9263-8
- Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., et al. (2013). Human Connectome Project informatics: quality control, database services, and data visualization. *Neuroimage* 80, 202–219. doi:10.1016/j.neuroimage.2013.05.077
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi:10.1385/NI.5:11
- Masolo, C., Borgo, S., Gangemi, A., Guarino, N., and Oltramari, A. (2003). *WonderWeb Deliverable D18, Ontology Library (Final)*. Technical Report. Trento: LOA-ISTC, CNR.
- Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: theory and rationale for its development. The International Consortium for Brain Mapping (ICBM). *Neuroimage* 2, 89–101. doi:10.1006/nimg.1995.1012
- Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* 34, 1993–2024. doi:10.1109/TMI.2014.2377694
- Michel, F., Gaignard, A., Ahmad, F., Barillot, C., Batrancourt, B., Dojat, M., et al. (2010). Grid-wide neuroimaging data federation in the context of the NeuroLOG project. *Stud. Health Technol. Inform.* 159, 112–123.
- Ooi, C., Bullmore, E. T., Wink, A.-M., Sendur, L., Barnes, A., Achard, S., et al. (2009). CamBAfx: workflow design, implementation and application for neuroimaging. *Front. Neuroinformatics* 3:27. doi:10.3389/neuro.11.027.2009
- Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi:10.1038/nn.3818
- Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Front. Neuroinform.* 6:9. doi:10.3389/fninf.2012.00009
- Rimal, B. P., Choi, E., and Lumb, I. (2009). “A taxonomy and survey of cloud computing systems. INC, IMS and IDC, 2009. NCM'09,” in *Fifth International Joint Conference on IEEE* (Seoul: IEEE), 44–51.
- Roland, P. E., and Zilles, K. (1994). Brain atlases – a new research tool. *Trends Neurosci.* 17, 458–467. doi:10.1016/0166-2236(94)90131-7
- Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., et al. (2011). COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. *Front. Neuroinform.* 5:33. doi:10.3389/fninf.2011.00033
- Shepherd, G. M., Mirsky, J. S., Healy, M. D., Singer, M. S., Skoufos, E., Hines, M. S., et al. (1998). The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data. *Trends Neurosci.* 21, 460–468. doi:10.1016/S0166-2236(98)01300-9
- Sherif, T., Rioux, P., Rousseau, M. E., Kassis, N., Beck, N., Adalat, R., et al. (2014). CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* 8:54. doi:10.3389/fninf.2014.00054
- Styner, M., Lee, J., Chin, B., Chin, M. S., Commowick, O., Tran, H., et al. (2008). 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. *Midas J.* 11. Available at: <http://www.midasjournal.org/browse/publication/638>
- Temal, L., Dojat, M., Kassel, G., and Gibaud, B. (2008). Towards an ontology for sharing medical images and regions of interest in neuroimaging. *J. Biomed. Inform.* 41, 766–778. doi:10.1016/j.jbi.2008.03.002
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn Human Connectome Project: an overview. *Neuroimage* 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041
- Van Horn, J. D., and Gazzaniga, M. S. (2013). Why share data? Lessons learned from the fMRIDC. *Neuroimage* 82, 677–682. doi:10.1016/j.neuroimage.2012.11.010
- Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., et al. (2001). 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.* 356, 1323–1339. doi:10.1098/rstb.2001.0916
- Walport, M., and Brest, P. (2011). Sharing research data to improve public health. *Lancet* 377, 537–539. doi:10.1016/S0140-6736(10)62234-9
- Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2012). The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimers Dement.* 8, S1–S68. doi:10.1016/j.jalz.2011.09.172

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Barillot, Bannier, Commowick, Corouge, Baire, Fakhfakh, Guillaumont, Yao and Kain. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



BIOMIST: A Platform for Biomedical Data Lifecycle Management of Neuroimaging Cohorts

Marianne Allanic¹, Pierre-Yves Hervé^{2,3,4}, Cong-Cuong Pham⁵, Myriam Lekkal^{1,2,3,4,5}, Alexandre Durupt⁶, Thierry Brial¹, Arthur Grioche¹, Nada Matta⁶, Philippe Boutinaud¹, Benoit Eynard⁶ and Marc Joliot^{2,3,4*}

¹CADESIS Group, Courbevoie, France, ²GIN, IMN, UMR 5293, CNRS, Bordeaux, France, ³GIN, IMN, UMR 5293, Université de Bordeaux, Bordeaux, France, ⁴GIN, IMN, UMR 5293, CEA, Bordeaux, France, ⁵Sorbonne Universités, Université de Technologie de Compiègne, Department of Mechanical Systems Engineering UMR7337 Roberval CNRS, Compiègne, France, ⁶Université de Technologie de Troyes, ICD, UMR CNRS 6281, Troyes, France

The data management needs of the neuroimaging community are currently addressed by several specialized software platforms, which automate repetitive data import, archiving and processing tasks. The BIOMedical Imaging Semantic data management (BIOMIST) project aims at creating such a framework, yet with a radically different approach: the key insight behind it is the realization that the data management needs of the neuroimaging community—organizing the secure and convenient storage of large amounts of large files, bringing together data from different scientific domains, managing workflows and access policies, ensuring traceability and sharing data across different labs—are actually strikingly similar to those already expressed by the manufacturing industry. The BIOMIST neuroimaging data management framework is built around the same systems as those that were designed in order to meet the requirements of the industry. Product Lifecycle Management (PLM) systems rely on an object-oriented data model and allow the traceability of data and workflows throughout the life of a product, from its design to its manufacturing, maintenance, and end of life, while guaranteeing data consistency and security. The BioMedical Imaging—Lifecycle Management data model was designed to handle the specificities of neuroimaging data in PLM systems, throughout the lifecycle of a scientific study. This data model is both flexible and scalable, thanks to the combination of generic objects and domain-specific classes sourced from publicly available ontologies. The data integrated management and processing method was then designed to handle workflows of processing chains in PLM. Following these principles, workflows are parameterized and launched from the PLM platform onto a computer cluster, and the results automatically return to the PLM where they are archived along with their provenance information. Third, to transform the PLM into a full-fledged neuroimaging framework, we developed a series of external modules: DICOM import, XML form data import web services, flexible graphical querying interface, and SQL export to spreadsheets. Overall, the BIOMIST platform is well suited for the management of neuroimaging cohorts, and it is currently used for the management of the BIL&GIN dataset (300 participants) and the ongoing magnetic resonance imaging-Share cohort acquisition of 2,000 participants.

Keywords: data management, neuroscience, neuroimaging, provenance, product lifecycle management, workflow

OPEN ACCESS

Edited by:

Michel Dojat,
INSERM, France

Reviewed by:

Alex Pappachen James,
Nazarbayev University, Kazakhstan
Camille Maumet,
University of Warwick, UK

*Correspondence:

Marc Joliot
marc.joliot@u-bordeaux.fr

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 01 September 2016

Accepted: 22 December 2016

Published: 30 January 2017

Citation:

Allanic M, Hervé P-Y, Pham C-C,
Lekkal M, Durupt A, Brial T,
Grioche A, Matta N, Boutinaud P,
Eynard B and Joliot M (2017)
BIOMIST: A Platform for Biomedical
Data Lifecycle Management of
Neuroimaging Cohorts.
Front. ICT 3:35.
doi: 10.3389/fict.2016.00035

INTRODUCTION

Provenance Complexity in Neuroimaging Studies

Cognitive neuroscience is multidisciplinary “by its very nature” (Van Horn et al., 2001) and relies on a large set of complementary approaches for probing brain function and behavior. Different combination of methods, such as computerized experimental psychology, magnetic resonance imaging (MRI), electro and magneto encephalography (EEG/MEG), functional near-infrared spectroscopy, eye tracking, genetics, etc., can be used during a scientific project and require an active interaction between many specialties—physics, medicine, mathematics, and engineering among others. Resulting data are complex, and neuroscience researchers have to deal with many data sources, natures, and types of processing (Goble and Stevens, 2008).

One can only expect the heterogeneity of the tools and data formats involved in research to increase over time. With more and more studies—neurogenetic, neuroepidemiology, and longitudinal—requiring large cohorts and therefore producing huge amounts of data in a multicentric context. Besides, these large-scale studies may need to be aggregated into meta-analyses to reach the adequate level of statistical power, given the staggering number of hypotheses being tested. This implies the frequent reuse of pre-existing data, for validation of new findings. In addition, the high cost of data (both acquisition and processing) and the need for reproducibility make data reuse and sharing a necessity (Yarkoni et al., 2010; Poline et al., 2012).

The information of what a piece of data is, when, where, and how it was produced, why and for whom it was performed is called *provenance*—the origin and history of a set of data (Simmhan et al., 2005). The provenance in BioMedical Imaging studies is complex: acquisition devices and parameters impact raw data, processing algorithm, parameters, and tools impact on derived data, processing input traceability is intricate. All this information is required to be able to reproduce scientific results and also to share data and understand how specific data were obtained.

Sharing study data between scientific teams—inside and outside the institutions that produced the data—implies to ensure consistency of data and their provenance on one side, and data security on the other side, particularly on studies involving human subjects.

The lifecycle of a study can be described by four stages: (1) study specifications define the purpose of the study, what data will be acquired, stored, and analyzed, (2) raw data are acquired with appropriate devices and following protocols, (3) derived data are generated from raw data by analytical means, and (4) results are published and the data may be shared with the community. **Figure 1** summarizes the links between the stages with examples of data at each stage along with required provenance information.

Existing Systems for the Management of the Provenance of Neuroimaging Studies

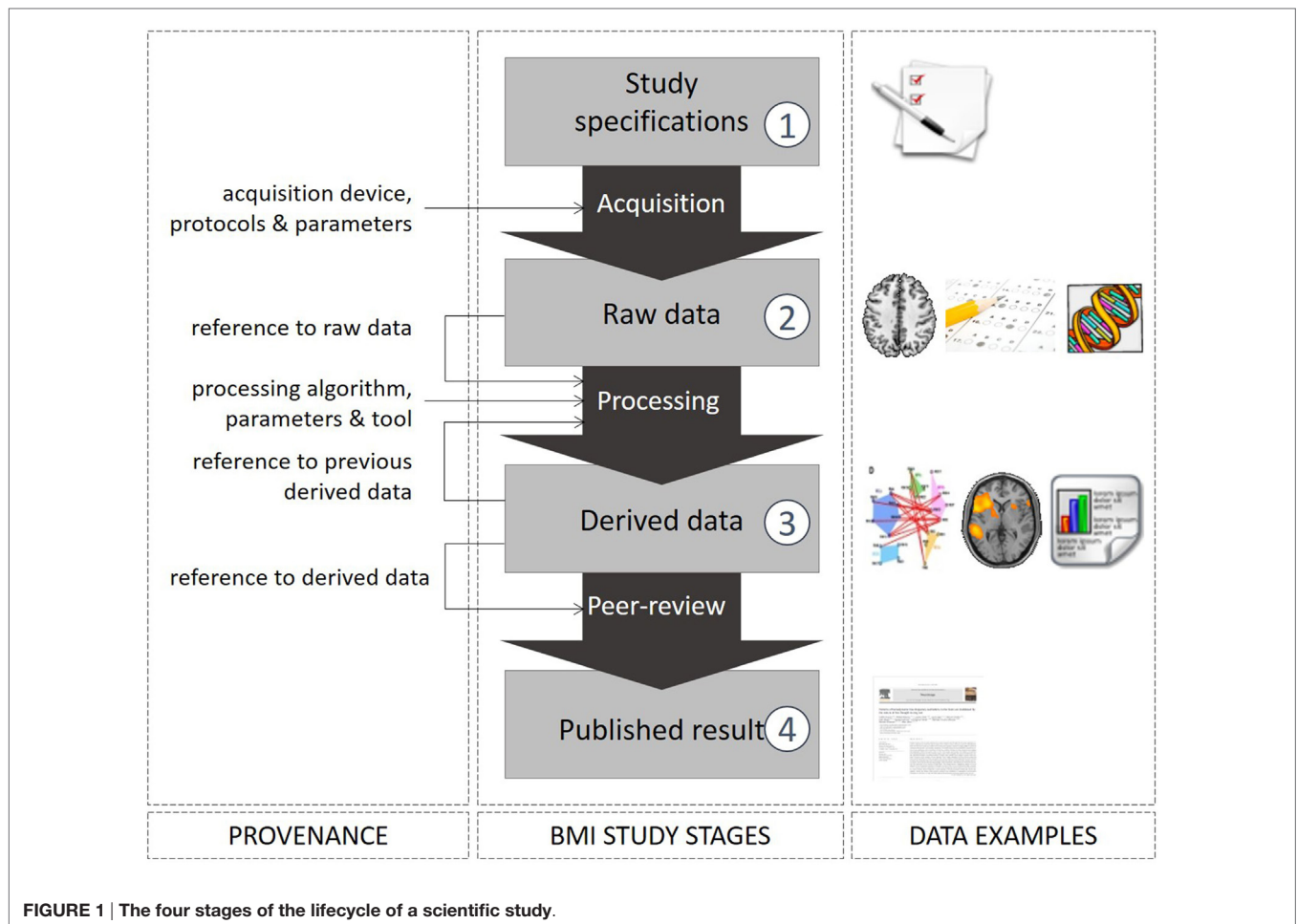
So far, this challenging need for neuroscience data sharing has been met by the emergence of dedicated systems, especially for modalities that were made affordable to researchers because they

were so widely used in hospitals, and this chiefly applies to MRI. In this case, the best solution was to build upon the pre-existing medical standard, namely, Digital Imaging and COmmunications in Medicine (DICOM), with the development of research-dedicated picture and archival communication systems (PACS). Compared with traditional clinical PACS, neuroimaging data management systems can manage research projects involving large sets of subjects instead of being confined to the individual patient, storing data from other sources than DICOM entities and controlling access to the data in a fine-grained way. They also include procedures to clean patient health information from the data to comply with human research ethical norms, visual and/or automated quality control procedures, and are capable of interacting with computing clusters or workflow managers for data processing.

Existing neuroimaging data management systems so far—XNAT (Marcus et al., 2007a), LORIS (Das et al., 2012), COINS (Scott et al., 2011), IDA (Crawford et al., 2016), MIDAS (Kitware Inc.), HID (Keator et al., 2016), NIDB (Book et al., 2013), SHANOIR (Barillot et al., 2015), etc.—were implemented using the standard web technologies, in the form of J2EE or PHP web applications, with a browser-based graphical frontend and a relational database backend, and some also provide means to automate interactions through application programming interfaces (APIs; REST or SOAP). Such web systems leverage DICOM libraries such as dcm4che or DICOM toolkit to implement at least a DICOM receiver and offer separate upload services for non-DICOM data, over HTTP. This scalable web architecture makes it possible to serve brain imaging and associated data to distant users over the web or store data in the cloud, as best exemplified with XNAT at the Human Connectome Project. Naturally, with this multiplication of like-minded, yet idiosyncratic web applications for neuroimaging data management, came the need for database federation and interoperability, and for a common lexicon across different systems, such as shared ontologies (Gupta et al., 2008).

A detailed comparison of 18 neuroimaging data management systems is presented in (Allanic et al., 2017). Criteria of comparison are:

- Type of managed data: which disciplines (imaging, genetics, psychology, clinical, etc.) and which level of data (raw, derived, and published) can be managed in the system. Most of the existing data management systems focuses on one or two levels (raw and derived or derived and published) and most of them manage only imaging data [except HIS (Keator et al., 2009), LORIS (Das et al., 2011), XNAT (Marcus et al., 2007b), and fMRIDC (Van Horn et al., 2001)].
- Provenance strategy: how is the provenance described and made available to enable data sharing and reuse. It appears that data provenance is sometimes more precise and complete in systems managing published results, as users must provide additional metadata that describe how data were produced to be allowed to submit their data (Fox et al., 2005); openfMRI (Poldrack et al., 2013), fMRIDC, and BrainMap (Fox and Lancaster, 2002) are good examples.
- Data model flexibility: how the system can be adapted to new types of data, new protocols. Few data management systems



allow to customize their data model; among them REDCap, COINS (Scott et al., 2011), XNAT, CVT (Gerhard et al., 2011), NiDB (Book et al., 2013), DFBIdb (Adamson and Wood, 2010), and Neurolog (Dojat et al., 2011).

- Integration of processes and existing tools: how pipelines, quality workflow, and visualization software can be integrated to the system. Some neuroimaging data management systems allow to launch pipelines and to visualize results directly from the database interface.

There is to our knowledge no existing data management system that allows to manage and to analyze study data from study specifications to publication; we aim at providing such an environment.

Product Lifecycle Management (PLM) Systems: A Key to Provenance Management

The main assumption in our work is to reuse a proven data management system designed for manufacturing industry to the management of data from neuroimaging studies at every stage, ensuring full provenance.

Regarding data management, the manufacturing industry is confronted with the same issues as neuroimaging: heterogeneous

product data must be tracked throughout the product lifecycle—product requirement, design, manufacturing, maintenance, and end of life. Products are made from the collaboration of multi-disciplinary teams, not always working on the same site. PLM system has been designed since the 1990s to answer the needs of the manufacturing industry and enable the storage, versioning, and collaborative work on computer-aided design (CAD) data, with a strong focus on traceability. The aim of PLM systems could be summarized by providing the right data at the right person and at the right moment: they facilitate collaborative and concurrent work, in addition to multi-sites data sharing, answering the imperative need to exchange data seamlessly between various geographic locations within a worldwide company (Kiritzis et al., 2003; Terzi et al., 2010).

Although the design of PLM software is not oriented toward neuroimaging data, or any kind of scientific data in particular, their inherent properties make them a very compelling IT solution for scientific laboratories, and neuroimaging labs in particular (Allanic et al., 2017).

Outlines of the Paper

We present in the paper the BIOMedical Imaging SemanTic data management (BIOMIST) platform, whose aim is to respond to the need of data management, sharing, reuse, and reproducibility

of the neuroimaging domain by ensuring automated provenance tracking throughout the lifecycle of a study and access to analysis software in a unique environment.

The targets of the BIOMIST platform are new neuroimaging studies from small (100 subjects) to medium (5,000 subjects) cohort, with multimodal, longitudinal, and multi-sources acquisitions requiring complex pipelines, quality controls, and efficient access management.

Section “Design: The BIOMIST Platform” presents the BIOMIST platform and the integration of its components. The technical details of the implementation of the platform are developed in Section “Implementation.” The benefits of the platform were tested on the BIL&GIN dataset and the I-Share study: results are presented in Section “Application.” This paper closes with a discussion and leads for future work toward the BIOMIST platform in Section “Discussion.”

DESIGN: THE BIOMIST PLATFORM

This section presents the BIOMIST platform, whose purpose is to manage heterogeneous data of neuroimaging cohorts, from study specifications to published results, in order to ensure data reproducibility, sharing, and reuse. Section “Design Method” explains our design method, and then sections “Key Principles of PLM,” “The BMI-LM Data Model to Manage Data and Provenance,” “Mapping Strategy for Data Import,” “The DIMP Method for Integration of Processing Pipelines,” and “Querying Strategies” develop the characteristics of each component of the platform: the core PLM system is customized by the BioMedical Imaging—Lifecycle Management (BMI-LM) data model, data are imported into the PLM thanks to mapping strategies and processed with the data integrated management and processing (DIMP) method, to end with, users query data managed by the PLM through two interfaces, graphical and Open Database Connectivity (ODBC). **Figure 2** shows the integration of the components of the BIOMIST platform.

Design Method

To understand the concerns of daily neuroimaging research work and the associated data management issues, we studied the literature and interviewed the staff of a representative neuroimaging laboratory (GIN, from the University of Bordeaux, France). Ongoing projects at this laboratory rely on structural and functional MRI acquisitions performed over hundreds of participants, as well as smaller scale task-based functional MRI projects. Over the 2006–2009 period, this group designed its own relational database (GINdb, based on SQL technology) in order to manage experiments: processing data, subject data and paths to files stored on disks of their IT system (Joliot et al., 2009).

Eleven members of the research group (eight tenured researchers, two research engineers, and one *post doc*) were interviewed, by small groups of two or three people to avoid group effects. They were asked to express their needs: what was missing in GINdb and what would be their ideal system. They mainly highlighted that the data model should feel natural for the users, especially regarding the queries, and that it should be flexible enough to allow future changes. Besides, they would like to launch analyses

batch directly from the database and to label data with one or several statuses, such as “valid exam” or “checked data.”

From these interviews and the review of the literature, four main axes are defined:

1. Provenance: manages all the data generated during a study, from its specifications to published results, and track the associated provenance to be able to share and reuse data optimally. The PROVenance Data Model (PROV-DM) standard is developed by the World Wide Web (W3) consortium to help exchanging data, a main objective is to comply with it.
2. Heterogeneity: accepts all data formats and manage the concepts of the disciplines involved in a neuroimaging study.
3. Integration: allows automated data import, processing launch, data analysis, and visualization from the platform.
4. Flexibility: allows data model changes without consequences on existing data to handle new data format, as well as semantic changes, evolution of acquisition protocols.

To validate the resulting BIOMIST platform, we tested it with two use cases from the GIN: (1) the 300 subjects BIL&GIN and (2) the I-Share study. Results are presented in Section “Application.”

Key Principles of PLM

PLM systems supports multisite sharing and collaborative work, by managing product data throughout its lifecycle along with advanced access management features that guarantee data security and with file and database replication mechanisms that allows multisite collaboration even through low latency or low-bandwidth networks.

Product Lifecycle Management systems do not only manage data (i.e., documents/files + metadata) but concepts, thanks to its object-oriented data model. Concepts at every phase of the product lifecycle are represented by objects instantiated as *items* whose versions are tracked. *Items* can be classified with a fully flexible hierarchy of concepts and vocabulary. Any kind of file types and formats are allowed and are stored in objects called *datasets*. Every event on an item is tracked: it is possible to know who created, modified, updated or validated it, when and why. Automated or manual workflows can be launched by users from the system; these workflows can be customized and can be used to implement a process with validation from several users (e.g., validating an acquired dataset) and to perform automated actions on *items* or *datasets* (create new version, add status, update metadata, comment, classify, etc.). A typical application in manufacturing industry would be a workflow that follows validations of a design change in a product. Query facilities complete the features of PLM systems: queries can be customized, both to retrieve items and datasets and to generate reports. Data can be accessed from the web and visited directly into the PLM interface, as soon as a suitable visualization software is integrated, or downloaded on users' computer, automatically opened in the right software. For managing large set of data, the PLM infrastructure includes various replication strategies that enables access to sites that may have low latencies or low-bandwidth network connections.

Data security is ensured in PLM systems through their infrastructure and an advanced module for access management.

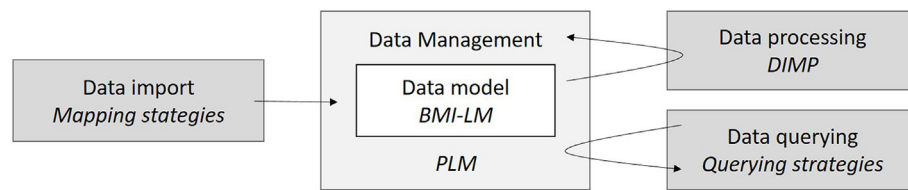


FIGURE 2 | The components of the BIOMedical Imaging Semantic data management platform and their integration, arrows indicating the direction of data flows. Each component is developed in indicated subsections.

The infrastructure of PLM systems is composed of four tiers (resource, enterprise, web application, and client tiers) that are presented in **Figure 3**. In resource tier, the SQL database manages data instances and metadata, and one or several volumes contain data files that may be encrypted according to users' needs. This organization implies that data (files and instances) can only be accessed through a client who ensures data consistency. An account is required to connect to the client: users are associated to *roles* and belong to *group* and *projects*, which determine their level access to the data stored in the PLM system (none, read, write, export, promotion, validation, etc.).

The compatibility of PLM features with the four axes required for neuroimaging data management—that were highlighted in Section “Design Method”—is presented in **Table 1**. The basic features of PLM systems allow (1) to fulfill context and traceability of the provenance axis, (2) to manage every data types and formats, which fulfill part of the heterogeneity axis, and (3) the integration of visualization software and the possibility to connect to external software, web services, etc. These features do not cover all parts of the perimeter of the four axes. Therefore, we developed a data model to complete provenance, heterogeneity, and flexibility axes, as the data model of a PLM system can be easily modified.

The BMI-LM Data Model to Manage Data and Provenance

The stages of a neuroimaging study can be modeled as a cycle that constitutes the lifecycle of a research study, from study specifications to published results (see **Figure 1**).

First, the BMI-LM developed for the BIOMIST platform is presented from its two aspects: generic objects (see Generic Objects to Manage Heterogeneity) and specific classes (see Specific Classes to Bring Flexibility). To end with, the BMI-LM data model is compared with PROV-DM specifications (see Conceptual Equivalence Between the BMI-LM Data Model and the PROV-DM Standard).

Generic Objects to Manage Heterogeneity

The BMI-LM data model is composed of generic objects representing concepts related to a study. The 17 generic concepts (see **Table 2** below) are divided into three categories:

1. *Definition objects*: they described how *result objects* were obtained and can be reused from one study to another. They are part of the provenance strategy.

2. *Result objects*: they store data of a study, raw and derived, in shape of datasets (files) and metadata.
3. *Ambivalent objects*: depending on the context, these objects can be used as a *definition object* or a *result object*. They are part of the provenance strategy.

The generic objects are presented in **Table 3** according to their category and their stage in the study lifecycle. **Figure 4** presents a UML model of BMI-LM with the relationships between objects and related cardinalities.

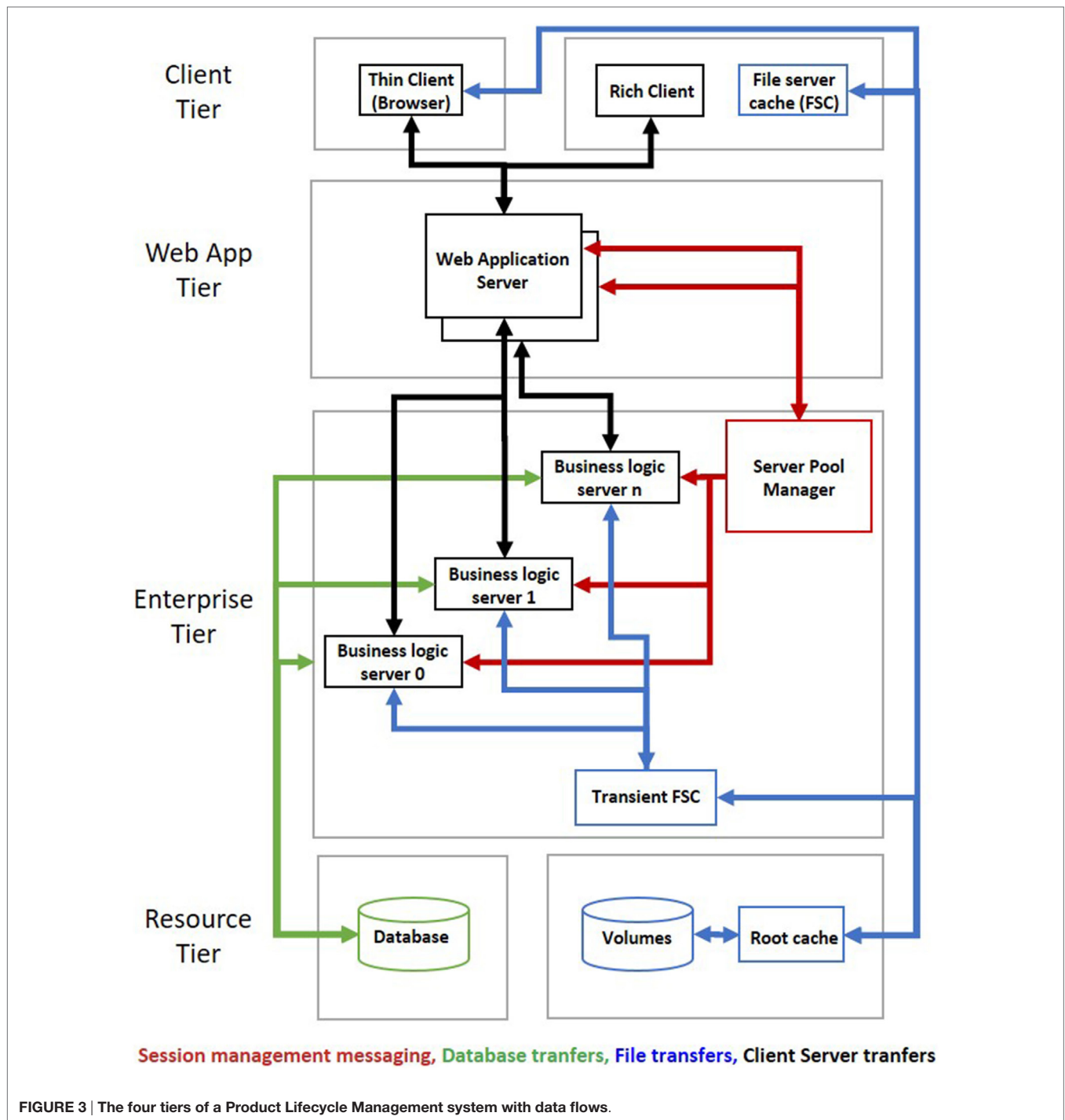
Specific Classes to Bring Flexibility

To enable flexibility in the semantic definition of the objects, “classes” may be associated with instances of the data model. A class in the context of the PLM system is a name (hopefully with a meaning for the end user: names were issued from ontologies of the application domain, see Section “Domain Classification for Neuroimaging”). A class has typed attributes that allows values to be associated with items. All the classes are organized in a standard inheritance hierarchy tree and attributes are inherited. Every item of the BMI-LM data model can be classified, and the root structure of the classification is organized by object categories: *definition branch*, *result branch*, and *ambivalent branch*, which are themselves divided into subcategories. The classes play the role of subtypes of objects; for example, an *exam result* object can be classified as an imaging, psychology, or genetic examination.

The different domains involved in neuroimaging studies do not use the same vocabularies, as well as acquisitions and processing tools. Such information is stored in the attributes of the classes, so a classification is domain dependent. The highest level of the classification (the main categories) will be used in every deployment of BMI-LM, the lower-level branches may be deployed where needed; and new classes/attributes may be easily created.

Conceptual Equivalence between the BMI-LM Data Model and the PROV-DM Standard

A representation of provenance is proposed by the World Wide Web (W3) consortium, who develop standards to support the expansion of the web. According to the PROV-DM standard, the provenance is defined “as a record that describes people, institutions, entities and activities involved in producing a piece of data or thing in the world” (Moreau and Missier, 2013). An entity can be physical, numeric, or conceptual. An activity occurs on a time period and act with or on one or many entities. This includes



consumption, processing, transformation, modification, using, or generation of entities. An agent is responsible in the execution of an activity. Entities, activities, and agents are modeled by seven relationships, which are given in **Figure 5A**.

Figure 5B shows how the BMI-LM data model and the PROV-DM standard are equivalent in a conceptual way: result objects are entities, definition objects are activities and some PLM features (users, workflows) are agents.

Mapping Strategy for Data Import

The strategy for data import is essential to ensure that the BIOMIST platform will be integrated as a study data management tool. Import processes must stay flexible and easy enough for any data format or acquisition process. In order to set up automatically the provenance, a mapping between the data to import and the data model of the platform must be efficient. First, we present two key principles of our mapping strategy to import

TABLE 1 | Features of Product Lifecycle Management (PLM) systems and the BioMedical Imaging—Lifecycle Management (BMI-LM) data model against the four axes required for the management of neuroimaging studies.

	Provenance	Heterogeneity	Integration	Flexibility
PLM	Context (PROV:Agents) Traceability (PROV:Entity)	Data types Formats	Visualization software	
BMI-LM	Identification (PROV:Activity)	Multidisciplinary		Evolution of research protocols Integration of new disciplines

The compliance with PROVenance Data Model standard is indicated for the provenance axis.

TABLE 2 | Generic objects of the BioMedical Imaging—Lifecycle Management (BMI-LM) data model.

Generic object	Definition
Acquisition result	Indivisible period of data acquisition
Acquisition definition	Description of an acquisition protocol
Acquisition device	Description of the device used during an examination
Bibliographical reference	Published paper
Data unit result	Single acquired piece of data
Data unit definition	Definition of a piece of data
Exam result	Continuous line of acquisitions
Exam definition	Examination protocol
Processing result	Instance of a processing chain
Processing definition	Definition of a processing chain
Processing unit result	Derived data
Processing unit definition	Definition of a processing to compute derived data
Processing parameters	Set of parameters of a processing unit
Reference data	Pattern computed from derived data
Software tool	Description of a piece of software used to compute derived data
Study	Research study
Study subject	Subject in the context of a study
Subject	Unique subject in the database
Subject group	Group of study subjects

TABLE 3 | Generic objects of the BioMedical Imaging—Lifecycle Management data model according to study stages and categories.

Study stages	Definition objects	Result objects	Ambivalent objects
Specification		Study	
Raw data	Subject Exam definition Acquisition definition Data unit definition Acquisition device	Study subject Exam result Acquisition result Data unit result	
Derived data	Processing definition Processing unit definition Processing parameters Software tool	Processing result Processing unit result	Reference data Subject group
Published results			Bibliography reference

data, and then, this strategy is exemplified for the import of form and DICOM data.

Key Mapping Principles

To import data with complete provenance, its context must be known—at least the project and the subject it belongs to, its future

owner—and its definition. For the BIOMIST platform, it means that the PLM system must know what kind of item to create (result item), how to classify it, and how to link it with existing items in the database (definition items and other result items).

Our strategy is to define an XML structure to map imported data and its associated metadata to an item of the data model, a class associated with the item and class attributes. An example of XML mapping is given as Part S1 in Supplementary Material: a DICOM series is imported as a data unit in an existing exam and in a new acquisition.

The XML mapping file is associated to definition items (e.g., an exam definition item since this particular mapping is specific to this examination protocol), with two objectives in mind: to understand how the data was imported and to reuse the mapping for another study.

Form Data Import

A form is a set of simply typed data (set of answers, tracings, parameters, etc.) that needs to be acquired for every subject in a study. For instance, it may be the result of a behavioral survey, or an electronic case report form. The definition of the form is an *Acquisition Definition* item, and the questions are defined by *Data Unit Definition* items. Therefore, the result of the import of a form for a subject is an *Acquisition Result* item with all the related *Data Unit Result* (the answer by a subject to a question).

DICOM Import

Digital Imaging and Communications in Medicine (DICOM) is a worldwide used protocol for exchanging data between imaging modalities, archival systems, and visualization workstations (Mildenberger et al., 2002). A DICOM instance usually contains images to which is associated a series of attributes (tags), selected from a dictionary described in part three of DICOM standard specifications. The standard tags that are used by imaging devices to store modality-specific imaging parameters, patient, institution, and device information, as well as date and time information. Beside the standard fields, the DICOM standard allows for proprietary fields in dedicated parts of the DICOM header. A same DICOM tag will not have the same meaning depending on the vendor, and vendor-specific dictionaries are required. Our mapping strategy allows tackling this issue as the definition of import mapping from DICOM attributes dictionary to BIOMIST classification attributes dictionary can be adjusted for every exam definition if needed.

A basic mapping between equivalent concepts of the DICOM and the BMI-LM data model is given in **Table 4**. The main

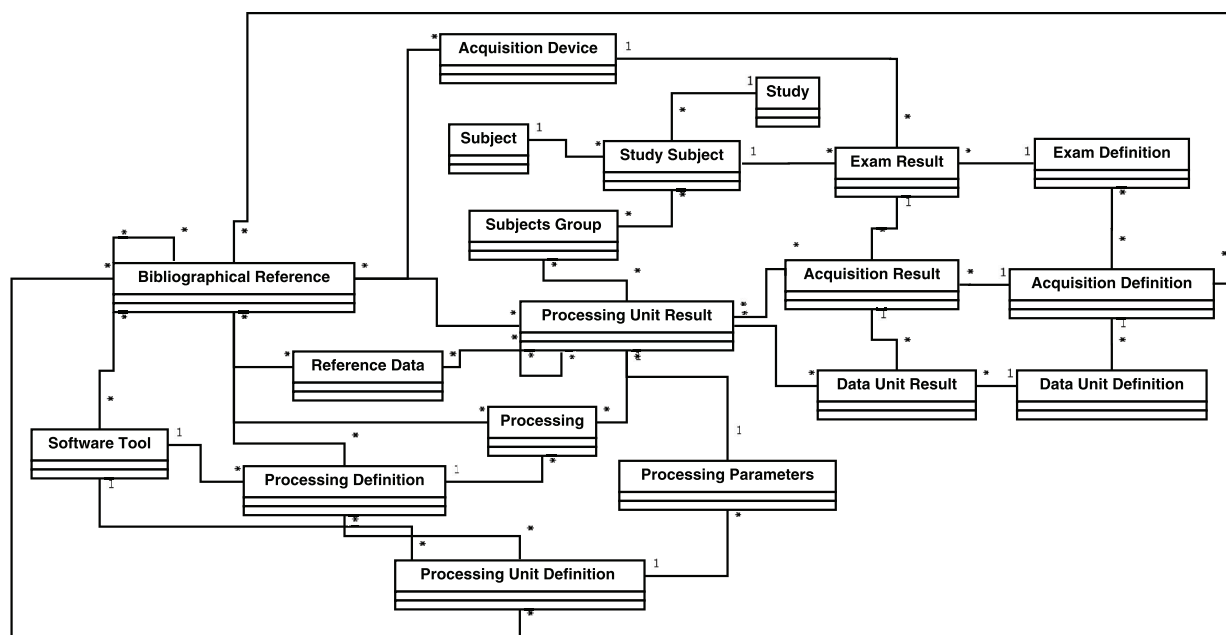


FIGURE 4 | UML representation of the generic objects of the BioMedical Imaging—Lifecycle Management data model.

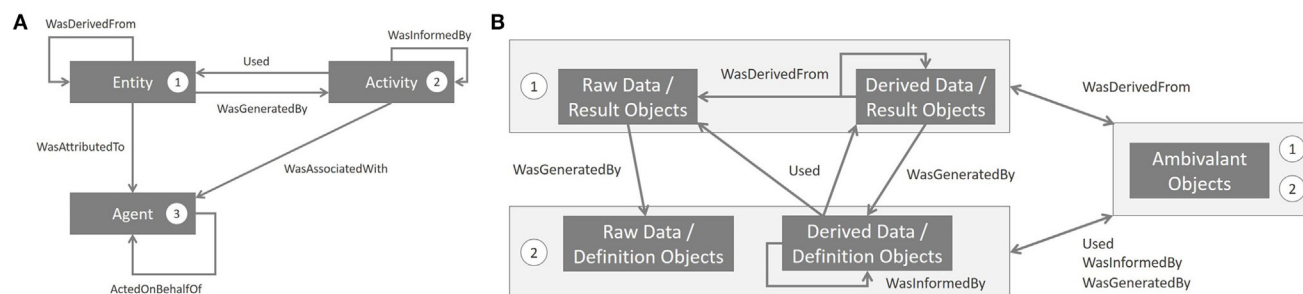


FIGURE 5 | Links between the PROV standard and the BioMedical Imaging—Lifecycle Management (BMI-LM) data model. **(A)** Organization of the Provenance Data Model (PROV-DM) standard developed by the W3 consortium (<http://www.w3.org/TR/2013/REC-prov-dm-20130430/>). Numbers represents categories shown in panel **(B)**. **(B)** PROV relationships between categories of objects of the BMI-LM data model. The *Agent* concept of PROV-DM is not mentioned on the figure, because Product Lifecycle Management (PLM) features naturally fulfill agent provenance: *Agent* is represented by a PLM user (real person or robot), *WasAssociatedWith* by a workflow order, *ActedOnBehalf* by the project organization, and *WasAttributedTo* by the owner of the resulting data (*Entity*).

difficulty we face is that there is no equivalent for the definition objects in the DICOM standard: if a same scan generates five DICOM different images series, we get five seemingly unrelated DICOM series. In order to tell the PLM system which series derive from which acquisition, we first have to group the DICOM series derived from a single scan, based on the contents of several different DICOM attributes.

The DIMP Method for Integration of Processing Pipelines

Studies in neuroimaging require complex pipelines for the processing of images: registration, segmentation, temporal or spatial filtering, etc. The pipelines may include many different steps and algorithms, parameters, and software that are regularly

evolving as research progresses. Their structure varies according to the image acquisition techniques employed and the nature of the endpoints that are needed to test the studies hypotheses. The neuroimaging community has developed elaborate pipeline management systems, such as LONI pipeline (Rex et al., 2003; Dinov et al., 2010) or Nipype (Gorgolewski et al., 2011). With such systems, Command Line Interfaces tools are wrapped by structures describing each of their inputs, options flags and outputs, and storing the name of the executable, enabling the software to build proper command lines. These structures can be linked together into a processing graph with a node representing a processing unit and an edge representing an input and output relationships. The graph is then analyzed to optimize the parallelization of jobs on grid computers.

TABLE 4 | Basic mapping between DICOM protocol and BioMedical Imaging—Lifecycle Management (BMI-LM) data model concepts.

DICOM	BMI-LM
Patient	Study subject object
Study	Exam result object
Series	Data unit result object
Set of series from the same scan	Acquisition result object
–	Definition objects

Some neuroimaging software suites, such as XNAT, come with an integrated pipeline management system, by allowing users to launch processing pipelines directly from the database. In this case, imaging sessions are launched one by one. When processing data in large batches, it is more convenient to push and pull the data in and out of the database (Schwartz et al., 2012). However, if the pipelines are launched externally, the inputs and parameters become more difficult to track. In order to ensure research reproducibility, traceability of statistical models used for prediction, data sharing with peers and data reuse, the provenance information of the processing pipelines must be properly managed. Because of the complexity of pipelines, provenance information has to be generated automatically by the pipeline management system and then stored in the database. We developed the DIMP method with these two objectives in mind: ensuring full provenance and facilitating the launch of processing pipelines by users.

Specifying the Inputs to an Image Processing Pipeline

To launch a pipeline, users must select: (1) the items to process, (2) a processing pipeline to apply, and (3) parameter settings. The multiplicity of the parameters involved in image processing in neuroimaging studies create a major issue: all the parameters involved in the generation of the derived data need to be tracked to ensure the reproducibility of results, both on same data and on new data. Furthermore, in longitudinal imaging studies, subjects undergo imaging sessions regularly over a long period of time (up to several years), and exactly the same processing chains must be applied so that the data can be compared. Users may also want to store concurrent versions of the derived data, differing over a few processing parameters or processing steps to understand their impact on the results.

To implement this functionality, one needs to add a generic object to the BMI-LM data model: the *Workflow Input* object. Its role is to gather all the definition items needed to launch a processing pipeline: the processing pipeline itself (object: *Processing Definition*), processing parameters for every step (object: *Processing Parameters*), and the definitions of input data (objects: *Data Unit Definition* for raw data, *Processing Unit Definition* for derived data). These last data are crucial: they allow the PLM system to query the right data, for the subjects selected by the user. **Figure 6** shows how using a *Workflow Input* object is particularly valuable to reproduce same processing chain several times on new data (acquisitions on the fly, longitudinal studies, new studies).

Stages of Integrated Processing in a PLM System

The main objective of the DIMP method is to ensure quality provenance of derived data by reducing manual operations from users: data resulting from processing chains are automatically linked to input data, definition of processing chain, and parameters. The DIMP method is defined by the following stages:

Initialization

1. (User) build or identify a workflow input
2. (User) launch integrated processing workflow
 - o Select workflow input
 - o Select subjects

Workflow execution

3. (PLM system) query input data
4. (PLM system) export in working folder
 - o Input data
 - o Definition of the pipeline
 - o Parameters of the pipeline and processing nodes (processing parameters items)
5. (Computer cluster) launch the pipeline script stored in the definition object representing the pipeline. This script parameterizes and executes processing operations.

Traceability operations

6. (PLM system) upload resulting data
 - o Create corresponding result objects
 - o Link result objects to its input data (raw or derived) and *definition objects* (pipeline structure and parameters)
7. (PLM system) sends an email notification: data are ready

Integration of Existing Neuroimaging Pipeline Engines

Processing pipelines are executed outside of the PLM system, typically on a computer cluster. Existing neuroimaging workflow management systems can therefore be used to execute the pipelines on any software libraries that can be launched in command lines. When manual processing is needed (such as expert delineation of brain structure), it is easy to checkout any dataset, modify it or create a new dataset, and send the results back to the PLM. Indeed, this corresponds to how CAD engineers work.

To facilitate user's work, the definition objects of the processing pipeline can be generated through software tools, which extract the relevant information from pipeline specification files and facilitate the specification derived data annotations.

Querying Strategies

Efficiently storing data and managing provenance is not sufficient to ensure that data can be reused: the platform also should enable easy data querying. One major issue preventing from data access is user's knowledge and understanding of the data model: as provenance is complex so are the queries. Therefore, getting to know the different concepts is time-consuming to occasional users. A query is defined both by the search criteria and the formatting of

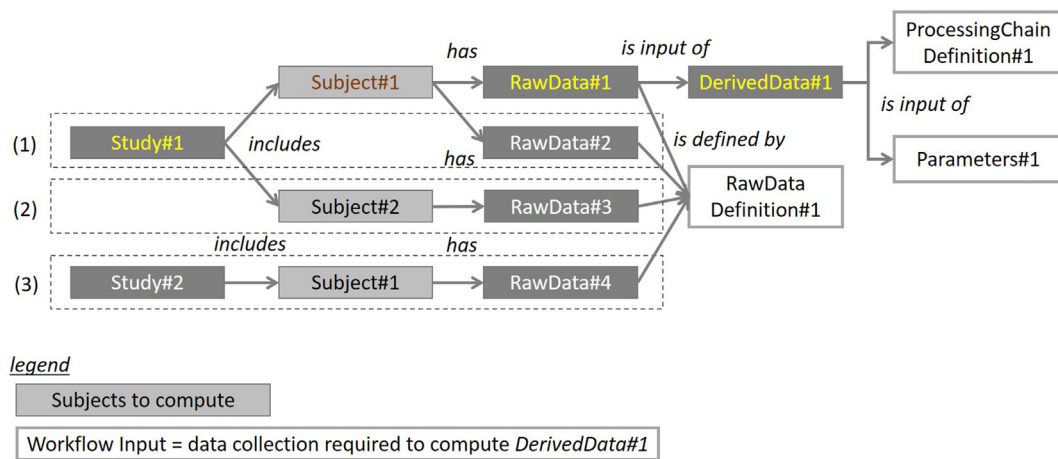


FIGURE 6 | Diagram illustrating how the *Workflow Input* object can be used to reproduce an existing workflow on three use cases: (1) analysis of data associated with a new time point in a longitudinal study, (2) analysis of data from a new subject of the same study, and (3) analysis of data from a new subject of the different study (different study same processing chain). Definition of input data (raw data in the figure, but it could be derived data), definition of processing chain, and parameters are collected in a *Workflow Input* object by the user. When a processing chain has to be computed again on new data (new acquisition or new subject), the *Workflow Input* object is reused and the targeted subjects are given to the system to query corresponding input data. For use case (1), which represents a longitudinal study, appropriate raw data are found by excluding data that has already been computed with the processing chain.

the retrieved data. The BIOMIST platform provides researchers with an intuitive way to retrieve data through a graphical interface (see Graphical Querying Interface). With this interface, queries are designed using concepts and relationships. Consolidated data can also be obtained through ODBC connectivity (see Report Building).

Graphical Querying Interface

Even if the neuroimaging community shares many standards, each research group—not to say each researcher—uses its own vocabulary to label its data. Besides, neuroimaging is a multidisciplinary domain and each discipline has its own concepts and ways of using data. In this context, it is difficult for researchers to query an unknown or an occasionally accessed database, because they are neither familiar with the data model nor with the semantics behind it (Pham et al., 2016). In the BIOMIST platform, to facilitate the query definition process of various kinds of users—occasional/regular, experienced/inexperienced, or the ones who come from different disciplines—we propose a graphical and user-oriented query approach.

For the “user-oriented” aspect, the proposed query approach is composed of three levels of abstraction—lowest, intermediate, and highest corresponding with three kinds of users: technical users, regular users, and occasional or non-technical, inexperienced users, respectively. At the *lowest level*, technical users, who have a good understanding of the way data structured, can directly select business objects in the data model to create a query. For instance, the *Acquisition Result* object is used to query all acquired data during the data acquisition process.

At the *intermediate level*, regular users, who manipulate frequently with data and have a certain understanding about them, are provided with a more abstract hierarchy of data classes. A class can have attributes and is named accordingly to the data it

represent. Regular users could easily find their interesting data from one or many classes. For example, in the “Imaging Result” class, users could find all acquired imaging data like “EEG,” “MEG,” “MR,” and “PET” data. Some relations between classes can be defined to help users make more complex queries on multiple kinds of data.

The *highest level* is dedicated to inexperienced and non-technical users who have no knowledge about the data model and classification. We use ontologies and its graphical representation to facilitate the query making process of these users. The ontology is defined as “an explicit, formal specialization of a shared conceptualization” (Studer et al., 1998) and can be used to provide an explicit representation of domain knowledge and semantic relations between data in the database that is easily understood by inexperienced users. Without needing to understand the underlying data structure, inexperienced users express their queries with ontological concepts. For instance, the “imaging-acquisition-data” concept from OntoNeuroLog ontology (Gibaud et al., 2011) is used to query all acquired imaging data. The query formulated with ontologies is then translated into a formal query over data sources by using a set of mappings. Each mapping is an association between an ontological concept and the database schema. The set of defined mappings is then exported and implemented in the query transformation module of the PLM system.

For the “visual” aspect, playing the role of an external cognitive support to understand complexity (Keller and Tergan, 2005), graphical visualizations are used at the three levels to facilitate users’ query making process. All objects of the data model, classes of the classification, or ontological concepts are represented in a browsing tree while all eventual relationships between them (objects versus objects, etc.) are represented in an intuitive, interactive graphical zone to help users quickly and easily define their queries.

For example, at the highest level, a user starts by navigating through concept tree to select an interesting concept. When a concept is selected, the graph highlights all its relationships with the other concepts; user can select one of these concepts and add it into the query in order to make a query condition. This process is repeated until the query is defined completely. During the making process, the query formulated by the user is graphically represented to provide an illustrated visualization of all selected concepts and query conditions. At the end, this query is translated into one executable query by a query processor. The query results are displayed on the same interface, under the shape of a graph (nodes for resulting objects, edges for relationships).

Report Building

In neuroimaging, more and more studies include meta-analysis. For example, both supervised and unsupervised classification algorithms are typically used for discovering correlation between biomarkers extracted from brain images and behavioral observations or extract hidden structures (Abraham et al., 2014). The building of such data files prepared for analysis is quite fastidious because of the multiple sources of data. Furthermore, beside classification, deep learning algorithms (LeCun et al., 2015) are raising more and more interest in the neuroimaging research community since they begin to show a real potential on analyzing flexible and high-dimensional data, which is their main advantage. To exploit these heterogeneous data in a machine learning context, we designed a data mapping that consists of exporting neuroimaging data classification from the PLM, to a database server that most statistical analysis softwares should be able to address. The connexion between the PLM database and the database structured for statistical analyses is enabled with ODBC, a standard API (Signore et al., 1995).

IMPLEMENTATION

PLM Choice and Customization

The BMI-LM data model has been implemented in the PLM software Teamcenter (v10.6) developed by Siemens Industries Software, which has a commercial license. Information about Teamcenter architecture and technical details can be found in Teamcenter documentation: Teamcenter system administration (Siemens PLM Software, 2015b) and Teamcenter access manager (Siemens PLM Software, 2015a). Besides, Siemens PLM Software published a white paper on security management in Teamcenter (Siemens PLM Software, 2011). CIMdata, a leading independent global consulting and research authority toward PLM, wrote a white paper focused on Teamcenter as a unified platform that describes its functionalities (CIMdata, 2010). A type of Teamcenter objet is created for each object of the BMI-LM model, so that the four stages of a neuroimaging study are supported. Data are attached to object instances through dataset objects. The object instances are linked through typed relationships as defined in the BMI-LM data model. Teamcenter proposes a classification feature, which is often used in manufacturing industry to classify products in families.

Teamcenter PLM system is easily customizable to fit users' needs: data model, data formats, workflows, access management, queries, integrated visualization and analysis tools, and interface. These make Teamcenter a backbone that can be adapted to the specific features of new domains (processes, formats, tools, etc.).

The organization feature of Teamcenter is used to model users' groups and roles, which are required to design access rules to the data. Four roles are defined to access data inside of a study: principal investigator (can view all data of the project and edit all instances), data administrator (can view some data of project, can create and edit instances of objects, and can manage relationships between instances), editor (can view some data of the project, can edit instances of objects), and guest (can view some data of the project). The amount of data viewed and editable for each role can be defined.

Three data vaults that store files are set up with different backup strategies, according to data value:

- Raw data: this vault is the most valuable, as it contains all acquisition and study data. During acquisition or import campaigns, daily backup.
- Derived data: valuable too, but as these data can be computed again thanks to provenance storage and because the volume may be very big, the backup is occasional.
- Definition data: this vault is the lightest, as it contains only the data from definition objects. The backup strategy is high, as these data are crucial. Domain classification for neuroimaging.

Domain Classification for Neuroimaging

The definition of a classification requires a substantial investment in time and expertise. Some ontologies have already been designed and used by the neuroscience and neuroimaging communities (Temal et al., 2008). Therefore, defining the neuroimaging classification on existing organized knowledge seems relevant. Besides, the use of existing ontologies allows future data sharing between the PLM system and existing neuroimaging databases. Ontologies can be used as a mediation model between the data models of two databases. Aside of ontologies, standardized and partly aligned lexicons also exist, such as NeuroLex¹ and DICOM that can provide class attributes. In a PLM system, class attributes are stored in a dictionary. Classes are stored in a hierarchical tree and can receive any number of attributes from the dictionary. We imported classes from OntoNeuroLog (Gibaud et al., 2011) ontologies for the classification branches that deal with image acquisition (image examination, acquisition, and data unit definitions) and image processing (processing unit definitions, imaging datasets). We based the subject-related branch of the classification on QIBO (Buckler et al., 2013). MRI parameter attributes (parameters such as the echo time) were imported from the DICOM lexicon (Clunie, 2000). Currently, we use attributes in the experimental psychology classes to store labels from the cognitive atlas (Poldrack et al., 2011) or cognitive paradigm

¹<http://neurolex.org>.

(Turner and Laird, 2012) ontologies, as those seemed too large to be imported fully in the classification.

The classification that is used in the BIOMIST platform in its current state is available as Part S2 in Supplementary Material, in a mindmap format that can be viewed with the Freemind software.²

Data and Software Integration

Data Import

We developed a DICOM/Teamcenter interface that relies on the dcm4che java DICOM library. This way, the PLM server can act as a C-STORE service class provider (a DICOM archive), as well as a query/retrieve service class provider. It is therefore able to interact with existing PACS instances and DICOM viewing workstations. As XNAT, we rely on an intermediary gateway to comply with the defined PLM access management policies during query/retrieve operations. We also use web services to import other types of data (i.e., non-imaging data): for instance, to import the resting-state debriefing questionnaires, a web service receives the data from a LimeSurvey³ server and imports it into the PLM database.

Data Processing

As neuroimaging pipeline engines are now very mature, there was no need to develop a new one for the BIOMIST project. To implement the DIMP method, we chose the Nipype⁴ (Gorgolewski et al., 2011) pipeline engine, because it is simple to extend, flexible (written in Python), able to deal with many grid schedulers. Since this software originates from the neuroimaging community, it has a very rich catalog of interfaces for neuroimaging Command Line Interfaces tools [AFNI (Cox, 1996), ANTS (Klein et al., 2009), SPM (Ashburner, 2012), Freesurfer (Fischl, 2012), FSL (Jenkinson et al., 2012), etc.]. When running a job on a computer cluster, there are two different aspects to take into account: the command line to be executed (what are the inputs and options?) and the way the scheduler is going to handle it (how much memory, time or CPUs do we need?). The former is the domain of specific command line wrappers (i.e., the Nipype interfaces); the latter is the domain of generic processing node properties. We use the Teamcenter classification system to account for both. Accordingly, we developed python tools to import the existing Nipype interfaces, which describe the input and outputs of each command line tool, within the PLM classification as *processing parameter* classes. Based on these tools, we also developed tools to import entire Nipype workflows in the PLM (*processing definition*, *processing unit definitions*, and *processing parameters* items) and build the associated *workflow input* items.

Data Querying

The querying interface was implemented as a Javascript web client that connect to Teamcenter through a web service. The interface is composed of several windows, displaying information to build the query: the domain ontology, the relationships,

the related classification, the criteria of the query chosen so far, and the query path itself. A view of the web querying interface is presented in Figure 7.

For the implementation of consolidated data files for statistical analysis, we took advantage of the PLMXQuery tool that is an approach for querying and exporting data from PLM (Sriti and Boutinaud, 2012). The concept of this approach is to make the PLM content seen as a XML document, in order to benefit from XML-related technologies, in particular XPath and XQuery, which are standard languages working on XML structures. XQuery scripts are used to browse PLM content (items, classification data, dataset contents, etc.) and to convert that data to any desired format. It can be used to create or update anything from a Hive table to a CSV file. It is currently used nightly to update data tables containing information about ongoing MRI acquisition that are accessed by researchers through ODBC connectivity for analysis with the JMP statistical software.

Example of Workflow: Raw Data Quality Check

Teamcenter PLM system allows creating easily workflows of operations. We present an example of workflow that is used to control the quality of new imaging raw data. Figure 8 shows the steps of the workflow:

1. Start of workflow: the workflow is initiated with raw data to control.
2. Automated quality control of raw data imaging parameters against those stored in the definition items.
3. A temporary status is assigned depending on control results.
4. The data manager (technical expert) is notified by email that there are new imaging data to control.
5. The data manager controls new imaging data.
6. The final status is set on new imaging data. If this status is "validated," then the raw data would be involved in new workflows, such as processing workflows.

Speed of Access and Computing

Teamcenter PLM system is an efficient system to query and retrieve managed data. Data relationships are browsed as a graph and therefore query complexity is equal to graph browsing complexity. During the DIMP method, input data are queried and retrieved on computing grid and output data are imported when computation is done. Speed of data retrieving, as well as the speed of data import, is dependent of computing grid network performances. Besides, speed of data computing is dependent of computing grid performances and analysis tools chosen.

Licensing of the BIOMIST Platform

The conceptual data model is published and freely available to the community, as well as methods and functioning principles. The core of the BIOMIST platform is Teamcenter PLM system, which has a commercial license and academic licenses that are available for education and research purposes. Any analysis or visualization tool can be integrated with Teamcenter, whatever their type of license. We plan to release the TeamCenter business model files (which are meant for the TeamCenter Business

²<http://freemind.sourceforge.net>.

³<http://limesurvey.org>.

⁴<http://nipy.org>.

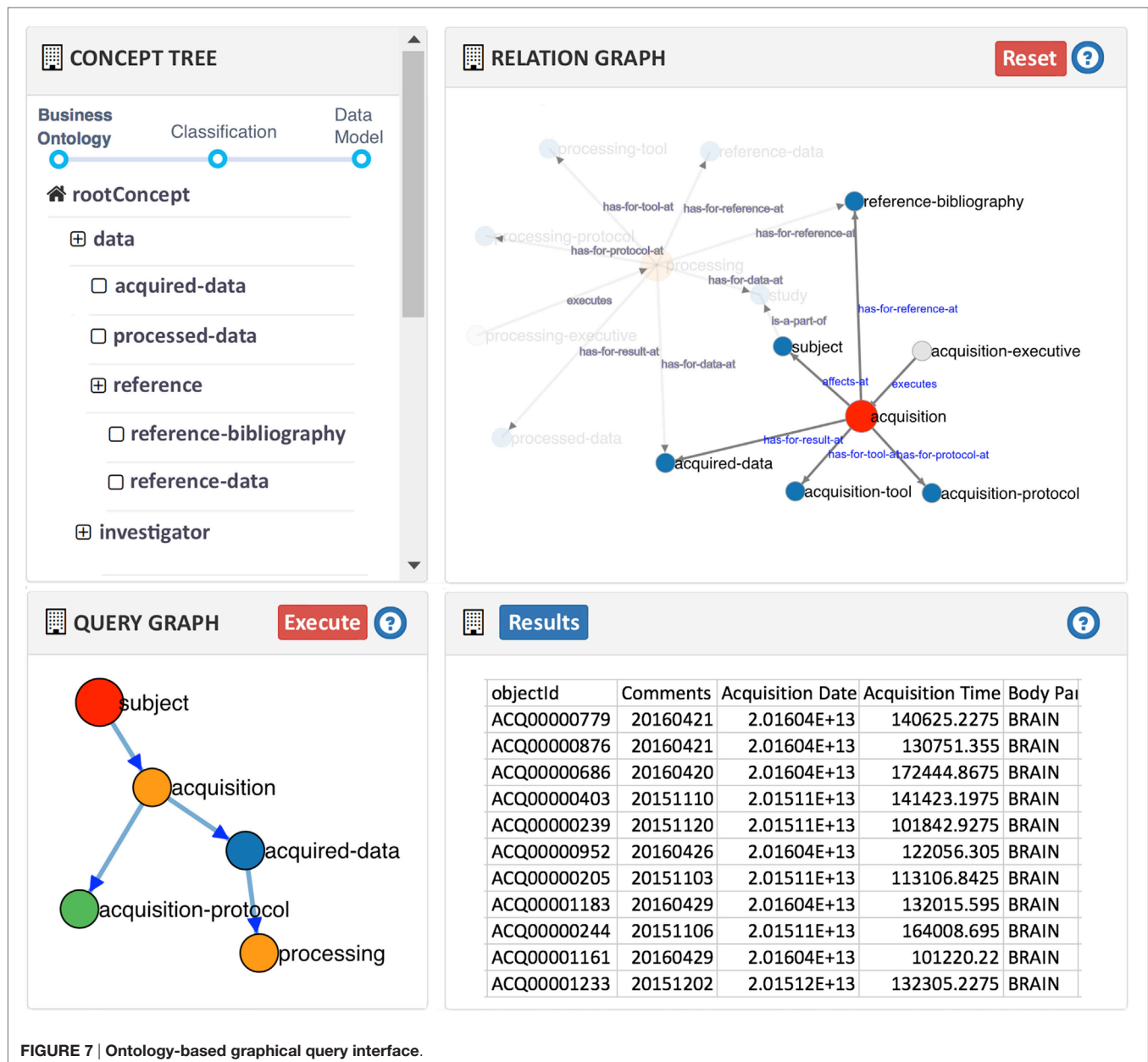


FIGURE 7 | Ontology-based graphical query interface.

