The Brain Imaging Data Structure (BIDS) Specification

v. 1.0.0-rc1

The specification is now frozen (no major changes are expected) to let developers start working with it. New features and possible big changes will be incorporated in the next release.

Browse example datasets:
https://drive.google.com/folderview?id=0B2JWN60ZLkgkN0dMTVQ1QU1IUEk&usp=sharing
Download example datasets:
https://drive.google.com/folderview?id=0B2JWN60ZLkgkMGlUY3B4MXZIZW8&usp=sharing

Table of contents

Introduction

Motivation

Neuroimaging experiments result in complicated data that can be arranged in many different ways. So far there is no consensus how to organize and share data obtained in neuroimaging experiments. Even two researchers working in the same lab can opt to arrange their data in a different way. Lack of consensus (or a standard) leads to misunderstandings and time wasted on rearranging data or rewriting scripts expecting certain structure. Here we describe a simple and easy to adopt way of organising neuroimaging and behavioural data. By using this standard you will benefit in the following ways:

- It will be easy for another researcher to work on your data. To understand the organisation of the files and their format you will only need to refer them to this document. This is especially important if you are running your own lab and anticipate more than one person working on the same data over time. By using BIDS you will save

time trying to understand and reuse data acquired by a graduate student or postdoc that has already left the lab.
- There is a growing number of data analysis software packages that can understand data organised according to BIDS.
- Databases such a OpenfMRI.org accept datasets organised according to BIDS. If you ever plan to share your data publicly (nowadays some journals require this) you can minimize the additional time and energy spent on publication, and speed up the curation process by using BIDS to structure and describe your data right after acquisition.
- There are validation tools that can check your dataset integrity and let you easily spot missing values.

BIDS is heavily inspired by the format used internally by OpenfMRI.org. While working on BIDS we consulted many neuroscientists to make sure it covers most common experiments, but at the same time is intuitive and easy to adopt. The specification is intentionally based simple file formats and folder structures to reflect current lab practices and make it accessible to wide range of scientists coming from different backgrounds.[1]

Definitions

A **MRI session** is a set of MRI image acquisition procedures in which imaging data is measured from a participant. Sessions are equivalent to visits in a longitudinal study. During one session, one more more **MRI runs** may be acquired, each of which is a specific measure of brain structure or function that can be characterized by the type of information that the scan was designed to characterize. For example, different imaging scan acquisition types are used to characterize different aspects of brain structure, such as T1 weighted, T2 weighted and diffusion weighted MRI scans.
A session is therefore a set of coherent MRI acquisitions done in similar fashion across subjects, and, in general, subjects will stay in the scanner during a session. If for some reason a subject has to leave the scanner room and then be re-positioned on the scanner bed, the set of MRI acquisitions will still be considered as a session and match sessions acquired in other subjects. If several sessions are planned and performed on all -or most- subjects, often in the case of some intervention between sessions (e.g. training, etc), then several sessions would be defined (eg. pre and post behavioural session).

For functional runs, in addition to the options available for measuring different aspects of brain function (such as blood flow, blood volume, and blood oxygenation level), scans vary based on the experimental perturbations that are performed during or between scans to elicit specific changes in brain activity.
In some cases, these perturbations may be tasks in which a participant is presented with specific stimuli or asked to perform a function that is designed to elicit brain activity, they may be interventions such as exposure to drugs, or the participant may be asked to lie quietly. For simplicity, all of the different experimental perturbations are referred to as **tasks**[2], and a **task**

[1] For review of other data organization standards in neuroimaging see Appendix.
[2] Please mind that according to this definition “resting state” is considered a task.

run is a single consecutive measurement of the participant during the perturbation. Depending on the nature of the experiment, the participant may be exposed to several different perturbations during a task run, or may receive different perturbations in different task runs. A **protocol** provides an end-to-end specification of an MRI scan including the parameters used to acquire the scan and the task that was performed during the scan. For scanning acquisition, the protocol contains a list of the various scanning parameters that were the existing folder structure following common sense used in order to optimize the measurement for the type of information that the scan was designed to characterize. For the task, the protocol consists of any instructions given to the participant along with the timing of the various experimental events (stimuli and responses) that occurred during the task scan. A **condition** is a specific category of events, an **onset** is the time at which the event began, and the **duration** is the amount of time that the event lasted.

Compulsory, optional, and additional

The following standard describes a way of arranging data and writing down metadata for a subset of neuroimaging experiments. Some aspects of the standard are compulsory. For example a particular file name format is required when saving structural scans. Some aspects are regulated but optional. For example a T2 volume does not need to be included, but when it is available is should be saved under a particular file name specified in the standard.
This standard aspires to describe a majority of datasets, but acknowledges that there will be cases that do not fit. In such cases one can include additional files and subfolders to the existing folder structure following common sense. For example one may want to include eye tracking data in a vendor specific format that is not covered by this standard. The most sensible place to put it is next to the continuous recording file with the same naming scheme but different extensions. The solutions will change from case to case and publicly available datasets will be reviewed to include common data types in the future releases of the BIDS spec.

Raw vs. derived data

BIDS in its current form is designed to described raw (unprocessed) data. During analysis such data will be transformed and partial as well as final results will be saved. Derivatives of the raw data should be kept separate from the raw data. This way one can protect the raw data from accidental changes by file permissions. In addition it is easy to distinguish partial results from the raw data and share the latter. Even though this specification does not go into details of recommending a particular naming scheme for different data derivatives (correlation maps, brain masks, contrasts maps etc.) we recommend keeping them in a separate “derivatives” folder with a similar folder structure as presented below for the raw data. For example:
`derivatives/sub-01/ses-pre/mask.nii.gz`.

The Inheritance Principle

Any metadata file (.json, .bvec, .tsv, etc.) may be defined at any directory level. The values from the top level are inherited by all lower levels unless they are overridden by a file at the lower level. For example, tasks.tsv may be specified at the participant level, setting TR to a specific

value. If one of the series has a different TR than the one specified in that file, another tasks.tsv file can be placed within that specific series directory specifying the TR for that specific run.

Extensions

BIDS specification can be extended. Currently Model and hypothesis specification is one of such extensions.

File Format specification

Imaging files

Imaging data are stored using the NIfTI file format, preferably compressed NIfTI files (.nii.gz), either 1.0 or 2.0 version. Imaging data should be converted to the NIfTI format using a tool that provides as much of the NIfTI header information (such as orientation and slice timing information) as possible. Since the NIfTI standard offers limited support for the various image acquisition parameters available in DICOM files, we also encourage users to provide additional meta information extracted from DICOM files in a sidecar JSON file (with the same filename as the .nii.gz file, but with .json extension). This metadata can be easily extracted using DCMStack DICOM to NIFTI converter (https://github.com/moloney/dcmstack). Provided validator will check for conflicts between the JSON file and the data recorded in the NIFTI header.

Tabular files

Other files are saved as tab delimited values (.tsv) files, aka csv files where commas are replaced by tabs. Tabs need to be true tab characters and not series of space characters. Therefore to avoid confusion files should not include more than one adjacent space character. Each TSV file need to start with a header line listing the names of all columns (with the exception for physiological and other continuous acquisition data - see below for details). Names need to be separated with tabs. String values containing tabs should be escaped using double quotes.
Missing values should be coded as “n/a”.

