



The "*My data organization is as good as yours*" fallacy

MaRBEL meeting
February 2025



Christophe Phillips
c.phillips@uliege.be



Program

- ▶ Data acquisition & processing
- ▶ Issues with “my” data organization
- ▶ Brain Imaging Data Structure, aka. BIDS
- ▶ Take home message

Data workflow...



Acquisition:

- #subjects
- #modalities
- #sessions/visits



Processing:

- Spatial processing
- Statistical analysis

“Simplest” fMRI project:

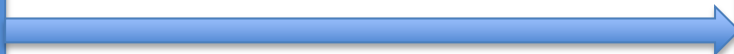
- ▶ 20-40 subjects: young & healthy + demographics (age, sex, handedness)
- ▶ 1 fMRI session + 1 anatomical MRI
- ➔ 1h/subject & data acquisition by 1 person in 1 month
- ➔ “simple” processing by 1 person for 1 paper

Data workflow...



Acquisition:

- #subjects
- #modalities
- #sessions/visits



Processing:

- Spatial processing
- Statistical analysis

“Real” MRI project:

- ▶ 20-200 subjects: thorough phenotype + full neuropathology evaluation + ...
- ▶ Several modalities: MRIs, actimetry, EEG, blood/saliva,...
- ▶ Several sessions/visits: over days to months
- ➔ >6h/subject & data acquisition by many persons over >1 year
- ➔ **data heterogeneity & asynchronous acquisition !**

My great MRI project...



Experiment time line (excl. ethics, insurance, funding, recruiting,...):

- ▶ Devise experimental protocol (stimuli/conditions, #groups, #subjects, etc.)
- ▶ Scan subjects, over some time: weeks, months, years
- ▶ Accumulate data on disk à la "**My data organization is as good as yours**"
- ▶ Process data, i.e. create processed data on disk
(+ potentially mess up original data)
- ▶ Publish, i.e. reformat *some* of the results (& keep/forget the others)
- ▶ “Move on and forget about it!”
(where is what?, what was published exactly?, how was it obtained?)

What next ? Reproducible results ? Reusable data ?

Efficient data processing



Everything is scripted!

- ➔ Need to organize data/metadata
- ▶ Data selection through “filters”
 - Loop over subjects → select subject specific data
 - Modality specific steps → select specific modalities
- ▶ All parameters findable & accessible
- ▶ Preserve original data, raw or derived at previous step

Data/metadata MUST be carefully organized & labelled,
i.e. human and computer readable

Data workflow...



“Real” MRI project:

- ▶ 20-200 subjects: thorough phenotype + full neuropathology evaluation + ...
- ▶ Several modalities: MRIs, actimetry, EEG, blood/saliva,...
- ▶ Several sessions/visits: over days to months
- ➔ >6h/subject & data acquisition by many persons over >1 year
- ➔ **data heterogeneity & asynchronous acquisition !**



Program

- ▶ Data acquisition & processing
- ▶ Issues with “my” data organization
- ▶ Brain Imaging Data Structure, aka. BIDS
- ▶ Take home message

Raw NIfTI MRI data



- ▶ All files in one folder, `.nii` and `.json`
- ▶ One **scanner Id** per acquisition session
- ▶ Protocols identified by “**Series number**”
→ unreliable and unclear!
- ▶ Select **series of images** manually or through “filters” on indexes
→ error prone!
- ▶ Processing output in the same folder (sometimes) or files modified!
→ messy and risky!

s02438/			
nii			
f2438	0004	00001	000001-01.json
f2438	0004	00001	000001-01.nii
f2438	0004	00002	000002-01.json
f2438	0004	00002	000002-01.nii
f2438	0004	00003	000003-01.json
f2438	0004	00003	000003-01.nii
f2438	0004	00004	000004-01.json
f2438	0004	00004	000004-01.nii
f2438	0004	00005	000005-01.json
f2438	0004	00005	000005-01.nii
f2438	0004	00006	000006-01.json
f2438	0004	00006	000006-01.nii
f2438	0004	00007	000007-01.json
f2438	0004	00007	000007-01.nii
f2438	0004	00008	000008-01.json
f2438	0004	00008	000008-01.nii
f2438	0004	00009	000009-01.json
f2438	0004	00009	000009-01.nii
f2438	0004	00010	000010-01.json
f2438	0004	00010	000010-01.nii
f2438	0005	00001	000001-01.json
f2438	0005	00001	000001-01.nii
f2438	0005	00002	000002-01.json
f2438	0005	00002	000002-01.nii
f2438	0005	00003	000003-01.json

Rename and organize the data “my way”



- ▶ Subjects' label and indexing ? → s23, HC23/AD18, ...
 - ▶ Folder for each modality ?
 - functional MRI → fmri, func, funcMRI, functional,...
 - anatomical MRI → amri, anat, struct, sMRI,...
 - ▶ Image metadata ?
 - Which parameters ? E.g. echo & repetition time, slice order/time,...
 - Which units ? Seconds or milliseconds ?
 - Which name ? TE/TR, EchoT/RepT, Techo/Trepetition,...
- ➔ From raw JSON file, “on the fly” vs. “extract and save aside” ?