Model IDE) under a GPLv3 license, using a web-based version control management service. This will still require that users have access to a TeamCenter license, however. We hope to have provided enough details in the article so that the model as described here can also be re-implemented using open-source software.

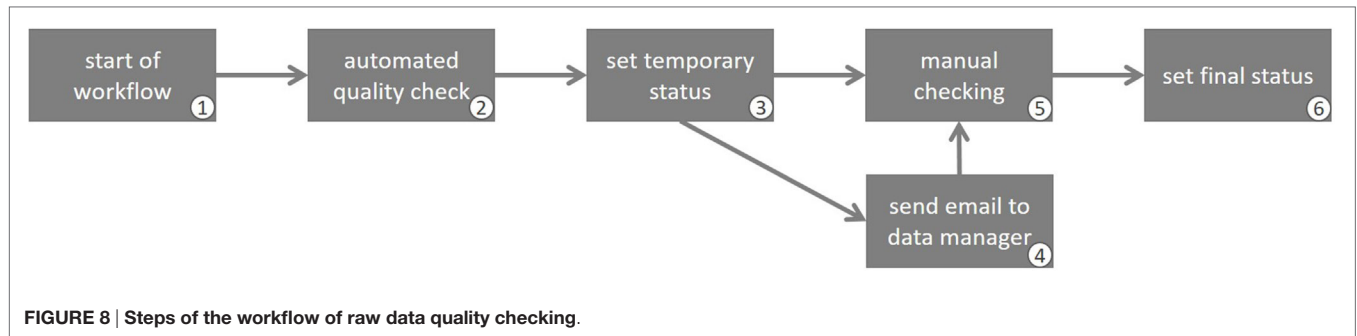
We plan to open in the middle of 2017 the platform to researchers through collaborative scientific projects with the GIN. We plan to open in the middle of 2017 the platform to researchers through collaborative scientific projects with the GIN. For those projects, researchers of both groups will decide the sharing of their respective data in relation with the goals of the collaborative study. For projects that are not in the field of scientific expertise of GIN, Ginesis-lab (joint venture project

between GIN and Cadesis) intends to launch another system to give researchers an access to the functionalities of the platform. Researcher groups interested are welcome to contact the corresponding author.

APPLICATION

Study of Brain Network Connectivity on the BIL&GIN Dataset

The GIN first Brain Imaging Laterality (BIL&GIN1) dataset is composed of 300 subjects, balanced by gender and handedness, and was acquired between 2009 and 2011 (Mazoyer et al., 2016). MRI resting-state images are segmented with a 384-region atlas and connectivity by pair of regions is measured.



The use case tested on the BIOMIST platform with the BIL&GIN dataset stands in six steps (illustrated in **Figure 9**):

1. *Acquisition of raw data*: the BIL&GIN dataset is imported from the GINdb database of the GIN laboratory (Mazoyer et al., 2016).
2. *Processing of individual data*: a pipeline that computes functional connectivity between regions of the brain is automatically launched with the DIMP workflow.
3. *Creation of analysis groups*: groups of subjects are queried according to research assumption based on subjects' characteristics. The chosen criteria are: age, gender, and declared handedness, stated in ranges.
4. *Processing of group data*: a pipeline computes median functional connectivity for each group, creates from these data a MDG and computes a constraint layout to help the visualization analysis. All these processing operations are performed with the DIMP workflow.
5. *Visual browsing of complex graphs*: the resulting MDG is analyzed in an integrated visual browser.
6. *Publication of results*: the paper presenting the results of the MDG analysis would be written with a versioning history and linked to the data used for the analysis, which enables the replication of the procedures involved.

The BIL&GIN dataset, stored in a SQL-based database, was imported into the BIOMIST platform through a scripts that converted SQL tables into PLMXML files readable by Teamcenter PLM. **Figure 10** shows raw data of a subject from the BIL&GIN dataset in the BIOMIST platform: the subject has two exams, one fMRI resting-state exam with three acquisitions (resting-state, anatomical, debriefing form) and one exam about subject's individual characteristics.

Imaging raw data were processed with the DIMP method, with four workflows: (1) preprocessing workflow (registration, segmentation), (2) workflow to compute individual adjacency matrices of functional connectivity, (3) workflow to build group adjacency matrices, and (4) workflow to compute and analyze dynamic graphs from group adjacency matrices. **Figure 11** shows how the final dynamic graph is obtained from individual adjacency matrices of functional connectivity.

The study of resting-state networks with MDGs on the BIL&GIN dataset is currently under process.

Ongoing Cohort Acquisition Campaign

The MRI-Share study is a subpart of the i-Share epidemiological study on students' health.⁵ As many as 2,000 students are expected to undergo an MRI protocol including structural, diffusion, and multiband resting-state acquisitions on a recent 3-T scanner.

The MRI-Share study is particularly suited to test the BIOMIST platform, as it is a multidisciplinary study: resting-state fMRI acquisitions are followed by a debriefing questionnaire (Delamillieure et al., 2010) and other psychological data and genetics acquisitions. Because of the high number of subjects, batch data processing, as implemented with the DIMP method, is mandatory.

The acquisition campaign started in November 2015. Up to 10 subjects participate every day in the study from Tuesday to Friday, every week. At the time of writing, 1,200 subjects have participated. The import of a typical MRI-Share DICOM study (about 2.5 Go of data and 3,300 instances) into the BIOMIST database takes an average of 7 min and 56 s, with a SD of 221.7 s (3 min and 41.7 s). The daily acquisitions are imported every night, through an intermediary PACS system (dcm4chee) and a web service.

DISCUSSION

The BIOMIST platform is designed to manage, share, and reuse data from neuroimaging studies. Provenance is tracked throughout the four stages of the lifecycle of a study, whatever data type or format, thanks to:

- PLM systems that naturally enable collaborative work and lifecycle management in a secure environment.
- The BMI-LM data model that supplements PLM features by introducing the concepts of a neuroimaging study and by allowing future semantic changes and evolutions of research practices. The data model enables the traceability of the data in ways similar to PROV-DM standard from W3C.
- Mapping strategies that allow automated data import, such as DICOM files or forms.
- The DIMP method that allows to launch processing pipelines and to retrieve automatically the resulting data; existing workflow engines and processing software can be integrated.

⁵<http://www.i-share.fr/>.

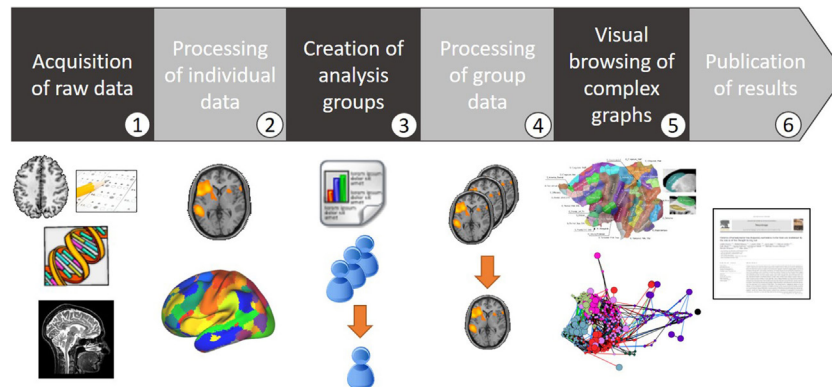


FIGURE 9 | The six steps of the use case on the BIL&GIN dataset.

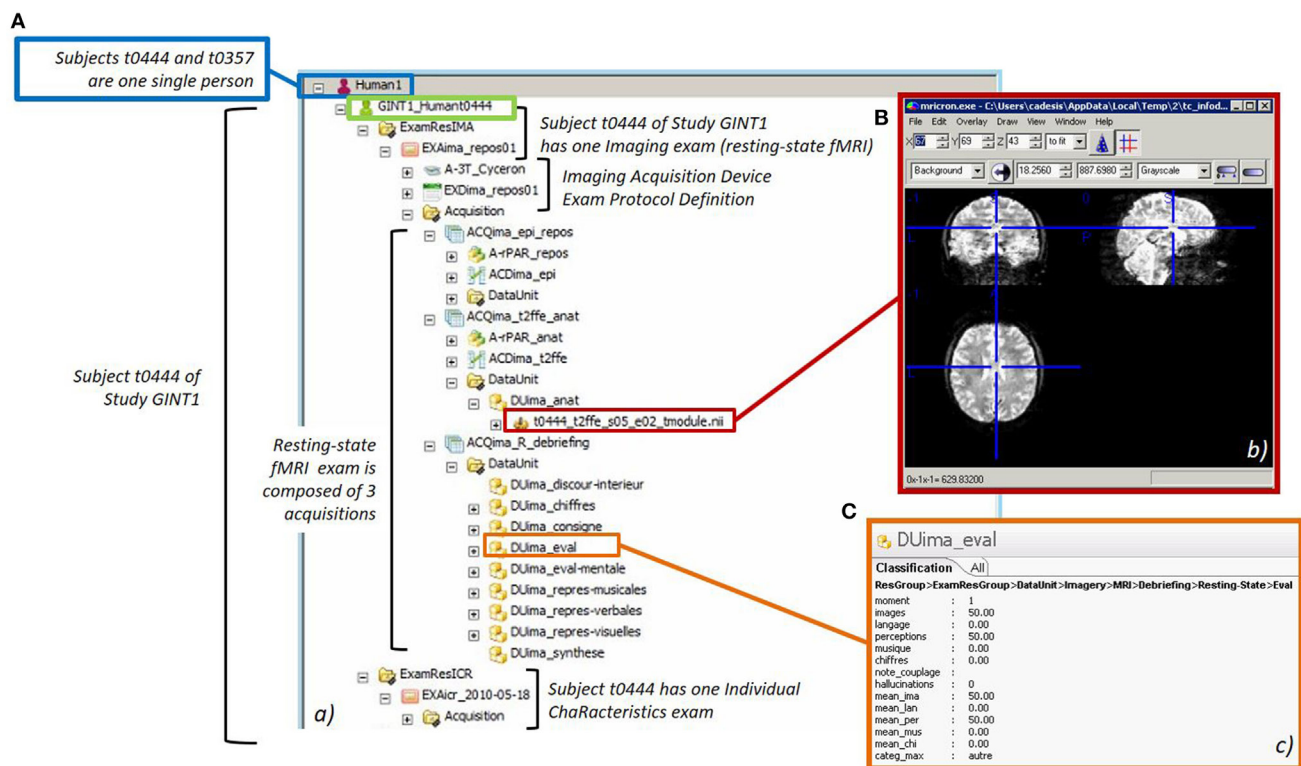


FIGURE 10 | Step 1 of the use case with the BIOMIST platform. (A) Raw data of a subject identified t0444 from the BIL&GIN dataset in Teamcenter client. Nifti anatomical image (B) and resting-state debriefing form (C) are displayed.

- A graphical query-building interface accessible to occasional users and report building to perform statistical analyses.
- Easy integration of visualization and processing tools.

The BIOMIST platform is currently used for the management of the BIL&GIN dataset (300 participants) and the ongoing longitudinal MRI-Share cohort acquisition of 2,000 participants, and its target is new neuroimaging studies from small (100 subjects)

to medium (5,000 subjects) cohort, with multimodal, longitudinal and multi-source acquisitions requiring complex pipelines, quality controls, and efficient access management. The studies managed on the BIOMIST platform are still ongoing; therefore, the BMI-LM has not been validated on the fourth stage of a study (published results).

The BIOMIST platform distinguishes from existing neuroimaging data management systems by providing in one environment:

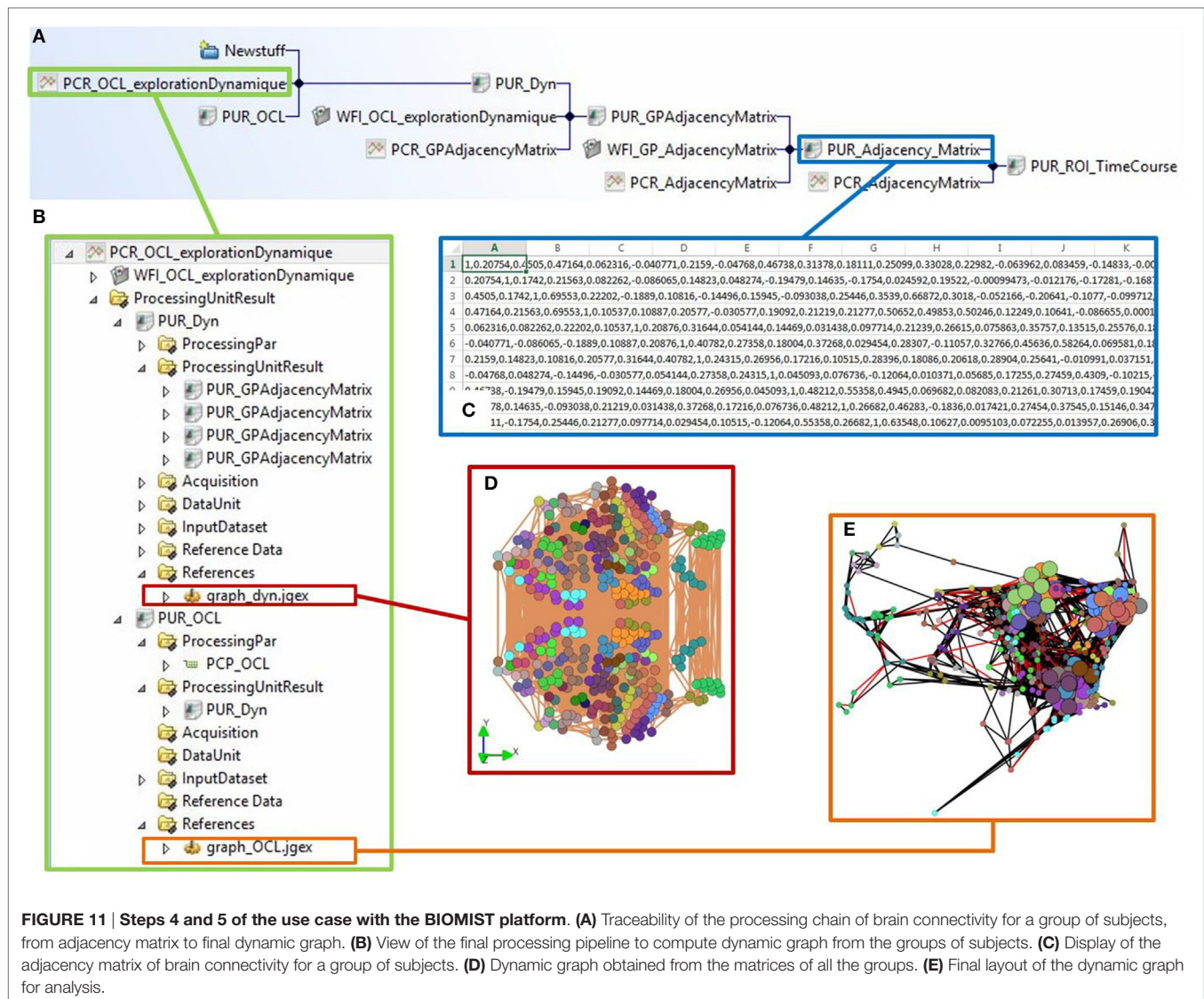


FIGURE 11 | Steps 4 and 5 of the use case with the BIOMIST platform. (A) Traceability of the processing chain of brain connectivity for a group of subjects, from adjacency matrix to final dynamic graph. **(B)** View of the final processing pipeline to compute dynamic graph from the groups of subjects. **(C)** Display of the adjacency matrix of brain connectivity for a group of subjects. **(D)** Dynamic graph obtained from the matrices of all the groups. **(E)** Final layout of the dynamic graph for analysis.

(1) study data management throughout study lifecycle, (2) heterogeneous data management, and not only imaging, (3) managing provenance in order to enable data sharing and reuse, (4) allowing data processing and analysis inside the platform, with users' regular software tools, and (5) providing a secured access to preserve data consistency and confidentiality. One current disadvantage of the BIOMIST platform is the necessity to train a specialized data manager in order to maintain the system, because it is complex with many possibilities of personalization.

One of the main objectives in designing the platform was to enable the use of existing neuroimaging tools and community standards: data formats, workflow engines, processing and visualization software, and ontologies. To foster data sharing through the community, it would also be relevant to bridge PLM systems with web-based archival systems such as XNAT or such as PubMed in order to link bibliography management of the BMI-LM model with the most complete bibliography database in medical field. Mediation between databases is possible through

ontologies. Some work has already been done on this topic in the neuroimaging community (Ashish et al., 2010). Although classes from ontologies are being used in the BIOMIST platform for the neuroimaging data classification and the graphical querying interface, richer semantics would improve the management of relationships between the different objects in PLM systems (Assouroko et al., 2012). For instance, the mapping for data import could rely on an ontology-based description, rather being described in a XML file. Therefore, future work on the BIOMIST platform will focus on application of ontologies within PLM systems for improved interoperability, reusing, and simplified data management.

Moreover, in order facilitate data exchange between the BIOMIST platform and existing neuroimaging data management systems, we plan to develop a feature to export data provenance in PROV-DM format.

GIN users' feedback also highlighted that the eclipse-based graphical user interface of the deployed PLM system would be

unsuitable to them, because of the daunting numbers of sub-windows and menus that reduce the implicit use of the system; a simplified and more adequate user interface is being developed, intended for occasional users. Due to the nature of neuroimaging research work, the relationships between database objects are complex, so the ability to navigate among data is critical. However, current PLM systems do not propose a satisfactory relation browser or viewer, and they exhibit shortcomings in terms of data visualization and analysis, all the more as complex and heterogeneous data are managed (Allanic et al., 2014). Therefore, a major concern in the upcoming work on the BIOMIST platform is to visualize data relationships, using a visual graph representation, in order to improve the browsing and the visualization of data and provenance in PLM systems.

With the current querying facilities of the BIOMIST platform, users can build and retrieve data reports for statistical analysis. One of our main goals is now to integrate more tightly analytical tools, such as deep learning algorithms on large, multimodal heterogeneous data. The objective is to be able to extract knowledge after analyzing correlations between inter individual variables (age, gender, education, handedness, etc.) and brain structures, in order to provide additional information for a better understanding of brain organization and its mechanisms and also to be able to make predictive assumptions about some neurological pathologies.

AUTHOR CONTRIBUTIONS

All the authors participated to the redaction of the paper. They are members of the BIOMIST project consortium, directed by PB (principal investigator), NM, BE, and MJ. MA worked on Sections “The BMI-LM Data Model to Manage Data and Provenance,” “The DIMP Method for Integration of Processing Pipelines,” “Implementation,” and “Application.” P-YH worked on Sections

“Mapping Strategy for Data Import,” “The DIMP Method for Integration of Processing Pipelines,” “Implementation,” and “Application”; C-CP worked on Sections “Querying Strategies” and “Implementation”; ML worked on Sections “Querying Strategies” and “Implementation”; AD worked on Sections “The BMI-LM Data Model to Manage Data and Provenance,” “Querying Strategies,” and “Implementation”; TB worked on Sections “Mapping Strategy for Data Import,” “The DIMP Method for Integration of Processing Pipelines,” “Querying Strategies,” “Implementation,” and “Application”; AG worked on Sections “Mapping Strategy for Data Import,” “The DIMP Method for Integration of Processing Pipelines,” and “Implementation.”

ACKNOWLEDGMENTS

The authors would like to thank Nicolas Boulic, Jérôme Cornet, and Olivier Menuel from Cadesis, and Christophe Delalande from the GIN, who supported technically their work.

FUNDING

The work presented in the paper was supported by the Agence Nationale de la Recherche (ANR) founded BIOMIST (no. ANR-13-CORD-0007) and Ginesis-Lab project (no. ANR16-LCV2-0006-01). This study also benefited from ABACI, a project supported by a public grant from ANR in the context of the Investments for the Future Program, referenced ANR-10-LABX-57 and named TRAIL.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <http://journal.frontiersin.org/article/10.3389/fict.2016.00035/full#supplementary-material>.

REFERENCES

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformatics* 8:14. doi:10.3389/fninf.2014.00014
- Adamson, C. L., and Wood, A. G. (2010). DFBIdb: a software package for neuroimaging data management. *Neuroinformatics* 8, 273–284. doi:10.1007/s12021-010-9080-z
- Allanic, M., Durupt, A., Eynard, B., Joliot, M., Brial, T., and Boutinaud, P. (2014). “Towards an enhancement of relationships browsing in mature PLM systems,” in *IFIP International Conference on Product Lifecycle Management* (Berlin Heidelberg: Springer), 345–354. Available at <http://www.scopus.com/inward/record.url?eid=2-s2.0-84919388938&partnerID=tZOtx3y1>
- Allanic, M., Pierre-Yves, H., Alexandre, D., Marc, J., Philippe, B., and Eynard, B. (2017). PLM as a strategy for the management of heterogeneous information in bio-medical imaging field. *Int. J. Info. Technol. Manag.* 16, 1. doi:10.1504/IJITM.2017.080950
- Ashburner, J. (2012). SPM: a history. *Neuroimage* 62, 791–800. doi:10.1016/j.neuroimage.2011.10.025
- Ashish, N., Ambite, J. L., Muslea, M., and Turner, J. A. (2010). Neuroscience Data integration through mediation: an (F)BIRN case study. *Front. Neuroinformatics* 4:118. doi:10.3389/fninf.2010.00118
- Assouroko, I., Ducellier, G., Eynard, B., and Boutinaud, P. (2012). “Semantic relationship knowledge management and reuse in collaborative product development,” in *9th International Conference on Product Lifecycle Management* (Québec: Springer), 13.
- Barillot, C., Bannier, E., Commowick, O., Corouge, I., Baire, A., Fakhfakh, I., et al. (2016). Shanoir: applying the software as a service distribution model to manage brain imaging research repositories. *Front. ICT* 3, 25. doi:10.3389/fict.2016.00025
- Book, G. A., Anderson, B. M., Stevens, M. C., Glahn, D. C., Assaf, M., and Pearlson, G. D. (2013). Neuroinformatics database (NiDB) – a modular, portable database for the storage, analysis, and sharing of neuroimaging data. *Neuroinformatics* 11, 495–505. doi:10.1007/s12021-013-9194-1
- Buckler, A. J., Liu, T. T., Savig, E., Suzek, B. E., Rubin, D. L., and Paik, D. (2013). Quantitative imaging biomarker ontology (QIBO) for knowledge representation of biomedical imaging biomarkers. *J. Digit. Imaging* 26, 630–641. doi:10.1007/s10278-013-9599-2
- CIMdata. (2010). *Teamcenter “Unified” – Siemens PLM Software’s Next Generation PLM Platform White Paper*. Ann Arbor: CIMdata.
- Clunie, D. A. (2000). *DICOM Structured Reporting*. Bangor, PA: PixelMed Publishing.
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* 29, 162–173. doi:10.1006/cbmr.1996.0014
- Crawford, K. L., Neu, S. C., and Toga, A. W. (2016). The image and data archive at the laboratory of neuro imaging. *Neuroimage* 124(Pt B), 1080–1083. doi:10.1016/j.neuroimage.2015.04.067

- Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., and Evans, A. C. (2011). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinformatics* 5:37. doi:10.3389/fninf.2011.00037
- Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., and Evans, A. C. (2012). LORIS: a web-based data management system for multi-center studies. *Front. Neuroinformatics* 5:37. doi:10.3389/fninf.2011.00037
- Delamillieure, P., Doucet, G., Mazoyer, B., Turbelin, M. R., Delcroix, N., Mellet, E., et al. (2010). The resting state questionnaire: an introspective questionnaire for evaluation of inner experience during the conscious resting state. *Brain Res. Bull.* 81, 565–573. doi:10.1016/j.brainresbull.2009.11.014
- Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., et al. (2010). Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PLoS ONE* 5:e13070. doi:10.1371/journal.pone.0013070
- Dojat, M., Pélérini-Issac, M., Ahmad, F., Barillot, C., Batrancourt, B., Gaignard, A., et al. (2011). NeuroLOG: a framework for the sharing and reuse of distributed tools and data in neuroimaging. *Organ Hum Brain Mapp. HBM* 11, 2–5.
- Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021
- Fox, P. T., Laird, A. R., Fox, S. P., Fox, P. M., Uecker, A. M., Crank, M., et al. (2005). BrainMap taxonomy of experimental design: description and evaluation. *Hum. Brain Mapp.* 25, 185–198. doi:10.1002/hbm.20141
- Fox, P. T., and Lancaster, J. L. (2002). Mapping context and content: the BrainMap model. *Nat. Rev. Neurosci.* 3, 319–321. doi:10.1038/nrn789
- Gerhard, S., Daducci, A., Lemkaddem, A., Meuli, R., Thiran, J.-P., and Hagmann, P. (2011). The connectome viewer toolkit: an open source framework to manage, analyze, and visualize connectomes. *Front. Neuroinformatics* 5:3. doi:10.3389/fninf.2011.00003
- Gibaud, B., Kassel, G., Dojat, M., Batrancourt, B., Michel, F., Gaignard, A., et al. (2011). NeuroLOG: sharing neuroimaging data using an ontology-based federated approach. *AMIA Annu. Symp. Proc.* 2011, 472–480.
- Goble, C., and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *J. Biomed. Inform.* 41, 687–693. doi:10.1016/j.jbi.2008.01.008
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinformatics* 5:13. doi:10.3389/fninf.2011.00013
- Gupta, A., Bug, W., Marengo, L., Qian, X., Condit, C., Rangarajan, A., et al. (2008). Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics* 6, 205–217. doi:10.1007/s12021-008-9033-y
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi:10.1016/j.neuroimage.2011.09.015
- Joliet, M., Delcroix, N., Zago, L., Vigneau, M., Crivello, F., Simon, G., et al. (2009). “GINdb: portable database for the storage and processing of human functional brain imaging data,” in *Proceedings of the 16th Annual Meeting of the Organization for Human Brain Mapping*, Barcelona, Spain.
- Keator, D. B., Marcus, D., and Murphy, S. (2009). A national human neuroimaging collaborative enabled by the biomedical informatics research network (BIRN). *NIH Public Access* 12, 162–172. doi:10.1109/TITB.2008.917893.A
- Keator, D. B., van Erp, T. G. M., Turner, J. A., Glover, G. H., Mueller, B. A., Liu, T. T., et al. (2016). The function biomedical informatics research network data repository. *Neuroimage* 124(Pt B), 1074–1079. doi:10.1016/j.neuroimage.2015.09.003
- Keller, T., and Tergan, S.-O. (2005). “Visualizing knowledge and information: an introduction,” in *Knowledge and Information Visualization* (Berlin, Heidelberg: Springer), 1–23.
- Kiritsis, D., Bufardi, A., and Xirouchakis, P. (2003). Research issues on product lifecycle management and information tracking using smart embedded systems. *Adv. Eng. Info.* 17, 189–202. doi:10.1016/j.aei.2004.09.005
- Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M., et al. (2009). Evaluation of 15 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 46, 1–62. doi:10.1016/j.neuroimage.2008.12.037
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi:10.1038/nature14539
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007a). The extensible neuroimaging archive toolkit. *Neuroinformatics* 5, 11–33. doi:10.1385/NI:5:1:11
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007b). The extensible neuroimaging archive toolkit and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi:10.1385/NI:5:1:11
- Mazoyer, B., Mellet, E., Perchey, G., Zago, L., Crivello, F., Jobard, G., et al. (2016). BIL&GIN: a neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage* 124, 1225–1231. doi:10.1016/j.neuroimage.2015.02.071
- Mildnerberger, P., Eichelberg, M., and Martin, E. (2002). Introduction to the DICOM standard. *Eur. Radiol.* 12, 920–927. doi:10.1007/s003300101100
- Moreau, L., and Missier, P. (2013). *PROV-DM: the PROV Data Model*. Southampton: University of Southampton.
- Pham, C. C., Durupt, A., Matta, N., and Eynard, B. (2016). “Knowledge sharing using ontology graph-based: application in PLM and bio-imaging contexts,” in *Product Lifecycle Management in the Era of Internet of Things: 12th IFIP WG 5.1 International Conference, PLM 2015*, October 19–21, 2015, Vol. 467 (Doha, Qatar: Springer), 238.
- Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinformatics* 7:12. doi:10.3389/fninf.2013.00012
- Poldrack, R. A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., et al. (2011). The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinformatics* 5:17. doi:10.3389/fninf.2011.00017
- Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Front. Neuroinformatics* 6:9. doi:10.3389/fninf.2012.00009
- Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The LONI pipeline processing environment. *Neuroimage* 19, 1033–1048. doi:10.1016/S1053-8119(03)00185-X
- Schwartz, Y., Barbot, A., Thyreau, B., Frouin, V., Varoquaux, G., Siram, A., et al. (2012). PyXNAT: XNAT in Python. *Front. Neuroinformatics* 6:12. doi:10.3389/fninf.2012.00012
- Scott, A., Courtney, W., Wood, D., de la Garza, R., Lane, S., King, M., et al. (2011). COINS: an innovative informatics and neuroimaging tool suite built for large heterogeneous datasets. *Front. Neuroinformatics* 5:33. doi:10.3389/fninf.2011.00033
- Siemens PLM Software. (2011). *Teamcenter Security Management – White Paper*. Plano: Siemens PLM Software.
- Siemens PLM Software. (2015a). *Teamcenter v10.6.1 Access Management*. Plano: Siemens PLM Software.
- Siemens PLM Software. (2015b). *Teamcenter v10.6.1 System Administration*. Plano: Siemens PLM Software.
- Signore, R., Stegman, M. O., and Creamer, J. (1995). *The ODBC solution: Open database connectivity in distributed environments*. New York, NY: McGraw-Hill, Inc.
- Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance techniques technical report IUB-CS-TR618. *Science* 47405, 1–25. doi:10.1145/1084805.1084812
- Sriti, M. F., and Boutinaud, P. (2012). “PLMXQuery: towards a standard PLM querying approach,” in *IFIP Advances in Information and Communication Technology* (Berlin, Heidelberg: Springer), 379–388.
- Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data Knowl. Eng.* 25, 161–197. doi:10.1016/S0169-023X(97)00056-6
- Temal, L., Dojat, M., Kassel, G., and Gibaud, B. (2008). Towards an ontology for sharing medical images and regions of interest in neuroimaging. *J. Biomed. Inform.* 41, 766–778. doi:10.1016/j.jbi.2008.03.002
- Terzi, S., Bouras, A., Dutta, D., Garetti, M., and Kiritsis, D. (2010). Product lifecycle management – from its history to its new role. *Int. J. Prod. Lifecycle Manag.* 4, 360–389. doi:10.1504/IJPLM.2010.036489
- Turner, J. A., and Laird, A. R. (2012). The cognitive paradigm ontology: design and application. *Neuroinformatics* 10, 57–66. doi:10.1007/s12021-011-9126-x
- Van Horn, J. D., Grethe, J. S., Kostelec, P., Woodward, J. B., Aslam, J. A., Rus, D., et al. (2001). 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.* 356, 1323–1339. doi:10.1098/rstb.2001.0916

Yarkoni, T., Poldrack, R. A., Van Essen, D. C., and Wager, T. D. (2010). Cognitive neuroscience 2.0: building a cumulative science of human brain function. *Trends Cogn. Sci.* 14, 489–496. doi:10.1016/j.tics.2010.08.004

Conflict of Interest Statement: CADESIS company is distributor of two PLM solutions: Teamcenter published by Siemens Industries Software and Windchill published by PTC.

Copyright © 2017 Allanic, Hervé, Pham, Lekkal, Durupt, Brial, Grioche, Matta, Boutinaud, Eynard and Joliot. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



ArchiMed: A Data Management System for Clinical Research in Imaging

Emilien Micard^{1*}, Damien Husson^{1,2}, CIC-IT Team¹ and Jacques Felblinger^{1,2}

¹ CIC-IT 1433, INSERM, Université de Lorraine, CHRU de Nancy-France, Nancy, France, ² IADI U947, INSERM, Université de Lorraine, Nancy, France

Context: There is a great need in clinical research with imaging to collect, to store, to organize, and to process large amount of varied data according to legal requirements and research obligations. In practice, many laboratories or clinical research centers working in imaging domain have to manage innumerable images and their associated data without having sufficient information technology skills and resources to develop and to maintain a robust software solution. Since conventional infrastructure and data storage systems for medical image such as “Picture Archiving and Communication System” may not be compatible with research needs, we propose a solution: ArchiMed, a complete storage and visualization solution developed for clinical research.

Material and methods: ArchiMed is a service-oriented server application written in Java EE™, which is integrated into local clinical environments (imaging devices, post-processing workstations, others devices, etc.) and allows to safely collect data from other collaborative centers. It ensures all kinds of imaging data storage with a “study-centered” approach, quality control, and interfacing with mainstream image analysis research tools.

Results: With more than 10 millions of archived files for about 4TB stored with 116 studies, ArchiMed, in function for 5 years at CIC-IT¹ of Nancy-France, is used every day by about 60 persons, among whom are engineers, researchers, clinicians, and clinical trial project managers.

Keywords: clinical research, infrastructure, data storage system, imaging, database, web services, centralized resources, data sharing

INTRODUCTION

One main challenge of clinical research with imaging is the need to collect, to store, to organize, and to process large amount of various data according to legal requirements and research obligations.

In practice, most of labs and Contract Research Organizations (CRO) manage many studies or protocols at the same time with several classes of contributors, including radiologists, researchers, physicians, and project managers who perform different kinds of data analysis. These contributors have to manage innumerable images and their associated data without having sufficient IT skills and resources to develop and to maintain a robust software solution.

¹ CIC-IT: Clinical investigation center for Technology and Innovation.

OPEN ACCESS

Edited by:

Michel Dojat,
INSERM, France

Reviewed by:

Alex Pappachen James,
Nazarbayev University, Kazakhstan
Avan Suinasiaputra,
University of Auckland, New Zealand

*Correspondence:

Emilien Micard
e.micard@chru-nancy.fr

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 31 August 2016

Accepted: 05 December 2016

Published: 20 December 2016

Citation:

Micard E, Husson D, CIC-IT Team
and Felblinger J (2016) ArchiMed:
A Data Management System for
Clinical Research in Imaging.
Front. ICT 3:31.
doi: 10.3389/fict.2016.00031

The first approach to solve the important problem is to use conventional clinical data storage systems for medical image such as “Picture Archiving and Communication System” (PACS) (Choplin et al., 1992; van de Wetering et al., 2006). Those systems, available in most of clinical centers are designed to allow to store and transfer only Digital Imaging and Communications in Medicine (DICOM)² images.

The first observation that can be made is that those clinical data storage system offer “patient centered” data structure. Such structure does not match specific research needs in terms of data usage, large-scale processing, and connection with mainstream image analysis research tools. It does not allow fine-tune rights and restriction management for each user according to different studies, either.

DICOM is the most used medical image format but has some limitation for research. It is required in research to store and process various other formats like image in MR raw data,³ physiological signals (Odille et al., 2008), analysis results, or segmentations.

Moreover, taking into account the diversity of research contributors and the big data quantity in clinical research context, every proposed system must be user friendly and respond quickly. More than just a question of appearance, the user experience of graphical user interface is the key point for clinicians or researchers who want to search data efficiently, import a large number of images, or load a dataset into the software that they work with every day.

Thus, we need a solution that is able to store all kinds of files (DICOM, raw data, etc.) with metadata, allows “batch processing” for large-scale studies and has to be fully interoperable within a research environment. This means:

- Easy and safe data transfer from/to local clinical environments;
- Easy and safe data import from the outside (e.g., multicenter trials);
- Easy access to the data from mainstream image analysis research tools.

Considering legal requirements of clinical research, in order to comply with the local law, which the data hosting institutes have to follow (e.g., MR-001 CNIL⁴ reference methodology),⁵ such a system has to ensure data confidentiality using study based access restriction and built-in de-identification (Kushida et al., 2012; Tucker et al., 2016). Take the French law as an example, according to article R. 1123-61—*decree of August 29, 2008 (French Public Health Code)*,⁶ clinical centers should ensure the long-term

conservation of data for at least 15 years after the end of study. This means that not only the storage system must guarantee file and database integrity (Tucker et al., 2016) but also it must offer a quality insurance process to check data validity before integration.

Others emerging research picture archiving system solutions like Shanoir⁷ or CATI⁸ are specialized in neuroimaging data management. These solutions are not really adapted for multi organs and many other file formats storage.

For all the aforementioned reasons, it is understood that all these sensitive data must be stored inside a safe, centralized, and isolated system, effectively excluding short-term data support like CD/DVD/USB-keys and non-secured shared location such as network drives, external hard drives, or common public cloud storages.

In response to all these needs and requirement, we introduce ArchiMed, a complete, centralized, and modular storage and visualization solution developed for clinical research. Designed with a “study-centered” approach, which better fits the research workflow and organization needs, the server application developed in Java™ EE is fully integrated into the local clinical environment (imaging devices, post-processing workstations, others devices, etc.), and is able to safely collect data from collaborative centers and ensures all kinds of data storage, quality control, and interfacing with mainstream image analysis research tools.

MATERIALS AND METHODS

General Description and Software Architecture

ArchiMed is based on a three-tier architecture.⁹ It has been designed to be a service oriented application to integrate environments with multiple clients/users (Figure 1).

The server side of the application is implemented in Java-EE™¹⁰ (Goncalves, 2009) to be deployable on any operating system with Java™. It is hosted on a local network and running on an open source Glassfish application server.¹¹

Data layer and underlying database is currently deployed on a MySQL™ Relational Database Management System (RDBMS) but is also compatible with any other database management system (e.g., Oracle, Microsoft SQL Server, etc.); thanks to Object Relational Mapping.¹² This high-level abstraction technique creating virtual object database can be used from within the programming language independently of the host RDBMS. Business logic layer is one component of the server part that manages how data can be created, displayed, stored, and changed. Some

² «DICOM Homepage», <http://medical.nema.org/>.

³ Image raw data: image signal before any treatment or process. In MRI context, raw data contain “Fourier transform” of the MR image measured, before any reconstruction or filtering.

⁴ Commission nationale de l’informatique et des libertés (CNIL), <https://www.cnil.fr/>.

⁵ Méthodologie de référence MR-001 pour les traitements de données personnelles opérées dans le cadre des recherches biomédicales, <https://www.cnil.fr/sites/default/files/atoms/files/mr-001.pdf>.

⁶ Article R1123-61 <https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006072665&idArticle=LEGIARTI000006908442&dateTexte=&categorieLien=cid>.

⁷ Shanoir (Sharing Neuroimaging Resources), <http://www.shanoir.org/>.

⁸ CATI Neuroimaging, <http://cati-neuroimaging.com>.

⁹ 3-tier architecture is a client-server typically composed of a presentation tier, a domain logic tier, and a data storage tier. This is the most used architecture for service oriented applications.

¹⁰ Java EE at a glance, <http://www.oracle.com/technetwork/java/javasee/overview/index.html>.

¹¹ Glassfish Application Server, <https://glassfish.java.net/>.

¹² Mapping Objects to Relational Databases: O/R Mapping In Detail, <http://www.agiledata.org/essays/mappingObjects.html>.

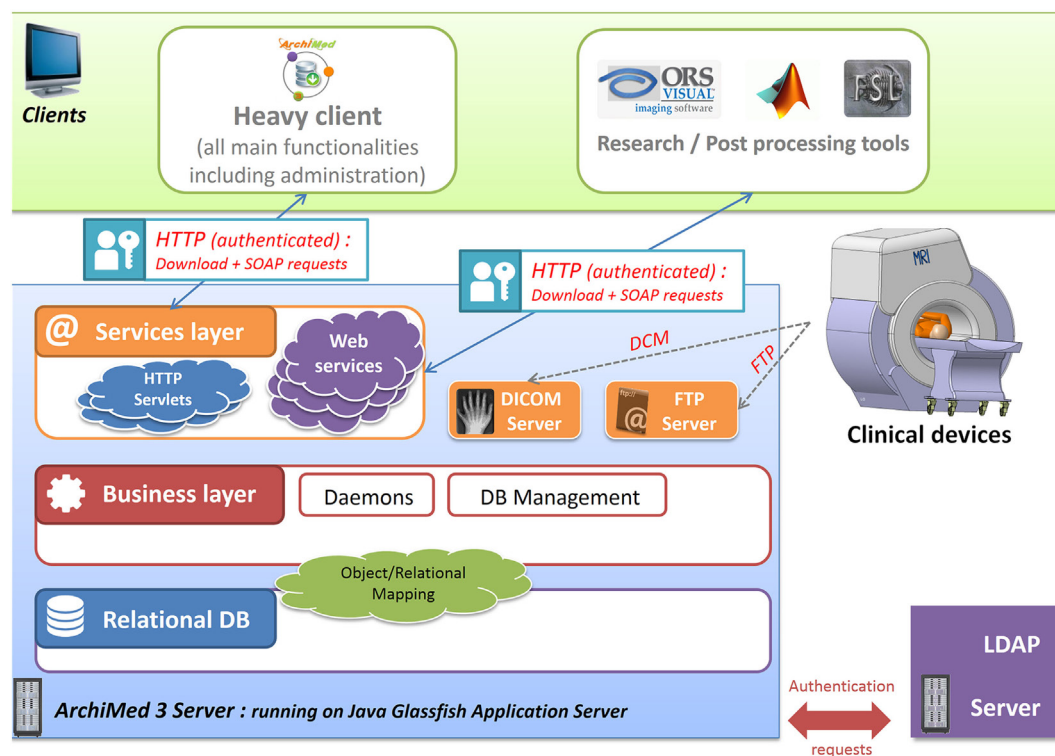


FIGURE 1 | ArchiMed three-tier architecture: describe server 3 layers architecture (data, logic/business, and services) and interactions with client applications and devices.

background tasks (daemons) are dedicated to database cleaning, new image importation, and users action queue management.

The web container layer or service layer, only accessible as authenticated user, provides HTTP SOAP web services¹³ for database querying and HTTP Servlet¹⁴ to perform file download/upload. FTP¹⁵ (Postel and Reynolds, 1985) and DICOM services are also available to transfer files from/to clinical devices (scanners, PACS) or post-processing workstations.

All these services are reachable *via* any third-party application to access and exploit data locally and also by a built-in Java™ heavy client to administrate, to browse, to visualize, and to manage data from ArchiMed. Multi-OS and secured; this client is always up to date using a version check and auto-update mechanism.

Regarding authentication requirements, ArchiMed can be linked to any existing LDAP¹⁶ (Koutsonikola and Vakali, 2004; Zeilenga, 2006)-based user directory system (like active directory) and consequently users can use their own operating system login credentials.

Multiprocessing, Access Speed and Throughputs

Knowing the big amount of data to manage (up to 20,000 images for a standard fMRI¹⁷ exam or up to 2 GB of files for a heart exam with raw data and physiological records, etc.), it is critically important to optimize network stream and parallel access to data.

Every consultation request to ArchiMed server is stateless and treated over HTTP protocol as an independent transaction that is unrelated to any previous request and is executed in a separated process. The downloading rate of data stream is predominantly limited by local area network throughput; server load is estimated as insignificant compare to network flow.

To avoid inconsistency in case of parallel contradictory operations, user actions (data import, transfer, delete, and modification) and every “Create, Update, and Delete” query that involves data change are managed in queues executed asynchronously in transactional¹⁸ context.

¹³Web Services Architecture, <https://www.w3.org/TR/ws-arch/>.

¹⁴Java Servlet Technology Overview, <http://www.oracle.com/technetwork/java/javasee/servlet/index.html>.

¹⁵File Transfer Protocol, <https://tools.ietf.org/html/rfc959>.

¹⁶Lightweight Directory Access Protocol (LDAP): the Protocol, <https://tools.ietf.org/html/rfc4511>.

¹⁷fMRI: functional magnetic resonance imaging or functional MRI (fMRI) is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting changes associated with blood flow.

¹⁸Transactional processing is designed to maintain a database or file system's integrity ensuring that interdependent operations on the system are either all completed successfully or all canceled successfully.

Data Structure

Database and inherent file system are designed with a “study-centered” approach, which better fits the research workflow and organization needs.

ArchiMed data tree is organized around four major node types inherited from DICOM standard: study, exam, series, and file (**Figure 2**).

- Study node: regroupes exams for a study/protocol. Associated information of this node type is
 - Study code: a unique ID of the study
 - Study description
 - Stated investigators and authorized users
 - ...
- Exam node: groups all data linked to one case of the study. Each case corresponds to an inclusion in the protocol. Associated information of this node type is
 - Exam code: a unique ID of the case corresponding to the subject identification number inside the protocol
 - Exam description
 - Exam date, time
 - Last access date
 - ...
- File type/modality node: groups all data with the same file format (file type). In most case, there is a file type for each modality (scanner-type) like DCM_MR (MRI DICOMs), DCM_CT (CT DICOMs), and AEC (physiological signal customized file format).
- Series node: may regroup file of a specific acquisition/sequence. Associated information of this node type is
 - Series number
 - Series description (acquisition sequence name)
 - ...
- File node: single file/image node. As the leaf of the data tree, this node is the representation of the data file that physically presents in the file system. Associated information of this node type is
 - File URL: the address of the file
 - Insertion date
 - Specific metadata: many other metadata depending on the file type (acquisition parameters, voxel size, matrix dimensions, sampling frequency, etc.)

All these metadata and node information extracted from file headers are inserted into the database during the insertion process using customized rules for each different file type.

It is possible to add a new customer recognized format or file type into our system by programmatically defining which metadata need to be extracted from the file and how to read them. Then database dynamically adapt its structure to integrate this new type.

Some meta information are generics and some others are common to all types (Exam code, location URL, dates, etc.), while some others are specific to a certain file type (e.g., echo time and repetition time for MR image data, rescale slope, and rescale intercept for CT image data, customized comment for result or physiological data file, etc.) (**Figure 3**).

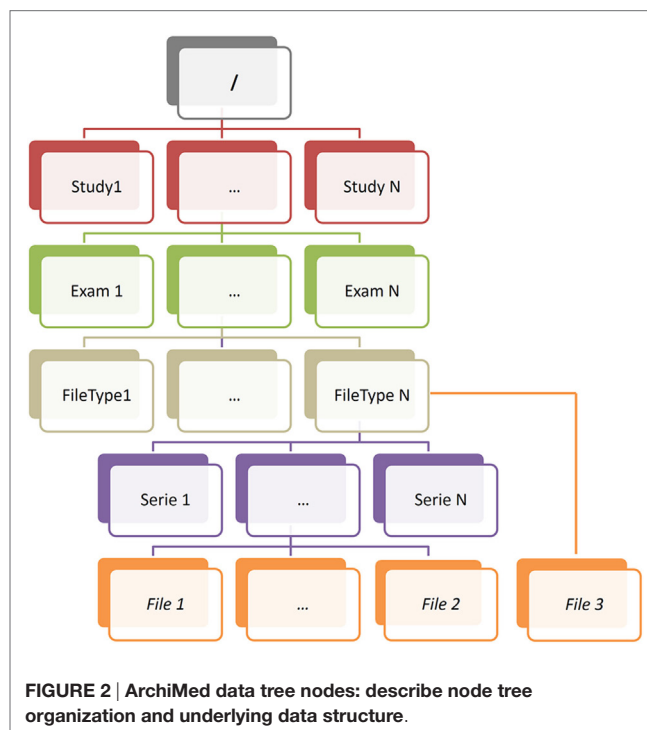


FIGURE 2 | ArchiMed data tree nodes: describe node tree organization and underlying data structure.

Derived data generated by specific local tools such as reconstruction and post-processing software can be inserted inside ArchiMed as “Result” files and linked to the initials data by sharing the same exam node and series node (relationship between data are intrinsically defined inside the database schema). Every result file type can embed other analysis file (CSV, segmentations file, MESH, etc.).

Integrity and Quality Control

Above all, it is important to underline that for security reasons, in order to keep full local control of data. ArchiMed has been design to be hosted on a local network and not accessible from outside (*via* internet network). This considerably limits the risk of intrusion and subsequent data loss or damage.

To avoid unfortunate deletion, move, or file corruption, ArchiMed does not allow direct access to file systems. Data are only accessible *via* authenticated HTTP requests¹⁹ (Fielding et al., 1999), which greatly limits direct access to physical files in order to reduce human factor error.

A “recycle bin” temporary storage retains data deleted by users for several days before permanently erasing them from the file system and ArchiMed storage eases built-in standard backup and archiving system.

As previously stated, ArchiMed can support usual user and group management *via* a standard LDAP connection. It is, therefore, possible to integrate it into an environment with existing right management system (such as Microsoft Windows Active

¹⁹Hypertext Transfer Protocol, <https://www.rfc-editor.org/info/rfc2616>.

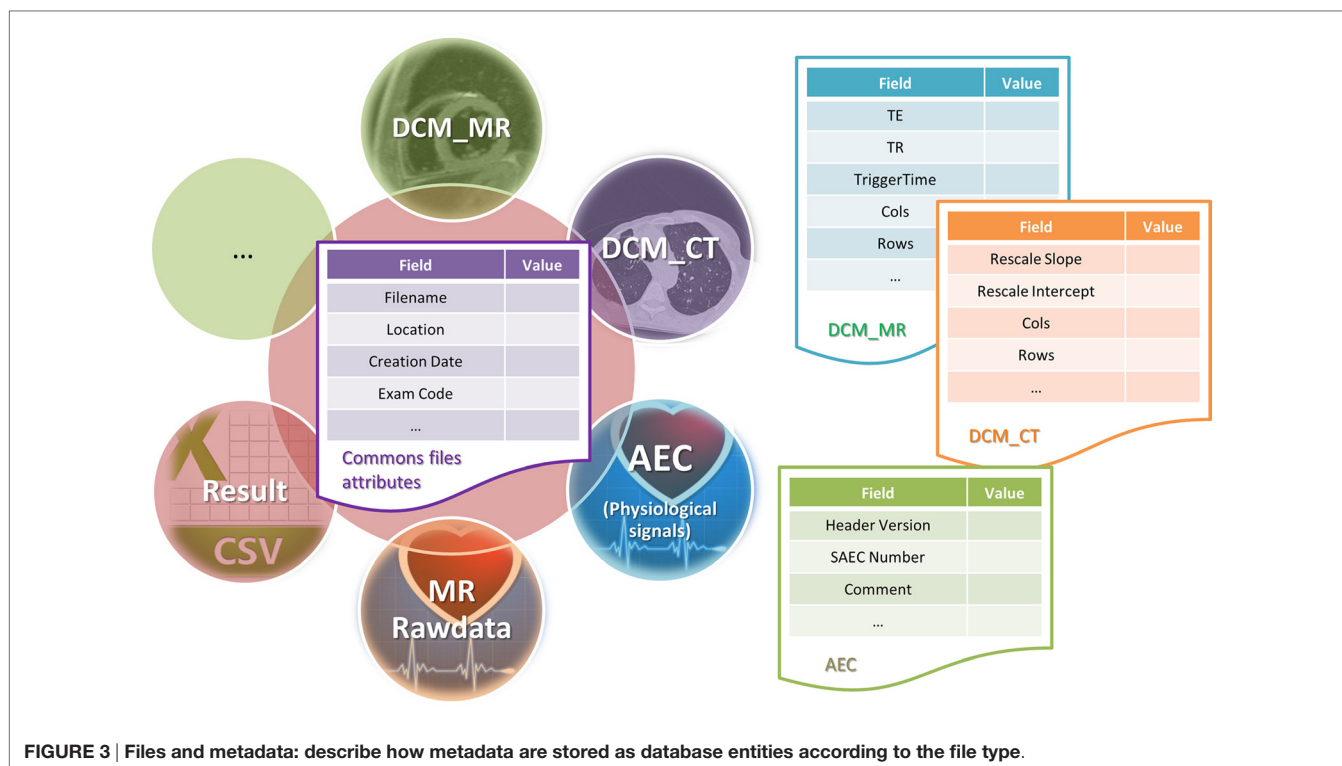


FIGURE 3 | Files and metadata: describe how metadata are stored as database entities according to the file type.

Directory²⁰ or any other LDAP based user's directory). This way, as a member of a group, a user will be able to perform different categories of actions, such as visualization, adding, removing, and downloading (view). It allows a flexible and sufficiently sensitive right management and data access for every nodes (study, exam, series, file). By default, there are four levels of rights with different access privileges for each of them

- Limited user: view only,
- Standard user: view and import (data insertion),
- Power user: view, import, update, and delete (into recycle bin),
- Administrator: view, import, update, delete, permanently delete, and configuration tasks.

Finally, keeping in mind that only human experts can definitely validate or reject acquired data before inserting into the database, we separated the data integration into two steps

- In the first step, data are automatically transferred in a temporary archive. At this point, users, who are in charge of data transfer, must check data validity by checking file information and going through all the images.
- In the second step, data are “transferred” to the final storage, assigned to a specific study, and eventually de-identified.

Although more time is needed than a single-pass automatic import, we believe that these two steps are necessary to prevent image insertion error.

Confidentiality

In addition to action user right explained above, administrator can grant access to a study to users. Thereby, only clinicians, researchers, engineer, and project managers identified in the protocol of the study can access to the data of this study. At this level, there is no consideration of group; access grant to a study is allowed individually.

Data files from different study are physically separated in different cache directories and totally inaccessible without authentication.

ArchiMed offers a built-in de-identification²¹ (Kushida et al., 2012) functionality that automatically replaces or erases identifying fields (name, date of birth, acquisition center, etc.) from data headers before or after import.

In accordance to legal requirements of clinical research and because research centers are not supposed to keep any other trace of patients/volunteers across the studies, exam code unique ID is the only available information about the case inclusion inside ArchiMed database, multicenter data storage.

In multicenter clinical trial context, the main source or error in collected data are due to bad de-identification (removed fields, missing data) or bad electronic support quality (corrupted CD, wrong data, etc.). Since ArchiMed server is hosted on a local network not accessible from outside, it cannot replace non-secure CD transfers and safely collect data from other collaborative centers. Therefore, we developed Eureka, a secure transfer tools

²⁰Microsoft Active Directory, <https://technet.microsoft.com/en-us/library/cc977985.aspx>.

²¹De-identification (DICOM in general), [https://wiki.nci.nih.gov/display/Imaging/De-identification+\(DICOM+in+general\)](https://wiki.nci.nih.gov/display/Imaging/De-identification+(DICOM+in+general)).

coupled with ArchiMed, which allows external centers to send de-identified data through internet into ArchiMed (Figure 4).

The secondary goal of this auxiliary application is to standardize and to keep control on data de-identification and transfer while allowing centers' operators to visualize and check image before sending.

Interoperability

To connect ArchiMed to the clinical environment, a DICOM transfer protocol is implemented (Figure 4). Behaving like a DICOM node, it can receive/send images from/to clinical devices (e.g., MR, CT, etc.), PACS, or workstation. Specific files like MRI raw data are sent *via* FTP.

Web services and Servlets technologies make ArchiMed interoperable on HTTP. This means that every application can query the database and upload/download files.

Moreover Plugin Development Toolkit (PDK) provides Java developers with the tools necessary to create plugins that extend the functionalities of ArchiMed client application. Developed plugins can be global or associated to a specific node (exams, series, and files).

RESULTS

Current version of ArchiMed is in function for 5 years at CIC-IT of Nancy-France and used every day by about 60 people, among

whom are researchers, clinicians, and clinical trial project managers for local or multicenter studies.

- Studies: 116 (including 7 multicenter studies)
- Stored exams: ~10,000
- Stored files: ~10,000,000
- Disk size: ~4 TB
- Database size: ~7 GB

Figure 5 shows data import activity and amount of data for all studies.

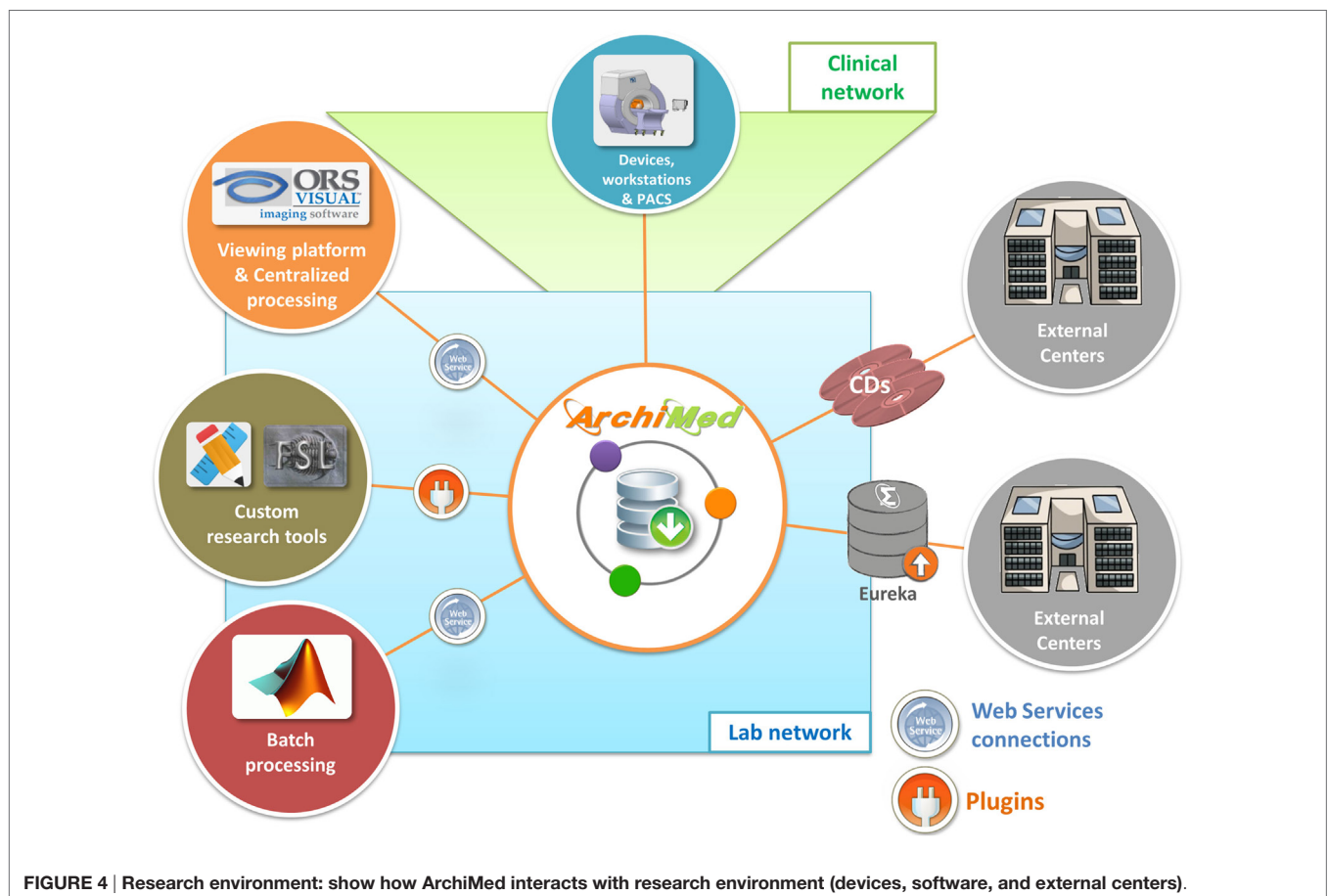
Studies Examples

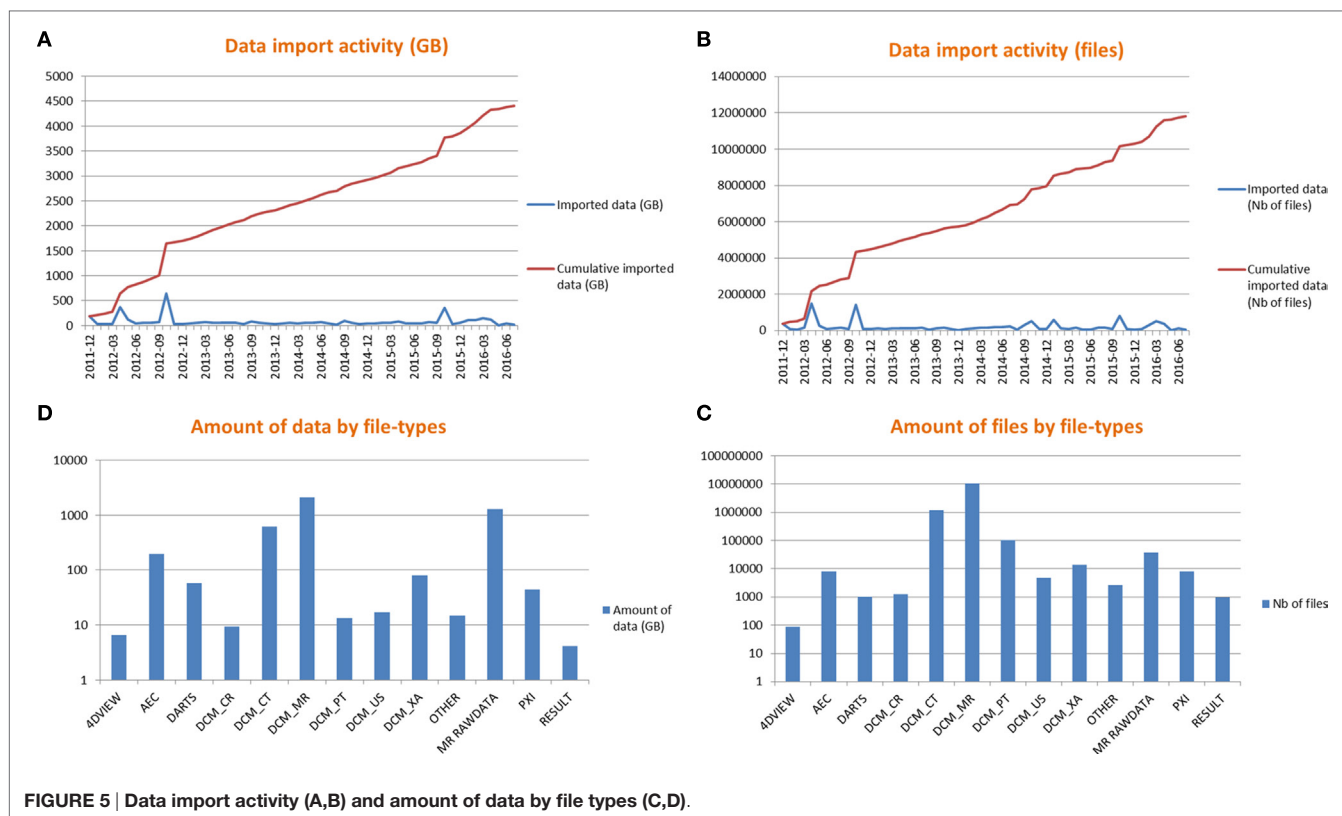
“MRI Methodology” CIC-IT Protocol (Local Study Code: 2008-0003)

Started in 2008 (before the deployment of the current ArchiMed version) and still active, this research protocol, designed to improve MRI technology and sequences, represents the most important activity using ArchiMed.

