Shareable Science:

Standards as a driving force in open, reproducible research

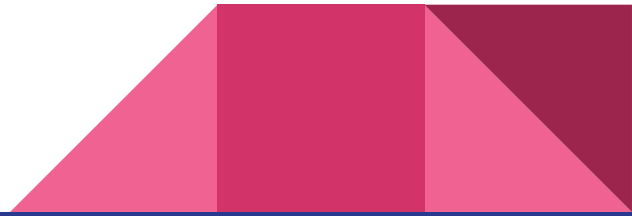
Elizabeth DuPre
McGill QLS612
13 May, 2020

Science today

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

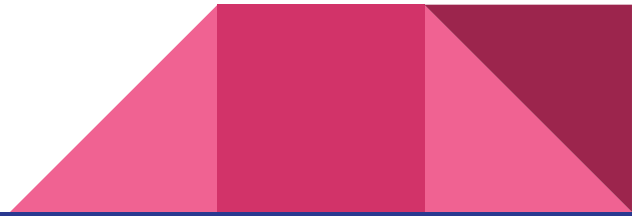
An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

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



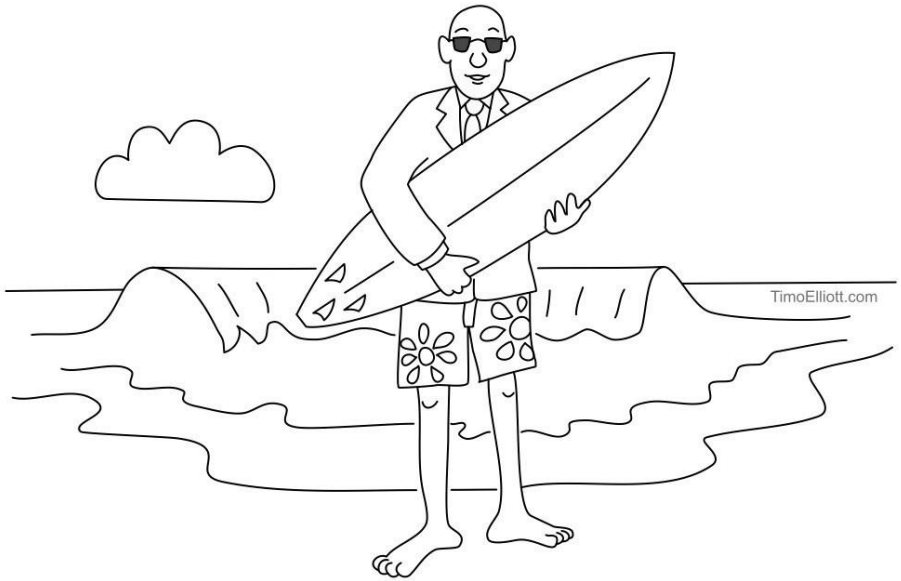
Community-based project management standards

