Toward an Open Science Ecosystem in Neuroimaging

Russ Poldrack
Stanford University

https://poldrack.github.io/talks-OpenScienceEcosystem/
Transparency is essential for reproducibility

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>Same</td>
</tr>
<tr>
<td>Different</td>
<td>Different</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
“we can distill Claerbout’s insight into a slogan:
An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of...
Why neuroimaging is a best-case scenario for open science

- Magnetic resonance imaging (MRI) is the primary tool for studying human brain structure and function
- MRI data are digital end-to-end
  - From MRI scanner to automated analysis
  - Usually zero/few manual analysis steps
A false start for fMRI data sharing
A false start for fMRI data sharing

This letter comes from a group of scientists who are publishing papers using fMRI to understand the links between brain and behavior. We are writing in reaction the recent announcement of the creation of the National fMRI Data Center (www.fmridc.org). In the letter announcing the creation of the center, it was also implied that leading journals in our field may require authors of all fMRI related papers accepted for publication to submit all experimental data pertaining to their paper to the Data Center. ...

We are particularly concerned with any journal's...
2010: The year data sharing broke in neuroimaging

Toward discovery science of human brain function

Data sharing is becoming the norm in neuroimaging

Anonymous senior researcher:
“OHBM has been taken over by the open science zealots!”

Milham et al., *Nature Communications*, 2018
An open ecosystem for retrospective data sharing

- Neurosynth.org: Open database of published neuroimaging coordinates
- Neurovault.org: Open archive for neuroimaging results
- OpenNeuro.org: Open...
Maximally open sharing

- Data shared under maximally permissive data use agreements:
  - Neurosynth: Open Data Commons Open Database License v1.0
  - Neurovault: CC0
  - OpenNeuro: CC0
- All data available programmatically via web API

- CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as
Neurosynth: Sharing activation coordinates

- Brain activity is reported in a (somewhat) standardized coordinate system

Table 1
Regions that showed a condition × time interaction in the ANOVA analysis

<table>
<thead>
<tr>
<th>No.</th>
<th>Region</th>
<th>Hemisphere</th>
<th>BA</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>mm³</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Middle/superior temporal gyrus</td>
<td>L</td>
<td>21/22/37</td>
<td>-52</td>
<td>-54</td>
<td>9</td>
<td>13257</td>
</tr>
<tr>
<td>2</td>
<td>Inferior frontal gyrus</td>
<td>L</td>
<td>45/46/37</td>
<td>-49</td>
<td>26</td>
<td>6</td>
<td>2781</td>
</tr>
<tr>
<td>3</td>
<td>Posterior cerebellum</td>
<td>L</td>
<td>9/8</td>
<td>-19</td>
<td>-79</td>
<td>-38</td>
<td>2214</td>
</tr>
<tr>
<td>4</td>
<td>Dorsomedial PFC</td>
<td>L</td>
<td>10</td>
<td>-37</td>
<td>49</td>
<td>15</td>
<td>3051</td>
</tr>
<tr>
<td>5</td>
<td>Left anterior PFC</td>
<td>L</td>
<td>10</td>
<td>-42</td>
<td>-58</td>
<td>47</td>
<td>3132</td>
</tr>
<tr>
<td>6</td>
<td>Inferior parietal cortex</td>
<td>L</td>
<td>6</td>
<td>-43</td>
<td>0</td>
<td>50</td>
<td>1485</td>
</tr>
<tr>
<td>7</td>
<td>Dorsal premotor cortex</td>
<td>L</td>
<td>17</td>
<td>-10</td>
<td>-95</td>
<td>-2</td>
<td>378</td>
</tr>
<tr>
<td>8</td>
<td>Lingual gyrus</td>
<td>L</td>
<td>21/22/37</td>
<td>52</td>
<td>-40</td>
<td>5</td>
<td>16470</td>
</tr>
<tr>
<td>9</td>
<td>Middle/superior temporal gyrus</td>
<td>R</td>
<td>45/46</td>
<td>51</td>
<td>28</td>
<td>6</td>
<td>2241</td>
</tr>
<tr>
<td>10</td>
<td>Inferior frontal gyrus</td>
<td>R</td>
<td>23</td>
<td>-78</td>
<td>-34</td>
<td>2808</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Posterior cerebellum</td>
<td>R</td>
<td>9</td>
<td>53</td>
<td>29</td>
<td>405</td>
<td></td>
</tr>
</tbody>
</table>
Creating meta-analytic maps

