


Maintaining large datasets at scale

Christopher J. Markiewicz
Stanford University





(L)

NEURO

SEARCHSUPPORTDOCUMENTATIONSign in

A free and open platform for validating and sharing BIDS-compliant **MRI** (</search/modality/mri>), **PET** (</search/modality/pet>), **MEG** (</search/modality/meg>), **EEG** (</search/modality/eeg>), and **iEEG** (</search/modality/ieeg>) data

64,393 Participants

1,508 Public Datasets

Browse by Modalities

Or

Search

MRI

PET

MEG

EEG


iEEG

NIRS

The BRAIN Initiative

- ~130TB of public data (180TB total)
- ~1300 private datasets

https://effigies.github.io/distribits-2025

OpenNEURO



MRI

A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks

[\(/search/modality/mri\)](/search/modality/mri)

Follow6

Bookmark11

- Files
[\(/datasets/ds005256/versions/1.1.0\)](/datasets/ds005256/versions/1.1.0)
- Download
[\(/datasets/ds005256/versions/1.1.0/download\)](/datasets/ds005256/versions/1.1.0/download)
- Metadata
[\(/datasets/ds005256/versions/1.1.0/metadata\)](/datasets/ds005256/versions/1.1.0/metadata)

BIDS Validation

1888 WARNINGS

Valid

Analyze

Clone

README

Overview

A multimodal fMRI dataset unifying naturalistic processes with a rich array of experimental tasks

Abstract

Cognitive neuroscience has advanced significantly due to the availability of openly shared datasets. Large sample sizes, large amounts of data per person, and diversity in tasks and data types are all desirable, but are difficult to achieve in a single dataset. Here, we present an open dataset with N = 101 participants and 6 hours of scanning per participant, with 6 multifaceted cognitive tasks including 2 hours of naturalistic movie viewing. This dataset’s combination of ample sample size, extensive data per participant, more than 600 iso hours worth of data, and a wide range of experimental conditions — including cognitive, affective, social, and somatic/interoceptive tasks — positions it uniquely for probing important questions in cognitive neuroscience.

Canonical location

Clones of this dataset could be found at different locations. But the original location is at the OpenNeuro archive: <https://openneuro.org/datasets/ds005256> (<https://openneuro.org/datasets/ds005256>) .

OpenNeuro Accession Number
ds005256

Authors

Heejung Jung, Maryam Amini, Bethany J. Hung, Eilis I. Murphy, Patrick Sadil, Yaroslav O. Halchenko, Bogdan Petre, Zizhuang Miao, Philip A. Kragel, Xiaochun Han, Mickela O. Heilicher, Michael Sun, Owen G. Collins, Martin A. Lindquist, Tor D. Wager

Available Modalities

MRI [\(/search/modality/mri\)](/search/modality/mri)

Versions

1.1.0

Created: 2025-03-13

Versions

Tasks

alignvideo, faces, narratives, shortvideo, fractional, social

Uploaded by

By clicking "I Agree", I affirm that I have the appropriate institutional permissions to receive de-identified data for secondary data analysis, and that neither I nor my collaborators will attempt to reidentify individuals whose data are contained in downloads from OpenNeuro. Further, if for any reason the identity of participants contained in downloads from OpenNeuro become known to me I will make no effort to recontact such participants and will provide immediate notice to OpenNeuro staff.

I Agree

OpenNeuro goals

High availability

- Files are published to S3 (`exporttree=yes` `versioning=yes`)
- Datasets are published to GitHub
 - Versions are lightweight tags

No need to hit OpenNeuro servers to retrieve data

Data integrity

- Routine `git annex fsck` on server and S3 remotes
- Cold storage backups
- Public mirrors of git repositories

Detecting problems

With >2500 datasets, supplied by users, managed by an evolving code base, problems arise. How to detect them?

Sync failures with GitHub are often a good indicator:

```

1  ▫ scripts/check-github-sync
2  2025-10-03 19:00:49 [error    ] Missing latest tag          dataset=ds004021 tag=1.0.1
3  2025-10-03 19:00:50 [error    ] Missing latest tag          dataset=ds004212 tag=3.0.0
4  2025-10-03 19:00:51 [error    ] GraphQL query error
5  2025-10-03 19:00:51 [warning  ] mismatch: b7b87f8(2.0.3) != c995770 dataset=ds004516 tag=2.0.3
6  2025-10-03 19:00:51 [error    ] Missing latest tag          dataset=ds004639 tag=1.0.0
7  2025-10-03 19:00:52 [warning  ] mismatch: 2b53548(1.0.2) != 8222107 dataset=ds004837 tag=1.0.2
8  2025-10-03 19:00:52 [error    ] GraphQL query error
9  2025-10-03 19:00:54 [error    ] Missing latest tag          dataset=ds005293 tag=1.0.0
10 2025-10-03 19:00:54 [error    ] Missing latest tag          dataset=ds005381 tag=1.1.0
11 2025-10-03 19:00:56 [warning  ] mismatch: 554e144(1.0.0) != 8f0cfb1 dataset=ds006111 tag=1.0.0
12 2025-10-03 19:00:56 [error    ] Missing latest tag          dataset=ds006319 tag=1.0.0
13 Fetching ds006801 _____ 1496/1502
14 Checking ds006468 _____ 1496/1502

```

Typically a missing tag means an incomplete export or large files outside the annex.

