# Introduction to BIDS

Dragan Rangelov

# Share the love

- You can find all materials for this event on GitHub

github.com/draran/BIDS_tutorial

# Brain Imaging Data Structure - BIDS

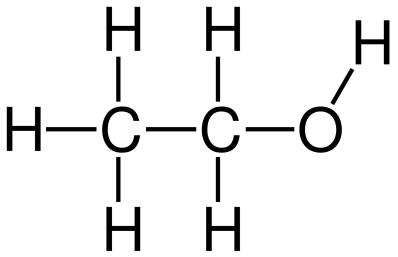
- **bids-specification.readthedocs.io**
- The current stable version is 1.4.0
- Specifies standards for:
  - File naming
  - Organising folders
  - Recording metadata
- Outcomes:
  - Increases reproducibility
  - Facilitates re-use of data


ZK
14

# Broader context – community perspective

- Facilitates open science
  - Open exchange and re-use of ideas, methods, data and analyses algorithms
  - With appropriate citations (you can cite datasets, or you will be able to)
- Using a well-defined vocabulary
- Similar to the naming system in organic chemistry
  - Name defines the nature of the named entity
  - And implies many of its properties

ETHANOL - booze
Two carbon atoms = ETH
Single link between them = AN
Hydroxyl group = OL



# Narrow context – individual/lab perspective

- PhD students
  - Don’t have to re-invent the wheel (s001_eeg.txt, vp_ABX_eye.tsv.gz, etc.)
- Post-docs
  - Can substantially speed-up their workflow by using automated pipelines
- Lab heads
  - Can hire a new person with minimal costs in knowledge-transfer
- All
  - Can quickly archive data and analyses and share them with the community
  - Can re-use freely (as they already do), but in a well-defined way

# Bad stuff

- Takes time to get familiar with
- Requires change in the existing workflows
  - Minimal for PhDs
  - Substantial for post-docs
  - Lab heads – well …
- There are still no standards for some types of data/data analyses
  - Heavy focus on “classic” neuroimaging – fMRI
  - But, there are groups developing analogous standards for all sorts of data

# How does it work

# What's in a name? All

- Folder structure:
  - Source data – the original data, frozen in time
  - Raw data – unprocessed data, transformed in a more convenient data format
  - Derivatives – all analysed data
  - Scripts – data analyses scripts
  - Export – my addition, location to export figures, tables, etc., to be used in publications (manuscript, presentations, posters)
- README.txt – free form description of the dataset
- CHANGES.txt – history of changes to the dataset
- dataset_description.json - structured description of the dataset

# Short primer on JSON data format

- JavaScript Object Notation (or Jason)
- Language-independent format, a text file with ”.json” extension
- That is organised in a specific way:
  - Attribute-value pairs, “task”: “Stroop”
- Such files can be easily parsed by most (all?) programming languages
- And read in as a language-specific data structure:
  - MATLAB – structure
  - Python – dictionary
  - R - data frames (lists, too)

# JSON data format

- What you cannot do with JSON
  - Cannot evaluate statements, e.g., x = cos(theta)
  - Cannot check for syntax errors (easily)
  - Because – it's a simple text file
- What you can do with JSON
  - Store anything that can be converted to strings (which is a lot/if not all)
  - Can organise structured metadata easily

# Raw data

- Modality-specific data
  - Will depend on the nature of the acquired data
  - Behaviour, fMRI, EEG/MEG, eye-movements, questionnaires, …
- Metadata – json files
  - fMRI – hardware specifics, sequence parameters, etc. …
  - EEG - hardware specifics (filters, sampling frequency, manufacturer, etc.)
  - You get the idea
- This is boring and tedious (if not impossible) to create by hand
- There are tools that automate this process (dcm2niix, mne-python)

# Task data – also part of raw data

- Used in task-based neuroimaging (broadly defined)
- As opposed to resting-state neuroimaging
- Describe timing and other properties recorded during the imaging session:
  - Essentially list of triggers (for EEG) or events (for fMRI)
    - Onset, duration
  - BUT, you can add other data as well – eye-data, reaction times, etc.
  - Which later on facilitates cross-modal analyses

# Physiological and other continuous recordings

- Eye position / Respiration / heart beat / GSR

# File formats

- Imaging data:
  - MRI – nifti
  - MEG – no consensus
  - (i)EEG – edf, biosemi, brainvision, eeglab
- Other data:
  - Compressed tab-separated values (tsv.gz) text files

# Organisation of files and folders

- Clear separation of data and metadata:
  - fMRI data file (e.g., nifti)
  - Accompanying (sidecar) json file
  - Matching file names, different extensions
- The file names are prescribed using fields containing pairs of values, separated by a hyphen “-”:
  - Subject field: sub-<subID>
  - Session field: ses-<sesID>
  - Site field: unspecified
  - Task field: task-<taskID>
  - Modality field: keyword, e.g., ‘bold’, ‘physio’, ‘events’
- Different fields are separated by an underscore “_”.

This is an example of the folder and file structure. Because there is only one session, the session level is not required by the format. For details on individual files see descriptions in the next section:

```
sub-control01/
    anat/
        sub-control01_T1w.nii.gz
        sub-control01_T1w.json
        sub-control01_T2w.nii.gz
        sub-control01_T2w.json
    func/
        sub-control01_task-nback_bold.nii.gz
        sub-control01_task-nback_bold.json
        sub-control01_task-nback_events.tsv
        sub-control01_task-nback_physio.tsv.gz
        sub-control01_task-nback_physio.json
        sub-control01_task-nback_sbref.nii.gz
    dwi/
        sub-control01_dwi.nii.gz
        sub-control01_dwi.bval
        sub-control01_dwi.bvec
    fmap/
        sub-control01_phasediff.nii.gz
        sub-control01_phasediff.json
        sub-control01_magnitude1.nii.gz
        sub-control01_scans.tsv

    code/
        deface.py
    derivatives/
    README
    participants.tsv
    dataset_description.json
    CHANGES
```

Additional files and folders containing raw data may be added as needed for special cases. They should be named using all lowercase with a name that reflects the nature of the scan (e.g., `calibration`). Naming of files within the directory should follow the same scheme as above (e.g., `sub-control01_calibration_Xcalibration.nii.gz`)

# Metadata inheritance principle

- Each dataset inherits more general metadata:
  - If all participants performed a single task, e.g., “Stroop”, then the task description applies to all of them
  - Single task.json file
- This can be overridden at any level:
  - If there were two paradigm versions, one more general, the other a fall-back option, this can be overwritten on an individual basis.
- Inheritance is made easy, as the file naming is cumulative.

# The best reason to use BIDS - automated pipelines

- FMRIPREP
  - Uses only open software (nipype) in python
  - Fully automated data pre-processing pipeline (with some parameter selection)
  - Relies on BIDS metadata and data organisation

# Some useful/inspiring links

- openneuro.org – storage of datasets (successor of openfmri)
- datalad.org – continuous versioning/integration of datasets
- cognitiveatlas.org – knowledgebase (ontology) of cognitive science
  - Concepts (e.g., abstractive reasoning, etc.)
  - Tasks (e.g., backward masking, etc.)
  - Disorders
  - Theories
- cogpo.org – cognitive paradigm ontology (specifies tasks in cognitive sciences)

# Demo