- Automated Coordinate Extraction
  - Automatically extracts activation tables from fMRI papers for 17 journals
  - Current database has 14,371 papers (with full text)
  - 84% sensitivity, 97% specificity against manual database (SumsDB)
  - Meta-analytic maps created for each

Yarkoni et al., 2011, *Nature Methods*
working memory

An automated meta-analysis of 1091 studies
Classification

- Working mem.: P = 78%
- Emotion: P = 64%
- Pain: P = 87%

Select highest probability

"Pain"
Decoding brain activity patterns using Neurosynth
Example of Neurosynth usage

Situating the default-mode network along a principal gradient of macroscale cortical organization

Daniel S. Margulies\textsuperscript{a,1}, Satrajit S. Ghosh\textsuperscript{b,c}, Alexandros Goulas\textsuperscript{d}, Marcel Falkiewicz\textsuperscript{a}, Julia M. Huntenburg\textsuperscript{a,e}, Georg Langs\textsuperscript{f,g}, Gleb Bezgin\textsuperscript{h}, Simon B. Eickhoff\textsuperscript{i,j}, F. Xavier Castellanos\textsuperscript{k,l}, Michael Petrides\textsuperscript{m}, Elizabeth Jefferies\textsuperscript{n,o}, and Jonathan Smallwood\textsuperscript{n,o}

- Identified gradients of functional organization across the cortex
- Used Neurosynth to identify the most common terms associated with each gradient
Neurovault: Sharing neuroimaging results

- The results of most neuroimaging studies are images with statistical estimates at each voxel
- Neurovault.org is an open archive for these results

Gorgolewski et al., 2015, *Frontiers in Neuroinformatics*
• Collections
  - A set of images (such as all images from a particular paper) can be uploaded as a collection
  - Each collection receives a persistent identifier

Preprocessed Consortium for Neuropsychiatric Phenomics dataset

Related article: http://doi.org/10.12688/f1000research.11964.2

Source data:

Citation guidelines

If you use the data from this collection please include the following persistent identifier in the text of your manuscript:


This will help to track the use of this data in the literature. In addition, consider also citing the paper related to this collection.
• **Image browser**
  
  - Individual images can be browsed and downloaded
  
  - A number of analysis tools can also be applied
  
  - Each image also receives a persistent identifier
Example of Neurovault usage

Comprehensive decoding mental processes from Web repositories of functional brain images

Romuald Menuet\textsuperscript{5,6}, Raphael Meudec\textsuperscript{1,2,3,6}, Jérôme Dockès\textsuperscript{4}, Gael Varoquaux\textsuperscript{1,2,3} & Bertrand Thirion\textsuperscript{1,2,3,6}

Scientific Reports | (2022) 12:7050
OpenNeuro: Sharing raw and processed neuroimaging data

A free and open platform for validating and sharing BIDS-compliant MRI, PET, MEG, EEG, and iEEG data

29,064 Participants 761 Public Datasets

Browse by Modalities

Or

Search

Validation Using BIDS

The Brain Imaging Data Structure (BIDS) is an emerging standard for the organization of neuroimaging data.

Want to contribute to BIDS? Visit the Google discussion group to contribute.

OpenNeuro Runs on DataLad

Want to access OpenNeuro datasets with DataLad? Visit the dataset collection on Github.
Simply sharing data is not sufficient. It must be shared in a way that makes it useful!
It’s easy to share data badly

Data Sharing and Management Snafu in 3 Short Acts

- I received the data, but when I opened it up it was in hexadecimal
- Yes, that is right
- I cannot read hexadecimal
- You asked for my data and I gave it to you. I have done what you asked.
Brain Imaging Data Structure (BIDS)

- A community-based open standard for neuroimaging data
  - A file organization standard
  - A metadata standard
The development of BIDS

- January 2015
  - Initial stakeholder meeting at Stanford (funded by INCF)
  - Initiated development of a draft standard
- September 2015
  - Draft standard posted to BIDS web site with 22 example datasets
  - Solicited feedback from community
- June 2016
BIDS Principles

- Adoption is crucial
  - Keep it as similar to existing practices as possible
    - Don’t let technology override usability!
  - Focus on engaging the community
- Don’t reinvent the wheel
  - Use existing standards when possible
- 80/20 rule
The importance of automated validation

<table>
<thead>
<tr>
<th>Summary</th>
<th>Available Tasks</th>
<th>Available Modalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 Files, 18.42kB</td>
<td>rhyme judgment</td>
<td>bold</td>
</tr>
<tr>
<td>13 - Subjects</td>
<td></td>
<td>T1w</td>
</tr>
<tr>
<td>1 - Session</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Your dataset is not a valid BIDS dataset.
BIDS Extensions