Example:[3]

```
onset duration   response_time   correct    stop_trial go_trial
200   20         0               n/a        n/a        n/a
```

Key/value files (dictionaries)
Javascript Object Notation Style (JSON) files will be used for storing key/value pairs. Extensive documentation of the format can be found here: http://json.org/. Several editors have built-in support for JSON syntax highlighting that aids manual creation of such files. An online editor for JSON with built-in validation is available at: http://jsoneditoronline.org. JSON files should be in either ASCII or UTF-8 encoding.

[3] Note that to make the display clearer, the second row does contain two successive tabs, which should not happen in an actual BIDS tsv file.

Example:

```
{
    "RepetitionTime": 3.0,
    "Instruction": "Lie still and keep your eyes open"
}
```

Subject naming

Subjects (participants participants) should be assigned unique labels. Labels can consist of letters and/or numbers. If numbers are used we strongly encourage zero padding (“01” instead of “1” if you have more than nine subjects) to make alphabetical sorting more intuitive.

Units

Elapsed time should be expressed in seconds. Please note that some DICOM parameters have been traditionally expressed in milliseconds. Those need to be converted to seconds.

Frequency should be expressed in Hertz.

Describing dates and timestamps:

- Date time information should be expressed in the following format YYYY-MM-DDThh:mm:ss (ISO8601 date-time format). For example: 2009-06-15T13:45:30
- Time stamp information should be expressed in the following format: 13:45:30
- Dates can be shifted by a random number of days for privacy protection reasons. To distinguish real dates from shifted dates always use year 1900 or earlier when including shifted years. For longitudinal studies please remember to shift dates within one subject by the same number of days to maintain the interval information. Example: 1867-06-15T13:45:30
- Age should be given as the number of years since birth at the time of scanning (or first scan in case of multi session datasets). Using higher accuracy (weeks) should be avoided due to privacy protection.

Directory structure (single session example)

This is an overview 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
    - Additional files and folders 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”)
- **code**
    - deface.py
- tasks.json
- participants.tsv
- dataset_description.json
- README
- CHANGES

Detailed file descriptions

Dataset description

Template: `dataset_description.json`, `README`, and `CHANGES`

A JSON file describing the dataset. Following fields are mandatory:

- `Name`: name of the dataset
- `License`: what license is this dataset distributed under? The use of license name abbreviations is suggested for specifying a license. A list of common licenses with suggested abbreviations can be found in appendix III.
- `Authors`: List of individuals who contributed to the creation/curation of the dataset
- `Acknowledgements`: who should be acknowledge in helping to collect the data
- `HowToAcknowledge`: Instructions how researchers using this dataset should acknowledge the original authors. This field can also be used to define a publication that should be cited in publications that use the dataset,

- `Funding`: sources of funding (grant numbers)
- `ReferencesAndLinks`: a list of references to publication that contain information on the dataset, or links.

Example:

```
{
    "Name": "The mother of all experiments",
    "License": "CC0",
    "Authors": ["Ramon y Cajal"],
    "Acknowledgements": "say here what are your acknowledgments",
    "HowToAcknowledge": "say here how you would like to be
acknowledged",
    "Funding": "list your funding sources",
    "ReferencesAndLinks": "a paper / resource to be cited when using
the data"
}
```

In addition a free form text file (`README`) describing the dataset in more details should be provided. Version history of the dataset (describing changes, updates and corrections) can be provided in a form of a `CHANGES` text file. This file should follow the following the CPAN Changelog convention: http://search.cpan.org/~haarg/CPAN-Changes-0.400002/lib/CPAN/Changes/Spec.pod. `README` and `CHANGES` files should be either i ASCII or UTF-8 encoding.

Example:

```
1.0.1 2015-08-27

  - Fixed slice timing information.

1.0.0 2015-08-17

  - Initial release.
```

Anatomy imaging data

Template: code/*

Source code of scripts used to prepare the dataset (for example if it was anonymized or defaced) can be stored here. There are no limitations or recommendation on the language and or code organization of these scripts at the moment.

Anatomy imaging data

Template:

`sub-<participant_label>/`

```
anat/
      sub-<participant_label>[_acq-<label>][_rec-<label>][_run-<
      index>]_T2w.nii.gz
```

Anatomical (structural) data acquired for that participant. Possible modalities include, but are not limited to:

Name	Suffix	Description
T1 weighted	_T1w	
T2 weighted	_T2w	
T1 map	_T1map	quantitative T1 map (likewise for T2)
T2 map	_T2map	quantitative T2 map
FLAIR	_FLAIR	
Proton density	_PD	
Combined PD/T2	_PDT2	
Inplane T1	_inplaneT1	T1-weighted anatomical image matched to functional acquisition
Inplane T2	_inplaneT2	T2-weighted anatomical image matched to functional acquisition
Angiography	_angio	
Defacing mask	_defacemask	Mask used for defacing
Susceptibility Weighted Imaging (SWI)	_SWImagandphase	Magnitude and corresponding phase images of the SWI

Several scans of the same modality can be acquired and will be indexed with a suffix: _run-01, _run-02, _run-03 etc. When there is only one scan of a given type the suffix can be omitted. Please note that diffusion imaging data is stored elsewhere (see below). Metadata corresponding to each anatomical NifTi file can be provided in a JSON file with the same name but .json extension.
The optional “acq-<label>” key/value pair correspond to a custom label user may use to distinguish different set of parameters used for acquiring the same modality. For example this should be used when a study includes two T1w images - one full brain low resolution and and one restricted field of view but high resolution. In such case two files could have the following names: `sub-01_acq-highres_T1w.nii.gz` and `sub-01_acq-lowres_T1w.nii.gz`,

however the user is free to choose any other label than “highres” and “lowres” as long as they are consistent across subjects and sessions.
Similarly the optional “rec-<label>” key/value can be used to distinguish different reconstruction algorithms (for example ones using motion correction).

Task (including resting state) imaging data

Template:

```
sub-<participant_label>/
       func/
              sub-<participant_label>_task-<task_label>[_acq-<label>][_run-<index>]_bol
              d.nii.gz
```

Imaging data acquired during BOLD imaging. This includes but is not limited to task based fMRI **as well as resting state fMRI (i.e. rest is treated as another task)**. For task based fMRI a corresponding task events (see below) file must be provided. For multiband acquisitions, one can also save the single-band reference image as type “sbref” (e.g. sub-control01_task-nback_sbref.nii.gz).

Each task has a unique label consisting of letters and/or numbers (other characters including spaces and underscores are not allowed). Those labels has to be consistent across subjects and sessions.
If more than one run of the same task has been acquired a key/value pair: _run-01, _run-02, _run-03 etc. can be used. If only one run was acquired the suffix can be omitted. In the context of functional imaging a run is defined as the same task, but a different set of stimuli (for example randomized order) and participant responses.
The optional “acq-<label>” key/value pair correspond to a custom label user may use to distinguish different set of parameters used for acquiring the same task. For example this should be used when a study includes two resting state images - one single band and one multiband. In such case two files could have the following names: `sub-01_task-rest_acq-singleband_bold.nii.gz` and `sub-01_task-rest_acq-multiband_bold.nii.gz`, however the user is free to choose any other label than “singleband” and “multiband” as long as they are consistent across subjects and sessions.