Rename and organize the data “my way”



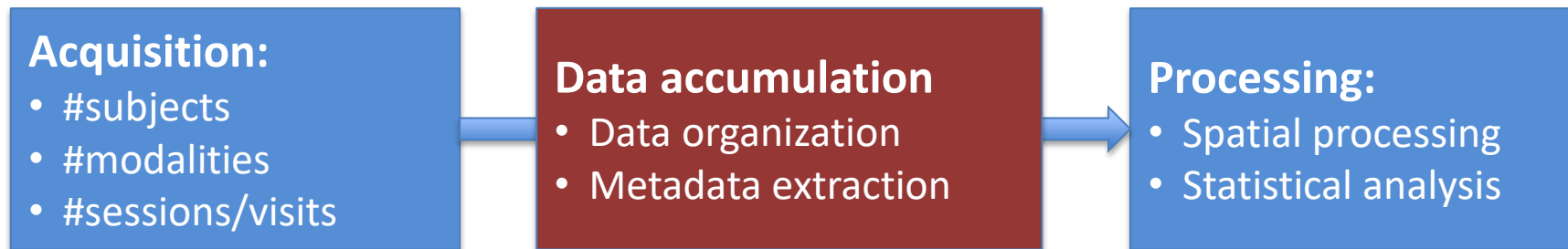
- ▶ Other data,
 - demographic & behavioural information
 - stimuli timing & responses in fMRI
 - b-value & vectors in DW-MRI
- ▶ Saved/stored in
 - Distinct Excell files ? On “another computer” ?
 - PhD/postdoc/PI/lab-assistant’s head ?
- ➔ Do we have ALL the information ? Unique and clear ?
- ➔ Script must be tailored to “my way” organization
- ➔ Difficult to share/reuse across datasets & people

“My organization” wish...



- ▶ **Between colleagues**
 - Similar dataset structure with small adjustments
 - “Quick & dirty” code to experiment
 - “Clean & documented” code for regular/final processing
 - Well-defined (& relative) path and file names
 - All metadata extracted from dataset, i.e. “one place”
- ▶ **Between institutions/open source**
 - Some more documentation
 - Limited or no hard-coded paths
 - Issues/bugs follow up
 - Increased flexibility for (local) data specificities

Data workflow...



Need to

- ▶ know what data **and metadata** are needed, e.g. acquisition parameters for all data, task description and subject's responses, subjects' parameters,...
- ▶ convert/extract data and metadata → explicit & easy to find

How to organize and name all these ? **Define a nomenclature!**



Program

- ▶ Data acquisition & processing
- ▶ Issues with “my” data organization
- ▶ Brain Imaging Data Structure, aka. BIDS
- ▶ Take home message

Some IT wisdom



"DRY - Don't Repeat Yourself –

Every piece of knowledge must have a single, unambiguous, authoritative representation within a system."

- [Andy Hunt](#) & [Dave Thomas](#)

"Data dominates.

If you've chosen the right data structures and organized things well, the algorithms will almost always be self-evident. Data structures, not algorithms, are central to programming."

– [Rob Pike](#) in 1989

Brain Imaging Data Structure



- ▶ Community effort
 - Started in 2015
 - Current version 1.10.0
- ▶ Human readable
 - Minimized curation
 - Error checking & reduction
- ▶ Computer readable
 - Optimized usage of data
 - Analysis software
 - Development of automated tools


scientific **data**

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

[nature](#) > [scientific data](#) > [articles](#) > article

Article | [Open access](#) | Published: 21 June 2016

The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

[Krzysztof J. Gorgolewski](#) , [Tibor Auer](#), [Vince D. Calhoun](#), [R. Cameron Craddock](#), [Samir Das](#), [Eugene P. Duff](#), [Guillaume Flandin](#), [Satrajit S. Ghosh](#), [Tristan Glatard](#), [Yaroslav O. Halchenko](#), [Daniel A. Handwerker](#), [Michael Hanke](#), [David Keator](#), [Xiangrui Li](#), [Zachary Michael](#), [Camille Maumet](#), [B. Nolan Nichols](#), [Thomas E. Nichols](#), [John Pellman](#), [Jean-Baptiste Poline](#), [Ariel Rokem](#), [Gunnar Schaefer](#), [Vanessa Sochat](#), [William Triplett](#), ... [Russell A. Poldrack](#) [+ Show authors](#)

