

From Chaos to Structure: Designing MR Datasets for Reuse and Longevity

ESMRMB
October 2026



Program

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

Data workflow... “Simplest” fMRI project



Acquisition:

- #subjects
- #modalities
- #sessions/visits



Processing:

- Spatial processing
- Statistical analysis

- ▶ 20 subjects: young & healthy + demographics (age, sex, handedness)
- ▶ 1 fMRI session + 1 anatomical MRI
- ➔ 1h/subject & data acquisition by 1 person in 1 month
- ➔ “simple” workflow from data acquisition to results
- ➔ easy to understand and manage

Data workflow... “Real” MRI project



Acquisition:

- #subjects
- #modalities
- #sessions/visits



Processing:

- Spatial processing
- Statistical analysis

- ▶ 50-200 subjects: thorough phenotyping + full neuropsych evaluation + ...
 - ▶ Several modalities: MRIs, actimetry, EEG, blood/saliva,...
 - ▶ Several sessions/visits, over days to months
-
- ➔ data acquisition by N persons over months to year
 - ➔ complex workflow from data acquisition to results
 - ➔ complicated to understand and hard to manage

Efficient data processing



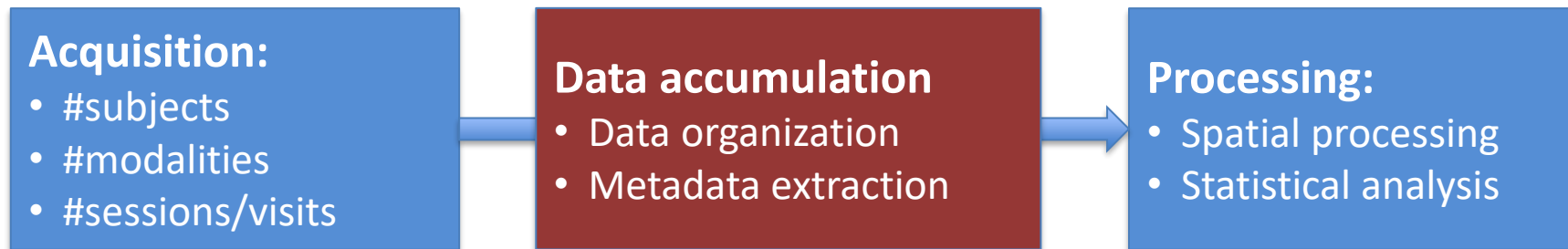
Everything is scripted!

- ➔ Need to organize data ...and metadata
- ▶ Data structured unequivocally
- ▶ All parameters/metadata findable & accessible
- ▶ Preserve original data, raw or derived at previous step

Data & metadata MUST be carefully curated,

i.e. human and computer readable (structure & names)

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 ?



Program

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



- Common format for different manufacturers but...**one file per slice!**
- 100's of official metadata fields
- ...but **manufacturer and protocol private fields!**

- Open-format 3D and 4D (multi-volumes) data,
- ...but **(almost) no acquisition metadata stored**
- Full DICOM header can be dumped into sidecar `.json` file, incl. cryptic manufacturer and protocol private fields

[illegible]

Raw NIfTI MRI data... in one folder



What is what ?

- ▶ One **scanner Id** per acquisition session
→ issue if multiple sessions!
- ▶ Protocols identified by “**Series number**” & **metadata**
→ unreliable and unclear!
- ▶ Select **series of images** on indexes,
manually or through “filters”
→ error prone!

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			

- ➔ From the raw .json file ?

[illegible]

Do we have ALL the information ?



► Other data

- demographic & behavioural information
- stimuli timing & responses in fMRI
- b-value & vectors in DW-MRI
- ...

