BIDS Extension Proposal BIDS-MEGA (BEP035): Modular extensions for individual participant data mega-analyses with non-compliant derivatives

version 0.1.0

Extension moderators/leads: Giuseppe Gallitto <giuseppe.gallitto@uk-essen.de>, Balint Kincses <balint.kincses@uk-essen.de>, Christian Büchel <buechel@uke.de>, Ulrike Bingel <ulrike.bingel@uk-essen.de>, Tor Wager <tor.d.wager@dartmouth.edu>, Tamas Spisak <tamas.spisak@uni-due.de>

Contributors: Thomas Nichols <thomas.nichols@bdi.ox.ac.uk>, Chris Markiewicz <markiewicz@STANFORD.EDU>, Anderson Winkler <anderson.winkler@nih.gov>, Cyril Pernet <cyril.pernet@ed.ac.uk>, Remi Gau <remi.gau@gmail.com>, Yaroslav Halchenko <yoh@dartmouth.edu>, Taylor Salo <tsalo006@fiu.edu>, Kay Robbins <kay.robbins@utsa.edu>

Example Datasets:

https://udue.de/BEP035examples

Contents

Outline

Proposal Structure

Contributor Guidelines

Relationship between BIDS-MEGA and BIDS v1.7

Terminology

Module A. The meta-BIDS directory
- Challenge
- Solution Concept

Module B. The BIDS mapper sidecar
- Challenge
- Solution Concept

Module C: Mega-entities and synergies with Modules A and B
- Challenge
- Solution Concept

User Stories

General Remarks
- Non-invasive consolidation vs. full consolidation
- Dataset Referencing for easy sharing and to avoid dataset proliferation
- Hierarchical Event Descriptors and BIDS-MEGA
- Links to BEP028 Provenance

Detailed Specification Proposals
- Module A: the meta-BIDS directory
- Module B: the BIDS mapper sidecar
- Module C: the BIDS mapper sidecar

Discussion History

List of proposed changes to the specification

Outline

This specification aims to extend the Brain Imaging Data Structure (BIDS) specification for individual p*articipant data (IPD) meta- and mega-analyses* of raw and first-level derivative data.

In contrast to coordinate-based meta-analyses of neuroimaging data, image-level individual-participant data (IPD) meta-analyses (Emmert et al, 2016, Zunhammer et al, 2018, Zugman et al, 2020; Zunhammer et al, 2021, He et al, 2021) require that the authors of different studies share participant-level (raw or derivative) data with the researchers conducting the mega-analysis. To maximize the authors' willingness-to-share (and thereby the data available for the mega-analysis), the formal/structural requirements for the data being shared must be as permissive as possible and the researchers conducting the mega-analysis must take over much of the data consolidation efforts from the authors of the single studies.

As a result, such mega-analyses must often deal with the high heterogeneity of the individual datasets, possibly encompassing anything from raw (nifti) data to BIDS-compliant and non-compliant derivative folders or even custom hand-picked collections of derivative files. Handling such heterogeneous multi-study datasets (including data consolidation) is currently not sufficiently covered by BIDS.

This modular extension proposal describes three *fully backward compatible* extensions to the BIDS specification and discusses how together they could provide full support for a wide variety of mega-analysis datasets.

Part A of the document proposes a new (optional) top-level mega-analysis dataset directory (strongly building on analogies with the BIDS-raw specification) that encompasses multiple study-level BIDS folders.

Part B proposes a new "mapping" mechanism for data and meta-data, complementary to that currently offered by BIDS, that allows seamless integration of non-BIDS compliant derivative folders.

Part C establishes synergies between the independent extensions A and B, so that together they allow lightweight, non-invasive consolidation of complex mega-analysis datasets, with maximal accessibility, human readability and without unnecessary dataset proliferation.

Altogether, BIDS-MEGA offers a set of extensions that may remove unnecessary overhead from the data-consolidation process of IPD mega-analysis datasets and thereby, enhances the applicability of BIDS in the case of analyses encompassing a highly heterogeneous collection of studies.

Proposal Structure

BIDS-MEGA (BEP035) follows a modular structure, with the idea that its modules can be discussed and reviewed in parallel so that they do not block each other in getting integrated into the official BIDS specification, allowing rapid integration of the most straightforward elements of the extension.
Specifically, **Module A** (the meta-BIDS directory) and **Module B** (The BIDS-mapper) are independent of each other and functional on their own. **Module C** builds on module A and B to establish the full support for IPD meta- and mega-analyses.
The proposal starts with brief descriptions of all three modules that follow the same layout: first the **Challenge** corresponding to the module is described, then the general **Solution Concept** is outlined (without concrete specification details). Finally, a reference to the **Detailed Specification** Proposal for the module is given.
All three modules are accompanied with numerous examples, to illustrate the Challenge and the Solution Concept and to serve as a basis for discussions.

In the end of the Challenge and the Solution Concept sections, there is a **Consensus Statement** block, listing the points on which the community needs to agree, in order to proceed. Contributors are expected to either “sign” the consensus statement, by leaving a comment with the text “agreed” or, alternatively specify any issue that prevents consensus about the proposal in its actual form.

Contributor Guidelines

This is a working document in draft stage, everyone is invited to leave comments. In order to become an ‘official’ contributor, i.e. to be listed as a co-author of the proposal document (and in the potential paper), the commenter must:

- contribute to this document by “signing”, or commenting on, at least one of the consensus statements.
- have a documented professional experience with meta-, mega- or multi-center analysis of neuroimaging data, non BIDS-compliant derivative data or comparable,
- disclose his/her name and affiliation.

The list of contributors will be maintained continuously by the extension leads, based on the above criteria. Contributors can ‘opt out’ from being publicly listed in the document/publication at any time.

Relationship between BIDS-MEGA and BIDS v1.7

BIDS-MEGA is a fully backward compatible extension of BIDS v1.7 and does not override any part of it. Wherever possible BIDS-MEGA applies terminology already used in BIDS v1.7. All extensions proposed by BIDS-MEGA follow the general ‘BIDS-philosophy’; from ensuring both human and machine readability to preferring minimalistic solutions with a ’gentle learning curve’. Wherever possible, the extensions in BIDS-MEGA are based on analogies to the original BIDS and harmonized with other BIDS Extension proposals (BEPs).