Some meta information about the acquisition needs to be provided in an additional JSON file. Required fields:

- `RepetitionTime`: The time in seconds between the beginning of an acquisition of one volume and the beginning of acquisition of the volume following it (TR). Please note that this definition includes time between scans (when no data has been acquired) in case of sparse acquisition schemes. This value needs to be consistent with the ‘pixdim[4]’ field (after accounting for units stored in ‘xyzt_units’ field) in the NIFTI header.
- `TaskName`: Name of the task (for resting state use the “rest” prefix). No two tasks should have the same name. Task label is derived from this field by removing all non alphanumeric ([a-zA-Z0-9]) characters.

Other recommended metadata.
MRI acquisition parameters are divided into several categories based on "A checklist for fMRI acquisition methods reporting in the literature" by Ben Inglis::

- Scanner Hardware
  - `Manufacturer`: Manufacturer of the equipment that produced the composite instances. Corresponds to http://neurolex.org/wiki/Category:Manufacturer
  - `ManufacturerModelName`: Manufacturers model name of the equipment that produced the composite instances. Corresponds to http://neurolex.org/wiki/Category:Manufacturers_Model_Name
  - `MagneticFieldStrength`: Nominal field strength of MR magnet in Tesla. Corresponds to http://neurolex.org/wiki/Category:Magnetic_Field_Strength
  - `HardcopyDeviceSoftwareVersion`: Manufacturer's designation of the software of the device that created this Hardcopy Image. Corresponds to http://neurolex.org/wiki/Category:Hardcopy_Device_Software_Version
  - `ReceiveCoilName`: Information describing the receiver coil (Note: This isn't a consistent field name across vendors. This name is the dcmstack output from a GE dataset, but it doesn't seem to be simply coded into the dcmstack output for a Siemens scan). This doesn't correspond to a term in the DICOM ontology
  - `GradientSetType`: It should be possible to infer the gradient coil from the scanner model. If not,*e.g.* because of a custom upgrade or use of a gradient insert set, then the specifications of the actual gradient coil should be reported independently.
  - `MRTransmitCoilSequence`: This is a relevant field if a non-standard transmit coil is used. Corresponds to http://neurolex.org/wiki/Category:MR_Transmit_Coil_Sequence
  - `MatrixCoilMode`: (If used) A method for reducing the number of independent channels by combining in analog the signals from multiple coil elements. There are typically different default modes when using un-accelerated or accelerated (*e.g.* GRAPPA, SENSE) imaging. [No corresponding Dicom ontology term?]
  - `CoilCombinationMethod`:Almost all fMRI studies using phased-array coils use root-sum-of-squares (rSOS) combination, but other methods exist. The image reconstruction is changed by the coil combination method (as for the matrix coil mode above), so anything non-standard should be reported. [No corresponding Dicom ontology term?]
- In-Plane Spatial Encoding
  - `ScanningSequence, SequenceVariant, ScanOptions, MRAcquisitionType, SequenceName`: These are parameters that define the pulse sequence. The labels are outputted by dcmstack, [No corresponding Dicom ontology terms?]
  - `NumberShots`:Most echo planar imaging uses one excitation pulse per slice. This parameter can be used to note if one or more excitation pulses are used [No corresponding Dicom ontology term?]

    - `ParallelReductionFactorInPlane`: The parallel imaging (e.g, GRAPPA) factor. Use the denominator of the fraction of k-space encoded for each slice. For example, 2 means half of k-space is encoded. Corresponds to http://neurolex.org/wiki/Category:Parallel_Reduction_Factor_In-plane.
    - `ParallelAcqusitionTechnique`: The type of parallel imaging used (e.g. GRAPPA, SENSE). Corresponds to http://neurolex.org/wiki/Category:Parallel_Acquisition_Technique.
    - `PartialFourier`: The fraction of partial Fourier information collected. Corresponds to http://neurolex.org/wiki/Category:Partial_Fourier
    - `PartialFourierDirection`: The direction where only partial Fourier information was collected. Corresponds to http://neurolex.org/wiki/Category:Partial_Fourier_Direction
    - `PhaseEncodingDirection`: Possible values: “x”, “y”, “z”, “x-”, “y-”, “z-” (corresponding to the first, second and third axis of the data in the NIFTI file together with the direction). Note that this is not the same as the DICOM term InPlanePhaseEncodingDirection which can have “ROW” or “COL” values. **This parameter is required if a corresponding fieldmap data is present.**
    - `EffectiveEchoSpacing`: The sampling interval also known as the dwell time; required for unwarping distortions using field maps; **expressed in seconds.** See here how to calculate it. **This parameter is required if a corresponding fieldmap data is present.**
- Timing Parameters:
    - `EchoTime`: The echo time (TE) for the acquisition, specified in seconds. May be a list of values separated for multi-echo acquisitions. **This parameter is required if a corresponding fieldmap data is present.** Corresponds to http://neurolex.org/wiki/Category:Echo_Time.
    - `SliceTiming`: The time at which each slice was acquired during the acquisition. Slice timing is not slice order - it describes the time (sec) of each slice acquisition in relation to the beginning of volume acquisition. It is described using a list of times (separated by spaces) referring to the acquisition time for each slice. The list goes through slices along the slice axis in the slice encoding dimension (see below). **This parameter is required for sparse sequences. In addition without this parameter slice time correction will not be possible.**
    - `SliceEncodingDirection`: Possible values: “x”, “y”, “z”, “x-”, “y-”, “z-”(corresponding to the first, second and third axis of the data in the NIFTI file). The axis of the NIFTI data along which a slices were acquired. This value needs to be consistent with the ‘slice_dim’ field in the NIFTI header. **Without this parameter slice time correction will not be possible.**
- RF & Contrast
    - `FlipAngle`: Flip angle for the acquisition, specified in degrees. Corresponds to: http://neurolex.org/wiki/Category:Flip_Angle.
- Slice Acceleration

    - `MultibandAccelerationFactor`: The multiband factor, for multiband acquisitions
- fMRI task information
    - `Instructions`: Text of the instructions given to participants before the scan. This is especially important in context of resting state fMRI and distinguishing between eyes open and eyes closed paradigms.
    - `TaskDescription`: Longer description of the task.
    - `CogAtlasID`: URL of the corresponding Cognitive Atlas Task term
    - `CogPOID`: URL of the corresponding CogPO term

When adding additional metadata please use camel case version of DICOM ontology terms whenever possible.

Example:

```
sub-control01/
      func/
            sub-control01_task-nback_bold.json
{
    "RepetitionTime": 3.0,
    "EchoTime": 0.03,
    "FlipAngle": 78,
    "SliceTiming": [0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8,
2.0, 2.2, 2.4, 2.6, 2.8],
    "MultibandAccellerationFactor": 4,
    "ParallelReductionFactorInPlane": 2,
    "PhaseEncodingDirection": "y"
}
```

If this information is the same for all participants, sessions and runs it can be provided in `task<task_number>_bold.json` (in the root directory of the dataset). However, if the information differs between subjects/runs it can be specified in the `sub-<participant_label>/func/sub-<participant_label>_task-<task_label>[_acq-<label>][_run-<index>]_bold.json` file. If both files are specified the one corresponding to a particular participant, task and run takes precedence.