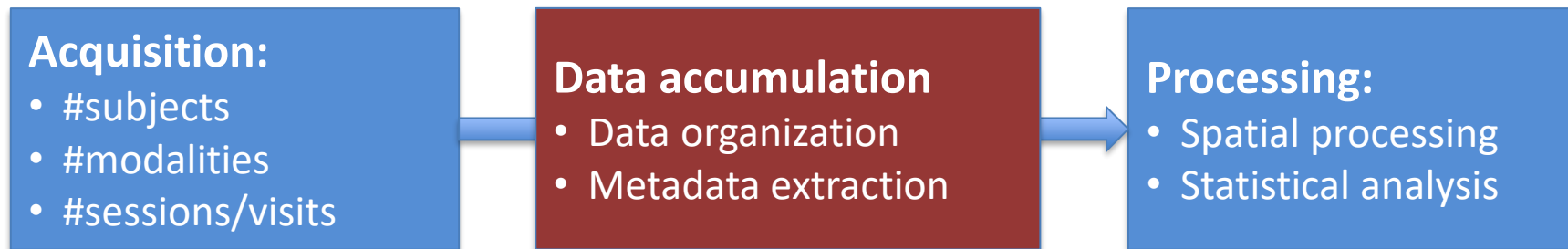
► Saved/stored in

- Distinct Excell files ? Another computer or other USB-drive ?
- PhD/postdoc/PI/lab-assistant's head ?

➔ Processing script must be tailored to “my data”

➔ Difficult to share/reuse across datasets & people

Data workflow...



Need to

- ▶ know what data **and metadata** are needed,
- ▶ convert/extract data and metadata → explicit & easy to find

How to organize and name all these ?

Define a nomenclature → rely on BIDS



Program

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

Brain Imaging Data Structure



- ▶ Community effort
 - Started in 2015
 - Current version 1.11.1
- ▶ 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.
 - Data types & acquisition parameters
 - Subject specific information
 - Experimental design & project information
 - Original raw data, intermediate/derivative data & final results
 - Processing information & derived data/results
 - Ownership & references

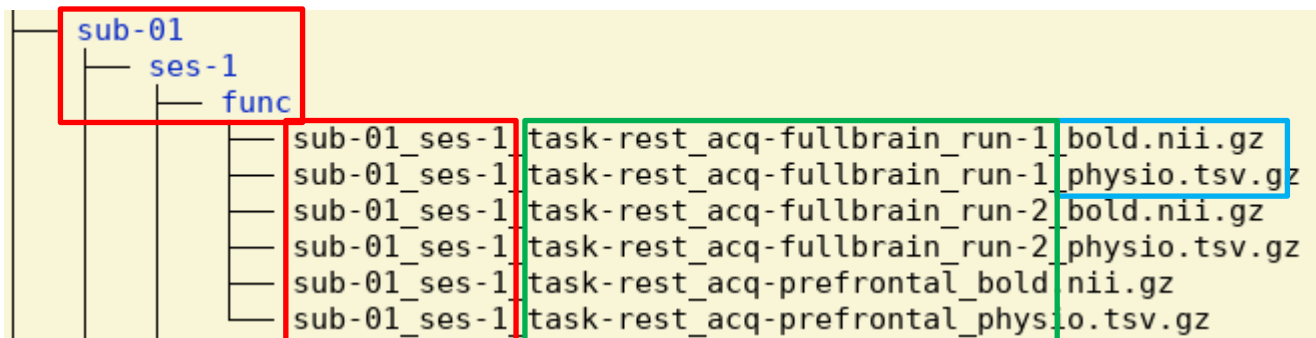
BIDS goal



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 or data)
- ***<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



► Stored in JSON file

► Strict definition: conventions and units

► Requirement levels:

- *REQUIRED* → needed to interpret data
- *RECOMMENDED* → will improve interpretation
- *OPTIONAL* → might be useful

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

```
{  
  "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 Diedrichsen, 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 (2018)
- ▶ Electroencephalography (EEG/iEEG) – 2019, Sci Data 6 (2019)
- ▶ Positron emission tomography (PET) – 2022, Sci Data 9 (2022)
- ▶ Quantitative MRI (qMRI) – 2022, Sci Data 9 (2022)
- ▶ Microscopy – 2022, Front Neurosci, 16 (2022)
- ▶ Magnetic Resonance Spectroscopy (MRS) – 2025, Sci Data 12 (2025)
- ▶ Derivatives – “work in progress”
- ▶ ...

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



Program

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

Take home message



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

Think BIDS & open data *by design*
...so plan your data structure before acquiring data !

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>



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

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

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`, `fmripipeline` – 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*