Terminology

Most core principles of the original BIDS-Raw specification are inherited by the BIDS-MEGA specification, though some special considerations and additional fields are noted below.
As in all versions of BIDS, in this specification the keywords "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

BIDS-MEGA uses the “required”, “recommended” and “optional” keywords throughout the document, including the description of json fields:

- REQUIRED: essential to be BIDS compliant (i.e. MUST as per RFC2199)
- RECOMMENDED: gives a warning if not present (i.e. SHOULD as per RFC2199)
- OPTIONAL: no warning if missing (i.e. MAY as per RFC2199)

As in BIDS-Raw, the following apply:

- All specifications of paths need to use forward slashes.
- The inheritance principle applies: any metadata file (.json, .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 details see BIDS-Raw (The Inheritance Principle).

The data structure-related terminology used in this specification is either inherited from the current version of the BIDS specification or can be considered as a natural extension of it. Below we list some of the most important concepts for this proposal:

- **raw data**: unprocessed or minimally processed (e.g. file format conversion) data; source of derivative data (see below), as described here.
- **derivative data**: data generated from raw data by various analysis pipelines/neuroimaging software tools, as also described in here.
- **non-compliant derivative**: derivative data stored inside of the derivatives folder of a BIDS dataset but in a format that does not (fully) comply with the BIDS specification, as also described here and here.
- **mega-analysis-level derivatives**: derivative files that store the results of a mega-analysis, generated from raw or processed (derivative) individual participant data by various analysis pipelines.
- **first-level analysis**: analysis of raw data of a single participant, producing first-level derivatives.
- **second-level analysis**: analysis of first-level derivatives, produces results representative of the study population.
- **third-level analysis**: analysis of second-level derivatives to produce mega-analysis-level derivatives (a common approach in mega-analysis).
- **entities**: key-value pairs used for specifying meta-data throughout the BIDS-specification, as described here and here.
- **data consolidation**: the process of taking data from disparate (possibly independent) sources, cleaning it up, and combining it in a single location.

Module A. The meta-BIDS directory

Challenge

The current version of BIDS (v1.7) recommends two options for storing multi-site datasets: (i) treat each center as a separate dataset, (ii) combine centers into a single dataset. While these solutions might satisfy the requirements for homogeneous, centrally orchestrated multi-site datasets (e.g. prospective multi-center trials), currently, none of these options seem to be feasible for heterogeneous multi-center datasets consisting of data from independent studies, (referred to as "multi-study" datasets, see **Ex A1**).
Specifically, option (i) handles centers as totally independent datasets, with no specification for storing overarching metadata and no directory to host multi-center (mega-analysis level) derivatives.
Option (ii), on the other hand, forces all studies to share certain meta-data (having a single dataset_description.json). While this might be a feasible approach for multi-site projects with centrally orchestrated harmonization measures (e.g. prospective multi-site trials), it may be suboptimal for heterogeneous datasets consisting of largely independent studies (e.g. mega-analyses with retrospective data collection). Namely, with this approach it is difficult to reference single centers as individual datasets and to resolve conflicts arising from between-center heterogeneity. Moreover it is unclear where to store center-specific derivatives (e.g. within-center statistical summaries as often used in meta-analyses). This approach does also not explicitly support one-to-many relations between single-center and multi-center datasets; if data from a single center is included in multiple multi-center datasets, it requires storing duplicate variants of the data, that may be slightly different due to project specific data consolidation requirements and lead to unnecessary dataset proliferation.

Example A1: We want to analyze IPD first-level activation maps from 3 independent studies together. The analysis consists of creating study-level (2nd-level) summary maps and then constructing a mean activation map across all studies (3rd-level). The studies used similar but not identical experimental procedures that we want to keep track of and account for in the 3rd-level analysis. We also want to make sure that the authors and the necessary references for the single studies are acknowledged in the dataset and the funding sources, license information and EthicsApprovals are properly stored for each study.

Consensus Statement A1.
I agree that storing heterogeneous "multi-study" datasets is not fully covered by BIDS v1.7.0.

Solution Concept

BEP035A proposes to represent heterogeneous multi-study datasets as a 'BIDS directory of BIDS directories' (the meta-BIDS directory). The proposed approach reflects the natural hierarchy of such datasets and fits well to meta-analysis ('analysis of analyses') approaches. The meta-BIDS directory provides straightforward analogies: studies in the meta-BIDS directory are analogous to participants in the regular single-study BIDS directories. Specifically the top-level meta-bids directory MUST contain regular BIDS directories (one per

study), and a dataset_description.json file, SHOULD contain a studies.tsv (analogous to participants.tsv) and MAY contain a studies.json (analogous to participants.json) and a derivatives folder (e.g for mega-analysis derivatives). These analogies make such a meta-BIDS dataset easily accessible for all users who have prior experiences with single-study BIDS directories.

Solution for example A1

Directory-tree hosting the dataset example A1.

```
my_megaanalysis
├── dataset_description.json
├── studies.json
├── studies.tsv
├── derivatives
│    ├── nimare-0.0.10
│    :
│
├── study-doe2012
│     ├── dataset_description.json
│     ├── participants.json
│     ├── participants.tsv
│     ├── derivatives
│     ├── sub-001
│     ├── sub-002
│     :
│
├── study-mustermann2017
│     ├── dataset_description.json
│     ├── participants.json
│     ├── participants.tsv
│     ├── derivatives
│     ├── sub-001
│     ├── sub-002
│     :
│
├── study-smith2015
:     ├── dataset_description.json
      ├── participants.json
      ├── participants.tsv
      ├── derivatives
      ├── sub-001
      ├── sub-002
      :
```

For more details see the [detailed specification proposal BEP035A].

Consensus Statement A2.
I agree that an optional top-level meta-BIDS folder, as described in BEP035A v0.1.0, would be a useful extension of the BIDS specification, to handle multi-study datasets.

Module B. The BIDS mapper sidecar

Challenge

The BIDS specification (v1.7) provides a comprehensive set of so-called entities (key-value pairs) for specifying values for a certain type of meta-data (stored either in file and directory names, in json sidecars or in the file participant.tsv).
However, in certain datasets, (i) some data files might not be explicitly assigned to any meta-data (a typical example is the case of non-compliant derivative folders **Ex. B1-2**) (ii) the precise terminology used for the key-value pairs might vary across studies or (iii) storage of non-standard, project-specific meta-data might be needed. Such cases can be problematic not only in multi-study settings, but in regular, single-study BIDS datasets with derivatives produced by commonly used software tools, too. Module B focuses on case (i), but the proposed solution is easily extendable to tackle cases (ii) and (iii), see Module C for details.

Examples
Ex. B1 Non BIDS-compliant derivative folder as output by popular software tools:
study-01/derivatives/freesurfer-7.2/sub-001/mri/aseg.mgz
This file should be linked to e.g. the key-value pairs: 'space-fsaverage' (see BIDS appendix) and possibly the suffix 'dseg' (according to BEP011).