Task events

Template:

```
sub-<participant_label>/
      func/
            <matches>_events.tsv
```

Where `<matches>` corresponds to task file name. For example:

```
sub-control01_task-nback
```

The purpose of this file is to describe timing and other properties of events recorded during the scan. Those can be both stimuli presented to the participant, or participant responses. Each task events file requires a corresponding task imaging data file. The first two columns are mandatory and should contain “onset” and “duration” in that order. Both should be expressed in seconds (not milliseconds). Timing should be measured from the beginning of the acquisition of the first volume in the corresponding task imaging data file. Therefore negative numbers in “onset” are allowed[4] (however “duration” needs to be always positive and greater than zero). If response time was measured it should be included as the “response_time” column also expressed in seconds. An arbitrary number of additional columns can be added. Those allow describing other properties of events that could be later referred in modelling and hypothesis extensions of BIDS.

Example:

```
sub-control01/
      func/
            sub-control01_task-nback_events.tsv
onset       duration    trial_type response_time
1.2         0.6         go         1.435
5.6         0.6         stop       1.739
```

References to existing databases can also be encoded using additional columns. Example 2 includes references to the Karolinska Directed Emotional Faces (KDEF) database[5]:

Example2:

```
sub-control01/
      func/
            sub-control01_task-nback_events.tsv
onset       duration    trial_type identifier database   response_time
1.2         0.6         afraid     AF01AFAF   kdef       1.435
5.6         0.6         angry      AM01AFAN   kdef       1.739
5.6         0.6         sad        AF01ANSA   kdef       1.739
```

Physiological and other continuous recordings

Template:

```
sub-<participant_label>/
      func/
            <matches>[_recording-<label>]_physio.tsv.gz
```

and

[4] For example in case there is an in scanner training phase that begins before the scanning sequence has started events from this sequence should have negative onset time counting down to the beginning of the acquisition of the first volume.

[5] http://www.emotionlab.se/resources/kdef

```
sub-<participant_label>/
      func/
            <matches>[_recording-<label>]_physio.json

sub-<participant_label>/
      func/
            <matches>[_recording-<label>]_stim.tsv.gz
```
and
```
sub-<participant_label>/
      func/
            <matches>[_recording-<label>]_stim.json
```
Optional: yes

Where `<matches>` corresponds to task file name. For example: `sub-control01_task-nback`. If the same continuous recording has been used for all subjects (for example in the case where they all watched the same movie) one file can be used and placed in the root directory. For example: `task-movie_stim.tsv.gz`

Physiological recordings such as cardiac and respiratory signals and other continuous measures (such as parameters of a film or audio stimuli) can be specified using two files: a gzip compressed TSV file with data (without header line) and a JSON for storing start time, sampling frequency, and name of the columns from the TSV. Please note that in contrast to other TSV files this one does not include a header line. Instead the name of columns are specified in the JSON file. This is to improve compatibility with existing software (FSL PNM) as well as make support for other file formats possible in the future. Start time should be expressed in seconds with relation of the time of acquisition of first volume in the series (negative values are allowed). Sampling frequency should be expressed in Hz. The following naming conventions should be used:

- `cardiac`: continuous pulse measurement
- `respiratory`: continuous breathing measurement
- `trigger`: continuous measurement of the scanner trigger signal

Any combination of those three can be included as well as any other potentially stimuli related continuous variables (such as low level image properties in a video watching paradigm).

Physiological recordings should use the `_physio` suffix and signals related to the stimulus should use `_stim` suffix.

More than one continuous recording file can be included (with different sampling frequencies). In such case use different labels. For example: `_recording-contrast`, `_recording-saturation`.

Example:

```
sub-control01/
      func/
            sub-control01_task-nback_physio.tsv.gz (after
            decompression)
34          110             0
44          112             0
23          100             1

sub-control01/
      func/
            sub-control01_task-nback_physio.json
{
    "SamplingFrequency": 100.0,
    "StartTime": -22.3,
    "Columns": ["cardiac", "respiratory", "trigger"]
}
```

Diffusion imaging data

Template:

```
sub-<participant_label>/
      dwi/
            sub-<participant_label>[_acq-<label>][_run-<index>]_dwi.ni
            i.gz

sub-<participant_label>/
      dwi/
            sub-<participant_label>[_acq-<label>][_run-<index>]_dwi.bv
            al

sub-<participant_label>/
      dwi/
            sub-<participant_label>[_acq-<label>][_run-<index>]_dwi.bv
            ec

sub-<participant_label>/
      dwi/
            sub-<participant_label>[_acq-<label>][_run-<index>]_dwi.js
            on
```

Diffusion-weighted imaging data acquired for that participant. The optional “acq-<label>” key/value pair correspond to a custom label user may use to distinguish different set of

parameters. For example this should be used when a study includes two diffusion images - one single band and one multiband. In such case two files could have the following names: `sub-01_acq-singleband_dwi.nii.gz` and `sub-01_acq-multiband_dwi.nii.gz`, however the user is free to choose any other label than “singleband” and “multiband” as long as they are consistent across subjects and sessions.
The bvec and bval files in the FSL format[6]: The bvec files contain 3 rows with n space-delimited floating-point numbers (corresponding to the n volumes in the relevant Nifti file). The first row contains the x elements, the second row contains the y elements and third row contains the z elements of a unit vector in the direction of the applied diffusion gradient, where the i-th elements in each row correspond together to the i-th volume with [0,0,0] for non-diffusion-weighted volumes.

.bvec example:

```
0 0 0.021828 -0.015425 -0.70918 -0.2465
0 0 0.80242 0.22098 -0.00063106 0.1043
0 0 -0.59636 0.97516 -0.70503 -0.96351
```

The bval file contains the b-values (in s/mm^2) corresponding to the volumes in the relevant Nifti file), with 0 designating non-diffusion-weighted volumes, space-delimited.

.bval example:

```
0 0 2000 2000 1000 1000
```

.bval and .bvec files can be saved on any level of the directory structure and thus define those values for all sessions and/or subjects in one place (see Inheritance principle).

In case multiple identical sequences with varying phase encoding direction were acquired (for distortion correction) a JSON file should be provided. The file should contain phase encoding direction (`[PA, AP, RL, LR, HF, FH]`) for use in b0 distortion correction. In addition the Total Readout Time specified as time in seconds between from the center of the first echo to the center of the last echo (see here and here how to calculate it).

.json example:

```
{
    "PhaseEncodingDirection": "PA",
    "TotalReadoutTime": 0.000095,
}
```

Fieldmap data

Data acquired to correct for B0 inhomogeneities can come in different forms. The current version of this standard considers four different scenarios. Please note that in all cases fieldmap

[6] http://fsl.fmrib.ox.ac.uk/fsl/fsl4.0/fdt/fdt_dtifit.html

data can be linked to a specific scan(s) it was acquired for by filling the “IntendedFor” field in the corresponding JSON file. For example:

```
{
   "IntendedFor": "func/sub001-task-motor_bold.nii.gz"
}
```

The IntendedFor field consists of one or more relative paths stripped from the “sub-<label>”. Here’s an example with multiple target scans:

```
{
   "IntendedFor": ["ses-pre/func/sub001-task-motor_run-01_bold.nii.gz",
                   "ses-post/func/sub001-task-motor_run-01_bold.nii.gz"]
}
```

Multiple fieldmaps can be stored. In such case the “_run-01”, “_run-02” should be used.