- Stored exams: 1,246
- MR Dicom files: ~3,000,000 (~300 GB)
- AEC (physiological signals) files: ~4,000 (~100 GB)
- MR raw data: ~14,000 (~600 GB)

On daily practice, images and associated files acquired from clinical MRIs are directly sent into ArchiMed temporary storage





through DICOM protocol or FTP connections, before being validated, identified (if necessary), and transferred to the corresponding study cache by project manager.

THRACE (Local Study Code: 2009-007)

Started in 2009, this multicenter protocol has been designed to assess effectiveness of endovascular mechanical thrombectomy for acute ischemic stroke (Bracard et al., 2016). CIC-IT of Nancy-France is responsible of collecting, archiving, and analyzing of all images from all the 26 participating centers.

- Centers: 26
- Stored exams: 1,402
- MR Dicom files: ~300,000 (~100 GB)
- CT Dicom files: ~400,000 (~200 GB)
- XA Dicom files (angiography): ~10,000 (~80 GB)

DICOM images were sent to CIC-IT *via* CD in earlier days, which is the source of many practical problems (especially concerning de-identification, missing data, or corrupted files) with a significant impact in term of time and resources.

The THRACE project experience has motivated the creation of Eureka for transferring files from external centers.

User Interface

Installed on more than 50 computers, the Java client application is the most used way to access ArchiMed. It has been designed

to be easy to use with a comprehensive user interface similar to those commonly used in clinical imaging software applications (Figure 6).

Plugins and External Software Connections

Interoperable HTTP service interface is already used by different external applications, developed either in C++, Java™ or Matlab™ environments such as “ORS Visual ArchiMed Loader” (loading DICOM from ArchiMed to “ORS Visual™²²” viewing and processing platform) or “Matlab ArchiMed Connector” (loading Dicom or other file into Matlab™²³ by querying ArchiMed) (Figure 7).

To extend ArchiMed client application functionalities, more than 20 plugins have already been developed for specific processing, case report form, statistical analysis, external database filling, and connection with other software programs, etc. (Figure 8).

See Table 1 for examples of plugins that have been already built.

²²ORS – Radiology Software, PACS, DICOM Viewer and Medical Imaging, <http://www.theobjects.com/en/>.

²³MATLAB™ – MathWorks, <http://mathworks.com/products/matlab/>.

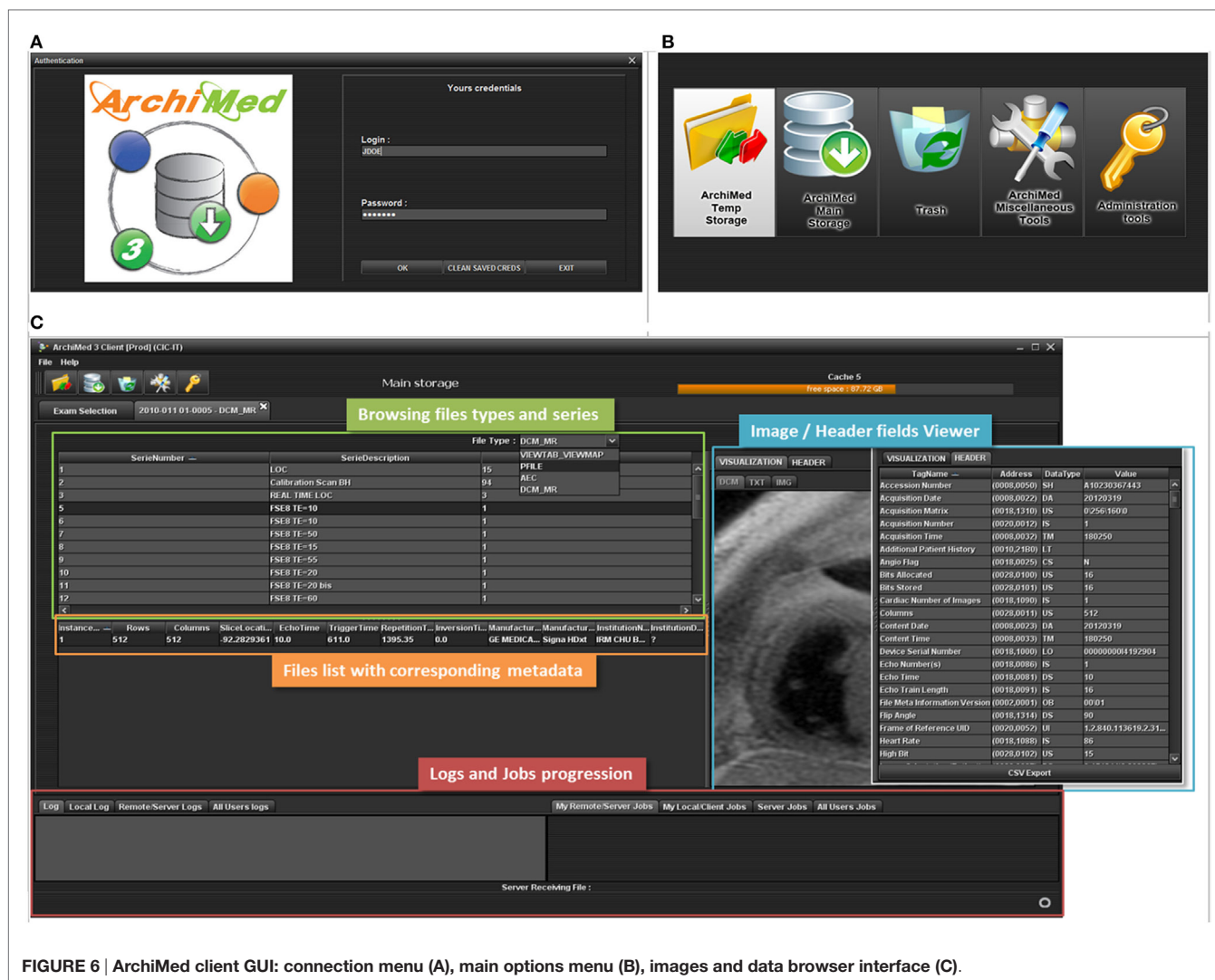


FIGURE 6 | ArchiMed client GUI: connection menu (A), main options menu (B), images and data browser interface (C).

DISCUSSION AND CONCLUSION

ArchiMed is a complete storage and visualization solution respecting legal requirements and research obligations. It is in function for 5 years at CIC-IT of Nancy-France and is a handy research data management system for our 60 staff, among whom are researchers, clinicians, and clinical trial project managers for local or multicenter studies.

Initially, based on the internal needs of our lab to safely store and easily access imaging data, it has met all our expectations and is used in more than 100 clinical protocols at CIC-IT of Nancy-France, France since 2011.

Collaborations

It turns out that the functional organization of clinical imaging data covered by ArchiMed is not only our own needs. Many labs and research centers specialized in imaging face the same issues

and are experiencing difficulties in finding a solution dedicated for research and meeting legal requirements for data preservation and confidentiality. This is why ArchiMed has won great attention from our partners in France. The very first external deployment of our ArchiMed system was with CIC-IT of Tours²⁴ under their request. Since 2015, ArchiMed has been under function for their local clinical protocols. More installations are currently being considered.

In the context of the French research infrastructure in imaging (France Life Imaging, FLI), we will interconnect all research data storage systems in France. The FLI-IAM²⁵ workgroup will use already existing data storage and information processing facilities

²⁴Clinical Investigation Center of Tours (France), <http://cic-it-tours.fr/>.

²⁵France Life Imaging – Information Analysis and Management (IAM) Node, <https://project.inria.fr/fli/en/>.

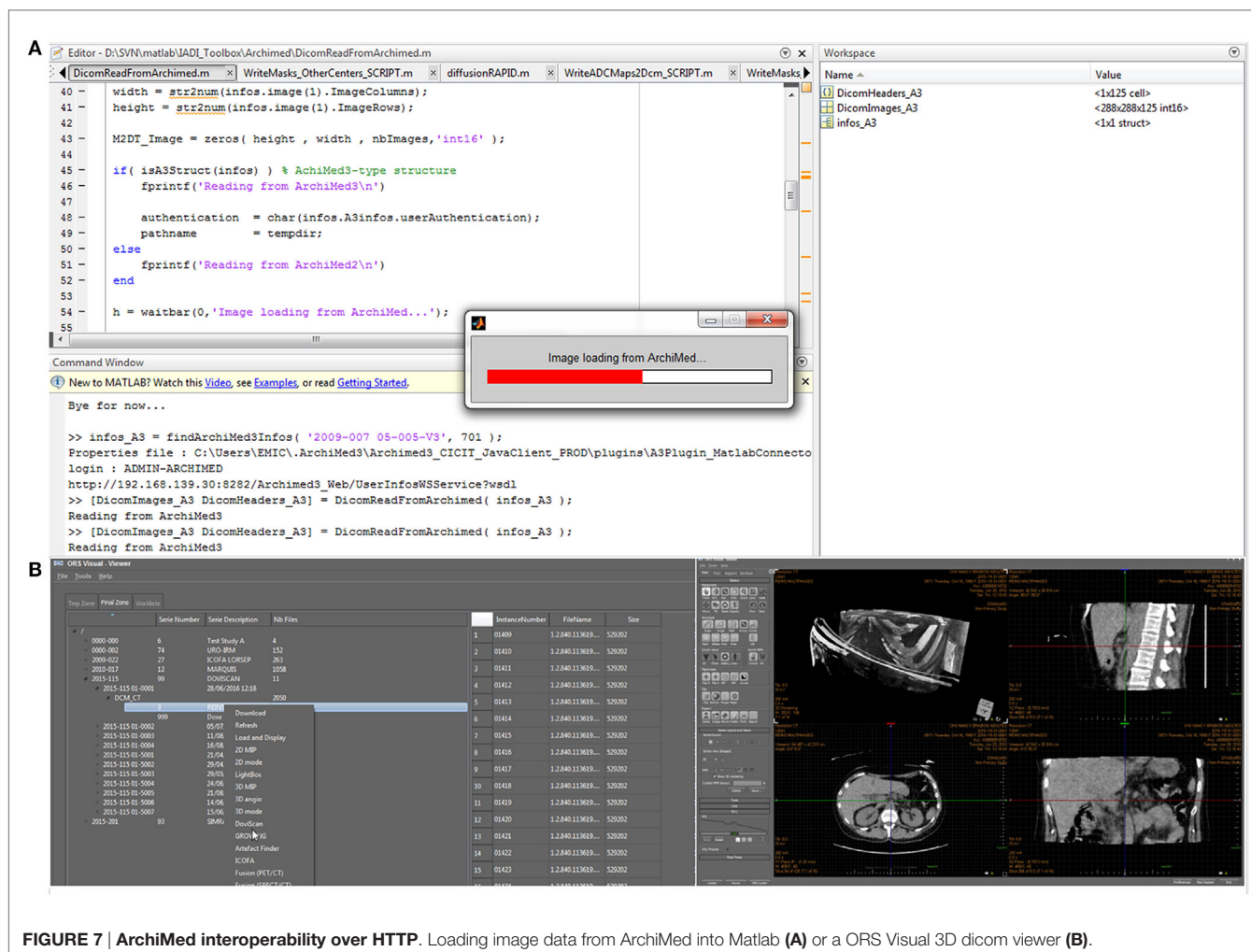


FIGURE 7 | ArchiMed interoperability over HTTP. Loading image data from ArchiMed into Matlab (A) or a ORS Visual 3D dicom viewer (B).

(including CATI,²⁶ Shanoir,²⁷ and ArchiMed) and will increase their capacities for the FLI infrastructure.

Current Limitations and Improvement About Metadata and Traceability

So far, metadata extraction and de-identification rules only depend on file types and are not related to a specific study. It can be problematic especially when we consider that some metadata can be relevant in a specific study context but unnecessary in another one (e.g., MR diffusion tensor directions will be an important factor for brain study while heart's beat parameter will be more relevant in cardiovascular study). For these reasons, we are thinking about building in a study profile which defines, by study, what header information will be extracted and available as metadata in the database and what header information will be deleted or de-identified.

Moreover, current version of our software only logs errors and information about data insertion/modification/delete (at exam node level). For recording user activities and tracking workflow better, it should be interesting to keep reports of data consultation/download by user and client application destination (processing tool, viewing platform, etc.).

Cloud

ArchiMed has been created to be a local solution, only accessible from a local area network by authorized and identified users. However, in the age of Cloud computing (Rimal et al., 2009), it will be proper to offer a secured and fully online version of ArchiMed. We are seriously considering upgrading it as a cloud service based on a Software as a service (SaaS)²⁸ (Levinson, 2007) model. Current service oriented architecture renders ArchiMed technically compatible with cloud infrastructure, but the impact

²⁶ CATI Neuroimaging, <http://cati-neuroimaging.com>.

²⁷ Shanoir (Sharing Neuroimaging Resources), <http://www.shanoir.org/>.

²⁸ SaaS (Software as a Service), https://en.wikipedia.org/wiki/Software_as_a_service.

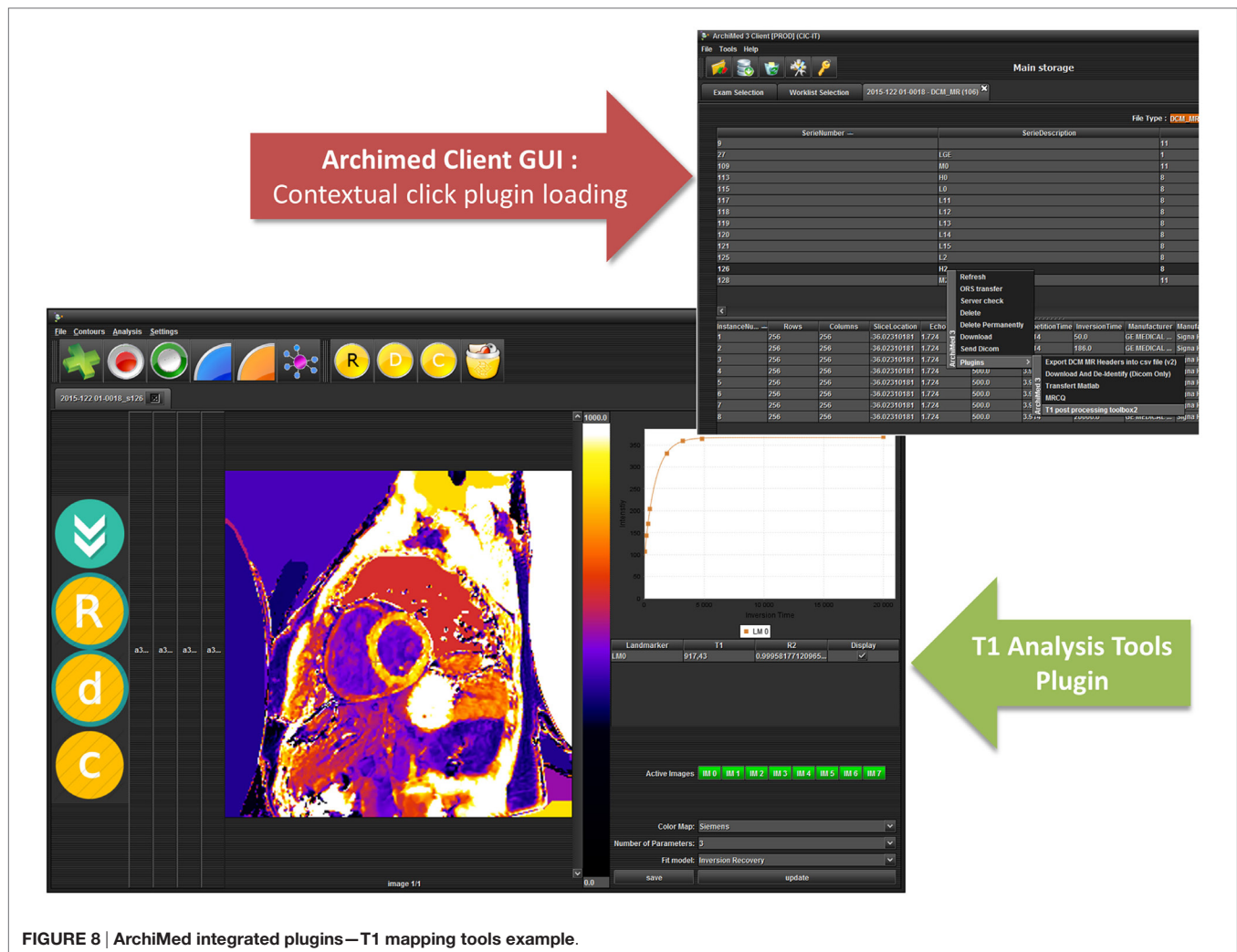


FIGURE 8 | ArchiMed integrated plugins—T1 mapping tools example.

TABLE 1 | Some existing ArchiMed plugins.

Plugin	Description
Dcm to Nifti converter	Convert a DICOM Series to Nifti format file
Matlab connector plugin	Connect with Matlab in two ways: push from ArchiMed client to Matlab or pull directly from server using web services
XXXX plugin	Exam electronic case report form and data analysis for XXXX protocol
PC analysis	Phase contrast—pulse wave velocity MR image analysis
FSL_FA	Compute ADC and FA maps from DICOM Series using FSL (brain imaging analysis)
DICOM export	Export images or movies from stored DICOM
MR quality control	Extract quality control parameters from a specific normalized protocol
DCM Series splitter	Split DICOM series according to specified criterions
App Launcher	Launch defined externals applications
XXXX pre-screening	Data check and reviewing plugin for XXXX protocol
Colormap display	Display DICOM images using different colormaps (useful for mapping)
Download and de-identify	Download DICOM files/series/exams with de-identifying custom tags

of external storage in terms of data security, confidentiality, and legal obligations have to be taken into account.

In conclusion, ArchiMed is a well-adapted research PACS. It is working for more than 5 years at CIC-IT of Nancy-France and perfectly matches with clinical research needs in terms of workflow, organization, legal requirement, and usability.

GLOSSARY

CNIL—Commission Nationale de l'Informatique et des Libertés (National Commission on Informatics and Liberty) is an independent French administrative regulatory body whose mission is to ensure that data privacy law is applied to the collection, storage, and use of personal data.

DICOM—Digital Imaging and Communications in Medicine is a standard for handling, storing, printing, and transmitting information in medical imaging. It includes a file format definition and a network communications protocol.

FTP—The File Transfer Protocol is a standard network protocol used to transfer computer files between a client and server on a computer network.

HTTP—The Hypertext Transfer Protocol is an application protocol for distributed, collaborative, hypermedia information systems.

Java™—Java is a general-purpose computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible.

Java-EE™—Java Platform, Enterprise Edition is a widely used enterprise computing platform developed under the Java Community Process.

LDAP—The Lightweight Directory Access Protocol is an open, vendor-neutral, industry standard application protocol for accessing and maintaining distributed directory information services over an Internet Protocol (IP) network.

MRI—Magnetic resonance imaging is a medical imaging technique used in radiology to image the anatomy and the physiological processes of the body.

PACS—Picture Archiving and Communication System is a medical imaging technology which provides economical storage and convenient access to images from multiple modalities (source machine types).

SOAP—Simple Object Access Protocol is a protocol specification for exchanging structured information in the implementation of web services in computer networks.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We acknowledge all members of CIC-IT of Nancy-France to taking part in the different tests and discussions regarding software improvements and debug.

REFERENCES

- Bracard, S., Ducrocq, X., Mas, J. L., Soudant, M., Oppenheim, C., Moulin, T., et al. (2016). Mechanical thrombectomy after intravenous alteplase versus alteplase alone after stroke (THRACE): a randomised controlled trial. *Lancet Neurol.* 15, 1138–47. doi:10.1016/S1474-4422(16)30177-6
- Choplin, R. H., Boehme, J. M., and Maynard, C. D. (1992). Picture archiving and communication systems: an overview. *Radiographics* 12, 127–129. doi:10.1148/radiographics.12.1.1734458
- Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., et al. (1999). *Hypertext Transfer Protocol – HTTP/1.1*. Available at: <https://www.rfc-editor.org/info/rfc2616>
- Goncalves, A. (2009). *Beginning Java EE 6 Platform with GlassFish 3: From Novice to Professional*, 1st Edn. APress.
- Koutsonikola, V., and Vakali, A. (2004). LDAP: framework, practices, and trends. *IEEE Internet Comput.* 8, 66–72. doi:10.1109/MIC.2004.44
- Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., and Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* 50(Suppl.), S82–S101. doi:10.1097/MLR.0b013e3182585355
- Levinson, M. (2007). *Software As a Service (SaaS) Definition and Solutions*. CIO. Available at: <http://www.cio.com/article/2439006/web-services/software-as-a-service-saas-definition-and-solutions.html>
- Odille, F., Vuissoz, P.-A., Marie, P.-Y., and Felblinger, J. (2008). Generalized reconstruction by inversion of coupled systems (GRICS) applied to free-breathing MRI. *Magn. Reson. Med.* 60, 146–157. doi:10.1002/mrm.21623
- Postel, J., and Reynolds, J. (1985). *File Transfer Protocol*. Available at: <https://tools.ietf.org/html/rfc959>
- Rimal, B. P., Choi, E., and Lumb, I. (2009). “A taxonomy and survey of cloud computing systems,” in *Fifth International Joint Conference on INC, IMS and IDC, 2009. NCM '09* (Seoul: IEEE), 44–51.
- Tucker, K., Branson, J., Dilleen, M., Hollis, S., Loughlin, P., Nixon, M. J., et al. (2016). Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med. Res. Methodol.* 16(Suppl. 1):77. doi:10.1186/s12874-016-0169-4
- van de Wetering, R., Batenburg, R., Versendaal, J., Lederman, R., and Firth, L. (2006). A balanced evaluation perspective: picture archiving and communication system impacts on hospital workflow. *J. Digit. Imaging* 19, 10–17. doi:10.1007/s10278-006-0628-2
- Zeilenga, K. (2006). *Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map*. Available at: <https://tools.ietf.org/html/rfc4510.html>

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Micard, Husson, CIC-IT Team and Felblinger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Cyberinfrastructure for Open Science at the Montreal Neurological Institute

Samir Das^{1,2*}, Tristan Glatard³, Christine Rogers^{1,2}, John Saigle², Santiago Paiva^{1,4}, Leigh MacIntyre^{1,2}, Mouna Safi-Harab^{1,2}, Marc-Etienne Rousseau^{1,2}, Jordan Stirling^{1,2}, Najmeh Khalili-Mahani^{1,2,3}, David MacFarlane^{1,2}, Penelope Kostopoulos^{1,2}, Pierre Rioux^{1,2}, Cecile Madjar⁵, Xavier Lecours-Boucher^{1,2}, Sandeep Vanamala², Reza Adalat^{1,2}, Zia Mohaddes^{1,2}, Vladimir S. Fonov^{2,4}, Sylvain Milot^{2,4}, Ilana Leppert^{2,4}, Clotilde Degroot², Thomas M. Durcan², Tara Campbell^{1,2}, Jeremy Moreau^{2,4}, Alain Dagher^{1,4}, D. Louis Collins^{2,4}, Jason Karamchandani², Amit Bar-Or², Edward A. Fon², Rick Hoge^{2,4}, Sylvain Baillet^{2,4}, Guy Rouleau² and Alan C. Evans^{1,2}

OPEN ACCESS

Edited by:

Michel Dojat,
Institut National de la Santé et de la
Recherche Médicale (INSERM),
France

Reviewed by:

Graham J. Galloway,
Translational Research Institute,
Australia
Xi-Nian Zuo,
Chinese Academy of Sciences, China
Bernard Gibaud,
Institut National de la Santé et de la
Recherche Médicale (INSERM),
France

*Correspondence:

Samir Das
samir.das@mcgill.ca

Received: 31 August 2016

Accepted: 01 December 2016

Published: 06 January 2017

Citation:

Das S, Glatard T, Rogers C, Saigle J, Paiva S, MacIntyre L, Safi-Harab M, Rousseau M-E, Stirling J, Khalili-Mahani N, MacFarlane D, Kostopoulos P, Rioux P, Madjar C, Lecours-Boucher X, Vanamala S, Adalat R, Mohaddes Z, Fonov VS, Milot S, Leppert I, Degroot C, Durcan TM, Campbell T, Moreau J, Dagher A, Collins DL, Karamchandani J, Bar-Or A, Fon EA, Hoge R, Baillet S, Rouleau G and Evans AC (2017) Cyberinfrastructure for Open Science at the Montreal Neurological Institute. *Front. Neuroinform.* 10:53. doi: 10.3389/fninf.2016.00053

¹ McGill Centre for Integrative Neuroscience, Montreal Neurological Institute, Montreal, QC, Canada, ² Montreal Neurological Institute, Montreal, QC, Canada, ³ Department of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada, ⁴ McConnell Brain Imaging Centre, Montreal Neurological Institute, Montreal, QC, Canada, ⁵ Douglas Mental Health University Hospital, Montreal, QC, Canada

Data sharing is becoming more of a requirement as technologies mature and as global research and communications diversify. As a result, researchers are looking for practical solutions, not only to enhance scientific collaborations, but also to acquire larger amounts of data, and to access specialized datasets. In many cases, the realities of data acquisition present a significant burden, therefore gaining access to public datasets allows for more robust analyses and broadly enriched data exploration. To answer this demand, the Montreal Neurological Institute has announced its commitment to Open Science, harnessing the power of making both clinical and research data available to the world (Owens, 2016a,b). As such, the LORIS and CBRAIN (Das et al., 2016) platforms have been tasked with the technical challenges specific to the institutional-level implementation of open data sharing, including:

- (1) Comprehensive linking of multimodal data (phenotypic, clinical, neuroimaging, biobanking, and genomics, etc.)
- (2) Secure database encryption, specifically designed for institutional and multi-project data sharing, ensuring subject confidentiality (using multi-tiered identifiers).
- (3) Querying capabilities with multiple levels of single study and institutional permissions, allowing public data sharing for all consented and de-identified subject data.
- (4) Configurable pipelines and flags to facilitate acquisition and analysis, as well as access to High Performance Computing clusters for rapid data processing and sharing of software tools.
- (5) Robust Workflows and Quality Control mechanisms ensuring transparency and consistency in best practices.
- (6) Long term storage (and web access) of data, reducing loss of institutional data assets.
- (7) Enhanced web-based visualization of imaging, genomic, and phenotypic data, allowing for real-time viewing and manipulation of data from anywhere in the world.

- (8) Numerous modules for data filtering, summary statistics, and personalized and configurable dashboards.

Implementing the vision of Open Science at the Montreal Neurological Institute will be a concerted undertaking that seeks to facilitate data sharing for the global research community. Our goal is to utilize the years of experience in multi-site collaborative research infrastructure to implement the technical requirements to achieve this level of public data sharing in a practical yet robust manner, in support of accelerating scientific discovery.

Keywords: neuroimaging, big data, open science framework, cyberinfrastructure, neuroscience, data sharing, bids, workflow

INTRODUCTION

The challenge of reproducibility in science (Campbell, 2016) has compelled the neuroscience research community to adopt new approaches to ensure scientific reliability without impeding innovation. The recent commitment by the Montreal Neurological Institute (MNI) to Open Science aims to improve replicability and transparency in research through collaboration, and in doing so, accelerate scientific discovery (Owens, 2016a,b).

The MNI's Open Science initiative calls for the free release of research data, findings, analytical tools, and publications from MNI-based researchers. Institutional sharing aims to prevent data loss, increase sample size and statistical power, and reduce acquisition costs by encouraging data re-use (thereby maximizing returns on public funding). In addition to these advantages, inviting external researchers to access these institutional resources will expand the reach and impact of research conducted at the institute (Poldrack and Gorgolewski, 2014).

Open Science initiatives have been spearheaded within the bioinformatics and neuroscience communities by groups such as the Center for Open Science (Asante et al., 2016), the Allen Institute (Koch and Jones, 2016), the Human Connectome Project (Van Essen et al., 2012), OpenfMRI (Poldrack et al., 2013), the Consortium for Reliability and Reproducibility (CoRR) (Zuo et al., 2014), and a multitude of independent data sharing and open-source academic software initiatives such as BrainHack (Craddock et al., 2016), Brainstorm (Baillet et al., 2011), SPM (Friston et al., 1994), FSL (Jenkinson et al., 2012), ADNI (Petersen et al., 2010), Nipype (Gorgolewski et al., 2011), and BigBrain (Amunts et al., 2013). At the same time, emerging definitions of common data sharing standards, practices, and formats are being established via BIDS (Gorgolewski et al., 2016), the Neuro-Imaging Data Model (NIDM) (Maumet et al., 2016), FAIR principles (Wilkinson et al., 2016) and even extending to data organization and citation strategies (Honor et al., 2016). Meanwhile, governments and funding agencies in the USA (National Institutes of Health, 2014; National Institute of Mental Health, 2015), Canada (Tri-Agency Statement of Principles of Digital Data Management, 2016), Europe (Horizon 2020, The Wellcome Trust, 2016) and elsewhere encourage and increasingly require research programs to establish data management and

sharing plans from the start of the research data lifecycle. Despite these efforts, such initiatives are frequently constrained to particular projects or focused collaborations rather than institutional initiatives, as the sharing of data often remains at the discretion of individual investigators whose technical resources and expertise in data infrastructure may be limited.

As the first leading academic research institution to develop an Open Science framework at the institutional level¹, the MNI's cyberinfrastructure platform will play a critical role in this initiative. To fulfill this vision, several key implementational challenges must be met, including policy, security, and ethics, as well as infrastructural design, software interoperability, data harmonization, validation, processing, and provenance capture. The solutions to these issues must adhere to open data sharing principles and respect domain-specific best practices (Honor et al., 2016; Nichols et al., 2016; Wilkinson et al., 2016).

For effective data sharing at an institutional level, it is imperative to use a cyberinfrastructure that can incorporate heterogeneous datasets acquired from multiple sources over time as well as across modalities – and to do so in a way that is robust. Data collected by investigators in multiple studies across the institute span diverse data types from many domains, including clinical/behavioral measures, biological samples from the MNI biobanking collections, genomic data, and a growing multimodal repository of brain imaging data. The institutional cyberinfrastructure housing these datasets must also be able to integrate workflows from all stages of the research data lifecycle, and interoperate with platforms that capture and disseminate large datasets.

To this end, the MNI has selected LORIS (Das et al., 2011) to serve as the core data management platform for this initiative, coupled to the CBRAIN distributed high-performance computing environment (Sherif et al., 2014). These two platforms, combined with embedded data visualization utilities (Sherif et al., 2015), constitute an “ecosystem” capable of supporting Open Science at an institutional level (Das et al., 2016).

¹Open Science (Open Access). HORIZON 2020, The EU Framework Programme for Research and Innovation. Retrieved from <https://ec.europa.eu> (Accessed on August 29, 2016).

This paper describes the ethical and policy challenges, the technical infrastructure used for storage and curation of the various data types, and the workflows and processing environment for the implementation of Open Science at the MNI.

METHODS

Four cornerstones of the MNI's Open Science framework and cyberinfrastructure are discussed below: (1) ethics (including subject privacy, consent and security), (2) multi-modal data entry, (3) workflows and quality control, and (4) high-performance data processing and software-systems interoperability.

Ethics, Privacy and Security

Embarking on the endeavor of institutional Open Science poses unique challenges, particularly with regard to respecting ethical guidelines. One critical component is that personally identifiable information (PII) of all subjects must be protected and the data itself must be de-identified and secured within the context of private and independent databases—but will also be reconcilable into a single subject record in the Open Science platform.

Since the creation of the first human cell-line (Lucey et al., 2009), the ethical considerations surrounding the distribution and use of human subject data have been manifold (Nelson, 2015). In accordance with local Quebec law and research ethics, informed consent must be obtained from subjects in order to collect and study tissue and data. The Canadian Tri-Council has also provided clear criteria to protect the privacy of subjects, and these criteria must be met in order for researchers to have access to sensitive data (Canadian Institutes of Health Research Natural Sciences and Engineering Research Council of Canada and Social Sciences and Humanities Research Council of Canada, 2014). Accordingly, a proposal was submitted and approved by the MNI Research Ethics Board (REB) for the Neuro OpenScience Clinical Biologic Imaging and Genetic Repository, or C-BIG-R, addressing the implementation of an infrastructure technically compatible with these ethics policies. A dual-level governance structure was created to oversee these ethical concerns via the REB as well as a newly-established "Tissue and Data" committee. The REB is tasked with the identification of best practices employed by comparable initiatives, and the Tissue and Data committee is responsible for determining what materials are deposited into the bank, the storage mechanisms, and how they can be accessed for research. Participating studies may profit from this governance model throughout the research data lifecycle, since matters of storage, security, inclusion, and exclusion criteria, disposal of samples etc., will already be covered by this ethical framework.

Data sharing at any level requires nuanced procedures and consent processes, and involves particular technological constraints. These technical considerations include how to share data (i) within a single study as well as (ii) between collaborating investigators, and finally (iii) at an institutional and public level such that subject data from multiple studies are linkable and queryable in a unified manner. From its

inception, the MNI's platform design allows researchers to first store and share data internally and privately, while ultimately allowing data to be selectively pushed to the public-facing platform for dissemination (**Figure 1**). Both de-identification and reconciliation of subject records must be carefully designed in view of the Open MNI platform.

De-identification of subject data is an integral requirement: the identifier must ensure privacy and ethically-compliant data sharing, while also preventing data duplication. For this purpose, a system of hashed identifiers has been designed to safeguard subject identity at every stage and prevent reconstructive subject identification. This process encodes identifying information and is incorporated into LORIS such that PII is never transmitted over a network; only the encoded information is used (**Figure 2**).

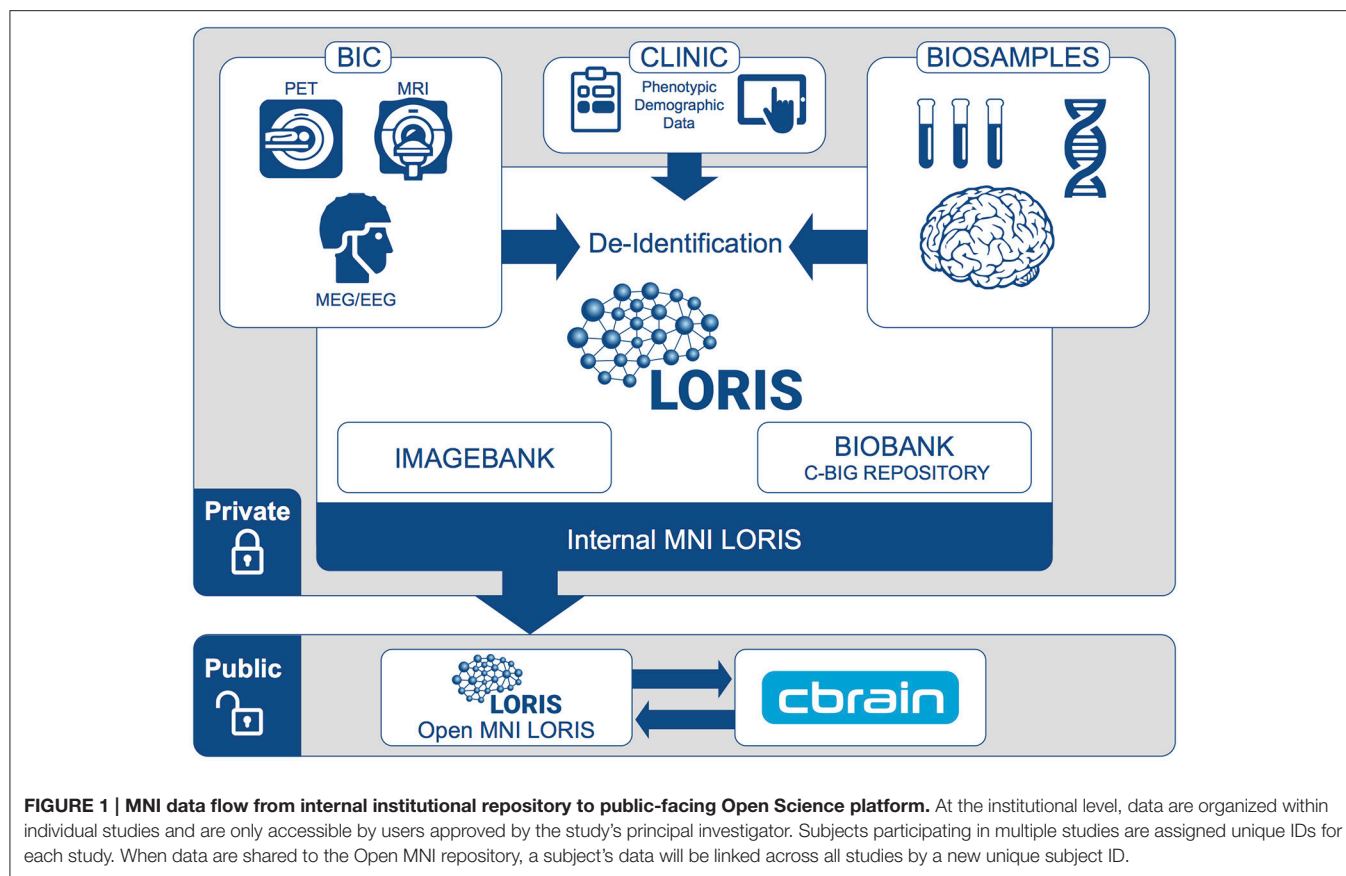
A one-way cryptographic hash function is employed to uniquely refer to individual subjects without revealing any of their identifying information. A given subject's first, middle and last names, date of birth and mother's maiden name are concatenated and passed through the PBKDF2² ("Password-Based Key Derivation Function 2") algorithm to generate a unique hash value, created by iteratively applying a SHA1³ (Secure Hash Algorithm 1) hashing function one million times. The resulting hashed value (a 125-character string) is then mapped onto a unique MNI-internal identifier (e.g., "StudyA1007"), distinctly generated for every study in which the subject is a participant. These study-specific identifiers can be disseminated without compromising the subject's privacy. The internal hash is only accessible by database administrators and is therefore also kept secret within the institution.

Research platforms or researchers that have access to a subject's private information will never store PII directly in the database; rather, they will automatically trigger this hashing function when registering subject data in LORIS. The function was selected for its efficiency given a sufficiently short execution time to perform mass registration of data, yet long enough such that brute-force attackers cannot identify subjects by repeated attempts to guess subject names. The entire process of hashing takes approximately 7 seconds on a current CPU.

Datasets can be shared (at the owner's discretion) by uploading to the public-facing Open MNI repository. The sharing process entails additional data curation steps for further de-identification, such as transforming images via de-facing to avoid identification based on facial features (Bischoff-Grethe et al., 2007). Another of these transformations is an encryption performed on the locally hashed identifiers. This encrypted hash is used to detect non-unique subjects for the sole purpose of avoiding redundancy (i.e., same subject appearing in different datasets). When a subject is determined to be unique within the Open Science repository, they are assigned a unique public ID which unifies their de-identified data from disparate studies.

²PBKDF2 is a key derivation function that applies a pseudo-random function to a specified input, repeating the process multiple times, to produce a derived key (<https://en.wikipedia.org/wiki/PBKDF2>).

³SHA-1 a cryptographic hash function designed as a one-way function to map data of arbitrary size to a fixed data size, making it unfeasible to invert. It is considered a U.S. Federal Information Processing Standard (<https://en.wikipedia.org/wiki/SHA-1>).



In the event that a subject revokes consent, a database administrator has the capability of removing that subject from the Open MNI LORIS database using the unique public subject ID. Upon revocation, the physical data as well as the computer records will be destroyed and deleted. However, any derived datasets or results obtained through the analysis of biospecimens and data for which consent has been withdrawn will not be destroyed. This process complies with NIH-NDA standards and methodology regarding Global Unique Identifiers (Johnson et al., 2010), and is explicitly outlined in the biobank consent form.

Loris Functionality: Multi-Modal Data Entry, Provenance, Storage, and Linking

The LORIS system (Das et al., 2011, 2016) was designed specifically for heterogeneous data acquisition, curation and dissemination. It is a web-based PHP/MySQL database, freely available on GitHub⁴ as open-source software. Its modular organization and support for multiple data modalities (including behavioral/clinical, neuroimaging, and genetic summary data) provide a flexible and robust platform for many types of multi-site studies and projects.

Within LORIS, data are organized based on subject profiles and longitudinal data-collection timepoints within a given study. After creating a de-identified profile of a subject, multiple modalities of data are associated to that subject and their

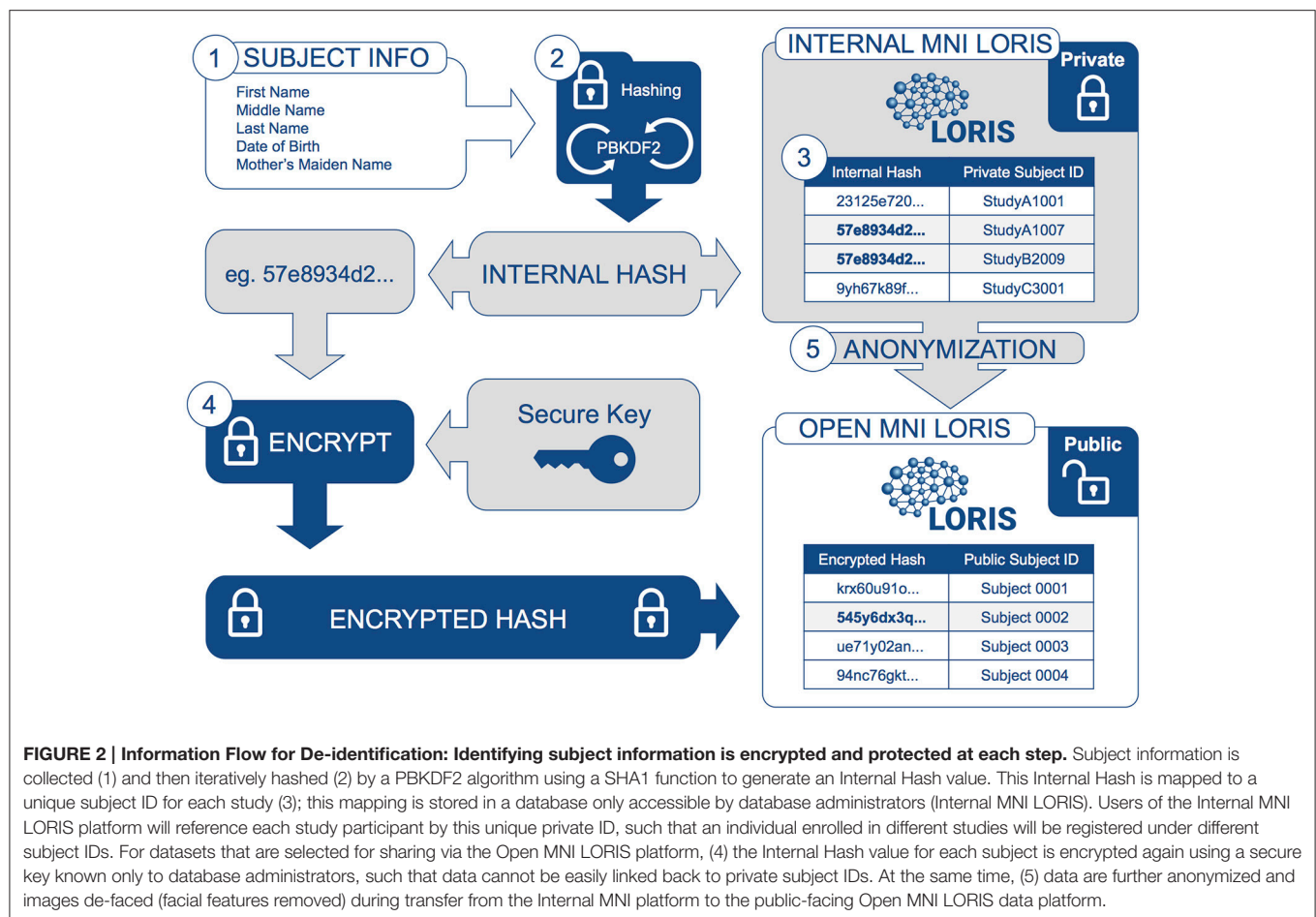
corresponding timepoints. For example, data collected at a particular subject timepoint may include the acquisition of MRI and PET volumes, a collection of biospecimens, and a variety of other clinical measures. All of this information is associated to the subject within LORIS and can be easily retrieved, reviewed, and exported.

Data can be imported into LORIS from external software systems, such as laboratory information management systems (LIMS) that handle sample registration, tracking, and storage. Such systems export data in various formats, demonstrate different data transfer capabilities, and implement varying configurations in their Application Programming Interfaces (APIs). To ensure interoperability across this diverse range of systems, a series of processing scripts have been created in order to bridge the gap between LORIS and the heterogeneous outputs of these platforms.

Importation of data is best illustrated through examples from two contexts: imaging volumes and biospecimen information. The transfer, insertion and processing of imaging data is performed via a sequence of open-source scripts⁵ native to the LORIS platform. These scripts form a software “pipeline” that is installed on the server to automate the pre-processing and insertion of imaging datasets. In addition, a web-based imaging uploader integrated with these server-side scripts handles image uploading, filename anonymization validation, and interactive

⁴<https://github.com/aces/Loris>

⁵<https://github.com/aces/Loris-MRI>



flagging of protocol verification checks. Once loaded in the database, imaging volumes become searchable and sortable in the Imaging Browser module. 3D visualization of volumes and morphological surfaces is natively embedded in the interface via the BrainBrowser⁶ tool used for quality control review of images (Sherif et al., 2015).

Another approach is presently being explored for LORIS to directly import multimodal data organized according to the emerging BIDS convention (Gorgolewski et al., 2016): data volumes would be pushed automatically from their respective acquisition sources (MRI scanners, PET cameras, MEG, and EEG arrays) into a central BIDS-compliant file system. This consists of structured folders containing raw and metadata information in simple JSON files. The new data entries would then be systematically imported and registered into the database after being detected by an automated daemon process that monitors further updates to the BIDS system.

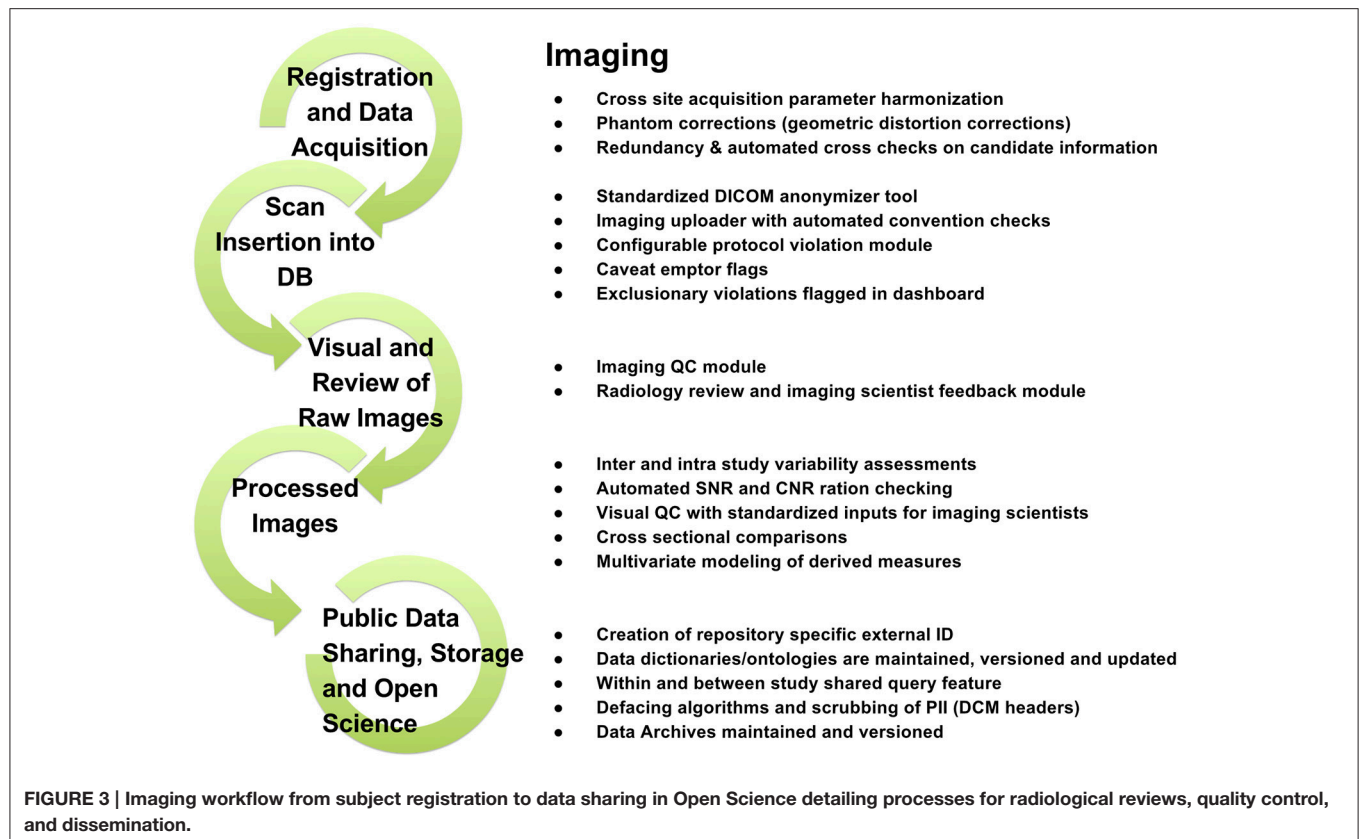
For biospecimen data, a similar automated workflow has been implemented. Biosamples are collected and processed in a lab, at which time information about the sample collected (e.g., sample type, date of collection, etc.) and its current status (e.g., stage of processing, storage location) are registered within a third-party

LIMS data system. Custom scripts are used to extract data based on archives of these data systems, simultaneously converting and normalizing the data for use within LORIS.

Once data are acquired and loaded in LORIS (through either manual data entry or automated pipeline scripts), researchers will be able to review and curate information using quality control tools and procedures assuring quality inputs to their analysis pipelines. Following data acquisition, review and curation, researchers can download, query, and disseminate datasets via LORIS' Data Querying Tool (DQT) which is built on a NoSQL framework (Katz et al., 2005) to enable fast and precise extraction of large datasets. Via the DQT, users can construct complex queries and apply custom filters in order to target populations and subsets of interest.

Common data description vocabularies are required to properly address the challenges of Open Science at a large scale. However, implementing a common vocabulary covering the range of concepts involved in studies conducted across the MNI will be a significant undertaking, and will be driven by the MNI's researchers as they seek to share their data in a common Open Science framework; convergence upon a usable solution will be challenging. LORIS is committed to the standardization of ontologies, and currently adopts a practical approach where (1) all the (DICOM) fields related to imaging data are preserved and

⁶<https://github.com/aces/brainbrowser>



made queryable, and (2) terms used for behavioral variables and biobanking studies are defined on a study-by-study basis, while their re-utilization is also promoted across studies, compliant (where possible) with conventions such as BIDS (Gorgolewski et al., 2016) or NDAR (Hall et al., 2012). Prospectively, LORIS plans to adopt ontologies under development by the NIDM initiative to formally and uniformly describe raw data, terms, workflows and derived data (Maumet et al., 2016), as well as open data citation standards such as those developed for neuroimaging (Honor et al., 2016). Further integration of domain-specific standards, such as MIABIS 2.0 developed for biobanking data by the BBMRI-ERIC network (Merino-Martinez et al., 2016), is a priority for integration of data dissemination formats for the Open Science platform.

Workflows and Quality Control for Imaging, Clinical/Behavioral and Biobanking

To support data review processes, multiple tiers of quality control tools are embedded in LORIS, enabling researchers to standardize data collection, which in turn facilitates reproducible results and compatible data-sharing in an Open Science environment. Validating the reliability of assessments for data collected at different sites and over time enables researchers to control for variability (Van Essen et al., 2013; Ducharme et al., 2015; Orban et al., 2015). **Figures 3–5** show domain-specific procedures that allow for data to be both standardized within a study and across studies in the context of Open Science for imaging (**Figure 3**),

biobanking (**Figure 4**), and clinical/behavioral (**Figure 5**) data collection.

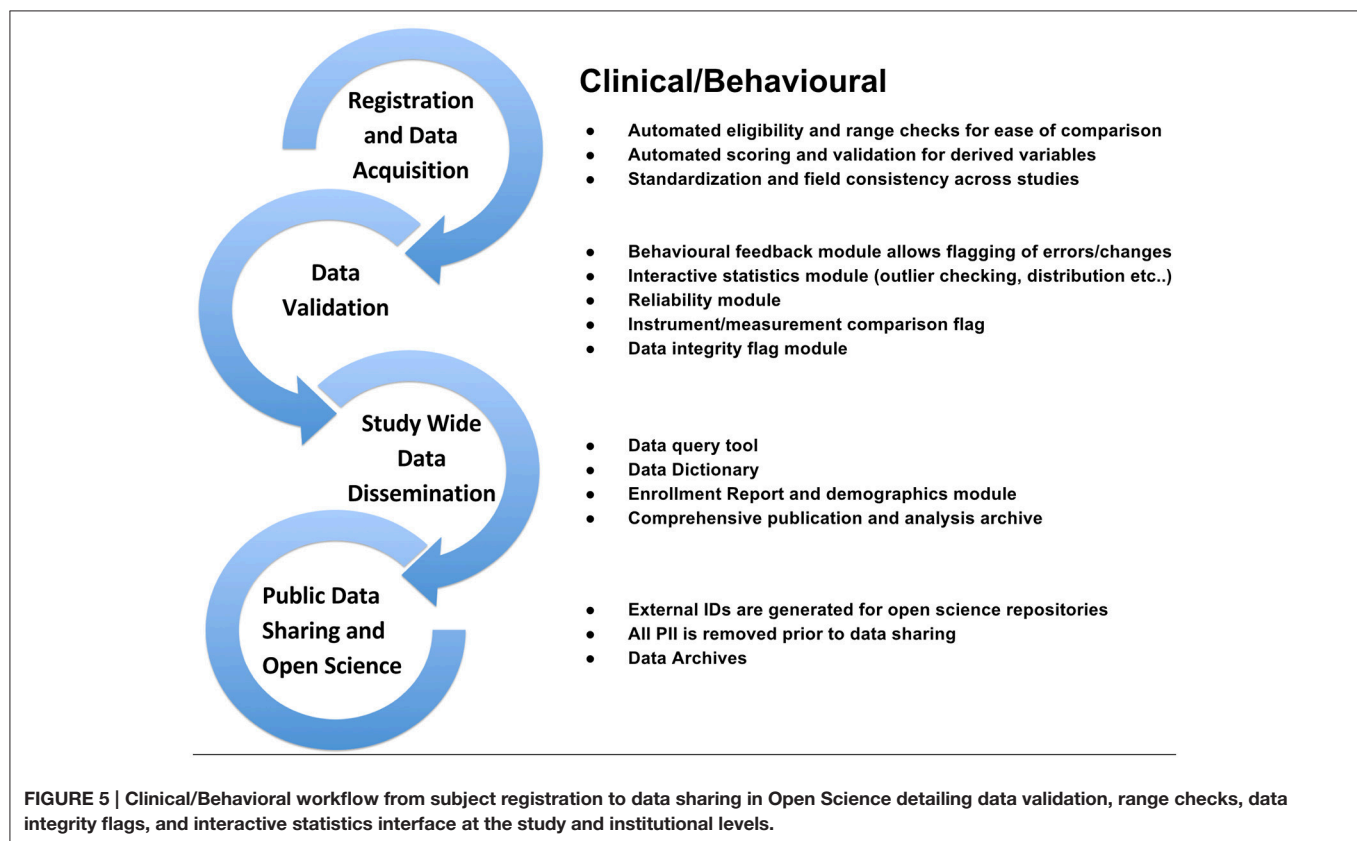
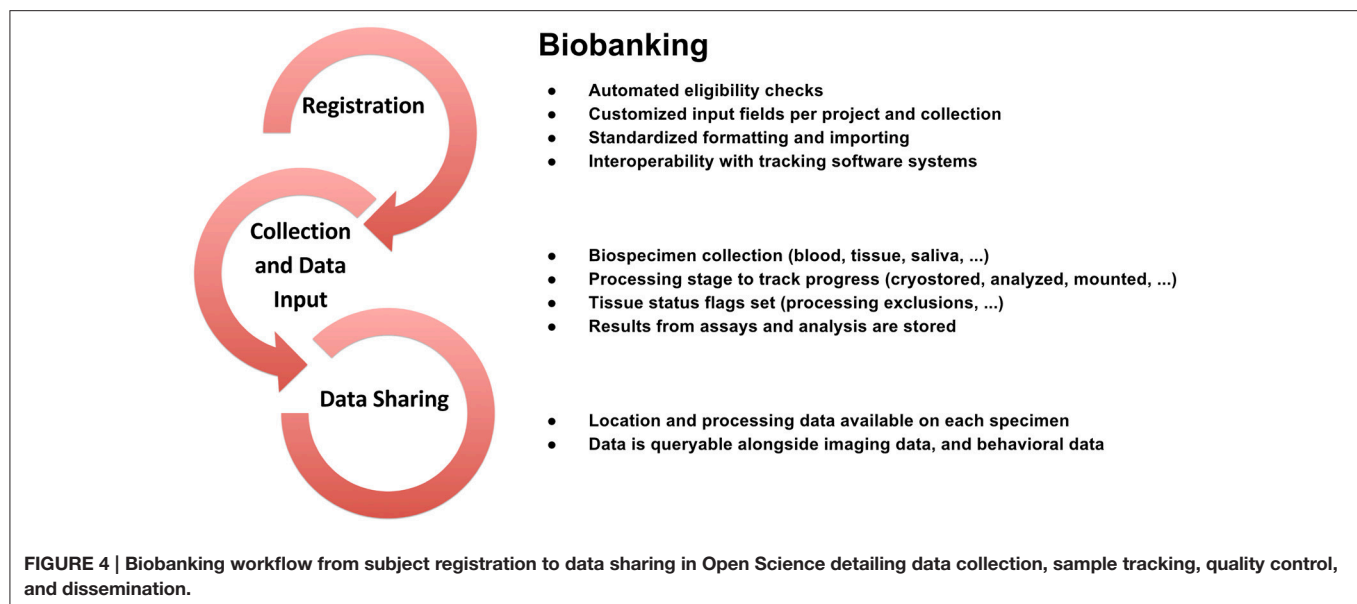
LORIS implements these new frameworks, techniques, and procedures, both automatic and manual, to ensure that the integrity, validity and reliability of data are not compromised from the collection stage through to data sharing.

High-Performance Data Processing

Open Science at the MNI is further facilitated by the interface between LORIS and CBRAIN's high performance computing (HPC) capabilities (Das et al., 2016). CBRAIN is a web-based collaborative research platform developed in response to the challenges raised by data-heavy, computationally-intensive neuroimaging research (Sherif et al., 2014). It offers transparent access to remote data sources, distributed computing sites, and an array of processing and visualization tools within a controlled, secure environment. The framework code is entirely open-source and available on GitHub⁷.