Ex. B2 Non BIDS-compliant, custom made derivative folders, e.g. a hand-picked collection of beta contrast images from 1st level task-fMRI processing.
study-01/derivatives/fsl-feat-3.3/sub-001_cope1.nii.gz
We should be able to map these files to e.g.: 'space-fsaverage', 'task-pain', 'sess-baseline', and, possibly, the Hierarchical Event Descriptor (HED) 'Sensory-event, Experimental-stimulus, Hot, Pain'.

See also: **Ex. C1-6** in Module C.

Consensus Statement B1.
I agree that working with non BIDS-compliant derivative folders and resolving meta-data discrepancies across various BIDS datasets is, to date, problematic.

Solution Concept

This challenge can be tackled by a dedicated (optional) mapper sidecar file, that can establish the required mappings between data and meta-data, without breaking the original data structure. The proposed 'mapper sidecar' should simply list the BIDS key-value pairs and the data files they need to be mapped to. See **Module C**, for more details on handling between-study discrepancies and custom meta-data with the mapper concept.

Solution to Ex. B1:
A dedicated (optional) json sidecar file can directly map the entity 'space-fsaverage' to the file study-01/derivatives/freesurfer-7.2/sub-001/mri/aseg.mgz:

study-01/derivatives/freesurfer-7.2/bids_mapper.json

```
{
        "File": "sub-001/mri/aseg.mgz"
        "Entity": "space-fsaverage_T1w_dseg"
}
```

Solution to Ex. B2:
A dedicated (optional) json sidecar file can directly map the entities 'space-native', 'task-pain', 'sess-baseline', as well as the required HED-tag (optionally) to: *study-01/derivatives/fsl-feat-3.3/sub-001_cope1.nii.gz.*

study-01/bids_mapper.json

```
{
        "File":"derivatives/fsl-feat-3.3/sub-*_cope*.nii.gz"
        "Entity": "space-individual_task-pain_sess-baseline"
        "HED": "Sensory-event, Experimental-stimulus, Hot, Pain"
}
```

For more detail, please refer to the detailed specification proposal for Module B, which among others, outlines how the mapper files can be made extremely powerful by adapting bash wildcards or regular expressions.

Consensus Statement B2.
I agree that the BIDS mapper sidecars, as proposed by BEP035 Module B (v0.1.0) provides a straightforward way of handling non BIDS-compliant (derivative) folders.

Module C: Mega-entities and synergies with Modules A and B

Challenge

Although the top-level meta-BIDS directory, proposed in **Module A**, may be sufficient on its own for multi-study datasets with high homogeneity (e.g. centrally orchestrated multi-center trials), it is not able to resolve all issues arising from larger between-study heterogeneity (e.g. IPD meta-analyses of independent studies). Characteristic examples are shown in **Ex. C1-5**.

Specifically, if multiple, independent studies are considered together, values of certain entities (key-value pairs) or certain pieces of participant information might be incompatible.

For instance, as highlighted by case (ii) in part B, entity values representing conceptually the same meta-data may be different from dataset to dataset due to differences in the naming conventions (See **Ex. C1-3** for key-value pairs and Ex. **C4-5** for participant information).

Moreover, as highlighted by case (iii) in part B, in a mega-analysis context, it may be desirable to store custom, project specific meta-data, that - as opposed to general-purpose meta-data - is not supposed to be covered by the BIDS specification. **Ex. C6** provides an illustration of such situations.

In the lack of clean conventions, keeping track of incompatible or custom meta-data may be detrimental to the accessibility of mega-analysis datasets and result in unnecessary dataset proliferation (multiple versions of the same dataset).

Examples

Ex. C1 Different names for (conceptually) the same task in task-fMRI:
meta-analysis/study-01/sub-001/func/sub-001_task-pain_bold.nii.gz
meta-analysis/study-02/sub-001/func/sub-001_task-heatpain_bold.nii.gz

Ex. C2 Different names for various runs in rsfMRI:
meta-analysis/study-01/sub-001/func/sub-001_task-rest_run-1_bold.nii.gz
meta-analysis/study-02/sub-001/func/sub-001_task-rest_run-baseline_bold.nii.gz

Ex. C3 Repeated measures might be represented with runs in one study and with sessions in another. An example extending Ex. C2:
meta-analysis/study-03/sub-001/func/sub-001_task-rest_session-baseline_bold.nii.gz

Ex. C4 The name for a certain type of between-subject information is given with different column names in the participants.tsv of different studies. E.g. Pharmacological treatment is referred to as ‘medication’, ‘drug’, or <name-of-drug> or simply “group” in different studies.

Ex. C5 Conceptually identical levels of a between subject factor are referred to with different names in the participants.tsv of different studies. E.g. a typical control group can be referred to as ‘control’, ‘ctr’, ‘healthy control’, ‘HC’, ‘neurotypical control’, ‘NTC’, ‘saline’, ‘treatment-as-usual’, ‘TAU’, etc.

Ex. C6 In the placebo metaanalysis consortium, first level task-based fMRI beta or contrast images have been collected from independent studies. Collected images can be ‘pain’ responses with or without placebo intervention (‘contrast-pain-control’, ‘contrast-pain-placebo’, respectively) as well as contrast images showing placebo related

activity (‘contrast-placebo’) that are already constructed by contrasting matching ‘pain-control’, ‘pain-placebo’ images.
In such a dataset, we might face this situation:

meta-analysis/study-01/derivatives/fsl-feat-3.3/subject001.feat/stats/cope1.nii.gz
must be labeled as ‘contrast-pain-control’
meta-analysis/study-01/derivatives/fsl-feat-3.3/subject001.feat/stats/cope2.nii.gz
must be labeled as ‘contrast-pain-placebo’
meta-analysis/study-02/derivatives/fsl-feat-3.3/subject001.feat/stats/cope1.nii.gz
must be labeled as ‘contrast-placebo’

Ex. C7: Redundancy in case of overlapping mega-analysis datasets: let’s suppose that we have two mega-analysis datasets: mega-analysis X on the neural mechanisms beyond placebo and mega-analysis Y on pain anticipation. An exemplary study by Doe at all investigated placebo analgesia, with a task-fMRI paradigm that included an anticipation period before the pain stimulation. When sharing datasets X and Y with collaborators the dataset corresponding to the study by Doe at al. will be duplicated and, possibly, stored in different versions.

Consensus Statement C1.
I agree that handling multiple, independent studies in BIDS (even when considering BEP035 Modules A and B) is challenging due to differences in naming conventions.

