

# Shareable Science:

*Standards as a driving force in open,  
reproducible research*

**Elizabeth DuPre**  
McGill QLS612  
13 May, 2020

# Science today





Olah & Carter (2017). Research Debt. *Distill.*

*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 instructions** which generated the figures.*

Buckheit and Donoho  
WaveLab and Reproducible Research, 1995

# An outline for this morning

- Planning with project management
- Sharing science: A role for standards
- Community-driven development

# Community-based project management standards

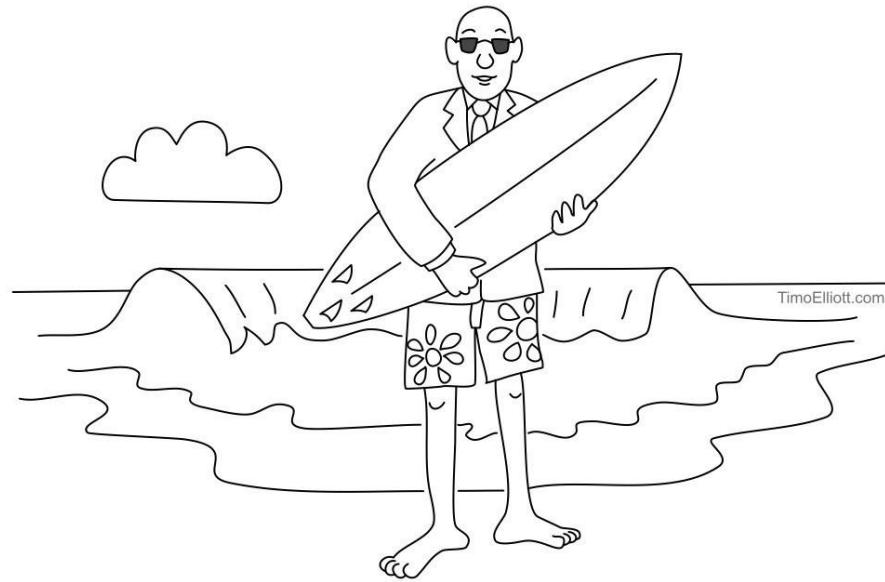
# Community-based project management standards

# Meet Professor Smith



With thanks to [Chris Gorgolewski](#)

