

BIDS: My data organization is as good as yours

ULiège Open Science Day

Christophe Phillips¹ Nikita BELIY¹

¹GIGA CRC, Liège University, Belgium

November 6, 2024



Introduction I: Data acquisition

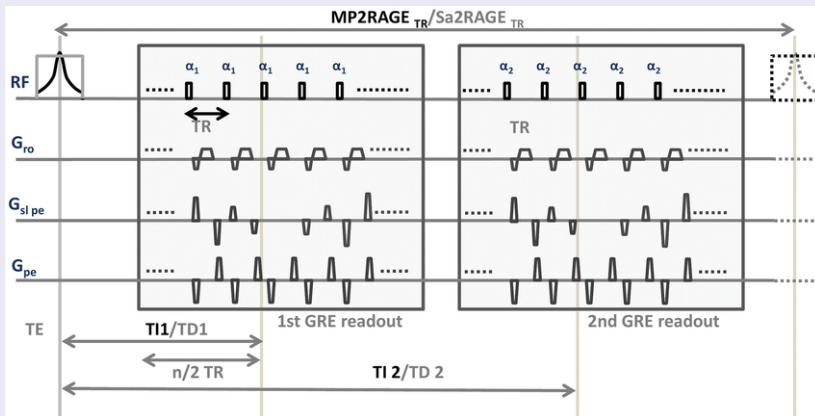
MRI acquisition

- MRI principles
 - B_0 Proton spin alignment along z -axis
 - Repeated RF pulses to move proton spins out of equilibrium
 - Disturbed protons generate a electro-magnetic field used to reconstruct image
- Pulse parameters and sequences define different modalities
 - anatomical images
 - functional images (time-dependent blood oxygenation)
 - diffusion images (water molecules)
- Acceleration and noise suppression techniques add complexity

Knowledge of parameters needed to interpret images

MRI acquisition

MP2RAGE sequence



Neuroimage. 2010 Jan 15;49(2):1271-81

DICOM data

Too much metadata

<https://www.dicomstandard.org/current>

- Created in 80s
- File format and TCP/IP protocols
- Common format for different manufacturers
- Format of out-of scanner raw image

Difficulties

- Image and metadata manipulation
- Hundreds of defined metadata fields
- Manufacturer private fields
- Protocol private fields

NIfTI

Not enough metadata

<https://nifti.nimh.nih.gov/>

A (Sort of) New Image Data Format Standard: NIfTI-1

Robert W Cox¹, John Ashburner², Hester Breman³, Kate Fissell⁴, Christian Haselgrove⁵, Colin J Holmes⁶, Jack L Lancaster⁷, David E Rex⁸, Stephen M Smith⁹, Jeffrey B Woodward¹⁰, Stephen C Strother¹¹

¹SSCC/NIMH/NIH/DHHS/Bethesda, ²FIL/London, ³Brain Innovation/Maastricht, ⁴U Pittsburgh/Pittsburgh, ⁵MGH/Charlestown, ⁶SIG/Mountain View, ⁷RIC/UTHSCSA/San Antonio, ⁸LONI/UCLA/Los Angeles, ⁹FMRIB/Oxford, ¹⁰Dartmouth College/Hanover, ¹¹U Minnesota/Minneapolis and NIFTI-DFWG Chair

- **NIFTI** = **N**euroimaging **I**nformatics **T**echnology **I**nitiative
 - NIH-sponsored working group to promote interoperability of functional neuroimaging software tools
- **DFWG** = **D**ata **F**ormat **W**orking **G**roup within NIFTI to deal with **data** interoperability
 - e.g., make it easier to interchange image (etc.) data between analysis packages
 - Near-term efforts: extend ANALYZE™-7.5 file format (**.hdr/.img** file pairs) to add features the DFWG agreed were highly desirable for fMRI analysis = **the NIFTI-1 format**
 - New features fit into unused/little-used ANALYZE fields

Current Status

- DFWG has approved NIFTI-1 format
- Major software packages (AFNI, BrainVoyager, FSL, SPM) agree to *read* NIFTI-1 files by **July 31, 2004** and to be able to *write* them by **Dec 31, 2004**
- NIFTI-1 specification is in the form of a very heavily commented C header file, laying out the fields and their interpretations:

<http://nifti.nimh.nih.gov/dfwg/>

OHBM 2004

- (Almost) no acquisition metadata stored
- Metadata dumped at conversion into JSON file