Solution Concept

Project-specific entities (mega-entities)
To address issues with incompatible or custom meta-data, we propose a way to define overarching ‘project-specific entities’ (for short: ‘mega-entities’, hinting their usefulness in mega-analyses) that can be freely linked to any data or meta-data. A project-entity can represent any general or project-specific concept that is valid in the whole scope of the dataset, e.g. across all studies in the multi-study dataset. mega-entities are in many ways analogous to built-in BIDS-entities, but they are defined ad-hoc, in any of the dataset_description.jsons (but typically in the meta-BIDS directory proposed in Part A):

```
"MegaEntities":
[{
            “Key”: <project-entity-key>,
            “Values” : [<list of possible values>],
            “Description”: <description>
}]
```

With some extensions, the mapper sidecar proposed in **Part B** can map mega-entities to:

- **BIDS key-value pairs**, solving **Ex C1-3**.
  A project-entity “TASK-PAIN” can be linked to the entity ‘task-pain’ in study-01 and ‘task-heatpain’ in study-02 in **Ex C1**.
  A project-entity called “CONDITION-BASELINE” can be linked to ‘run-1’ in study-01 and ‘run-baseline’ in study-02 in **Ex C2**.
  A project-entity called “CONDITION-BASELINE” can be linked to ‘session-baseline’ in study-02 in **Ex C3**.

- **participants.tsv columns**, solving **Ex C 4-5**.
  A project-entity key called “TREATMENT” can be linked to the corresponding column names in all studies in **Ex. C4** and **5.**
  A project-entity called “TREATMENT-CONTROL” can be linked to the different factor level names in all studies in **Ex. C4** and **5.**

- **data files**, solving **Ex C6**.
  The project-entity ‘CONTRAST’ with values ‘‘PAIN-CONTROL’, ‘PAIN-PLACEBO’ and ‘PLACEBO’’ can be assigned to the files, as needed.

See the detailed specification proposal for more details.

Solution to Ex. C1.
meta-analysis/bids_mapper.json

```
[
  {
      "MegaEntity": "TASK-PAIN"
      "Entity": “task-pain",
      “Scope”: “study01”
  },
  {
      "MegaEntity": "TASK-PAIN"
      "Entity": “task-nociception",
      “Scope”: “study02”
  }
]
```

Solution to Ex. C2 AND C3.
meta-analysis/bids_mapper.json

```
[
  {
      "MegaEntity": "VISIT-BASELINE"
      "Entity": “run-1",
      “Scope”: “study01”
  },
  {
```

```
        "MegaEntity": "TASK-PAIN"
        "Entity": “run-baseline",
        “Scope”: “study02”
    },
    {
        "MegaEntity": "TASK-PAIN"
        "Entity": “session-baseline",
        “Scope”: “study03”
    }
]
```

Solution to Ex. C4.
meta-analysis/bids_mapper.json

```
[
    {
        "MegaEntity": "TREATMENT",
        "ParticipantInfo": “drug",
        “Scope”: “study01”
    },
    {
        "MegaEntity": "TREATMENT"
        "ParticipantInfo": “group",
        “Scope”: “study02”
    }
]
```

Solution to Ex. C5.
meta-analysis/bids_mapper.json

```
[
    {
        "MegaEntity": "TREATMENT-CONTROL"
        "ParticipantInfo": “drug-saline",
        “Scope”: “study01”
    },
    {
        "MegaEntity": "TREATMENT-CONTROL"
        "ParticipantInfo": “group-control",
        “Scope”: “study02”
    }
]
```

Solution to Ex. C6.
meta-analysis/bids_mapper.json

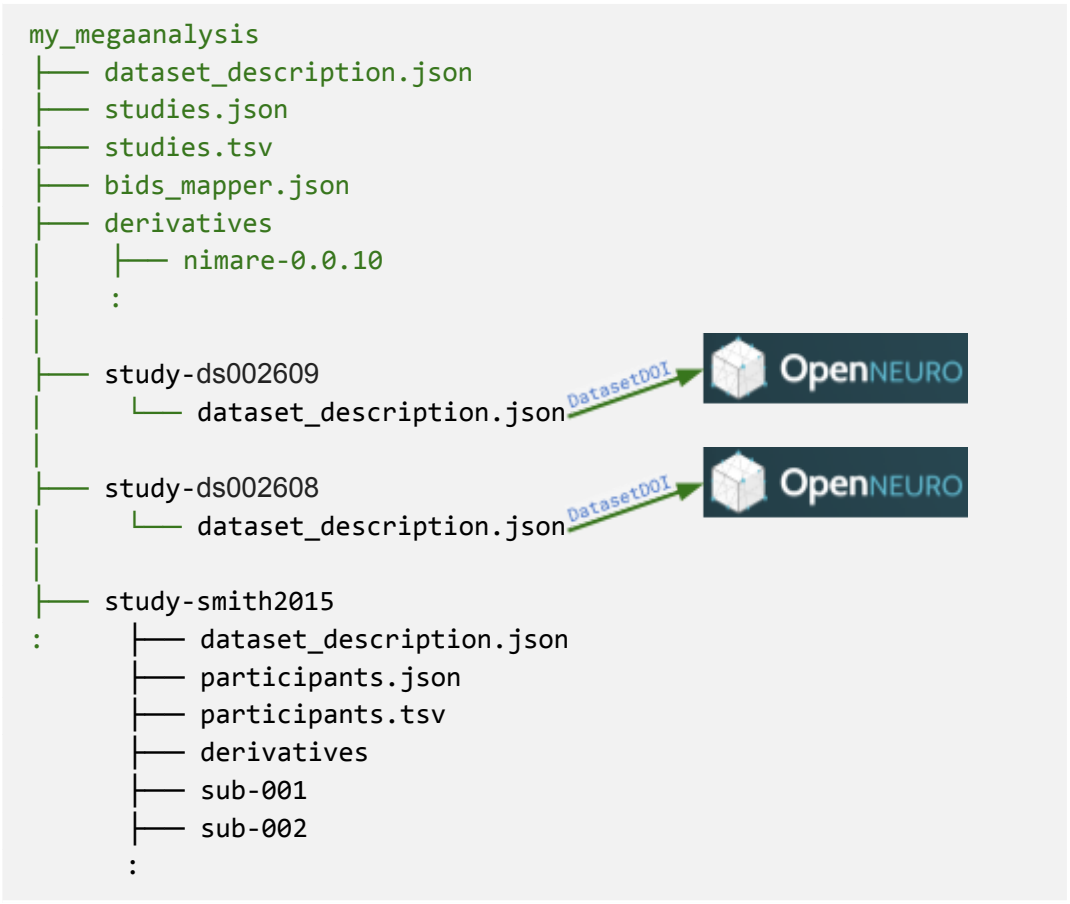
```
[
    {
        "MegaEntity": "CONTRAST-PAIN-CONTROL"
        "File": “derivatives/fsl-feat-3.3/subject001.feat/stats/cope1.nii.gz",
```

```
        "Scope": ["study01", "study3"]
    },
    {
        "MegaEntity": "CONTRAST-PAIN-PLACEBO"
        "File": "derivatives/fsl-feat-3.3/subject001.feat/stats/cope1.nii.gz",
        "Scope": ["study01", "study3"]
    },
    {
        "MegaEntity": "CONTRAST-PLACEBO"
        "File": "derivatives/fsl-feat-3.3/subject001.feat/stats/cope1.nii.gz",
        "Scope": "study02"
    }
]
```

Solution to Ex. C7.