Community-based project management standards

Meet Professor Smith

With thanks to *Chris Gorgolewski*

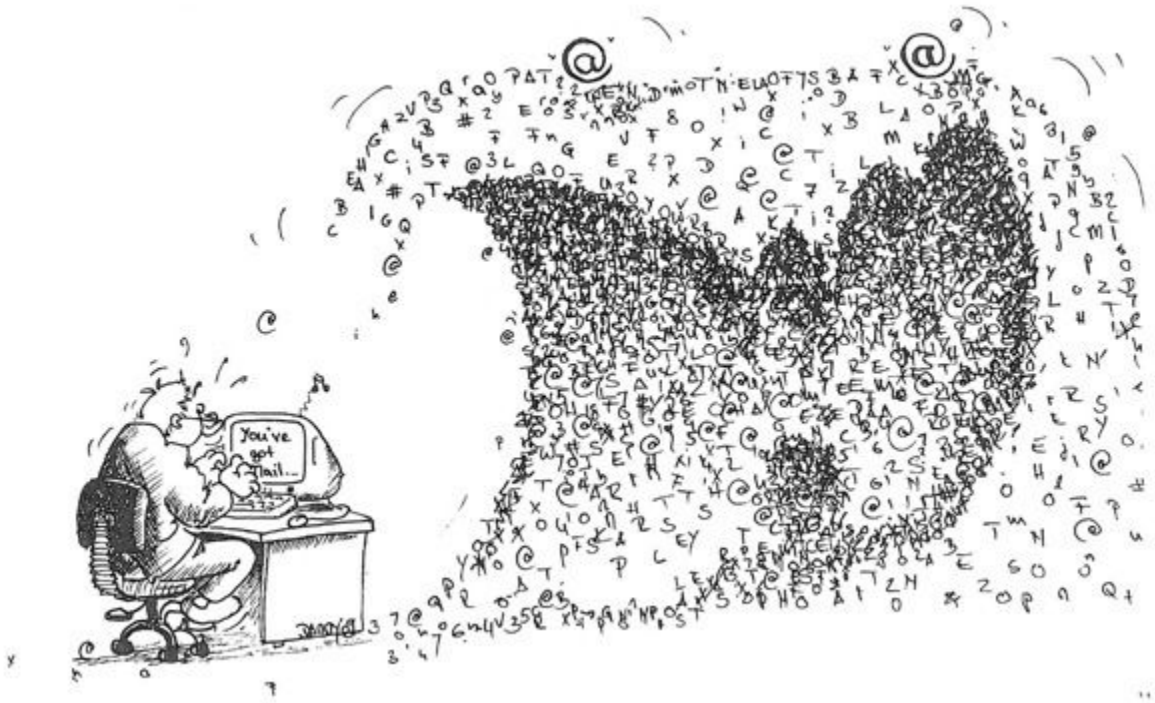
Meet Mike



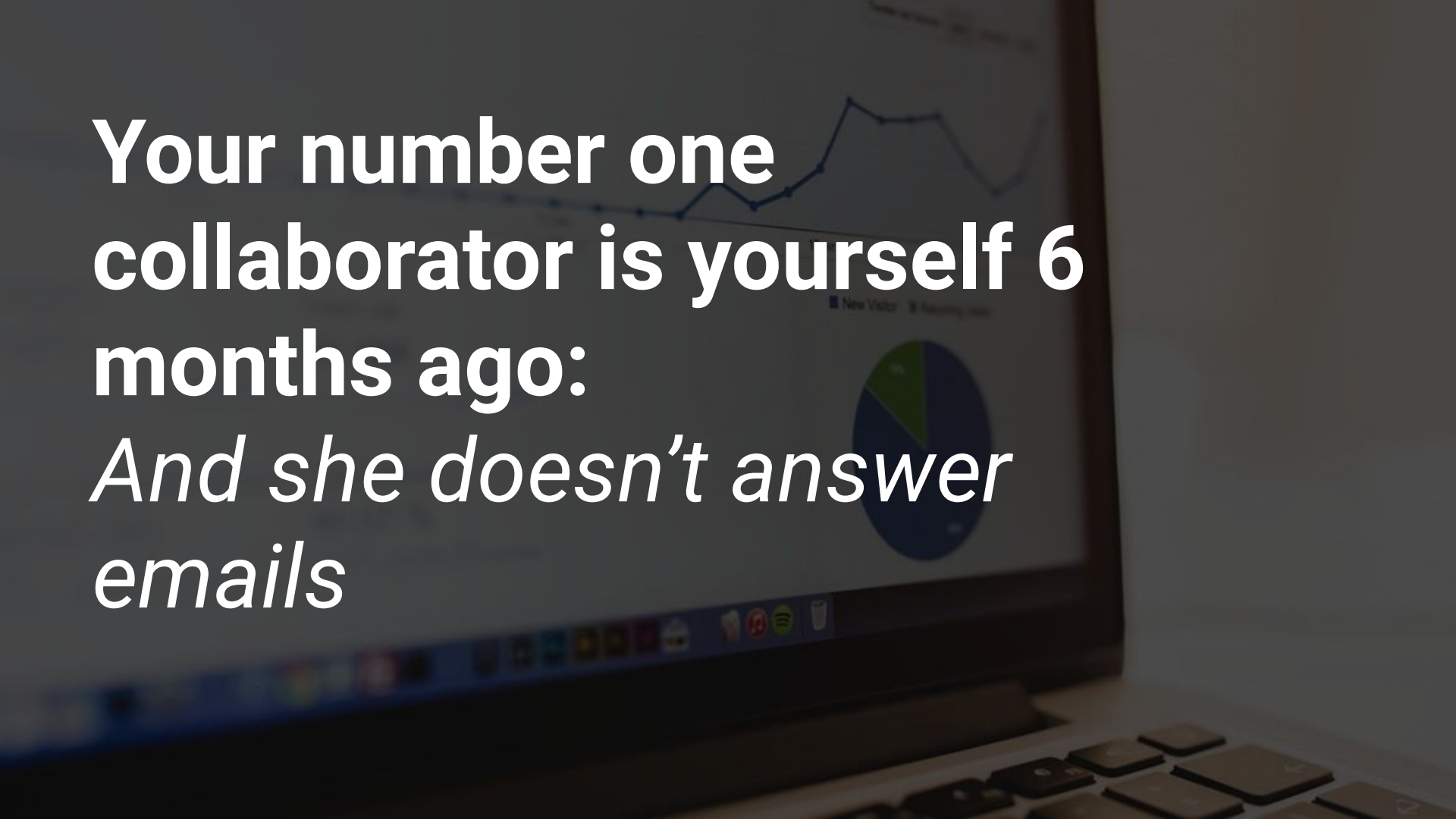
A Business Analyst that lives in California.

With thanks to *Chris Gorgolewski*

Meet Mike’s project folders



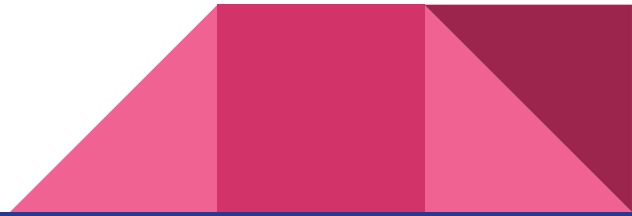
With thanks to *Chris Gorgolewski*



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

What is project management?

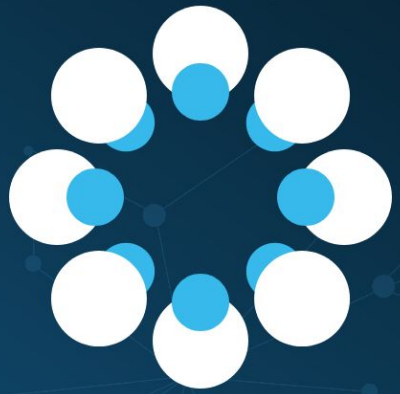
- A consistent organization for all components of your research project across its lifecycle
- Components can include:
    1. Data
    2. Code
    3. Documentation



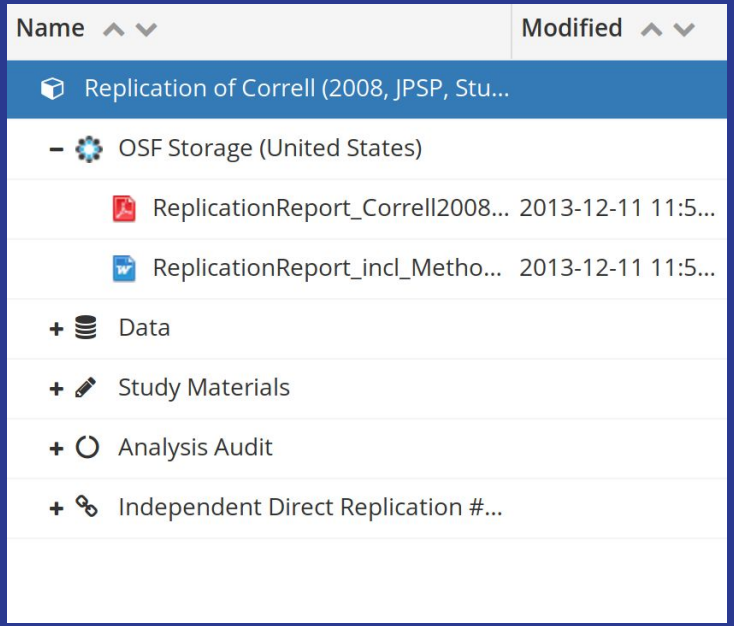
A (quick) case study

Open Science Framework

A scholarly commons to connect the entire research cycle



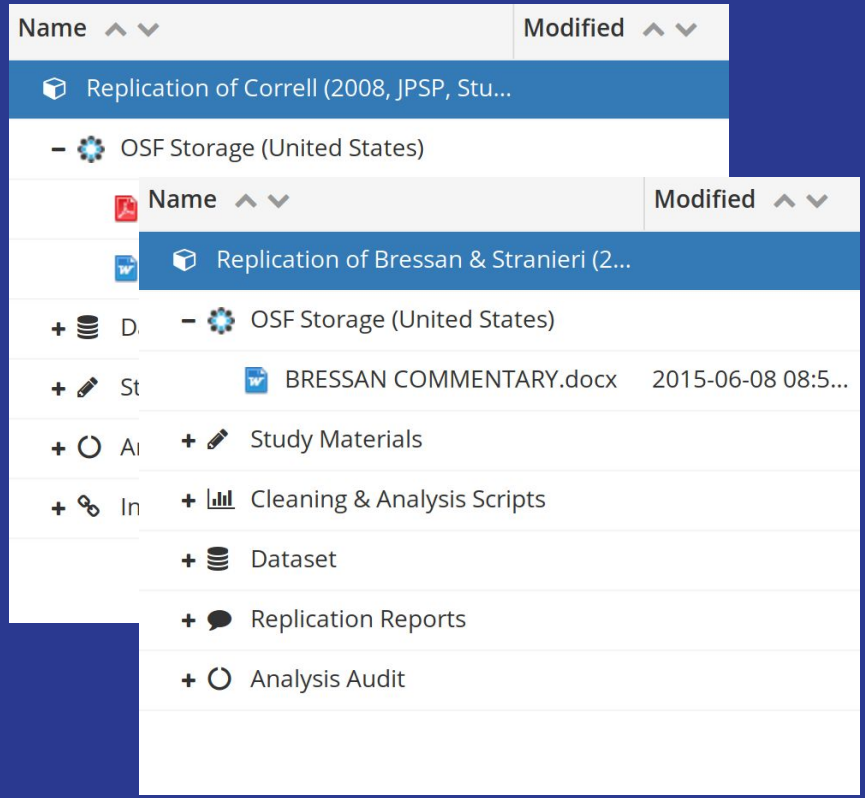
The Reproducibility Project



https://osf.io/fejxb/

The Reproducibility Project

… still did not have fully standardized project organization !



