ReproIn: automatic generation of shareable, version-controlled BIDS datasets from MR scanners

Matteo Visconti di Oleggio Castello[1], James E. Dobson[1], Terry Sackett[1], Chandana Kodiweera[1], James V. Haxby[1], Mathias Goncalves[2], Satrajit Ghosh[2], Yaroslav O. Halchenko[1]

[1]Dartmouth College, [2]MIT

OHBM 2018, Poster #2000

Introduction

The formalization of the Brain Imaging Data Structure (BIDS, Gorgolewski et al., 2016) made it easier for researchers to collaborate on shared data and to benefit from standardized processing using various BIDS-aware application.

At Dartmouth, following the philosophy that science should be open by design (Halchenko & Hanke, 2015), we automated the collection of neuroimaging data as a hierarchy of BIDS datasets right from the MR scanner.

Using BIDS standard allows investigators to immediately use BIDS-aware applications for data QA (e.g., bids-validator, MRIQC, Esteban et al. 2017) to catch obvious problems with data acquisition, and to automate preprocessing and analysis.

The entire process occurs in a Singularity container (Kurtzer et al., 2017), and all data is version-controlled with DataLad and git.

Our approach eliminates virtually any ambiguity in data provenance.

Setup

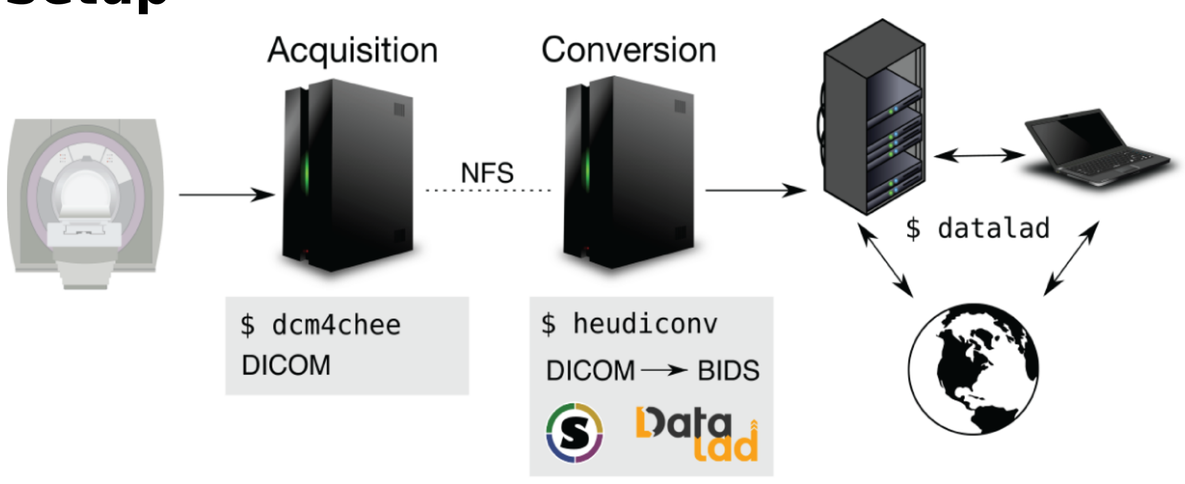


Figure 1. ReproIn workflow in use at the Dartmouth Brain Imaging Center.

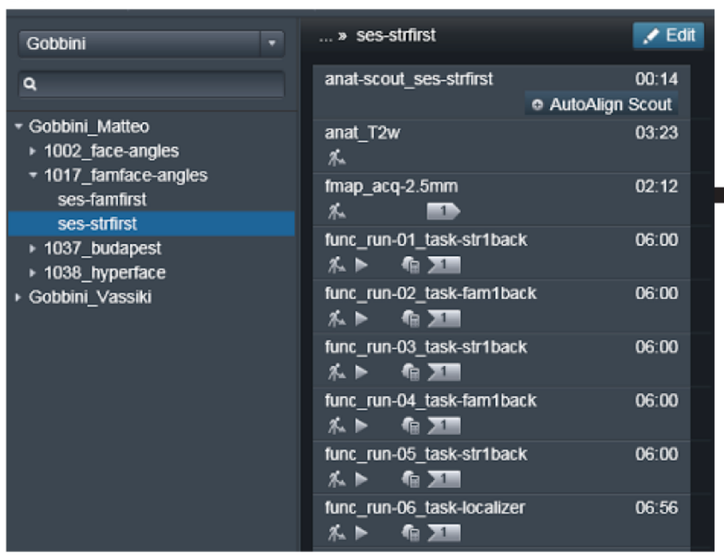
Current usage at Dartmouth Brain Imaging Center