[Scientific Data](#) **3**, Article number: 160044 (2016) | [Cite this article](#)

70k Accesses | 783 Citations | 105 Altmetric | [Metrics](#)

Brain Imaging Data Structure



- ▶ Fixed specific file & (sub)folder naming
- ▶ Fixed specific file organization in subfolders
- ▶ Complete representation of data set, incl.
 - Experimental design & project information
 - Subject specific information
 - Data types & acquisition parameters
 - Original raw data, intermediate/derivative data & final results
 - Processing information & derived data/results
 - Ownership & references

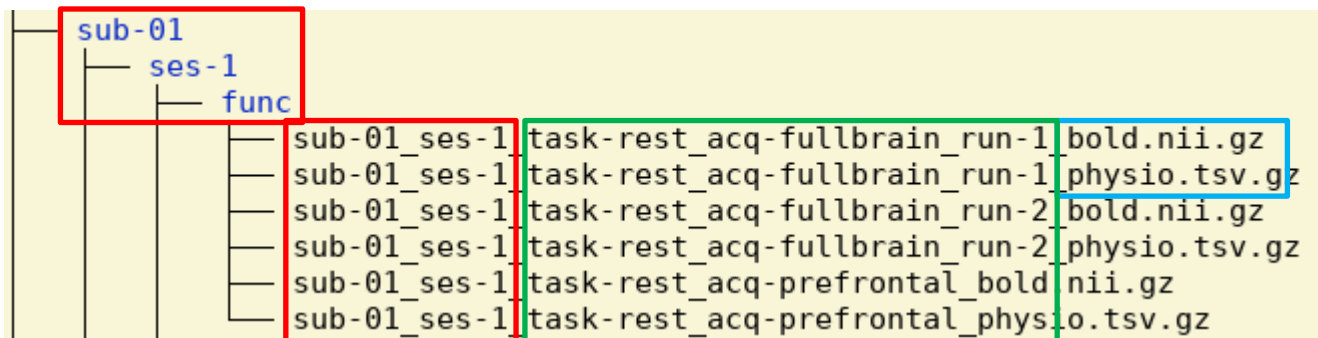
BIDS goal



With only the data and metadata in `>MyExperiment` folder, be able to

- **understand** the whole experiment & data,
- **check** data consistency,
e.g. same acquisition parameters, missing modality in 1 subject,...
- **reprocess** (automatically) the whole dataset,
e.g. generate all the results from my papers or test a new tool,
- **reuse** any part of the data, e.g. for another project, and/or **share** with others.

BIDS naming principles



► Directories:

- *sub-**<label>***: per subject
- *ses-**<label>***: per session (optional)
- ***<data type>***: group of different types of data

► Names:

- *sub-**<label>*** and *ses-**<label>***
- ***<suffix>***: modality ("kind" of image)
- ***<entity>-**<label>*****: acquisition parameters or properties of image(s)

BIDS data files



All images → NIfTI files, `.nii`

3D volume or 4D for “series”, e.g. fMRI and DWI

(actually, zipped `.nii` are preferred)

Meta-data

- ▶ array → “tab-separated value” files, `.tsv`
- ▶ key/value pair → JSON files, `.json`
- ▶ b-values/vectors → text file, `.bvec/.bval`
- ▶ “description” files → text & Markdown file, `Readme.txt`, `changes.txt`, `description.md`

All open-format file types !