https://osf.io/fejxb/

https://osf.io/blcj6/

Teaching Integrity in Empirical Research

Project TIER

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

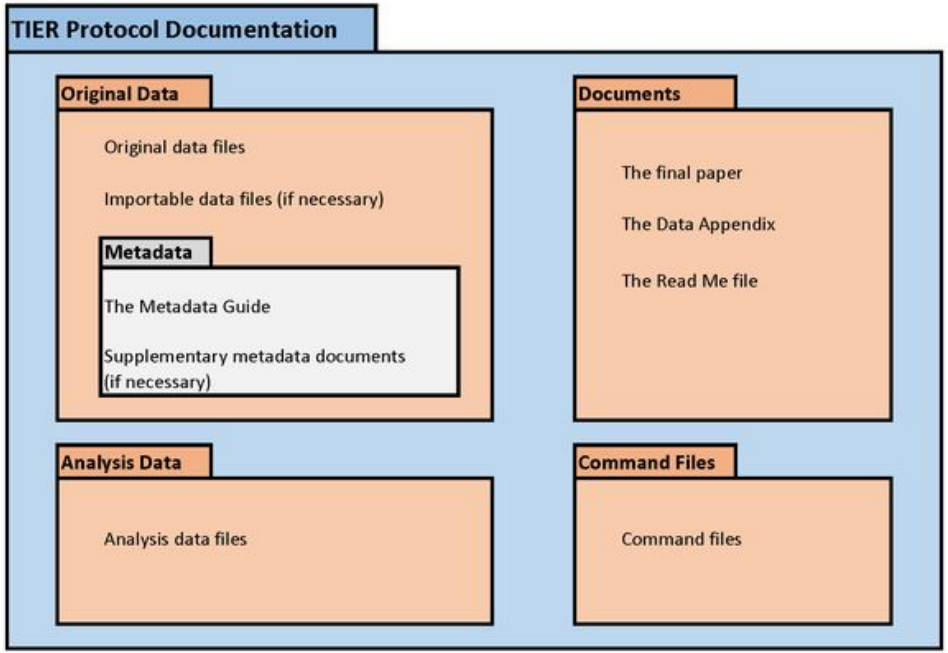
Your original data files **serve as a record** of the data you began the project with.

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

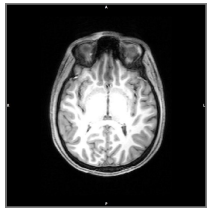
Documentation to understand the study

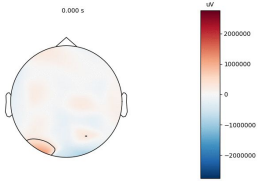
- A copy of your final paper
- Your Data Appendix
- Your README file

https://www.projecttier.org/



The TIER Specification

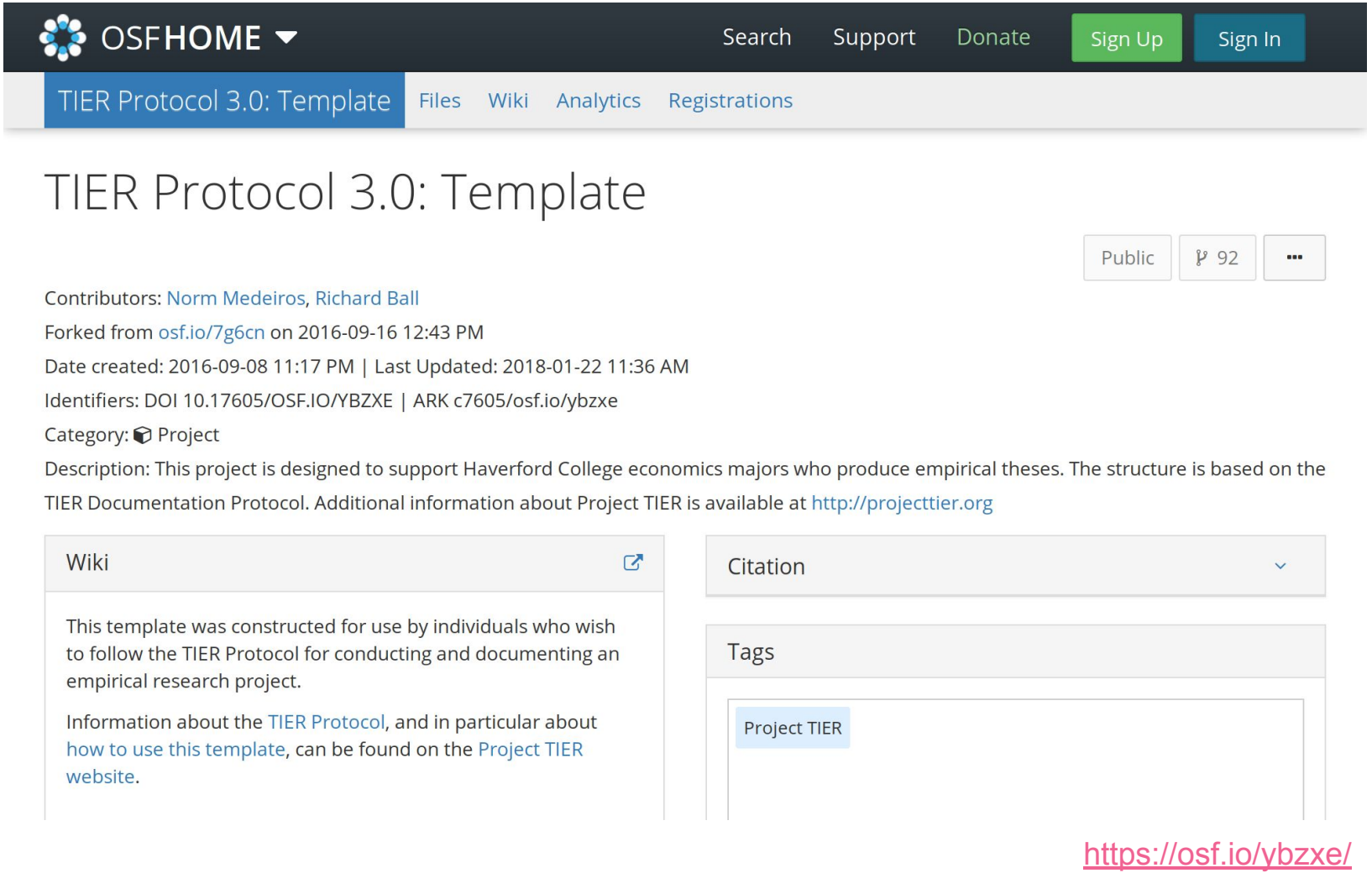




TIER Protocol Documentation

- Original Data
  - Original data files
  - Importable data files (if necessary)
  - Metadata
    - The Metadata Guide
    - Supplementary metadata documents (if necessary)
- Documents
  - The final paper
  - The Data Appendix
  - The Read Me file
- Analysis Data
  - Analysis data files
- Command Files
  - Command files

[The TIER Specification]