When storing the bids_mapper.json sidecar in the mega-BIDS folder, i.e. directly at the top level, the single study folders are completely unchanged (non-invasive data consolidation). In this case one can share the mega-analysis dataset so that some of the study-folders contain only a dataset_description.json file and the dataset_description.json provides all details (doi, version) to access the dataset (e.g. from openneuro).

This allows effective sharing of mega-analysis datasets and avoiding duplicate datasets and dataset proliferation.



Consensus Statement C2.
I agree with the concept of Mega-entities for meta-data harmonization across multiple studies, as described BEP035 Module C (v0.1.0).

User Stories

The proposed extensions cover a large variety of use cases. In many cases, proper storage of a mega-analysis dataset might not require all the proposed features. Some examples, from less, to more complex:

1. Analysis of raw data from a centrally orchestrated multi-site study: in such studies, sequences, experimental procedures and naming conventions are often already harmonized to a large degree. Although the current specification (v1.7) might already cover this situation (see here), module A might provide a useful, clean alternative for such datasets (e.g. studies.csv, mega-analysis level derivatives). Module B and C are typically not required in such homogenous datasets.
2. Analysis of relatively homogenous raw data from independent studies: a typical example for this use-case is the pooled analysis of raw resting state fMRI data. Such data might be relatively simplistic in terms of data consolidation due to the simplicity of the paradigm. However, combining the datasets into a single BIDS folder, as recommended in the current specification (v1.7) may still require certain consolidation steps (e.g. handling identical participant IDs in two studies) that most probably results in changes of the original datasets (causing dataset proliferation) and might face licensing issues. This can be prevented with module A. Dataset referencing (see below) may be especially useful in such cases, by allowing easy sharing of potentially overlapping collections of studies.
3. Analysis of uniformly generated derivatives from centrally orchestrated studies: a typical example for this use-case is the analysis of anatomical derivative data, as generated by e.g. Freesurfer. Analyzing derivatives in IPD mega-analysis is often motivated by the unavailability of raw data (e.g. due to limitations of the ethical approval). While BIDS, and several BEPs, provide certain specifications for derivative data, the output of several popular software tools is, and expected to remain, incompatible with BIDS. For such cases, module B allows mapping BIDS-style meta-data to derivative data files and, thereby, improves the overall coherency of the dataset.
4. Analysis of heterogeneous raw data from independent studies: a typical example is a mega analysis of raw task-based fMRI data, already stored in BIDS. In this case, module B and Cs can resolve problems arising from different naming conventions (e.g. for task names). Dataset referencing (see below) may be especially useful in this case, too.
5. The “dirty” use case: literature-based retrospective mega analyses may manifest as a mix of the above cases. For instance, the Placebo Imaging Consortium involves behavioral data, raw anatomical and resting state functional data (use case 2) and first-level activation maps, generated from task-based fMRI with various software tools. Some studies involve only one modality (e.g. only anatomical), others multiple modalities. Raw data is often already available publicly (e.g. openneuro), in other cases the authors simply share a hand-picked collection of files. Modules A, B and C, together aim to cover even these complex use cases.

General Remarks

Non-invasive consolidation vs. full consolidation

While storing the bids_mapper.json sidecar in the mega-BIDS folder, i.e. directly at the top level allows fully non-invasive data consolidation, it can also serve as an input for automatic data-consolidation approaches that might create datasets that are fully compatible with the single-center BIDS specification and do not need bids_mapper.json sidecars anymore.

Dataset Referencing for easy sharing and to avoid dataset proliferation

As demonstrated in **Example C7**, the proposed extensions can take advantage of the uniform resource indicator (URI) in the study-specific dataset_description.json (as provided by the original specification) to reference datasets, without actually storing them. Based on the URI, the specified version of the given study can be fetched (e.g. from openneuro). Note that, in this case, all information regarding data consolidation must be stored in the mega-BIDS folder, in the files 'dataset_description.json' and 'bids_mapper.json'.
Such a lightweight representation prevents dataset proliferation, allows seamless sharing and storage-efficient handling of overlapping mega-analysis dataset.

Hierarchical Event Descriptors and BIDS-MEGA

The proposed extensions display a synergy with using Hierarchical Event Descriptors (HEDs) for tasks/events. Datasets already using HED-annotations can be seamlessly integrated into the mega-BIDS structure. Where needed, HED-annotations can be added post-hoc to certain files via bids_mapper.json sidecar, e.g. for first-level beta-images (generated from task-based fMRI) corresponding to a given type of event.

Links to BEP028 Provenance

The proposed extensions are fully compatible with, and complementary to, BEP028 'Provenance'. The single studies can have their own provenance structure and can be merged into arbitrary mega-analysis datasets, without any interference between their provenance structures, as both file and dataset level BIDS-Provs naturally retain their scopes. The whole mega-analysis, as well as its derivatives (mega-analysis/derivatives) can also be equipped with provenance, naturally extending the provenance graph. Furthermore, provenance information might provide input for the mega-analysis. In this case, the bids-mapper can be used to resolve discrepancies in the naming of *jsonld* files, stemming from e.g. different naming convention of the BIDS entities.