1.5 TB of data converted (DICOM + BIDS)

40 datasets for 13 PIs + weekly QA

http://datasets.datalad.org/?dir=/dbic/QA

Scanner



DICOM

```
001-anat-scout_ses-strfirst
005-anat_T2w
  000001.dcm
  000002.dcm
  ...
006-fmap_acq-2.5mm
007-fmap_acq-2.5mm
008-func_run-01_task-str1back
011-func_run-01_task-str1back
018-func_run-02_task-fam1back
025-func_run-03_task-str1back
032-func_run-04_task-fam1back
039-func_run-05_task-str1back
046-func_run-06_task-localizer
```

$ heudiconv

DataLad

BIDS

```
anat
  sub-sid000005_ses-strfirst_T2w.json
  sub-sid000005_ses-strfirst_T2w.nii.gz
fmap
  sub-sid000005_ses-strfirst_acq-25mm_magnitude1.json
  sub-sid000005_ses-strfirst_acq-25mm_magnitude1.nii.gz
  ...
func
  sub-sid000005_ses-strfirst_task-fam1back_run-02_bold.json
  sub-sid000005_ses-strfirst_task-fam1back_run-02_bold.nii.gz
  sub-sid000005_ses-strfirst_task-fam1back_run-02_events.tsv
  ...
sub-sid000005_ses-strfirst_scans.tsv
```

```
$ git grep TODO

CHANGES:TODOs:
README:TODO: Provide description for the dataset ...
dataset_description.json: "Acknowledgements": "TODO...",
dataset_description.json:   "TODO:",
dataset_description.json: "DatasetDOI": "TODO: ...
task-fam1back_bold.json:  "CogAtlasID": "TODO",
task-fam1back_bold.json:  "TaskName": "TODO: full task name",
task-localizer_bold.json:  "CogAtlasID": "TODO",
task-localizer_bold.json:  "TaskName": "TODO: full task name",
task-str1back_bold.json:  "CogAtlasID": "TODO",
task-str1back_bold.json:  "TaskName": "TODO: full task name",
...
```

Figure 2. Example of the naming scheme at the scanner console, the resulting DICOM folder structure, and the automatically converted BIDS dataset. The user can quickly find what needs to be added by executing a simple `git grep TODO`, since the entire dataset is under git and DataLad control.

Key components

1. Naming scheme at the scanner

We defined a consistent naming scheme for subjects (anonymized), studies and sequences at the scanner (see Figure 2). This naming scheme was created to flexibly accommodate all use-cases at the Dartmouth Brain Imaging Center (DBIC).

2. Heuristic for HeuDiConv

We created a heuristic definition for HeuDiConv to incrementally convert collected data from DICOM into BIDS without human intervention. The heuristic automates

- identification of the location for a particular accession within the hierarchy of studies
- identification of the session for multi-session studies
- identification and annotation of the canceled runs so they could later be reviewed and removed from the study dataset repository, while allowing to revert back in case of mistakes thanks to git/DataLad.

The automation eliminates manual interactions with the acquired data during conversion, and only requires the user to validate the dataset using the BIDS validator at the end to detect possible anomalies.

3. DataLad

DataLad provides a system for version control, meta-data annotation, and access to the data. DataLad allows to

- incrementally update local clones of the dataset as more data comes in, while incorporating changes done locally (e.g., to BIDS metadata files such as `dataset_description.json`, `*_events.tsv`)
- annotate files with possibly sensitive information as non-distributable
- flexibly fetch or upload portions of the dataset to the processing machines (e.g., only anatomicals for FreeSurfer parcellation)
- incorporate acquired datasets into a larger study, lab or institutional repository as git sub-modules for version control and provenance

Work in progress: come help us!

- anonymization of conversion date recorded in git/git-annex
- configurable customizations
  - to provide ability to map names for studies not following ReproIn convention
  - to re-create datasets from scratch using stored DICOMs
- automated monitoring of new data and conversion
- automated reslicing of electrophysiology data in BIDS format

Resources

http://reproin.repronim.org
Singularity/Docker definition files; walkthrough; documentation

http://datalad.org
Management and version control for data, code, containers

https://github.com/nipy/heudiconv
Flexible DICOM converter based on heuristics

References

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., ... Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Scientific Data.

Halchenko, Y. O., & Hanke, M. (2015). Four aspects to make science open "by design" and not as an after-thought. GigaScience.

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PloS One.

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PloS One.

Funding

OHBM Merit Abstract Award, Neukom Institute Travel Grant, Dartmouth Graduate Student Council Travel Grant, Dartmouth Center for Cognitive Neuroscience, NIH #1P41EB019936-01A1 (ReproNim), NSF #1429999 (DataLad)