# Meet Mike

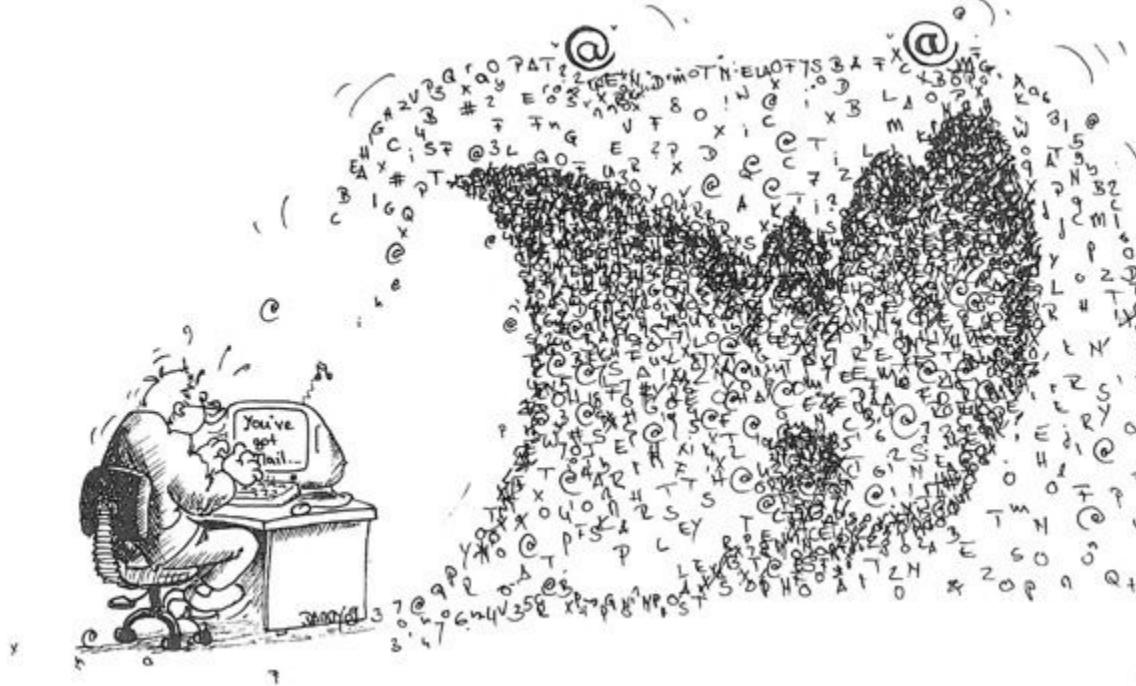


TimoElliott.com

*A Business Analyst that lives in California.*

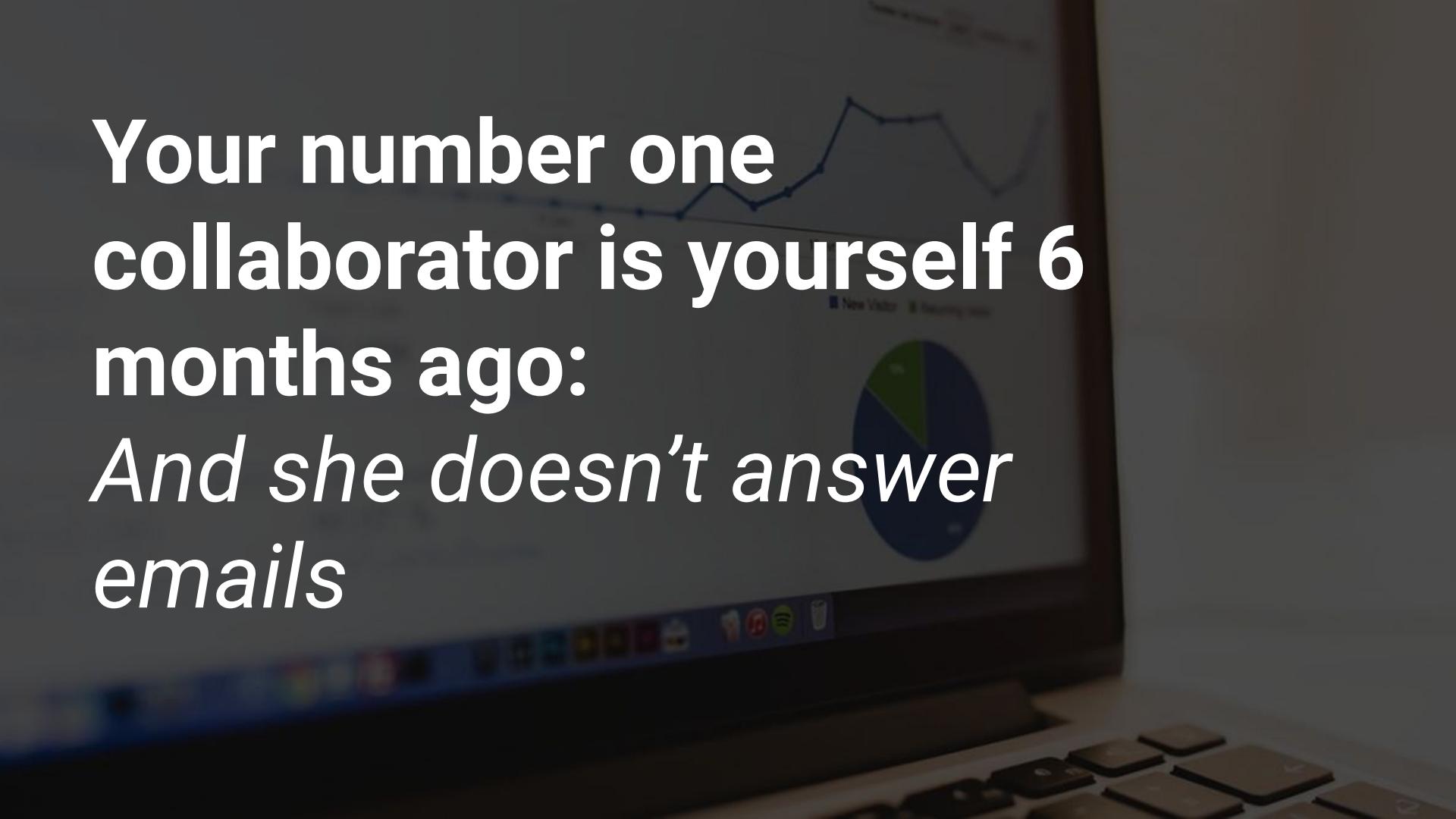
With thanks to [Chris Gorgolewski](#)

# Meet Mike's project folders



With thanks to [Chris Gorgolewski](#)

Your number one  
collaborator is yourself 6  
months ago:  
*And she doesn't answer  
emails*

A laptop screen is visible in the background, showing a line graph with a blue line and some text at the top right. Below the graph is a pie chart divided into two main sections: a large dark blue section and a smaller green section. The laptop's taskbar is visible at the bottom, showing various icons. A dark, semi-transparent rectangular overlay covers the majority of the slide, containing the white text of the main message.

# What is project management?

- A consistent organization for all components of your research project across its lifecycle
- Components can include:
  1. Data
  2. Code
  3. Documentation

A (quick) case study

# Open Science Framework

A scholarly commons to connect the entire research cycle



# The Reproducibility Project

Name	Modified
Replication of Correll (2008, JPSP, Stu...	
- OSF Storage (United States)	
ReplicationReport_Correll2008... 2013-12-11 11:5...	
ReplicationReport_incl_Metho... 2013-12-11 11:5...	
+ Data	
+ Study Materials	
+ Analysis Audit	
+ Independent Direct Replication #...	

<https://osf.io/fejxb/>

# The Reproducibility Project

... still did not have fully  
standardized project organization !

The screenshot shows the OSF Storage interface with two main sections. The top section lists a project titled 'Replication of Correll (2008, JPSP, Stu...)' which is associated with 'OSF Storage (United States)'. The bottom section lists another project titled 'Replication of Bressan & Stranieri (2...)' which is also associated with 'OSF Storage (United States)'. Both projects contain several items: 'BRESSAN COMMENTARY.docx' (modified 2015-06-08 08:5...), 'Study Materials', 'Cleaning & Analysis Scripts', 'Dataset', 'Replication Reports', and 'Analysis Audit'. There are also collapsed sections for 'Dataset', 'Study Materials', and 'Analysis Audit'.