CBRAIN promotes Open Science in several ways by providing: (1) web access to a wide range of data processing pipelines, (2) an API open to other systems such as LORIS, (3) a full provenance trail of software versions, processing logs and all data manipulations, (4) strong security features, (5) a mechanism of tool containers and descriptors to facilitate the integration and open distribution of new analysis tools/pipelines (Glatard et al., 2015), and (6) connections to new private or shared data sources

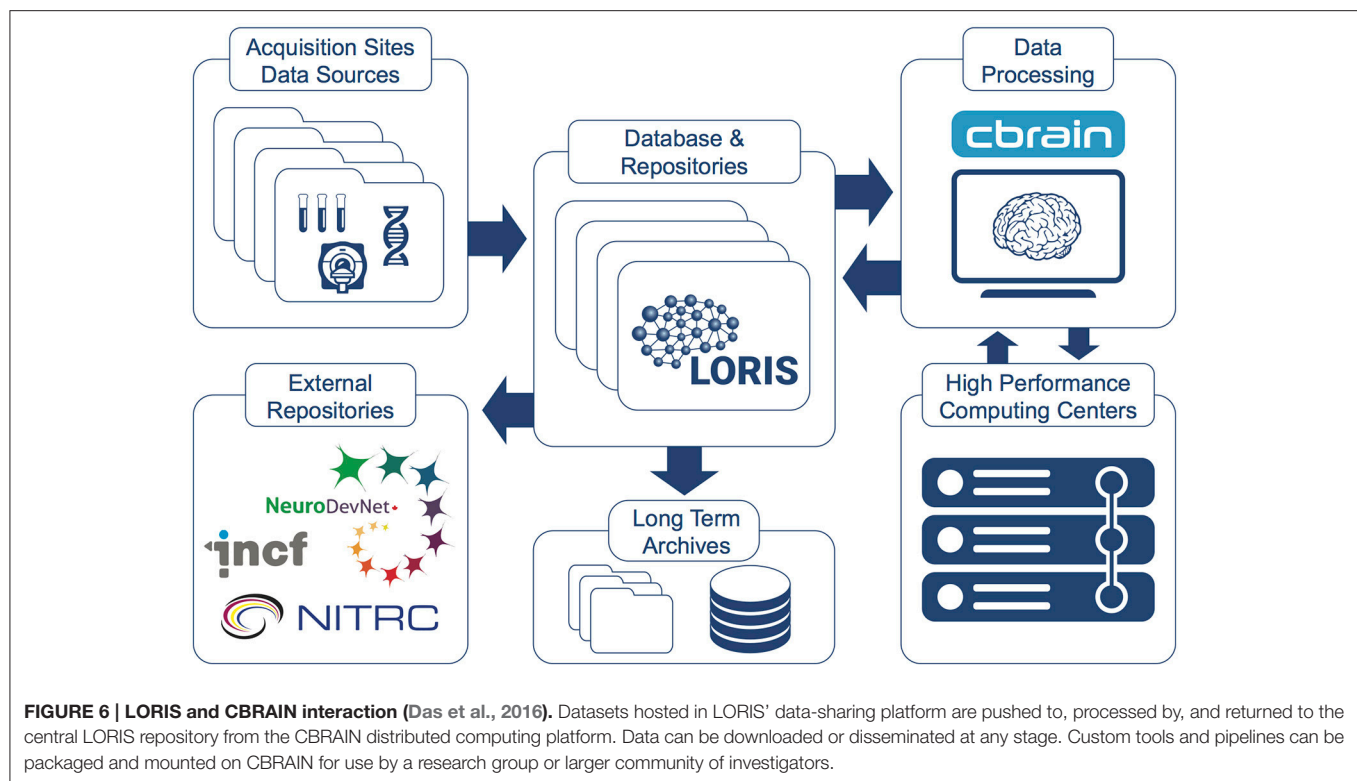
⁷<https://github.com/aces/cbrain>



for research groups. An overview of CBRAIN's integration with LORIS is shown in **Figure 6** and further detailed in "The MNI data-sharing and processing ecosystem" (Das et al., 2016).

While LORIS stores and manages the data gathered and distributed by the institute, the CBRAIN platform provides an interface to the tools and high-performance computing and

processing capabilities needed by the researchers. CBRAIN and LORIS each have APIs and can be connected such that data files managed by LORIS can be transferred to CBRAIN and processed on its computing ecosystem. When submitted workloads are completed, the resulting data files can be transferred back and registered in LORIS under the proper subject profile. This



eliminates the complexity of manual multi-site data transfers and saves the researchers from having to deal with the peculiarities of each computing center (e.g., queuing system and library environment, site policies, number of usable cores per nodes, queue limits, downtime, etc.).

A built-in mechanism allows for extensive provenance recording of any entity managed by CBRAIN, in particular for all operations on files, tasks, user groups, data, and computing resources, as well as the full standard and error logs provided by the analysis tools during processing. This audit trail is essential to ensure future reproducibility of results, and is also useful for troubleshooting and debugging.

CBRAIN's capabilities integrate well with the institutional requirements of privacy when dealing with files that are not yet openly releasable. All CBRAIN data traffic to and from the high performance computing centers is encrypted. Secure connections between authorized resources are transient, and temporary files can be configured to be automatically purged after processing is finished. Fine-grained access rights can be defined on any data file via user groups. Strict access permissions can also be defined for complete data servers, for analysis tools and for computation sites.

Extensibility is an important component of the CBRAIN architecture, and includes software and processing pipelines, data sources and data formats, and computational backends. Researchers can provide different software packages that make a vast number of processing tools available to authorized users. Standardized processing pipelines can be integrated either by writing dedicated CBRAIN plugins, or by leveraging the

open Boutiques⁸ framework (Glatard et al., 2015). Boutiques provides a high-level specification to describe command-line tools without writing any code, and to install these tools uniformly on computing systems through Docker⁹ containers. CBRAIN is designed to provide a generic data processing framework, accepting different data-types from various sources as determined by the data-processing software. This is achieved by the creation of data models that associate each data-type with its own processing software and corresponding visualization tool. Finally, CBRAIN provides a meta-scheduler and adaptors to common cluster systems (PBS, Torque, SGE, MOAB, LSF, Amazon EC2 or simple UNIX prompt submission) in order to extend the computational backends needed to process large amounts of data through these diverse processing pipelines.

Currently, CBRAIN deploys Docker containers on a 20,000-node computing cluster provided by Compute Canada, and on Amazon Elastic Compute Cloud (EC2) using its cloud support plugin. Several data analysis tools and processing pipelines are currently deployed in these clusters (CIVET, FreeSurfer, FSL, etc.), and data models for viewing and processing common file types (csv, txt) and various neuroimaging data formats are defined (MINC, NIFTI, BIDS). In the future, other types of containers, for instance Singularity¹⁰, can further facilitate sharing of new tools in an Open Science context. Other scheduling systems can be easily added using the modularity

⁸<http://boutiques.github.io>

⁹<https://docker.com>

¹⁰<http://singularity.lbl.gov>

of the resource access framework to further extend the computational backend.

RESULTS

The cyberinfrastructure for Open Science at the MNI consists of three primary components: the technical infrastructure that facilitates acquisition, storage, querying, processing, and data analysis; the workflows, procedures and best practices associated with data integrity and privacy at each step; and the data themselves.

Technical Infrastructure

Numerous large-scale projects have already employed LORIS for multi-site use (Evans and Brain development cooperative group, 2006; Wolff et al., 2012; Amunts et al., 2013; Paolozza et al., 2014; Foster et al., 2015; Orban et al., 2015), and several institutions have chosen or planned for LORIS as their institutional infrastructure (e.g., PERFORM Centre at Concordia University, University of Edinburgh's Brain Research Imaging Centre). LORIS is used across 150 acquisition sites in numerous countries with over 500 instruments, over 75,000 variables, and 40 TB of data.

The CBRAIN service deployed at the MNI¹¹ currently provides over 460 collaborators in 20 countries with web access to several systems, including six clusters of the Compute Canada¹² high-performance computing infrastructure (totalling more than 100,000 computing cores and 40PB of disk storage) and Amazon EC2. Presently, CBRAIN transiently stores about 10 million files representing over 50TB distributed over 42 servers. 56 data processing tools are integrated and over 340,000 processing batches have been submitted since 2010.

Workflows

One of the most important aspects in constructing large-scale data sharing initiatives is the incorporation of properly-designed user workflows, which are vital to ensuring effective usability and viability. Creating software that provides a seamless user experience for a subset of functionalities is a widely understood best practice; however, incorporating diversified workflows into a complex infrastructure, such as institutional Open Science, requires more than wizardry in programming or knowledge of the latest code libraries.

To that end, detailed workflows have been created to facilitate procedures involved in acquiring, storing, and analyzing neuroscience data including clinical, imaging, genetic, and biobanking information. These workflows, outlined in the Methods section of this paper (Figures 3–5), are designed to improve consistency within studies and are critical in an Open Science model across studies. Such procedures help ensure consistency and compliance with data collection standards (i.e., naming, data collection, and imaging pipelines), and coupled with proper and intuitive data organization, provide the foundation of data sharing, for easier interoperability between

software systems. Consistent application of such workflows also serves to reduce time spent manually identifying and addressing variability in data formatting. These systems are augmented by a comprehensive set of previously-discussed QC procedures ensuring validation of data and flagging of data for correction. As imaging, clinical, or biospecimen information proceeds from registration through analysis, these streamlined workflows save significant time and energy for researchers as well as developers, all while producing a robustly documented and well-validated dataset.

The Data

Various data types are stored in LORIS including phenotypic, clinical, demographic, imaging, and genomic data. The MNI's Open Science platform will initially consist of contributions of imaging and biobanking data from two key institutional resources. Within the MNI, biospecimens will be housed and tracked in the institutional biobank component of the C-BIG Repository. Neuroimaging data will also be contributed to the C-BIG Repository by researchers using the MNI's McConnell Brain Imaging Centre (BIC) Imagebank platform. The resulting unified repository (see Table 1) will serve the MNI with an enriched data platform, providing multi-modal data querying via the DQT, and enabling visualizations and analyses of more complex datasets (European Society of Radiology, 2015).

Imagebank Infrastructure

In its pilot phase, the MNI's Imagebank will serve as a central repository of scans primarily collected at the BIC's MRI unit. Scans transferred to the Imagebank server will be loaded through a series of software scripts into LORIS, and automatically made available for download through the Imagebank's web-based browser interface. This repository allows all images, whether raw or processed, to be available for visualization, quality control, and download/export. Currently, this database links to a compressed archive of every MRI dataset sent to the server, which will grow considerably as the infrastructure is further deployed and usage grows. Expansion for other imaging modalities across the MNI, such as PET and MEG (Niso et al., 2016), is underway. Imaging volumes stored in this LORIS-based repository can be pushed to CBRAIN for image processing and returned in an automated manner into the Imagebank.

TABLE 1 | C-BIG repository overview.

MNI C-BIG Centralized LORIS Repository		
Type	Description	Data
Imagebank	Multi-modal, raw/processed neuroimaging data	MRI, PET, MEG, EEG, Spectroscopy
Biobank	Biospecimen data	Blood, saliva, skin, muscle & nerve biopsies, whole brains, cerebrospinal fluid
Genetic	Summary genetic data	SNPs, CNVs, CpG, GWAS
Phenotypic	Behavioral, clinical data	Instruments, Assessments, Questionnaires

Data types and description of data that will be stored in the MNI's C-BIG Repository.

¹¹<https://portal.cbrain.mcgill.ca>

¹²<http://www.computeCanada.ca>

In addition to storing, processing, archiving, and retrieving data, investigators will have the option of releasing their scans to the Open MNI platform in accordance with institutional ethical and policy constraints as discussed in the Methods section.

Biobank Infrastructure

Biosamples or biospecimens collected from subjects at the MNI are stored within an infrastructure of freezers and labs. This physical infrastructure, together with the software modules within LORIS which retrieve and process data related to these biospecimens, are collectively referred to as “The Biobank.” Biosample types collected on-site include blood, saliva, skin, muscle, and nerve biopsies, whole brains, and cerebrospinal fluid. LORIS logs specimen information - including sample type, specimen quantity and availability, methodology employed, and so on - beginning at the stage of collection and initial storage and continuing through successive stages of analysis in the research data lifecycle. During these stages, samples may also be located offsite in any number of collaborating institutions or facilities, such as the Genome Quebec Innovation Centre. Results from the assays and analysis performed on these specimens are stored in LORIS.

Both qualitative and quantitative outputs - such as cell counts, protein expression, or diagnostic information—can be captured for each biospecimen. Precisely which input fields are used depends on the study and can be extended and customized on a per-project and/or per-methodology basis. All of these data are queryable in conjunction with clinical/behavioral data which are also stored in LORIS.

LORIS contains a wide range of data collected from physical biospecimens, including skin, blood and saliva. In addition to these common sample types, a key strength of the MNI biobank is enabling access to data obtained via complex, invasive or rare procedures, such as muscle, brain and nerve biopsies, cerebrospinal fluid, and whole brain specimens. Information and analyses collected by one researcher (including data acquisition log files, observations, models, outcomes, etc.) can be added to the biobank for review and reuse by others. In providing access to a large online dataset, LORIS greatly facilitates optimal use and data re-analysis of rare specimens. This has clear benefits for the acceleration of new discoveries in neuroscience.

DISCUSSION

Open Science, at an institutional level, is a concept that has not yet been widely adopted across the scientific community. In tandem with the deployment of a robust cyberinfrastructure, key enhancements to organizational practices are necessary for Open Science to truly proliferate. Beginning with obtaining subject consent for data sharing, protecting subject privacy and complying with ethical regulations, there are challenges in ensuring that all such considerations are executed properly, securely, and effectively.

For an institution to go completely open, it requires considerable buy-in from investigators who will share data and tools, and a comprehensive institutional policy contingent upon full support and leadership across the organization. Naturally there are some risks and challenges associated in the adoption of an Open Science framework. On an individual level, researchers may be concerned about the ownership of data they have generated, or autonomy over their research findings. However the realities of any such risks are far outweighed by increasing the outreach of the research and the number of citations (Piwowar and Vision, 2013) and recognition that is attributed to shared data, as initiatives such as ADNI (Petersen et al., 2010), the Human Connectome Project (Van Essen et al., 2012), ABIDE (Di Martino et al., 2014), FCP (Biswal et al., 2010), ADHD (ADHD-200 Consortium, 2012), OpenfMRI (Poldrack et al., 2013), and CoRR (Zuo et al., 2014) have demonstrated. From an institutional perspective, there is often a fear that foregoing potential patent royalties will result in lost revenue and recognition of innovation (David, 2004). However, open access initiatives can result in greater funding opportunities, increased efficiency, and greater institutional recognition (Poldrack and Gorgolewski, 2014).

The MNI's commitment to move toward an Open Science model of data sharing (Owens, 2016a,b) leverages the benefits of increased access to datasets in sample sizes and variability while advancing the data lifecycle toward enriching exploratory analyses and hypothesis formulation, which allows for new questions to be asked. Increased sample size and sample variation also improves reproducibility and reliability of inference testing as well as publication quality and impact. While simply releasing data under an Open Science context does not in itself address all the concerns regarding reproducibility (such as selective reporting and analysis, processing pipeline deviations, proper documentation, etc.), it does push toward principles of replicability by pressuring for improved descriptions and provenance, allowing for increased analysis and re-analysis, and facilitating collaborative quality control and validation (Zuo et al., 2014; Zuo and Xing, 2014).

It is important to note that by facilitating collaborations through data sharing, the cost of entry for many researchers will be lowered (Edwards et al., 2009; Abboud, 2016; Owens, 2016a,b), thus maximizing the return on public science funding and research investments (Poldrack and Gorgolewski, 2014). Emerging interoperability between specialized data systems, such as XNAT (imaging, Marcus et al., 2007), REDCap (clinical/behavioral, Harris et al., 2009) and LIMS systems, as well as LORIS, will also serve to lower technical barriers to the federation of datasets across modalities and repositories.

Another important consideration for Open Science at the MNI is its foundation on an established software infrastructure—i.e., the combination of LORIS and CBRAIN—that has been already operational for several years. Over the lifecycle of these applications, these platforms have been designed and developed in close collaboration with researchers and have grown according to their needs and goals. This infrastructure is used internationally, operating across the full life-cycle of data-sharing (i.e., acquisition to analysis), and is proven to be scalable for large-scale datasets. This wealth of experience is key to the

cyberinfrastructure of the Open Science initiative as it addresses many of the major hurdles that this endeavor could involve. However, as the first of its kind, the MNI's institutional Open Science initiative has necessitated the addition of the following features and functionalities.

In LORIS:

- (1) A complete de-identification mechanism has been developed that allows publication of data beyond the usual confines of a particular study, while at the same time ensuring ethics and privacy.
- (2) Support for several data modalities is being added, including PET, EEG/MEG, and biosamples. This is of particular importance since the range of modalities used at an institutional level is much wider than in a single project.
- (3) Quality control tools have been extended and made more robust, based on 15 years of experience in a number of data acquisition project lifecycles.

In CBRAIN:

- (1) Tighter integration with the LORIS database to allow for compute-intensive processing of Open Data.
- (2) Streamlined account creation process and handling of access permissions, so that various user profiles can be easily handled by administrators. This will be particularly important when the MNI's Open Science initiative reaches its full potential, as users with a wide range of profiles are expected to access the data and to have various processing requirements.
- (3) Facilitated tool integration, so that external researchers could contribute their tool to the CBRAIN ecosystem without expert knowledge of its internal mechanisms.

CONCLUSION

Open Science is a simple concept that masks a daunting set of ethical, conceptual, and technical challenges. As the scale of scientific data collection and scope of discovery increase with technological advancement, the promise of collaboration through Open Science presents a potential solution to limits faced by institution-based science, including statistical power and resource constraints. This Open Science cyberinfrastructure at the MNI, comprised of the LORIS and CBRAIN platforms, intends to increase transparency in data curation, dissemination and analysis, reduce data loss, facilitate innovation and collaboration, and efficiently accelerate the discovery and the application of neuroscience

at the Montreal Neurological Institute and across the greater research community.

AUTHOR CONTRIBUTIONS

SD, TG, MR, AE—Contributed to the writing of this paper, contributed to the infrastructure, contributed to conceptualization of the initiative, contributed to policy. JK, AB, RH, EF, GR—Contributed to the writing of this paper, contributed to conceptualization of the initiative, contributed to policy. CR, JSA, SP, DM, JST, PR, SM, PK—Contributed to the writing of this paper, contributed to the infrastructure, contributed to conceptualization of the initiative. LM, MS, VF, IL, TC—Contributed to the writing of this paper, contributed to the infrastructure. CM, ZM, XL, DC—Contributed to the infrastructure, contributed to conceptualization of the Initiative. AD, DC, SB—Contributed to the writing of this paper, contributed to conceptualization of the initiative. CD, SV—contributed to conceptualization of the initiative, contributed to policy. RA, NM, TD, JM—contributed to conceptualization of the initiative.

ACKNOWLEDGMENTS

The authors thank the following people for their sustained efforts toward building the MNI Open Science cyberinfrastructure: Justin Kat, Nicolas Brossard, Ted Strauss, Stella Lee, Gregory Luneau, Rida Abou-Haidar, Wang Shen, Tarek Sherif, Nicolas Kassis, Claude Lepage, Carolina Makowski, Natacha Beck, Robert Vincent, Derek Lo, Lindsay Lewis, Guiomar Niso, Pierre-Emmanuel Morin, Alden Woodward, Pamela Patterson, Christopher Steel, Elizabeth Bock, Jean-Francois Malouin, Deepak Sharma, Rishabh Tandon, Hohai Phuok Truong. This work has been made possible with the support of Canadian Institutes of Health Research (CIHR), Canadian Foundation for Innovation (CFI), The National Sciences and Engineering Council of Canada (NSERC Research Technology & Instruments Grant 69910 to the McConnell Brain Imaging Centre, and Discovery Grant 436355-13), The Fonds du Recherche du Quebec - Sante, Brain Canada (Platform Support Grant to the McConnell Brain Imaging Centre), National Institutes of Health (2R01EB009048-05 to the MEG Unit, McConnell Brain Imaging Centre), CANARIE, Compute Canada (Research Portals and Platforms support to the McConnell Brain Imaging Centre), the Irving Ludmer Family Foundation and the Ludmer Centre for Neuroinformatics and Mental Health. The Montreal Neurological Institute's Open Science initiative has been made possible by the support of the Larry and Judy Tanenbaum family.

REFERENCES

- Abboud, A. (2016). "Principle vs. Practice in Open Science Data-Sharing Consortia," *4th Annual Conference on Governance of Emerging Technologies: Law, Policy, and Ethics* (2016). Available online at: <http://conferences.asucollegeoflaw.com> (Accessed on August 30, 2016).
- ADHD-200 Consortium (2012). The ADHD-200 Consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* 6:62. doi: 10.3389/fnsys.2012.00062
- Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M. É., et al. (2013). BigBrain: an ultrahigh-resolution 3D human brain model. *Science* 340, 1472–1475. doi: 10.1126/science.1235381
- Asante, K., Barbour, E., Barker, L., Benjamin, M., Bowman, S., Boughton, A., et al. (2016). *Open Science Framework*. Available online at: <http://osf.io/4znzp> (Accessed on November 17th, 2016).

- Baillet, S., Friston, K., and Oostenveld, R. (2011). Academic software applications for electromagnetic brain mapping using MEG and EEG. *Comput. Intell. Neurosci.* 2011:972050. doi: 10.1155/2011/972050
- Bischoff-Grethe, A., Ozyurt, I. B., Busa, E., Quinn, B. T., Fennema-Notestine, C., Clark, C. P., et al. (2007). A technique for the deidentification of structural brain MR images. *Hum. Brain Mapp.* 28, 892–903. doi: 10.1002/hbm.20312
- Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107
- Campbell, P. (2016). Reality check on reproducibility. *Nature* 533, 437. doi: 10.1038/533437a
- Canadian Institutes of Health Research Natural Sciences and Engineering Research Council of Canada and Social Sciences and Humanities Research Council of Canada (2014). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*.
- Craddock, C. R., Margulies, D. S., Bellec, P., Nolan Nichols, B., Alcauter, S., Barrios, F. A., et al. (2016). Brainhack: a collaborative workshop for the open neuroscience community. *GigaScience* 5:16. doi: 10.1186/s13742-016-0121-x
- Das, S., Glatard, T., MacIntyre, L. C., Madjar, C., Rogers, C., Rousseau, M. E., et al. (2016). The MNI data-sharing and processing ecosystem. *NeuroImage* 124, 1188–1195. doi: 10.1016/j.neuroimage.2015.08.076
- Das, S., Zijdenbos, A. P., Harlap, J., Vins, D., and Evans, A. C. (2011). LORIS: A web-based data management system for multi-center studies. *Front. Neuroinform.* 5:37. doi: 10.3389/fninf.2011.00037
- David, P. A. (2004). Can “Open Science” be protected from the evolving regime of IPR protections? *J. Institutional Theor. Econ.* 160. Available online at: <http://www.jstor.org/stable/40752435> (Accessed on December 17, 2016).
- Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alarats, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78
- Ducharme, S., Albaugh, M. D., Nguyen, T. V., Hudziak, J. J., Mateos-Pérez, J. M., Labbe, A., et al. (2015). Trajectories of cortical thickness maturation in normal brain development—The importance of quality control procedures. *NeuroImage* 125, 267–279. doi: 10.1016/j.neuroimage.2015.10.010
- Edwards, A. M., Bountra, C., Kerr, D. J., and Willson, T. M. (2009). Open access chemical and clinical probes to support drug discovery. *Nat. Chem. Biol.* 5, 436–440. doi: 10.1038/nchembio0709-436
- European Society of Radiology (ESR) (2015). ESR position paper on imaging biobanks. *Insights Imaging* 6, 403–10. doi: 10.1007/s13244-015-0409-x
- Evans, A. C., and Brain development cooperative group (2006). The NIH MRI study of normal brain development. *NeuroImage* 30, 184–202. doi: 10.1016/j.neuroimage.2005.09.068
- Foster, N. E., Doyle-Thomas, K. A., Tryfon, A., Ouimet, T., Anagnostou, E., Evans, A. C., et al. (2015). Structural gray matter differences during childhood development in autism spectrum disorder: a multimetric approach. *Pediatr. Neurol.* 53, 350–359. doi: 10.1016/j.pediatrneurol.2015.06.013
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., Frackowiak, R. S. J., et al. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402
- Glatard, T., Da Silva, R. F., Boujelben, N., Adalat, R., Beck, N., Rioux, P., et al. (2015). Boutiques: an application-sharing system based on Linux containers. *Front. Neurosci. Conf. Abstr. Neuroinform.* 46, 17–35. doi: 10.3389/conf.fnins.2015.91.00012
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Front. Neuroinform.* 5:13. doi: 10.3389/fninf.2011.00013
- Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure: a standard for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44
- Hall, D., Huerta, M. F., McAuliffe, M. J., and Farber, G. K. (2012). Sharing heterogeneous data: the national database for autism research. *Neuroinformatics* 10, 331–339. doi: 10.1007/s12021-012-9151-4
- Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* 42, 377–381. doi: 10.1016/j.jbi.2008.08.010
- Honor, L. B., Haselgrove, C., Frazier, J. A., and Kennedy, D. N. (2016). Data citation in neuroimaging: proposed best practices for data identification and attribution. *Front. Neuroinform.* 10:34. doi: 10.3389/fninf.2016.00034
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *NeuroImage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Johnson, S. B., Whitney, G., McAuliffe, M., Wang, H., McCreedy, E., Rozenblit, L., et al. (2010). Using global unique identifiers to link autism collections. *J. Am. Med. Inform. Assoc.* 17, 689–695. doi: 10.1136/jamia.2009.002063
- Katz, D., Lehnardt, J., Slater, N., Christopher Lenz, J., Anderson, C., Davis, P., et al. (2005). *CouchDB*. Available online at: <http://couchdb.apache.org> (Accessed on August 31, 2016).
- Koch, C., and Jones, A. (2016). Big science, team science, and open science for neuroscience. *Neuron* 92, 612–616. doi: 10.1016/j.neuron.2016.10.019
- Lucey, B. P., Nelson-Rees, W. A., and Hutchins, G. M. (2009). Henrietta Lacks, HeLa cells, and cell culture contamination. *Arch. Pathol. Lab. Med.* 133, 1463–1467. doi: 10.1043/1543-2165-133.9.1463
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007). The extensible neuroimaging archive toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics* 5, 11–34. doi: 10.1385/NI:5:1:11
- Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). NIDM-Results: a Neuroimaging Data Model to share brain mapping statistical results. *bioRxiv* 2016:041798. doi: 10.1038/sdata.2016.102
- Merino-Martinez, R., Norlin, L., van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., et al. (2016). Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core). *Biopreserv. Biobank.* 14, 298–306. doi: 10.1089/bio.2015.0070
- National Institute of Mental Health (2015). *Data Sharing Expectations for Clinical Research Funded by NIMH*. Available Online at: <https://grants.nih.gov/grants/guide/notice-files/NOT-MH-15-012.html> (Accessed on August 24, 2016).
- National Institutes of Health (NIH) (2014). *Final NIH Genomic Data Sharing Policy. Federal Register (79 FR 51345)*. Available Online at: <https://www.federalregister.gov> (Accessed on August 24, 2016).
- Nelson, G. S. (2015). “Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification,” in *SAS Global Forum Proceedings 2015*. Available Online at: <http://support.sas.com> (Accessed on August 28, 2016).
- Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2016). Best practices in data analysis and sharing in neuroimaging using MRI. *bioRxiv* doi: 10.1101/054262
- Niso, G., Rogers, C., Moreau, J. T., Chen, L. Y., Madjar, C., Das, S., et al. (2016). OMEGA: the open MEG archive. *NeuroImage* 124(Pt B), 1182–1187. doi: 10.1016/j.neuroimage.2015.04.028
- Orban, P., Madjar, C., Savard, M., Dansereau, C., Tam, A., Das, S., et al. (2015). Test-retest resting-state fMRI in healthy elderly persons with a family history of Alzheimer’s disease. *Sci. Data* 2:150043. doi: 10.1038/sdata.2015.43
- Owens, B. (2016a). Data sharing: access all areas. *Nature* 533, S71–S72. doi: 10.1038/533S71a
- Owens, B. (2016b). Montreal institute going ‘open’ to accelerate science. *Sci. News* doi: 10.1126/science.aae0265
- Paolozza, A., Treit, S., Beaulieu, C., and Reynolds, J. N. (2014). Response inhibition deficits in children with Fetal Alcohol Spectrum Disorder: relationship between diffusion tensor imaging of the corpus callosum and eye movement control. *NeuroImage Clin.* 5, 53–61. doi: 10.1016/j.nicl.2014.05.019
- Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., et al. (2010). Alzheimer’s Disease Neuroimaging Initiative (ADNI). *Neurology* 74, 201–209. doi: 10.1212/WNL.0b013e3181cb3e25
- Piowar, H. A., and Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ* 1:e175. doi: 10.7717/peerj.175
- Poldrack, R. A., Barch, D. M., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012
- Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi: 10.1038/nn.3818

- Sherif, T., Kassis, N., Rousseau, M. É., Adalat, R., Evans, A. C., et al. (2014). CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* 8:54. doi: 10.3389/fninf.2014.00054
- Sherif, T., Kassis, N., Rousseau, M.-É., Adalat, R., and Evans, A. C. (2015). BrainBrowser: distributed, web-based neurological data visualization. *Front. Neuroinform.* 8:89. doi: 10.3389/fninf.2014.00089
- Tri-Agency Statement of Principles of Digital Data Management (2016). *Science.gc.ca, the Government of Canada's official Science Portal*. Available online at: <http://www.science.gc.ca> (Accessed on August 29, 2016).
- The Wellcome Trust (2016). *Policy on Data Management and Sharing*. Available online at: <https://wellcome.ac.uk> (Accessed on December 31, 2016).
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., et al. (2012). The human connectome project: a data acquisition perspective. *Neuroimage* 62, 2222–2231. doi: 10.1016/j.neuroimage.2012.02.018
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18
- Wolff, J. J., Gu, H., Gerig, G., Elison, J. T., Styner, M., Gouttar, S., et al. (2012). Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *Am. J. Psychiatry* 169, 589–600. doi: 10.1176/appi.ajp.2011.11091447
- Zuo, X. N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* 1:140049. doi: 10.1038/sdata.2014.49
- Zuo, X. N., and Xing, X. X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neurosci. Biobehav. Rev.* 45, 100–118. doi: 10.1016/j.neubiorev.2014.05.009

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer BG and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2017 Das, Glatard, Rogers, Saigle, Paiva, MacIntyre, Safi-Harab, Rousseau, Stirling, Khalili-Mahani, MacFarlane, Kostopoulos, Rioux, Madjar, Lecours-Boucher, Vanamala, Adalat, Mohaddes, Fonov, Milot, Leppert, Degroot, Durcan, Campbell, Moreau, Dagher, Collins, Karamchandani, Bar-Or, Fon, Hoge, Baillet, Rouleau and Evans. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Fastr: A Workflow Engine for Advanced Data Flows in Medical Image Analysis

Hakim C. Achterberg^{1*}, Marcel Koek¹ and Wiro J. Niessen^{1,2}

¹Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC, Rotterdam, Netherlands, ²Imaging Science & Technology, Faculty of Applied Sciences, Delft University of Technology, Delft, Netherlands

With the increasing number of datasets encountered in imaging studies, the increasing complexity of processing workflows, and a growing awareness for data stewardship, there is a need for managed, automated workflows. In this paper, we introduce Fastr, an automated workflow engine with support for advanced data flows. Fastr has built-in data provenance for recording processing trails and ensuring reproducible results. The extensible plugin-based design allows the system to interface with virtually any image archive and processing infrastructure. This workflow engine is designed to consolidate quantitative imaging biomarker pipelines in order to enable easy application to new data.

Keywords: workflow, pipeline, data processing, provenance, reproducible research, distributed computing, data flow, Python

OPEN ACCESS

Edited by:

Florence Forbes,
INRIA, France

Reviewed by:

Suyash P. Awate,
Indian Institute of Technology
Bombay, India
Avan Suinesiaputra,
University of Auckland, New Zealand

*Correspondence:

Hakim C. Achterberg
h.achterberg@erasmusmc.nl

Specialty section:

This article was submitted to
Computer Image Analysis,
a section of the journal
Frontiers in ICT

Received: 15 April 2016

Accepted: 03 August 2016

Published: 24 August 2016

Citation:

Achterberg HC, Koek M and
Niessen WJ (2016) Fastr:
A Workflow Engine for
Advanced Data Flows in
Medical Image Analysis.
Front. ICT 3:15.
doi: 10.3389/fict.2016.00015

1. INTRODUCTION

In medical image analysis, most methods are no longer implemented as a single executable, but as a workflow composed of multiple programs that are run in a specific order. Each program is executed with inputs that are predetermined or resulting from the previous steps. With increasing complexity of the methods, the workflows become more convoluted and encompass more steps. This makes execution of such a method by hand tedious and error-prone, and makes reproducing the exact chain of processing steps in subsequent studies challenging. Therefore, solutions have been created that are based on scripts that perform all the steps in the correct order.

In population imaging, data collections are typically very large and are often acquired over prolonged periods of time. As data collection is going on continuously, the concept of a “final” dataset is either non-existent or defined after a very long follow up time. Commonly, analyses on population imaging datasets, therefore, define intermediate cohorts or time points. To be able to compare intermediate cohorts, all image analysis methods need to produce consistent results over time and should be able to cope with the ever growing size of the population imaging. Therefore, the process of running analysis pipelines on population imaging data needs to be automated to ensure consistency and minimize errors.

When different population imaging cohorts are combined in multi-center imaging studies or imaging biobanks (e.g., ADNI (Mueller et al., 2005), OASIS (Marcus et al., 2007b), The Heart-Brain Connection (van Buchem et al., 2014) and BBMRI-NL2.0¹) where data are often acquired from different scanners, the challenge of ensuring consistency and reliability of the processing results also calls for automated processing workflows.

Traditionally, this is accomplished by writing scripts created specifically for one processing workflow. This can work well, but generally the solutions are tailor-made for a specific study and software environment. This makes it difficult to apply such a method to different data or on a

¹<http://www.bbmr.nl>

different infrastructure than originally intended. With evolving computational resources, in practice this approach is, therefore, not reproducible and difficult to maintain. Additionally, for transparency and reproducibility of the results, it is very important to know exactly how the data were processed. To accomplish this, a comprehensive data provenance system is required.

Writing a script that takes care of all the aforementioned issues is a challenging and time consuming task. However, many of the components are generic for any type of workflow and do not have to be created separately for each workflow. Workflow management systems can be used to address these issues. These systems help formalize the workflow and can provide features, such as provenance as part of the framework, removing the need to address these for every separate workflow.

For our use cases, we desire a workflow management system that works with the tools found in the domain of image analysis, can handle advanced data flows (explained more in detail in section 2.3), has strong provenance handling, can handle multiple version of tools, flexible execution backend, and can be embedded in our infrastructure. There are already a number of workflow systems available, but none of them fit all our criteria (see Table 1).

The most notable open-source, domain-specific workflow system that we are aware of is Nipype (Gorgolewski et al., 2011), which is aimed at creating a common interface for a variety of neuroimaging tools. It also features a system for creating workflows. The tool interfaces of Nipype are elaborate, but Nipype only tracks the version of tools, but does not manage it. This means the system is only aware of the currently installed version of the tool, and cannot offer multiple versions simultaneously.

LONI pipeline (Rex et al., 2003; Dinov et al., 2010) and CBrain (Sherif et al., 2015) also have been developed for the domain of medical image analysis. They include workflow engines, but these systems are part of larger environments that includes data management and processing backends. This makes it difficult to integrate in our infrastructure. Furthermore, LONI is closed-source, which makes it even more difficult to integrate it.

The XNAT storage system also has a related workflow system called XNAT pipeline engine (Marcus et al., 2007a). The pipeline engine is integrated nicely with the XNAT storage system and works with simple data flows. However, it does not handle advanced data flows and does not provide tool versioning.

Besides the workflow systems specific for the domain of medical image analysis, there are a number of other notable workflow systems that are either domain-independent or have been created for a different domain. Taverna (Oinn et al., 2006) and KNIME (Berthold et al., 2008, 2009) are well-known and mature workflow management systems. These systems are domain-independent, but mostly used in the bioinformatics field. Their support for local binary targets is limited and, therefore, not suitable for using most medical imaging analysis tools. KNIME needs tools to be created with their API and Taverna is mostly focused on web services.

Finally, Galaxy (Goecks et al., 2010) is a web-based workflow system for bioinformatics. It is mainly focused on next-generation sequencing (NGS). It has a large repository of tools, web interface, and large support in their domain. However, the system is not designed for batch processing and it does not support complex data-flows.

We developed an image processing workflow framework for creating and managing processing pipelines: Fastr. The framework is designed to build workflows that are agnostic to where the input data are stored, where the resulting output data should be stored, where the steps in the workflow will be executed, and what information about the data and processing needs to be logged for data provenance. To allow for flexible data handling, the input and output of data are managed by a plugin-based system. The execution of the workflow is managed by a pluggable system as well. The provenance system is a built-in feature that ensures a complete log of all processing steps that led to the final result.

In the following section, we discuss the design of Fastr. In Section 3, we present the resulting software. Finally, we discuss related work and future directions in section 4.

2. DESIGN

The Fastr workflow design follows similar principles as flow-based programming (Morrison, 2010). This paradigm defines applications as a network of black boxes, with predefined connections between the black boxes that indicate the data flow. The black boxes can be reordered and reconnected to create different workflows. However, it should be noted that other aspects of the paradigm are not met, so our design can at most be considered to have flow-based programming aspects.

TABLE 1 | A overview of workflow systems and the important features of each.

Workflow software						
Name	Open-source	Language	Data flow	Tools	Tool versioning	Citation
CBrain	Yes	Ruby	Simple	Binaries	Yes	Sherif et al. (2015)
Fastr	Yes	Python	Advanced	Binaries	Yes	
Galaxy	Yes	Python	Simple	Binaries	Yes	Goecks et al. (2010)
KNIME	Yes	Java	Advanced	Wrappers for Java, Python, Perl code	No	Berthold et al. (2008, 2009)
LONI pipeline	No	Java	Advanced	Binaries	Yes	Rex et al. (2003), Dinov et al. (2010)
Nipype	Yes	Python	Advanced	Binaries	No	Gorgolewski et al. (2011)
Taverna	Yes	Java	Advanced	Webservices	No	Oinn et al. (2006)
XNAT pipeline engine	Yes	Java	Simple	Binaries	No	Marcus et al. (2007a)

The column Data Flow can have the value simple or advanced. Simple means the workflow system supports only sequential data flows whereas advanced indicates support for more complex data flows (e.g., the data flows in Section 2.3).

In Fastr, the workflow is described as a *Network*, which is a directional acyclic graph. The *Nodes* of this *Network* are based on templates that we call *Tools*. These *Nodes* can be interpreted as the black boxes from the flow-based programming paradigm. In the next subsection, we will discuss the *Tools* in more detail. After that, we will describe the *Network* and its components in more detail using an example from medical image analysis.

2.1. Tools

In Fastr, the *Tools* are the *blueprints* for the *Nodes*; they describe the input, output, and behavior of the *Node*. The *Tools* are composed of three main parts: general metadata, a target, and an interface. The *Tools* are stored as XML or JSON files. An example of a simple *Tool* that adds two list of integers element-wise is given in Listing 1. The general metadata contains information about the *Tool*, such as id, version, author, and license. The target describes how to set the execution environment properly, e.g., by setting the correct search path to use a specific version of the software. The interface describes the inputs and outputs of a *Tool* and how the *Tool* executes given a set of inputs and outputs.

The tools are specified in a schema. This schema validates the internal python data structures (after conversion from XML or JSON) and is specified as a JSON schema. The schemas are located in the source code. There is a schema for the general *Tool*² and a schema for the *FastrInterface*.³ Other types of *Interfaces* can also be defined by their own data schema files.

Listing 1. The XML code that defines the *AddInt Tool*. Note that though it might seem the two author entries are redundant or conflicting, the first one states the author of the *Tool* description file, whereas the second states the author of the underlying command (*addint.py* in this case).

```
<tool id="AddInt" name="Add two integers"
  ↪ version="1.0">
<description>Add two integers together.
</description>
<authors>
  <author name="Hakim Achterberg"
    ↪ email="h.achterberg@erasmusmc.nl"
    ↪ url="http://www.bigr.nl/people/HakimAchterberg"/>
</authors>
<command version="0.1" url="">
  <targets>
    <target os="*" arch="*"
      ↪ interpreter="python" paths="."/>
      ↪ bin="addint.py"/>
  </targets>
```

```
<description>
  addint.py value1 value2
  output=value1+value2
</description>
<authors>
  <author name="Marcel Koek"
    ↪ email="m.koek@erasmusmc.nl"
    ↪ url="http://www.bigr.nl/people/MarcelKoek"/>
</authors>
</command>
<repository/>
<interface>
  <inputs>
    <input id="left_hand" name="left hand"
      ↪ value" datatype="Int" prefix="--
      ↪ in1" cardinality="1-*" repeat_
      ↪ prefix="false" required="true"/>
    <input id="right_hand" name="right
      ↪ hand value" datatype="Int"
      ↪ prefix="--in2"
      ↪ cardinality="as:left_hand"
      ↪ repeat_prefix="false"
      ↪ required="true"/>
  </inputs>
  <outputs>
    <output id="result"
      ↪ name="Resulting value"
      ↪ datatype="Int" automatic="True"
      ↪ cardinality="as:left_
      ↪ hand" method="json"
      ↪ location="^RESULT=(.*)$">
    <description>The summation of
      ↪ left_hand and right_hand.
    </description>
  </output>
</outputs>
</interface>
</tool>
```

The content of the interface tag depends on the class of *Interface* used. The default *Interface* class in Fastr creates a call to a command-line program given the set of *Inputs* and *Outputs*. In the example, there are two inputs and one output. In Fastr, the minimal information required for an *Interfaces* to function is the id, cardinality and data type for each *Input* and *Output*. The cardinality is the number of values a sample contains (e.g., an argument requiring a point in 3D space, represented by three float values, would have a cardinality of 3).

In Fastr, there is a notion of datatypes: each input and output has a (set of) data types it accepts or produces. The datatypes in Fastr are plugins that, in the simplest form, only need to expose their id, but can be extended to include functionality, such as validators and handlers for multi-file data formats. Data types can be simple values or point to files.

Fastr checks if the datatypes of a linked input and output are (or at least can be) compatible. In addition, data types can be

²https://bitbucket.org/bigr_erasmusmc/fastr/src/default/fastr/resources/schemas/Tool.schema.json

³https://bitbucket.org/bigr_erasmusmc/fastr/src/default/fastr/resources/schemas/FastrInterface.schema.json

grouped, which is useful for groups of programs using a common (io) library (for example, programs created with The Insight Segmentation and Registration Toolkit⁴ (Yoo et al., 2002) can read/write a number of images formats that we grouped together in a pseudo-datatype).

2.2. Networks

After Tools are defined, a workflow can be created by linking a set of Tools that results in a Network. Once a Network is defined, it can be executed. **Figure 1** shows a graphic representation of an atlas-based segmentation workflow, using the image registration software Elastix (Klein et al., 2010). Elastix can register two images by optimizing the transformation applied to a moving image to match it to a fixed reference image.

There are different classes of Nodes: normal Nodes (gray blocks in **Figure 1**), Source Nodes (green), Constant Nodes (purple), and Sink Nodes (blue). Data enter the Network through a Source Node and leave the Network through a Sink Node. A Constant Node is similar to the Source Nodes, but has its data defined as part of the Network. When a Network is executed, the data for the Source Nodes and Sink Nodes has to be supplied. The specifics of the Source Nodes and Sink Nodes will be discussed in section 2.4. The normal Nodes process the data as specified by the Tool.

The data flow in the Network is defined by links (the arrows in **Figure 1**). A link is a connection between the output of a Node and the input of another Node. A link can manipulate the flow of the data, which will be discussed in section 2.3.

The Nodes and links in the Network form a graph from which the dependencies can be determined for the execution

order. Since all Nodes are black-boxes that can operate independently of each other, this allows for Nodes to be executed in parallel as long as the input dependencies are met.

2.3. Data Flow

In Fastr, a sample is defined as the unit of data that are presented to an input of a Node for a single job. It can be a simple scalar value, a string, a file, or a list of the aforementioned types. For example, in the *addint* Tool presented in Listing 1, the *left_hand* and *right_hand* inputs of the Tool are required to be (lists of) integers. The *result* output will generate a sample that contains a list of integers. As the cardinality of *right_hand* and *result* are defined to be the same as the *left_hand*, they will all have to same length.

Fastr can handle multiple samples on a specific input. **Figure 2** shows examples of how Fastr handles inputs with multiple samples and in which output samples this results. The inputs and output names are abbreviated as lh for *left_hand*, rh for *right_hand* and res for *result*. In **Figure 2A**, we present the simplest situation, in which one sample with one value is offered to each input and one sample with one value is generated. In **Figure 2B**, the *left_hand* and *right_hand* inputs have one sample with two values. The result is a sample with two values, as one result value is created per input value.

To facilitate batch processing, a Node can be presented with a collection of samples. These collections are multi-dimensional arrays of samples. In **Figure 2C**, we depict a situation where three additions are performed. Three samples are offered to the *left_hand* input and one sample is offered to the *right_hand* input. This results in three samples: each sample of the *left_hand* input was used in turn, whereas the samples for the *right_hand* were considered constant. In **Figure 2D**, there are three samples for the *left_hand* and *right_hand* inputs. The result is again three samples,

⁴www.itk.org

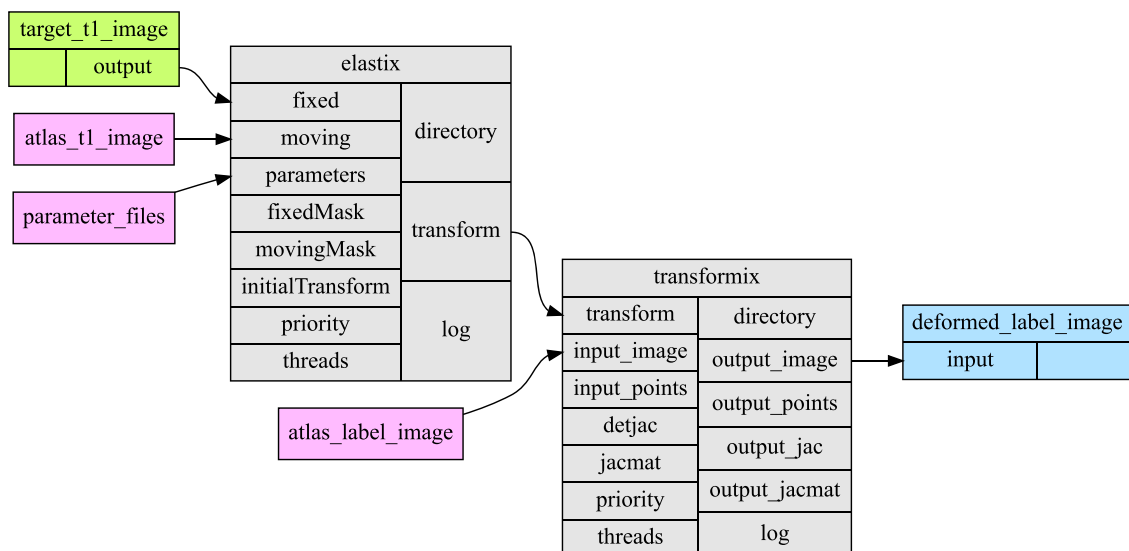


FIGURE 1 | Example Network representing a single atlas-based segmentation workflow implemented using the open source Elastix image registration software. Green boxes are Source Nodes, purple Constant Nodes, gray normal Nodes, and blue Sink Nodes. Each Node contains two columns: the left column represents the inputs, the right column represents the outputs of the Node. The arrows indicate links between the inputs and outputs. This image was generated automatically from the source code.

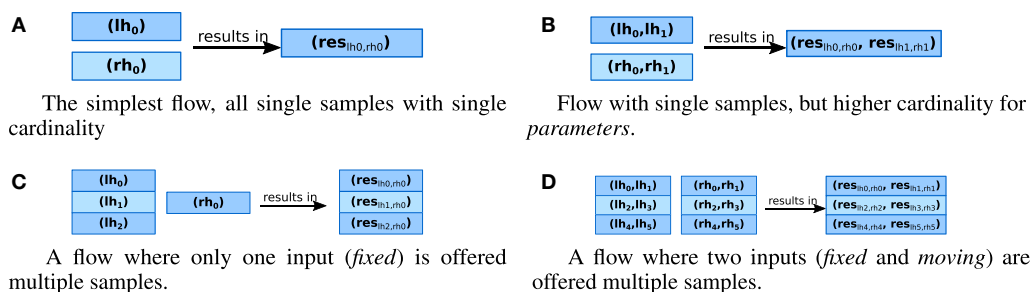


FIGURE 2 | Illustration of the data flows in a Node. Each rectangle is a sample, and a block of rectangles represents a sample collection. The value is printed in each rectangle, where the commas separate multiple values. The samples lh are offered to the *left_hand* input, the sample rh to *right_hand* input. The sample res is generated for the *result* output. The subscript of sample res indicates which input samples were used to generate the result.

as now each pair of samples from *left_hand* and *right_hand* inputs was taken.

This is useful for simple batch processing where a task should be repeated a number of times for different input values. However, in certain situation (e.g., multi-atlas segmentation), it is required to register every fixed image to every moving image. To simplify this procedure, Fastr can switch from pairwise behavior to cross product behavior. In **Figure 3**, this is depicted graphically. Every combination of *left_hand* and *right_hand* sample is used for registration and the result is a two-dimensional array of transformation samples that in turn contain two transformations each.

Sometimes a Tool outputs a sample with a higher cardinality that should be treated as separate samples for further processing, or conversely a number of samples should be offered as a single sample to an input (e.g., for taking an average). For this, Fastr offers two flow directives in data links. The first directive is *expand*, which indicates that the cardinality is to be transformed into a new dimension. This is illustrated in the left side of **Figure 4**. The second directive is *collapse*, which indicates one or more dimensions in the sample array should be collapsed and combined into the cardinality. This process is illustrated in the right side of **Figure 4**. These flow directives allow for more complex dataflows in a simple fashion and enable users to implement MapReduce type of workflows.

2.4. Data Input and Output

The starting points of every workflow are *Source Nodes*, in which the data are imported into the *Networks*. Similarly, the endpoints of every workflow are the *Sink Nodes*, which export the data to the desired location. When a *Network* is constructed only the data type for the *Source Nodes* and *Sink Nodes* needs to be defined. The actual definition of the data is done at runtime using uniform resource identifiers (URI).

Based on the URI scheme, the retrieval and storage of the data will be performed by a plugin. Consider the following two example URIs:

```
vfs://mount/some/path/file1.txt
xn timer://xn timer.example.com/data/archive/projects/sandbox/subj...
```

The schemes (in red) of these URI indicate by which plugin the retrieval or storage of the data is handled. For the first

URI, *vfs* indicates that the URI will be handled by the Virtual File System plugin. For the second URI, *xn timer* indicates that the URI will be handled by the XNAT storage plugin. These plugins implement the methods to actually retrieve and store the data. The remainder of the URI is handled by the plugin, so the format of the schemes URI format is defined by the plugin developer.

Plugins can also implement a method to expand a single URI into multiple URIs based on wildcards or searches. In the following example, URIs we use wildcards (shown in blue) to retrieve multiple datasets in one go:

```
xn timer://xn timer.example.com/search?projects=test
&subjects=s[0-9]...
vfsregex://tmp/network_dir/.*/.*/__fastr_result__.pickle.gz
```

The XNAT storage plugin has a direct storage as well as search URI scheme defined. The VFS regular expression plugin uses the *regex* filter to generate a list of matching *vfs* URIs. This illustrates that a plugin can expand a url into urls of a different type, and the newly generated urls will be handled by the appropriate plugin.

The use of URIs makes the *Network* agnostic to the location and storage method of the source and target data. Also, it allows easy loading of large amounts of resources using wildcards, csv files or search queries.

Currently, Fastr includes plugins for input/output from the (virtual) file system, csv files and XNAT. New plugins can be created easily as there are only a few methods that need overwriting. It is also possible to make plugins that can only read data, only write data, or only perform search queries. This allows users to create plugins purely for reading or writing.

Fastr does not include a credential store or other solution for authentication. For all *Network* based input/output plugins (e.g., the XNAT plugin) a netrc file stored in the user's home directory is used for authentication. However, for running Fastr on a grid without a shared network drive this might lead to problems.

2.5. Execution

The Fastr framework is designed to offer flexible execution of jobs. The framework analyzes the workflow and creates a list of

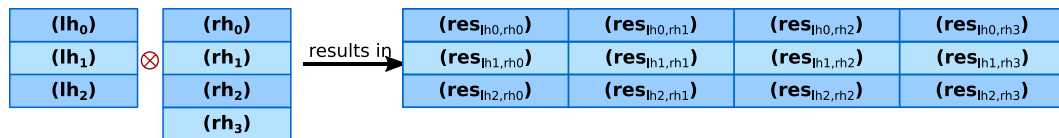


FIGURE 3 | Illustration of the data flows in a node that has multiple input groups. The default operator creates a new sample for each combination of input groups.



FIGURE 4 | Collapsing and expanding flows. The start situation on the left expands to the situation in the middle after which data collapses the first dimension. Note that in the middle situation there is an empty place in the sample collection (top right). This is possible due to a sparse array representation of the sample collections. This results in two samples with different cardinality in the right-most situation.

jobs, including dependencies, that need to be executed. Then it dispatches the jobs to an execution plugin. The plugins can run jobs locally or dispatch them to an execution system, such as a cluster, grid, or cloud. A different plugin can be selected for each run allowing for easy switching of the execution backend.

The Fastr execution system consists of a number of components that work together in a layered fashion (see **Figure 5**). The execution starts when the *Network* execute method is called. We will call the machine on which the *Network* execution is started the *Submit Host*.

Fastr analyzes the *Network* and divides it in chunks that can be processed further. For each chunk, the *Network* determines in what order the *Nodes* have to be processed and then executes the *Nodes* in the correct order. When a *Node* is executed, it analyzes the samples on each input and creates a job for each combination input (as specified by the data flow directives).

Jobs contain all information needed to run a single task (e.g., input/output arguments, *Tool* used, etc). The jobs are then dispatched by an execution plugin. The plugin can run the job remotely (e.g., on a compute cluster or cloud) or locally (in which case the *Submit Host* and *Execution Host* are the same).

Jobs are executed on the *Execution Host*, and during this step the arguments are translated from urls to actual paths/values. Subsequently, the *Tool* sets the environment for execution according to the target specification and invokes the interface. The interface executes the actual *Tool* commands. Once the interface returns its results, they are validated and the paths in the results are translated back into urls.

Once the job execution is finished, the execution plugin will trigger a callback on the *Submit Host* that reads the job result and updates the *Network* accordingly. If a chunk is finished, the *Network* will process the next chunk, using the updated information. If all chunks are finished, the *Network* execution is done.

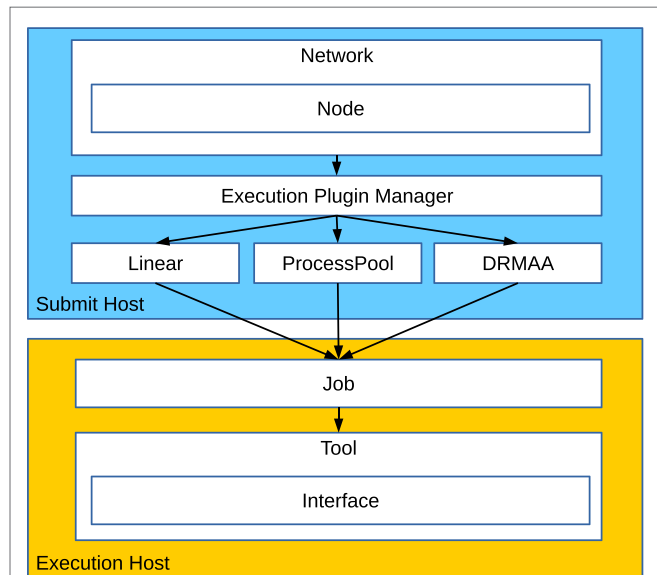


FIGURE 5 | An overview of the execution components in Fastr. The *Network* controls the main execution, it sorts the *Nodes* required and executes those, resulting in a list of jobs to be run. The jobs are dispatched via an execution plugin. The job is then executed. On execution, all arguments are translated to values and paths that the *Tool* can use. The *Tool* then sets the environment and, finally, calls the *Interface* for the actual running of the underlying task.

Currently, Fastr supports functional plugins for processing locally and on a cluster (using the DRMAAv1 API⁵). Future plugins will focus on flexible middleware for grid/cluster/cloud, like Dirac,⁶ that offer support for a wide range of systems. For creating a new plugin, five methods need to be implemented: an

⁵<http://www.drmaa.org>

⁶<http://diracgrid.org>

initialization and a cleanup method as well as methods for queuing, releasing and canceling a job.

2.6. Provenance

Data provenance is a built-in feature of Fastr and is based on an implementation of the W3C *PROV-DM: Prov Data model recommendation* (Belhajjame et al., 2013). Fastr records all relevant data during execution and ensures that for every resulting file a complete data provenance document is included. The standard format of a provenance document is PROV-N, which can be serialized to PROV-JSON or PROV-XML.

In **Figure 6**, the three base classes and the properties of how they relate to each other are illustrated. For Fastr, *Networks*, *Tools*, and *Nodes* are modeled as agents, jobs as activities and data objects as entities. The relating properties are naturally valid for our workflow application. The hierarchy and topology of the *Network* follows automatically from the relating properties between the classes, but in order to make the provenance document usable for reproducibility, extra information is stored as attributes on the classes and properties. For every *Tool*, the version is stored. For every data sample, the value or file path and a checksum is stored. For every job, the start and end time of execution, the stdout and stderr logs are stored, the end status (success, success with warnings, failed, etc.), and an exhaustive description of the execution environment.

2.7. Visualization

To give the user insight in the data flow through the *Network*, it is possible to visualize the *Network* using graphviz (Gansner and North, 2000). The figures in this paper that show examples of *Networks* (**Figures 1 and 7**) are generated automatically by Fastr. Fastr plots the *Tool* as a collection of inputs and outputs and draws the links between them.

Because Fastr allows for more advanced data flows, there is a few visualization options that can aid users in validating the data flow. First, the color of a link changes if the flow in the Link is different. Second, there is an option to draw the dimension sizes in a *Network*. This shows the number of dimension and the expected size (as symbols). A simple example of the visualization of a more advanced dataflow is given in **Figure 7**.

3. EVALUATION

A functional version of Fastr is available from https://bitbucket.org/bigr_erasmusmc/fastr. Fastr is open-source and free to use (under the Apache license 2.0). The framework is written in Python and easy to install using the python package index (`pip install`)⁷ or using the included `setuptools` from the source distribution. Fastr is platform independent and runs on Linux, Mac, and Windows environments. However, Linux support is much more stable, since that is the platform used in most processing environments.

Documentation is available at <http://fastr.readthedocs.io>; it includes a quick start tutorial, a user manual and a developer reference of the code. The documentation is built using Sphinx.

The Fastr software is composed of core modules and plugins. The core modules implement the networking, data flow, and interfacing with the plugins. The plugins provide the data input/output, and execution functionality. Fastr is tested for code quality using both unit tests and functional tests. The unit tests are limited to the core modules and ensure the integrity of the core on a fine grained level. The functional testing covers the building and execution of small *Networks*. The functional tests validate the functional requirements of Fastr. Both the unit and the functional

⁷<https://pypi.python.org/pypi/fastr>

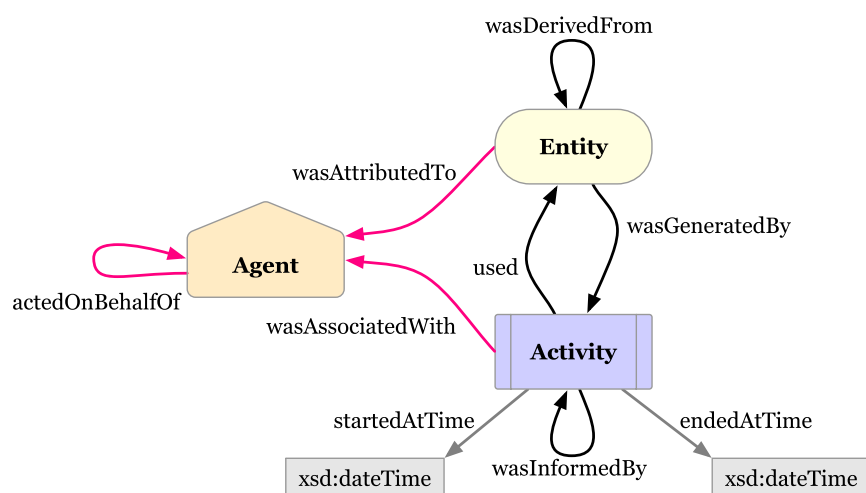


FIGURE 6 | The three base classes of the provenance data model with their relating properties. The agents are orange pentagons, the entities are yellow ovals and the activities are depicted as blue squares. This image is copied from PROV-O: The PROV Ontology. Copyright © 2015 W3C® (MIT, ERCIM, Keio, Beihang). <http://www.w3.org/Consortium/Legal/2015/doc-license>

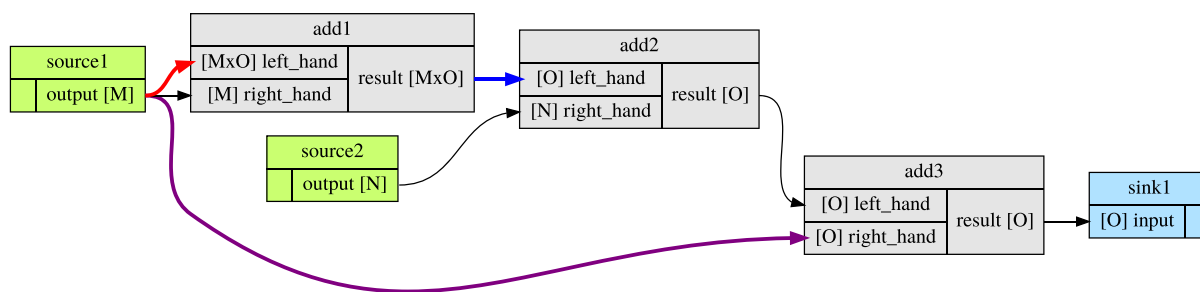


FIGURE 7 | An example of flow visualization. The colored arrows indicate the flow directive in the link: red for expand, blue for collapse, and purple for a combination of both. After each input and output, the dimensions are printed in square bracket. In this workflow, the dimensions N and O should match, but the system can only validate this at runtime.

tests are performed continuously using the continuous integration framework Jenkins.⁸

Currently, we are using Fastr for a number of workflows for several single-center and multi-center studies. For example, the Rotterdam Scan Study (Ikram et al., 2011), containing over 12,000 scan sessions, uses an analysis pipeline implemented in Fastr for the preprocessing, tissue type segmentation, white matter lesion segmentation and lobes segmentation of brain MR images (see Figure 8). The data are fetched from the archive and is processed in a cluster environment. The resulting data are stored in an image archive.