- BIDS was originally focused on structural/functional MRI data
- BIDS extension process allows extension of the standard through BIDS Extension Proposals (BEPS) initiated by the community
  - Patterned after the Python Enhancement Proposal (PEP) process

11 Completed BEPs:

<table>
<thead>
<tr>
<th>BEP #</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEP001</td>
<td>Quantitative MRI (qMRI)</td>
</tr>
<tr>
<td>BEP003</td>
<td>Common Derivatives</td>
</tr>
<tr>
<td>BEP005</td>
<td>Arterial Spin Labeling (ASL)</td>
</tr>
<tr>
<td>BEP006</td>
<td>Electroencephalography (EEG)</td>
</tr>
<tr>
<td>BEP007</td>
<td>Hierarchical Event Descriptor (HED) Tags</td>
</tr>
<tr>
<td>BEP008</td>
<td>Magnetoencephalography (MEG)</td>
</tr>
</tbody>
</table>
The growing usage of BIDS: An example

- MRIQC Web API
  - Crowdsourced database of MR QC metrics
  - QC metrics from ~375K unique BOLD scans and ~280K T1w scans as of June 2022
  - Publicly available: https://mriqc.nimh.nih.gov/

SCIENTIFIC DATA

Crowdsourced MRI quality metrics and expert quality annotations for training of humans and machines

Received: 19 September 2018
Accepted: 12 March 2019

Oscar Esteban, Ross W. Blair, Dylan M. Nielsen, Jan C. Varada, Sean Marrett, Adam G. Thomas, Russell A. Poldrack & Krzysztof J. Gorgolewski

https://poldrack.github.io/talks-OpenScienceEcosystem/
BIDS enables a growing open-source software ecosystem
BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods

Krzysztof J. Gorgolewski¹*, Fidel Alfaro-Almagro², Tibor Auer³, Pierre Bellec⁴,⁵, Mihai Capota⁶, M. Mallar Chakravarty⁷,⁸, Nathan W. Churchill⁹, Alexander Li Cohen¹⁰, R. Cameron Craddock¹¹,¹², Gabriel A. Devenyi⁷,⁸, Anders Eklund¹³,¹⁴,¹⁵, Oscar Esteban¹, Guillaume Flandin¹⁶, Satrajit S. Ghosh¹⁷,¹⁸, J. Swaroop Guntupalli¹⁹, Mark Jenkinson², Anisha Keshavan²⁰, Gregory Kiar²¹,²², Franziskus Liem²³, Pradeep Reddy Raamana²⁴,²⁵, David Raffelt²⁶, Christopher J. Steele⁷,⁸, Pierre-Olivier Quirion¹⁵, Robert E. Smith²⁶, Stephen C. Strother²⁴,²⁶, Gaël Varoquaux²⁷, Yida Wang⁶, Tal Yarkoni²⁸, Russell A. Poldrack¹

PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi.1005209  March 9, 2017

- Containerized applications that can be run on a BIDS dataset
  - Containers provide ease of use as well as better reproducibility
fMRIprep: Robust preprocessing of fMRI data

**Anatomical preprocessing**

- **Brain atlas**
  - Default: MN152 nonlinear asymmetric v003ac
  - Brain extraction is performed on the reference T1w image

- **Spatial normalization**
  - Non-linear, spatial alignment to the brain atlas

- **Brain tissue segmentation**
  - The brain-extracted image is classified into CSF, GM and WM

- **Surface reconstruction**
  - Surfaces of the cortical sheet are reconstructed from the anatomical information (T1w reference, T2w)

**Functional preprocessing**

- **Alignment to T1w reference**
  - Registers activity in BOLD voxels to anatomical location

- **Spatial normalization**
  - Non-linear, spatial alignment to the brain atlas

- **Brain tissue segmentation**
  - The brain-extracted image is classified into CSF, GM and WM

- **Surface reconstruction**
  - Surfaces of the cortical sheet are reconstructed from the anatomical data

**Preprocessing steps**

- **Fuse & Conform**
  - All T1w images are aligned and averaged to form a 3D reference image
  - NIT headers are checked for validity

- **INU Correction**
  - The T1w reference is run through the N4 algorithm to correct for intensity nonuniformity (INU)

- **Skull-stripping**
  - Atlas-based brain extraction is performed on the reference T1w image

- **BOLD run**
  - One run of one task (or resting-state) time-series of blood-oxygen level (BOLD) measurements

- **Generate reference & brain mask**
  - Time-points showing non-steady state artifacts (e.g. excess of T1 contrast) are aligned and averaged to generate a reference image in native space