---

<https://osf.io/fejxb/>

<https://osf.io/blcj6/>

# Teaching Integrity in Empirical Research



# Project TIER

## Raw Data

A copy of **every original data file** from which you extract any of data used in your study.

Your original data files **serve as a record** of the data you began the project with.

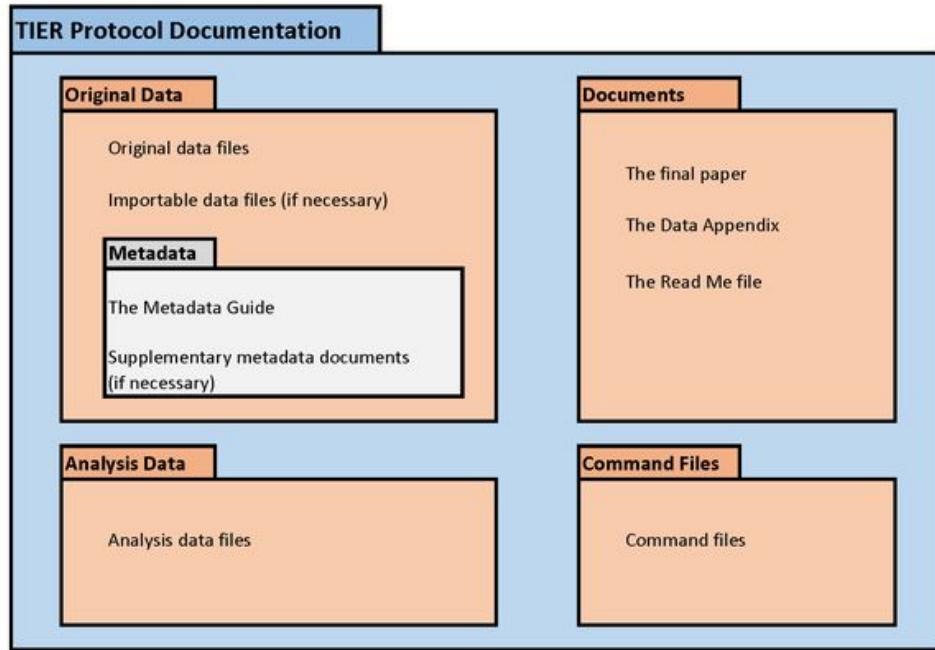
## Analysis Code + Analysis Data

One or more files containing code used for the study... should **execute all the data processing and analysis necessary to replicate the study** and reproduce the reported results

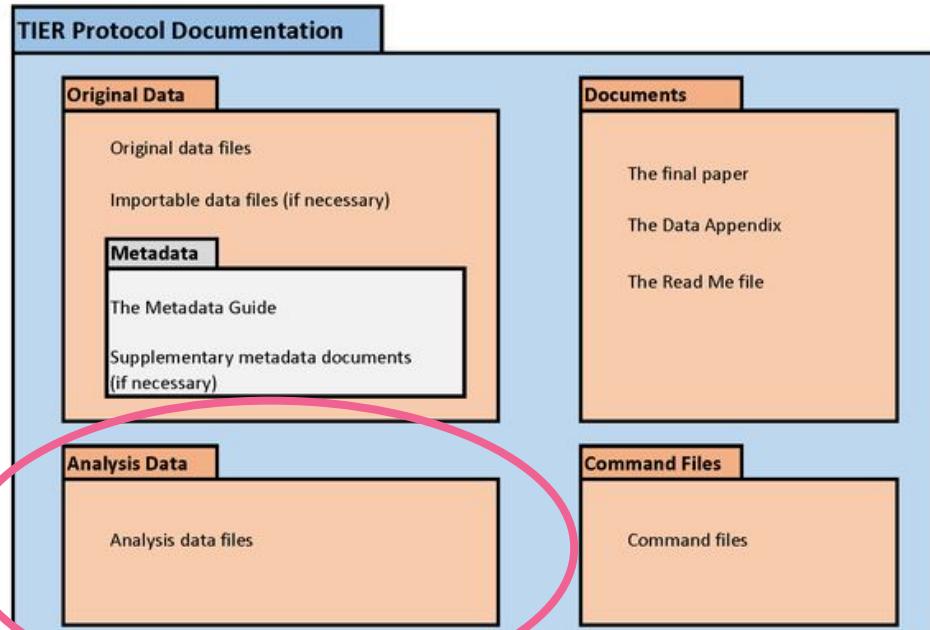
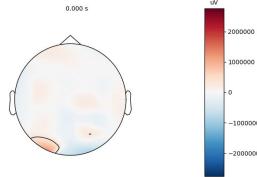
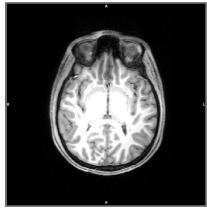
## Documents

**Documentation to understand the study**

- A copy of your final paper
- Your Data Appendix
- Your README file



*The TIER Specification*



*The TIER Specification*



# TIER Protocol 3.0: Template

Public

92

...