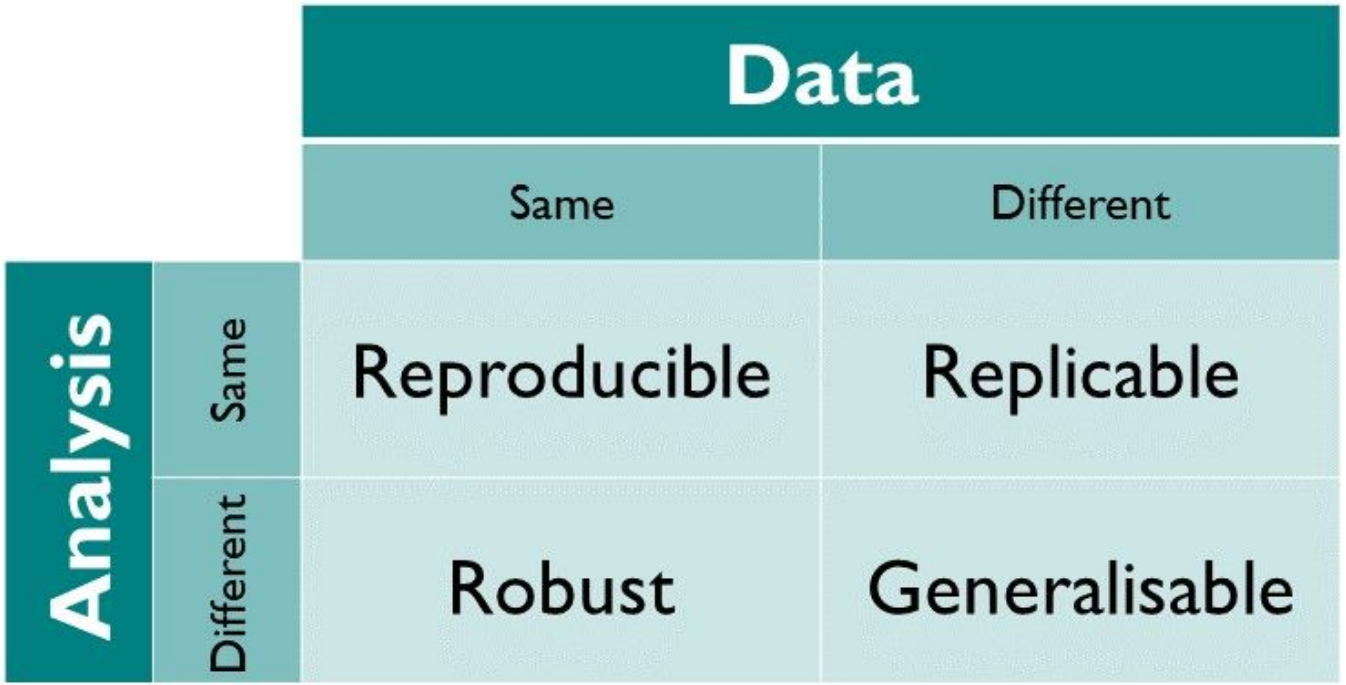
https://osf.io/ybzxe/

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

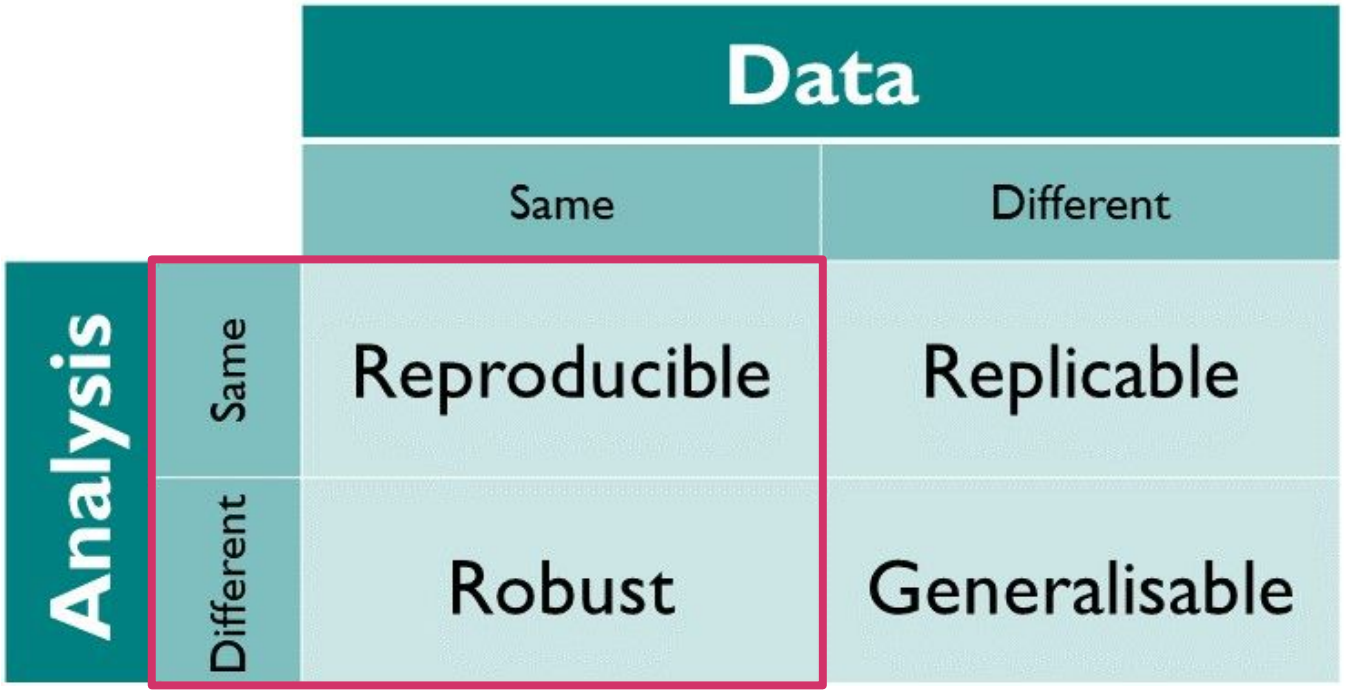
Community-based project management standards

Reproducibility as a roadmap



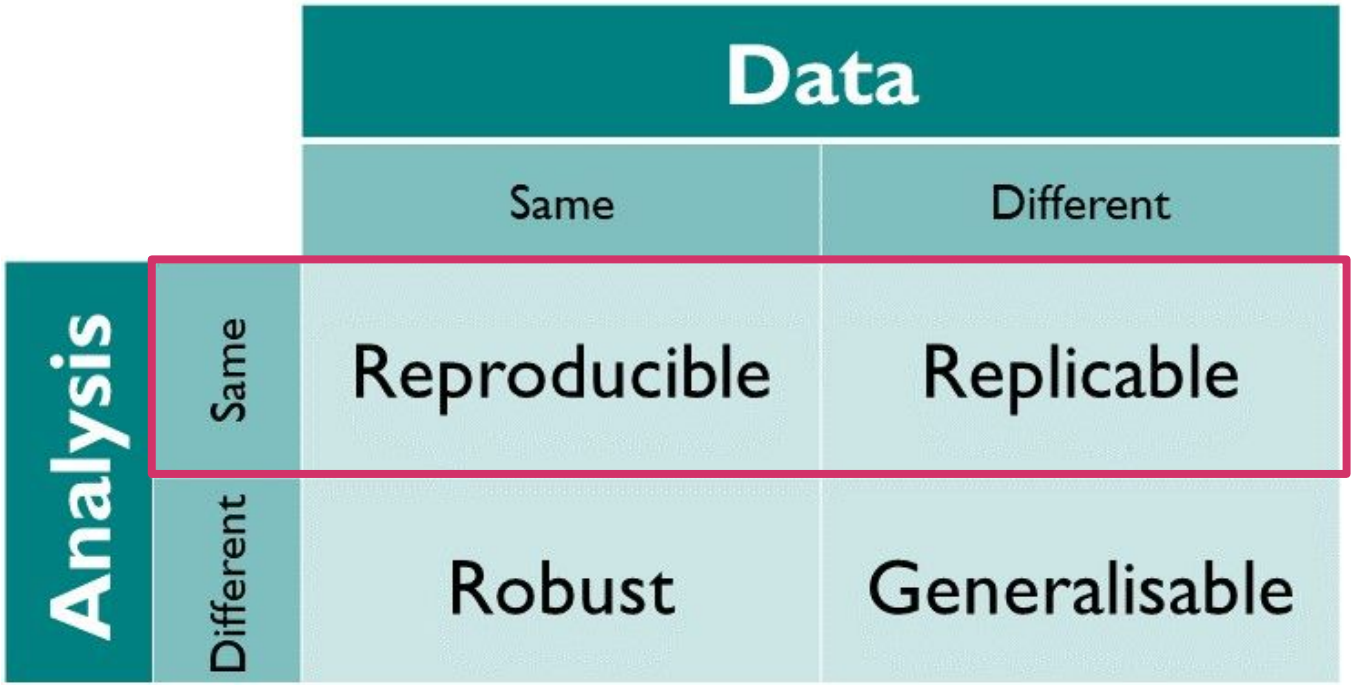
The Turing Way. doi: 10.5281/zenodo.3233853