- **Estimation of head-motion**
  - Parameters representing bulk head motion (due to involuntary drift, swallowing, etc.) of each timepoint with respect to the reference are estimated

- **Slice-timing correction**
  - (Optional) When the acquisition time of 2D axial slices of a given timepoint is unavailable, temporal dynamics are estimated and all slices resampled to the mid-timepoint of that TR

- **Sample on surface**
  - Sample the BOLD signal on the surfaces reconstructed from the anatomical data

- **Sample in native**
  - Perform a “one-shot” resampling of the BOLD signal in atlas-space, concatenating all pertinent transformations

- **Sample in template**
  - Resample the BOLD signal in atlas-space, concatenating all pertinent transformations

- **Define nuisance modules**
  - Calculate and store nuisance regressors such as noise

- **Calculate regressors and fit**
  - Fit the functional data to the nuisance covariates
MRIQC: MRI quality control for BIDS data

The individual reports show the calculated IQMs and metadata in the summary, and a series of image mosaics and plots designed for the visual assessment of images.

Data points in the scatter plots of the group report can be clicked to open the corresponding individual report. This feature is particularly useful to identify low-quality datasets visually.
Templates and atlases are commonly used in neuroimaging.

There is a significant lack of clarity in the use of these templates.

- There are numerous versions of the widely used “MNI template.”
- Different versions of the MNI template are often used interchangeably.

TemplatesFlow: FAIR Sharing of Neuroimaging Templates

https://poldrack.github.io/talks-OpenScienceEcosystem/
OpenNeuro: A BRAIN Initiative archive for BIDS data

- Supports sharing of any validated BIDS dataset
These filters return 194 datasets:

**The Reading Brain Project L1 Adults**
Uploaded by: Chenshan Gu on 2022-01-07 - 10 months ago  |  Updated: 2022-01-05 - 10 months ago

MODALITY: MRI

TASKS: read task, rest

OPENNEURO ACCESSION NUMBER: ds003974  |  SESSIONS: 1  |  PARTICIPANTS: 52  |  PARTICIPANTS’ AGES: N/A  |  SIZE: 46.67GB  |  FILES: 893

https://poldrack.github.io/talks-OpenScienceEcosystem/
Each shared dataset is versioned and receives a persistent identifier (DOI)
Any valid BIDS dataset can be shared via OpenNeuro
A free and open platform for validating and sharing BIDS-compliant MRI, PET, MEG, EEG, and iEEG data

29,064 Participants 761 Public Datasets
A free and open platform for validating and sharing BIDS-compliant MRI, PET, MEG, EEG, and IEEG data

29,064 Participants  761 Public Datasets
Upload Dataset

Step 1: Select Files  Step 2: Validation  Step 3: Metadata  Step 4: Accept Terms

To protect the privacy of the individuals who have been scanned, we require that all scan data be defaced before publishing a dataset.

Select a BIDS dataset to upload

Select folder
We found 3 warnings in your dataset. You are not required to fix warnings, but doing so will make your dataset more BIDS compliant. Continue or fix the issues and select folder again.

Warning 1:
You should define 'SliceTiming' for this file. If you don't provide this information slice time correction will not be possible. 'Slice Timing' is the time at which each slice was acquired within each volume (frame) of the acquisition. Slice timing is not slice order -- rather, it is a list of times containing the time (in seconds) of each slice acquisition in relation to the beginning of volume acquisition.

Warning 2:
http://poldrack.github.io/talks-OpenScienceEcosystem/
Not all subjects/session/runs have the same scanning parameters.
Upload Dataset

By uploading this dataset to OpenNeuro I agree to the following conditions:

I am the owner of this dataset and have any necessary ethics permissions to share the data publicly. This dataset does not include any identifiable personal health information as defined by the Health Insurance Portability and Accountability Act of 1996 (including names, zip codes, dates of birth, acquisition dates, etc). I agree to destroy any key linking the personal identity of research participants to the subject codes used in the dataset.

I agree that this dataset will become publicly available under a Creative Commons CC0 license after a grace period of 36 months counted from the date of the first snapshot creation for this dataset. You will be able to apply for up to two 6 month extensions to increase the grace period in case the publication of a corresponding paper takes longer than expected. See FAQ for details.

This dataset is not subject to GDPR protections.

Generally, data should only be uploaded to a single data archive in the rare cases where it is necessary to upload the data to two databases (such as the NIMH Data Sharing Policy).
This dataset has not been published! Before it can be published, please create a version.