Fastr has been used to run this workflow on new batches of subjects since mid 2015. Its performance has proven to be very stable as the workflow always succeeded. The overhead is limited as the Fastr workflow engine uses only a fraction of the resource compared the underlying Tools.

4. DISCUSSION

With Fastr we created a workflow system that allows users to rapidly create workflows. The simple access to advanced features makes Fastr suitable for both simple and complex workflows. Workflows created with Fastr will automatically get data provenance, support for execution on various computational resources, and support for multiple storage systems. Therefore, Fastr speeds up the development cycle for creating workflows and minimizes the introduction of errors.

Fastr offers a workflow system that works with tools that can really be black boxes, they do not need to implement a specific API as long as their inputs and outputs can be defined. Fastr can manage multiple versions of tools, as we believe it is important to be able to keep an environment where all the old versions of tools are available for future reproducibility of the results. Additionally, it provides provenance records for every result for reproducibility of the experiments. Batch processing and advanced data flows are at the core of Fastr's design. Fastr communicates with processing backends and data providers via plugins allowing interoperability with other components of research infrastructures.

4.1. Workflow Languages

Most workflows systems and languages are simpler with respect to data flow. However, there are two languages that have features similar to that of Fastr. Taverna, using the SCUFL2 language, has a concept of a dot product or cross product for input ports. This is equivalent to the use of input groups in Fastr. Also the MOTEUR (Glatard et al., 2008) system, using the GWENDIA (Montagnat et al., 2009) language, has the same cross product and dot product concepts.

A main difference between Fastr and the other two languages is that Fastr describes the data as N-D arrays, and a cross product increases the number of dimensions, whereas GWENDIA and SCUFL2 follow the list (of lists) principle. Of course, a list of lists can be seen as a 2D array, but that is not used by the aforementioned languages.

There is also the recent effort of the Common Workflow Language, CWL (Amstutz et al., 2016). The CWL includes a specification for tools and workflows. The CWL has a support for an optional scatter directive. This allows a cross product type of behavior. However, this is not part of main specification, but rather an optional feature.

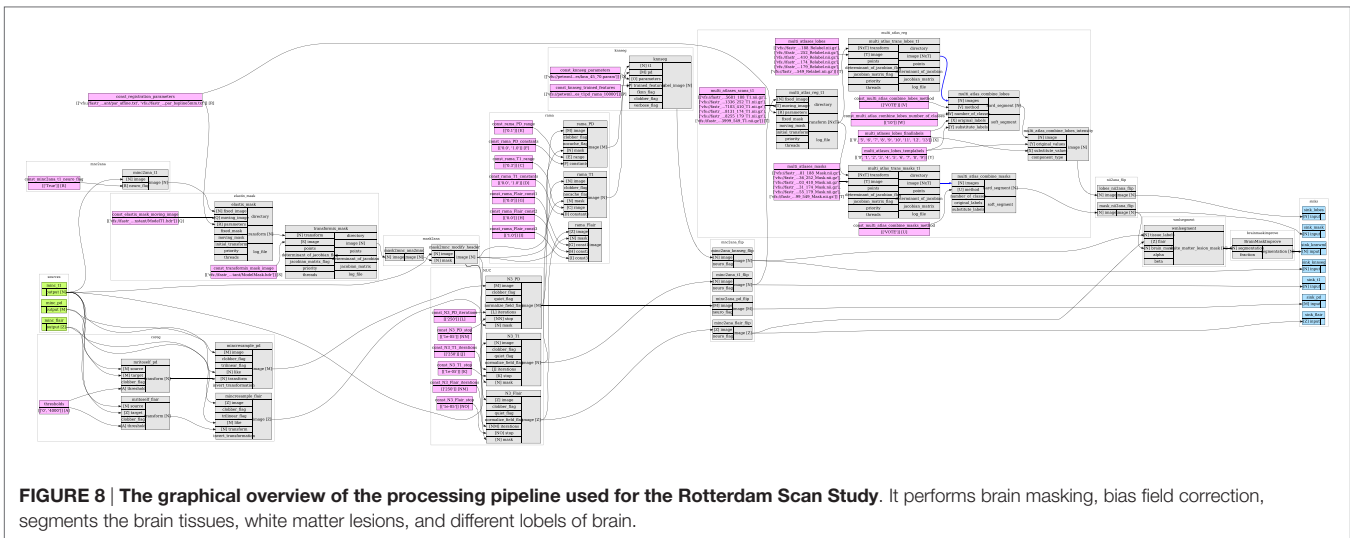
4.2. Limitations

The Fastr workflow system has been created with some clear goals, but there are also some limitations in the design. First of all, our design is created with automated processing workflows in mind and there is no support for interactive steps in the workflow. This is a design choice and there are no plans to address this issue.

Maybe the largest drawback of Fastr is that as a new system the amount of Tools available is limited. The Tool wrappers and interfaces are very flexible, but compared to systems as LONI pipeline and Nipype there is a lack of resources. This is a problem any new system faces and we believe that in time this issue will be resolved.

A similar issue is the limited number of execution backend plugins. The system is plugin based and has the potential to support almost any computational resource, but currently only supports local execution and cluster environments. We will add new plugins whenever a project requires one, but do not aim to create many additional plugins on the short term. For grid execution, this could be more challenging due to the lack of credentials

⁸<https://jenkins.io>



management in Fastr. Currently, we do not facilitate advanced credential storage, which is often an important requirement in grid computing.

The system is currently completely command-line based and offers no graphical user interface (GUI). Since the focus of Fastr is batch processing, the target environments are mostly headless. It is good practice to completely decouple core functionality from the user interface, especially when running in headless environments. Therefore, we decided to spend our time on creating a solid workflow engine before creating a GUI. We believe that the tooling can always be added and improved later, but that the core design limitations are generally harder to solve in the future. We plan on adding more (graphical) tools that provide more convenient user interaction in the future.

And finally, we are not satisfied with our current test code coverage. We have test for some core functionality, but the code coverage of the unit tests on the low side. This is partially offset by the functional testing, but we feel we should improve the test code coverage to avoid technical debt.

4.3. Future Directions

Because of the differences in design philosophy, Fastr and Nipype are complementary in focus: Fastr is created for managed workflows and has tools and interfaces as a necessity, whereas the interfaces are the primary focus of Nipype. Considering that there are many interfaces available for Nipype, we created a prototype NipypeInterface in Fastr, which allows `Tools` in Fastr to use Nipype for the interface. This is still experimental and there are still some limitations because Nipype and Fastr have incompatible data type systems.

Another option to increase the amount of tools available is to start supporting Boutiques.⁹ Boutiques are an application repository with a standard packaging of tools, so that they can be used on multiple platforms. The boutiques applications are somewhat similar to Fastr `Tools`, as they describe the inputs and output

in a JSON file. Additionally, the underlying binaries, scripts, and data are all packaged, versioned, and distributed using Docker¹⁰ containers. It would require to either rewrite the boutique inputs/outputs into a Fastr interface or to create a new interface class for Boutiques.

Although the CWL at the moment is as far as we know not used in the medical imaging domain, we think that support for the CWL is an important future feature for Fastr as we fully support the idea to have a common standard language. Support for CWL tools in Fastr could possibly use a new interface class, but the support for workflows would probably need to be an import/export that transcribes workflows from CWL to Fastr and back.

For reproducibility, it is important to be able to re-run analyses in exactly the same conditions. Currently, Fastr supports environment modules to keep multiple versions of software available at the same time. However, the same version of the software can still be different based on underlying libraries, compiler used, and the OS. Virtual Machines or Linux Containers offer a solution to this problem. Linux containers, such as Docker and LXC, are often seen as a light-weight alternative to Virtual Machines. They ensure that the binaries and underlying libraries are all managed, but they use the kernel of the host OS. We plan to add support for Docker containers to make it easier to share tools and improve reproducibility further.

For continuous integration, we have a Jenkins (see text footnote 8) continuous integration server that runs our tests nightly. Additionally, we use SonarQube¹¹ for inspecting code quality, technical debt, and code coverage. We are aiming for each release to increase the code coverage and to decrease the technical debt.

Finally, we are working on more (web-based) tooling around Fastr to make it easier to visualize, develop, and debug `Networks` and to inspect the results of a run (including provenance information).

⁹<http://boutiques.github.io/>

¹⁰<https://www.docker.com>

¹¹<http://www.sonarqube.org/>

GLOSSARY

API – An application programming interface, a set of functions and protocols that allow the creation of applications that access the features another application or service.

Cardinality – The number of elements in a grouping. For Fastr specifically, this is the number of elements contained in a sample.

Code coverage – A measure indicating what part of the code is covered by a test suite. This is often expressed as a percentage of the total lines of code.

JSON – JavaScript Object Notation is an open data format that is used often in client-server communication and uses human readable text to present data in key-value pairs.

Linux Containers – Virtualization for running multiple isolated linux systems on one Linux kernel on the operating system level.

MapReduce – A programming model for processing large datasets. Typically, it consists of a *Map* operation on the elements and a *Reduce* operation that aggregates the elements into a final result.

Population imaging – Population imaging is the large-scale acquisition and analysis of medical images in controlled population cohorts. Population imaging aims to find imaging biomarkers that allow prediction and early diagnosis of diseases and preventive therapy.

Provenance – Report of the origin and operations that has been done on an object.

technical debt – A concept in programming that reflects the extra work that is the results of using quick solutions instead of the proper solution.

XML – eXtensible Markup Language is a human readable markup language for encoding documents.

AUTHOR CONTRIBUTIONS

The Fastr workflow engine described in the article was designed and implemented primarily by HA and MK under supervision of WN. The manuscript was written primarily by HA and MK, and was revised by WN. All of the authors approved this work for publication.

ACKNOWLEDGMENTS

We would like thank Coert Metz and Fedde van der Lijn for helping us with the creating the very first prototype of the system. We are grateful to the people from the BIGR group who were the first to test the system, give valuable feedback, and enough patience to allow us to improve to Fastr further. We would like to thank Esther Bron for proofreading the paper and giving us valuable feedback. Finally, we would like to thank all people who have contributed in any way to Fastr.

FUNDING

This work was supported by the following projects: The Heart Brain Connection Consortium, supported by the Netherlands Cardiovascular Research Initiative (CVON2012-06); Population Imaging Infrastructuur in Medical Delta, supported by European Regional Development Fund (Kansen voor West) and co-financed by the province of South-Holland; and BBMRI-NL2.0 (see text footnote 1).

REFERENCES

- Amstutz, P., Andeer, R., Chapman, B., Chilton, J., Crusoe, M. R., Guimer, R. V., et al. (2016). *Common Workflow Language, Draft 3*. doi:10.6084/m9.figshare.3115156.v2
- Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., et al. (2013). *PROV-DM: The PROV Data Model. Recommendation, W3C*. Available at: <http://www.w3.org/TR/2013/REC-prov-dm-20130430/>
- Berthold, M. R., Cebon, N., Dill, F., Gabriel, T. R., Kötter, T., Meinel, T., et al. (2008). *KNIME: The Konstanz Information Miner*. Springer.
- Berthold, M. R., Cebon, N., Dill, F., Gabriel, T. R., Kötter, T., Meinel, T., et al. (2009). Knime-the konstanz information miner: version 2.0 and beyond. *ACM SIGKDD Explor. Newslett.* 11, 26–31. doi:10.1145/1656274.1656280
- Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., et al. (2010). Neuroimaging study designs, computational analyses and data provenance using the Ioni pipeline. *PLoS ONE* 5:e13070. doi:10.1371/journal.pone.0013070
- Gansner, E. R., and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* 30, 1203–1233. doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.3.CO;2-E
- Glatard, T., Montagnat, J., Lingrand, D., and Pennec, X. (2008). Flexible and efficient workflow deployment of data-intensive applications on grids with moteur. *Int. J. High Perform. Comput. Appl.* 22, 347–360. doi:10.1177/1094342008096067
- Goecks, J., Nekrutenko, A., Taylor, J., and The Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86. doi:10.1186/gb-2010-11-8-r86
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5:13. doi:10.3389/fninf.2011.00013
- Ikram, M. A., van der Lugt, A., Niessen, W. J., Krestin, G. P., Koudstaal, P. J., Hofman, A., et al. (2011). The rotterdam scan study: design and update up to 2012. *Eur. J. Epidemiol.* 26, 811–824. doi:10.1007/s10654-011-9624-z
- Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluijm, J. P. (2010). elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi:10.1109/TMI.2009.2035616
- Marcus, D. S., Olsen, T. R., Ramaratnam, M., and Buckner, R. L. (2007a). The extensible neuroimaging archive toolkit. *Neuroinformatics* 5, 11–33. doi:10.1385/NI:5:1:11
- Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007b). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J. Cogn. Neurosci.* 19, 1498–1507. doi:10.1162/jocn.2007.19.9.1498
- Montagnat, J., Isnard, B., Glatard, T., Maheshwari, K., and Fornarino, M. B. (2009). “A data-driven workflow language for grids based on array programming principles,” in *Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, WORKS '09* (New York, NY: ACM), 7:1–7:10.
- Morrison, J. P. (2010). *Flow-Based Programming, 2nd Edition: A New Approach to Application Development*. Paramount, CA: CreateSpace.
- Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., et al. (2005). Ways toward an early diagnosis in Alzheimers disease: the Alzheimers Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement.* 1, 55–66. doi:10.1016/j.jalz.2005.06.003
- Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., et al. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurr. Comput.* 18, 1067–1100. doi:10.1002/cpe.993
- Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The Ioni pipeline processing environment. *Neuroimage* 19, 1033–1048. doi:10.1016/S1053-8119(03)00185-X

- Sherif, T., Rioux, P., Rousseau, M.-E., Kassis, N., Beck, N., Adalat, R., et al. (2014). CBRAIN: a web-based, distributed computing platform for collaborative neuroimaging research. *Front. Neuroinform.* 8:54. doi:10.3389/fninf.2014.00054
- van Buchem, M. A., Biessels, G. J., Brunner la Rocca, H. P., de Craen, A. J., van der Flier, W. M., Ikram, M. A., et al. (2014). The Heart-Brain Connection: a multidisciplinary approach targeting a missing link in the pathophysiology of vascular cognitive impairment. *J. Alzheimers Dis.* 42, S443–S451. doi:10.3233/JAD-141542
- Yoo, T. S., Ackerman, M. J., Lorensen, W. E., Schroeder, W., Chalana, V., Aylward, S., et al. (2002). Engineering and algorithm design for an image processing Api: a technical report on ITK-the insight toolkit. *Stud. Health Technol. Inform.* 85, 586–592. doi:10.3233/978-1-60750-929-5-586

Conflict of Interest Statement: WN is cofounder, part-time Chief Scientific Officer, and stockholder of Quantib BV. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Achterberg, Koek and Niessen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Reproducible Large-Scale Neuroimaging Studies with the OpenMOLE Workflow Management System

Jonathan Passerat-Palmbach^{1*}, Romain Reuillon², Mathieu Leclaire², Antonios Makropoulos¹, Emma C. Robinson¹, Sarah Parisot¹ and Daniel Rueckert¹

¹ BioMedIA Group, Department of Computing, Imperial College London, London, UK, ² Institut des Systemes Complexes Paris Ile de France, Paris, France

OPEN ACCESS

Edited by:

Michel Dojat,
INSERM, France

Reviewed by:

Zhengyi Yang,
The University of Queensland,
Australia
K Jarrod Millman,
University of California, Berkeley, USA

*Correspondence:

Jonathan Passerat-Palmbach
j.passerat-palmbach@imperial.ac.uk

Received: 22 April 2016

Accepted: 01 March 2017

Published: 22 March 2017

Citation:

Passerat-Palmbach J, Reuillon R, Leclaire M, Makropoulos A, Robinson EC, Parisot S and Rueckert D (2017) Reproducible Large-Scale Neuroimaging Studies with the OpenMOLE Workflow Management System. *Front. Neuroinform.* 11:21. doi: 10.3389/fninf.2017.00021

OpenMOLE is a scientific workflow engine with a strong emphasis on workload distribution. Workflows are designed using a high level Domain Specific Language (DSL) built on top of Scala. It exposes natural parallelism constructs to easily delegate the workload resulting from a workflow to a wide range of distributed computing environments. OpenMOLE hides the complexity of designing complex experiments thanks to its DSL. Users can embed their own applications and scale their pipelines from a small prototype running on their desktop computer to a large-scale study harnessing distributed computing infrastructures, simply by changing a single line in the pipeline definition. The construction of the pipeline itself is decoupled from the execution context. The high-level DSL abstracts the underlying execution environment, contrary to classic shell-script based pipelines. These two aspects allow pipelines to be shared and studies to be replicated across different computing environments. Workflows can be run as traditional batch pipelines or coupled with OpenMOLE's advanced exploration methods in order to study the behavior of an application, or perform automatic parameter tuning. In this work, we briefly present the strong assets of OpenMOLE and detail recent improvements targeting re-executability of workflows across various Linux platforms. We have tightly coupled OpenMOLE with CARE, a standalone containerization solution that allows re-executing on a Linux host any application that has been packaged on another Linux host previously. The solution is evaluated against a Python-based pipeline involving packages such as scikit-learn as well as binary dependencies. All were packaged and re-executed successfully on various HPC environments, with identical numerical results (here prediction scores) obtained on each environment. Our results show that the pair formed by OpenMOLE and CARE is a reliable solution to generate reproducible results and re-executable pipelines. A demonstration of the flexibility of our solution showcases three neuroimaging pipelines harnessing distributed computing environments as heterogeneous as local clusters or the European Grid Infrastructure (EGI).

Keywords: high performance computing, reproducibility, pipeline, large datasets, parameter exploration, neuroimaging, workflow systems

1. INTRODUCTION

1.1. Problem

Larger sample sizes increase statistical power by reducing the variance of the sampling distribution. With large datasets like the Human Connectome Project¹ (HCP) now freely available, one of the reasons why large studies are not more often conducted is the tremendous amount of computing power required. Distributed computing can offer this processing power but it can be hard to set up a distributed experiment for non-computer scientists.

Another important aspect to increase the quality and impact of scientific results is their capacity to be reproduced, especially by a different scientist. Researchers are more and more encouraged to share their experiments and the source code that led to the results they present. In order to be usable by other researchers, experiments have to be organized in a certain way.

Researchers are thus faced with two major problems in order to produce top quality studies: the necessity to provide a reproducible experimental protocol, and the technical challenge to upscale their implemented solutions to cope with large datasets. The whole solution must be made available in a relatively standard way so that other groups can pick up the experiment and re-run against their own set of resources and data.

What is the best way to describe experiments so that they can easily be reproduced by other researchers? Workflow, or pipelines, are a common way to model scientific problems involving different tools along multiple distinct stages. Although some initiatives try to unify workflow description (Amstutz et al., 2016), a majority of researchers still compose their pipelines using plain shell scripts. This approach makes it very hard to share the resulting pipelines, as shell scripts are strongly tied to their definition environment. Scripting languages are perfectly satisfying for workflow definition as long as they offer the readability and guided design that a high-level programming language does.

However, can we simply rely on a high-level scripting language to distribute the workload resulting from a pipeline? *Ad hoc* solutions to submit jobs to a local cluster are very efficient to quickly run an experiment. However, they cannot manage job resubmissions on unexpected failures, and are very unlikely to manage several computing environments. The resulting pipeline is once again not suitable to share with other researchers using another computing environment. A very good example in a widely distributed software package is FSL² (FMRIB Software Library), which ships with pipelines that can only be delegated to a Sun Grid Engine (SGE) cluster.

Some applications might show more complicated than others to distribute in view of the complex set of dependencies they require for their execution. The DevOps community has tackled the problem of complex application deployments with an increasing use of software containers, the most famous solution being Docker. However, scientific computing environments are often designed as High Performance Computing (HPC) clusters, and cannot be customized for each user's needs. Cutting-edge

containerization solution such as Docker are not available on these platforms, most of the time for security reasons as they require administrator privileges. While this is not a problem to empower the owner of a virtual machine with such privileges, HPC administrators are reluctant to grant such powers to researchers.

In order to build reproducible experiments at large scale, we thus need three elements:

- a simple access to large scale HPC/cloud environments
- a high-level formalism, such as workflows, to express the experiment in a portable way
- a standalone container platform that do not require administrator privileges at any point of its execution chain

In this paper, we introduce how the OpenMOLE (Reuillon et al., 2013) workflow management system can be paired up with the user-level archiver CARE (Janin et al., 2014) to address these problems in the context of large medical imaging studies.

1.2. Proposed Solution

OpenMOLE is a generic workflow management solution not targeting a particular community. It allows users to embed their own application, rather than limiting them to a set of pre-packaged tools made available for a specific usage. Although this approach requires more involvement from the user's side, it also gives them more flexibility. Further down the line, a pipeline solution tailored for a specific field might not be suitable for multidisciplinary studies. In the specific case of neuroimaging projects, it is not rare to also collect genetics data in order to combine it with the information extracted from the images.

Reproducibility and sharing of OpenMOLE workflows start with its Domain Specific Language (DSL) that is used to describe the workflow steps and connections. The OpenMOLE DSL is an embedded DSL, written as a set of extensions to the Scala programming language. As a superset to Scala, it benefits from all the constructs available in this high-level programming language and harnesses Scala's strong type system to make workflow descriptions more meaningful and less error-prone. As a Scala application, OpenMOLE runs in the Java Virtual Machine (JVM) runtime. This makes it agnostic to its underlying Operating System (OS) and is another step toward sharing OpenMOLE workflows from one user to another, regardless of their work environment.

OpenMOLE is built with a strong focus toward the distribution of a pipeline workload to remote computing environments. Pipelines defined within the OpenMOLE framework are totally decoupled from the environments on which they are executed. This allows running the same pipeline on different environments without modifying the definition of the pipeline itself. On top of that, OpenMOLE was designed to enable a fine granularity of distribution. Individual tasks, or groups of tasks, can be deployed to different computing environments. This is particularly useful when a task of the pipeline requires specific devices such as GPUs to run, while the rest of the pipeline can be distributed to classic CPUs.

¹<http://humanconnectome.org/>.

²<http://fsl.fmrib.ox.ac.uk>.

This work presents the integration of CARE archives as a new foundation to make tasks re-executable on the various computing environments supported by OpenMOLE. The CARE toolkit (Janin et al., 2014) provides a standalone containerization solution that does not need administrator privileges to re-execute on target hosts. While this perfectly fits our requirements for a solution in par with HPC environments' constraints, CARE cannot be used on its own to provide a standard format of exchange for scientific applications. It has not been built with this kind of applications in mind and focuses on providing low-level elements ensuring re-executability of a command line on any other Linux machine. However, its possibilities can be harnessed to form the base of a new OpenMOLE task re-executable on multiple environments.

Medical imaging pipelines are ideal candidates to evaluate our solution as they typically involve an heterogeneous software ecosystem. These software pieces usually come with a broad set of dependencies that are hard to track manually. They also manipulate large datasets that cannot be embedded in the software container and have to be transferred separately to the execution node running the current stage of the pipeline. The same remark applies to the pipeline's results as can be seen in Parisot et al. (2015) for instance.

1.3. Related Work

1.3.1. Generic Workflow Engines

Like OpenMOLE, other initiatives made the choice not to target a specific community. Kepler (Altintas et al., 2004) was one of the first general-purpose scientific workflow systems, recognizing the need for transparent and simplified access to high performance computing platforms more than a decade ago. Pegasus (Deelman et al., 2005) is a system that initially gained popularity for mapping complex workflows to resources in distributed environments without requiring input from the user.

PSOM (Pipeline System for Octave and Matlab) (Bellec et al., 2012) is a workflow system centered around Matlab/Octave. Although this is certainly a good asset for this userbase, it revolves around Matlab, a proprietary system. This hinders by definition sharing workflows to the wider community and reduces the reproducibility of experiments.

1.3.2. Community-Tailored Workflow Engines

On the other hand, some communities have seen the emergence of tailored workflow managers. For example, the bioinformatics community has developed Taverna (Oinn et al., 2004) and Galaxy (Goecks et al., 2010) for the needs of their community.

In the specific case of the neuroimaging field, two main solutions emerge: NiPype (Gorgolewski et al., 2011) and LONI (Rex et al., 2003). NiPype is organized around three layers. The most promising one is the top-level common interface that provides a Python abstraction of the main neuroimaging toolkits (FSL, SPM, ...). It is extremely useful to compare equivalent methods across multiple packages. NiPype also offers pipelining possibilities and a basic workload delegation layer only targeting the cluster environments SGE and PBS. Workflows are delegated

to these environments as a whole, without the possibility to exploit a finer grain parallelism among the different tasks.

The LONI Pipeline provides a graphical interface for choosing processing blocks from a predefined library to form the pipeline. It supports workload delegation to clusters preconfigured to understand the DRMAA API (Tröger et al., 2012).

However, the LONI Pipeline displays limitations at three levels. First, the format used to define new nodes is XML (eXtensible Markup Language), and assumes the packaged tools offer a well-formed command line and its input parameters. On this aspect, the Python interfaces forming NiPype's top layer is far superior to LONI pipeline's approach. Second, one might also regret the impossibility to script workflows, to the best of our knowledge.

The third and main drawback of the LONI pipeline is in our opinion its restrictive licensing, which prevents an external user to modify and redistribute the modifications easily. Previous works in the literature have shown the importance of developing and releasing scientific software under Free and Open Source licenses (Stodden, 2009; Peng, 2011). This is of tremendous importance to enable reproducibility and thorough peer-reviewing of scientific results.

Finally, we have recently noted another effort developed in Python: FastR³ (Achterberg et al., 2015). It is designed around a plugin system that enables connecting to different data sources or execution environments. At the moment, execution environments can only be addressed through the DRMA (Distributed Resource Management Application) API but more environments should be provided in the future.

1.3.3. Level of Support of HPC Environments

Table 1 lists the support for various HPC environments in the workflow managers studied in this section. It also sums up the features and domains of application for each tool.

To the best of our knowledge, we are not aware of any workflow engine that targets as many environments as OpenMOLE, but more importantly that introduces an advanced service layer to distribute the workload. When it comes to very large scale infrastructures such as grids and clouds, sophisticated submission strategies taking into account the state of the resources as well as implementing a level of fault tolerance must be available. Most of the other workflow engines offer service delegation layers that simply send jobs to a local cluster. OpenMOLE implements expert submission strategies (job grouping, over submission, ...), harnesses efficient middlewares such as Dirac, and automatically manages end-to-end data transfer even across heterogeneous computing environments.

Compared to other workflow processing engines, OpenMOLE promotes a zero-deployment approach by accessing the computing environments from bare metal, and copies on-the-fly any software component required for a successful remote execution. OpenMOLE also encourages the use of software components developed in heterogeneous programming

³<http://www.fastr.eu/>.

TABLE 1 | Summary table of the features, HPC environments supported and domains of application of various workflow managers.

Workflow engine	Local multi-processing	HPC support	Grid support	Cloud support
Galaxy ⁴	Yes	DRMAA clusters	No	No (manual cluster deployment)
Taverna ⁵	Yes	No	No	No
FastR	Yes	DRMAA clusters	No	No
LONI ⁶	No	DRMAA clusters	No	No (manual cluster deployment)
NiPype	Yes	PBS/Torque, SGE	No	No
Kepler ⁷	Yes	PBS, Condor, LoadLeveler	Globus	No
Pegasus ⁸	No (need local Condor)	Condor, PBS	No	No (manual cluster deployment)
PSOM	Yes	No	No	No
OpenMOLE	Yes	Condor, Slurm, PBS, SGE, OAR	<i>Ad hoc</i> grids, gLite/EMI, Dirac, EGI	EC2 (fully automated) ⁹

Workflow engine	Scripting support	GUI	Generic/Community	License
Galaxy	No	Yes	Bioinformatics	AFL 3.0
Taverna	No	Yes	Bioinformatics	Apache 2.0
FastR	Python	No	Neuroimaging	BSD
LONI	No	Yes	Neuroimaging	Proprietary (LONI)
NiPype	Python	No	Neuroimaging	BSD
Kepler	Partly with R	Yes	Generic	BSD
Pegasus	Python, Java, Perl	No	Generic	Apache 2.0
PSOM	Matlab	No	Generic	MIT
OpenMOLE	Domain Specific Language, Scala	Yes	Generic	AGPL 3

Information was drawn from the web pages in footnote when present, or from the reference paper cited in the section otherwise.

languages and enables users to easily replace the elements involved in the workflow.

1.4. Main Contributions

This paper puts the light on OpenMOLE's new features enabling large-scale pipelines to be reproducible while distributed to a large range of computing environments.

We first describe the three main elements from the OpenMOLE platform: (1) the DSL to design meaningful, reusable workflows, (2) the integration and simple access to a wide range of High Performance Computing (HPC) environments, and (3) the embedded parameter exploration methods (Section 2).

As evoked in the introduction, distributing an application can be troublesome. We list the potential issues encountered when distributing a typical medical imaging pipeline in Section 3. We then justify the solution chosen to enable re-executability and sharing of experiments in Section 3.2, and detail its implementation in OpenMOLE in Section 3.3.

This solution is evaluated with a workflow exploring the performance of different parameter initializations for decoding fMRI acquisitions from a canonical dataset (Haxby et al., 2001) (Section 4). The decoder is taken from the NiLearn

tutorials (Abraham et al., 2014) and demonstrates how a workflow made of a complex combination of Python and native binary dependencies can be successfully reproduced on different computing platforms without any prior knowledge regarding the state of their software stack. This study demonstrates the potential of this work to process a well-known dataset for which the performance and validity of the pipeline can be evaluated.

As a case-study, we finally detail three neuroimaging pipelines managed by OpenMOLE and the different benefits brought by the platform and its software ecosystem (Section 5).

2. WHAT IS OPENMOLE?

Scientific experiments are characterized by their ability to be reproduced. This implies capturing all the processing stages leading to the result. Many execution platforms introduce the notion of workflow to do so (Barker and Van Hemert, 2008; Mikut et al., 2013). Likewise, OpenMOLE manipulates workflows and distributes their execution across various computing environments.

A workflow is a set of tasks connected through transitions. From a high level point of view, tasks comprise inputs, outputs and optional default values. Tasks describe what OpenMOLE should execute and delegate to remote environments. They embed the actual applications to study. Depending on the kind of program (binary executable, Java...) to embed in OpenMOLE, the user chooses the corresponding task. Tasks execution depends on inputs variables, which are provided by the dataflow. Each task

⁴<https://wiki.galaxyproject.org/>.

⁵<https://taverna.incubator.apache.org/introduction/taverna-features>.

⁶<http://pipeline.loni.usc.edu/explore/features/>.

⁷<https://code.kepler-project.org/code/kepler-docs/trunk/outreach/documentation/shipping/2.5/UserManual.pdf>.

⁸https://pegasus.isi.edu/documentation/execution_environments.php.

⁹<https://github.com/adraghici/openmole/tree/aws-env>.

produces outputs returned to the dataflow and transmitted to the input of consecutive tasks. OpenMOLE exposes entry points to inject data in the dataflow (*sources*) and extract useful results at the end of the experiment (*hooks*).

As shown in **Figure 1**, OpenMOLE revolves around three main elements: the *Applications*, the exploration *Methods* and the support of *Massively parallel environments*. These three components are put together in a common DSL to describe the workflows.

We will give a quick overview of these different components in the subsections. For more details regarding the core implementation and features of OpenMOLE, interested readers can refer to Reuillon et al. (2010, 2013, 2015a) and the OpenMOLE website (Reuillon et al., 2015b).

2.1. A DSL to Describe Workflows

According to Barker and Van Hemert (2008), workflow platforms should not introduce new languages but rely on established ones. OpenMOLE's DSL is based on the high level Scala programming language (Odersky et al., 2004).

OpenMOLE's DSL introduces new operators in the Scala programming language to manage the construction and execution of the workflow. The advantage of this approach lies in the fact that workflows can exist even outside the OpenMOLE environment. As a high-level language, the DSL can be assimilated to an algorithm described in pseudo-code, easily understandable by most scientists. Moreover, it denotes all the types and data used within the workflow, as well as their origin. This reinforces the capacity to reproduce workflow execution both within the OpenMOLE platform or using another tool.

The philosophy of OpenMOLE is *test small* (on a local computer) and *scale for free* (on remote distributed computing environments). The DSL supports all the Scala constructs and provides additional operators and classes especially designed to compose workflows. OpenMOLE workflows expose implicit parallel aspects of the workload that can be delegated to distributed computing environments in a transparent manner.

2.2. Distributed Computing Environments

OpenMOLE helps delegate the workload to a wide range of HPC environments including remote servers (through SSH), clusters (supporting the job schedulers PBS, SGE, Slurm, OAR, and Condor), computing grids running the gLite/EMI middleware (through the WMS, CREAM and DIRAC entry points) and Amazon Elastic Compute Cloud (EC2). Support to these environments is implemented in GridScale¹⁰, a Free and Open Source Scala library.

Building on top of GridScale's as a service layer, OpenMOLE's simple workflow description is quite convenient to determine the computing environment best suited for a workflow. Switching from one environment to another is achieved by modifying a single line in the script. The granularity of the implementation allows each task of the workflow to be assigned to a different execution environment. This feature proves very useful when considering the limited availability of a particular resource (shared cluster) or its suitability to process a particular problem

(necessity to be processed on a GPU or another type of hardware accelerator).

The final workflow description can thus connect tasks using different software components but also running on heterogeneous execution environments thanks to GridScale's large support of HPC platforms.

The execution platform of OpenMOLE has proved to be robust enough to manage no less than half a billion instances (Schmitt et al., 2015) of a task delegated to the European Grid Infrastructure (EGI).

2.3. Exploration Methods

OpenMOLE has been designed with distributed parameter space exploration as a core use case (Reuillon et al., 2013). First its DSL comprehends a high level representation of design of experiments¹¹, which is concise and expressive. For instance expressing the exploration a full-factorial combination on a discrete parameter i , a continuous one x , a set of files f in a directory and replicate the experiment 10 times with randomly generated seeds s is expressed as shown in Listing 1:

```
val i = Val[Int]
val x = Val[Double]
val f = Val[File]
val s = Val[Long]
val exploration =
  ExplorationTask(
    (i in (0 to 10)) x
    (x in (0.0 to 100.0 by 10.0)) x
    (f in (workDirectory / "inputs")) x
    (s in (UniformDistribution[Long]()) take 10))
)
```

Listing 1 | Sampling example in OpenMOLE.

OpenMOLE also proposes advanced design of experiments with better coverage properties such as the low discrepancy Sobol sequence¹² and the Latin Hypercube Sampling (LHS)¹³. These sampling methods have been widely used for model exploration and are also adapted to evaluate other classes of parametric algorithms.

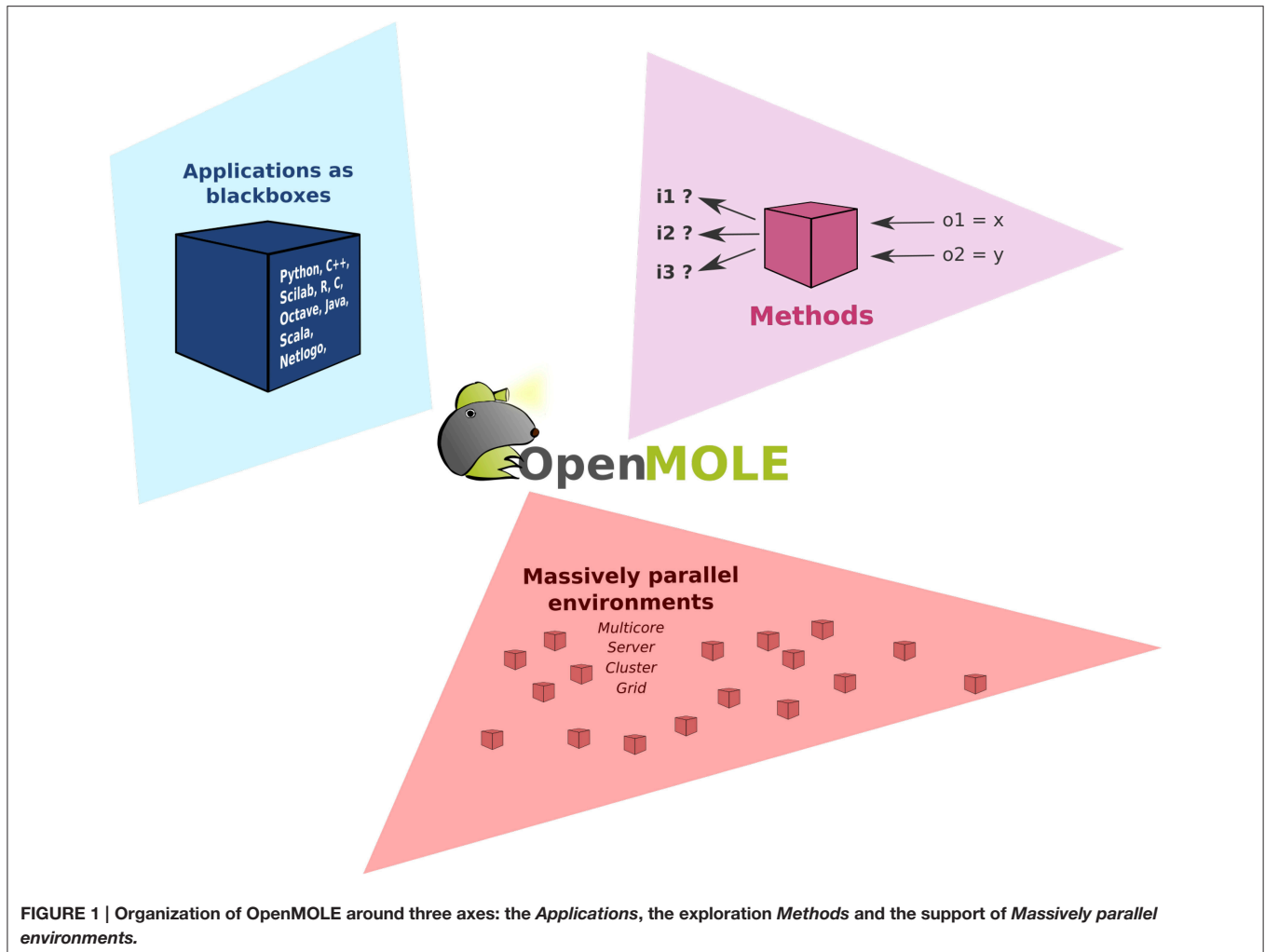
In addition to these classical a priori sampling methods, OpenMOLE generic formalism is a prolific playground to develop innovative exploration methods based on iterative refinement of the sampling. In these methods the results (*outputs*) of the explored program are taken into account in order to generate additional samples at interesting locations in the parameter space. These exploration methods are aimed to better comprehend the behavior of an application, or to finely tune parameters.

Several state-of-the-art iterative methods have been developed, evaluated and made available through OpenMOLE (multi-objective calibration (Schmitt et al., 2015), calibration profile (Reuillon et al., 2015c), Pattern Space Exploration Chérel et al., 2015; Cottineau et al., 2015a) and more are being developed such as the model family method (Cottineau et al., 2015b). Implementations of Evolutionary Algorithms (EA) techniques

¹¹http://www.openmole.org/current/Documentation_Language_Samplings.html.

¹²https://en.wikipedia.org/wiki/Sobol_sequence.

¹³http://en.wikipedia.org/wiki/Latin_hypercube_sampling.



taken from the literature such as (Deb et al., 2002) are also available.

Integrating these methods into OpenMOLE makes them available to a wide range of use cases (modeling, algorithm benchmarking, parameter tuning and testing applications...). The methods pair up perfectly with OpenMOLE as they are inherently parallel algorithms that can be distributed. The exploration methods elements of OpenMOLE thus benefit from the wide range of distributed computing environments available in the platform.

3. THE CHALLENGES OF DISTRIBUTING APPLICATIONS

3.1. Problems and Classical Solutions

Let us consider all the dependencies introduced by software bundles explicitly used by the developer. They can take various forms depending on the underlying technology. Compiled binary applications will rely on shared libraries, while interpreted languages such as Python will call other scripts stored in packages.

These software dependencies become a problem when distributing an application. It is very unlikely that a large number of remote hosts are deployed in the same configuration as a researcher's desktop computer. Actually, the larger the pool of distributed machines, the more heterogeneous they are likely to be.

If a dependency is missing at runtime, the remote execution will simply fail on the remote hosts where the requested dependencies are not installed. An application can also be prevented from running properly due to incompatibilities between versions of the deployed dependencies. This case can lead to silent errors, where a software dependency would be present in a different configuration and would generate different results for the studied application.

Silent errors break Provenance, a major concern of the scientific community (Miles et al., 2007; MacKenzie-Graham et al., 2008). Provenance criteria are satisfied when an application is documented thoroughly enough to be reproducible. This can only happen in distributed computing environments if the software dependencies are clearly described and available.

Some programming environments provide a solution to these problems. Compiled languages such as C and C++ offer to build a static binary, which packages all the software dependencies. Some applications can be very difficult to compile statically. A typical case is an application using a closed source library, for which only a shared library is available.

Another approach is to rely on an archiving format specific to a programming language. The most evident example falling into this category are Java Archives (JAR) that embed all the Java libraries an application will need.

A new trend coming from recent advances in the software engineering community is embodied by Docker. Docker has become popular along with other DevOps techniques to improve efficiency of software engineers. It enables shipping an application within a so-called container that will include the application and its required set of dependencies. Containers can be transferred just like an archive and re-executed on another Docker engine. Docker containers run in a sandboxed virtual environment but they are not to be confound with virtual machines. They are more lightweight as they don't embed a full operating system stack. The use of Docker for reproducible research has been tackled in Boettiger (2014) and Chamberlain et al. (2014).

The main drawback of Docker is that it implies deploying a Docker engine on the target host. Having a Docker engine running on every target host is an unlikely hypothesis in heterogeneous distributed environments such as computing grids. It is also impossible to deploy a Docker engine on the fly as its execution requires administrator privileges. Such privileges are not granted to end-users on HPC infrastructures at the heart of most scientific computing experiments. This is only the case in a fully-controlled environment, most of the time a cloud-based deployment where the user controls his own virtual machines.

The last option is to rely on a third-party application to generate re-executable applications. The strategy consists in collecting all the dependencies during a first execution in order to store them in an archive. This newly generated bundle is then shipped to remote hosts instead of the original application. This is the approach championed by tools like CDE (Guo, 2012), ReproZip (Chirigati et al., 2013), or CARE (Janin et al., 2014).

Considering all these aspects, the OpenMOLE platform has for long chosen to couple with tools providing standalone packages. While CDE was the initial choice, recent requirements in the OpenMOLE user community have led the development team to switch to the more flexible CARE. The next section will detail why OpenMOLE relies on CARE to package applications.

3.2. Why Should I CARE?

The first step toward spreading the workload across heterogeneous computing elements is to make the studied application executable on the largest number of environments. We have seen previously that this could be difficult with the entanglement of complex software environments available nowadays. For instance, a Python script will run only in a particular version of the interpreter and may also make use of binary dependencies. The best solution to make sure the execution will run as seamlessly on a remote host as it does

on the desktop machine of the scientist is to track all the dependencies of the application and ship them with it on the execution site.

OpenMOLE used to provide this feature through a third-party tool called CDE (Code, Data, and Environment packaging) (Guo, 2012). CDE creates archives containing all the items required by an application to run on any recent Linux platform. To do so, it tracks all the files that interact with the application and creates the base archive. At the time of writing, CDE appears not to be maintained anymore, the last significant contribution to the main source tree dating back from 2012¹⁴.

The only constraint regarding CDE is to create the archive on a platform running a Linux kernel from the same generation as those of the targeted computing elements. As a rule of thumb, a good way to ensure that the deployment will be successful is to create the CDE package from a system running Linux 2.6.32. Many HPC environments run this version, as it is the default kernel used by science-oriented Linux distribution, such as Scientific Linux and CentOS.

CARE on the other hand presents more advanced features than CDE. CDE actually displays the same limit than a traditional binary run on a remote host: i.e., the archive has to be generated on a platform running an old enough Linux kernel, to have a maximum compatibility with remote hosts. CARE lifts this constraint by emulating missing system calls on the remote environment. Thus, an application packaged on a recent release of the Linux kernel will successfully re-execute on an older kernel thanks to this emulation feature. CARE is, to the best of our knowledge, the only standalone solution ensuring re-execution on any Linux host, regardless of the original packaging host and without requiring administrator privileges.

We have also noted ReproZip (Chirigati et al., 2013) as a promising packaging solution. ReproZip's most interesting feature is to produce a package that can be re-run against different backends. Standalone archives can be extracted as plain folders, and then re-executed in a chrooted environment using the target host's environment and installed packages. Another option is to install them in the host system as a package in the case of a Debian-based Operating System. Although they don't require any pre-installed software, these solutions cannot ensure a successful re-execution due to low-level incompatibilities between the packaging and extraction environments. Other extraction solutions for ReproZip offer to run in a Vagrant virtual machine or a Docker container. However, none of these solution fit our design assumptions to exploit arbitrary environments without having to deploy anything beforehand.

The next section will describe how OpenMOLE integrates CARE seamlessly, as a first-class citizen in the DSL.

3.3. Combining OpenMOLE with CARE

Different types of tasks co-exist in OpenMOLE workflows, each embedding a different kind of application. Portable applications packaged with CARE are handled by the CARETask. Packaging an application is done once and for all by running the original application against CARE. CARE's re-execution mechanisms

¹⁴<https://github.com/pgbovine/CDE/commit/219c41590533846de12d7c5cca3f34a.ac471aae7>, last accessed 12-nov-16.

allow changing the original command line when re-running an application. This way we can update the parameters passed on the command line and the re-execution will be impacted accordingly. As long as all the configuration files, libraries, and other potential dependencies were used during the original execution, there is no need to package the application multiple times with different input parameters. To ensure all the initial execution conditions are captured, the environment variables defined in the session are also stored in the archive and populated on re-execution.

The newly packaged archive is the first argument expected by the CARETask. The second argument corresponds to a modified command line, updating the original call to specify a different parameter combination for each instance. The CARETask performs two actions: it first extracts the CARE archives by executing *archive.tgz.bin* (the archive is a self-extracting executable). The actual re-execution can then take place in the freshly unarchived work directory. Note that for each execution of the CARETask, any command starting with/is relative to the root of the CARE archive's filesystem, and any other command is executed in the current directory. The current work directory defaults to the original packaging directory.

Figure 2 represents the interactions between the CARE archive and the CARETask in OpenMOLE.

The CARETask can be customized to fit the needs of a specific application. For instance, some applications disregarding standards might not return the expected 0 value upon successful completion. The return value of the application is used by OpenMOLE to determine whether the task has been successfully executed, or needs to be re-executed. Setting the boolean flag *errorOnReturnValue* to false will prevent OpenMOLE from re-scheduling a CARETask that has reported a return code different from 0. The return code can be saved in a variable using the *returnValue* setting.

Another default behavior is to print the standard and error outputs of each task in the OpenMOLE console. Such raw prints might not be suitable when a very large number of tasks is involved or that further processing are to be performed on the outputs. A CARETask's standard and error outputs can be assigned to OpenMOLE variables and thus injected in the dataflow by summoning respectively the *stdout* and *stderr* actions on the task.

When packaging an application with CARE, we make sure of excluding any input data from the archived files. CARE allows this with the option *-p*. Data can later be reinjected in the archive from OpenMOLE using the *inputFiles* directive. This directive accepts OpenMOLE variables that describe a set of files to be used as parameters. This means that each instance of a CARETask will see a different input data in its archive's filesystem. The task instance's work directory will thus contain the extracted application supplemented by the specific input data files that were previously discarded from the packaging stage. In this configuration, input data are perfectly decoupled from the application and can be manipulated using OpenMOLE's advanced parameter exploration methods, before being injected to the appropriate task.

Files that are not part of the exploration can also be made available within the CARETask's filesystem using either the *hostFiles* or *resources* directives.

Listing 2 demonstrates the elements of the CARETask described in this section.

```
// Declare the variable
val output = Val[String]
val error = Val[String]
val value = Val[Int]
val file = Val[File]

// Any task
val pythonTask =
  CARETask("hello.tgz.bin", "python hello.py
    /data/fileA.txt") set (
    stdout := output,
    stderr := error,
    returnValue := value,
    inputFiles += (file, "myFile$value.txt"),
    hostFiles += ("/home/user/fileA.txt",
      "/data/fileA.txt")
  )
```

Listing 2 | Example of a CARETask using a file from the host injected in the archive.

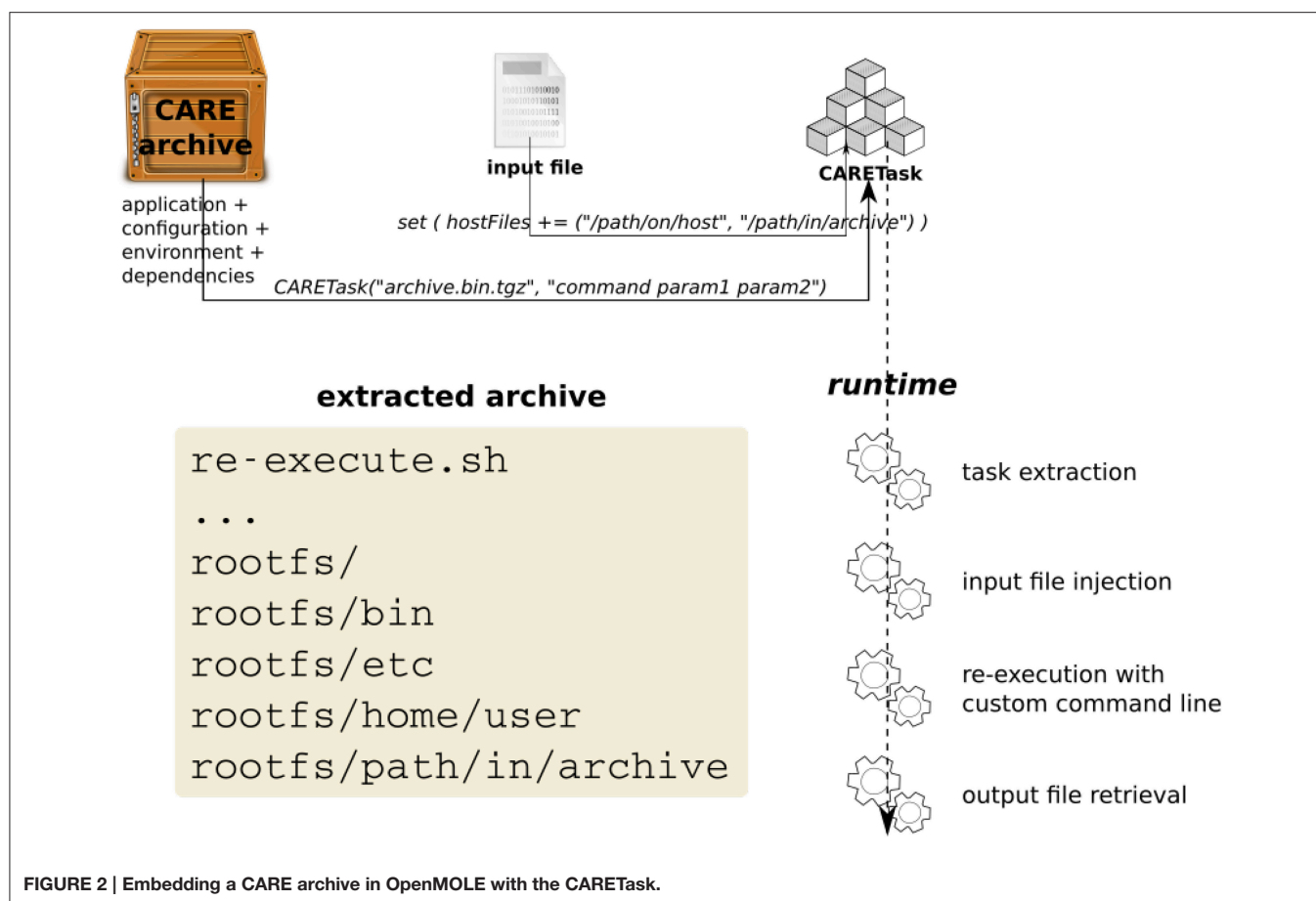
The support of CARE as a first-class citizen in the platform added to existing OpenMOLE features enforces provenance in workflows at two levels. Not only the workflows are defined using a platform agnostic language, but we can now ship standalone archives containing re-executable applications for each stage of the pipeline.

Integrating CARE in OpenMOLE has enhanced the scope of potential applications for CARE, which was initially designed as a tool to create comprehensive bug reports. The development efforts made in OpenMOLE over the past few months have propelled CARE in the range of potential solutions to enable reproducibility in scientific experiments. This integration layer was necessary to bridge the gap between CARE and the scientific community, in order to provide a simple interaction with the end-user.

The next section will show how the CARETask can help explore a canonical dataset on a heterogeneous set of computing infrastructures, and create a reproducible workflow describing the experiment.

4. EVALUATION OF THE REPRODUCIBILITY OF A NEUROIMAGING WORKFLOW

We will evaluate the reproducibility enabled by the CARETask using an fMRI decoder on the Haxby dataset (Haxby et al., 2001). The goal of this experiment is to show that a pipeline intended to run on a local machine and requiring a set of preinstalled dependencies can be re-executed on various distributed computing environments using the CARETask. It validates the choice of the CARE technology to package applications and demonstrates the OpenMOLE integration that enables CARE to be used to reproduce scientific experiments.



4.1. Parameter Space Exploration of a Classifier

This experiment is based on a tutorial¹⁵ for the NiLearn package (Abraham et al., 2014). The example compares different classifiers on a visual object recognition decoding task using the Haxby dataset (Haxby et al., 2001).

The Haxby dataset consists in the fMRI activity recorded for 6 subjects exposed to various stimuli from different categories. The example evaluates the performance of different parameter initialization of a logistic regression classifier to predict the category the subject is seeing from the fMRI activity. Significant prediction shows that the signal in the region contains information about the corresponding category.

We have slightly modified the online example to focus on well-known classifier: the logistic regression. In the NiLearn tutorial, two input parameters vary for this algorithm. The same parameter ranges are tested for this classifier as detailed in Table 2. In order to obtain comparable results, we have set the seed of the pseudorandom number generator used in the logistic regression to 0.

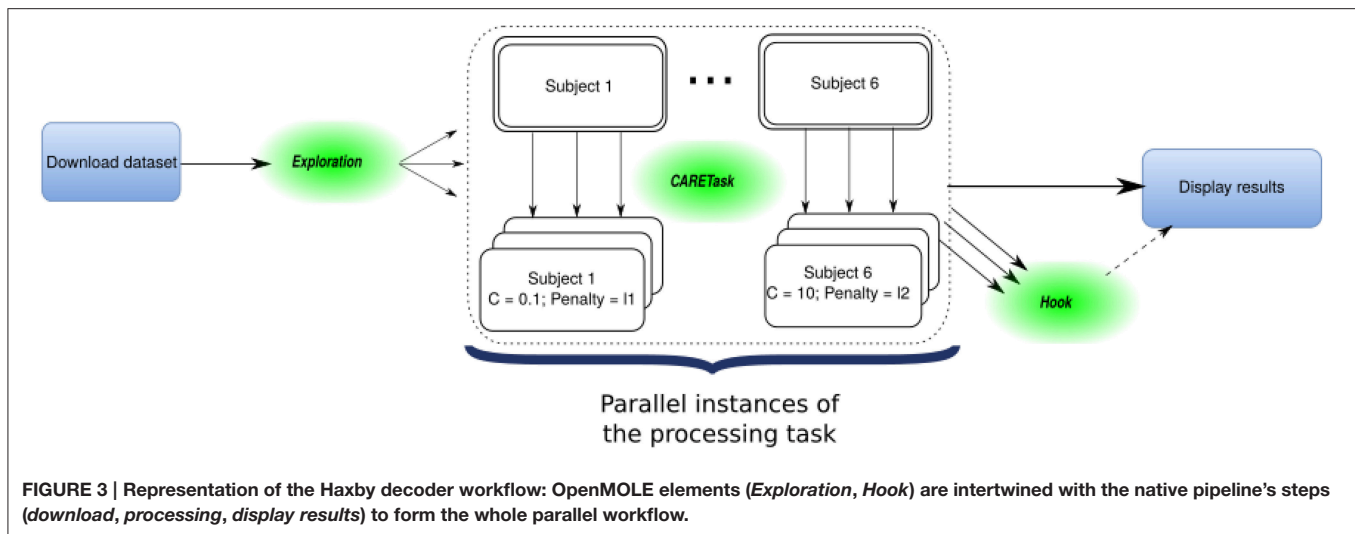
The OpenMOLE workflow for this experiment is made of multiple tasks running both locally and on remote execution

TABLE 2 | Parameters and their values for the Logistic Regression classifier.

Parameter	Range	Description
C	{0.1; 0.5; 1; 5; 10; 50; 100}	Inverse of regularization strength
Penalty	{11; 12}	Norm used in the penalization
Seed	0	Seed initializing the Pseudorandom number generator

nodes as depicted in Figure 3. The initial task asks NiLearn to download the whole dataset from an online repository. An `ExplorationTask` then determines the parameter space that will be explored in parallel by OpenMOLE. The processing task takes a specific tuple of initialization parameters for the logistic regression from the exploration, along with a single subject as in the original example. Each instance of the processing task computes a leave-one-out cross-validated score for the logistic regression classifier initialized with the given input parameters. Result files are retrieved using the OpenMOLE hook mechanism from the remote execution node. They contain a serialized data structure with the results of the processing task stored in Python's pickle format. The collected results are aggregated on the host machine and plotted locally in a separate PNG file per subject.

¹⁵https://nilearn.github.io/auto_examples/02_decoding/plot_haxby_different_estimators.html, last accessed on 12-nov-16.



Input and result files are automatically transferred and passed to the next task, regardless of their format by OpenMOLE's internal mechanisms.

4.2. Testing the Reproducibility

The experiment aims at testing the reproducibility of the whole workflow on each of the platforms described in **Table 3**. The workflow is considered successfully reproduced when generating the exact same result from one machine to another. This for two reasons:

- The seed of the PseudoRandom Number Generator (PRNG) was set to the same value (0) for each instance of the parameter exploration and across the execution environments. This disables any stochastic variability in the results;
- The floating precision reported in the original version of the tutorial is low enough (two digits) so that the underlying hardware does not impact the final results.

The ensemble of Python scripts taken from the NiLearn tutorial to form the workflow steps were packaged as a single CARE archive on the host labeled *Personal machine* in **Table 3**. There is no need to know about the packaged tool in details, or to manually track its software dependencies. Only the input and output data (results) locations must be known so that they can be excluded from the archive. Input data and results are dynamically injected and extracted at runtime from and to the OpenMOLE dataflow. This perfectly fits OpenMOLE's definition of a workflow as a set of connected black boxes only communicating with the external world through their inputs and outputs.

The archive embeds the following Python packages installed in a virtual environment along with their own binary dependencies:

- matplotlib (1.5.1)
- nibabel (2.0.2)
- nilearn (0.2.5)
- numpy (1.11.1)
- pip (8.1.2)

- scikit-learn (0.17.1)
- scipy (0.17.1)
- virtualenv (15.0.2)

The only common aspect between the platforms in **Table 3** is that their Operating System (OS) runs Linux as a kernel.

The heterogeneity in Java Runtime Environment (JRE) versions is solved by OpenMOLE shipping with its own JRE (OpenJDK 1.8.0) to execute on remote machines. It has been built against a 2.6.32 Linux kernel in order to ensure it re-executes successfully on the largest possible range of Linux platforms.

The execution time is only reported here as a marker of successful re-execution on the given platform. Multiple parameters can explain the variability from one environment to another, the most obvious being the different availability of the required resources.

Table 4 reports the prediction scores resulting from running the pipeline on the first subject of the dataset. The prediction scores obtained are similar to those obtained in the tutorial for equivalent parameters (ex: $C = 11$, $p = 50$), down to the second decimal.

An even more interesting aspect of this technique is that we obtained identical results from one environment to another, across all the platforms described in **Table 3**. In order to switch the execution of the processing task from one environment to another, only one line was impacted in the workflow. File transfers are managed by OpenMOLE as well as data injection at the right location of the CARE pseudo file system. This is shown in Listing 3 and is further detailed in specific case studies in Sections 5.1 and 5.2.

```
val processing = CARETask(workDirectory /
  "haxby_example.tgz.bin",
  s"""python processing.py ${dataFolder} $$subjectID
    $$C $$penalty""")
) set (
  (inputs, outputs) += (subjectID, C, penalty),
  inputFiles += (dataFolder),
  outputFiles += ("classifiers_scores.pkl", resultFile)
)
```

```
)
val slurm = SLURMEnvironment("jpassera",
    "predict5.doc.ic.ac.uk")
val pbs = PBSEnvironment("jpassera",
    "login.cxl.hpc.ic.ac.uk")

createDirs — download — exploData —<
    (processing on slurm hook pickleHook) >—
    exploResults —< plot
```

Listing 3 | Data injection and environment switching in the Haxby workflow.