Case 1: Phase difference image and at least one magnitude image

Template:

```
sub-<participant_label>/
      fmap/
            sub-<label>_phasediff.nii.gz

sub-<participant_label>/
      fmap/
            sub-<label>_phasediff.json

sub-<participant_label>/
      fmap/
            sub-<label>_magnitude1.nii.gz
```

(optional)

```
sub-<participant_label>/
      fmap/
            sub-<label>_magnitude2.nii.gz
```

This is a common output for build in fieldmap sequence on Siemens scanners. In this particular case the sidecar JSON file has to define the Echo Time Difference parameter which is defined as Echo Time of the second phase image acquisition minus Echo Time of the first phase image acquisition. For example:

```
{
    "EchoTimeDifference": 0.00246,
    "IntendedFor": "func/sub001-task-motor_bold.nii.gz"
}
```

Case 2: Two phase images and two magnitude images

Template:

```
sub-<participant_label>/
      fmap/
            sub-<label>_phase1.nii.gz

sub-<participant_label>/
      fmap/
            sub-<label>_phase1.json

sub-<participant_label>/
      fmap/
            sub-<label>_phase2.nii.gz

sub-<participant_label>/
      fmap/
            sub-<label>_phase2.json

sub-<participant_label>/
      fmap/
            sub-<label>_magnitude1.nii.gz

sub-<participant_label>/
      fmap/
            sub-<label>_magnitude2.nii.gz
```

Similar to the case above, but instead of a precomputed phase difference map two separate phase images are presented. The two sidecar JSON file need to specify corresponding EchoTime values. For example:

```
{
    "EchoTime": 0.00746,
    "IntendedFor": "func/sub001-task-motor_bold.nii.gz"
}
```

Case 3: A single, real fieldmap image (showing the field inhomogeneity in each voxel)

Template:

```
sub-<participant_label>/
      fmap/
            sub-<label>_fieldmap.nii.gz
```

```
sub-<participant_label>/
      fmap/
            sub-<label>_fieldmap.json
```

In some cases (for example GE) the scanner software will output precomputed fieldmap denoting the B0 inhomogeneities. In this case the sidecar JSON file need to include the units of the fieldmap. The possible options are: “Hz”, “rad/s”, or “Tesla”. For example:

```
{
    "Units": "rad/s",
    "IntendedFor": "func/sub001-task-motor_bold.nii.gz"
}
```

Case 4: Multiple phase encoded directions (topup)

Template:

```
sub-<participant_label>/
      fmap/
            sub-<label>_dir-<index>_epi.nii.gz

sub-<participant_label>/
      fmap/
            sub-<label>_dir-<index>_epi.json
```

This technique combines two or more Spin Echo EPI scans with different phase encoding directions. In such a case, the phase encoding direction is specified in the corresponding JSON file as one of: “x”, “y”, “z”, “x-”, “y-, “z-”. For these differentially phase encoded sequences, one also needs to specify the Total Readout Time defined as the time (in seconds) from the center of the first echo to the center of the last echo (see here and here how to calculate it). For example

```
{
    "PhaseEncodingDirection": "y-",
    "TotalReadoutTime": 0.000095,
    "IntendedFor": "func/sub001-task-motor_bold.nii.gz"
}
```

Scans key file

Template:

```
sub-<participant_label>/
      sub-<participant_label>_scans.tsv
```

Optional: yes

The purpose of this file is to describe timing and other properties of each imaging acquisition sequence (each run .nii.gz file) within one session. Each .nii.gz file should be described by at most one row. Relative paths to files should be used under a compulsory “filename” header. If acquisition time is included it should be under “acq_time” header. Datetime should be expressed in the following format `2009-06-15T13:45:30` (year, month, day, hour (24h), minute, second; this is equivalent to the RFC3339 “date-time” format, time zone is always assumed as local time). For anonymization purposes all dates within one subject should be shifted by a randomly chosen (but common across all runs etc.) number of days. This way relative timing would be preserved, but chances of identifying a person based on the date and time of their scan would be decreased. Dates that are shifted for anonymization purposes should be set to a year 1900 or earlier to clearly distinguish them from unmodified data. Shifting dates is recommended, but not required.
Additional fields can include external behavioural measures relevant to the scan. For example vigilance questionnaire score administered after a resting state scan.

Example:

```
filename                                          acq_time
func/sub-control01_task-nback_bold.nii.gz  1877-06-15T13:45:30
func/sub-control01_task-motor_bold.nii.gz  1889-06-15T13:55:33
```

Participant key file

Template:
(single session case)
`participants.tsv`

The purpose of this file is to describe properties of participants such as age, handedness, gender etc. In case of single session studies this file has one compulsory column “participant_id” followed by a list of optional columns describing participants. Each participant needs to be described by one and only one row.

Single session example:

```
participant_id  age   sex   group
sub-control01   34    M     control
sub-control02   12    F     patient
sub-patient01   33    F     control
```

Longitudinal studies with multiple sessions (visits)

Multiple sessions (visits) are encoded by adding an extra layer of directories and file names in the form of “ses-<session_label>”. Session label can consist only of alphanumeric characters

[a-zA-Z0-9] and should consistent across subjects. If numbers are used in session labels we recommend using zero padding (for example ses-01, ses-11 instead of ses-1, ses-11). This makes results of alphabetical sorting more intuitive. Acquisition time of session can be defined in the sessions file (see below for details).

The extra session layer should be added for all subjects if at least one subject in the dataset has more than one session. Skipping the session layer for some subjects in the dataset is not allowed.

- **sub-control01**
  - **ses-predrug**
    - **anat**
      - sub-control01_ses-predrug_T1w.nii.gz
      - sub-control01_ses-predrug_T1w.json
      - sub-control01_ses-predrug_T2w.nii.gz
      - sub-control01_ses-predrug_T2w.json
    - **func**
      - sub-control01_ses-predrug_task-nback_bold.nii.gz
      - sub-control01_ses-predrug_task-nback_bold.json
      - sub-control01_ses-predrug_task-nback_events.tsv
      - sub-control01_ses-predrug_task-nback_cont-physio.tsv.gz
      - sub-control01_ses-predrug_task-nback_cont-physio.json
      - sub-control01_ses-predrug_task-nback_sbref.nii.gz
    - **dwi**
      - sub-control01_ses-predrug_dwi.nii.gz
      - sub-control01_ses-predrug_dwi.bval
      - sub-control01_ses-predrug_dwi.bvec
    - **fmap**
      - sub-control01_ses-predrug_phasediff.nii.gz
      - sub-control01_ses-predrug_phasediff.json
      - sub-control01_ses-predrug_magnitude1.nii.gz
    - sub-control01_ses-predrug_scans.tsv
  - **ses-postdrug**
    - **func**
      - sub-control01_ses-postdrug_task-nback_bold.nii.gz
      - sub-control01_ses-postdrug_task-nback_bold.json
      - sub-control01_ses-postdrug_task-nback_events.tsv
      - sub-control01_ses-postdrug_task-nback_cont-physio.tsv.gz
      - sub-control01_ses-postdrug_task-nback_cont-physio.json
      - sub-control01_ses-postdrug_task-nback_sbref.nii.gz
    - **fmap**
      - sub-control01_ses-postdrug_phasediff.nii.gz
      - sub-control01_ses-postdrug_phasediff.json