Reproducibility as a roadmap



The Turing Way. doi: 10.5281/zenodo.3233853

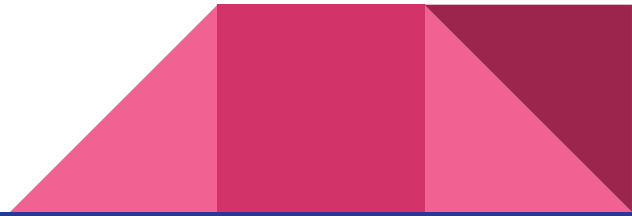
Reproducibility as a roadmap

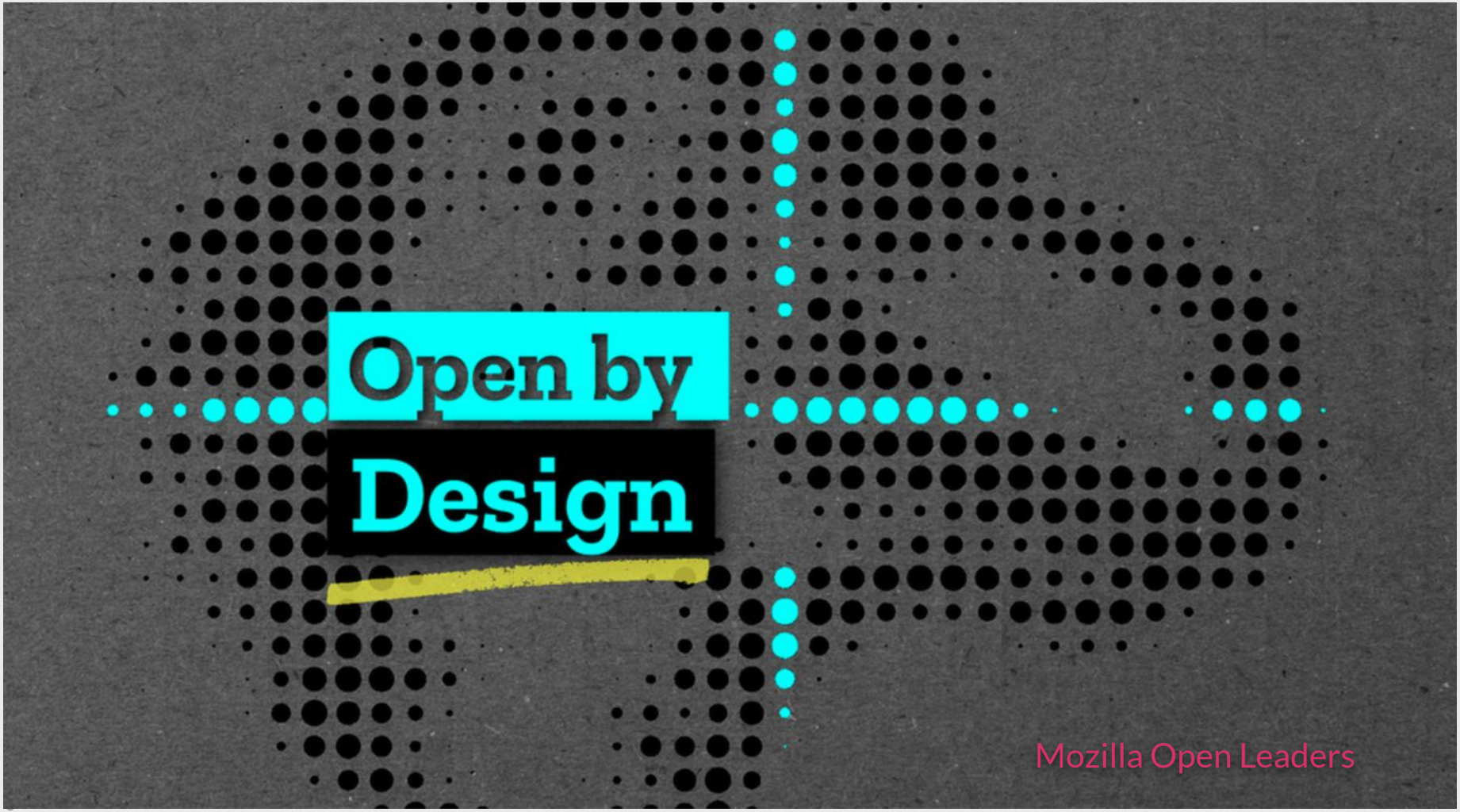


The Turing Way. doi: 10.5281/zenodo.3233853

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?

