BIDS-prov: a provenance framework for BIDS

Rémi Adon, Stefan Appelhoff, Tibor Auer, Laurent Guillo, Yaroslav Halchenko, David Keator, Christopher Markiewicz, Thomas Nichols, Jean-Baptiste Poline, Satrajit Ghosh, et al.

BIDS-prov: a provenance framework for BIDS

Rémi Adon[1], Stefan Appelhoff[2], Tibor Auer[3], Laurent Guillo[4], Yaroslav O. Halchenko[5], David Keator[6], Christopher J. Markiewicz[7], Thomas E. Nichols[8], Jean-Baptiste Poline[9], Satrajit Ghosh[10]*, Camille Maumet[11]*
* equal contributions

1. Inria, Univ Rennes, Inserm, CNRS, Rennes, France - *Contributions*: Software, Writing Original draft
2. Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
3. School of Psychology, University of Surrey, Guildford, UK - *Contributions*: Software, Writing – Review & Editing
4. Univ Rennes, Inria, CNRS, IRISA, France - *Contributions*: Supervision
5. Dartmouth College, NH, USA - *Contributions*: Writing – Review & Editing
5. Department of Psychiatry and Human Behavior, University of California, Irvine. - *Contributions*: Conceptualization, Writing – Review & Editing
7. Stanford University, Stanford, CA, USA - Contributions: Writing - Review & Editing
8. Big Data Institute BuildingBig Data Institute Building, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, UK - *Contributions*: Conceptualization, Writing – Review & Editing
9. McGill University, Montreal, Canada, and UC Berkeley, CA, USA - *Contributions*: Conceptualization, Review & Editing
10. Massachusetts Institute of Technology, Cambridge, MA, USA - *Contributions*: Conceptualization, Methodology, Writing - Review & Editing
11. Inria, Univ Rennes, Inserm, CNRS, Rennes, France - *Contributions*: Conceptualization, Methodology, Writing Original draft, Supervision

Introduction

Differences in details of analysis pipelines have a non negligible impact on neuroimaging results: end-to-end pipelines (Botvinik-Nezer et al., 2020), neuroimaging software package (Bowring et al., 2018), software versions (Gronenschild et al., 2012) and operating system (Glatard et al., 2015) have all been shown to introduce some level of variations. In order to interpret and compare scientific results as well as enable data reuse, researchers need a precise description of a hierarchy of data manipulation and transformations steps from original data to a finding. This description or ‘provenance’ includes information about data, software (versions, parameters, etc.) and people.

The Brain Imaging Data Structure (BIDS) (Gorgolewski et al., 2016) has been well adopted in the neuroimaging community and provides structured file hierarchies with JSON metadata files to represent many different aspects of a brain datasets including: raw data across modalities (MRI, EEG, iEEG) but also some derived data. However, BIDS does not capture details of transformations within a BIDS dataset (e.g., DICOMs to BIDS files, BIDS derivatives). Here we review our work on building a formal provenance framework for BIDS.

Methods

This work was developed as part of a collaborative effort under BIDS Extension Proposal (BEP) 28 (BIDS-PROV). BIDS-PROV uses semantic web technologies and is based on the W3C PROV family of specification that defines three main types of objects:

- Agent: a piece of software or an individual that can transform data
- Entity: a digital input/output, e.g. a file on disk
- Activity: a process applied on some entities (inputs) to produce other entities (outputs).

BIDS-PROV adds provenance information to the BIDS structure in the form of sidecar JSON-LD files, that can be generated by the processing software itself or post-hoc. The generic model can represent any pipeline regardless of the tools that were used. The semantics are described using controlled vocabularies consistent with the Neuroimaging Data Model (NIDM), which includes BIDS terminologies. This supports queries of experimental metadata and computational workflows used to generate scientific results.

Results

The BIDS-Prov specification is available at: https://bids.neuroimaging.io/bep028.

An overview of the proposed model is provided in Fig. 1. Depending on the available contextual information, BIDS-PROV can be used to describe provenance at different levels of granularity. For instance, when directly using a neuroimaging software package, BIDS-PROV could be used to describe each call to a module and its input / outputs. In another use case, in which data processing would be done using a docker container, BIDS PROV could be used to describe a much simpler pipeline of a single Activity and linking to the container image.

Examples on real datasets are available at: https://github.com/bids-standard/BEP028_BIDSprov/.

Conclusions

Here, we propose enriching the BIDS specification with a standardized representation of provenance for any BIDS dataset. The representation is human-readable and can capture different levels of granularity, from an entire workflow to individual steps. This fine-grained tracking of low level details contextualizes our understanding of study results with respect to data generation, computational processes, and human decisions. We applied this representation onto existing examples produced by three widely-used software packages. Finally, we encourage high-level visualisation / querying on these graphs, to make validation and information retrieval accessible to both experts and non-experts.