Contributors: Norm Medeiros, Richard Ball

Forked from [osf.io/7g6cn](https://osf.io/7g6cn) on 2016-09-16 12:43 PM

Date created: 2016-09-08 11:17 PM | Last Updated: 2018-01-22 11:36 AM

Identifiers: DOI 10.17605/OSF.IO/YBZXE | ARK c7605/osf.io/ybzxe

Category: Project

Description: This project is designed to support Haverford College economics majors who produce empirical theses. The structure is based on the TIER Documentation Protocol. Additional information about Project TIER is available at <http://projecttier.org>

## Wiki



This template was constructed for use by individuals who wish to follow the TIER Protocol for conducting and documenting an empirical research project.

Information about the [TIER Protocol](#), and in particular about [how to use this template](#), can be found on the [Project TIER website](#).

## Citation



## Tags

Project TIER

# Project management is for **everyone**

- Whether working alone or in a team, project management is crucial to sustaining a research project
- But how can we describe our organization to other research groups?
- Our goal is to automatically aggregate information
  - For the purposes of this introduction, I'll focus on collaborating across data sets

# Community-based project management standards

# Reproducibility as a roadmap

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

# Reproducibility as a roadmap

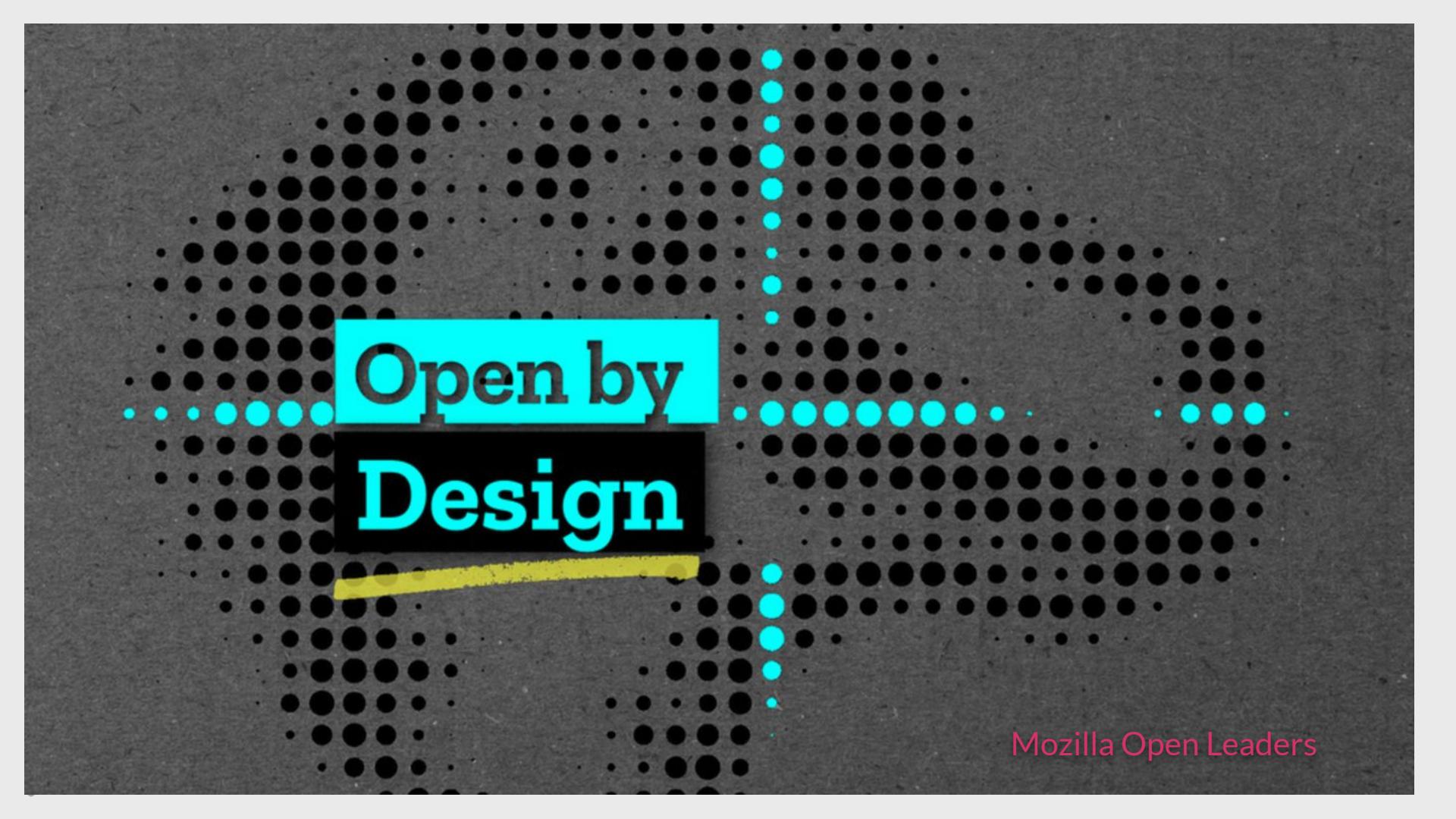
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

# Reproducibility as a roadmap

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

# A quick detour: What if you can't share

- There are barriers preventing us from sharing:
  - Participant consent to share (or lack thereof)
  - Code licensing
  - Trademarked stimuli
  - Time !
- In these cases, why should you care about having your research conform to standards?