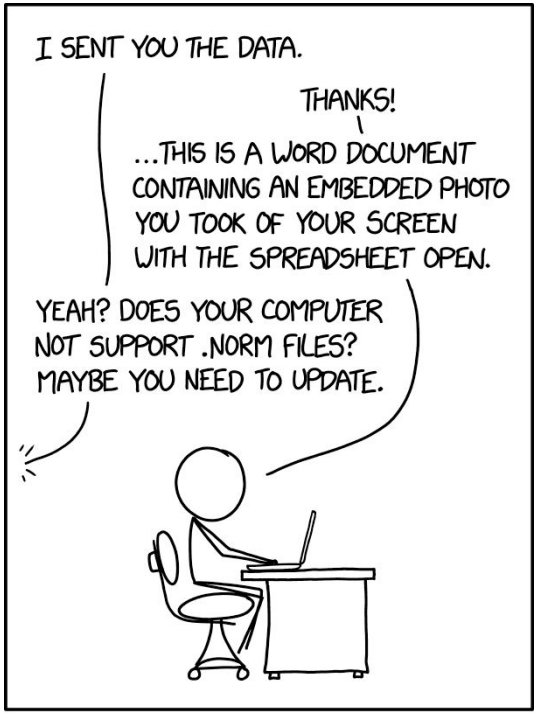



Open by
Design
Mozilla Open Leaders

“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

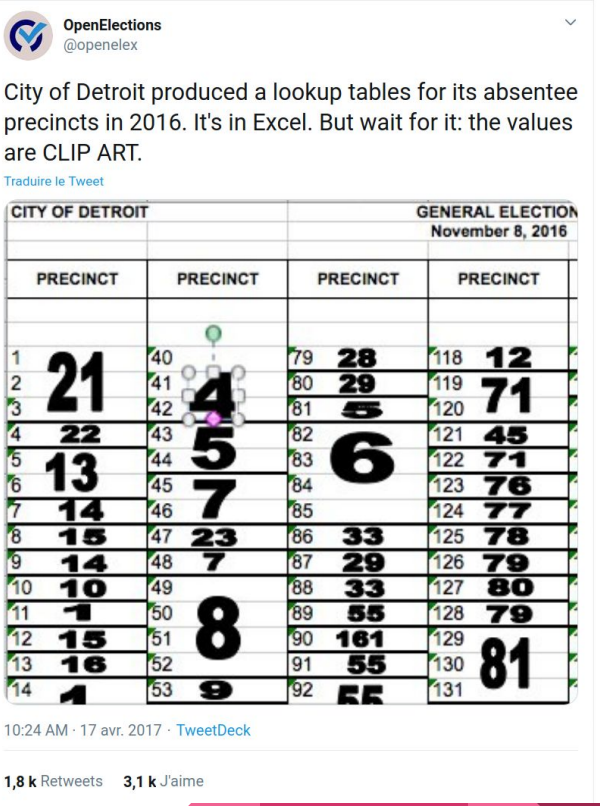


SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

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



We want to be FAIR

https://ogsl.ca/en/fair-principles

FAIR: Findable

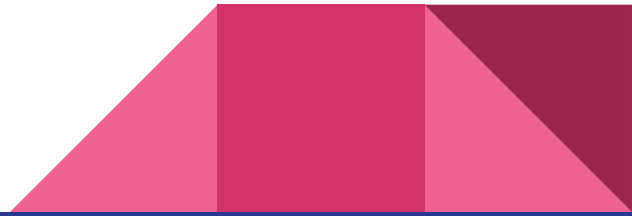
Findable: the first step in (re)using data is to find them!

- 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

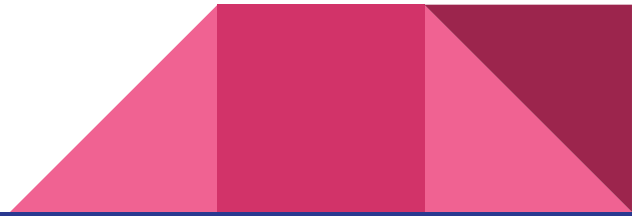
- A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
- A2. Metadata are accessible, even when the data are no longer available



FAIR: Interoperable

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

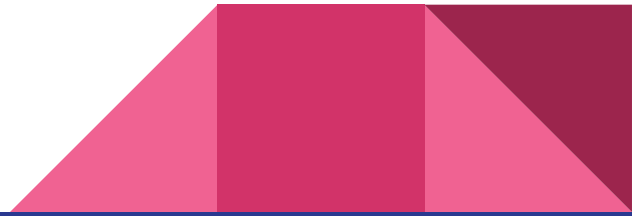
- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data



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

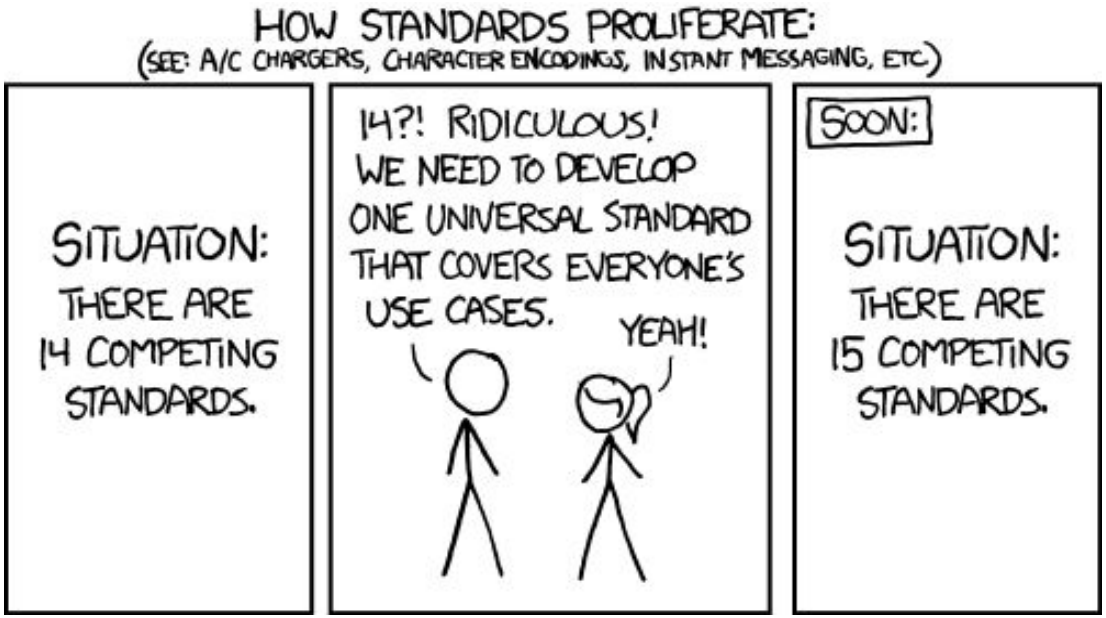


FAIR is fuzzy

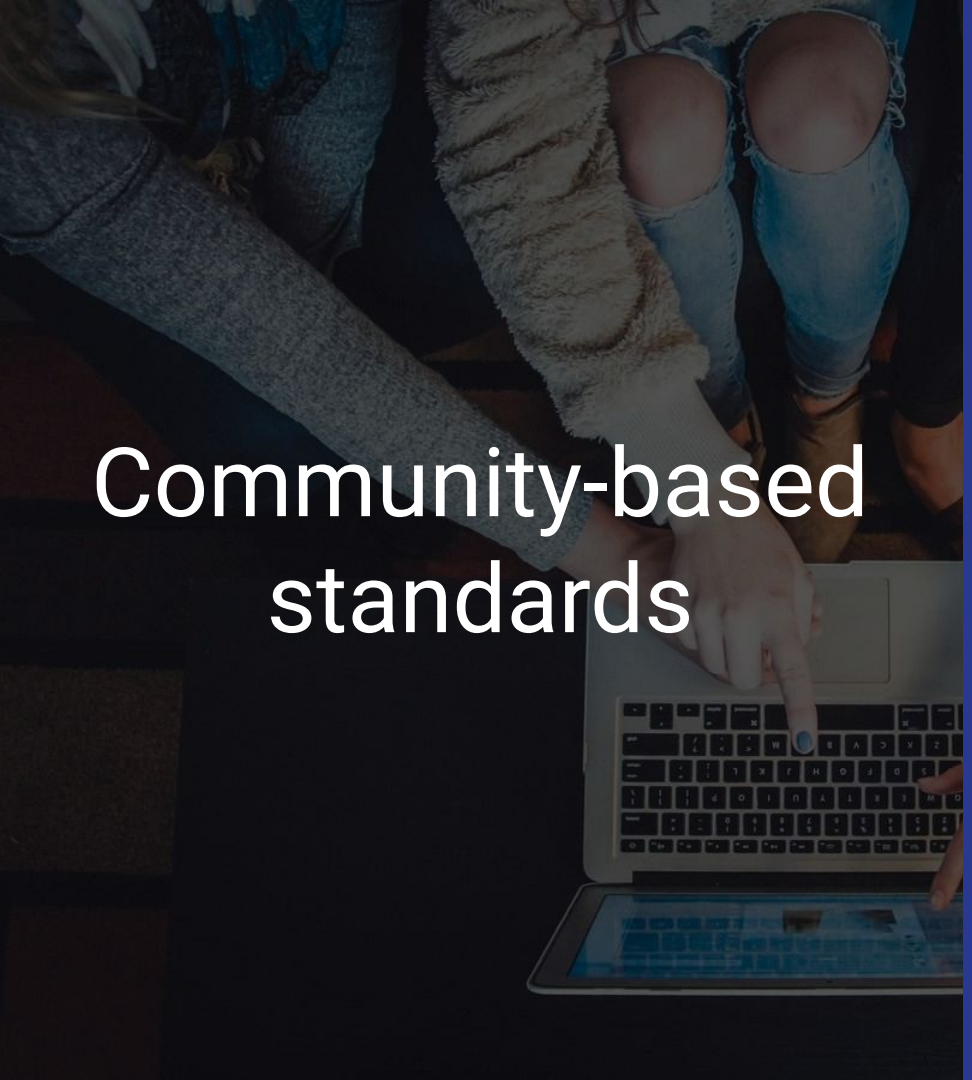
- This is by design ! FAIR is a general set of principles that should guide research across domains
  - What does it mean to have “rich metadata” ? Is that different for neuroimaging vs genomics research ?
- Each discipline needs its own FAIR standards