Detailed Specification Proposals

Module A: the meta-BIDS directory

```
<mega-analysis-dir>
├── dataset_description.json
├── studies.json
├── studies.tsv
├── derivatives
│   ├── <pipeline>[-<variant>]
│   :
│
├── study-<id1>
│     ├── dataset_description.json
│     ├── participants.json
│     ├── participants.tsv
│     ├── derivatives
│     │   └── <pipeline>[-<variant>]
│     ├── sub-001
│     ├── sub-002
│     :
│
├── study-<id2>
:     ├── dataset_description.json
      ├── participants.json
      ├── participants.tsv
      ├── derivatives
      │   └── <pipeline>[-<variant>]
      ├── sub-001
      ├── sub-002
      :
```

To link all studies in the mega-analysis, BIDS-MEGA extends the BIDS-raw specification with a new OPTIONAL top-level dataset directory (with its own dataset_description.json), encompassing all study-level BIDS directories (as nested datasets, with their own study-level dataset_description.json files) that are part of the mega-analysis. This folder implements the `mega-analysis level`, that is, it encompasses all data and meta-data for the mega-analysis. The structure of the proposed top-level directory builds on strong analogies with the ordinary 'study level' BIDS directories; studies at the mega-analysis level behave very similarly to subject in the study-level (e.g. participants.tsv -> studies.tsv, sub-001 -> study-001). See section 3.3 *File formats and json sidecars* for a detailed specification of data and meta-data files.

Study-level BIDS directories within the proposed top-level mega-analysis folder MUST fully comply with the BIDS-raw specification and can also be considered as standalone datasets. If the top-level directory exists, the name of the study directories MUST follow the

template study-<studylabel>, where study label is an alphanumeric label as defined in BIDS.

The results of the mega-analysis SHOULD be placed into a (newly proposed) 'derivatives' subfolder of this top-level mega-analysis dataset directory. Organizing the contents of this mega-analysis level derivative folder is analogous to how the BIDS-raw specification organizes the study-level derivatives. That is, a mega-analysis pipeline will typically have a dedicated sub-directory under which it stores all of its outputs. As at the study level (BIDS-raw), different components of a pipeline can also be stored under different subfolders. All pipeline folders can be considered as a standalone dataset (i.e. they SHOULD contain a dataset_description.json). For the naming of these directories, it is RECOMMENDED to use the format <pipeline>-<variant> in cases where it is anticipated that the same pipeline will output more than one variant. For the sake of consistency, the subfolder name SHOULD be the *GeneratedBy.Name[GeneratedBy.Version]* field(s) in data_description.json of the given pipeline-folder, optionally followed by a hyphen and a suffix.
Example path:

```
placebo-metaanalysis/derivatives/pain-GIV
```

If the top-level mega-analysis directory exists, it MUST contain the following data and meta-data files:

- ***dataset_description.json***: the mega-analysis level dataset description sidecar. File format is identical to the study-level and derivative pipeline-level dataset_description.json files (see: BIDS dataset description).
- :**studies.tsv** a tab separated file, which is analogous to participants.tsv of BIDS-raw, but on the study-level, i.e it contains data corresponding to the single studies. If the top-level mega-analysis directory does exist, this file is RECOMMENDED to exist too. If it exists, it MUST contain a column named "study_id" which MUST consist of study-<label> values corresponding to the name of the study-specific folders and identifying one row for each study. Each study MUST be described by one and only one row. Note that some data can be considered as meta-data within studies but becomes data at the mega-analysis level (e.g. MagneticFiledStrength). The file 'studies.tsv' provides a great flexibility; it MAY contain arbitrary columns for data that is considered as of interest in the mega-analysis. Whenever possible, tthe names of optional columns are RECOMMENDED to re-use the the key names listed in the sections 'Modality agnostic files' and 'modality specific files' in the BIDS-raw specification (e.g. 'RepetitionTime'). If demographic/behavioral variables that would normally reside in participants.tsv are not available for all participants in a study, but study-level aggregated information is available (e.g. from the publication text), columns called mean_<variable> or ratio_<variable> should be added for continuous or factor variables, respectively. We RECOMMEND to create such columns for all meta-data which would normally be available in the sidecars of the BIDS-raw specification but - for at least one study - the meta-data is unavailable.

This can typically happen if:

- only derivative data is available for certain studies (i.e. meta-data for raw data is also unavailable)
- some data is not available at the participant-level, (i.e. participants.tsv is incomplete)

study_ID	mean_age	ratio_female	Manufacturer	MagneticFieldStrength	tfMRI_RepetitionTime	stimulus_duration
study1	27	0.47	Siemens	3T	2000	100
study2	23.5	0.60	GE	1.5T	2600	5200
study3	n/a	0.55	Philips	3T	3000	6000
study4	38.6	0.41	Siemens	3T	3000	12000

In the above example, we have two columns providing aggregated descriptions of participant-level demographic data (mean age and the percentage of female participants) and two related to study-wise imaging parameters (scanner manufacturer and field strength) and two imaging-related columns (repetition time - TR - and stimulus_duration). Note that if raw data is available for a study, sequence parameters might be redundantly stored, but for studies where only derivative data is available for the mega-analysis, this might be the only location this information is stored.

> Note that, as mega-analyses themselves, studies.tsv is inherently goal-oriented, i.e. the choice of variables to be represented at this level may seem somewhat arbitrary. Conceptually, as participants.tsv is related to the second-level analyses, studies.tsv should be used to store data required by third-level analyses. However, his analogy is limited, as mega-analyses do not necessarily involve third-level analyses. A rule of thumb could be to store data here in an 'on-demand' manner, e.g. if participant-level information is not available in at least one of the studies (e.g. in "classical" meta-analysis, age might not be available on the participant-level, only as summary statistics, from the corresponding publication). Studies.tsv is optional and might not be needed at all in many applications.

studies.json: this json sidecar is OPTIONAL to accompany the studies.tsv file. In analogy with participants.json of BIDS-raw, it explains the columns of studies.tsv. if it exists, it MUST present the structure below, starting with a <column name> object.

Key name	Requirement level	Data type	Description
<column name>	REQUIRED	object	Name of the studies.tsv's column to describe.

This object contains the following keys, in analogy with the participants.json sidecar:

Key name	Requirement level	Data type	Description
LongName	OPTIONAL	string	Long (unabbreviated) name of the column.
Description	RECOMMENDED	string	Description of the column.
Levels	RECOMMENDED	object of strings	For categorical variables: An object of possible values (keys) and their descriptions (values).
Units	RECOMMENDED	string	Measurement units. SI units in CMIXF formatting are RECOMMENDED (see: tabular files).
TermURL	RECOMMENDED	string	URL pointing to a formal definition of this type of data in an ontology available on the web.