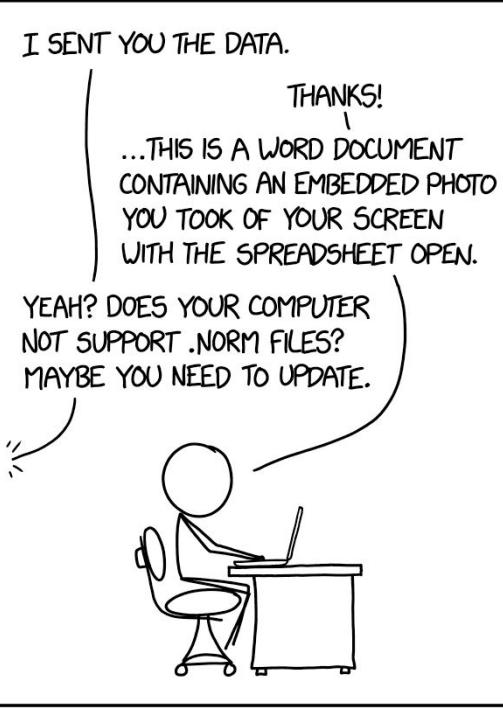
**Open by  
Design**

Mozilla Open Leaders

# “Open by design” in research

- If we design research to be shared openly, we make it easier to share later
  - Use open consent forms, such as [Open Brain Consent](#)
  - Use standard licensing -- it's not just for code ! Check out [Choosealicense.com](#)
  - Understand copyright laws when choosing your stimuli
  - ... Everything is easier if you start at the beginning !
- And you can take advantage of open work throughout your project's life cycle

For more on this, see: [DOI: 10.1186/s13742-015-0072-7](#)



XKCD

# What's in a standard ?

- How can we make it easier for colleagues to work with our data ?
- Human-readable is a low-bar !
- What we need is machine-readable data



OpenElections  
@openelex

City of Detroit produced a lookup tables for its absentee precincts in 2016. It's in Excel. But wait for it: the values are CLIP ART.

[Traduire le Tweet](#)

CITY OF DETROIT		GENERAL ELECTION November 8, 2016			
PRECINCT	PRECINCT	PRECINCT	PRECINCT	PRECINCT	
1	21	40	79	28	118 12
2	4	41	80	29	119 71
3	42	81	82	5	120
4	22	43	83	45	121 45
5	13	44	84	71	122 71
6	7	45	85	6	123 76
7	14	46	86	77	124 77
8	15	47	23	33	125 78
9	14	48	7	29	126 79
10	10	49	87	33	127 80
11	1	50	88	55	128 79
12	15	51	89	161	129
13	16	52	90	55	130 81
14	1	53	91	55	131
		53	92	55	

10:24 AM · 17 avr. 2017 · [TweetDeck](#)

1,8 k Retweets 3,1 k J'aime

# We want to be FAIR



# FAIR: Findable

**Findable:** the first step in (re)using data is to find them!

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

# FAIR: Accessible

**Accessible:** there is a means to access the data, either with authentication and authorization (if necessary), or freely if data is openly available

- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
- A2. Metadata are accessible, even when the data are no longer available

# FAIR: Interoperable

**Interoperable:** data can be integrated with other data, applications, or workflows

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

# FAIR: Re-usable

**Reusable:** data are well-described so that they can be used in different settings

- R1. Meta(data) are richly described with accurate and relevant attributes

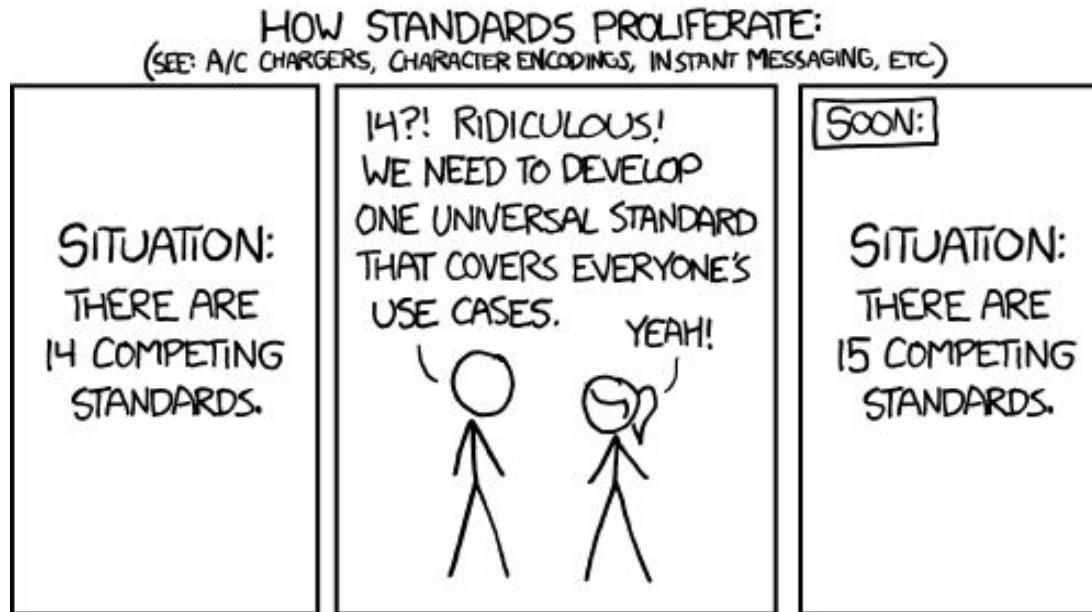
# FAIR is fuzzy

- This is by design ! FAIR is a general set of principles that should guide research across domains
  - What does it mean to have “rich metadata” ? Is that different for neuroimaging vs genomics research ?
- Each discipline needs its own FAIR standards