Community-based project management standards

Houston, we have a problem



XKCD



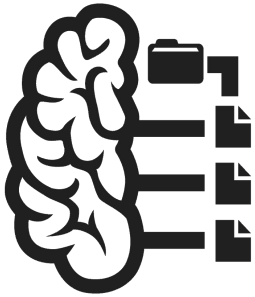
Community-based standards

- 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



BIDS

BRAIN IMAGING DATA STRUCTURE

NEURODATA WITHOUT BORDERS

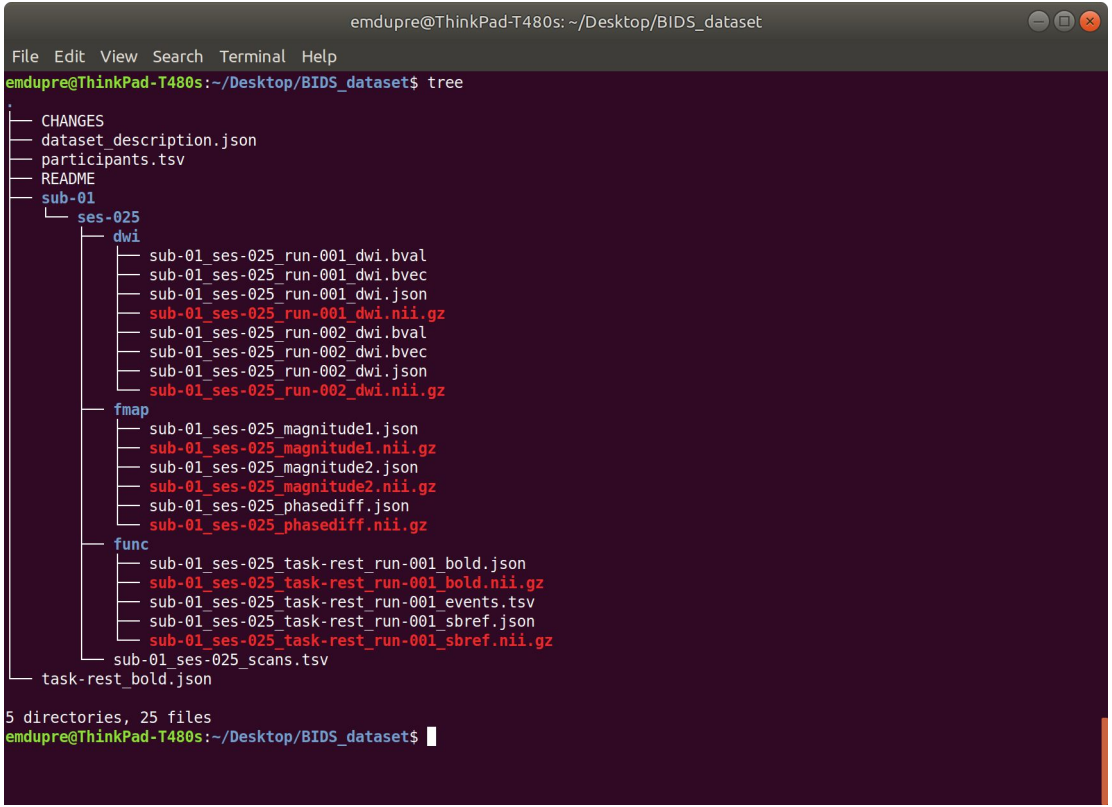
Brain Imaging Data Structure

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.

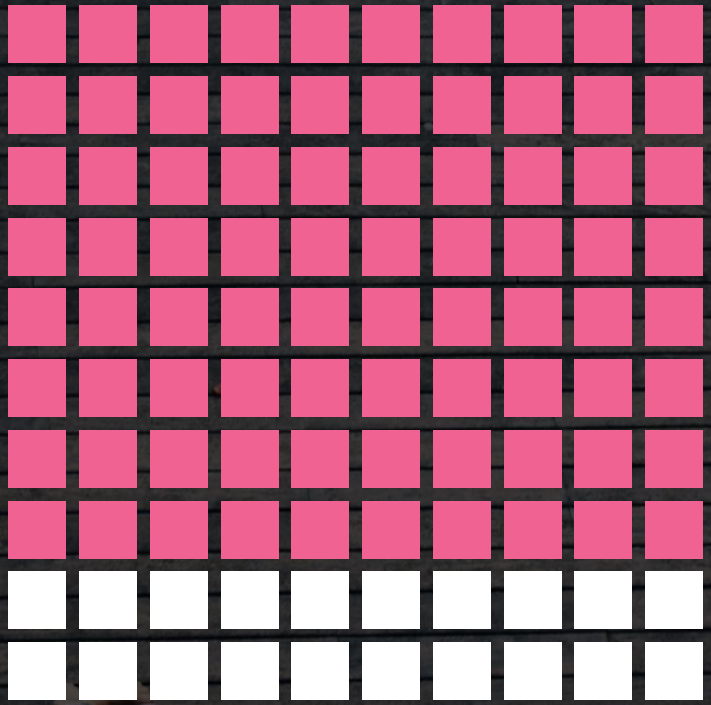
To get started, check out the introduction. If you'd like more information on how to adapt your own datasets to match the BIDS specification, we recommend exploring the bids-specification starter kit.