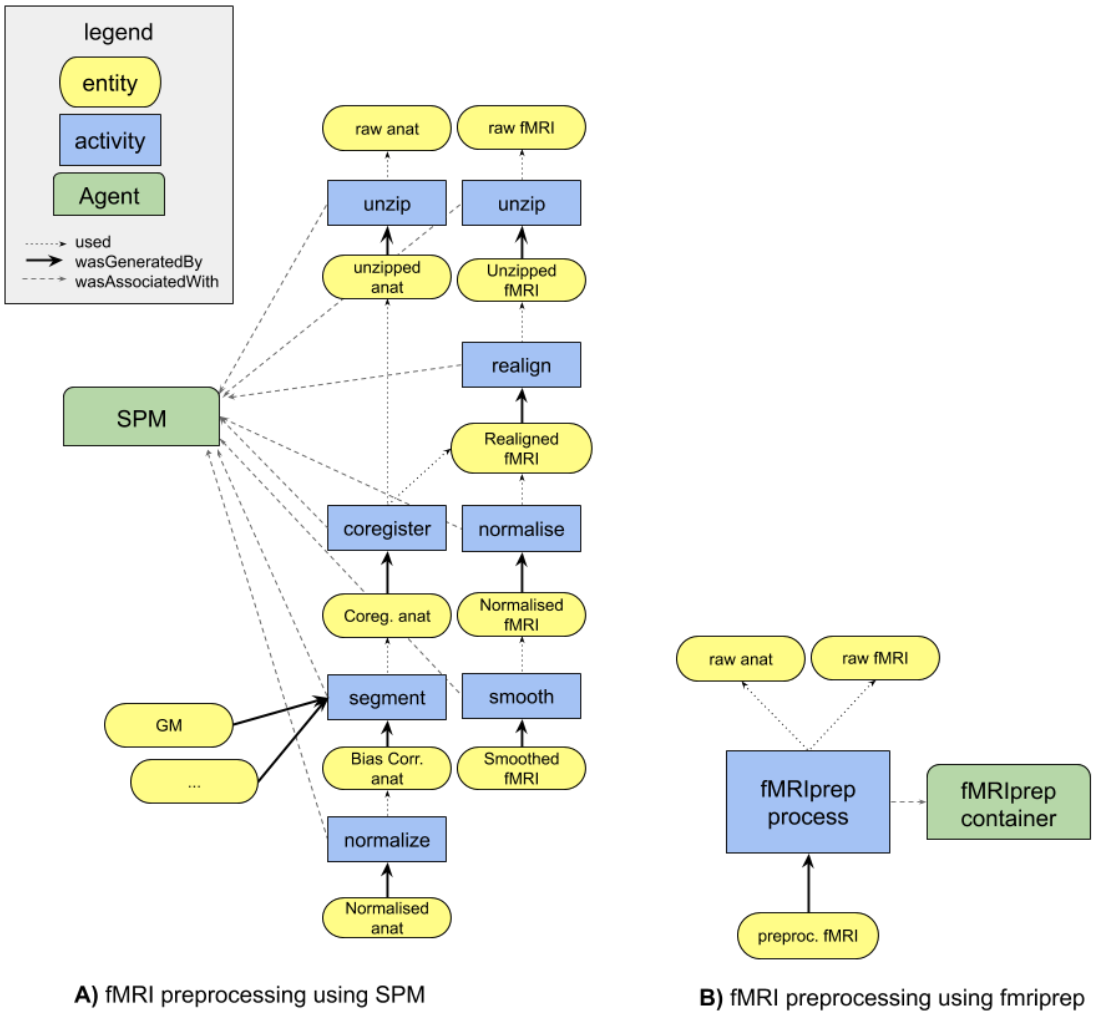

Figure 1: Two examples of BIDS-PROV graphs with A) a detailed and B) a compact graph representing the derived data provenance.

References

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Bowring, A., Maumet, C., & Nichols, T. (2018). *Exploring the Impact of Analysis Software on Task fMRI Results*. https://doi.org/10.1101/285585

Glatard, T., Lewis, L. B., Ferreira da Silva, R., Adalat, R., Beck, N., Lepage, C., Rioux, P., Rousseau, M.-E., Sherif, T., Deelman, E., Khalili-Mahani, N., & Evans, A. C. (2015). Reproducibility of neuroimaging analyses across operating systems. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00012

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). BIDS. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Gronenschild, E. H. B. M., Habets, P., Jacobs, H. I. L., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements. *PLoS ONE*, *7*(6), e38234. https://doi.org/10.1371/journal.pone.0038234