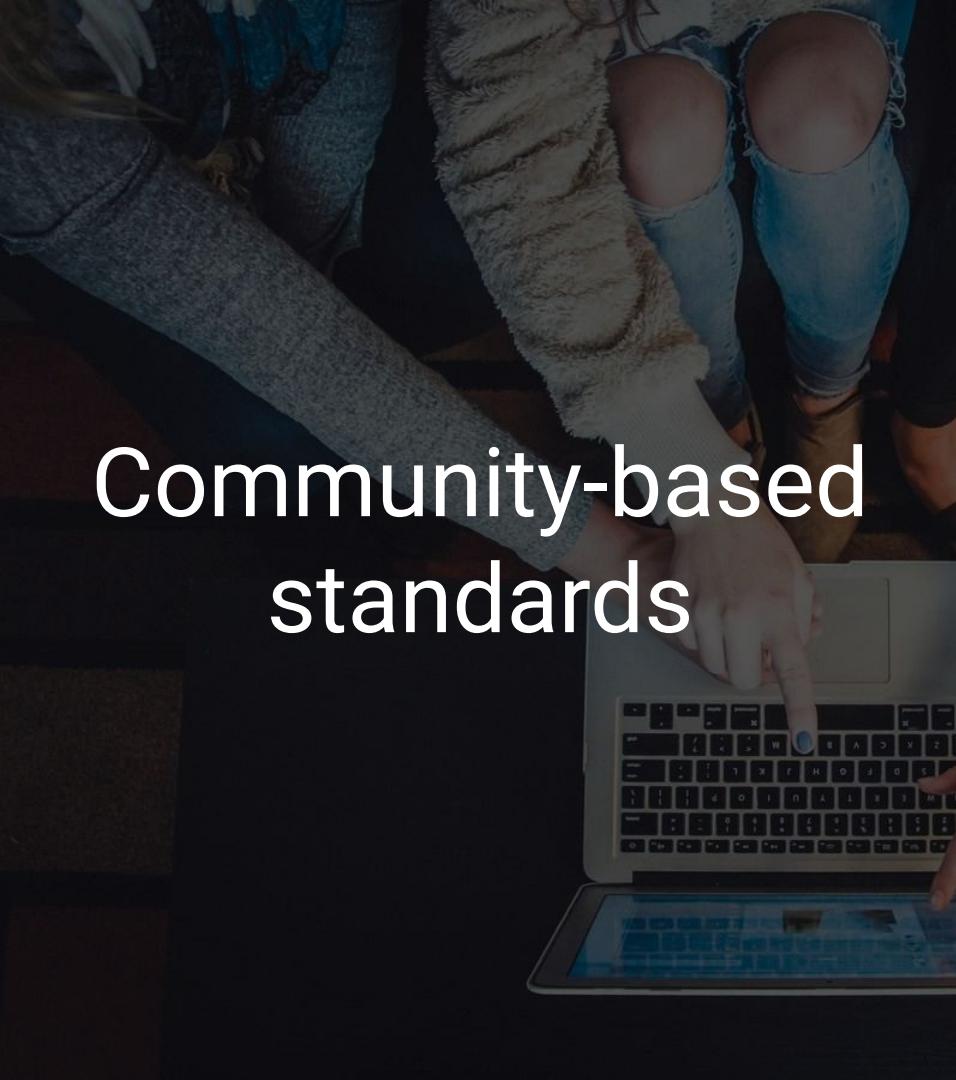


# Community-based project management standards

# Houston, we have a problem



XKCD

A photograph showing a person from the waist down, sitting at a desk. They are wearing a light-colored, fuzzy-textured jacket over a dark top, and blue jeans. Their hands are visible on a silver laptop keyboard. The background is dark.

# Community-based standards

- Developed **openly**
- Strive for **consensus** in decision-making
- Designed to **empower** and equip community members

---

<https://mozilla.github.io/open-leadership-training-series>

# Existing community-based standards for neuroscience data:

*BIDS and NWB*



# Brain Imaging Data Structure



**Brain Imaging Data Structure  
v1.2.0**[The BIDS Specification](#) ▾[The BIDS Starter Kit](#)

# The Brain Imaging Data Structure

This site serves as an online resource to see the current state of the Brain Imaging Data Structure (BIDS) specification. It contains information about the core specification, as well as many modality-specific extensions.

To get started, [check out the introduction](#). If you'd like more information on how to adapt your own datasets to match the BIDS specification, we recommend exploring the [bids-specification starter kit](#).