This experiment demonstrated OpenMOLE's ability to efficiently delegate the workload of a real-world pipeline to an heterogeneous set of computing environments. Coupling CARE and OpenMOLE in the CARETask enables experiments to be designed on a personal machine using any toolkit or programming language. Experiments can then be distributed to remote environments regardless of the availability of the tools they depend on, or the ability of the user to install new software components on the target environment (as illustrated by the *Administrator Privileges* column in **Table 3**).

On a side note, this experiment has shown that the genericity of the OpenMOLE platform was not a barrier to exploit field-specific tools in a workflow, NiLearn in this case. By focusing on providing a high-level workflow formalism and simplifying the access to HPC environments, this experiment has shown OpenMOLE was flexible enough to address the needs of the neuroimaging community while respecting their popular software ecosystem.

Finally, this experiment has highlighted the role the CARETask could play in producing reproducible results and re-executable pipelines. Section 5 will now feature the CARETask in combination with the DSL and various computing environments throughout three real-world examples of neuroimaging pipelines.

5. CASE STUDIES

The source code and required material for the three case studies is not part of the OpenMOLE market place¹⁶ due to license restrictions induced by some of the binary dependencies. It is however available in its own repository¹⁷ and contains entries presented as they would be on the original market place. For the sake of clarity, this section will only highlight the parts relevant with the use case.

5.1. Multiple Environments in the Same workflow

The first workflow preprocesses the input data as necessary for a brain parcellation algorithm. Brain parcellation is an essential task for the construction of brain connectivity networks, which has the potential to provide new insights into the brain's organization. Brain parcellation aims at regrouping brain regions that have similar connectivity profiles to the rest of the brain, so as to construct connectivity networks of tractable dimension for subsequent analysis.

¹⁶<https://github.com/openmole/openmole-market>.

¹⁷<https://github.com/openmole/frontiers2016>.

The method proposed in Parisot et al. (2015) uses diffusion Magnetic Resonance Imaging (dMRI) data and structural connectivity to drive the parcellation task. dMRI provides an indirect measurement of the brain's structural connectivity (white matter fiber tracts), by measuring the anisotropy of water molecules in the brain. Several processing steps are required in order to recover the white matter tracts and consequently parcellate the brain from dMRI volumes. In Parisot et al. (2015), the data is processed using FSL's bedpostX and probtrackX (Behrens et al., 2007; Jbabdi et al., 2012), which estimate the fibres orientations at each voxel and perform probabilistic tractography respectively. Both methods are very time consuming. On high quality data such as the HCP database¹⁸, BedpostX takes approximately a week on CPU and 3 h on GPU, while ProbtrackX runs for approximately 30 h. In order to process a large group of subjects for group-wise analysis in a reasonable amount of time, it is necessary to use BedpostX's GPU-enabled version (Hernández et al., 2013) and process the subjects in parallel.

This workflow benefits from OpenMOLE's capacity of delegating different tasks of the pipeline to different computing environments. In this workflow, the first tasks runs a GPU-enabled version of the FSL bedpostX tool (Hernández et al., 2013) while the rest of the workflow is executed on CPU. We thus leverage two distinct computing environments to delegate the workload of this workflow. Listing 4 highlights the section of the workflow description declaring two environments and connecting them with the corresponding tasks.

```
/// Execution environments configuration

// cluster environment with GPU computing facilities
val SLURMgpu =
    SLURMEnvironment(
        "jpassera",
        "predict5.doc.ic.ac.uk",
        queue = "gpu",
        gres = List(Gres("gpu", 1)),
        memory = 15000
    )

// default cluster environment
val SLURMcpu =
    SLURMEnvironment(
        "jpassera",
        "predict5.doc.ic.ac.uk",
        queue = "long",
        memory = 15000
    )

/// Connect the tasks with transitions and run the
    workflow
exploIDsTask —< (bpTask on SLURMgpu) — trajectoryTask
—
exploHemispheresTask —< (ptTask on SLURMcpu)
```

Listing 4 | Multiple environments used by the parcellation preprocessing workflow. The bpTask task requires a GPU to run so it is assigned to the SLURMgpu environment, whereas ptTask can run on traditional CPUs. Both SLURMxxx environments are ubiquitous declinations of the same Slurm cluster, with different requirements.

It is worth noting that the required authentications to connect to the environment do not have to appear in the workflow

¹⁸<https://db.humanconnectome.org>.

TABLE 3 | Different configurations employed in the reproducibility experiment.

Denomination	Resource manager/Scheduler	CPUs	Execution time	Operating system	Linux kernel
<i>Personal machine</i>	None	4 cores	20'36"	Debian 8	4.6.0-1-amd64
<i>Desktop machine</i>	SSH	8 cores	28'14"	Ubuntu 14.04	3.13.0-91-generic
<i>Lab's private cluster</i>	Slurm	312 cores	14'50"	Ubuntu 14.04	3.13.0-63-generic
<i>College wide cluster</i>	PBS	13,558 cores	48'25"	Red Hat Enterprise Linux Server release 6.7	2.6.32-573.12.1.el6.x86_64
<i>European Grid Infrastructure (EGI)</i>	EMI/gLite	650,000 cores	27'15"	CentOS 6/Scientific Linux	2.6.32-642.6.2.el6.x86_64

Denomination	File system	Python version	Java runtime environment	Administrator privileges
<i>Personal machine</i>	Permanent	2.7.12	OpenJDK 1.8.0_91	Yes
<i>Desktop machine</i>	Shared, permanent	2.7.6	OpenJDK 1.7.0_101	Yes
<i>Lab's private cluster</i>	Shared, permanent	2.7.6	OpenJDK 1.7.0_101	No
<i>College wide cluster</i>	Temporary	2.6.6	OpenJDK 1.7.0_101	No
<i>European Grid Infrastructure (EGI)</i>	Shared, temporary	2.7.8	OpenJDK 1.6.0_40	No

description, but are specified once and for all to the platform. Authentications are from then on encrypted and stored in the user's preferences folder.

It is valid in the OpenMOLE syntax for the same remote host to appear in different environment blocks. This ubiquity in environments enables specifying different settings for the same computing host, for example different memory requirements, or devices in the present case. This feature goes along with the ability of each task to run on a separate environment to increase the finer parallelism granularity in the workflow.

Environments are only associated with the tasks at the final stage of the workflow description when tasks are also interconnected. The workflow could be shared without the environments and remain syntactically correct. Users familiar with other computing environments can simply replace the environment declaration by the one of their choice, all in a single location.

5.2. Sharing a Pipeline with the Community

The second workflow in this study segments a collection of developing brain images using the Draw-EM software. Draw-EM¹⁹ (Developing brain Region Annotation With Expectation-Maximization) is an open-source software for neonatal segmentation based on the algorithm proposed in Makropoulos et al. (2014). The algorithm performs atlas-based segmentation to divide the neonatal brain MRI into 87 regions. The different parts of the workflow are:

- Data pre-processing. The original MRI is brain-extracted to remove non-brain tissue and corrected for intensity inhomogeneity.
- Initial tissue segmentation. A spatio-temporal tissue atlas is registered to the brain MRI. The MRI is segmented into the different tissue types with an Expectation-Maximization scheme that combines an intensity model of the image with the tissue priors of the atlas.

- Structural atlas registration. Structural atlases (20 in total) are registered to the subject MRI with a multi-channel registration technique. The original intensity image and the GM probability map are used as different channels of the registration.
- Structure priors computation. The prior probability maps of the different structures are computed based on the local similarity of the transformed atlases with the input MRI.
- Label segmentation. The MRI is segmented into the different structures with a consequent Expectation-Maximization scheme.
- Post-processing. The segmented labels are merged in different granularities to further produce the final tissue segmentations and different hemispheres of the brain. Temporary files used for the computations are removed.

The software is used in collaboration between two teams, and potentially more when data from the developing HCP get publicly released. This workflow is a good example of common use cases evoked in introduction to this work. Here we are faced with two problems when we want to share the pipeline with collaborators: making the description portable from one system to another, and ensuring that the applications that form each stage can be re-executed on another environment.

A first excerpt from this workflow in Listing 5 shows how OpenMOLE interacts with CSV files to explore a fixed parameter space. The notion of samplings in OpenMOLE is flexible enough to traverse a parameter space described in a CSV file or using the more complex methods listed in Section 2.3.

```
val subjectID = Val[String]
val age = Val[Int]

val explo = ExplorationTask(
  CSVSampling(workDirectory/"ages.csv") set (
    columns += subjectID,
    columns += age,
    separator := ' '
  )
)
```

Listing 5 | CSV file exploration using samplings.

¹⁹<https://github.com/MIRTK/DrawEM>.

TABLE 4 | Average prediction scores out of 12 leave-one-out cross validations (\pm standard deviation) for subject 1 of the Haxby dataset.

Penalty	C	Bottle	Cat	Chair	Face	House	Scissors	Scrambledpix	Shoe
/1	000.10	0.175096102728 (± 0.231049939141)	0.379731749159 (± 0.276586523275)	0.158451539359 (± 0.164116793889)	0.604156173217 (± 0.230013422495)	0.848084821382 (± 0.193749012866)	0.233506609807 (± 0.244159557883)	0.66680287676 (± 0.147995188894)	0.375028443778 (± 0.269241140013)
		0.451689065765 (± 0.150073309605)	0.62791249587 (± 0.244023131922)	0.406748456287 (± 0.13915694051)	0.732189944558 (± 0.16331860702)	0.87213622291 (± 0.173805716785)	0.44593597263 (± 0.257571429727)	0.716080161668 (± 0.174958432456)	0.493902257873 (± 0.193535971087)
		0.47399243887 (± 0.148313849961)	0.632733430141 (± 0.22733488965)	0.440552878726 (± 0.137281494554)	0.734600107495 (± 0.133217998162)	0.891882988013 (± 0.133412666949)	0.429409863592 (± 0.28355944954)	0.703740403044 (± 0.175397314428)	0.487136011563 (± 0.18609138877)
		0.471410676788 (± 0.164574051317)	0.619848767217 (± 0.219134957488)	0.445594322982 (± 0.0886794828788)	0.73255493045 (± 0.118718190623)	0.888374216083 (± 0.134171749036)	0.432704249094 (± 0.268164483186)	0.69471103372 (± 0.203315889402)	0.482796210813 (± 0.189916530226)
	010.00	0.466838744958 (± 0.166579054815)	0.638928250112 (± 0.219737279322)	0.444088450235 (± 0.0842821140037)	0.732606643443 (± 0.122777408669)	0.892686605904 (± 0.118406687757)	0.405431521822 (± 0.281777460975)	0.71285195265 (± 0.210434007693)	0.497537258804 (± 0.192609613885)
		0.489227398669 (± 0.173950833724)	0.63862354636 (± 0.182743896393)	0.455303030303 (± 0.120853053014)	0.688123601676 (± 0.109180148231)	0.857546710256 (± 0.117807127625)	0.416753246753 (± 0.263269896373)	0.764532755937 (± 0.201940516096)	0.486474730818 (± 0.18853099241)
		0.478975007701 (± 0.188926437971)	0.673136147956 (± 0.164701994218)	0.478630692661 (± 0.156000687925)	0.648015275109 (± 0.157277843991)	0.830941774901 (± 0.166878589573)	0.437724466891 (± 0.231837298803)	0.755797787415 (± 0.208210996079)	0.495224735512 (± 0.181335061366)
	100.00	0.419064747547 (± 0.196075499695)	0.529790472655 (± 0.214583690108)	0.540197885259 (± 0.177491061481)	0.524839160021 (± 0.174635484257)	0.607328524302 (± 0.219564887192)	0.503213203538 (± 0.157493925525)	0.775192036147 (± 0.182461202755)	0.511069835451 (± 0.190309164145)
		0.442440703126 (± 0.20440149609)	0.541560090043 (± 0.206602126667)	0.545476902154 (± 0.17485799816)	0.540376138138 (± 0.184370175442)	0.633534946986 (± 0.230112025184)	0.514000952751 (± 0.151190492209)	0.790346387359 (± 0.185784023473)	0.492847276932 (± 0.165133380392)
		0.43321401391 (± 0.196947156928)	0.5356102353 (± 0.203785520359)	0.539036006956 (± 0.168881360193)	0.549934036724 (± 0.188614078434)	0.63394509057 (± 0.223754917472)	0.511795894766 (± 0.147923692585)	0.779956776969 (± 0.181218528886)	0.506150386029 (± 0.168495305593)
		0.437276734917 (± 0.19981805444)	0.539067074353 (± 0.204503694746)	0.536326639695 (± 0.178953853726)	0.561171120546 (± 0.199103541317)	0.639533423003 (± 0.224359546867)	0.51426105273 (± 0.152538374463)	0.772198269399 (± 0.182761066327)	0.505117174666 (± 0.166622660548)
/2	010.00	0.4366835824451 (± 0.200058165728)	0.53423394452 (± 0.204880105531)	0.535061891372 (± 0.177213495836)	0.563763615878 (± 0.199772413453)	0.639533423003 (± 0.224359546867)	0.514457769593 (± 0.154085991243)	0.769480878095 (± 0.182828486106)	0.507298139645 (± 0.166133356219)
		0.438753630834 (± 0.206114028623)	0.542474425035 (± 0.21163564676)	0.531695098097 (± 0.174964145627)	0.561135198439 (± 0.200185974598)	0.644376756606 (± 0.226195394616)	0.495566262135 (± 0.144364998604)	0.769480878095 (± 0.182828486106)	0.504692100888 (± 0.167107109882)
		0.438753630834 (± 0.206114028623)	0.546178279103 (± 0.206459632778)	0.530134975891 (± 0.173661532191)	0.561135198439 (± 0.200185974598)	0.643647391194 (± 0.226021956671)	0.495953500594 (± 0.144393468122)	0.764509986917 (± 0.183050686489)	0.503501624698 (± 0.167085408715)
	005.00								
	000.50								

TABLE 5 | Description of the parameters optimized for the MSM tool.

Parameter	Dimensionality	Range	Description
Lambda	3	[0.00001, 100.0]	Weights the contribution of the regularizer relative to the similarity force.
sigma_in	3	[2; 10]	Sets the input smoothing: this changes the smoothing kernel's standard deviation
Iterations	3	[3; 5]	Controls the number of iterations at each resolution.

A single CARE archive was prepared containing the necessary material for all the tasks of the original pipeline (available from Draw-EM's repository²⁰). We have noticed that generating one archive per task generally leads to a large amount of duplicated binaries and shared libraries from one archive to another. When the different tasks of a pipeline share the same dependencies, it is thus more efficient to gather all of them in a unique archive. This strategy leverages OpenMOLE's file replication mechanisms better and reduces the amount of data transferred to remote environments.

The generated CARE archive is then integrated using CARETasks in the OpenMOLE workflow, and fed with input data files stored on the host machine. The command used in the original pipeline is reused as is to build the CARETask and accepts the parameters explored by the sampling in Listing 5. The resulting CARETask is presented in Listing 6.

```
val packagingDirectory =
  "/homes/am411/vol/MIRTK-develop/MIRTK/Packages/DrawEM/
  scripts/v1.1"

val preprocess = CARETask(
  workDirectory/"careArchives/drawem-bundle.tgz.bin",
  packagingDirectory + "/preprocess.sh ${subjectID}
  $age"
) set (
  (inputs, outputs) += (subjectID, age),
  hostFiles += (workDirectory.toString + "/data/T2",
    packagingDirectory + "/T2")
)
```

Listing 6 | The preprocessing CARETask extracted from the Draw-EM pipeline. Input data files are injected from the host system and parameters *subjectID* and *age* taken from the CSV sampling in Listing 5.

As this pipeline is meant to be shared and labeled with a specific version, the fact that CARE archives are not as flexible as Docker turns from a drawback to an advantage as it makes it simpler to ship to the end-user. All the parameterizable parts of the pipeline are handled by the OpenMOLE script, and the pipeline can still be customized by inserting new tasks. Still, any user downloading the OpenMOLE workflow along with the associated CARE archives will be able to reproduce the same experiments that have been performed by the packager, or to reuse the same pipeline for further experiments and comparisons. It is important to note that the data necessary to run the pipeline are not included in the shipped CARE archives.

5.3. Advanced Parameter Tuning Methods

This third workflow performs parameter optimization for cortical surface registration. In this example, cortical surface

alignment is performed using the Multimodal Surface Matching tool (MSM) (Robinson et al., 2013); developed as part of the HCP to enable between subject alignment of multiple different types of cortical surface features (for example functional activations and cortical folding). Registration is optimized to maximize the ratio of feature similarity relative to surface warp distortions.

Here, we study a simplified version of the parameter optimization. The workflow consists in optimizing the value of nine parameters of the MSM tool for a fixed pair of subjects. The parameters explored can be found in Table 5.

In order to find the optimal values for these parameters, we need to compute a fitness function that we will try to minimize using our methods. The fitness function estimates a distortion metric and is computed within its own OpenMOLE task as in Listing 7.

Now, Listing 8 shows how the NSGA-II (Deb et al., 2002) could be initialized to optimize this problem in OpenMOLE.

```
val subjectID = Val[String]

val fitness = CARETask(
  workDirectory/
    "estimate_metric_distortion_anat.tgz.bin",
  "/usr/bin/fsl/MSM/estimate_metric_distortion \
  /home/user/data/${subjectID}/L.white.IC06.native.
  surf.gii \
  /home/user/data/${subjectID}/L.reg.surf.gii \
  /home/user/data/${subjectID}/L.areal_ -abs")
  set (
    inputs += (subjectID),
    stdOut += metric
  )
```

Listing 7 | The result metric is retrieved from the standard output (the command lines have been simplified for the sake of readability).

```
val evolution =
  SteadyStateEvolution(
    algorithm =
      NSGA2(
        // Define the population size: 100
        mu = 100,
        // Define the inputs and their respective
        variation bounds.
        genome = Seq(
          Sequence(lambdas, 0.00001, 100.0, size=3),
          Sequence(sigmaIn_opt, 2.0, 10.0, size=3),
          Sequence(iterations_opt, 3.0, 50.0, size=3),
        ),
        // Define the objectives to minimize.
        objectives = Seq(metric)
      ),
    // Define the fitness evaluation
    // Define the parallelism level
    // Terminate after 1000 evaluations
    evaluation = fitness ,
```

²⁰<https://github.com/MIRTK/DrawEM/blob/c98022a5b78ee99bef5d329fc23f57f9c15b1a5f/pipelines/neonatal-pipeline-v1.1.sh>.

```

parallelism = 10,
termination = 1000
)

...

(evolution on env)

```

Listing 8 | Initialization of the NSGA-II algorithm with the parameters to optimize according to the fitness function from Listing 7. Multi-dimensional parameters are seamlessly handled by the algorithm.

Advanced exploration methods are computationally greedy, but are well suited for parallelization on distributed computing environments. This exploration can also benefit from OpenMOLE's workload delegation by using the `on` keyword seen in Listing 4. This shows that exploration methods fit well in the OpenMOLE ecosystem and can benefit from the other components of the platform, such as the computing environments.

6. CONCLUSION

In this paper, we have shown the ability of the OpenMOLE scientific workflow engine to provide reproducible pipelines that can be shared and distributed on any Linux based environment.

We have seen that the OpenMOLE DSL provided a high-level description of experiments that can be shared and reused by scientists on any platform with a JVM. The newly added `CARETask` offers a solution to ensure Linux-based application can be packaged and re-executed seamlessly on another Linux host without the need to obtain administrator privileges. This criterion was necessary to target HPC environments, a de-facto choice to distribute experiments in the scientific world.

Extensions to the OpenMOLE DSL led to a fine integration of CARE in the framework. Archives only contain binaries and their dependencies, leaving the data to process to be injected in the archive's pseudo-filesystem at runtime from the dataflow. This results in a solution that can be shared from one machine to another, from the description of the pipeline to the applications composing its steps, with the single assumption that it will be re-executed on a Linux host.

Our experiments have reported successful re-executions with the distributed computing environments supported by OpenMOLE. In particular, Section 4 has shown that results obtained from a pipeline with complex software dependencies could be identically reproduced on an heterogeneous set of Linux computing environments.

REFERENCES

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8:14. doi: 10.3389/fninf.2014.00014
- Achterberg, H., Koek, M., and Niessen, W. (2015). "Fastr: a workflow engine for advanced data flows," in *1st MICCAI Workshop on Management and Processing of Images for Population Imaging* (Munich), 39–46.

Medical imaging pipelines were a perfect testbed for our solution, as they are composed of very diverse software tools. A description of case studies inspired from real-world medical imaging solutions has illustrated the suitability of the solution to handle reproducible medical imaging experiments at large scale. Problems such as enabling finer grain parallelism in pipelines, enhancing pipeline sharing with the community, and automatic parameter tuning are three of the concerns that can be encountered by researchers tackling large-scale medical imaging studies. We have addressed these topics through OpenMOLE implementations of three inhouse neuroimaging pipelines. They have showcased various features of the OpenMOLE platform that can help sharing and reproducing pipelines.

OpenMOLE, as well as all the tools forming its ecosystem, are free and open source software distributed under the Affero General Public License version 3 (AGPLv3). This allows anyone to contribute to the main project, or build extensions on top of it.

Future releases of the OpenMOLE platform will strengthen the support of cloud computing environments, with a particular attention given to Amazon EC2. As major datasets become publicly available in the Amazon cloud, moving neuroimaging studies to the cloud is necessary to explore whole datasets. Reproducible OpenMOLE workflows are a valuable addition to the set of tools available to the community in order to set up ambitious experiments.

AUTHOR CONTRIBUTIONS

JP has led this work, drafted the initial version of the manuscript, and is an active contributor to the OpenMOLE project. RR is the leader of the OpenMOLE project and a main developer. ML is a main developer of the OpenMOLE project and has created the graphical user interface. ER, AM, and SP are the original authors of the pipelines presented as case studies. DR has taken part in the inception and conception phases of this work. All authors have revised and agreed on the content of the manuscript.

FUNDING

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 319456.

- Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., and Mock, S. (2004). "Kepler: an extensible system for design and execution of scientific workflows," in *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on* (Santorini: IEEE), 423–424.
- Amstutz, P., Andeer, R., Chapman, B., Chilton, J., Crusoe, M. R., Guimerà, R. V., et al. (2016). *Common Workflow Language, Draft 3*.
- Barker, A., and Van Hemert, J. (2008). "Scientific workflow: a survey and research directions," in *Parallel Processing and Applied*

- Mathematics* (Gdansk: Springer), 746–753. doi: 10.1007/978-3-540-68111-3_78
- Behrens, T., Berg, H. J., Jbabdi, S., Rushworth, M., and Woolrich, M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: what can we gain? *Neuroimage* 34, 144–155. doi: 10.1016/j.neuroimage.2006.09.018
- Bellec, P., Lavoie-Courchesne, S., Dickinson, P., Lerch, J. P., Zijdenbos, A. P., and Evans, A. C. (2012). The pipeline system for Octave and Matlab (PSOM): a lightweight scripting framework and execution engine for scientific workflows. *Front. Neuroinform.* 6:7. doi: 10.3389/fninf.2012.00007
- Boettiger, C. (2014). An introduction to Docker for reproducible research, with examples from the R environment. arXiv preprint arXiv:1410.0846.
- Chamberlain, R., Invenchere, L., and Schommer, J. (2014). *Using Docker to Support Reproducible Research*. Technical Report 1101910, figshare, 2014.
- Chérel, G., Cottineau, C., and Reuillon, R. (2015). Beyond corroboration: strengthening model validation by looking for unexpected patterns. *PLoS ONE* 10:e0138212. doi: 10.1371/journal.pone.0138212
- Chirigati, F., Shasha, D., and Freire, J. (2013). “ReproZip: using provenance to support computational reproducibility,” in *Proceedings of the 5th USENIX conference on Theory and Practice of Provenance (TaPP)*.
- Cottineau, C., Chapron, P., and Reuillon, R. (2015a). Growing models from the bottom up. An evaluation-based incremental modelling method (EBIMM) applied to the simulation of systems of cities. *J. Artif. Soc. Soc. Simulat.* 18:9. doi: 10.18564/jasss.2828
- Cottineau, C., Reuillon, R., Chapron, P., Rey-Coyrehourcq, S., and Pumain, D. (2015b). A modular modelling framework for hypotheses testing in the simulation of urbanisation. *Systems* 3, 348–377. doi: 10.3390/systems3040348
- Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evol. Comput. IEEE Trans.* 6, 182–197. doi: 10.1109/4235.996017
- Deelman, E., Singh, G., Su, M.-H., Blythe, J., Gil, Y., Kesselman, C., et al. (2005). Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Sci. Progr.* 13, 219–237. doi: 10.1155/2005/128026
- Goecks, J., Nekrutenko, A., Taylor, J., and others (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11:R86. doi: 10.1186/gb-2010-11-8-r86
- Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., et al. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front. Neuroinform.* 5:13. doi: 10.3389/fninf.2011.00013
- Guo, P. (2012). CDE: a tool for creating portable experimental software packages. *Comput. Sci. Eng.* 14, 32–35. doi: 10.1109/MCSE.2012.36
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736
- Hernández, M., Guerrero, G. D., Cecilia, J. M., García, J. M., Inuggi, A., Jbabdi, S., et al. (2013). Accelerating fibre orientation estimation from diffusion weighted magnetic resonance imaging using GPUs. *PLoS ONE* 8:e61892. doi: 10.1371/journal.pone.0061892
- Janin, Y., Vincent, C., and Durauffort, R. (2014). “CARE, the comprehensive archiver for reproducible execution,” in *Proceedings of the 1st ACM SIGPLAN Workshop on Reproducible Research Methodologies and New Publication Models in Computer Engineering* (Edinburgh: ACM), 1.
- Jbabdi, S., Sotiropoulos, S. N., Savio, A. M., Graña, M., and Behrens, T. E. J. (2012). Model-based analysis of multishell diffusion MR data for tractography: how to get over fitting problems. *Magn. Reson. Med.* 68, 1846–1855. doi: 10.1002/mrm.24204
- MacKenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., and Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage* 42, 178–195. doi: 10.1016/j.neuroimage.2008.04.186
- Makropoulos, A., Gousias, I., Ledig, C., Aljabar, P., Serag, A., Hajnal, J., et al. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Trans. Med. Imaging* 33, 1818–1831. doi: 10.1109/TMI.2014.2322280
- Mikut, R., Dickmeis, T., Driever, W., Geurts, P., Hamprecht, F. A., Kausler, B. X., et al. (2013). Automated processing of Zebrafish imaging data: a survey. *Zebrafish* 10, 401–421. doi: 10.1089/zeb.2013.0886
- Miles, S., Groth, P., Branco, M., and Moreau, L. (2007). The requirements of using provenance in e-science experiments. *J. Grid Comput.* 5, 1–25. doi: 10.1007/s10723-006-9055-3
- Odersky, M., Altherr, P., Cremet, V., Emir, B., Maneth, S., Micheloud, S., et al. (2004). *An Overview of the Scala Programming Language*. Technical Report IC/2004/64, EPFL Lausanne, Switzerland.
- Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 3045–3054. doi: 10.1093/bioinformatics/bth361
- Parisot, S., Arslan, S., Passerat-Palmbach, J., Wells, W. M. III., and Rueckert, D. (2015). “Tractography-driven groupwise multi-scale parcellation of the cortex,” in *Information Processing in Medical Imaging*, eds S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso (Springer), 600–612.
- Peng, R. D. (2011). Reproducible research in computational science. *Science* 334, 1226–1227. doi: 10.1126/science.1213847
- Reuillon, R., Chuffart, F., Leclaire, M., Faure, T., Dumoulin, N., and Hill, D. (2010). “Declarative task delegation in OpenMOLE,” in *High Performance Computing and Simulation (hpccs), 2010 International Conference on* (Caen: IEEE), 55–62.
- Reuillon, R., Leclaire, M., and Passerat-Palmbach, J. (2015a). “Model Exploration Using OpenMOLE - a workflow engine for large scale distributed design of experiments and parameter tuning,” in *IEEE High Performance Computing and Simulation Conference 2015* (Amsterdam: IEEE), 1–8.
- Reuillon, R., Leclaire, M., and Passerat-Palmbach, J. (2015b). *OpenMOLE Website*.
- Reuillon, R., Leclaire, M., and Rey-Coyrehourcq, S. (2013). OpenMOLE, a workflow engine specifically tailored for the distributed exploration of simulation models. *Future Gen. Comput. Syst.* 29, 1981–1990. doi: 10.1016/j.future.2013.05.003
- Reuillon, R., Schmitt, C., De Aldama, R., and Mouret, J.-B. (2015c). A new method to evaluate simulation models: the calibration profile (CP) algorithm. *J. Artif. Soc. Soc. Simul.* 18:12. doi: 10.18564/jasss.2675
- Rex, D. E., Ma, J. Q., and Toga, A. W. (2003). The LONI pipeline processing environment. *Neuroimage* 19, 1033–1048. doi: 10.1016/S1053-8119(03)00185-X
- Robinson, E. C., Jbabdi, S., Andersson, J., Smith, S., Glasser, M. F., Van Essen, D. C., et al. (2013). “Multimodal surface matching: fast and generalisable cortical registration using discrete optimisation,” in *Information Processing in Medical Imaging* (Asilomar, CA: Springer), 475–486.
- Schmitt, C., Rey-Coyrehourcq, S., Reuillon, R., and Pumain, D. (2015). Half a billion simulations: evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model. arXiv preprint arXiv:1502.06752.
- Stodden, V. (2009). The legal framework for reproducible scientific research: licensing and copyright. *Comput. Sci. Eng.* 11, 35–40. doi: 10.1109/MCSE.2009.19
- Tröger, P., Brobst, R., Gruber, D., Mamonski, M., and Templeton, D. (2012). *Distributed Resource Management Application API Version 2 (DRMAA)*.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Passerat-Palmbach, Reuillon, Leclaire, Makropoulos, Robinson, Parisot and Rueckert. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Exploring fMRI Results Space: 31 Variants of an fMRI Analysis in AFNI, FSL, and SPM

Ruth Pauli^{1*}, Alexander Bowring¹, Richard Reynolds², Gang Chen², Thomas E. Nichols^{1,3} and Camille Maumet¹

¹ Warwick Manufacturing Group, University of Warwick, Coventry, UK, ² Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA, ³ Department of Statistics, University of Warwick, Coventry, UK

Keywords: data sharing, functional MRI, provenance, fMRI analysis, neuroimaging

BACKGROUND

Data sharing is becoming a priority in functional Magnetic Resonance Imaging (fMRI) research, but the lack of a standard format for shared data is an obstacle (Poline et al., 2012; Poldrack and Gorgolewski, 2014). This is especially true for information about data provenance, including auxiliary information such as participant characteristics and task descriptions. The three most commonly used analysis software packages [AFNI¹ (Cox, 1996), FSL² (Jenkinson et al., 2012), and SPM³ (Penny et al., 2011)] broadly conduct the same analysis, but differ in how fundamental concepts are described, and have a myriad of differences in the pre-processing and modeling steps. The practical consequence is that sharing analyzed data is further complicated by the idiosyncrasies of the particular software used.

The Neuroimaging Data Model [NIDM⁴ (Keator et al., 2013; Maumet et al., 2016)] is an initiative from the International Neuroinformatics Coordinating Facility (INCF⁵) that addresses these practical barriers through the development of a standard format for neuroimaging data. Ultimately, NIDM will provide a standard format that can handle data that has been processed in any of the common software packages. In order to achieve this, the development of NIDM requires publicly available derived data that covers all the major use cases in the main software programs.

The purpose of the current work was to produce a set of results of mass univariate fMRI analyses using the most common software packages: AFNI, FSL, and SPM [which between them cover 80% of published fMRI analyses (Carp, 2012)], utilizing publicly available data from OpenfMRI⁶ (Poldrack et al., 2013). The analyses ('variants') presented in this paper cover the most common options available in each software package at each analysis stage, from different Hemodynamic response function (HRF) basis functions through to group-level tests. The tests are arranged so that readers can compare the closest equivalent variants across software packages. In particular, these tests will be useful for comparing the results from default test settings across software packages.

While this collection of analyses was chosen for their relevance to the NIDM project, it also addresses a gap in the literature where publicly available processed data is concerned. Specifically, while there are published comparisons of different processing pipelines, the data are not publicly available (Carp, 2012) or are for resting state fMRI only (Bellec et al., 2016). Others have shared raw data but lack analysis results (Hanke et al., 2014) or do not include

OPEN ACCESS

Edited by:

David N Kennedy,
University of Massachusetts Medical
School, USA

Reviewed by:

Christine Cong Guo,
QIMR Berghofer Medical Research
Institute, Australia
Jiaojian Wang,
University of Electronic Science
and Technology of China, China

*Correspondence:

Ruth Pauli
r.pauli@warwick.ac.uk

Received: 15 April 2016

Accepted: 22 June 2016

Published: 05 July 2016

Citation:

Pauli R, Bowring A, Reynolds R,
Chen G, Nichols TE and Maumet C
(2016) Exploring fMRI Results Space:
31 Variants of an fMRI Analysis
in AFNI, FSL, and SPM.
Front. Neuroinform. 10:24.
doi: 10.3389/fninf.2016.00024

¹<http://afni.nimh.nih.gov/>

²<http://fsl.fmrib.ox.ac.uk/fsl>

³<http://www.fil.ion.ucl.ac.uk/spm/>

⁴<http://nidm.nidash.org>

⁵<http://www.incf.org>

⁶<https://openfmri.org>

comparisons across multiple software packages [e.g., analyses in the The Human Connectome Project⁷ (Van Essen et al., 2013) are performed with FSL only]. Shared raw data is a useful resource, but we argue that shared processed data is also important, both to provide a basis of cross-software comparisons and to create a benchmark for testing of automated provenance software. The dataset presented in this paper is a contribution toward this omission in the literature.

METHODS

Data Source

Data were downloaded from OpenfMRI's BIDS-compliant ds000011 dataset⁸ between 09/02/2016 and 15/02/2016. A full description of the paradigm is in the original paper (Foerde et al., 2006). The first task was a training exercise in which participants counted high tones in a series of high-pitched and low-pitched tones ('tone counting' condition), and then selected a number that represented the number of high tones (the 'tone counting probe' condition, referred to as 'probe' hereafter). We modeled both the tone counting and probe conditions, using tone counting as the effect of interest ([1 0] contrast with implicit baseline). Single-subject tests were conducted with data from subject 01 only, while group-level tests were run with all 14 subjects. Analyses were conducted in AFNI, SPM12, and FSL.

In AFNI, single-subject variants were conducted using the `uber_subject.py` interface, which generates and runs two scripts: `cmd.ap.sub_001` and `proc.sub_001`. Other variants did not require changing options in the interface, so were run directly from the command line, using a copy of the default `cmd.ap.sub_001` script. Scripts for group-level tests were created manually.

For each of the SPM variants, a `batch.m` file conducting the full analysis (using dependencies across processing steps) was created and run with the Batch Editor GUI.

FSL-specific variants were modeled using FSL's FMRI Expert Analysis Tool⁹, where a `fsf` file for the complete analysis was created using the FEAT GUI.

Pre-defined Settings

In this section, settings held constant over variants (e.g., drift modeling) are described for each of the packages. These pre-defined settings (including pre-processing) were identical for each variant.

Pre-processing

As slice-time information was not available for this study, this step was not considered in the pre-processing.

In AFNI, pre-processing was conducted using the default settings in the AFNI `uber_subject.py` graphical interface. First, the BOLD images were rigidly aligned to the skull-stripped

anatomical T1-weighted image using a negative local correlation cost function. Next, the anatomical image was registered to standard space using the AFNI default 'Colin brain' (TT_N27+tlrc) Talairach space template (Holmes et al., 1998) with an affine transformation and weighted least squares cost function. Head motion correction was performed by rigid body registration of each BOLD volume to the third volume, also using weighted least squares cost function; all three transformations were concatenated to allow a single resampling with cubic interpolation. In addition, volumes presenting an estimated motion greater than 0.3 mm (as estimated at 85% of the distance to the cortical envelope¹⁰) compared to the previous scan were censored from the first level regression. The BOLD images were smoothed with a 4 mm Full Width at Half Maximum (FWHM) Gaussian smoothing kernel, and each voxel was scaled to have a mean value of 100 across the run with values larger than 200 truncated to that value.

In FSL, pre-processing was conducted using the Brain Extraction Tool (BET)¹¹ and the default options of the FEAT GUI. First, the anatomical image was skull-stripped. To correct for motion, each volume from the BOLD images was first registered rigidly to the middle volume using a normalized correlation cost function and linear interpolation (MCFLIRT¹² tool). After 6 mm FWHM spatial smoothing and global scaling to set median brain intensity to 10,000, first level fMRI model fitting took place in the subject space. The mean realigned fMRI data was rigidly registered to the brain extracted anatomical image using a correlation ratio cost function, followed by affine registration of the anatomical to MNI space (as defined by the ICBM MNI 152 non-linear 6th generation template image), also with correlation ratio cost function. For group or second level fMRI modeling, the preceding registration parameters were composed to directly resample first level contrast estimates and their variance into standard space with trilinear interpolation.

In SPM, pre-processing was conducted with the Batch Editor GUI. To correct for motion, a two-step rigid body registration procedure was performed with a least squares cost function; each volume from the BOLD images was first registered to the first volume, and then registered to the mean of the aligned images ('Realign: Estimate & Reslice' function, cubic spline interpolation). The anatomical T1-weighted image was then rigidly registered to the mean BOLD image with a mutual information cost function ('Coregister: Estimate' function, cubic spline interpolation). Segmentation, bias field correction and non-linear registration of the anatomical image to standard space ("unified segmentation") were then conducted ('Segment' function); instead of a simple cost function, this process uses a model incorporating tissue class, bias field and spatial deformations to best fit the T1 image data. The estimated deformation field was then used for warping the realigned BOLD images (cubic spline interpolation) and bias corrected anatomical image to MNI space (as defined by the average image of 549

⁷<http://www.humanconnectome.org>

⁸<https://drive.google.com/folderview?id=0B2JWN60ZLkgMGIUY3B4MXZIZW8&usp=sharing>

⁹<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT>

¹⁰<https://afni.nimh.nih.gov/afni/community/board/read.php?p?1,149511,149513#msg-149513>

¹¹<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/BET>

¹²<http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT>

of the subjects from the IXI dataset¹³, cf. `spm_template.man` for more details) using spline interpolation ('Normalise: Write' function). Finally, the normalized realigned BOLD images were smoothed using a 6 mm FWHM Gaussian smoothing kernel ('Smooth' function). Global scaling (1 value for whole 4D dataset) was used to set the mean brain intensity to target value of 100¹⁴.

Data Analysis

As specified by the tone counting task, for each software package, the subject-level design matrix included at least two regressors ("tone counting" and "probe").

By default AFNI adds nine additional regressors in the design matrix: an intercept, two to model slow signal drifts using a second-order polynomial, and six motion regressors (three rotations, three shifts), resulting in a design matrix with 11 columns.

By default SPM adds a discrete cosine transform basis to the linear model to account for drift. The default cutoff of 128 s with this 208 s acquisition allowed three regressors. With an intercept, the model has six regressor parameters, though the drift basis columns are not displayed to the user.

In FSL, slow signal drifts were removed from the data and modeled with a Gaussian-weighted running line smoother with bandwidth parameter 60 s¹⁵, a reduction from the software default value of 100 s, since this is recommended for event-related

¹³<http://www.brain-development.org/>

¹⁴Due to an over-sized brain mask used for global mean computation, SPM's intracerebral mean tends to be under-estimated, resulting in a scaled brain mean intensity of 200 or more instead of 100. For more see: http://blogs.warwick.ac.uk/nichols/entry/spm_plot_units/

¹⁵This parameter is only approximately the FWHM of the smoother's Gaussian, since FSL uses 2.0 instead of 2.335 in the FWHM-to-sigma conversion. For more see: <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1109&L=FS&P=R30849>

TABLE 1 | Folder names for variants in each software package (columns 3–5), for each variant type (columns 1–2).

	Variant name	AFNI	FSL	SPM
Model	First-level regression	Default afni_default	Default fsl_default	Default spm_default
	Second level: 1 sample <i>t</i> -test with ordinary least squares	afni_group_ols	fsl_group_ols	spm_group_ols
	Second level: 1 sample <i>t</i> -test with weighted least squares	afni_group_wls	fsl_group_wls	spm_group_wls
Hemodynamic response function (HRF)	Gamma difference	afni_hrf_gammadiff	fsl_hrf_gammadiff	Default spm_default
	Gamma	Default afni_default	Default fsl_default	NA
	FIR/TENT Basis Function	afni_hrf_tent	fsl_hrf_fir	spm_hrf_fir
Threshold	voxel-wise uncorrected $p \leq 0.001$	Default* afni_default	Default* fsl_default	Default* spm_default
	voxel-wise uncorrected $t \geq 4$	afni_thr_voxelunc4	NA	spm_thr_voxelunc4
	voxel-wise (peak-wise) FWE $p \leq 0.05$	NA	fsl_thr_voxelfwep05	spm_thr_voxelfwep05
	voxel-wise FDR $p \leq 0.05$	afni_thr_voxelfdrp05	NA	spm_thr_voxelfdrp05
	Cluster-wise uncorrected $k \geq 10$, cluster-defining threshold $p \leq 0.001$	afni_thr_clustunc10	NA	spm_thr_clustunc10
	Cluster-wise FWE $p \leq 0.05$, cluster-defining threshold $p \leq 0.001$	afni_thr_clustfwep05	fsl_thr_clustfwep05	spm_thr_clustfwep05
Contrast type	<i>t</i> -contrast	Default** afni_default	Default** fsl_default	Default** spm_default
	<i>f</i> -contrast	afni_con_f	fsl_con_f	spm_con_f
Cluster connectivity	6-Connected: faces	Default afni_default	NA	NA
	18-Connected: faces and edges	afni_clustconn_18	NA	Default spm_default
	26-Connected: faces, edges, and corners	afni_clustconn_26	Default fsl_default	NA
Hypothesis type	One-tailed test	afni_alt_onesided	Default fsl_default	Default spm_default
	Two-tailed test	Default afni_default	NA	NA

Default tests are marked in bold in the table. *For group-level analyses, the default threshold is cluster-wise $p \leq 0.05$ FWE-corrected. **For analyses using flexible basis functions to model the HRF the default test is an *f*-test with an identity matrix of size equal to the number of basis.

designs¹⁶. The design matrix has only two columns, which are mean centered.

Variants

Users can specify from a range of options at each stage in the analysis processing pipeline. A one-factor-at-a-time design was used to run tests with these different options. In each of the analysis packages, at each processing stage (Column 1, **Table 1**.) a single variant was labeled as a default (Columns 3–5, **Table 1**.). This default variant was usually the same in each software package, except for software-defined defaults, which were left unchanged (e.g., HRF).

The variants are presented below. Apart from the aspect that had been explicitly changed for that variant, all other stages of the processing pipeline were kept the same as the default analysis. Using this method, at least one analysis was conducted for each possible variant at every stage of the processing pipeline in AFNI, SPM, and FSL.

HEMODYNAMIC RESPONSE FUNCTION

Gamma Function

The HRF was modeled using a Gamma function.

Difference of Gamma Functions

The HRF was modeled using the difference of two Gamma functions. This is the default in SPM (SPM's canonical HRF). In FSL, the Double-Gamma HRF option was used (with a phase 0).

Flexible Factorial Basis Functions

The HRF was modeled using a finite impulse response (FIR) basis set or a set of TENT functions. In FSL, three basis FIR functions spread over 15 s were defined. In SPM, 10 basis FIR functions spread over 20 s were defined. In AFNI, eight (respectively, seven) TENT functions were defined with a 0 s start and a duration of 12 s (respectively, 14 s) for the tone counting (respectively, the probe) regressor.

MODEL VARIANTS

First-Level Regression

The default analysis for all software packages was a single-subject *t*-test on the tone counting contrast.

Second Level: 1 Sample *t*-test Estimated with Ordinary Least Squares

A one-sample group *t*-test with ordinary least square estimation was performed on the tone counting contrast over the 14 participants.

1 Sample *t*-test Estimated with Weighted Least Squares

A one-sample group *t*-test with weighted least square estimation was performed on the tone counting contrast over the 14 participants.

THRESHOLD

Voxel-Wise Uncorrected $p \leq 0.001$

Results were thresholded with a voxel-wise threshold of $p \leq 0.001$ uncorrected for multiple comparisons.

Voxel-Wise $t \geq 4$

Results were thresholded with a voxel-wise threshold of $t \geq 4$.

Voxel-Wise (Peak-Wise) FWE $p \leq 0.05$

Results were thresholded with a voxel-wise threshold of family-wise error rate $p \leq 0.05$ with correction for multiple comparisons.

Voxel-Wise FDR $p \leq 0.05$

Results were thresholded with a voxel-wise threshold of false discovery rate $p \leq 0.05$ correction for multiple comparisons.

Cluster-Wise $k \geq 10$

Results were thresholded with a cluster-wise threshold of 10 voxels. Clusters were defined using a cluster-forming threshold of $p \leq 0.001$ uncorrected for multiple comparisons.

Cluster-Wise FWE $p \leq 0.05$

Results were thresholded with a cluster-wise threshold of family-wise error rate $p \leq 0.05$ with correction for multiple comparisons. Clusters were defined using a cluster-forming threshold of $p \leq 0.001$ uncorrected for multiple comparisons.

CONTRAST

t-test

The default analysis for all software packages was a *t*-test on the tone counting contrast.

f-test

An *f*-test on the tone counting contrast was performed.

CLUSTER CONNECTIVITY

6-Connected

Neighboring voxels had faces touching. Under this definition a voxel can have up to six nearest neighbors. This is the default in AFNI.

18-Connected

Neighboring voxels were defined as those with faces or edges touching. Under this definition a voxel can have up to 18 nearest neighbours. This is the default in SPM.

¹⁶http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FEAT/UserGuide#FEAT_Basics

26-Connected

Neighboring voxels had faces, edges, or corners touching. Under this definition a voxel can have up to 26 nearest neighbors. This is the default in FSL.

ALTERNATIVE HYPOTHESIS

One-Tailed Test

A one-tailed test looking at positive effects was performed. This is the default in SPM and FSL.

Two-Tailed Test

A two-tailed test looking at positive and negative effects was performed. This is the default in AFNI.

RESULTS

Figure 1 shows the tone counting group level results from a one-sided test FWE-corrected $p \leq 0.05$ cluster-wise inference with a $p \leq 0.001$ uncorrected cluster forming threshold. Despite differences in smoothing, the unthresholded maps show the same general pattern of activation. Thresholded maps from SPM and FSL (6 mm FWHM smoothing) were most similar, while AFNI (4 mm FWHM smoothing) presented a smaller number of active voxels. Aside from smoothing, an important difference with AFNI is the inclusion of motion regressors in the first level model; this is good statistical practice but can reduce sensitivity if the subject motion is correlated with the regressor of interest.

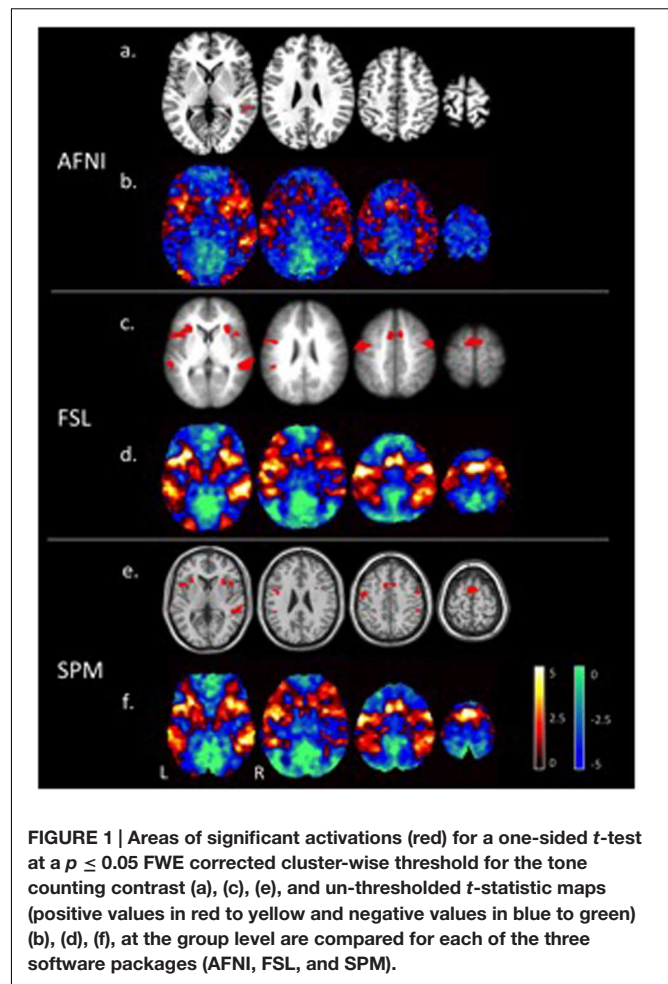
In comparing software packages, there is naturally a tension between exact matching of parameter sets versus use of recommended defaults. The analyses presented here remain faithful to the default settings where possible. Consequently, the differences between packages can be difficult to interpret fully. We plan future work with carefully matched analyses, in order to further elucidate the differences between software packages using many more datasets.

Finally, it is important to emphasize that these tests are not intended to demonstrate the superiority of any particular software package over the others. Each package has its strengths and weaknesses. For example, SPM cannot compute cluster-size inference by FWE p -value, while FSL cannot specify inference by uncorrected cluster threshold. AFNI has immense flexibility in this regard, but the sheer number of options available can make it difficult for the user to judge how to proceed. Ultimately, choice of software will usually be a matter of personal preference for the researcher.

Data Sharing

The dataset is named fMRI Results Comparison Library and can be available at: http://warwick.ac.uk/tenichols/fmri_results. The folder names for each variant are provided in **Table 1**.

For the AFNI variants, each folder contains scripts for running the analysis (cmd.ap.sub_001 and proc.sub_001 for



single-subject tests), scripts for specifying the threshold levels (batch.sh; these are not standard AFNI output, and are included so that future users can run the analysis without manually setting thresholds in the interface), the thresholded dataset (Clust_mask+tlrc) that contains significant clusters, and files that AFNI outputs automatically, saved in the sub_001.results folder.

For the FSL variants, each directory contains the complete FEAT output, which includes the FEAT setup file (design.fsf), motion correction reporting (mc/directory), low-res stats outputs (stats/directory), standard space registration outputs (reg/directory), resampling of stats images into standard space (reg_standard/stats/directory), and time series plots (tsplot/directory). All.html files of the FEAT report are also included.

For the SPM variants, each directory contains a batch.m file to run the analysis, as well as the SPM.mat file containing the design specification. NIFTI files for the regressors, contrasts, and thresholded results are included, and the results report obtained from the analysis has been printed in.pdf format.

Finally, a README.md file is contained in every variant directory, giving a description of the variant and data used in the test.

Recommended Uses

The dataset includes all the necessary scripts and files for future users to replicate the analyses exactly as they were carried out here. This is especially useful for those seeking quick comparisons between different processing options (both within and between software packages). In AFNI, this also removes the need to enter threshold or cluster information manually via the interface. In addition, the dataset and accompanying information in this paper should be useful for novice neuroimagers seeking clear descriptions and examples of basic tests to guide them in their own research.

AUTHOR CONTRIBUTIONS

Analyses were conducted by AB, CM, and RP, with guidance from GC, RR, CM, and TN. The paper was drafted by AB and RP, and written by AB, CM, RP, TN, and RR.

REFERENCES

- Bellec, P., Chu, C., Chouinard-Decorte, F., and Margulies, D. S. (2016). The neuro bureau ADHD-200 preprocessed repository. *bioRxiv*. doi: 10.1101/037044
- Carp, J. (2012). On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments. *Front. Neurosci.* 6:149. doi: 10.3389/fnins.2012.00149
- Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res. Int. J.* 162–73. doi: 10.1006/cbmr.1996.0014
- Foerde, K., Knowlton, B. J., and Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11778–11783. doi: 10.1073/pnas.0602659103
- Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., et al. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. *Sci. Data* 1:140003. doi: 10.1038/sdata.2014.3
- Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., and Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *J. Comput. Assisted Tomogr.* 22, 324–333. doi: 10.1097/00004728-199803000-00032
- Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Keator, D. B., Helmer, K., Steffener, J., Turner, J. A., Van Erp, T. G. M., Gadde, S., et al. (2013). Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage* 82, 647–661. doi: 10.1016/j.neuroimage.2013.05.094
- Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., et al. (2016). NIDM-results: a neuroimaging data model to share brain mapping statistical results. *bioRxiv*. doi: 10.1101/041798

FUNDING

The INCF provided funding for the use of Git Large File Storage (Git-LFS). AB, CM, and TN were supported by the Wellcome Trust. RP was supported by the BBSRC-funded Midlands Integrative Biosciences Training Partnership (MIBTP), and RR and GC were supported by the NIMH and NINDS Intramural Research Programs (ZICMH002888) of the NIH/HHs, USA.

ACKNOWLEDGMENTS

We gratefully acknowledge Karin Foerde, Barbara Knowlton, and Russell Poldrack, who provided their ds000011 data to OpenfMRI. We would like to thank Russell Poldrack for giving advice on re-using the data. We also gratefully acknowledge Krzysztof Gorgolewski, William Triplett, and others at OpenfMRI for their feedback.

- Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Cambridge, MA: Academic press.
- Poldrack, R. A., Barch, D. N., Mitchell, J. P., Wager, T. D., Wagner, A. D., Devlin, J. T., et al. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front. Neuroinform.* 7:12. doi: 10.3389/fninf.2013.00012
- Poldrack, R. A., and Gorgolewski, K. J. (2014). Making big data open: data sharing in neuroimaging. *Nat. Neurosci.* 17, 1510–1517. doi: 10.1038/nn.3818
- Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., et al. (2012). Data sharing in neuroimaging research. *Front. Neuroinform.* 6:9. doi: 10.3389/fninf.2012.00009
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Pauli, Bowring, Reynolds, Chen, Nichols and Maumet. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



SEGMA: An Automatic SEGmentation Approach for Human Brain MRI Using Sliding Window and Random Forests

Ahmed Serag^{1*}, Alastair G. Wilkinson², Emma J. Telford¹, Rozalia Pataky¹, Sarah A. Sparrow¹, Devasuda Anblagan^{1,3}, Gillian Macnaught⁴, Scott I. Semple^{4,5} and James P. Boardman^{1,3}

¹ MRC Centre for Reproductive Health, University of Edinburgh, Edinburgh, UK, ² Department of Radiology, Royal Hospital for Sick Children, Edinburgh, UK, ³ Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK, ⁴ Clinical Research Imaging Centre, University of Edinburgh, Edinburgh, UK, ⁵ Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, UK

OPEN ACCESS

Edited by:

David N. Kennedy,
University of Massachusetts Medical
School, USA

Reviewed by:

Hidetoshi Ikeno,
University of Hyogo, Japan
Frithjof Kruggel,
University of California, Irvine, USA

*Correspondence:

Ahmed Serag
a.f.serag@gmail.com

Received: 31 August 2016

Accepted: 05 January 2017

Published: 20 January 2017

Citation:

Serag A, Wilkinson AG, Telford EJ, Pataky R, Sparrow SA, Anblagan D, Macnaught G, Semple SI and Boardman JP (2017) SEGMA: An Automatic SEGmentation Approach for Human Brain MRI Using Sliding Window and Random Forests. *Front. Neuroinform.* 11:2. doi: 10.3389/fninf.2017.00002

Quantitative volumes from brain magnetic resonance imaging (MRI) acquired across the life course may be useful for investigating long term effects of risk and resilience factors for brain development and healthy aging, and for understanding early life determinants of adult brain structure. Therefore, there is an increasing need for automated segmentation tools that can be applied to images acquired at different life stages. We developed an automatic segmentation method for human brain MRI, where a sliding window approach and a multi-class random forest classifier were applied to high-dimensional feature vectors for accurate segmentation. The method performed well on brain MRI data acquired from 179 individuals, analyzed in three age groups: newborns (38–42 weeks gestational age), children and adolescents (4–17 years) and adults (35–71 years). As the method can learn from partially labeled datasets, it can be used to segment large-scale datasets efficiently. It could also be applied to different populations and imaging modalities across the life course.

Keywords: brain, MRI, large-scale, life-course, sliding window, random forests, classification, tissue segmentation

INTRODUCTION

During early life, the brain undergoes significant morphological and functional changes, the integrity of which determines long-term neurological, cognitive and psychiatric functions (Tamnes et al., 2013). For instance, a wide range of problems including autism spectrum disorder, poor cognitive aging, stroke and neurodegenerative diseases of adulthood may have early life origins (McGurn et al., 2008; Shenkin et al., 2009; Hill et al., 2010; Wardlaw et al., 2011; Stoner et al., 2014). Improved understanding of cerebral structural changes across the life course may be useful for studying early life determinants and atypical trajectories that underlie these common problems.

Quantitative volumes from brain structural magnetic resonance imaging (MRI) acquired at different stages of life offer the possibility of new insight into cerebral phenotypes of disease, biomarkers for evaluating treatment protocols, and improved clinical decision-making and diagnosis. The literature presents a clear distinction between methods developed for different ages partly because the computational task is determined by properties of the acquired data and these

are age-dependent (Cabezas et al., 2011; Despotovic et al., 2015; Išgum et al., 2015). For example, the infant brain presents challenges to automated segmentation algorithms developed for adult brain due to: wide variations in head size and shape in early life, rapid changes in tissue contrast associated with myelination, decreases in brain water, changes in tissue density, and relatively low contrast to noise ratio between gray matter (GM) and white matter (WM). Therefore, automated segmentation tools for modeling structure over years are limited, and this hampers research that would benefit from robust assessment of the newborn to the adult trajectory.

With regard to methodology, approaches for automatic segmentation of brain MRI can be classified into unsupervised (Cai et al., 2007; Leroy et al., 2011; Weglinski and Fabijanska, 2011; Gui et al., 2012) or supervised (Van Leemput et al., 2001; Fischl et al., 2002; Ashburner and Friston, 2005; Prastawa et al., 2005; Song et al., 2007; Altaye et al., 2008; Weisenfeld and Warfield, 2009; Shi et al., 2010; Kuklisova-Murgasova et al., 2011; Makropoulos et al., 2012; Serag et al., 2012b; Cardoso et al., 2013; Chérél et al., 2015; Moeskops et al., 2015; Wang et al., 2015; Loh et al., 2016) approaches. Supervised approaches have proven to be very successful in medical image segmentation (Aljabar et al., 2009; Lötjönen et al., 2010; Coupé et al., 2011; Rousseau et al., 2011; Kaba et al., 2014). However, as they rely on labeled training data (or atlases) to infer the labels of a test scan, most existing supervised approaches require a large number of training datasets to provide a reasonable level of accuracy and they usually carry a high computation cost due to their requirement of non-linear registrations between labeled data and the test scan (Iglesias and Sabuncu, 2015).

To address these challenges, here we describe a method for automatic brain segmentation of MR images, called **SEGMA** (**SEG**mentation **MA**pproach). **SEGMA** differs from current supervised approaches in the following ways. First, **SEGMA** uses a sparsity-based technique for training data selection by selecting training data samples that are “uniformly” distributed in the low-dimensional data space, and hence eliminates the need for target-specific training data (Serag et al., 2016). Second, **SEGMA** uses linear registration to provide an accurate segmentation (mainly to ensure the same orientation and size for all subjects). This is useful because it reduces computation time compared with most supervised methods which require non-linear registrations between the training images and the target image. Finally, **SEGMA** uses a machine learning classification based on random forests (Breiman, 2001) where a class label for a given test voxel is determined based on its high-dimensional feature representation. In addition to incorporating more information into the feature set (compared with methods that use voxel intensity information only), we use a sliding window technique that moves over all positions in the test image and classifies all voxels inside the window at once, instead of assigning labels on a voxel by voxel basis. This technique has the advantage of speeding-up the classification process while minimizing misclassifications compared with methods that use a global classifier (Iglesias et al., 2011; Vovk et al., 2011; Zikic et al., 2014). The feature extraction framework is illustrated in **Figure 1**.

MATERIALS AND METHODS

Data And Image Acquisition

The study includes brain imaging data from 179 subjects, spanning the ages of 0–71 years, from three MRI datasets.

Dataset I

The first dataset contained MR images from 66 infants: 56 preterms (mean post-menstrual age [PMA] at birth 29.23 weeks, range 23.28–34.84 weeks) were acquired at term equivalent age (mean PMA 39.84 weeks, range 38.00–42.71 weeks), and 10 healthy infants born at full term (>37 weeks' PMA). None of the infants had focal parenchymal cystic lesions. Participants of the newborns dataset were recruited to a larger study using MRI to study the effect of preterm birth on brain growth and long-term outcome. Ethical approval was granted by the National Research Ethics Service (South East Scotland Research Ethics Committee) and NHS Research and Development, and informed written parental consent was obtained.

A Siemens Magnetom Verio 3T MRI clinical scanner (Siemens Healthcare GmbH, Erlangen, Germany) and 12-channel phased-array head coil were used to acquire: [1] T1-weighted (T1w) 3D MPRAGE: TR = 1650 ms, TE = 2.43 ms, inversion time = 160 ms, flip angle = 9 degrees, acquisition plane = sagittal, voxel size = $1 \times 1 \times 1 \text{ mm}^3$, FOV = 256 mm, acquired matrix = 256×256 , acceleration factor (iPAT) = 2; [2] T2-weighted (T2w) SPACE STIR: TR = 3800 ms, TE = 194 ms, flip angle = 120 degrees, acquisition plane = sagittal, voxel size = $0.9 \times 0.9 \times 0.9 \text{ mm}^3$, FOV = 220 mm, acquired matrix = 256×218 . The image data used in this manuscript are available from the BRAINS repository (Job et al., 2017) (<http://www.brainsimagebank.ac.uk>).