OpenNeuro Accession Number
ds004338

Authors
Xue, G., Russell A. Poldrack

Available Modalities
MRI

Version

New Changelog

https://poldrack.github.io/talks-OpenScienceEcosystem/
This dataset was obtained from the Openfmri project (http://www.openfmri.org). Accession #: ds003 Description: Rhyme judgment

Release history: 10/06/2011: initial release 3/21/2013: Updated release with QA information 2/18/2016: Updated orientation information in nifti headers for improved left-right determination

This dataset is made available under the Public Domain Dedication and License v1.0, whose full text can be found at http://www.opendatacommons.org/licenses/pddl/1.0/. We hope that all users will follow the ODC Attribution-Share-Alike Community Norms (http://www.opendatacommons.org/norms/odc-by-sa/); in particular, while not legally required, we request that users of the data will acknowledge the Openfmri project and NSF Grant DCI-1131441 (R. Poldrack). All in any publications.
The growth of OpenNeuro

Total number of datasets:
- 300 datasets
- 400 datasets
- 500 datasets
- 600 datasets
- 700 datasets

https://poldrack.github.io/talks-OpenScienceEcosystem/
The diversity of OpenNeuro datasets

Number of tasks

Datatype | #
---|---
mri - anat | 597
mri - func | 521
eeg | 120
mri - dwi | 67
meg | 30
ieeg | 17
beh | 13

Species | #
---|---
Human | 676
Mouse | 20
Rat | 12
NHP | 2
phantoms | 1
Juvenile pigs | 1
Human, Mouse | 1

updated from Markiewicz et al, 2021, *eLife*
Figure 5. Published reuses of OpenNeuro datasets, split by the type of reuse. Note that the final bar includes only reuses identified through June 2021.

Markiewicz et al, 2021, *eLife*
Processing of OpenNeuro data

brainlife.io: processing of MRI data

Data Summary

Subject: sub-001
Session(s): 1, Run(s): 1
Data size: 32 channels, 298k frames
Acceptable channel counts: 100.0% (32 of 32)
Acceptable data points channels: 90.9% (271k of 298k)
Source quality metric based on independent component: 48.4%

Scalp channel log spectra

Sample scalp channel data (mid 2 seconds)
Example of OpenNeuro reuse

- A challenge for decoding brain activity from fMRI data is that most datasets are very small.
- We used OpenNeuro to train a “foundation model”.
  - A pre-trained model that can be used as a starting point for decoding models on smaller datasets.
- We pre-train models on broad fMRI data from OpenNeuro: 11,980 experimental subjects

Challenges to open sharing

- All OpenNeuro MRI datasets must be defaced
  - To reduce risk of reidentification
- There is increasing risk that subjects might be reidentified even after defacing using advanced face recognition systems + face imputation tools (Schwartz et al., 2021)
- If the risk continues to rise, it may become necessary to move away from open sharing
  - This would be a huge loss for researchers, research participants, and the world
Keys to success in neuroimaging data sharing

- **Data are digital end-to-end**
  - Minimizes manual steps in the process
- **Standardized file formats and data standards**
  - Makes data immediately usable by anyone
  - Reduces burden of curation and preparation
- **Demonstrated scientific utility**
- **Numerous success stories**
Lessons learned

- Community buy-in is essential
  - Mandates put in place before the community is ready can backfire
    - Unless they have overwhelmingly powerful advocates, as in genomics
  - Important that sharing advocates are members of community and eat their own dog food

https://poldrack.github.io/talks-OpenScienceEcosystem/
Lessons learned

- Keep it simple and as close to standard practice as possible
  - Overengineered solutions have generally failed
  - If there are more than 2 acronyms...

http://poldrack.github.io/talks-OpenScienceEcosystem/
Lessons learned

- Don’t let the perfect be the enemy of the good
  - 20% of the effort will cover 80% of the datasets - focus on these!
  - There is a long tail of edge cases with loud advocates

https://poldrack.github.io/talks-OpenScienceEcosystem/
Conclusions

- The field of neuroimaging has built an ecosystem for open science and data sharing
- Infrastructure is critical to ease friction
- Community engagement has been key to adoption
- Need to keep the tools as close as possible to current practice
Meta-analytic decoding using Neurosynth

- Given 2+ terms, can determine which is most likely given the data
- Naive Bayes classifier: assumes that all features (voxels) are independent; selects the most probable class
- Can apply this to any activation map—studies, individual subjects, etc.

• Cross-validated classification of all studies in database
• Select 25 high-frequency terms
• Pairwise classification: how well can we distinguish between the presence of each pair of terms?

Yarkoni et al., 2011, *Nature Methods*