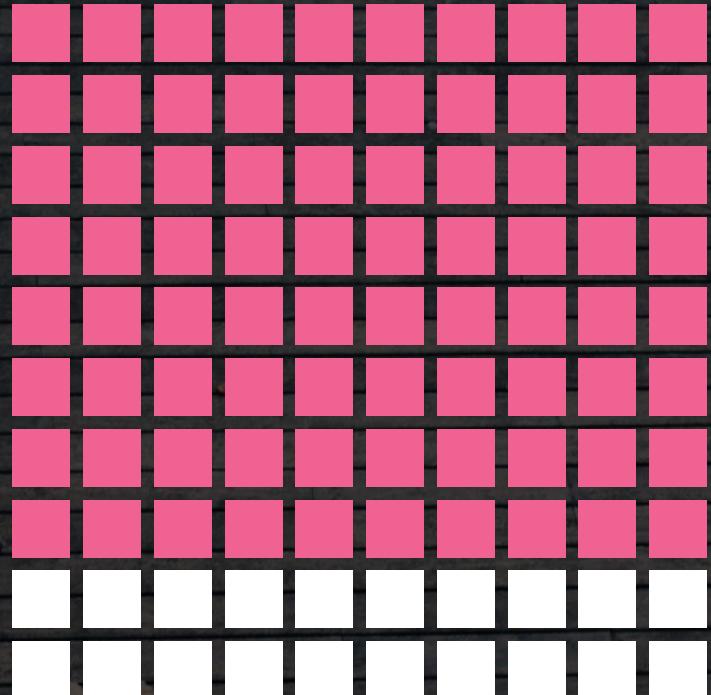


# Example BIDS dataset layout

```
emdupre@ThinkPad-T480s: ~/Desktop/BIDS_dataset$ tree
.
├── CHANGES
├── dataset_description.json
├── participants.tsv
└── README
    └── sub-01
        └── ses-025
            ├── dwi
            │   ├── sub-01_ses-025_run-001_dwi.bval
            │   ├── sub-01_ses-025_run-001_dwi.bvec
            │   ├── sub-01_ses-025_run-001_dwi.json
            │   ├── sub-01_ses-025_run-001_dwi.nii.gz
            │   ├── sub-01_ses-025_run-002_dwi.bval
            │   ├── sub-01_ses-025_run-002_dwi.bvec
            │   ├── sub-01_ses-025_run-002_dwi.json
            │   ├── sub-01_ses-025_run-002_dwi.nii.gz
            ├── fmap
            │   ├── sub-01_ses-025_magnitude1.json
            │   ├── sub-01_ses-025_magnitude1.nii.gz
            │   ├── sub-01_ses-025_magnitude2.json
            │   ├── sub-01_ses-025_magnitude2.nii.gz
            │   ├── sub-01_ses-025_phasediff.json
            │   ├── sub-01_ses-025_phasediff.nii.gz
            ├── func
            │   ├── sub-01_ses-025_task-rest_run-001_bold.json
            │   ├── sub-01_ses-025_task-rest_run-001_bold.nii.gz
            │   ├── sub-01_ses-025_task-rest_run-001_events.tsv
            │   ├── sub-01_ses-025_task-rest_run-001_sbref.json
            │   ├── sub-01_ses-025_task-rest_run-001_sbref.nii.gz
            │   └── sub-01_ses-025_scans.tsv
            └── task-rest_bold.json
5 directories, 25 files
emdupre@ThinkPad-T480s: ~/Desktop/BIDS_dataset$
```

# The 80 / 20 rule

Focus on the 80% use case to enhance clarity and collaboration



[Code](#)[Issues 47](#)[Pull requests 8](#)[Projects 2](#)[Wiki](#)[Insights](#)[Filters](#) is:issue is:open[Labels 24](#)[Milestones 0](#)[New issue](#) 47 Open  46 Closed[Author](#)[Labels](#)[Projects](#)[Milestones](#)[Assignee](#)[Sort](#)

- ! [ENH] Update task and continuous recording metadata for EEG/MEG/iEEG EEG MEG iEEG  
#209 opened 7 days ago by effigies

4

- ! Proposal: BEP PRs should come from branches on the main repository community  
#203 opened 13 days ago by effigies

- ! BEP Template community  
#201 opened 15 days ago by effigies

1

- ! [FIX] Improvements to the "entity table" good first issue help wanted  
#200 opened 18 days ago by choldgraf

- ! [Infra] Add dates to releases in Spec infrastructure  
#199 opened 19 days ago by sappelhoff

- ! [ENH] Proposal for multidimensional array file format  
#197 opened 20 days ago by tyarkoni
- ! MRI: new metadata tag "ElementSpatialSize" (and may be more) to facilitate access to .nii.gz header information  
#196 opened 21 days ago by yarikoptic





You?

We need your expertise !  
Join the [BIDS community](#)  
[on GitHub](#)

# Welcome to the BIDS Starter Kit



## How to get started with the Brain Imaging Data Structure

A community-curated collection of tutorials, wikis, and templates to get you started with creating BIDS compliant datasets.

[BIDS Homepage](#) | [Wiki](#) | [Standard](#) | [Tutorials](#) | [Chat](#) | [Forum](#)

**Click to view the intro video!**



# Checking in on Professor Smith



With thanks to [Chris Gorgolewski](#)



PUBLIC  
DASHBOARD

SUPPORT

FAQ

SIGN IN



# OpenNEURO

A free and open platform for sharing MRI,  
MEG, EEG, iEEG, and ECoG data



Sign in with Google



Sign in with ORCID

Browse 219 Public Datasets

[OpenNeuro.org](https://OpenNeuro.org)

# Recent Activity

## Most Active

 **UCLA Consortium for Neuropsychiatric Phenomics LA5c Study**  
**20,329**

 **The Midnight Scan Club (MSC) dataset**  
**19,376**

 **Visual object recognition**  
**10,341**

 **Forrest Gump**  
**9,449**

 **Classification learning**  
**8,599**

[View More...](#)