BIDS, metadata files



Key name	Requirement Level	Data type	Description
EchoTime	RECOMMENDED, but REQUIRED if corresponding fieldmap data is present, or the data	number or array of numbers	The echo time (TE) for the acquisition, specified in seconds. Corresponds to DICOM Tag 0018, 0081 Echo Time (please note that the DICOM term is in milliseconds not seconds). The data

- ▶ Stored in JSON file
- ▶ Strict definition: conventions and units
- ▶ Requirement levels:
 - *REQUIRED*: needed to interpret data
 - *RECOMMENDED*: will improve interpretation
 - *OPTIONAL*: might be useful

```
{  
  "CogAtlasID": "https://www.cognitiveatlas.org/id/trm_4c8a834779883",  
  "EchoTime": 0.017,  
  "EffectiveEchoSpacing": 0.0003333262223739227,  
  "PhaseEncodingDirection": "j-",  
  "RepetitionTime": 3.0,  
  "SliceEncodingDirection": "k",  
}
```

BIDS, Modality agnostic (top-level) files



- ▶ Participants description, `participants.tsv/.json`

participant_id	sex	age	number	handedness
sub-01	F	29	17	100
sub-02	F	23	6	100
sub-03	M	25	18	86
sub-04	M	26	8	100

.json file describes each variable (unit, range, possible value,...)

- ▶ Dataset description, `dataset_description.json`

```
{
  "Name": "Processed MPM qMRI aging data",
  "BIDSVersion": "1.8.0",
  "DatasetType": "derivative",
  "Authors": [
    "Martina F. Callaghan",
    "Christophe Phillips"
  ],
  "Acknowledgements": "Elaine Anderson, Marinella Cappelletti, Rumana Chowdhury, Joern Diedirchsen, Thomas H. B. Fitzgerald, and Peter Smittenaar as part of multiple cognitive neuroimaging studies performed at the Wellcome Centre for Human Neuroimaging",
  "HowToAcknowledge": "Please cite this paper: https://doi.org/10.1016/j.neurobiolaging.2014.02.008",
  "ReferencesAndLinks": [
    "https://doi.org/10.1016/j.neurobiolaging.2014.02.008",
    "Callaghan et al., Neurobiology of Aging, 2014."
  ],
}
```

BIDS Extension Proposals



Various modalities, **same principles**

- ▶ Magnetoencephalography (MEG) – 2018, Sci Data 5, 180110 (2018)
- ▶ Electroencephalography (EEG/iEEG) – 2019, Sci Data 6, 102 & 103 (2019)
- ▶ Positron emission tomography (PET) – 2022, Sci Data 9, 65 (2022)
- ▶ Quantitative MRI (qMRI) – 2022, Sci Data 9, 517 (2022)
- ▶ Microscopy – 2022, Front Neurosci, 16 (2022)
- ▶ Magnetic Resonance Spectroscopy (MRS) – 2024, (publication soon)
- ▶ Derivatives – “work in progress”.

https://bids.neuroimaging.io/get_involved.html

BIDS development & adoption



BIDS is community driven and broadly accepted

- ▶ Clear use cases & solving a common end user problem
- ▶ Low technical barrier to entry
- ▶ Maturity and the size of the field (30 years of NI)
- ▶ Open doors without “death by consensus”

But...

- ▶ Data conversion remains challenging!
- ▶ Lack of a (complete) machine-readable standard
- ▶ Challenges of “BIDS Extension Proposals” management



Program

- ▶ Data acquisition & processing
- ▶ Issues with “my” data organization
- ▶ Brain Imaging Data Structure, aka. BIDS
- ▶ Take home message

Human perspective



▶ **BAD (or not so convenient) day**

- Considerable effort to organize data
- Sometimes confusing and contradictory descriptions
- Need to integrate all acquisition data
- **Need careful planning before acquiring data**

▶ **GOOD day**

- Easy to retrieve information
- Easy to run pipelines
- Easy to share data, e.g. your colleague (or yourself in 2 years)

Computer perspective



▶ **GOOD day**

- Easy to retrieve data and metadata
`bids-matlab`, `pybids` – query based data retrieval
- Easy to patch errors
- Easy to write pipelines
`qmri`, `fmriprep` – query based data retrieval
- Modular composition (“BIDS in, BIDS out”)

▶ **NOT SO GOOD day**

- Rare case of missing metadata → improvise & patch
- Cases of modalities not included in BIDS → improvise
- No strict regulation of pipeline outputs (derivatives) → improvise

Take home message



Data curation is a pain, really.
...but it saves you from more pain later on !

Think BIDS & open data *by design*.

References

- ▶ BIDS specifications → “what & how to”, <https://bids.neuroimaging.io/>
- ▶ BIDS data → “example & re-use”, <https://openneuro.org/>
- ▶ BIDS-fication tool → “let’s do it for real”,
 - Dcm2bids (only MRI), <https://cdnis-brain.readthedocs.io/dcm2bids/>
 - BIDScoin (mostly imaging), <https://bidscoin.readthedocs.io/en/stable/>
 - BIDSme (multi-modal), <https://github.com/CyclotronResearchCentre/bidsme>

Thank you for your attention!