Introduction II: Data usage

Naive data organisation

CRC experience

- Subject's files in one folder
 - Not always
- Can be thousands of files in folder
 - In complex studies
- Protocols identified by "Series number"
 - Unreliable and unclear
- Selection by regexp or manually
 - Error prone
- Processing output in the same folder (sometimes)

```
s /
├── nii
│   ├── f -0004-00001-000001-01.json
│   ├── f -0004-00001-000001-01.nii
│   ├── f -0004-00002-000002-01.json
│   ├── f -0004-00002-000002-01.nii
│   ├── f -0004-00003-000003-01.json
│   ├── f -0004-00003-000003-01.nii
│   ├── f -0004-00004-000004-01.json
│   ├── f -0004-00004-000004-01.nii
│   ├── f -0004-00005-000005-01.json
│   ├── f -0004-00005-000005-01.nii
│   ├── f -0004-00006-000006-01.json
│   ├── f 0004 00006-000006-01.nii
│   ├── f 0004 00007-000007-01.json
│   ├── f 0004 00007-000007-01.nii
│   ├── f 0004 00008-000008-01.json
│   ├── f 0004 00008-000008-01.nii
│   ├── f 0004 00009-000009-01.json
│   ├── f 0004 00009-000009-01.nii
│   ├── f 0004 00010-000010-01.json
│   ├── f 0004 00010-000010-01.nii
│   ├── f 0005 00001-000001-01.json
│   ├── f 0005 00001-000001-01.nii
│   ├── f 0005 00002-000002-01.json
│   ├── f 0005 00002-000002-01.nii
│   └── f 0005 00003-000003-01.json
```

Sharing data

Running own pipeline on foreign data

- Dataset under their hand
- No or minimal effort to organize/document
- Often raw data without any organization
- Difficulty to obtain additional information

Sharing data

Running own pipeline on foreign data

- Dataset under their hand
- No or minimal effort to organize/document
- Often raw data without any organization
- Difficulty to obtain additional information

Personal experience (pathological cases)

- Non-organized data
 - Just one big folder with all files
- Converted data without any metadata
- Different protocol from what is claimed

Sharing code

Running foreign pipeline on own data

- Between colleagues
 - Similar dataset structure with small adjustments
 - **Low quality code** for personal use
 - **Little to no documentation**
 - **Hard-coded paths and file names** in surprising places
 - **Cryptic metadata retrieval**
even author don't remember what and why
 - **Hard-coded metadata**

Sharing code

Running foreign pipeline on own data

- Between colleagues
 - Similar dataset structure with small adjustments
 - **Low quality code** for personal use
 - **Little to no documentation**
 - **Hard-coded paths and file names** in surprising places
 - **Cryptic metadata retrieval**
even author don't remember what and why
 - **Hard-coded metadata**
- Between institutions/open source
 - Some effort on documentation
 - Limited or no hard-coded paths
 - Issues/bugs follow up
 - **Maybe incompatible data structure**

BIDS: Brain Imaging Data Structure

BIDS

Brain Imaging Data Structure

<https://bids.neuroimaging.io>

- Community effort
 - Started in 2015
 - Current version 1.10.0
- Human readable
 - Minimized curation
 - Error 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](#)

Sci Data **3**, 160044 (2016)

File naming rules

Findable

GitHub: [bids-examples](#)

```
sub-01
├── ses-1
│   └── func
│       ├── sub-01_ses-1_task-rest_acq-fullbrain_run-1_bold.nii.gz
│       ├── sub-01_ses-1_task-rest_acq-fullbrain_run-1_physio.tsv.gz
│       ├── sub-01_ses-1_task-rest_acq-fullbrain_run-2_bold.nii.gz
│       ├── sub-01_ses-1_task-rest_acq-fullbrain_run-2_physio.tsv.gz
│       ├── sub-01_ses-1_task-rest_acq-prefrontal_bold.nii.gz
│       └── sub-01_ses-1_task-rest_acq-prefrontal_physio.tsv.gz
```

Directories:

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

Names:

- *<suffix>*: defines modality ("kind" of image)
- *<entity>-<label>*: defines acquisition parameters of image

Metadata definitions

Interoperable

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

Metadata definitions

Interoperable

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

Sidecar JSON file

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

Modality agnostic (top-level) files

Reusable

- *dataset_description.json*
 - Dataset name, BIDS version, authors, DOI, etc...
- *README(.md, .txt, .rts)*
 - Free text detailed description and notes on dataset
- *CITATION.cff/CHANGES/LICENSE*

Modality agnostic (top-level) files

Reusable

- *dataset_description.json*
 - Dataset name, BIDS version, authors, DOI, etc...
- *README(.md, .txt, .rts)*
 - Free text detailed description and notes on dataset
- *CITATION.cff/CHANGES/LICENSE*

participants.tsv – participants description

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

Expandable to other data types

BIDS Expansion Proposals

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

- Magnetoencephalography (MEG) – 2018
Sci Data 5, 180110 (2018)
- Electroencephalography (EEG/iEEG) – 2019
Sci Data 6, 103 (2019), Sci Data 6, 102 (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)
- Near-Infrared Spectroscopy (NIRS) – 2023
PsyArXiv. doi:10.31219/osf.io/7nmcp
- Magnetic Resonance Spectroscopy (MRS) – 2024
(publication forthcoming)

Conclusion:
Why it works

Human perspective

Worst thing ever!

- Considerable effort to organize data
- Sometimes confusing and contradictory descriptions
- Need to integrate all acquisition data

Human perspective

Worst thing ever!

- Considerable effort to organize data
- Sometimes confusing and contradictory descriptions
- Need to integrate all acquisition data

Best thing ever!

- Easy to retrieve information
- Easy to run pipelines

Computer perspective

Best thing ever!

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

Computer perspective

Best thing ever!

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

Worst thing ever!

- Rare case of non-BIDS metadata
- Cases of modalities not included in BIDS
- No strict regulation of pipeline outputs (derivatives)

Thanks for your attention!