Example BIDS dataset layout



The 80 / 20 rule

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



bids-standard / **bids-specification**

Unwatch releases 33 | Star 77 | Fork 47

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
- **Proposal: BEP PRs should come from branches on the main repository** community — 4
  #203 opened 13 days ago by effigies
- **BEP Template** community — 1
  #201 opened 15 days ago by effigies
- **[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

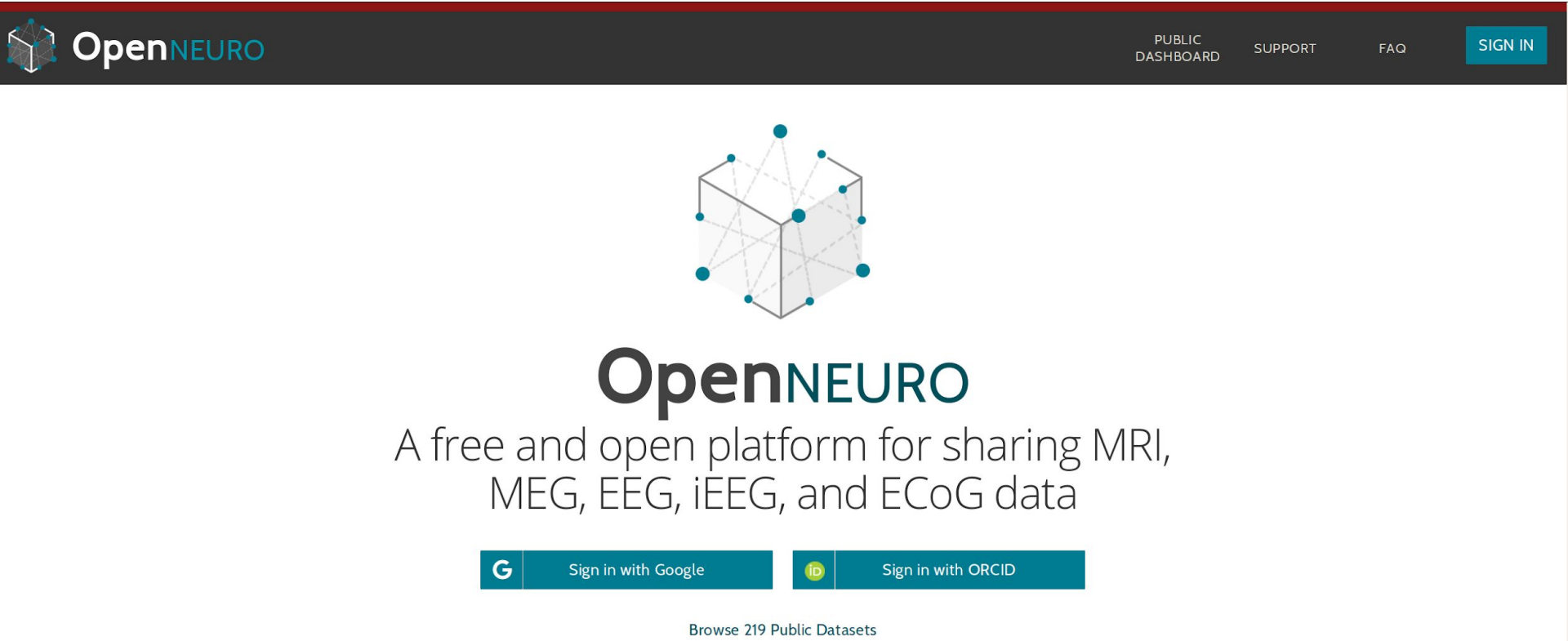
A community-curated collection of tutorials, wikis, and templates to get you started with creating BIDS compliant datasets.

BIDS Homepage | Wiki | Standard | Tutorials | Chat | Forum

Click to view the intro video!

Checking in on Professor Smith

With thanks to *Chris Gorgolewski*



OpenNeuro.org

Recent Activity

Most Active

20,329	UCLA Consortium for Neuropsychiatric Phenomics LA5c Study
19,376	The Midnight Scan Club (MSC) dataset
10,341	Visual object recognition
9,449	Forrest Gump
8,599	Classification learning

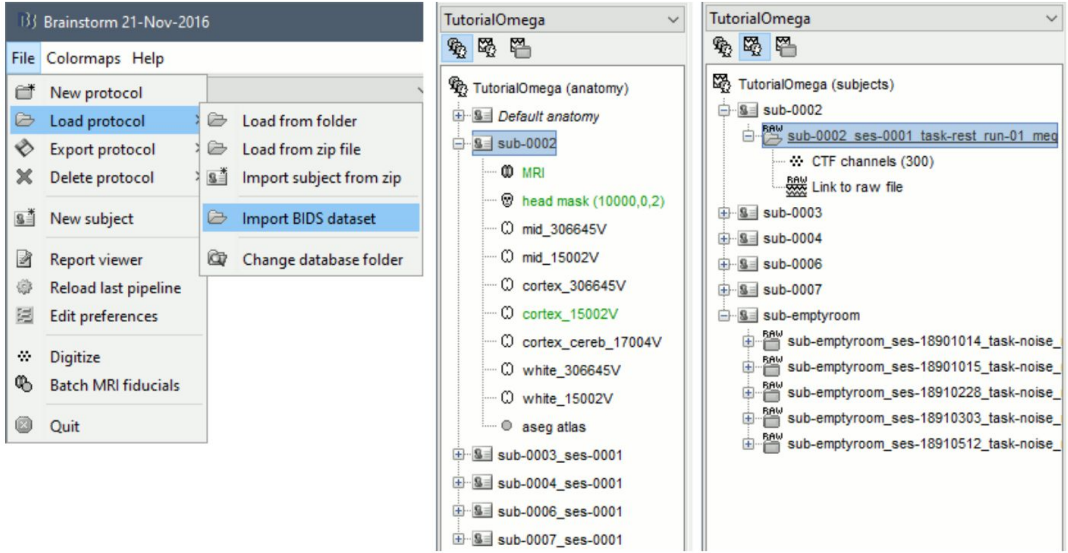
View More...

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

- 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

BIDS Apps

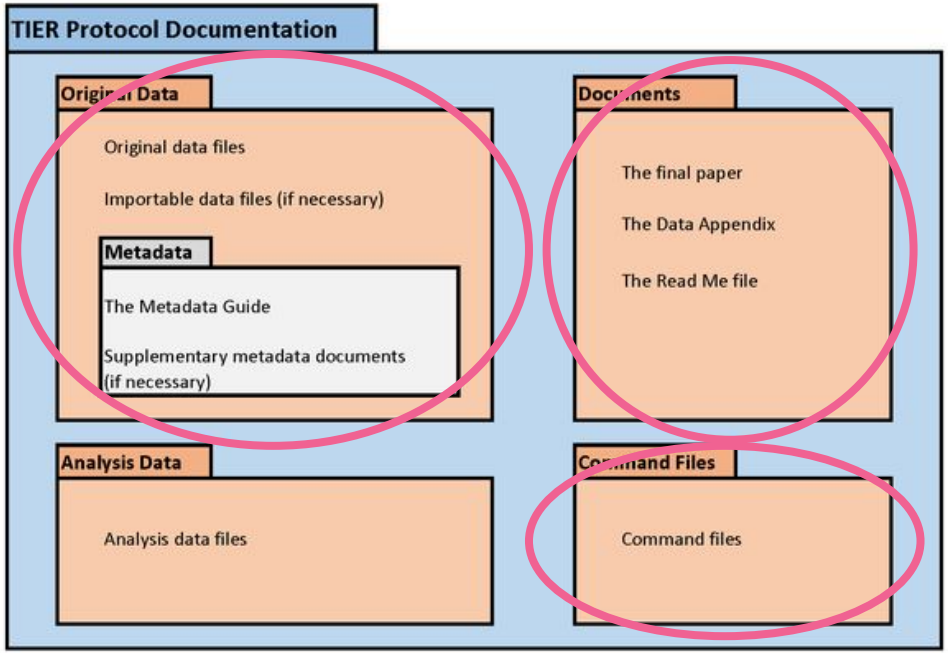
portable neuroimaging pipelines that understand BIDS datasets

About Tutorials Apps

Professor Smith (circa 2030)

With thanks to *Chris Gorgolewski*

Reproln



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