Reference tissue segmentations for the dataset were generated using an Expectation-Maximization algorithm with tissue priors provided by the atlas from (Serag et al., 2012a,c). Ground truth accuracy of reference neonatal segmentations was evaluated by a radiologist experienced in neonatal brain MRI, who concluded that they were all plausible representations of anatomical classes. Quantitative evaluation of the reference segmentations was performed against manual segmentations from 9 subjects chosen at random. For each subject, three slices (those numbered as 25th percentile, median and 75th percentile of the slices containing brain tissue) were segmented. In order to remove bias toward any particular anatomical plane, three subjects were segmented in the axial plane, three in the coronal plane, and three in the sagittal plane. The quantitative analyses indicated high agreement for all tissues (mean Dice coefficient of 92%).

Dataset II

The second dataset contained T1w MRI scans and corresponding manual expert segmentation of 32 structures from 103 subjects (mean age 11.24 years, range 4.20–16.90 years) publicly available from the Child and Adolescent NeuroDevelopment Initiative (CANDI) at University of Massachusetts Medical School (Frazier et al., 2008; Kennedy et al., 2012) (http://www.nitrc.org/projects/candi_share). The data originates from four diagnostic groups: healthy controls ($N = 29$), schizophrenia spectrum ($N =$

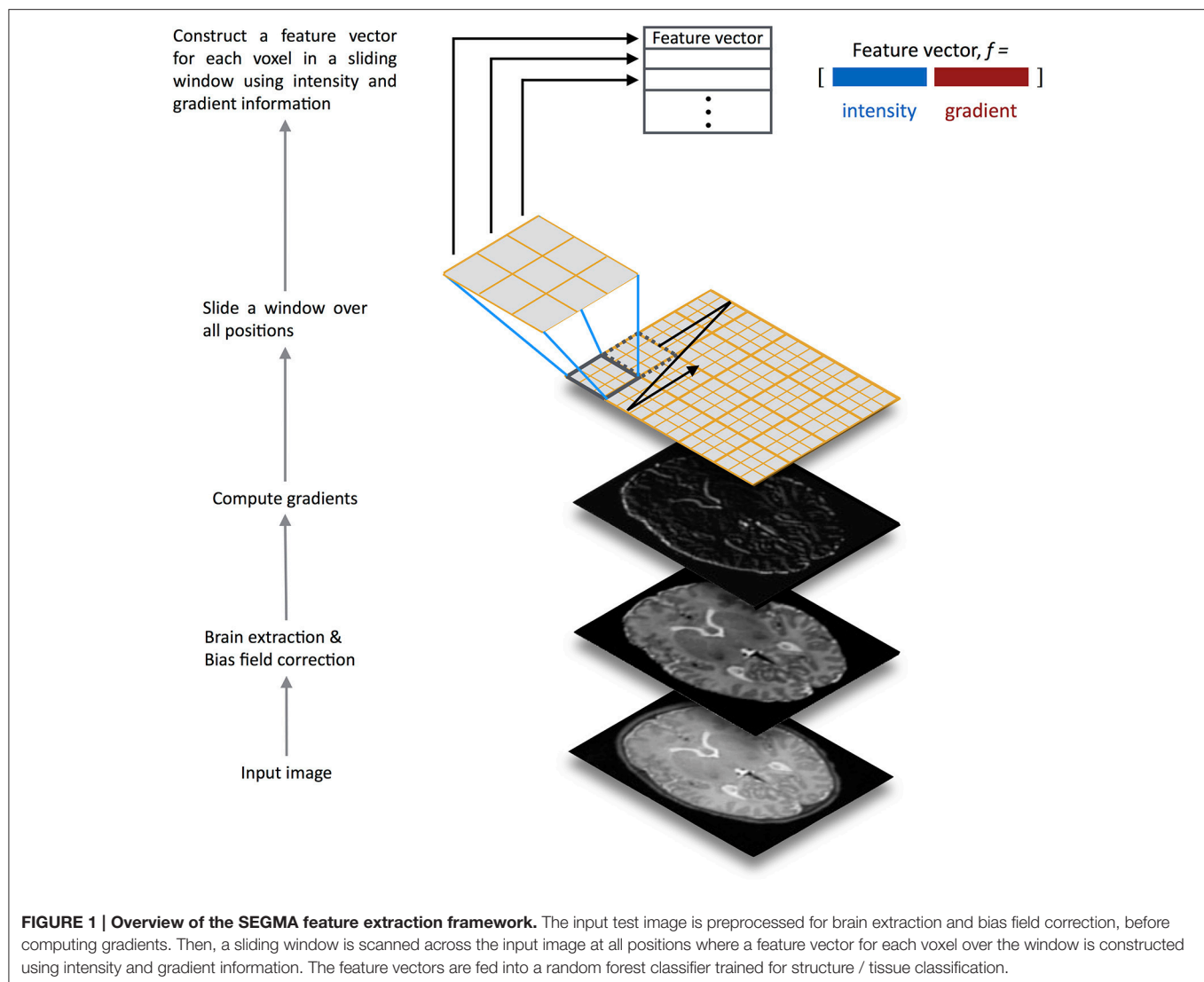


FIGURE 1 | Overview of the SEGMA feature extraction framework. The input test image is preprocessed for brain extraction and bias field correction, before computing gradients. Then, a sliding window is scanned across the input image at all positions where a feature vector for each voxel over the window is constructed using intensity and gradient information. The feature vectors are fed into a random forest classifier trained for structure / tissue classification.

20), Bipolar Disorder ($N = 35$), and Bipolar Disorder with psychosis ($N = 19$). The T1w images were acquired using a 1.5T Signa scanner (GE Medical Systems, Milwaukee, USA) with the following parameters: a three-dimensional inversion recovery-prepared spoiled gradient recalled echo coronal series, number of slices = 124, prep = 300 ms, TE = 1 min, flip angle = 25 degrees, FOV = 240 mm^2 , slice thickness = 1.5 mm, acquisition matrix = 256×192 , number of excitations = 2.

Dataset III

The third dataset contained brain images and the corresponding manual expert segmentation of the whole brain into 32 structures from 18 healthy subjects including both adults and children; for the current study, we used only the adult data ($N = 10$, mean age 38, range 35–71 years). The dataset is publicly available from the Internet Brain Segmentation Repository (www.nitrc.org/projects/ibsr) as IBSR v2.0 (Rohlfing, 2012). The T1w images were acquired using the following parameters: scanner/scan parameters unspecified, acquisition plane = sagittal, number of

slices = 128, FOV = $256 \times 256 \text{ mm}$, voxel size = $0.8\text{--}1.0 \times 0.8\text{--}1.0 \times 1.5 \text{ mm}^3$.

Preprocessing

For brain extraction, we used the brain masks which are provided with each dataset; except dataset I which was brain extracted using ALFA (Serag et al., 2016). All images from all datasets were corrected for intensity inhomogeneity using the N4 method (Tustison et al., 2010).

Training Data

The number of training examples often must be limited due to the costs associated with procuring, preparing and storing the training examples, and the computational costs associated with learning from them (Weiss and Provost, 2003). Therefore, we use in this work a sparsity-based technique to select a number of representative atlas images that capture population variability by determining a subset of n -dimensional samples that are “uniformly” distributed in the low-dimensional data space (Serag

et al., 2016). The technique works by first linearly registering (12 degrees of freedom) all images from each dataset to an appropriate common coordinate space, and image intensities are normalized using the method described by (Nyul and Udupa, 2000). For dataset I, the 40 weeks PMA template from the 4D atlas (Serag et al., 2012a) was used as the common space, which is the closest age-matched template to the mean age of the cohort, while datasets II and III were aligned to the common space defined by the International Consortium for Brain Mapping (ICBM) atlas (Mazziotta et al., 2001). Then, all N aligned images are considered as candidates for the subset of selected atlases. The closest image to the mean of the dataset is included as the first subset image. The consecutive images are selected sequentially, based on the distances to the images already assigned to the subset. Further details can be found in (Serag et al., 2016).

Features

We use machine learning to assign a label to all voxels in the test image, based on training a local classifier. Most existing methods for tissue classification only utilize information from voxel intensity, without considering other information. Here, in addition to voxel intensities, we incorporated various gradient-based features. Typically for each voxel v , a ten-dimensional feature vector \mathbf{f}_v is extracted:

$$\mathbf{f}_v = [I \quad I_x \quad I_y \quad I_z \quad r \quad \theta \quad \phi \quad I_{xx} \quad I_{yy} \quad I_{zz}]^T \quad (1)$$

where I is the gray scale intensity value, I_x , I_y and I_z are the norms of the first order derivatives, and I_{xx} , I_{yy} and I_{zz} are the norms of the second order derivatives. The image derivatives are calculated through the filters $[-1 \ 0 \ 1]^T$ and $[-1 \ 2 \ -1]^T$. The gradient magnitude (r), azimuth angle (θ) and zenith angle (ϕ) are defined as follows:

$$r = \sqrt{I_x^2 + I_y^2 + I_z^2} \quad (2)$$

$$\theta = \tan^{-1} \left(\frac{I_y}{I_x} \right) \quad (3)$$

$$\phi = \cos^{-1} \left(\frac{I_z}{r} \right) \quad (4)$$

where $r \in [0, \infty)$, $\theta \in [0, 2\pi)$, and $\phi \in [0, \pi]$.

Random Forests

In the last decade, random forests (RF) (Breiman, 2001) became a popular ensemble learning algorithm, as they achieve state-of-the-art performance in numerous medical applications (Yi et al., 2009; Huang et al., 2010; Geremia et al., 2011; Mitra et al., 2014; Zikic et al., 2014; Tustison et al., 2015; Pereira et al., 2016). A RF ensemble classifier consists of multiple decision trees. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble. Typically, each tree is trained by combining “bagging” (Breiman, 1996) (where a random selection is made from the examples in the training set) and random selection of a subset of features (Ho, 1998), which construct a collection of decision trees exhibiting controlled variation.

A test sample is pushed down to every decision tree of the random forest. When the sample ends up in one leaf node, the label of the training sample of that node it is assigned to the test sample as tree decision. Then, the final predicted class for a test sample is obtained by combining, in a voting procedure, the predictions of all individual trees. More details on decision forests for computer vision and medical image analysis can be found in Criminisi and Shotton (2013).

Sliding-Window Based Classification

A sliding window is used to move over all possible positions in the test image, and for each window, the voxels inside the window are classified into different tissues or structures. The vector in equation (1) represents the test sample for one voxel in a window, where the number of test samples is equal to the window size w . The training samples come from the voxels of the aligned atlas images that are located at the same location as the voxels belonging to the test window. This means that the number of training samples per window is equal to $k \times w$, where k is the number of training atlases and w is the window size, e.g., $5 \times 5 \times 5$, or $7 \times 7 \times 7$, etc.

A local RF classifier is then used to assign each voxel in the test image to a segmentation class. **Figure 2** shows an example of classifying one test window. The SEGMA algorithm is summarized in **Algorithm 1**.

Algorithm 1. SEGMA algorithm

```

Set  $\mathbf{f}_v$  to represent a feature vector for a voxel  $v$ 
Set  $c_v$  to represent a segmentation class for a voxel  $v$ 
Set  $k$  to represent the number of training data
Set  $w$  to represent the sliding window size
for each window  $W$  do
    Construct the training data matrix  $\mathcal{T}_W^{Train} = \{\mathbf{f}_v | v = 1, \dots, k; v = 1, \dots, w\}$ 
    Train the  $RF_W$  classifier for window  $W$  using  $\mathcal{T}_W^{Train}$ 
    Construct the test data matrix  $\mathcal{T}_W^{Test} = \{\mathbf{f}_v | v = 1, \dots, w\}$ 
    Determine the labels  $c_v$  for all voxels inside the test window  $W$  by applying  $RF_W$  to  $\mathcal{T}_W^{Test}$ 
end

```

Evaluation

A leave-one-out cross-validation procedure was performed for every dataset. Each subject from a dataset in turn was left out as a test sample and the remaining subjects were used as the training data where a subset of k atlases is selected. The comparison between automatic (A) and reference (M) segmentations was performed using the Dice coefficient (DC) (Dice, 1945) which measures the extent of spatial overlap between two binary images, with range 0 (no overlap) to 1 (perfect agreement). The Dice values are expressed as a percentage and obtained using the following equation:

$$DC(A, M) = \frac{2|A \cap M|}{|A| + |M|} \times 100 \quad (5)$$

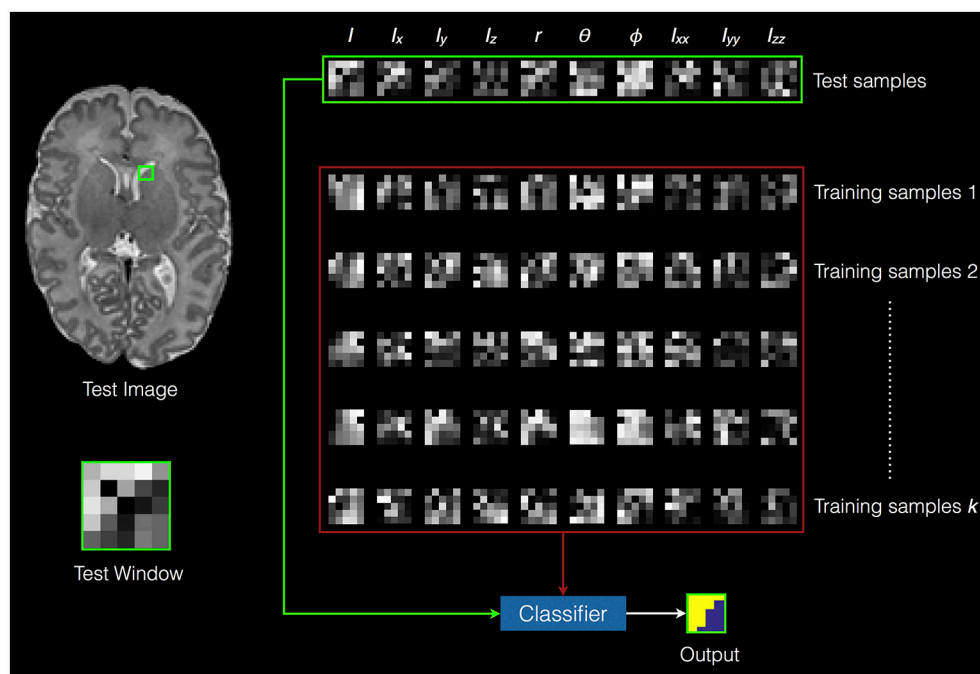


FIGURE 2 | An example of classifying one test window. The green square in the test image represents the test window. The green rectangle represents the extracted features from the test window (i.e., test samples). The red rectangle represents the extracted features from training data (i.e., training samples). The voxels inside the test window are classified into different classes based on training the random forest classifier using the training samples.

Comparison against Other Methods

We compared SEGMA against commonly used segmentation methods: Majority Vote (MV) (Rohlfing et al., 2004; Heckemann et al., 2006), Simultaneous Truth And Performance Level Estimation (STAPLE) (Warfield et al., 2004). The registration scheme for these methods is based on non-linear image deformation (Rueckert et al., 1999; Modat et al., 2010).

To compare SEGMA against other RF segmentation methods, we implemented a global RF classifier, similar to (Iglesias et al., 2011; Zikic et al., 2014), and experimented training it using intensity and gradient-based features, and intensity feature only. Non-linear registration was used as above to map the training images to the test image coordinate space, and the RF classifier was trained using 100,000 randomly sampled voxels from each training image.

Statistical Analyses

To test for differences between segmentation results, *t*-tests were used for normally distributed data, and Mann Whitney U was used to compare non-normal distributions (Shapiro-Wilk normality test was used). $P < 0.05$ were considered significant after controlling for Type I error using false discovery rate (FDR).

RESULTS

To evaluate segmentation performance across the life course, SEGMA was applied to three publicly available datasets that provide MR brain images at different stages of the life course:

neonatal period (38–42 weeks gestational age), childhood and adolescence (4–17 years), and adulthood (35–71 years). **Figure 3** shows examples of brain segmentation results across the life course, and **Figure 4** shows the resulting Dice coefficient (i.e., the agreement between the automatic and reference segmentations).

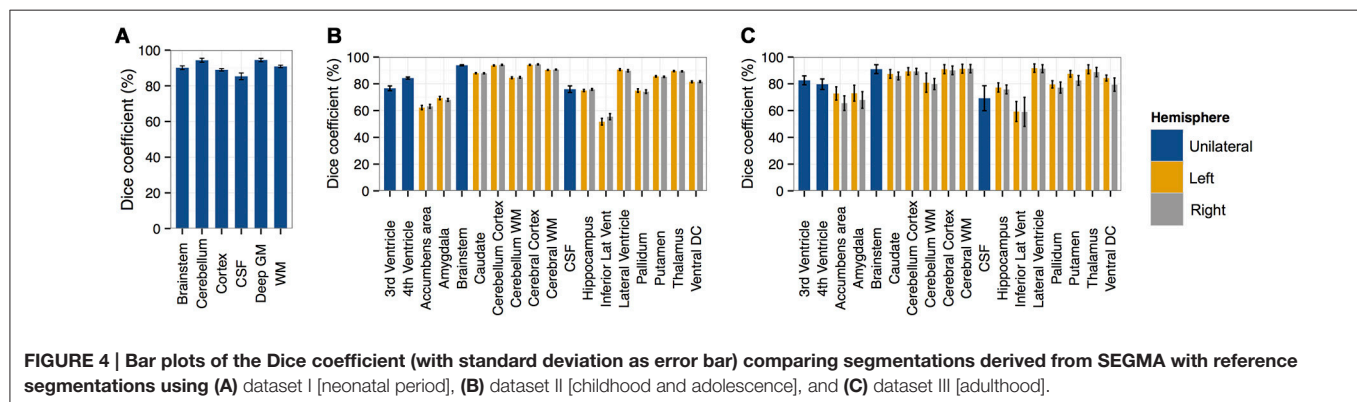
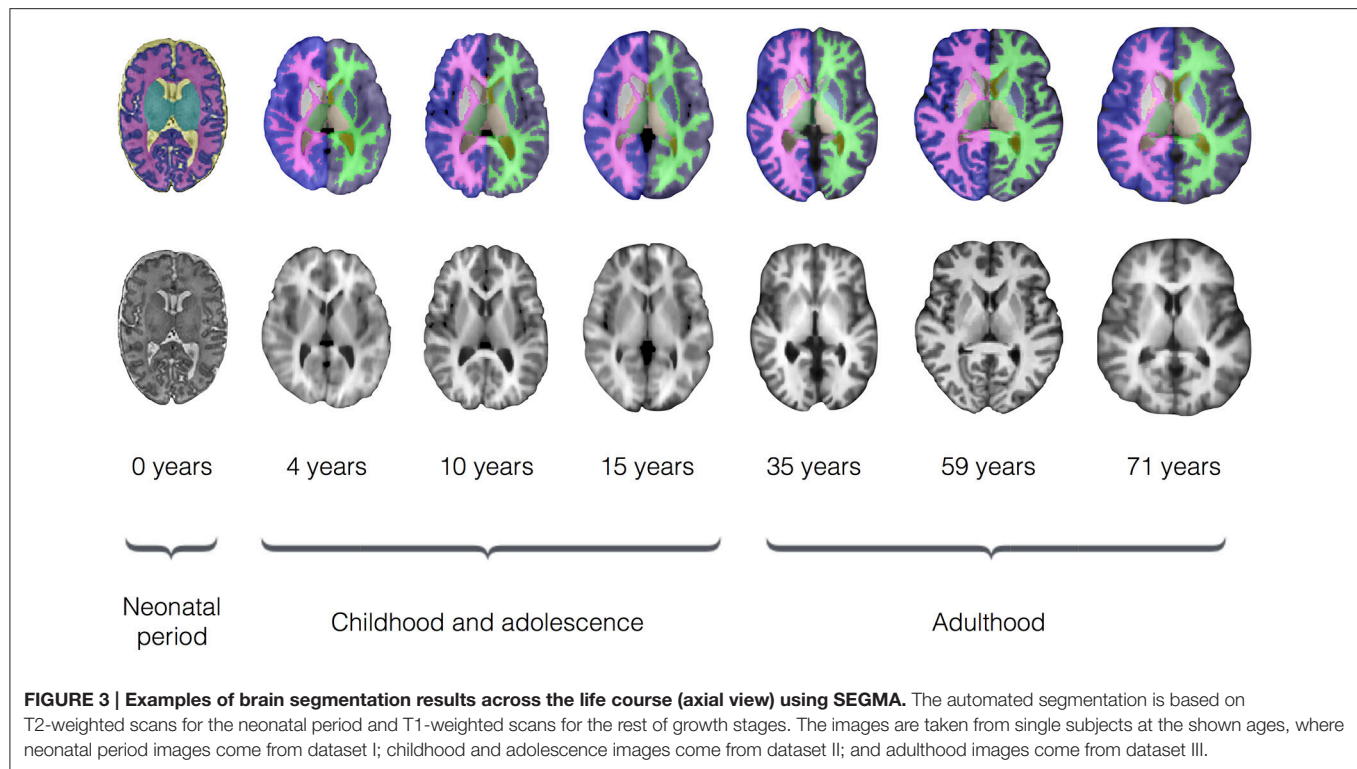
Brain Segmentation in Neonatal Period

We first applied the proposed segmentation method to a neonatal cohort (dataset I) consisting of 66 MR images and associated segmentation of the following tissues / structures: brainstem, cerebellum, cortex or GM, cerebrospinal fluid (CSF), deep GM and WM. Quantitative analyses (**Figure 4**) indicated high accuracy for all tissues and structures with a mean Dice coefficient of 91%.

The highest accuracies obtained for brainstem, cerebellum, deep GM, and WM with mean Dice coefficient of 90–94%, while cortex and CSF had average Dice coefficients of 89 and 85%, respectively.

Brain Segmentation in Childhood and Adolescence

To examine the performance of SEGMA in childhood and adolescence, we used 103 MR images from subjects aged 4–17 years (dataset II) with associated anatomical segmentation of 32 structures. Quantitative analyses (**Figure 4**) indicated high accuracy for all tissues and structures with a mean Dice coefficient of 86%. Nine structures had an average Dice coefficient higher than 90%, 7 structures had an average Dice coefficient



of 79–89%, and 2 structures had an average Dice coefficient of 51–67%.

Brains Segmentation in Adulthood

A dataset (dataset III) consisting of MR images and corresponding anatomical segmentation of 32 structures from 10 subjects (aged 38–71 years) was used to examine the performance of the segmentation algorithm in adulthood. Quantitative analyses (**Figure 4**) indicated high accuracy of 83%. Seven structures had an average Dice coefficient higher than 90%, 9 structures had an average Dice coefficient of 75–89%, and 2 structures had an average Dice coefficient of 49–57%.

Comparison against Other Methods

SEGMA was compared with two commonly used segmentation methods [Majority Vote (MV) (Rohlfing et al., 2004; Heckemann

et al., 2006), Simultaneous Truth And Performance Level Estimation (STAPLE) (Warfield et al., 2004)], and other RF-based segmentation methods. SEGMA improved overall segmentation accuracy compared with MV, STAPLE, global-RF-1 (trained using intensity and gradient features), and global-RF-2 (trained using intensity feature only); **Table 1** shows Dice coefficients averaged over all structures, generated by each segmentation method and applied to datasets I, II and III. ($P < 0.001$; after FDR correction).

Reproducibility

As dataset I (neonatal period) included T1-weighted (T1w) and T2-weighted (T2w) MR imaging, we used it to test the reproducibility of SEGMA across different MR modalities by segmenting the newborn brain using information from T1w and T2w data separately (**Figure 5**). SEGMA provided consistent

TABLE 1 | Dice coefficients averaged over all structures for datasets I, II, and III.

Dataset	SEGMA %	Global-RF-1 %	Global-RF-2 %	MV %	STAPLE %
I	90.68	85.29	84.22	86.97	87.01
II	86.05	78.98	74.90	81.75	79.17
III	82.56	78.75	76.02	77.13	77.54

SEGMA is compared with MV, STAPLE, global-RF-1, and global-RF-2.

segmentation results across different structural MRI modalities of the newborn brain. There was no statistically significant difference between mean Dice scores estimated from the two groups ($P = 0.8977$).

Influence of Parameters

We evaluated the influence of size of training data on segmentation accuracy, and found that increasing the size of the training data improves segmentation accuracy, evidenced by the increase in average Dice coefficient from 88% (7% training data) to 91% (30% training data) for neonates, and from 83% (5% training data) to 86% (20% training data) for children and adolescents. From our experiments, 5–10 training images were sufficient to yield accurate results.

Forest parameters such as tree depth and number of samples per leaf node were set according to previous work (Geremia et al., 2011; Zikic et al., 2014; Wang et al., 2015), and in this work, we only evaluated the influence of number of trees on segmentation accuracy. The number of trees in the forest characterizes the generalization power. As the number of trees becomes large, segmentation accuracy increases, but training time increases and a threshold value is reached after which further improvement is not achieved. In this work, number of trees was set to 10.

With regard to window size, the smaller the window, the longer the classification time. Hence, window size needs to be chosen carefully as it provides a balance between accuracy and speed. Therefore, in this paper, we select the window size as $5 \times 5 \times 5$.

Relative Importance of Features

As partial volume effects in neonatal brain MRI present challenges for automatic segmentation methods, we evaluated the influence of each of the features on segmentation accuracy of the neonatal brain (dataset I). This was done by dropping one or a group of the ten features and running segmentation with the remaining features (features of the same type were dropped together). Therefore, an approximation of relative importance of each feature was obtained. Our experiments show that dropping the intensity feature significantly hinders the segmentation accuracy (**Figure 6A**), whilst the accuracy is improved by incorporating gradient-based features. When all of the features are used, SEGMA yielded higher accuracy than each individual category ($P < 0.001$; after FDR correction). **Figure 6B** also shows an example of the automatic neonatal cortical GM segmentation and how the dropping of each of the ten features affects the segmentation accuracy.

We then analyzed the edge detection for various regions based on using all features (intensity combined with gradients) and gray scale intensity only. **Figure 7** shows that gradient-based features improved edge detection for various regions of the adult and neonatal brain.

Computation Time

One classification task on a 64-bit iMac[®] (Intel[®] Core i7 @ 3.5 GHz \times 4.32 GB RAM) takes 5–7 min. The classification has benefited much from the sliding window strategy used. This is because instead of performing the classification in a voxel-wise manner, this is done for a batch of voxels at once. Assuming a window size of $5 \times 5 \times 5$, the classification time is decreased by 125-folds. In addition, multi-core processing or computer clusters could greatly enhance the speed; and then one brain classification could be performed in about (or less than) 1 min.

DISCUSSION

In this article, we present a new method for MRI brain segmentation (SEGmentation Approach, SEGMA). SEGMA was evaluated on three different datasets (span the ages 0–71 years) that provide different challenges to the brain segmentation task, and accurate results were obtained at all stages of development.

The method is trained using partially labeled datasets where a relatively small number of manually labeled images from the population under study are sufficient to provide accurate results. It is possible that training the method with a larger dataset might increase the segmentation accuracy. However, our goal was to design a methodology that can provide an acceptable, yet high accuracy result using a small number of training images (and hence a low computation cost).

The relatively lower performance for CSF could be caused by its bordering with GM (which is a complex shape). The boundary between GM and CSF is especially difficult to identify inside the sulci, where it is often poorly visible. In addition, the relatively lower performance for the children and adolescence, and adult datasets compared with the neonatal dataset could be attributable to scanner strength. Yet, the results obtained are comparable with those obtained using other methods tested on the same datasets (Rousseau et al., 2011; Zikic et al., 2014).

SEGMA uses a local RF classifier (trained by information from neighboring voxels in the same window) to assign a label to each voxel, which makes it less susceptible to classification errors such as the partial volume misclassification on the CSF-GM and CSF-background boundaries (Kuklisova-Murgasova et al., 2011; Cardoso et al., 2013; Işgum et al., 2015; Moeskops et al., 2015). We chose to use random forests as the classification technique since they naturally handle multi-class classification problems and are accurate and fast (Huang et al., 2010; Geremia et al., 2011; Criminisi and Shotton, 2013). Also, the sliding window plays an important role in significantly speeding up the classification task (compared to voxel-wise approaches).

The method provides an accurate segmentation using only linear registration, which ensures the same orientation and size for all subjects. This is an advantage compared with most supervised methods, which require non-linear registrations

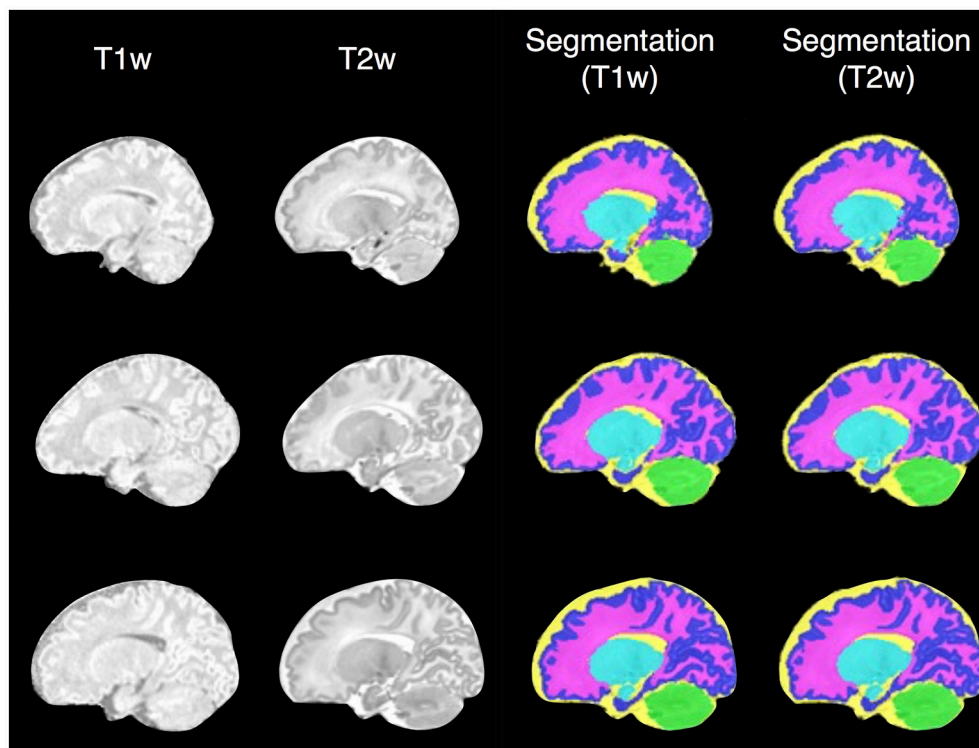


FIGURE 5 | Examples of SEGMA's output segmentation results (sagittal view) using T1-weighted (T1w) and T2-weighted (T2w) MR individually.

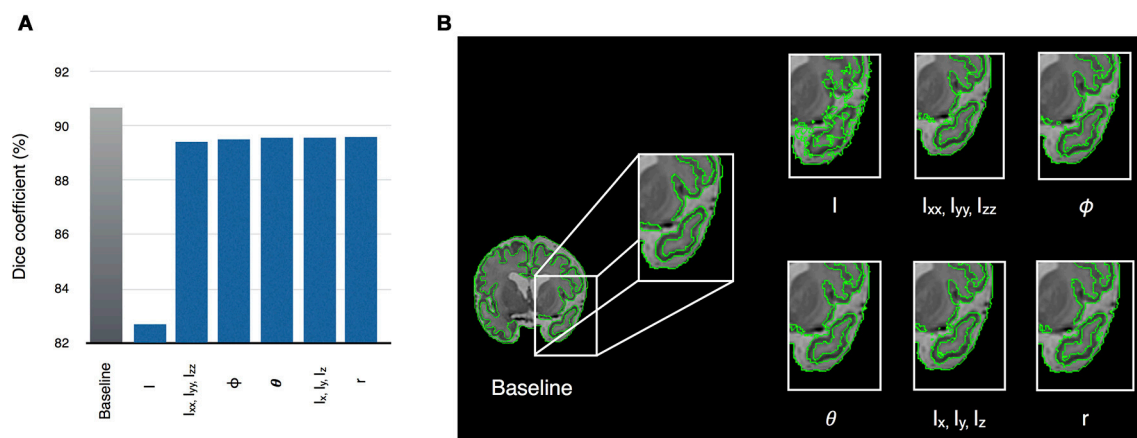


FIGURE 6 | (A) Relative importance of each of the ten features, expressed as the segmentation accuracy, on removing the feature from the feature vector. The leftmost bar shows a baseline value—Dice coefficient, when all features are used. (B) An example of the automatic segmentation of cortical GM (coronal view), which shows how the dropping of each of the ten features affects the segmentation accuracy. The baseline segmentation is obtained by using all features.

between the training images and the test image which increases segmentation time to several hours thereby compromising clinical utility (Iglesias and Sabuncu, 2015). SEGMA also has the advantage of providing an accurate segmentation using a single modality (which is important as the available data might be limited to one modality), and features that characterize object appearance and shape (intensity and gradients). However, the method is flexible and new features can easily be added to the high-dimensional feature vector.

To conclude, we present a method for segmentation of human brain MRI that is robust and provides accurate and consistent results across different age groups and modalities. As SEGMA can learn from partially labeled datasets, it can be used to segment large-scale datasets efficiently. The idea of SEGMA is generic and could be applied to different populations and imaging modalities across the life course. SEGMA is available to the research community at <http://brainsquare.org>.

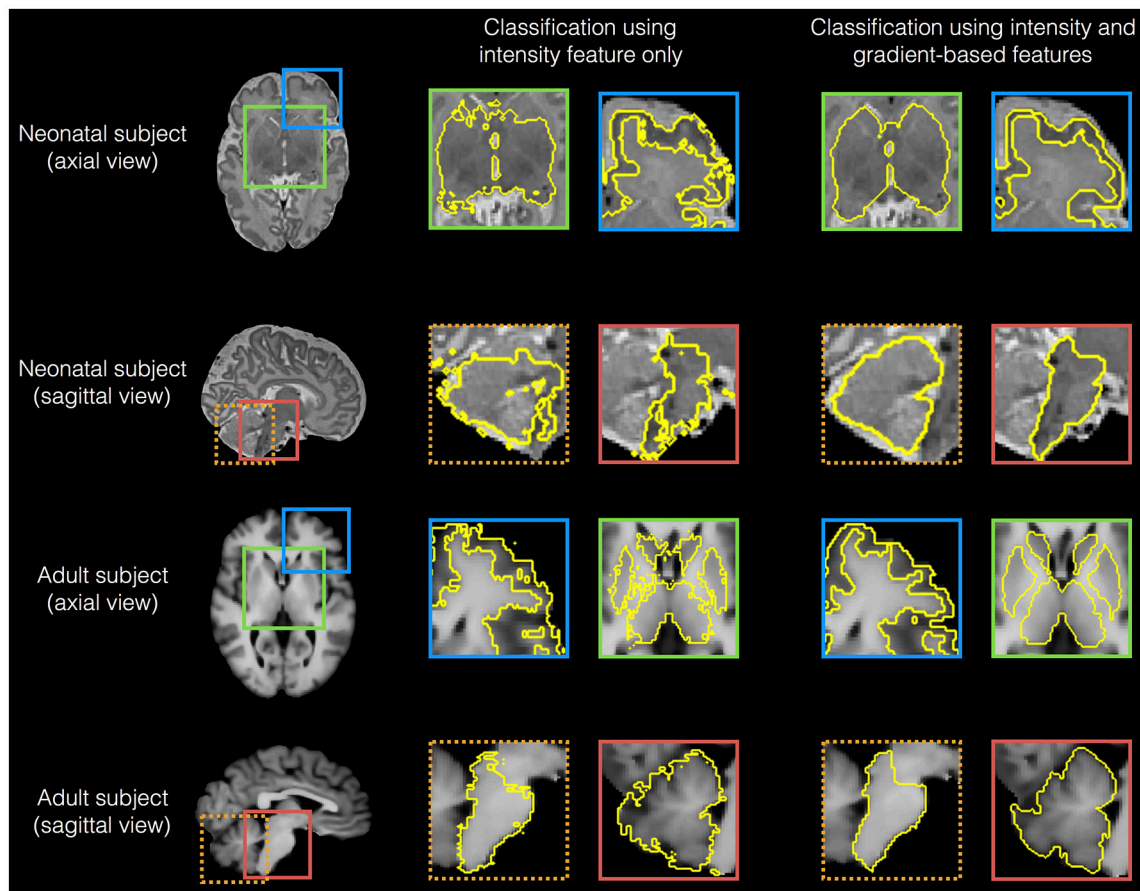


FIGURE 7 | Examples of edge detection for various regions (cortical gray matter, sub-cortical structures, brainstem and cerebellum) based on using all features (intensity combined with gradients) and intensity gray scale only, for a neonatal (dataset I) and an adult brain (dataset III).

AUTHOR CONTRIBUTIONS

AS designed and performed the experiments, and wrote the manuscript; AS, JB, and AGW analyzed output data; ET, RP, and SAS recruited patients; GM and SIS acquired imaging data. All authors approved the final submitted version, and agreed to be accountable for its content.

FUNDING

This work was supported by the Theirworld (<http://www.theirworld.org>), NHS Research Scotland, and NHS Lothian

Research and Development. This work was undertaken in the MRC Centre for Reproductive Health which is funded by the MRC Centre grant MR/N022556/1.

ACKNOWLEDGMENTS

We are grateful to the families who consented to take part in the study and to the nursing and radiography staff at the Clinical Research Imaging Centre, University of Edinburgh (<http://www.cric.ed.ac.uk>) who participated in scanning the infants.

REFERENCES

- Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage* 46, 726–738. doi: 10.1016/j.neuroimage.2009.02.018
- Altaye, M., Holland, S. K., Wilke, M., and Gaser, C. (2008). Infant brain probability templates for MRI segmentation and normalization. *Neuroimage* 43, 721–730. doi: 10.1016/j.neuroimage.2008.07.060
- Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018
- Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655
- Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., and Cuadra, M. B. (2011). A review of atlas-based segmentation for magnetic resonance brain images.

- Comput. Methods Programs Biomed.* 104, e158–e177. doi: 10.1016/j.cmpb.2011.07.015
- Cai, W. L., Chen, S. C., and Zhang, D. Q. (2007). Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recogn.* 40, 825–838. doi: 10.1016/j.patcog.2006.07.011
- Cardoso, M. J., Melbourne, A., Kendall, G. S., Modat, M., Robertson, N. J., Marlow, N., et al. (2013). AdaPT: an adaptive preterm segmentation algorithm for neonatal brain MRI. *Neuroimage* 65, 97–108. doi: 10.1016/j.neuroimage.2012.08.009
- Cherel, M., Budin, F., Prastawa, M., Gerig, G., Lee, K., Buss, C., et al. (2015). Automatic tissue segmentation of neonate brain MR images with subject-specific Atlases. *Proc. SPIE Int. Soc. Opt. Eng.* 9413. doi: 10.1117/12.2082209
- Coupé, P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *Neuroimage* 54, 940–954. doi: 10.1016/j.neuroimage.2010.09.018
- Criminisi, A., and Shotton, J. (2013). *Decision Forests for Computer Vision and Medical Image Analysis*. London; New York, NY: Springer.
- Despotovic, I., Goossens, B., and Philips, W. (2015). MRI segmentation of the human brain: challenges, methods, and applications. *Comput. Math. Methods Med.* 2015:450341. doi: 10.1155/2015/450341
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X
- Frazier, J. A., Hodge, S. M., Breeze, J. L., Giuliano, A. J., Terry, J. E., Moore, C. M., et al. (2008). Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizophrenia. *Schizophr. Bull.* 34, 37–46. doi: 10.1093/schbul/sbm120
- Geremia, E., Clatz, O., Menze, B. H., Konukoglu, E., Criminisi, A., and Ayache, N. (2011). Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *Neuroimage* 57, 378–390. doi: 10.1016/j.neuroimage.2011.03.080
- Gui, L., Lisowski, R., Faundez, T., Hüppi, P. S., Lazeyras, F., and Kocher, M. (2012). Morphology-driven automatic segmentation of MR images of the neonatal brain. *Med. Image Anal.* 16, 1565–1579. doi: 10.1016/j.media.2012.07.006
- Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage* 33, 115–126. doi: 10.1016/j.neuroimage.2006.05.061
- Hill, J., Dierker, D., Neil, J., Inder, T., Knutsen, A., Harwell, J., et al. (2010). A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. *J. Neurosci.* 30, 2268–2276. doi: 10.1523/JNEUROSCI.4682-09.2010
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844. doi: 10.1109/34.709601
- Huang, C., Ding, X., and Fang, C. (2010). “Head pose estimation based on random forests for multiclass classification,” in *20th International Conference on Pattern Recognition (ICPR)* (Istanbul), 934–937.
- Iglesias, J. E., Liu, C. Y., Thompson, P. M., and Tu, Z. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Trans. Med. Imaging* 30, 1617–1634. doi: 10.1109/TMI.2011.2138152
- Iglesias, J. E., and Sabuncu, M. R. (2015). Multi-atlas segmentation of biomedical images: a survey. *Med. Image Anal.* 24, 205–219. doi: 10.1016/j.media.2015.06.012
- Işgum, I., Benders, M. J., Avants, B., Cardoso, M. J., Counsell, S. J., Gomez, E. F., et al. (2015). Evaluation of automatic neonatal brain segmentation algorithms: the NeoBrainS12 challenge. *Med. Image Anal.* 20, 135–151. doi: 10.1016/j.media.2014.11.001
- Job, D. E., Dickie, D. A., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2017). A brain imaging repository of normal structural MRI across the life course: Brain Images of Normal Subjects (BRAIN). *Neuroimage* 144, 299–304. doi: 10.1016/j.neuroimage.2016.01.027
- Kaba, D., Wang, C., Li, Y., Salazar-Gonzalez, A., Liu, X. and Serag, A. (2014). Retinal blood vessels extraction using probabilistic modelling. *Health Inf. Sci. Syst.* 2:2. doi: 10.1186/2047-2501-2-2
- Kennedy, D. N., Haselgrove, C., Hodge, S. M., Rane, P. S., Makris, N., and Frazier, J. A. (2012). CANDIShare: a resource for pediatric neuroimaging data. *Neuroinform* 10, 319–322. doi: 10.1007/s12021-011-9133-y
- Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., et al. (2011). A dynamic 4D probabilistic atlas of the developing brain. *Neuroimage* 54, 2750–2763. doi: 10.1016/j.neuroimage.2010.10.019
- Leroy, F., Mangin, J. F., Rousseau, F., Glasel, H., Hertz-Pannier, L., Dubois, J., et al. (2011). Atlas-free surface reconstruction of the cortical grey-white interface in infants. *PLoS ONE* 6:e27128. doi: 10.1371/journal.pone.0027128
- Loh, W. Y., Connelly, A., Cheong, J. L., Spittle, A. J., Chen, J., Adamson, C., et al. (2016). A new MRI-based pediatric subcortical segmentation technique (PSST). *Neuroinformatics* 14, 69–81. doi: 10.1007/s12021-015-9279-0
- Lötjönen, J. M., Wolz, R., Koikkalainen, J. R., Thurfjell, L., Waldemar, G., Soininen, H., et al. (2010). Fast and robust multi-atlas segmentation of brain magnetic resonance images. *Neuroimage* 49, 2352–2365. doi: 10.1016/j.neuroimage.2009.10.026
- Makropoulos, A., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., Edwards, A. D., et al. (2012). Automatic tissue and structural segmentation of neonatal brain MRI using expectation-maximization. *MICCAI Grand Chall. Neonatal Brain Segmentation* 2012, 9–15. Available online at: <http://neobrain12.isi.uu.nl/pdf/Imperial.pdf>
- Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915
- McGurn, B., Deary, I. J., and Starr, J. M. (2008). Childhood cognitive ability and risk of late-onset Alzheimer and vascular dementia. *Neurology* 71, 1051–1056. doi: 10.1212/01.wnl.0000319692.20283.10
- Mitra, J., Bourgeat, P., Fripp, J., Ghose, S., Rose, S., Salvado, O., et al. (2014). Lesion segmentation from multimodal MRI using random forest following ischemic stroke. *Neuroimage* 98, 324–335. doi: 10.1016/j.neuroimage.2014.04.056
- Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., et al. (2010). Fast free-form deformation using graphics processing units. *Comput. Methods Programs Biomed.* 98, 278–284. doi: 10.1016/j.cmpb.2009.09.002
- Moeskops, P., Benders, M. J., Chit, S. M., Kersbergen, K. J., Groenendaal, F., de Vries, L. S., et al. (2015). Automatic segmentation of MR brain images of preterm infants using supervised classification. *Neuroimage* 118, 628–641. doi: 10.1016/j.neuroimage.2015.06.007
- Nyul, L. G., and Udupa, J. K. (2000). Standardizing the MR image intensity scales: making MR intensities have tissue-specific meaning. *Proc. SPIE Int. Soc. Opt. Eng.* 3976, 496–504. doi: 10.1117/12.383076
- Pereira, S., Pinto, A., Oliveira, J., Mendrik, A. M., Correia, J. H., and Silva, C. A. (2016). Automatic brain tissue segmentation in MR images using random forests and conditional random fields. *J. Neurosci. Meth.* 270, 111–123. doi: 10.1016/j.jneumeth.2016.06.017
- Prastawa, M., Gilmore, J. H., Lin, W., and Gerig, G. (2005). Automatic segmentation of MR images of the developing newborn brain. *Med. Image Anal.* 9, 457–466. doi: 10.1016/j.media.2005.05.007
- Rohlfing, T. (2012). Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging* 31, 153–163. doi: 10.1109/TMI.2011.2163944
- Rohlfing, T., Brandt, R., Menzel, R., and Maurer, C. R. Jr. (2004). Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *Neuroimage* 21, 1428–1442. doi: 10.1016/j.neuroimage.2003.11.010
- Rousseau, F., Habas, P. A., and Studholme, C. (2011). A supervised patch-based approach for human brain labeling. *IEEE Trans. Med. Imaging* 30, 1852–1862. doi: 10.1109/TMI.2011.2156806
- Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., and Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Trans. Med. Imaging* 18, 712–721. doi: 10.1109/42.796284
- Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A., et al. (2012a). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage* 59, 2255–2265. doi: 10.1016/j.neuroimage.2011.09.062

- Serag, A., Blesa, M., Moore, E. J., Pataky, R., Sparrow, S., Wilkinson, A. G., et al. (2016). Accurate Learning with Few Atlases (ALFA): an algorithm for MRI neonatal brain extraction and comparison with 11 publicly available methods. *Sci. Rep.* 6:23470. doi: 10.1038/srep23470
- Serag, A., Gousias, I. S., Makropoulos, A., Aljabar, P., Hajnal, J. V., Boardman, J. P., et al. (2012b). "Unsupervised learning of shape complexity: application to brain development," in *MICCAI Workshop on Spatio-Temporal Image Analysis for Longitudinal and Time-Series Image Data* (Nice).
- Serag, A., Kyriakopoulou, V., Rutherford, M. A., Edwards, A. D., Hajnal, J. V., Aljabar, P., et al. (2012c). A multi-channel 4D probabilistic Atlas of the developing brain: application to fetuses and neonates. *Ann. BMVA* 2012, 1–14. Available online at: <http://www.bmva.org/annals/2012/2012-0003.pdf>
- Shenkin, S. D., Bastin, M. E., Macgillivray, T. J., Deary, I. J., Starr, J. M., and Wardlaw, J. M. (2009). Birth parameters are associated with late-life white matter integrity in community-dwelling older people. *Stroke* 40, 1225–1228. doi: 10.1161/STROKEAHA.108.527259
- Shi, F., Yap, P. T., Fan, Y., Gilmore, J. H., Lin, W., and Shen, D. (2010). Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. *Neuroimage* 51, 684–693. doi: 10.1016/j.neuroimage.2010.02.025
- Song, Z., Awate, S. P., Licht, D. J., and Gee, J. C. (2007). Clinical neonatal brain MRI segmentation using adaptive nonparametric data models and intensity-based Markov priors. *Med. Image Comput. Comput. Assist. Interv.* 4791, 883–890. doi: 10.1007/978-3-540-75757-3_107
- Stoner, R., Chow, M. L., Boyle, M. P., Sunkin, S. M., Mouton, P. R., Roy, S., et al. (2014). Patches of disorganization in the neocortex of children with autism. *N. Engl. J. Med.* 370, 1209–1219. doi: 10.1056/NEJMoa1307491
- Tamnes, C. K., Walhovd, K. B., Dale, A. M., Østby, Y., Grydeland, H., Richardson, G., et al. (2013). Brain development and aging: overlapping and unique patterns of change. *Neuroimage* 68, 63–74. doi: 10.1016/j.neuroimage.2012.11.039
- Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie, Z., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908
- Tustison, N. J., Shrinidhi, K. L., Wintermark, M., Durst, C. R., Kandel, B. M., Gee, J. C., et al. (2015). Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (Simplified) with ANTsR. *Neuroinform* 13, 209–225. doi: 10.1007/s12021-014-9245-2
- Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., and Suetens, P. (2001). Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE Trans. Med. Imaging* 20, 677–688. doi: 10.1109/42.938237
- Vovk, A., Cox, R. W., Stare, J., Suput, D., and Saad, Z. S. (2011). Segmentation priors from local image properties: without using bias field correction, location-based templates, or registration. *Neuroimage* 55, 142–152. doi: 10.1016/j.neuroimage.2010.11.082
- Wang, L., Gao, Y., Shi, F., Li, G., Gilmore, J. H., Lin, W., et al. (2015). LINKS: Learning-based multi-source IntegratioN frameworkK for Segmentation of infant brain images. *Neuroimage* 108, 160–172. doi: 10.1016/j.neuroimage.2014.12.042
- Wardlaw, J. M., Bastin, M. E., Valdés Hernández, M. C., Maniega, S. M., Royle, N. A., Morris, Z., et al. (2011). Brain aging, cognition in youth and old age and vascular disease in the Lothian Birth Cohort 1936: rationale, design and methodology of the imaging protocol. *Int. J. Stroke* 6, 547–559. doi: 10.1111/j.1747-4949.2011.00683.x
- Warfield, S. K., Zou, K. H., and Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* 23, 903–921. doi: 10.1109/TMI.2004.828354
- Weglinski, T., and Fabijanska, A. (2011). "Brain tumor segmentation from MRI data sets using region growing approach," in *Proceedings of VIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH)* (Polyna), 185–188.
- Weisenfeld, N. I., and Warfield, S. K. (2009). Automatic segmentation of newborn brain MRI. *Neuroimage* 47, 564–572. doi: 10.1016/j.neuroimage.2009.04.068
- Weiss, G. M., and Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *J. Artif. Intell. Res.* 19, 315–354. Available online at: <https://www.jair.org/media/1199/live-1199-2209-jair.pdf>
- Yi, Z., Criminisi, A., Shotton, J., and Blake, A. (2009). Discriminative, semantic segmentation of brain tissue in MR Images. *Med. Image Comput. Comput. Assist. Interv.* 5762, 558–565. doi: 10.1007/978-3-642-04271-3_68
- Zikic, D., Glocker, B., and Criminisi, A. (2014). Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Med. Image Anal.* 18, 1262–1273. doi: 10.1016/j.media.2014.06.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Serag, Wilkinson, Telford, Pataky, Sparrow, Anblagan, Macnaught, Semple and Boardman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging

David Alexander Dickie^{1,2*†}, Susan D. Shenkin^{1,3,4†}, Devasuda Anblagan^{1,2,5}, Juyoung Lee⁶, Manuel Blesa Cabeza⁵, David Rodriguez^{1,2}, James P. Boardman⁵, Adam Waldman¹, Dominic E. Job^{1,2} and Joanna M. Wardlaw^{1,2,4*}

¹ Brain Research Imaging Centre, Neuroimaging Sciences, Centre for Clinical Brain Sciences, Royal Infirmary of Edinburgh, The University of Edinburgh, Edinburgh, UK, ² Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) Collaboration, Glasgow, UK, ³ Geriatric Medicine Unit, Royal Infirmary of Edinburgh, The University of Edinburgh, Edinburgh, UK, ⁴ Department of Psychology, Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh, UK, ⁵ MRC Centre for Reproductive Health, Queen's Medical Research Institute, Edinburgh, UK, ⁶ Graduate Training Centre of Neuroscience, International Max Planck Research School, University of Tübingen, Tübingen, Germany

OPEN ACCESS

Edited by:

David N. Kennedy,
University of Massachusetts Medical
School, USA

Reviewed by:

Xi-Nian Zuo,
Chinese Academy of Sciences, China
Frithjof Kruggel,
University of California, Irvine, USA

*Correspondence:

David Alexander Dickie
david.dickie@ed.ac.uk
Joanna M. Wardlaw
joanna.wardlaw@ed.ac.uk

[†]These authors have contributed
equally to this work.

Received: 25 August 2016

Accepted: 04 January 2017

Published: 19 January 2017

Citation:

Dickie DA, Shenkin SD, Anblagan D, Lee J, Blesa Cabeza M, Rodriguez D, Boardman JP, Waldman A, Job DE and Wardlaw JM (2017) Whole Brain Magnetic Resonance Image Atlases: A Systematic Review of Existing Atlases and Caveats for Use in Population Imaging. *Front. Neuroinform.* 11:1. doi: 10.3389/fninf.2017.00001

Brain MRI atlases may be used to characterize brain structural changes across the life course. Atlases have important applications in research, e.g., as registration and segmentation targets to underpin image analysis in population imaging studies, and potentially in future in clinical practice, e.g., as templates for identifying brain structural changes out with normal limits, and increasingly for use in surgical planning. However, there are several caveats and limitations which must be considered before successfully applying brain MRI atlases to research and clinical problems. For example, the influential Talairach and Tournoux atlas was derived from a single fixed cadaveric brain from an elderly female with limited clinical information, yet is the basis of many modern atlases and is often used to report locations of functional activation. We systematically review currently available whole brain structural MRI atlases with particular reference to the implications for population imaging through to emerging clinical practice. We found 66 whole brain structural MRI atlases world-wide. The vast majority were based on T1, T2, and/or proton density (PD) structural sequences, had been derived using parametric statistics (inappropriate for brain volume distributions), had limited supporting clinical or cognitive data, and included few younger (>5 and <18 years) or older (>60 years) subjects. To successfully characterize brain structural features and their changes across different stages of life, we conclude that whole brain structural MRI atlases should include: more subjects at the upper and lower extremes of age; additional structural sequences, including fluid attenuation inversion recovery (FLAIR) and T2* sequences; a range of appropriate statistics, e.g., rank-based or non-parametric; and detailed cognitive and clinical profiles of the included subjects in order to increase the relevance and utility of these atlases.

Keywords: brain mapping, MRI imaging, atlases as topic, brain, systematic review, aging, neurodevelopment, neurodegeneration

INTRODUCTION

Structural magnetic resonance imaging (MRI) brain atlases, frequently also referred to in the literature as templates, are important tools for research and, increasingly, clinical practice. Individual brain scans from several individuals can be combined to form a brain image bank, which can in turn be used to form a brain atlas—an anatomical representation of the brain showing group-wise or study population global or regional brain features.

The terms “brain atlas” and “brain template” have both been used commonly in the literature to date; while they may have different meanings in some situations, many papers do not make this clear but rather appear to use the terms interchangeably. Therefore, for the interests of this paper, we focus on using the term “atlas” but use both terms interchangeably. Atlases are derived by statistically summarizing, e.g., averaging, voxel-wise, regional, or global brain MRI measures from several individuals and they may be used in research as registration targets for functional activation, segmentation, and statistical mapping, for example in analysis of population imaging datasets (Good et al., 2001; Buckner et al., 2004; Avants et al., 2008). In the future, atlases may also be used in clinical practice as reference images to support diagnoses of age-related neurodegenerative disorders (Farrell et al., 2009); therefore their reliability and relevance to the clinical population on which they are being used is paramount.

Brain structure in old age and early life is different to brain structure in younger and middle-aged adults (Gur et al., 1991; Courchesne et al., 2000; Good et al., 2001; Sowell et al., 2003). For example, the developing brain presents specific challenges to atlas construction because of marked variations in head size and shape in early life, maturational processes leading to changes in signal intensity profiles (for example, reducing brain water content and increasing cell density over the perinatal period), relatively lower spatial resolution (cortical patterning at term birth is broadly similar to adult patterns but is approximately one third of the volume at adulthood), and lower contrast between tissue classes (Matsuzawa et al., 2001). In children >5 years, the brain is still developing at an accelerated rate. These issues invalidate the application of adult atlases to data acquired during development, because of misclassification of tissues and structures (Muzik et al., 2000; Yoon et al., 2009), and have led to the development of age-specific atlases for early life studies.

In older age the ventricles, particularly the lateral ventricles, and sulci spaces are generally larger, the gray matter and white matter atrophy in varying proportions, and white matter hyperintensities (WMH) are often present (Lemaitre et al., 2005; Dickie et al., 2015b, 2016b). These and the other many features of brain aging, e.g., lacunes, microbleeds and enlarged perivascular spaces, require specific T2-based sequences, such as fluid attenuated inversion recovery (FLAIR) and T2*, to be captured effectively (Wardlaw et al., 2013). Because of these differences in brain structure, the use of an atlas based on only younger subjects and a limited range of sequences can create a bias in life course population studies, e.g., systematic overexpansion (Buckner et al., 2004) or regional distortion of older brains. Even within restricted age bands brain structure is highly variable due to various factors such as ethnicity, medical

history, e.g., hypertension, smoking and cognition (Farrell et al., 2009; Wardlaw et al., 2011). Therefore, population brain atlases must include information on age, sex, ethnicity, relevant medical history, and cognitive testing to have broad uses and relevance. Further, brain atlases should be derived using statistical methods that effectively characterize the wide and irregular variance in brain structure across the life course (Dickie et al., 2013). Attempts to understand this variation and create brain atlases have increased exponentially with the advent of MR and other non-invasive imaging techniques but the origins of this pursuit extend back many thousands of years.

The gyral and sulcal pattern of the human brain is thought to have been first described in 3000 B.C. by Imhotep, an Egyptian “god” of medicine (Adelman and Smith, 1987). Although study of the structure of the brain continued for more than 4500 years, it was not until 1664 when Thomas Willis published *Cerebri Anatome* (“Anatomy of the Brain”) that robust methods for measuring brain structure started to be developed (O’connor, 2003). Willis directed novel autopsies of the brain in which it was first removed from the skull, in contrast to the traditional *in situ* dissections of the time, and then sliced from the base upwards. The slices were then viewed with a microscope and drawn by Christopher Wren (O’connor, 2003). These 350 year old drawings arguably represent the first attempt to create a brain atlas but more detailed atlases of the brains’ cyto- and myelo-architecture did not emerge until the late nineteenth/early twentieth century (Betz, 1874; Brodmann, 1909, 1994; Von Economo and Koskinas, 1925). Such atlases are useful to understand the distribution of tissue types and fibers, but they have little use in modern clinical practice. One of the first clinically relevant atlases was published by Talairach et al. (1967), who developed a 3D coordinate system to assist deep-brain surgery.

The subsequent Talairach and Tournoux atlas (Talairach and Tournoux, 1988) has become one of the most influential atlases in brain imaging (Evans et al., 2012). This atlas provides a standardized set of coordinates to determine specific sites within the brain. It has been used to describe the site of a biopsy, or to compare data from structural MRI, functional MRI (fMRI), SPECT, and PET studies. However, the Talairach and Tournoux atlas has been described as “woefully inadequate” (Toga and Thompson, 2007). The reasons for this, including that it was derived from a single fixed cadaveric brain from an elderly female with limited clinical information, have been listed by many and well-known since the atlases’ inception (Evans et al., 1993, 2012; Devlin and Poldrack, 2007). Indeed, they were noted in the original author’s foreword, “this method is valid with precision only for the brain under consideration” (Talairach and Tournoux, 1988), but this may not be commonly known amongst users of this and derived atlases, e.g., Montreal Neurological Institute (MNI)152 (Brett et al., 2001). Population brain atlases, many of which were descended from Talairach (Evans et al., 2012), may therefore be lacking in age-appropriate, clinically, and cognitively described subjects that were synthesized via appropriate image analysis and statistical methods. It is for this reason that we undertook the following systematic review to identify, collate, and describe existing structural MRI brain atlases.

In this review, we aim to summarize the currently available structural MRI brain atlases across the life span—published in journals and/or on the internet—for researchers in population based imaging. Following our review we discuss the practical, technical, and statistical considerations that should be borne in mind when using brain image atlases.

MATERIALS AND METHODS

We followed “Preferred reporting items for systematic reviews and meta-analyses (PRISMA)” reporting guidelines (Moher et al., 2009) in preparation of this manuscript. From October 2010 to April 2015, we systematically searched for “normal” brain structural MRI atlases. From April 2015 to August 2016, we supplemented this search with: hand searching of reference sections in previous review articles and records we included here (e.g., Mazziotta et al., 2001; Toga et al., 2006; Evans et al., 2012); periodical searching of Google with a subset of these terms; review of content alerts distributed by relevant journal articles, e.g., *NeuroImage* (<http://www.journals.elsevier.com/neuroimage/>), *Human Brain Mapping* [[http://onlinelibrary.wiley.com/journal/10.1002/\(ISSN\)1097-0193](http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0193)], and *Frontiers in Neuroscience* (<http://journal.frontiersin.org/journal/neuroscience>); and, finally, hand searching of neuroimaging data sharing initiatives *NeuroVault* (<http://neurovault.org/>) and *NITRC* (<http://www.nitrc.org/>). Two authors (DAD and JYL) independently and systematically searched PubMed (including MEDLINE; <http://www.ncbi.nlm.nih.gov/pubmed/>), and the internet using Google (<http://www.google.co.uk/>) and Google Scholar (<http://scholar.google.co.uk/>) with the terms: “Magnetic Resonance Imaging” or “Magnetic Resonance Image” or “Magnetic Resonance Images” or “MRI” or “MR” and “brain” and “template” or “atlas” or “stereotactic” or “stereotaxic” and “human.”