<https://github.com/OpenNeuroOrg/openneuro/blob/master/scripts/check-github-sync>

Wraps: `git ls-remote $REPO $TAG`

Problems and solutions

- Problem: File corruption in transit cannot be detected server-side
 - Solution: Upload client hashes files, creates a commit, and pushes
 - <https://github.com/OpenNeuroOrg/openneuro/blob/master/cli/src/worker/git.ts>
- Problem: Files can become un-annexed (typically race condition)
 - Solution: Replay commits since last publishable state, fixing as you go
 - <https://github.com/OpenNeuroOrg/openneuro/blob/master/scripts/reannex-to-tag.sh>
 - `git rebase -X theirs --exec $EXEC_SCRIPT $REF`
 - `git diff-tree [...] | git rm --cached --paths-from-file=- && git annex add && git commit`
- Problem: Older datasets have unversioned S3 remotes
 - Solution: Clear out S3 prefix, recreate the remote, and re-export
- Problem: Exports can be interrupted
 - Solution: Re-export tags and git push

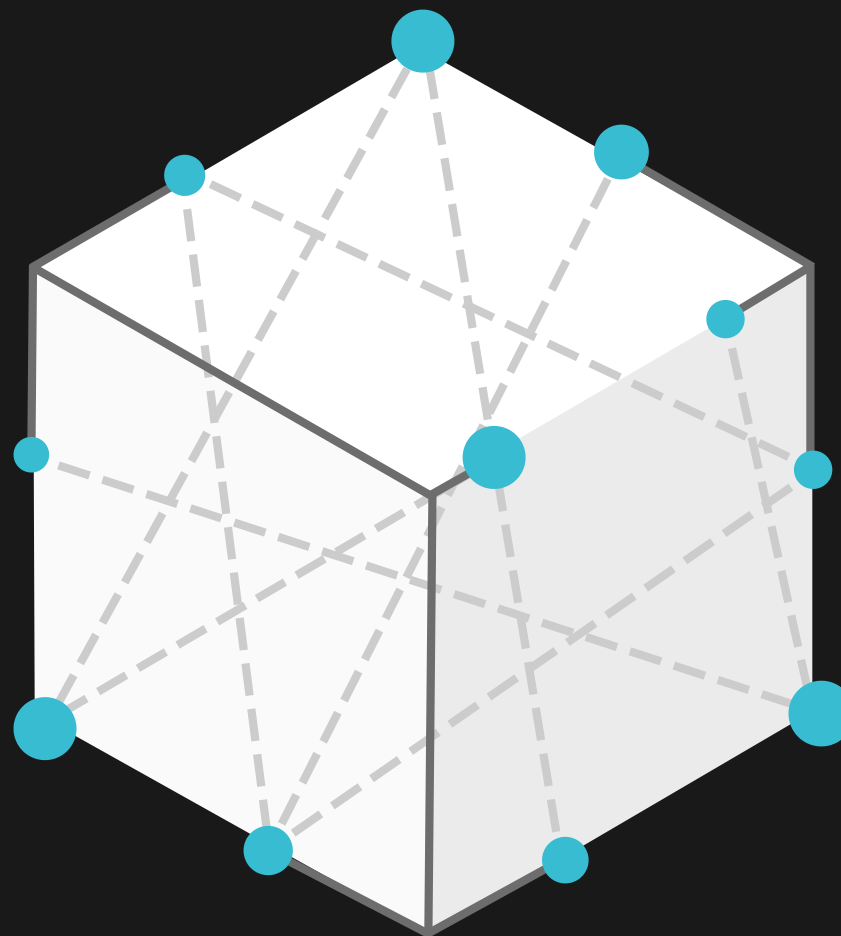
Quick, copy-on-write backups

Before potentially destructive operations, we make full backups. Using `git clone` and `rsync --link-dest`, this takes seconds.

```
1 % DS=ds005256
2 % git clone $DS{,.bak}
3 Cloning into 'ds005256.bak'...
4 done.
5 Updating files: 100% (28408/28408), done.
6 % rsync -a --ignore-existing --link-dest=../$DS $DS{,.bak}/.
7 % rsync -a --delete $DS{,.bak}/.
```

Less than 20s and 1GB needed to back up a 2TB dataset:

```
1 % du -sh $DS.bak
2 2.1T    ds005256.bak
3 % du -sch $DS*
4 2.1T    ds005256
5 570M    ds005256.bak
6 2.1T    total
7 % diff -r $DS*
```



OpenNEURO