Here we present an example of studies.json:

```
{
   "study_ID": {
      "LongName": "ID of the Study",
      "Description": "string representing the name of the study",

      },
   "mean_age": {
      "LongName": "participants mean age",
      "Description": "mean of the age of all participants in the
study",
   },
   "ratio_female": {
      "Description":"ratio of female participants in the study"
      }
   },
   "tfMRI_repetitionTime": {
      "Description":"TR for of the task-fMRI sequence that is the
source of the contrast-of-interest in the meta-analysis",
      "Units": "ms"
      }
   },
   "stimulus_duration": {
      "Description":"Duration of the stimulation in the task-fMRI
paradigm that is the source of the contrast-of-interest in the
meta-analysis",
      "Units": "ms"
      }
   }
}
```

Module B: the BIDS mapper sidecar

DISCLAIMER: The proposed mapper sidecar SHOULD NOT be applied for making derivative folders of future versions of various software tools/workflows BIDS-compatible. Any future software release is instead encouraged to implement a fully BIDS-compliant derivative specification following (and extending) the derivative specification of BIDS-raw. The primary application area of the BIDS mapper sidecars is to make retrospective derivative data from non BIDS-compliant software output (or hand-picked collections of derivative data) accessible within a BIDS dataset, as typically required for mega-analyses.

The BIDS mapper sidecar is an OPTIONAL json sidecar file that MAY be placed at any level in a BIDS directory structure. In case of mega-analyses, it is RECOMMENDED to be placed at the top-level mega-analysis directory (see Modules A and C), for non-invasive harmonization. The scope of a BIDS mapper sidecar spans to the containing directory and all subdirectories, or alternatively, to an explicitly specified directory and its subdirectories (to be specified by the Scope keyword, see below). Multiple BIDS mapper sidecars are possible in a single BIDS directory, in this case the inheritance principle applies.

The BIDS mapper sidecars MAY contain a single json object or a list of json objects, with the following keys:

Key name	Requirement level	Data type	Description
File	OPTIONAL	string or array of string File and FileRegExp are mutually exclusive.	Relative path to the data unit to be mapped, can contain UNIX-wildcards.
FileRegExp	OPTIONAL	string	Selects file(s) matching this regular expressions
Entity	OPTIONAL	string or array of string	BIDS-like key-value pair(s). If multiple, can be given in a single string separated by '_' or as an array of strings.
HED	OPTIONAL	string	Hierarchical Event Descriptor (HED) Tag. See Appendix III of the original specification for details.
Description	OPTIONAL	string	Plain-text description of the mapping.
Scope	OPTIONAL	string or array of string	The folder(s) in which the mapper is valid (default: the directory containing the json file)

File and FileRegExp are mutually exclusive.
This is to be extended in Module C.

Examples

Ex. B1.
study-01/derivatives/freesurfer-7.2/bids_mapper.json

```
{
        "File": ["sub-*/mri/aseg.mgz", "sub-*/mri/T1.mgz"]
        "Entity": "space-fsaverage_T1w_dseg"
}
```

Ex. B2.
study-01/bids_mapper.json

```
{
        "File":“derivatives/fsl-feat-3.3-1/sub-*_cope*.nii.gz”
        "Entity": "space-individual_task-pain_session-baseline"
}
{
        "File":“derivatives/fsl-feat-3.3-2/sub-*_cope*.nii.gz”
        "Entity": "space-individual_task-pain_session-day2"
}
```

todo: examles for selecting subset of subjects via bash wildcards.

- sub2[1-9]: sub21-29
- sub0{1,2,3,4,6,7,8,9}, 01-19, except 5
- sub0{!5}, same, supposedly

todo: elaborate regexp support!
(e.g. to capture subject labels, to fill in the sub-XY entities from the filename/path?)

See: https://regex101.com/r/xBDi1W/1

```
{
        "FileRegExp":"sub.*?-(?P<sublabel>.*?)_"
        "Entity": "sub-\k<sublabel>"
}
```

Module C: the BIDS mapper sidecar

To be able to define mega-entities, the specification of dataset_description.json must be extended with an additional (OPTIONAL) keyword:

Key name	Requirement level	Data type	Description
MegaEntities	OPTIONAL	array	array of project-entity objects

Mapped to a list of objects with the following keys:

Key name	Requirement level	Data type	Description
Key	REQUIRED	string	Name of the project-entity key
Values	OPTIONAL	array of strings	Possible values for this key
Description	OPTIONAL	string	Free-form description of the project-entity

An example of a meta-BIDS level dataset_description.json file with project-keys:

```
{
      "Name": "<mega-analysis name>",
      "BIDSVersion": "<version>",
      "DatasetType": "mega-analysis",
      "Authors": [
                  "<Author 1>",
                  "<Author 2>",
                  "<Author 3>"
      ],
      "ReferencesAndLink": "<reference article>",
      "MegaEntities":
          [{
                  “Key”: “TASK”,
                  “Values” : [“PAIN”],
                  “Description”:    “Task    paradigm    involving
painful simulation”
           },
           {
                 “Key”: “TREATMENT”,
                  “Values” : [“CONTROL”, “DRUG”],
                  “Description”: “Type of treatment”
          }]
      }
}
```

Project-keys MUST be unique and distinct from existing BIDS entities.

Module C extends the BIDS mapper sidecar so that it can handle mega-entities:

Key name	Requirement level	Data type	Description
File	OPTIONAL	string or array of string	Relative path to the data unit to be mapped, can contain UNIX-wildcards.
FileRegExp	OPTIONAL	string	Selects file(s) matching this regular expressions
Entity	OPTIONAL	string or array of string	BIDS-like key-value pair(s). If multiple, can be given in a single string separated by '_' or as an array of strings.
HED	OPTIONAL	string	Hierarchical Event Descriptor (HED) Tag. See Appendix III of the original specification for details.
MegaEntity	OPTIONAL	string or array of string	One or more mega-entities, as defined in the meta-BIDS folder's dataset_description.json If multiple, can be given in a single string separated by '_' or as an array of strings.
ParticipantInfo	OPTIONAL	string	key-value pair as string, corresponding to the contents of participant.tsv
EventInfo	OPTIONAL	string	key-value pair as string, corresponding to the contents of *_events.tsv files (all files in scope).
Scope	OPTIONAL	string or array of string	The folder(s) in which the mapper is valid (default: the directory containing the json file)
Description	OPTIONAL	string	Plain-text description of the mapping.

File and FileRegExp are mutually exclusive.

At least two of File, FileRegExp, Entity, HED, MegaEntity, ParticipantInfo SHOULD be given, otherwise no mapping is performed.