            - sub-control01_ses-postdrug_magnitude1.nii.gz
        - sub-control01_ses-postdrug_scans.tsv
    - sub-control01_sessions.tsv
- participants.tsv
- dataset_description.json
- README
- CHANGES

Sessions file

Template:

```
sub-<participant_label>/
      sub-<participant_label>_sessions.tsv
```

In case of multiple sessions there is an option of adding an additional participant key files describing variables changing between sessions. In such case one file per participant should be added. These files need to include compulsory “session_id” column and describe each session by one and only one row. Column names in per participant key files has to be different from group level participant key column names.

Multiple sessions example:

```
session_id      acq_time             systolic_blood_pressure
ses-predrug     2009-06-15T13:45:30  120
ses-postdrug    2009-06-16T13:45:30  100
ses-followup    2009-06-17T13:45:30  110
```

Appendix I: Glossary:

- **Session** is a continuous MRI image acquisition procedure in which imaging data is measured from a participant. During a session, more than one run can be acquired. Session correspond to visits in longitudinal studies.
- **Run** is an MRI data acquisition, a specific measure of brain structure or function that can be characterized by the type of information that the scan was designed to characterize. For example, different imaging scan acquisition types are used to characterize different aspects of brain structure, such as T1 weighted, T2 weighted and diffusion weighted MRI scans. For functional scans, in addition to the options available for measuring different aspects of brain function (such as blood flow, blood volume, and blood oxygenation level), runs vary based on the experimental perturbations that are performed during or between scans to elicit specific changes in brain activity.
- **Protocol** provides an end-to-end specification of an MRI scan including the parameters used to acquire the scan and the task that was performed during the scan. For scanning acquisition, the protocol contains a list of the various scanning parameters that were

used in order to optimize the measurement for the type of information that the scan was designed to characterize.
- **Event** is a stimulus presented to a participant or a direct or indirect participant response. Blocks of in fMRI tasks are also treated as event, but with longer duration.
- **Onset** is the time at which the event began expressed in seconds.
- **Duration** is the amount of time that the event lasted expressed in seconds.
- **Session** is one continuous visit of a subject to the scanning facility during which participant may enter the scanner multiple times.

Appendix II: Review of other standards

- XCEDE
- OntoNeuroLog
- XNAT
- OpenfMRI
- LORIS

Appendix III: Licenses

This section lists a number of common licenses for datasets and defines suggested abbreviations for use in the dataset metadata specifications

PD	Public Domain	No license required for any purpose; the work is not subject to copyright in any jurisdiction.
PDDL	Open Data Commons Public Domain Dedication and License	License to assign public domain like permissions without giving up the copyright: http://opendatacommons.org/licenses/pddl/
CC0	Creative Commons Zero 1.0 Universal.	Use this if you are a holder of copyright or database rights, and you wish to waive all your interests in your work worldwide: https://creativecommons.org/about/cc0

What is below is <u>not part of the specifications</u> : brainstorming / material / idea …

From Satra's mail:

if people are working on the openfmri extensions, here are things to consider:

- sparse/non-sparse acquisitions (sparse acquisitions require TR and TA and whether the acquisition was triggered by stimuli or whether acquisition was periodic)
    this can be handled through a is_sparse boolean key in scan_key

- metadata at:
  - dataset level
    - participant level
      - visit level
        (some demographics should go here things like age will change across visits, gender might change as well)
  - links to cognitive paradigms in tasks
    perhaps this goes into experiment.yaml/json/rdf

This section is a brainstorming on how do we go from the condition_key.tsv to the design matrix of the regression model. We thought about three options.

1) Keep the model description at a high level and have models following R or patsy (https://patsy.readthedocs.org/en/v0.1.0/formulas.html) but this may be constraining : many decisions are made on the parametrisation of the model without control of the user.
2) Design a low level description that would allow a precise specification of the design matrix from the condition_key.tsv.
3) Have both, low level is generated from high level. Users will in general specify a high level model, but can specify a low level model.

Another question is the first level versus second level.

- At the first level, regressors are obtained either from conditions (that are convolved with Hrf and subsampled), or are covariates already sampled at each TR. While we could have these two in the same file, I think it would simplify things to have two separate files. This also raise the question of the movement parameters. These are not yet computed when the data is submitted to openfMRI and thus require a symbolic notation.
- At the second level, regressors of dimension the number of participants should be specified (age, gender, scanner type, etc). If we want to allow for more complex models,

then we will need the full fledge model description (between / within factors, etc). Unlikely to be solved now (unless we just follow R conventions for specification of mixed models).

Intra series (run) model: A, B are columns found in the condition_key.

Operators:

1) + : A+B : create variable containing the onsets times of both A and B. From the example of SST: `GoTrial+StopTrial` will create a variable with all Go and Stop onsets.
2) * : A*B : use to create the interaction between conditions A and B.
3) (A < | > | <= | >= | == | <> | B) : element-wise create a boolean. Typically, create a boolean condition from a quantitative value (say, keep events that have RT less than 1s)
4) Hrf(A) :convolve by Hrf. Can specify if derivatives of the Hrf are used. Hrf(A, 'dt', 'dd') : 'dt' : with time derivative, ''dd': with dispersion derivative.
5) M(A) : mean center A. Use to mean center regressors.
6) #(A, B, C, ...) : orthogonalize A wrt to B; then C, ...
7) ** : A ** B: Used to create the full interaction space between factors with several levels.
8) []: slicing
9) {}: stack horizontally
10) I : Intercept

Example of models:
tsv will contain: var1, var2, RT, Diff, mvt_0..5
a1 = Hrf(var1)
a2 = Hrf(var2)
a1x2 = Hrf(var1*var2)

model = {a1, a2, a1x2, #(Hrf(var1*RT, {a1,I}) , mvt_0, .., mvt_5, I}

This reads : conditions 1, 2, their interaction, and the condition1 modulated by RT and orthogonalized wrt to a1 and I (ie mean centered).

Use Cases:
I've been thinking for some time that data hosts (i.e. XNAT) should provide image-level unique identifiers (such as DOI or other). Then, when a query generates a set of data for some other use, this new set contains a collection of all the identifiers of the source data, so that credit and utilization can be tracked more efficiently. While I'm open to discussion about the exact nature of this identifier, I do like to think of it in terms of the types of structure and functions that DOI's provide already, so I'll couch the use case in DOI terms until we replace the DOI with a different construct, just for the sake of concreteness. A DOI has a specified schema and set of fields that we can usurp for our data citation purposes: Author (person who collected the data), publisher (data host), etc. Included is a pointer to the actual data. While in the minimalist sense, this

pointer to the data could be a pointer to a NIfTI file (or zipped DICOM dataset). The DOI (or other unique specifier system) specification will not itself include all the metadata that a data host knows about the data. Thus, it would be better that the data pointer was a wrapper for the image and its known metadata. I think the BIDS structure (with our without its NIDM specification) would be a good target wrapper for this data pointer. One could also imagine XCEDE, or any of a number of other wrapper systems. Each dataset, via it's unique identifier, would also have a 'landing page' where the permitted metadata could be exposed, but access to the data itself regulated by a variety of data use agreements. This is in very close analogy to how DOI's for publications work.