## Recently Published

**The Reading Brain Project L1 Adults** **4 DAYS AGO**

**Parallel Adaptation of Symbols, Quantities, and Physical Size** **10 DAYS AGO**

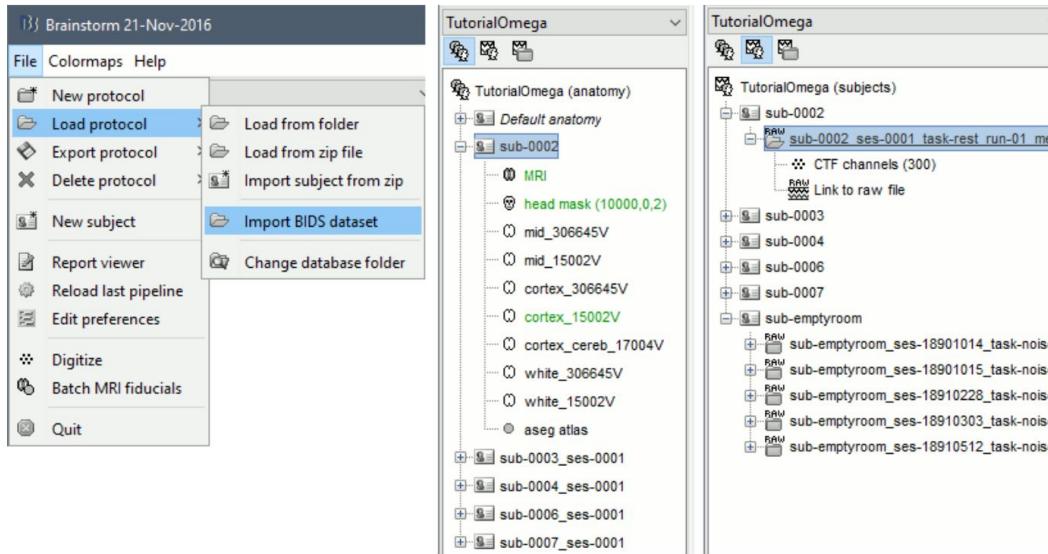
**A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex** **11 DAYS AGO**

**Spinal stimulation stepping and standing dataset** **17 DAYS AGO**

**Handedness and Symbolic Number Representation** **17 DAYS AGO**

Recent activity on [OpenNeuro.org](https://www.openneuro.org)

- Select the menu File > Load protocol > Import BIDS dataset > Select folder **OMEGA\_RestingState\_sample**.
- Keep the default values for all the questions that may be asked during the import process (eg. number of vertices in the cortex surfaces). Once done, you should be able to access the data for the 5 subjects in your database explorer: anatomy, and subject and noise recordings.



[https://neuroimage.usc.edu/brainstorm/Tutorials/RestingOmega#BIDS\\_specifications](https://neuroimage.usc.edu/brainstorm/Tutorials/RestingOmega#BIDS_specifications)



# BIDS Apps

---

portable neuroimaging pipelines that understand BIDS datasets

---

[About](#)

[Tutorials](#)

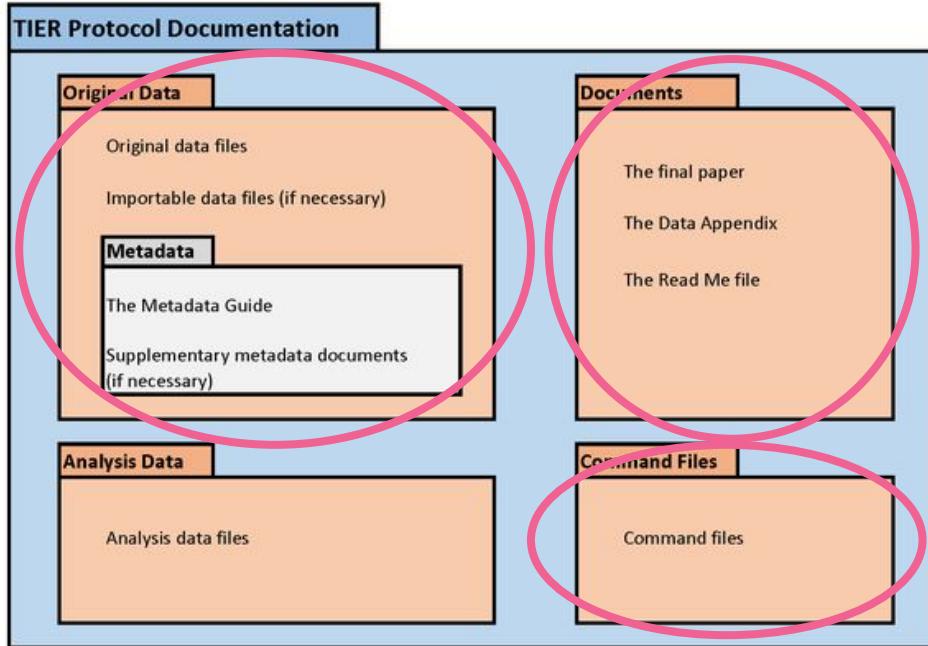
[Apps](#)

# Professor Smith (circa 2030)



With thanks to [Chris Gorgolewski](#)

ReproIn



RMarkdown,  
Jupyter

Cookiecutter

The TIER Specification

# Community-based project management standards



# Take-home ideas

- **Project management is for everyone**
- **Community-driven standards enable new kinds of science**



[@emdupre](https://twitter.com/emdupre)