Questions

- Matching only key but not value?
- threeway mapping (and more)
- better scope definition? (so that we don't need a separate mapping for all studies)

Discussion History

Q: How to accommodate a single study having multiple scanners, or multiple sequence variants, etc? - Anderson Winkler
A: In such cases the corresponding data must be available in the corresponding "participants.tsv", anyway.
In this case it should be investigated if representing that piece of information is neccessary at all. If yes, then "studies.tsv" should simply imply that this parameter is varying for this study.
The actual string to denote such "varying" parameters is to be agreed on. We will look up if there is any adaptable convention in any of the BEPs.
Input on this is welcome!
An alternative solution is (depending on the setting) to split the study into substudies and having multiple rows in studies.tsv. In this case of course consolidation is not non-invasive anymore.
In such questions it is useful to consider how BIDS works one level lower (as studies for BIDS-MEGA are analogous to participants in BIDS-raw).
For instance, if in an experiment a participant receives 10 stimuli, which are - in all but a few participants - delivered on the same side (left or right). But for a few participants, e.g. 5 was on the left and five was on the right side. How would participants.tsv look like in such a (probably not very well designed) study? I believe BIDS-raw allows multiple solutions, and so is BIDS-MEGA supposed to do, too.

Q: In this section (previous proposal version), is the idea to accommodate all types of potential multi-site analyses (meta-analysis with access to IPD, mega-analysis with tx of processed data, or mega with transfer of everything), but one at a time, or to allow a kind of mix-and-match? (please see this fig: https://onlinelibrary.wiley.com/doi/10.1002/hbm.25096#hbm25096-fig-0001)
I believe the former is better, even if more restrictive, because the latter may make statistics needlessly complicated. - Anderson Winkler
A: Original intention is to allow mixed situations and even having some of the derivatives in BIDS-compliant form, while others in the proposed “indexed” form.
In this sense, BIDS-MEGA should provide the same flexibility as BIDS-raw. BIDS-raw, in practice, allows (but does not encourage) a high heterogeneity of participant-level meta-data and a different level of preprocessing/analysis for various participants (at a given timepoint). In general, BIDS does allow “ill-posed” datasets to be valid BIDS datasets, too. BIDS-raw should not be more restrictive.
Responsibility regarding statistical validity should belong to the researcher and is out of the scope of this proposal. Although the proposed structure is intentionally tailored towards the overarching aims of the analysis, it is not supposed to put any restriction on *how* the analysis is done.
The top-level meta-analysis derivative foldes are meant to be highly analogous to those of the original BIDS specification, i.e. with minimal restrictions and the possibility of extension.
Allowing mixed situations might indeed result in undefined situations; exploring such cases is work in progress.

Q: In my lab's thinking about this (reference in the old version of proposal), we've wanted/needed to include info like this about each study, at the study level - including references, key authors to contact, e-mail address, and sharing permissions levels.
I wonder if we could consider that study-level info. -Tor Wager
A: Partly covered by BIDS-raw's study-level or derivative-level dataset_description.json

- Authors: authors of the dataset (not paper)
- Acknowledgements
- HowToAcknowledge
- Funding
- Ethics Approval
- ReferencesAndLinks
- DatasetDOI
- License

Plus: a README and a LICENSE can be added to any dataset (i.e. next to any dataset_description.json)
Currently not covered by BIDS: 'ContactPerson', i.e. whom to contact with questions related to the dataset (somewhat similar to the corresponding author of a paper). **Should this be added to the dataset_description.json?**

Q: Will we propose standards for how missing values are coded, e.g. in studies.tsv? -Tor Wager
A: Proposal: 'n/a' as im BIDS-raw

Q: What are some possible extreme cases worth thinking about, when challenging the current version of the proposal.
A:

1. Having mixed compliant and non-compliant derivatives
2. mega-analysis dataset with 2, 1 or 0? studies
3. play around with inheritance principle: any unexpected behavior?
4. ...

Q: Situation: for study-X we have access only to derivative data, in a non-compliant format. The folder contains one image per subject and a table (e.g. excel) with participant-level information. How to make participant-level information accessible in BIDS-MEGA in this case?
A: this is a case where we can sacrifice non-invasive data consolidation, it is most probably not badly needed for such a dataset, anyway. So the solution might be to add participants.tsv manually, as an extra file in the folder (and retain all other files). Subject-to-file mapping can than be solved with a bids_mapper.json

Q: BIDS-MEGA in a mega-analysis where the outcome-of-interest is a between-subject factor in some studies and a within-subject factor in others. Example: 'drug' is between-subject factor in study-1 (with groups 'control' and 'drug' in participans.tsv) and within-subject factor in study-2 (with BIDS sessions 'control' and 'drug').
A: We create a mega-entity for drug and assign it to the corresponding participants.tsv column in study-1 and the BIDS entity 'sess' in study-2, using the 'scope' filed.

Q: Is 'full standardization' to BIDS really insufficient?

(based on: https://github.com/bids-standard/bids-specification/issues/880)
A: In many cases that may be the way to go. BIDS-MEGA does not prevent the user from doing so. What's more, a BIDS-MEGA style non-invasive data consolidation can serve as an input for automatic scripts that do the full, 'invasive' consolidation. However, 'full standardization' into BIDS is in some cases not possible, simply because BIDS does not cover a specific case at all. With BIDS-MEGA, such datasets can also be seamlessly integrated into a mega-analysis.
See more here (related to the previous version of his BEP):
https://github.com/bids-standard/bids-specification/issues/880

Q: BIDS-MEGA can be abused, so that 'mapped' noncompliant folders subvert their "normal" BIDS-compliant siblings.
See more here (related to the previous version of his BEP):
https://github.com/bids-standard/bids-specification/issues/880
A: This is less of an issue on the 'end-user' side and requires more attention on the side of the developers of related software tools (that produce the non-compliant derivatives), whom we expect to understand the priority of producing fully BIDS-compliant output.

List of proposed changes to the specification

- Introduction:
  - briefly mention mega-analysis related functionality (Module C)
- Common Principles:
  - mention mega-entities here (C)
  - mention the bids-mapper here (B).
  - outline the possibility for the top-level mega-BIDS directory here (module A).
- Modality agnostic files:
  - extend the specification of dataset_descripion.json according to module C.
  - Add the specification of bids_mapper.json, according to B.
- Modality specific files: add reference to the HED-options of the BIDS-mapper (B).
- Derivatives: add a new section to introduce how the BIDS-mapper can be used with non-compliant derivatives (B).
- Longitudinal and multi-site studies: add new section to explain how mega-analysis are handled with modules A, B and C.
- Appendix: add references here.

4. Change-log

V0.1.0 2022-03-08:

Proposal completely revised to incorporate comments from the contributors and the BIDS maintainer team and converted to a modular format.
The old version v0.0.4 is available here.

V0.0.4 2021-09-21:
new “pending issue”

V0.0.3 2021-07-15:
incorporated remarks from Christian Büchel
fixed some typos

v0.0.2 2021-04-22:
more precise specification of columns in studies.tsv
added the section ‘Pending Issues’.

v0.0.1 2021-04-22:
Initial work on specification