October 2010–August 2016 was the time during which we conducted our search, there were no publication date restrictions on eligibility for inclusion and we included all normal MRI atlases of whole brain structures from across the lifespan. We included atlases with “anatomical” or “structural” sequences and probability maps, e.g., T1-, T2-, T2*-w, FLAIR-weighted images, and gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) probability maps. We did not include atlases solely of segmented regional structures (ROI), such as subcortical GM or individual cortical areas (e.g., Westbury et al., 1999; Ahsan et al., 2007), or histological sections (e.g., Eickhoff et al., 2005), but did include atlases that had whole brain and regional structures. We excluded: (1) non-human brain atlases, e.g., macaque; (2) diffusion or functional MRI connectively atlases without anatomical/structural components, e.g., JHU ICBM-DTI-81 and NTU-90 (Yeh and Tseng, 2011); (3) functional MRI brain atlases only, e.g., <http://www.brainmap.org/>; (4) records that described atlas methods only (e.g., Maldjian et al., 2003; Wilke et al., 2008; Van Leemput, 2009; Chen et al., 2012); and (5) atlases that included patients with known neurological or central nervous system disease, e.g., Alzheimer’s disease (Desikan et al., 2006; Loni, 2011).

We provide information reported in each structural MRI brain atlas on the number, age, and sex of participants; sequences collected; statistical derivation method; and clinical/cognitive data found.

RESULTS

We identified 543 potentially eligible records (**Figure 1**) of which 66 met inclusion criteria. Descriptions of each atlas are provided in **Table 1**.

We found 66 structural brain MRI atlases with a total of 10,354 subjects (median = 43, mean = 157, range = 1–2762), including European, North American, Chinese, Japanese, Korean, Indian, and Malay participants.

We identified 19 fetal, neonate and infant (0–5 years); six childhood (5–18 years); 23 young or middle aged adult (18–60 years); seven older adult (aged >60 years); and six life-course atlases including several age groups. Five atlases did not report the age of included subjects.

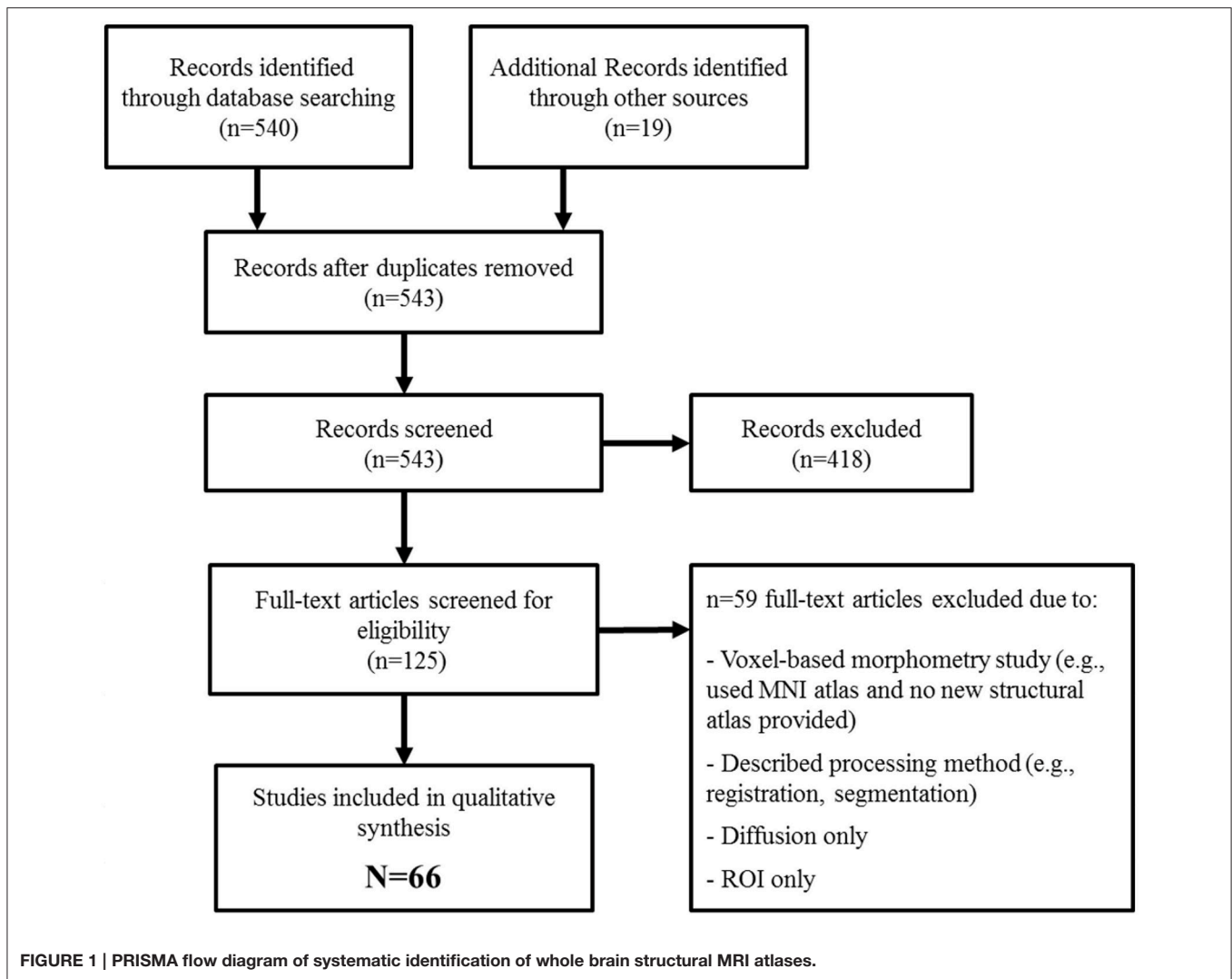
Twenty-seven atlases (41%) reported cognitive/clinical data but this was generally in summary form, e.g., “subjects had no history of neurological, psychiatric or other significant medical illnesses” (Lee et al., 2005) rather than summarized measures from individual subjects. One atlas of the elderly brain reported data on age, handedness, MMSE, education level, and proportion of hypertensive subjects (Lemaitre et al., 2005), but we found no atlas that reported a comprehensive battery of cognitive, medical, and demographic data that are increasingly found in large cohort studies (Wardlaw et al., 2011; Deary et al., 2012).

All atlases were based on T1, T2, and/or PD structural sequences. No atlas included FLAIR or T2* sequences. Almost all multiple subject atlases (except Farrell et al., 2009; Dickie et al., 2015a); were derived using parametric mean-based methods rather than non-parametric percentile ranks or ranges.

Some atlases used the same publicly available databases, e.g., Open Access Series of Imaging Studies (OASIS) data were used in at least two atlases (Dickie et al., 2015a; Richards et al., 2016). We were not able to quantify the subject overlap between atlases as subject identifiers were generally not provided. Ten atlases were based on a single subject. We identified 13 atlases (19.7%) that were developed by or descended from Talairach and Tournoux (labeled “T&T” in **Table 1**).

DISCUSSION

Brain atlases are an important resource for neuroanatomical definition and are often the basis for automated image analyses, which are likely to become increasingly used for population imaging studies. It is important that users are aware of the origins and assumptions underlying these atlases. We identified 66 whole brain structural MRI atlases with a total of 10,354 “normal” subjects from 15 weeks gestational age to 92 years. The number of subjects in each atlas was generally rather small (median = 43; mean = 157; range = 1–2762; $n \geq 100 = 18$; $n \geq 1000 = 3$) given that several hundreds or even thousands of subjects are required to represent population brain structure adequately



(Mazziotta et al., 2001; Toga, 2002; Toga et al., 2006; Evans et al., 2012). Only 622 subjects (6%) had measures of medical, cognitive, and demographic data to support their classification as normal (Lemaitre et al., 2005). Thirteen atlases (~20%) were descended from the Talairach and Tournoux atlas (Talairach and Tournoux, 1988), e.g., MNI, ICBM, and “Brain atlas for healthy elderly.”

Specific populations should be analyzed using an atlas derived from other subjects in that population, or a closely relevant population, otherwise systematic errors may be introduced, e.g., the overexpansion of atrophied brains registered to younger subject atlases (Buckner et al., 2004). Relevant to this, we suggest that the most appropriate atlas for a given study (should there be multiple atlases available with similar demographic, clinical, and cognitive profiles) is the one which requires the least amount of global or regional warping from native subject space to atlas space (and vice-versa). The consequences of various degrees of processing and warping individual subjects to an atlas space have previously been analyzed and discussed (Dickie et al., 2015a). The presence of cognitive deficits and medical conditions, e.g.,

vascular risk factors, also affect brain structure (Ritchie et al., 2015; Dickie et al., 2016b) and therefore it is essential for this information to be measured and tabulated in brain atlases. Although we appreciate that such depths of data may be difficult and expensive to acquire their strong influence on brain structure makes them imperative for understanding the appearance and structure of brain atlases. Medical, cognitive, and demographic data that may be useful in understanding the structure of atlases at different stages of life have been described previously (Job et al., 2016). Given the wide variation and features of brain structure across the life course (Good et al., 2001; Sowell et al., 2003; Allen et al., 2005; Raz et al., 2010), reliable studies, particularly at the extremes of life, require atlases with many more subjects including clinical and cognitive data and additional structural MRI sequences, e.g., T2-based sequences for measuring burden of small vessel disease (Wardlaw et al., 2013).

Such “big-data” approaches including a wide number of imaging sequences and supporting textual information have been successfully applied in studies with limited age ranges such as the “Human Connectome Project” which aims to map structural and

TABLE 1 | Whole brain structural MRI atlases (alphabetical order by name).

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
10–20 sensor placement system structural atlas (Kabdebon et al., 2014)	7.1 weeks	1 (M = 0; F = 1)	<ul style="list-style-type: none">• T1• T2• Tissue maps• ROI	<ul style="list-style-type: none">• Single subject	Not reported
4D dynamic probabilistic atlas of developing brains (Kuklisova-Murgasova et al., 2011)	36.6 ± 4.9 weeks GA	142 (M = 70; F = 72)	<ul style="list-style-type: none">• T2• Tissue maps	<ul style="list-style-type: none">• Voxel-wise weighted intensity averaging	Not reported
83 ROI 2-year old atlas (Gousias et al., 2008)	21.4–34.4 (24.8 ± 2.4) months	33 (M = 17; F = 16)	<ul style="list-style-type: none">• T1• T2• ROI	<ul style="list-style-type: none">• Single subjects	Not reported
A database of age-appropriate average MRI templates (Fillmore et al., 2015; Richards et al., 2016)	2 weeks–89 years*	2762	<ul style="list-style-type: none">• T1• T2• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
A multi-channel 4D probabilistic atlas of the developing fetal brain (Serag et al., 2012b)	29.6 ± 4.6 weeks GA	80	<ul style="list-style-type: none">• T1• T2• Tissue maps	<ul style="list-style-type: none">• Voxel-wise weighted intensity averaging	Not reported
A multi-modal map of human cerebral cortex (Glasser et al., 2016)	22–35 years	210 (M = 80; F = 130)	<ul style="list-style-type: none">• T1• T2• fmMRI• rfMRI	<ul style="list-style-type: none">• Group average parcellation	Not reported
A neonatal atlas template (Kazemi et al., 2007)	39–42 weeks GA	7 (M = 4; F = 3)	<ul style="list-style-type: none">• T1	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain (Habas et al., 2010)	20.57–24.71 weeks GA	20	<ul style="list-style-type: none">• SSFSE T2• Tissue maps• ROI	<ul style="list-style-type: none">• Voxel-wise averaging• Single subjects	Not reported
Adult brain maximum probability map: “Hammers adult atlases” (Hammers et al., 2003)	31.6 ± 9.9 years	30 (M = 15; F = 15)	<ul style="list-style-type: none">• T1• ROI	<ul style="list-style-type: none">• Voxel-wise probabilities	Reported
Age-specific MRI templates for pediatric neuroimaging (Sanchez et al., 2012a)	4.5–24 years*	1289 (M = 636; F = 653)	<ul style="list-style-type: none">• T1• T2/PD	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Allen Human Brain Atlas (Allen Institute for Brain Science, 2009)	24–57 years* (post-mortem)	8 (M = 6; F = 2)	<ul style="list-style-type: none">• T1• T2	<ul style="list-style-type: none">• Single subjects	Not reported
Automatic analysis of cerebral atrophy (Subsol et al., 1997)	37 years (mean)	10 (M = 10; F = 0)	<ul style="list-style-type: none">• T1• Ventricle map	<ul style="list-style-type: none">• Average and SD feature positions	Reported

(Continued)

TABLE 1 | Continued

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
Bayesian interference atlases (Van Leemput, 2009)		18	<ul style="list-style-type: none">• T1• T2• Tissue maps• ROI	<ul style="list-style-type: none">• Bayesian inference averaging	Not reported
Brain atlas for healthy elderly ^{T&T} (Lemaitre et al., 2005)	63–75 years	662 (M = 331; F = 331)	<ul style="list-style-type: none">• T1• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Brain Characterization Using Normalized Quantitative Magnetic Resonance Imaging (Wartjes et al., 2013)	26–67 (45 ± 11) years	31 (M = 14; F = 17)	<ul style="list-style-type: none">• R₁• R₂• PD	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
Brain Imaging of Normal Subjects (BRAINS) age-specific MRI atlases from young adults to the very elderly (Dickie et al., 2016a)	25–92 years*	225	<ul style="list-style-type: none">• T1• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Brain template for children from 2 weeks to 4 years age (Sanchez et al., 2012b)	8 days–4.4 years*	154 (M = 83; F = 71)	<ul style="list-style-type: none">• T1• T2/PD• ROI	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Brainetome atlas (Fan et al., 2016)	22–35 years	49 (M = 17; F = 32)	<ul style="list-style-type: none">• T1• T2• Diffusion• fMRI• ROI	<ul style="list-style-type: none">• Voxel-wise probabilities	Not reported
Cerefy brain atlas ^{T&T} (Nowinski, 2005)	60 years	1 (M = 0; F = 1)	<ul style="list-style-type: none">• Digitised Talairach plates• ROI	<ul style="list-style-type: none">• Single subject	Not reported
Chinese probabilistic atlas (Xing et al., 2013)	18–70 years*	1000	<ul style="list-style-type: none">• T1• T2• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Chinese ₅₆ ^{T&T} (Tang et al., 2010)	24.46 ± 1.81 years	56 (M = 56; F = 0)	<ul style="list-style-type: none">• T1• ROI	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Clinical toolbox ^{T&T} (Rorden et al., 2007)	72.9 ± 7.63years	50 (M = 18; F = 32)	<ul style="list-style-type: none">• T1• Tissue maps• CT	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
Consistent high-definition spatio-temporal atlas of the developing brain (Serag et al., 2012a)	28–44 (37.3 ± 4.8) weeks PMA	204	<ul style="list-style-type: none">• T1• T2	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
Construction of multi-region-multi-reference atlases (Shi et al., 2010)	1.3 ± 0.7 months	68 (M = 38; F = 30)	<ul style="list-style-type: none">• T2• Tissue maps• ROI	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported

(Continued)

TABLE 1 | Continued

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
Contributions to 3D Diffeomorphic Atlas Estimation: Application to Brain Images (Bossa et al., 2007)		19	• T1	• Voxel-wise averaging and SD	Not reported
Cortical gray matter of young adults (Luders et al., 2005)	25 ± 4 years	60 (M = 30; F = 30)	• T1 • Tissue maps • ROI	• Average and SD gyral locations	Not reported
Deformable Spatiotemporal MRI Atlas of the Fetal Brain (Gholipour et al., 2014)	26.14–35.86 (30.50 ± 3.05) weeks GA	40	• SSFSE	• Voxel-wise averaging	Not reported
Digital Pediatric Brain Structure Atlas (Shan et al., 2006)	9 years	1 (M = 0; F = 1)	• T1 • ROI	• Single subject	Reported
EvePM (Lim et al., 2013)	33 years	1 (M = 0; F = 1)	• T1 • Diffusion • ROI • susceptibility	• Single subject	Not reported
FreeSurfer “Destrieux” cortical atlas (Destrieux et al., 2010)	18–33 years	12 (M = 6; F = 6)	• T1 • ROI	• Vertex-wise probabilities	Not reported
Group-specific brain tissue probability map (Yoon et al., 2005)	26.07 ± 5.32 years	59 (M = 36; F = 23)	• T1 • Tissue maps • ROI	• Voxel-wise averaging	Reported
Harvard brain atlas (Shenton et al., 1995)	25 years	1 (M = 1; F = 0)	• T1 • ROI	• Single subject	Reported
Harvard-Oxford cortical and subcortical structural (Fmrib, 2008)	18–50 years	37 (M = 21; F = 16)	• T1 • ROI	• Voxel-wise probabilities	Not reported
Human cortical development map (Gogtay et al., 2004)	13.0 ± 4.8 years*	13 (M = 6; F = 7)	• T1 • GM map • ROI	• Average gyral locations	Reported
IOBM452 T&T (Lancaster et al., 2007)	20–40 years (27.8 ± 5.1) years	452	• T1 • T2 • Tissue maps • ROI	• Voxel-wise averaging	Not reported
Infant brain atlas (Altaie et al., 2009)	9–15 months	76 (M = 31; F = 45)	• T1 • Tissue maps	• Voxel-wise averaging	Not reported
Japanese pediatric standard brain (Uchiyama et al., 2013)	6–9 years	45 (M = 22; F = 23)	• T1	• Voxel-wise averaging	Reported

(Continued)

TABLE 1 | Continued

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
JHU-neonatal brain atlas (Oishi et al., 2011)	0–4 days	25 (M = 15; F = 10)	<ul style="list-style-type: none">• T1• T2• Diffusion	<ul style="list-style-type: none">• Voxel-wise averaging• Single subject	Not reported
Korean standard brain template (Lee et al., 2005)	18–77 (44.6 ± 19.4) years*	78 (M = 49; F = 29)	<ul style="list-style-type: none">• T1• F-18-FDG PET	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
LPBA40 ^{T&T} (Shattuck et al., 2008)	19–39 (29 ± 6) years	40 (M = 20; F = 20)	<ul style="list-style-type: none">• T1• Tissue maps• ROI	<ul style="list-style-type: none">• Voxel-wise averaging• Voxel-wise probabilities	Reported
Merged young- and old-adult atlas target: “Washington 711” ^{T&T} (Buckner et al., 2004)	49 years	24 (M = 9; F = 15)	<ul style="list-style-type: none">• T1	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Mindboggle-101 (Klein and Tourville, 2012)	19–61 years	101 (M = 57; F = 44)	T1 ROI	<ul style="list-style-type: none">• Single subjects	Not reported
MNI/ICBM 152 ^{T&T} (Mazziotta et al., 2001)	18–44 (24 ± 7) years	152 (M = 86; F = 66)	<ul style="list-style-type: none">• T1• T2/PD• Tissue maps• ROI	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
MNI 305 ^{T&T} (Evans et al., 1993)	23.4 ± 4.1 years	305 (M = 239; F = 66)	<ul style="list-style-type: none">• T1• Brain masks	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
MNI Pediatric atlases ^{T&T} (Fonov et al., 2011)	0–18.5 years*	324	<ul style="list-style-type: none">• T1• T2/PD• Tissue maps• Brain masks	<ul style="list-style-type: none">• Voxel-wise averaging and SD	Not reported
MNI-Colin27 ^{T&T} (Holmes et al., 1998; Aubert-Broche et al., 2006)		1 (M = 1; F = 0)	<ul style="list-style-type: none">• T1• T2/PD• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging (of repeated single subject scans)	Not reported
Neonatal brain atlas: “ALBERT” (Gousias et al., 2012)	39–45 (41) weeks PMA	5 (M = 3; F = 2)	<ul style="list-style-type: none">• T1• T2• ROI	<ul style="list-style-type: none">• Single subjects	Reported
Neonatal brain template of 1 week newborn (Hashicka et al., 2012)	5.6 ± 17.6 days	14 (M = 11; F = 3)	<ul style="list-style-type: none">• T2	<ul style="list-style-type: none">• Voxel-wise averaging• Single subjects	Not reported

(Continued)

TABLE 1 | Continued

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
Neonatal probabilistic models (Kazemi et al., 2008)	39–42 weeks	7 (M = 3; F = 4)	<ul style="list-style-type: none"> • T1 • Tissue maps 	<ul style="list-style-type: none"> • Voxel-wise averaging 	Not reported
Non-parametric percentile rank atlas of the aging brain (Dickie et al., 2015a)	55–90 years	98 (M = 40; F = 58)	<ul style="list-style-type: none"> • T1 • GM map 	<ul style="list-style-type: none"> • Voxel-wise non-parametric percentile ranking 	Reported
Normal Brain F-18 FDG-PET and MRI Atlas (Schifter et al., 1993)		1	<ul style="list-style-type: none"> • T1 • T2 • FDG-PET 	<ul style="list-style-type: none"> • Co-registration of within subject images 	Not reported
Normal reference MR images for aging brain (Farrell et al., 2009)	65–80 years*	79 (M = 61; F = 18)	<ul style="list-style-type: none"> • T1 • T2 	<ul style="list-style-type: none"> • Qualitative percentile ranking • Voxel-wise averaging 	Reported
NTU standard Chinese brain template (Jao et al., 2009)	19–42 (25.7) years	95 (M = 50; F = 45)	<ul style="list-style-type: none"> • T1 	<ul style="list-style-type: none"> • Voxel-wise averaging 	Reported
Parcellation of the Healthy Neonatal Brain into 107 Regions (Blesa et al., 2016)	39–47 ⁺¹ (42 ⁺²) weeks	33	<ul style="list-style-type: none"> • T1 • T2 • Diffusion • Tissue maps • ROI 	<ul style="list-style-type: none"> • Voxel-wise majority voting 	Reported
Population difference in brain among Chinese, Malay and Indian neonates (Bai et al., 2012)	5–17 days	177 (M = 94; F = 83)	<ul style="list-style-type: none"> • T2 • Diffusion 	<ul style="list-style-type: none"> • Voxel-wise averaging 	Reported
Population-Average, Landmark- and Surface-based (PALS) atlas (Van Essen, 2005)	18–24 years	12 (M = 6; F = 6)	<ul style="list-style-type: none"> • T1 • Cortical surface 	<ul style="list-style-type: none"> • Selected landmark averaging 	Not reported
Regional growth and atlas of the developing human brain (Makropoulos et al., 2016)	39 ⁺¹ (27 ⁺¹ –44 ⁺⁶) weeks PMA	338	<ul style="list-style-type: none"> • T1 • T2 • Tissue maps • ROI 	<ul style="list-style-type: none"> • Voxel-wise averaging 	Not reported
Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on T1-weighted MRI (Wu et al., 2016)	4–82 years*	90	<ul style="list-style-type: none"> • T1 • ROI 	<ul style="list-style-type: none"> • Hierarchical ontology 	Not reported
Spatial-temporal fetal atlas (Zhan et al., 2013)	15–22 weeks GA*	34 (M = 12; F = 22)	<ul style="list-style-type: none"> • T2 	<ul style="list-style-type: none"> • Voxel-wise averaging and SD 	Reported
SRI24 (Rohlfing et al., 2010)	19–84 (52 ± 5) years	24 (M = 12; F = 12)	<ul style="list-style-type: none"> • T1 • T2/PD • Diffusion • Tissue maps • ROI 	<ul style="list-style-type: none"> • Voxel-wise averaging 	Reported

(Continued)

TABLE 1 | Continued

Name	Age ^a	N (Sex)	Sequences/contents	Derivation method	Clinical/cognitive data
Symmetric atlas in normal older adults ^{T&T} (Grabner et al., 2006)	75 ± 6 years	153	<ul style="list-style-type: none">• T1• ROI	<ul style="list-style-type: none">• Voxel-wise averaging	Not reported
Talairach and Tournoux ^{T&T} (Talairach and Tournoux, 1988; Brett et al., 2001)	60 years	1 (M = 0; F = 1)	<ul style="list-style-type: none">• Histological slices• Photographs• Hand drawings• Stereotactic coordinates	<ul style="list-style-type: none">• Postmortem slicing• Photography• Drawing	Not reported
The human brain in 1700 pieces (Nowinski et al., 2012)		1 (M = 0; F = 1)	<ul style="list-style-type: none">• T1• 3D TOF• SWI• Diffusion• ROI	<ul style="list-style-type: none">• Single subject	Not reported
The pediatric template of brain perfusion (Avants et al., 2015)	7–18 years	120 (M = 59; F = 61)	<ul style="list-style-type: none">• T1• BOLD• Diffusion• pCASL• ROI• Tissue maps	<ul style="list-style-type: none">• Voxel-wise averaging	Reported
Three-dimensional digitized mono-subject anatomical template (Lalys et al., 2010)	45 years	1 (M = 1; F = 0)	<ul style="list-style-type: none">• T1• T2	<ul style="list-style-type: none">• Voxel-wise kappa-sigma clipping average (of repeated single subject scans)	Not reported
UNC Infant 0–1–2 atlases (Shi et al., 2011)	0–2 years	95 (M = 56; F = 39)	<ul style="list-style-type: none">• T1• T2• Tissue maps• ROI	<ul style="list-style-type: none">• Voxel-wise averaging• Voxel-wise majority voting (maximum probability)	Reported

Empty or partially empty cells indicate that we could not find relevant data in original manuscripts; ^aage is reported as in the original manuscript and is shown “range (mean ± SD)” if available; MRI, magnetic resonance imaging; SD, standard deviation; ROI, region of interest; PD, proton density; SWI, susceptibility weighted imaging; fMRI, task-based functional magnetic resonance imaging; rfMRI, resting-state functional magnetic resonance imaging; PMA, post-menstrual age; GA, gestational age; pCASL, pseudo continuous arterial spin labeled; BOLD, blood oxygen level-dependent; SSFSE, single shot fast spin echo; M, male; F, female; T&T, developed by or descended from Talairach and Tournoux.

functional connections in the healthy brain between ages 22 to 35 years (Van Essen et al., 2012) and UK Biobank (Miller et al., 2016). The challenge is to collect similarly rich and relevant data, including sequences such as T2* and FLAIR and vascular risk factor measures for appropriately characterizing cerebrovascular and cognitive development/aging effects on brain structure, at the extremes of life. An international collaborative and aggregative approach may be the best way of achieving this goal as was recently agreed by a panel of experts in structural brain mapping in 2014 (Job et al., 2016) and as is evidenced in similar efforts in functional imaging (Zuo et al., 2014). Although there are challenges to aggregating brain MRI from multiple centers/scanners, particularly in functional connectomics (Zuo and Xing, 2014), these issues have received great attention (e.g., Gountouna et al., 2010; Gradin et al., 2010) and the variability between scanners has often shown to be nominal compared to the great variability in brain structure among even people of the same age, gender, and cognitive status (Dickie et al., 2013; Ritchie et al., 2015; Miller et al., 2016).

High resolution structural MRI is increasingly used in population imaging to study brain development in fetal (pre-birth), neonatal (birth to 4 weeks corrected gestational age) and pediatric (1 month to 18 years) populations because of its utility to: provide quantitative measures of typical brain growth; map atypical growth following complications such as preterm birth, perinatal asphyxia and stroke; evaluate tissue effects of neuroprotective treatment strategies; identify the neural substrates of long-term neurodevelopmental impairments; and because it has potential to uncover early life origins of adult neurological and psychiatric disease. All of these applications benefit from the anatomic context provided by atlases.

There are challenges in analyzing structural images in early and late life. These begin during image acquisition and extend into image analysis. For example, infant participants are asleep during scanning while adults are usually awake; motion artifacts are generally low in mid-life but increase at the extremes of life; and heart and respiratory rates also vary greatly through life (Zuo et al., 2017). Brain structural patterns also vary greatly through life: in early life growth is rapid and head shape and size varies, with a changes in tissue composition and relatively low spatial resolution (Matsuzawa et al., 2001). In older people there is accelerated brain tissue loss, reduced cortical contrast, white matter disease, enlarged perivascular spaces, stroke infarcts, and microbleeds, among other features (Raz et al., 2010; Wardlaw et al., 2013; Dickie et al., 2016b). There have been several ($N = 19$) fetal, neonate, or infant (<age 5) atlases published, but our review found relatively limited age-specific childhood ($N = 6$: >5 and <18 years) and older adult atlases ($N = 7$: >60 years) compared to young/middle-aged adult atlases ($N = 23$). Despite their current under-representation in the literature, age-specific atlases in childhood, and old age may have important uses in research and clinical practice, such as providing targets for aiding classification and diagnoses of developmental and neurodegenerative diseases (Farrell et al., 2009; Dickie et al., 2013, 2014), particularly since better understanding of normal development, aging, and

dementia prevention are major focuses of many large population studies.

Most atlases we found were based on mean/parametric statistics and designed to provide a standard space for voxel-wise analyses or support tissue/ROI volume segmentation. In contrast, the “Normal reference MR images for the brain” atlas was based on qualitatively determined percentile ranks of brain volumes during normal aging and designed to support clinical diagnoses of whole brain volume loss in aging (65–70 and 75–80 year old) patients (Farrell et al., 2009). These clinical atlases are designed to “calibrate” differences in perception between neuroradiologists and have been of growing interest and in increased use since their inception in 2009 (Farrell et al., 2009; Hoggard, 2009; Job et al., 2016). Additionally, increased interest in use of computational automated image processing in clinical practice, e.g., to assess brain, hippocampus, or white matter lesion volumes, relies on availability of relevant and reliable age-relevant atlases. Atlases based on parametric statistics, e.g., mean and standard deviation, are not suitable to define the irregular brain volume distributions in old age (Dickie et al., 2013, 2015a). Therefore, non-parametric statistics were recently applied quantitatively to derive voxel-based percentile ranks and limits of normal aging GM, but this atlas was limited by the use of only T1 sequences and a wide age range (Dickie et al., 2015a). Further, work in developing non-parametric distributional representations of the brain, including a broad range of sequences in well-described (cognitively and medically) age-specific groups, may lead to clinically useful atlases for supporting diagnoses of developmental and neurodegenerative disease (Farrell et al., 2009; Wardlaw et al., 2013; Dickie et al., 2014).

The strengths of our review include the use of structured methods, that were reported following the PRISMA Guidelines (Moher et al., 2009), over ~6 years. We also conducted an exhaustive manual search of printed and online materials, and provided a structured evaluation of brain atlases according to pre-specified criteria. This allowed us to produce a holistic review of structural MRI brain atlases from across the life course in detail that we have not found previously. But despite these strengths, our review also has some limitations. The atlases we found were openly published, and identified through a formal search thus we may not have identified all relevant atlases, e.g., those described as part of larger studies (and therefore potentially not visible through traditional search methods) or those not published/openly accessible. We report data as described in the paper or website, and it is possible that additional data, e.g., on subjects' age, sex, clinical information, was collected and may have been published elsewhere. We did not contact authors for additional information. Further, we did not investigate potential uses for atlases beyond those described in the original manuscripts/sources. It could be that any one of these atlases may be modified to serve additional purposes. Related to this, we described the methods and uses of each atlas according to our interpretation of the source manuscripts/reference manuals, which may differ from the meaning intended by the original authors.

Notwithstanding these limitations, we have reviewed and described structural MRI brain atlases from across the life course and found that they were mostly of modest size with limited supporting subject information, developed with restricted image sequences for specific processing purposes, and that childhood and elderly populations were under-represented. We conclude that there is a continuing need for multi-sequence structural MRI, and the associated clinical, medical, and demographic data, collected in population imaging studies to be made widely available (with appropriate legal and ethical approvals) to create non-parametric brain atlases that adequately reflect the variability and features of brain changes throughout the life course. Brain image databanks, such as Brain Imaging in Normal Subjects (BRAINIS; <https://www.brainsimagebank.ac.uk/>; Job et al., 2016), should work together to maximize sample sizes, generalizability and optimize data use to benefit analyses in population imaging studies and in future clinical practice.

AUTHOR CONTRIBUTIONS

DAD and JL conducted systematic searches of the literature and internet. DAD, SS, JL, DA, MBC, and JB, conducted hand searching and reviewing of the literature and internet. DAD and SS wrote the manuscript. DAD, SS, JL, DA, MBC, JB, AW, DR, DJ, and JW edited the manuscript. DAD, SS, DR, DJ, and JW conceptualized and designed the study.

REFERENCES

- Adelman, G., and Smith, B. H. (1987). *Encyclopedia of Neuroscience*. Boston, MA: Birkhäuser.
- Ahsan, R. L., Allom, R., Gousias, I. S., Habib, H., Turkheimer, F. E., Free, S., et al. (2007). Volumes, spatial extents and a probabilistic atlas of the human basal ganglia and thalamus. *Neuroimage* 38, 261–270. doi: 10.1016/j.neuroimage.2007.06.004
- Allen Institute for Brain Science (2009). *Allen Human Brain Atlas*. Available online at: http://human.brain-map.org/mri_viewers/data (Accessed August 26, 2013).
- Allen, J. S., Bruss, J., Brown, C. K., and Damasio, H. (2005). Normal neuroanatomical variation due to age: the major lobes and a parcellation of the temporal region. *Neurobiol. Aging* 26, 1245–1260. doi: 10.1016/j.neurobiolaging.2005.05.023
- Altaye, M., Holland, S. K., Wilke, M., and Gaser, C. (2008). Infant brain probability templates for MRI segmentation and normalization. *Neuroimage* 43, 721–730. doi: 10.1016/j.neuroimage.2008.07.060
- Aubert-Broche, B., Evans, A. C., and Collins, L. (2006). A new improved version of the realistic digital brain phantom. *Neuroimage* 32, 138–145. doi: 10.1016/j.neuroimage.2006.03.052
- Avants, B. B., Duda, J. T., Kilroy, E., Krasileva, K., Jann, K., Kandel, B. T., et al. (2015). The pediatric template of brain perfusion. *Sci. Data* 2, 150003. doi: 10.1038/sdata.2015.3
- Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004
- Bai, J., Abdul-Rahman, M. F., Rifkin-Graboi, A., Chong, Y.-S., Kwek, K., Saw, S.-M., et al. (2012). Population differences in brain morphology and microstructure among Chinese, Malay, and Indian Neonates. *PLoS ONE* 7:e47816. doi: 10.1371/journal.pone.0047816
- Betz, W. (1874). Anatomischer nachweis zweier gehirncentra. *Zentralbl Med Wiss* 12, 578–595.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge and thank the following centres and funders. This work was carried out in The University of Edinburgh Brain Research Imaging Centre (BRIC; <http://www.bric.ed.ac.uk/>) within the Department of Neuroimaging Sciences. BRIC is part of the Scottish Imaging Network, A Platform for Scientific Excellence (SINAPSE) collaboration (<http://www.sinapse.ac.uk/>), funded by the Scottish Funding Council, Scottish Executive Chief Scientist Office, and the six collaborator Universities. Professor JW was funded by the Scottish Funding Council and Scottish Executive Chief Scientist Office through the SINAPSE collaboration. DAD was funded by a SINAPSE industrial collaboration (SPIRIT) Ph.D. scholarship, a Medical Research Council (MRC) scholarship, and the Tony Watson Scholarship bequest to The University of Edinburgh; and is currently funded by Innovate UK. Dr. DJ was funded by Wellcome Trust Grant 007393/Z/05/Z. Funding from Edinburgh and Lothians Health Foundation 53/311 and BBSRC Sparking Impact SI 2013-0210 is gratefully acknowledged. The University of Edinburgh Centre for Cognitive Aging and Cognitive Epidemiology (SS) is part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding from the Biotechnology and Biological Sciences Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Medical Research Council is also gratefully acknowledged.

- Blesa, M., Serag, A., Wilkinson, A. G., Anblagan, D., Telford, E. J., Pataky, R., et al. (2016). Parcellation of the healthy neonatal brain into 107 regions using atlas propagation through intermediate time points in childhood. *Front. Neurosci.* 10:220. doi: 10.3389/fnins.2016.00220
- Bossa, M., Hernandez, M., and Olmos, S. (2007). “Contributions to 3D diffeomorphic atlas estimation: application to brain images,” in *Proceedings of the 10th International Conference on Medical Image Computing and Computer-Assisted Intervention - Volume Part I* (Brisbane, QLD: Springer-Verlag).
- Brett, M., Christoff, K., Cusack, R., and Lancaster, J. (2001). Using the Talairach atlas with the MNI template. *Neuroimage* 13, S85. doi: 10.1016/S1053-8119(01)91428-4
- Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Großhirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Leipzig: Barth.
- Brodmann, K. (1994). *The Principles of Comparative Localisation in the Cerebral Cortex Based on Cytoarchitectonics*. London: Smith-Gordon.
- Buckner, R. L., Head, D., Parker, J., Fotenos, A. F., Marcus, D., Morris, J. C., et al. (2004). A unified approach for morphometric and functional data analysis in young, old, and demented adults using automated atlas-based head size normalization: reliability and validation against manual measurement of total intracranial volume. *Neuroimage* 23, 724–738. doi: 10.1016/j.neuroimage.2004.06.018
- Chen, T., Rangarajan, A., Eisenschenk, S. J., and Vemuri, B. C. (2012). Construction of a neuroanatomical shape complex atlas from 3D MRI brain structures. *Neuroimage* 60, 1778–1787. doi: 10.1016/j.neuroimage.2012.01.095
- Courchesne, E., Chisum, H. J., Townsend, J., Cowles, A., Covington, J., Egaas, B., et al. (2000). Normal brain development and aging: quantitative analysis at *in vivo* MR imaging in healthy volunteers. *Radiology* 216, 672–682. doi: 10.1148/radiology.216.3.r00au37672
- Deary, I. J., Gow, A. J., Pattie, A., and Starr, J. M. (2012). Cohort profile: the Lothian birth cohorts of 1921 and 1936. *Int. J. Epidemiol.* 41, 1576–1584. doi: 10.1093/ije/dyr197
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., et al. (2006). An automated labeling system for subdividing the human

- cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 968–980. doi: 10.1016/j.neuroimage.2006.01.021
- Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53, 1–15. doi: 10.1016/j.neuroimage.2010.06.010
- Devlin, J. T., and Poldrack, R. A. (2007). In praise of tedious anatomy. *Neuroimage* 37, 1033–1041. doi: 10.1016/j.neuroimage.2006.09.055
- Dickie, D. A., Job, D. E., Gonzalez, D. R., Shenkin, S. D., Ahearn, T. S., Murray, A. D., et al. (2013). Variance in brain volume with advancing age: implications for defining the limits of normality. *PLoS ONE* 8:e84093. doi: 10.1371/journal.pone.0084093
- Dickie, D. A., Job, D. E., Gonzalez, D. R., Shenkin, S. D., and Wardlaw, J. M. (2015a). Use of brain MRI atlases to determine boundaries of age-related pathology: the importance of statistical method. *PLoS ONE* 10:e0127939. doi: 10.1371/journal.pone.0127939
- Dickie, D. A., Job, D. E., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2016a). *Brain Imaging of Normal Subjects (BRAINS) Age-Specific MRI Atlases from Young Adults to the Very Elderly (v1.0)*, [dataset]. University of Edinburgh, Edinburgh Imaging, CCBS, BRAINS Imagebank. doi: 10.7488/ds/1369
- Dickie, D. A., Job, D. E., Sparrow, S., Piyasena, C., Wilkinson, G., Wardlaw, J. M., et al. (2014). “Preterm infant brain pathology revealed in individuals by voxel ranking against a normal term atlas,” in *Proceedings of the 20th Annual Meeting of the Organization for Human Brain Mapping* (Hamburg).
- Dickie, D. A., Karama, S., Ritchie, S. J., Cox, S. R., Sakka, E., Royle, N. A., et al. (2015b). Progression of white matter disease and cortical thinning are not related in older community-dwelling subjects. *Stroke* 47, 410–416. doi: 10.1161/STROKEAHA.115.011229
- Dickie, D. A., Ritchie, S. J., Cox, S. R., Sakka, E., Royle, N. A., Aribisala, B. S., et al. (2016b). Vascular risk factors and progression of white matter hyperintensities in the Lothian Birth Cohort 1936. *Neurobiol. Aging* 42, 116–123. doi: 10.1016/j.neurobiolaging.2016.03.011
- Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., et al. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* 25, 1325–1335. doi: 10.1016/j.neuroimage.2004.12.034
- Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., and Peters, T. M. (1993). “3D statistical neuroanatomical models from 305 MRI volumes,” in *IEEE Nuclear Science Symposium and Medical Imaging Conference* (San Francisco, CA), 1813–1817.
- Evans, A. C., Janke, A. L., Collins, D. L., and Baillet, S. (2012). Brain templates and atlases. *Neuroimage* 62, 911–922. doi: 10.1016/j.neuroimage.2012.01.024
- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., et al. (2016). The human Brainnetome atlas: a new brain atlas based on connective architecture. *Cereb. Cortex* 26, 3508–3526. doi: 10.1093/cercor/bhw157
- Farrell, C., Chappell, F., Armitage, P. A., Keston, P., MacLulich, A., Shenkin, S., et al. (2009). Development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. *Eur. Radiol.* 19, 177–183. doi: 10.1007/s00330-008-1119-2
- Fillmore, P. T., Phillips-Meek, M., and Richards, J. E. (2015). Age-specific MRI brain and head templates for healthy adults from twenty through eighty-nine years of age. *Front. Aging Neurosci.* 7:44. doi: 10.3389/fnagi.2015.00044
- Fmrib (2008). *Atlases Included with FSL*. Available online at: <http://www.fmrib.ox.ac.uk/fsl/data/atlas-descriptions.html> (Accessed March 25, 2011).
- Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., and Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* 54, 313–327. doi: 10.1016/j.neuroimage.2010.07.033
- Gholipour, A., Limperopoulos, C., Clancy, S., Clouchoux, C., Akhond-Asl, A., Estroff, J. A., et al. (2014). Construction of a deformable spatiotemporal mri atlas of the fetal brain: evaluation of similarity metrics and deformation models. *Med. Image Comput. Comput. Assist. Interv.* 17, 292–299. doi: 10.1007/978-3-319-10470-6_37
- Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., et al. (2016). A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171–178. doi: 10.1038/nature18933
- Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proc. Natl. Acad. Sci. U.S.A.* 101, 8174–8179. doi: 10.1073/pnas.0402680101
- Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N. A., Friston, K. J., and Frackowiak, R. S. J. (2001). A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 14, 21–36. doi: 10.1006/nimg.2001.0786
- Gountouna, V., Job, D., McIntosh, A., Moorhead, T., Lymer, G., Whalley, H., et al. (2010). Functional magnetic resonance imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage* 49, 552–560. doi: 10.1016/j.neuroimage.2009.07.026
- Gousias, I. S., Edwards, A. D., Rutherford, M. A., Counsell, S. J., Hajnal, J. V., Rueckert, D., et al. (2012). Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage* 62, 1499–1509. doi: 10.1016/j.neuroimage.2012.05.083
- Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., et al. (2008). Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. *Neuroimage* 40, 672–684. doi: 10.1016/j.neuroimage.2007.11.034
- Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. (2006). Symmetric atlas and model based segmentation: an application to the hippocampus in older adults. *Med. Image Comput. Comput. Assist. Interv.* 4191, 58–66. doi: 10.1007/11866763_8
- Gradin, V., Gountouna, V. E., Waiter, G., Ahearn, T. S., Brennan, D., Condon, B., et al. (2010). Between-and within-scanner variability in the CalBrain study n-back cognitive task. *Psychiatry Res. Neuroimaging* 184, 86–95. doi: 10.1016/j.psychres.2010.08.010
- Gur, R. C., Mozley, P. D., Resnick, S. M., Gottlieb, G. L., Kohn, M., Zimmerman, R., et al. (1991). Gender differences in age effect on brain atrophy measured by magnetic resonance imaging. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2845–2849. doi: 10.1073/pnas.88.7.2845
- Habas, P. A., Kim, K., Corbett-Detig, J. M., Rousseau, F., Glenn, O. A., Barkovich, A. J., et al. (2010). A spatiotemporal atlas of MR intensity, tissue probability and shape of the fetal brain with application to segmentation. *Neuroimage* 53, 460–470. doi: 10.1016/j.neuroimage.2010.06.054
- Hammers, A., Allom, R., Koepp, M., Free, S., Myers, R., Lemieux, L., et al. (2003). Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Hum. Brain Mapp.* 19, 224–247. doi: 10.1002/hbm.10123
- Hashioka, A., Kobashi, S., Kuramoto, K., Wakata, Y., Ando, K., Ishikura, R., et al. (2012). A neonatal brain MR image template of 1 week newborn. *Int. J. Comput. Assist. Radiol. Surg.* 7, 273–280. doi: 10.1007/s11548-011-0646-5
- Hoggard, N. (2009). Re: development and initial testing of normal reference MR images for the brain at ages 65–70 and 75–80 years. *Eur. Radiol.* 19, 1025–1025. doi: 10.1007/s00330-008-1230-4
- Holmes, C. J., Hoge, R., Collins, L., Woods, R., Toga, A. W., and Evans, A. C. (1998). Enhancement of MR images using registration for signal averaging. *J. Comput. Assist. Tomogr.* 22, 324. doi: 10.1097/00004728-199803000-00032
- Jao, T., Chang, C. Y., Li, C. W., Chen, D. Y., Wu, E., Wu, C. W., et al. (2009). “Development of NTU standard Chinese brain template: Morphologic and functional comparison with MNI template using magnetic resonance imaging,” in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Minneapolis, MN), 4779–4782.
- Job, D. E., Dickie, D. A., Rodriguez, D., Robson, A., Danso, S., Pernet, C., et al. (2016). A brain imaging repository of normal structural MRI across the life course: Brain Images of Normal Subjects (BRAINS). *NeuroImage* 144, 299–304. doi: 10.1016/j.neuroimage.2016.01.027
- Kabdebon, C., Leroy, F., Simmonet, H., Perrot, M., Dubois, J., and Dehaene-Lambertz, G. (2014). Anatomical correlations of the international 10–20 sensor placement system in infants. *Neuroimage* 99, 342–356. doi: 10.1016/j.neuroimage.2014.05.046
- Kazemi, K., Ghadimi, S., Abrishami-Moghaddam, H., Grebe, R., Gondry-Jouet, C., and Wallois, F. (2008). “Neonatal probabilistic models for brain, CSF and skull using T1-MRI data: Preliminary results,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Vancouver, BC), 3892–3895.
- Kazemi, K., Moghaddam, H. A., Grebe, R., Gondry-Jouet, C., and Wallois, F. (2007). A neonatal atlas template for spatial normalization of whole-brain

- magnetic resonance images of newborns: preliminary results. *Neuroimage* 37, 463–473. doi: 10.1016/j.neuroimage.2007.05.004
- Klein, A., and Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Front. Neurosci.* 6:171. doi: 10.3389/fnins.2012.00171
- Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., et al. (2011). A dynamic 4D probabilistic atlas of the developing brain. *Neuroimage* 54, 2750–2763. doi: 10.1016/j.neuroimage.2010.10.019
- Lalys, F., Haegelen, C., Ferre, J.-C., El-Ganaoui, O., and Jannin, P. (2010). Construction and assessment of a 3-T MRI brain template. *Neuroimage* 49, 345–354. doi: 10.1016/j.neuroimage.2009.08.007
- Lancaster, J. L., Tordesillas-Gutiérrez, D., Martínez, M., Salinas, F., Evans, A., Zilles, K., et al. (2007). Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* 28, 1194–1205. doi: 10.1002/hbm.20345
- Lee, J. S., Lee, D. S., Kim, J., Kim, Y. K., Kang, E., Kang, H., et al. (2005). Development of Korean standard brain templates. *J. Korean Med. Sci.* 20, 483–488. doi: 10.3346/jkms.2005.20.3.483
- Lemaire, H., Crivello, F., Grassiot, B., Alperovitch, A., Tzourio, C., and Mazoyer, B. (2005). Age- and sex-related effects on the neuroanatomy of healthy elderly. *Neuroimage* 26, 900–911. doi: 10.1016/j.neuroimage.2005.02.042
- Lim, I. A. L., Faria, A. V., Li, X., Hsu, J. T. C., Airan, R. D., Van Zijl, P. C. M., et al. (2013). Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: application to determine iron content in deep gray matter structures. *Neuroimage* 82, 449–469. doi: 10.1016/j.neuroimage.2013.05.127
- Loni (2011). *Alzheimer's Disease Template*. Available online at: <http://www.loni.usc.edu/atlas/> (Accessed).
- Luders, E., Narr, K. L., Thompson, P. M., Woods, R. P., Rex, D. E., Jancke, L., et al. (2005). Mapping cortical gray matter in the young adult brain: effects of gender. *Neuroimage* 26, 493–501. doi: 10.1016/j.neuroimage.2005.02.010
- Makropoulos, A., Aljabar, P., Wright, R., Hüning, B., Merchant, N., Arichi, T., et al. (2016). Regional growth and atlas of the developing human brain. *Neuroimage* 125, 456–478. doi: 10.1016/j.neuroimage.2015.10.047
- Maldjian, J. A., Laurienti, P. J., Kraft, R. A., and Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19, 1233–1239. doi: 10.1016/S1053-8119(03)00169-1
- Matsuzawa, J., Matsui, M., Konishi, T., Noguchi, K., Gur, R. C., Bilker, W., et al. (2001). Age-related volumetric changes of brain gray and white matter in healthy infants and children. *Cereb. Cortex* 11, 335–342. doi: 10.1093/cercor/11.4.335
- Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915
- Miller, K. L., Alfaro-Almagro, F., Bangarter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536. doi: 10.1038/nn.4393
- Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6:e1000097. doi: 10.1371/journal.pmed.1000097
- Muzik, O., Chugani, D. C., Juhász, C., Shen, C., and Chugani, H. T. (2000). Statistical parametric mapping: assessment of application in children. *Neuroimage* 12, 538–549. doi: 10.1006/nimg.2000.0651
- Nowinski, W. L. (2005). The cerefy brain atlases. *Neuroinformatics* 3, 293–300. doi: 10.1385/NI:3:4:293
- Nowinski, W. L., Chua, B. C., Qian, G. Y., and Nowinska, N. G. (2012). The human brain in 1700 pieces: design and development of a three-dimensional, interactive and reference atlas. *J. Neurosci. Methods* 204, 44–60. doi: 10.1016/j.jneumeth.2011.10.021
- O'Connor, J. P. B. (2003). Thomas Willis and the background to Cerebri Anatomie. *J. R. Soc. Med.* 96, 139–143. doi: 10.1258/jrsm.96.3.139
- Oishi, K., Mori, S., Donohue, P. K., Ernst, T., Anderson, L., Buchthal, S., et al. (2011). Multi-contrast human neonatal brain atlas: application to normal neonate development analysis. *Neuroimage* 56, 8–20. doi: 10.1016/j.neuroimage.2011.01.051
- Raz, N., Ghisletta, P., Rodrigue, K. M., Kennedy, K. M., and Lindenberger, U. (2010). Trajectories of brain aging in middle-aged and older adults: regional and individual differences. *Neuroimage* 51, 501–511. doi: 10.1016/j.neuroimage.2010.03.020
- Richards, J. E., Sanchez, C., Phillips-Meek, M., and Xie, W. (2016). A database of age-appropriate average MRI templates. *NeuroImage* 124(Pt B), 1254–1259. doi: 10.1016/j.neuroimage.2015.04.055
- Ritchie, S. J., Dickie, D. A., Cox, S. R., Valdes Hernandez, M. D. C., Corley, J., Royle, N. A., et al. (2015). Brain volumetric changes and cognitive ageing during the eighth decade of life. *Hum. Brain Mapp.* 36, 4910–4925. doi: 10.1002/hbm.22959
- Rohlfing, T., Zahr, N., Sullivan, E., and Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.* 31, 798–819. doi: 10.1002/hbm.20906
- Rorden, C., Bonilha, L., and Nichols, T. E. (2007). Rank-order versus mean based statistics for neuroimaging. *Neuroimage* 35, 1531–1537. doi: 10.1016/j.neuroimage.2006.12.043
- Sanchez, C. E., Richards, J. E., and Almli, C. R. (2012a). Age-specific MRI templates for pediatric neuroimaging. *Dev. Neuropsychol.* 37, 379–399. doi: 10.1080/87565641.2012.688900
- Sanchez, C. E., Richards, J. E., and Almli, C. R. (2012b). Neurodevelopmental MRI brain templates for children from 2 weeks to 4 years of age. *Dev. Psychobiol.* 54, 77–91. doi: 10.1002/dev.20579
- Schifter, T., Turkington, T. G., Berlangieri, S. U., Hoffman, J. M., Macfall, J. R., Pelizzari, C. A., et al. (1993). Normal brain F-18 FDG-PET and MRI anatomy. *Clin. Nucl. Med.* 18, 578–582. doi: 10.1097/00003072-199307000-00008
- Serag, A., Aljabar, P., Ball, G., Counsell, S. J., Boardman, J. P., Rutherford, M. A., et al. (2012a). Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage* 59, 2255–2265. doi: 10.1016/j.neuroimage.2011.09.062
- Serag, A., Kyriakopoulou, V., Rutherford, M., Edwards, A., Hajnal, J., Aljabar, P., et al. (2012b). A multi-channel 4D probabilistic atlas of the developing brain: application to fetuses and neonates. *Ann. BMVA*, 2012, 1–14.
- Shan, Z. Y., Parra, C., Ji, Q., Ogg, R. J., Zhang, Y., Laningham, F. H., et al. (2006). “A digital pediatric brain structure atlas from T1-weighted MR images,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1–6, 2006, Proceedings, Part II*, eds R. Larsen, M. Nielsen, and J. Sporring. (Berlin; Heidelberg: Springer Berlin Heidelberg), 332–339.
- Shattuck, D. W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K. L., et al. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage* 39, 1064–1080. doi: 10.1016/j.neuroimage.2007.09.031
- Shenton, M. E., Kikinis, R., McCarley, R. W., Saiviroonpom, P., Hokama, H. H., Robatino, A., et al. (1995). “Harvard brain atlas: a teaching and visualization tool,” in *Proceedings of the 1995 Biomedical Visualization* (Washington, DC), 10–17.
- Shi, F., Yap, P.-T., Fan, Y., Gilmore, J. H., Lin, W., and Shen, D. (2010). Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. *Neuroimage* 51, 684–693. doi: 10.1016/j.neuroimage.2010.02.025
- Shi, F., Yap, P.-T., Wu, G., Jia, H., Gilmore, J. H., Lin, W., et al. (2011). Infant Brain Atlases from Neonates to 1- and 2-Year-Olds. *PLoS ONE* 6:e18746. doi: 10.1371/journal.pone.0018746
- Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L., and Toga, A. W. (2003). Mapping cortical change across the human life span. *Nat. Neurosci.* 6, 309–315. doi: 10.1038/nn1008
- Subsol, G., Roberts, N., Doran, M., Thirion, J.-P., and Whitehouse, G. H. (1997). Automatic analysis of cerebral atrophy. *Magn. Reson. Imaging* 15, 917–927. doi: 10.1016/S0730-725X(97)00002-7
- Talairach, J., and Tournoux, P. (1988). *Co-planar Stereotactic Atlas of the Human Brain: 3-dimensional Proportional System: An Approach to Cerebral Imaging*. Stuttgart: Georg Thieme Verlag.

- Talairach, J., Szikla, G., Tournoux, P., Prosalenti, A., Bordas-Ferrier, M., Covello, L., et al. (1967). *Atlas D'Anatomie Stereotaxique du Telencephale*. Paris: Masson.
- Tang, Y., Hojatkashani, C., Dinov, I., Sun, B., Fan, L., Lin, X., et al. (2010). The construction of a Chinese MRI brain atlas: a morphometric comparison study between Chinese and Caucasian cohorts. *Neuroimage* 51, 33–41. doi: 10.1016/j.neuroimage.2010.01.111
- Toga, A. W. (2002). Neuroimage databases: the good, the bad and the ugly. *Nat. Rev. Neurosci.* 3, 302–309. doi: 10.1038/nrn782
- Toga, A. W., and Thompson, P. M. (2007). What is where and why it is important. *Neuroimage* 37, 1045–1068. doi: 10.1016/j.neuroimage.2007.02.018
- Toga, A. W., Thompson, P. M., Mori, S., Amunts, K., and Zilles, K. (2006). Towards multimodal atlases of the human brain. *Nat. Rev. Neurosci.* 7, 952–966. doi: 10.1038/nrn2012
- Uchiyama, H. T., Seki, A., Tanaka, D., Koeda, T., and Group, J. C. S. (2013). A study of the standard brain in Japanese children: morphological comparison with the MNI template. *Brain Dev.* 35, 228–235. doi: 10.1016/j.braindev.2012.04.005
- Van Essen, D. C. (2005). A population-average, landmark-and surface-based (PALS) atlas of human cerebral cortex. *Neuroimage* 28, 635–662. doi: 10.1016/j.neuroimage.2005.06.058
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., et al. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage* 62, 2222–2231. doi: 10.1016/j.neuroimage.2012.02.018
- Van Leemput, K. (2009). Encoding probabilistic brain atlases using Bayesian inference. *IEEE Trans. Med. Imaging* 28, 822–837. doi: 10.1109/TMI.2008.2010434
- Von Economo, C., and Koskinas, G. N. (1925). *Die Cytoarchitektonik der Hirnrinde des Erwachsenen Menschen*. Berlin: Julius Springer.
- Wardlaw, J. M., Bastin, M. E., Valdés Hernández, M. C., Mu-oz Maniega, S., Royle, N. A., Morris, Z., et al. (2011). Brain aging, cognition in youth and old age and vascular disease in the Lothian Birth Cohort 1936: rationale, design and methodology of the imaging protocol. *Int. J. Stroke* 6, 547–559. doi: 10.1111/j.1747-4949.2011.00683.x
- Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R., et al. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *Lancet Neurol.* 12, 822–838. doi: 10.1016/S1474-4422(13)70124-8
- Warntjes, J. B. M., Engström, M., Tisell, A., and Lundberg, P. (2013). Brain characterization using normalized quantitative magnetic resonance imaging. *PLoS ONE* 8:e70864. doi: 10.1371/journal.pone.0070864
- Westbury, C. F., Zatorre, R. J., and Evans, A. C. (1999). Quantifying variability in the planum temporale: a probability map. *Cereb. Cortex* 9, 392–405. doi: 10.1093/cercor/9.4.392
- Wilke, M., Holland, S., Altaye, M., and Gaser, C. (2008). Template-O-Matic: a toolbox for creating customized pediatric templates. *Neuroimage* 41, 903–913. doi: 10.1016/j.neuroimage.2008.02.056
- Wu, D., Ma, T., Ceritoglu, C., Li, Y., Chotianonta, J., Hou, Z., et al. (2016). Resource atlases for multi-atlas brain segmentations with multiple ontology levels based on T1-weighted MRI. *Neuroimage* 125, 120–130. doi: 10.1016/j.neuroimage.2015.10.042
- Xing, W., Nan, C., Zhentao, Z., Rong, X., Luo, J., Zhuo, Y., et al. (2013). Probabilistic MRI brain anatomical atlases based on 1,000 Chinese subjects. *PLoS ONE* 8:e50939. doi: 10.1371/journal.pone.0050939
- Yeh, F.-C., and Tseng, W.-Y. I. (2011). NTU-90: a high angular resolution brain atlas constructed by q-space diffeomorphic reconstruction. *Neuroimage* 58, 91–99. doi: 10.1016/j.neuroimage.2011.06.021
- Yoon, U., Fonov, V. S., Perusse, D., and Evans, A. C. (2009). The effect of template choice on morphometric analysis of pediatric brain data. *Neuroimage* 45, 769–777. doi: 10.1016/j.neuroimage.2008.12.046
- Yoon, U., Lee, J.-M., Koo, B. B., Shin, Y.-W., Lee, K. J., Kim, I. Y., et al. (2005). Quantitative analysis of group-specific brain tissue probability map for schizophrenic patients. *Neuroimage* 26, 502–512. doi: 10.1016/j.neuroimage.2005.01.056
- Zhan, J., Dinov, I. D., Li, J., Zhang, Z., Hobel, S., Shi, Y., et al. (2013). Spatial-temporal atlas of human fetal brain development during the early second trimester. *Neuroimage* 82, 115–126. doi: 10.1016/j.neuroimage.2013.05.063
- Zuo, X.-N., and Xing, X.-X. (2014). Test-retest reliabilities of resting-state fMRI measurements in human brain functional connectomics: a systems neuroscience perspective. *Neurosci. Biobehav. Rev.* 45, 100–118. doi: 10.1016/j.neubiorev.2014.05.009
- Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Sci. Data* 1, 140049. doi: 10.1038/sdata.2014.49
- Zuo, X.-N., He, Y., Betzel, R. F., Colcombe, S., Sporns, O., and Milham, M. P. (2017). Human connectomics across the life span. *Trends Cogn. Sci.* 21, 32–45. doi: 10.1016/j.tics.2016.10.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Dickie, Shenkin, Anblagan, Lee, Blesa Cabeza, Rodriguez, Boardman, Waldman, Job and Wardlaw. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read,
for greatest visibility



COLLABORATIVE PEER-REVIEW

Designed to be rigorous
– yet also collaborative,
fair and constructive



FAST PUBLICATION

Average 85 days from
submission to publication
(across all journals)



COPYRIGHT TO AUTHORS

No limit to article
distribution and re-use



TRANSPARENT

Editors and reviewers
acknowledged by name
on published articles



SUPPORT

By our Swiss-based
editorial team



IMPACT METRICS

Advanced metrics
track your article's impact



GLOBAL SPREAD

5'100'000+ monthly
article views
and downloads



LOOP RESEARCH NETWORK

Our network
increases readership
for your article

Frontiers

EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on