I would like to experiment with this, in the XNAT instance of NITRC-IR and would like to use a beta (alpha?) version of the BIDS specification with a minimal metadata set to represent 'simple' datasets of structure, diffusion and resting-state example cases. Even if this fails, I think it will be important to learn how it fails so that we can build a more robust way to foster these data unique identifiers.

More thoughts will be forming on this in a separate document for imaging data DOIs.
https://docs.google.com/document/d/1nH4MECcQ1G1W70vfkr7rACFBYRWnsJY4mAKlW3Obee0/edit

Proposal:
Multiple event files. Each needs onset and duration, but can include extra columns (different for each file). Events will have optional but unique suffixes. Devs can do events*.tsv.

On the storage of continuous "recordings" in TSV files (Michael Hanke)

Compression:
I propose to permit compressed storage for data files of continuous recordings. Nobody edits these things by hand of you have tens of thousands of lines and the storage/bandwidth benefits are significant. A quick test on my data shows that for cardiac/respiratory recordings the compressed file size is typically <10% (gzip) or <5% (XZ) of the original uncompressed TSV. GZIP support is already required for NIfTI IO, so it could be a good compromise.

Proposal: Switch from JSON to YAML

Here is a dataset description JSON file that I made by hand to get an impression what users would need to do:

```
{
	"Name": "studyforrest",
	"Description": "This dataset contains versatile brain imaging data for natural auditory stimulation and real-life cognition. It includes high-resolution functional
```

magnetic resonance (fMRI) data from 20 participants recorded at high field strength (7 Tesla) during prolonged stimulation with an auditory feature film ('Forrest Gump'). In addition, a comprehensive set of auxiliary data (T1w, T2w, DTI, susceptibility-weighted image, angiography) as well as measurements to assess technical and physiological noise components have been acquired. Participants were also scanned with a musical genre stimulation paradigm...",
"License": "PDDL",
"Authors": [
"Michael Hanke",
"Florian J. Baumgartner",
"Pierre Ibe",
"Falko R. Kaule",
"Stefan Pollmann",
"Oliver Speck",
"Wolf Zinke",
"Jörg Stadler",
"Annika Labs",
"Theresa Reich",
"Helene Schulenburg",
"Manuel Boennen",
"Mareike Gehrke",
"Madleen Golz",
"Benita Hartigs",
"Nico Hoffmann",
"Sebastian Keil",
"Malú Perlow",
"Anne Katrin Peukmann",
"Lea Noell Rabe",
"Franca-Rosa von Sobbe",
"Richard Dinga",
"Christian Häusler",
"J. Swaroop Guntupalli",
"Michael Casey"
],
"Acknowledgements": "We are grateful to the authors of the German 'Forrest Gump' audio description that made this study possible and especially Bernd Benecke for his support. We also want to thank Schweizer Radio und Fernsehen and Paramount Home Entertainment Germany for their permission to use the movie and audio description for this study. Thanks also go to Andreas Fügner and Marko Dombach for their help with developing the audio stimulation equipment, Renate Körbs for helping with scanner operations. Only open-source software was employed in this study. We thank their respective authors for making it publicly available.",
"HowToAcknowledge": "Please follow good scientific practice by citing the most appropriate publication(s) describing the aspects of this datasets that were used in a study.",
"Funding": "This research was funded by the German Federal Ministry of Education and Research (BMBF) as part of a US-German collaboration in computational neuroscience (CRCNS), co-funded by the BMBF and the US National Science Foundation (BMBF 01GQ1112; NSF 1129855). Michael Hanke was supported by funds from the German federal state of Saxony-Anhalt, Project: Center for Behavioral Brain Sciences.",

```
	"References": [
		"Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W. & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Scientific Data, 1:140003.",
		"Labs, A., Reich, T., Schulenburg, H., Boennen, M., Gehrke, M., Golz, M., Hartings, B., Hoffmann, N., Keil, S., Perlow, M., Peukmann, A. K., Rabe, L. N., von Sobbe, F.-R. & Hanke, M. (2015). Portrayed emotions in the movie “Forrest Gump”. F1000Research, 4:92.",
		"Hanke, M., Dinga, R., Häusler, C., Guntupalli, J. S., Casey, M., Kaule, F. R. & Stadler, S. (2015). High-resolution 7-Tesla fMRI data on the perception of musical genres – an extension to the studyforrest dataset. F1000Research, 4:174."
	],
	"VersionHistory": {
	"1": [
		"Initial release (Jan 22 2014)"
	],
	"2": [
		"Physiological data fixes and additions (Feb 20 2014)",
		"physiological data for all participants in original sampling rate (physio_pristine.txt.gz) was added",
		"physiological data for sub008 run005 was updated to strip leading data samples prior to the fixes MR",
		"trigger signal. Thanks to Christine Guo for the report",
		"rigger log for first MR trigger (only) was offset by one data sample (5-10ms)"
	],
	"3": [
		"Bug fixes, aggregate data (Jun 16 2014)",
		"Individual brain mask images from non-linear alignment had only the linear transformation component applied to them. Consequently the masks did not match the brain outline. This release contains the fixed brain mask images. No other data was modified.",
		"Include aggregate data (mean, median, min, max, std) for more than 200 regions-of-interests taken from the Harvard-Oxford cortical and sub-cortical atlas, as well as the Juelich brain atlas shipped with FSL. See the scripts in 'scripts/parcellation_ts/' in gumpdata source code repository for details. Data is available in CSV and HDF5 format."
	],
	"4": [
		"NIfTI header fixes (Sep 29 2014)",
		"task-nback BOLD images reported a wrong volume repetition time of ~24s in the NIfTI header field pixdim[4] (MRIConvert issue). This release updates the images to the correct value of 2.0s."
	],
	"5": [
		"task002 data added; emotion annotation, fixes (Apr 06 2015)",
		"New 7T fMRI data with music stimulation (task002).",
		"New group and per-subject EPI template images based on all available 7T data.",
		"Re-aligned task-nback BOLD data using the new templates.",
```

```
            "Updated, now frame-accurate annotation of movie scene boundaries.",
            "Frame accurate camera shot annotation for the audio-visual movie
(task003).",
            "Multi-observer emotion annotation for the audio-visual and audio-movie
stimulus (see Reich et al. F1000Research 2015)."
      ],
      "6": [
            "task002 bug fix (Jun 15 2015)",
            "Strip all samples prior to the first trigger from task002 physio data to
homogenize with task-nback.",
            "Physio data for task002, sub005 runs 3 and up were mis-numbered and
run008 was missing, due to a conversion error that did not take the aborted run into
account. All runs are available and correctly numbered now.",
            "Remove invalid log files for behavioral data of sub020, task002,
runs005-8. Subject aborted after run004.",
            "Add acquisition protocol for task002.",
            "Add task002 experiment implementation and stimulus files.",
            "Homogenize task002 stimulus file names with release experiment
implementation."
      ],
      "7": [
            "task002 stimulus auditory features (Jun 17 2015)",
            "Add extracted auditory features for task002 stimuli (incl. source code).
Thanks to Michael Casey for this contribution."
      ]
}
}
```

The raw file is here: http://kumo.ovgu.de/~mih/dataset_description.json

Overall it was an unpleasant experience. It took me almost an hour and 8 validator runs to get it JSON compliant. The most frustrating thing is the fact that JSON cannot handle multi-line strings, hence no changelog items without lists. Consequently, the missing line-break are not a mistake, but an effect of JSON's limitations.

Next I created the same description in YAML. Took less than a quarter of the time, and only 3 validator runs to figure out that YAML is more picky re indentation than Python. Below is how a full description would look like. Based on that I'd propose to reconsider the choice of JSON.

```
---
# one can have comments
Name:         studyforrest  # even here

# the '>' removes newlines from multi-line text and treats empty lines like
# paragraphs, similar to latex
Description: >
    This dataset contains versatile brain imaging data for natural auditory
    stimulation and real-life cognition. It includes high-resolution functional
```

magnetic resonance (fMRI) data from 20 participants recorded at high field strength (7 Tesla) during prolonged stimulation with an auditory feature film ('Forrest Gump'). In addition, a comprehensive set of auxiliary data (T1w, T2w, DTI, susceptibility-weighted image, angiography) as well as measurements to assess technical and physiological noise components have been acquired.

Participants were also scanned with a musical genre stimulation paradigm...

the empty lines between the fields are optional
License: PDDL

Authors:
- Michael Hanke
- Florian J. Baumgartner
- Pierre Ibe
- Falko R. Kaule
- Stefan Pollmann
- Oliver Speck
- Wolf Zinke
- Jörg Stadler
- Annika Labs
- Theresa Reich
- Helene Schulenburg
- Manuel Boennen
- Mareike Gehrke
- Madleen Golz
- Benita Hartigs
- Nico Hoffmann
- Sebastian Keil
- Malú Perlow
- Anne Katrin Peukmann
- Lea Noell Rabe
- Franca-Rosa von Sobbe
- Richard Dinga
- Christian Häusler
- J. Swaroop Guntupalli
- Michael Casey

Acknowledgements: >
We are grateful to the authors of the German 'Forrest Gump' audio description that made this study possible and especially Bernd Benecke for his support. We also want to thank Schweizer Radio und Fernsehen and Paramount Home Entertainment Germany for their permission to use the movie and audio description for this study. Thanks also go to Andreas Fügner and Marko Dombach for their help with developing the audio stimulation equipment, Renate Körbs for helping with scanner operations.

Only open-source software was employed in this study. We thank their respective authors for making it publicly available.

HowToAcknowledge: >
Please follow good scientific practice by citing the most appropriate publication(s) describing the aspects of this datasets that were used in a study.

Funding: >
This research was funded by the German Federal Ministry of Education and Research (BMBF) as part of a US-German collaboration in computational neuroscience (CRCNS), co-funded by the BMBF and the US National Science Foundation (BMBF 01GQ1112; NSF 1129855). Michael Hanke was supported by funds from the German federal state of Saxony-Anhalt, Project: Center for Behavioral Brain Sciences.

References:
- Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W. & Stadler, J. (2014). A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Scientific Data, 1:140003.
- Labs, A., Reich, T., Schulenburg, H., Boennen, M., Gehrke, M., Golz, M., Hartings, B., Hoffmann, N., Keil, S., Perlow, M., Peukmann, A. K., Rabe, L. N., von Sobbe, F.-R. & Hanke, M. (2015). Portrayed emotions in the movie “Forrest Gump”. F1000Research, 4:92.
- Hanke, M., Dinga, R., Häusler, C., Guntupalli, J. S., Casey, M., Kaule, F. R. & Stadler, S. (2015). High-resolution 7-Tesla fMRI data on the perception of musical genres – an extension to the studyforrest dataset. F1000Research, 4:174.

VersionHistory:
1:
- Initial release (Jan 22 2014)

2:
- Physiological data fixes and additions (Feb 20 2014)
- physiological data for all participants in original sampling rate (physio_pristine.txt.gz) was added
- physiological data for sub008 run005 was updated to strip leading data samples prior to the fixes MR trigger signal. Thanks to Christine Guo for the report
- trigger log for first MR trigger (only) was offset by one data sample (5-10ms)

3:
- Bug fixes, aggregate data (Jun 16 2014)
- Individual brain mask images from non-linear alignment had only the linear transformation component applied to them. Consequently the masks did not match the brain outline. This release contains the

```
      fixed brain mask images. No other data was modified.
    - Include aggregate data (mean, median, min, max, std) for more than
      200 regions-of-interests taken from the Harvard-Oxford cortical and
      sub-cortical atlas, as well as the Juelich brain atlas shipped with
      FSL. See the scripts in 'scripts/parcellation_ts/' in gumpdata
      source code repository for details. Data is available in CSV and
      HDF5 format.
  4:
    - NIfTI header fixes (Sep 29 2014)
    - task-nback BOLD images reported a wrong volume repetition time of ~24s
in
      the NIfTI header field pixdim[4] (MRIConvert issue). This release
      updates the images to the correct value of 2.0s.
  5:
    - task002 data added; emotion annotation, fixes (Apr 06 2015)
    - New 7T fMRI data with music stimulation (task002).
    - New group and per-subject EPI template images based on all available
      7T data.
    - Re-aligned task-nback BOLD data using the new templates.
    - Updated, now frame-accurate annotation of movie scene boundaries.
    - Frame accurate camera shot annotation for the audio-visual movie
      (task003).
    - Multi-observer emotion annotation for the audio-visual and audio-movie
      stimulus (see Reich et al. F1000Research 2015).
  6:
    - task002 bug fix (Jun 15 2015)
    - Strip all samples prior to the first trigger from task002 physio data
      to homogenize with task-nback.
    - Physio data for task002, sub005 runs 3 and up were mis-numbered and
      run008 was missing, due to a conversion error that did not take the
      aborted run into account. All runs are available and correctly
      numbered now.
    - Remove invalid log files for behavioral data of sub020, task002,
      runs005-8. Subject aborted after run004.
    - Add acquisition protocol for task002.
    - Add task002 experiment implementation and stimulus files.
    - Homogenize task002 stimulus file names with release experiment
      implementation.
  7:
    - task002 stimulus auditory features (Jun 17 2015)
    - Add extracted auditory features for task002 stimuli (incl. source
      code). Thanks to Michael Casey for this contribution.
```

Question:

How does this specification handle data sets where scans were collected at multiple sites? Does each site have its own directory hierarchy, etc? If not, and everything would go in the

same hierarchy, I suppose this information could be included in an optional additional JSON file, but it would be useful to have it included in file or directory names.

Ideas:

- provenance for attributes: added manually or inferred from DICOMS?