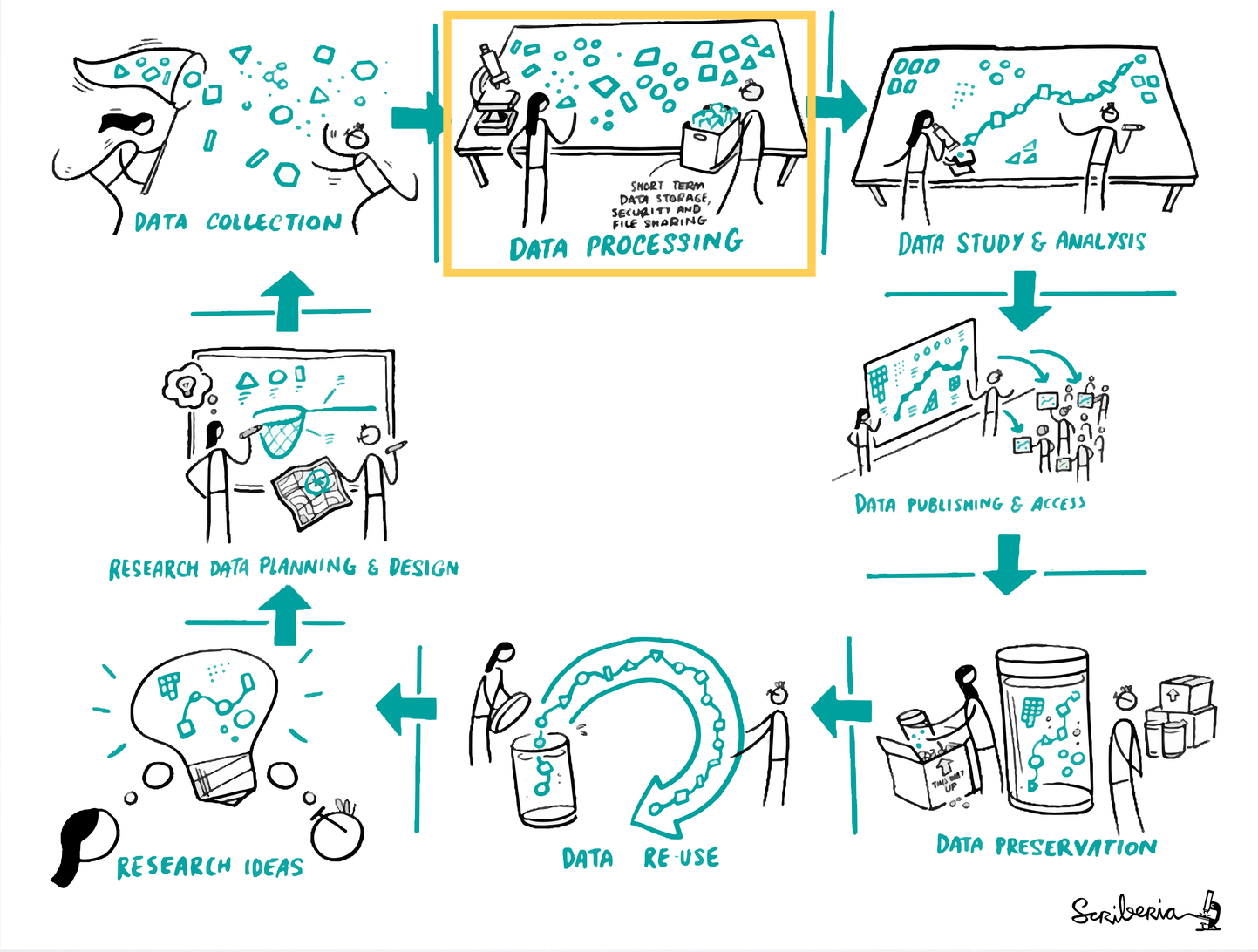
Data Management: Practicalities

Alaina Pearce

Project Lifecycle



Topics

- File Naming
- Directory Structures
- Metadata
- Version Control

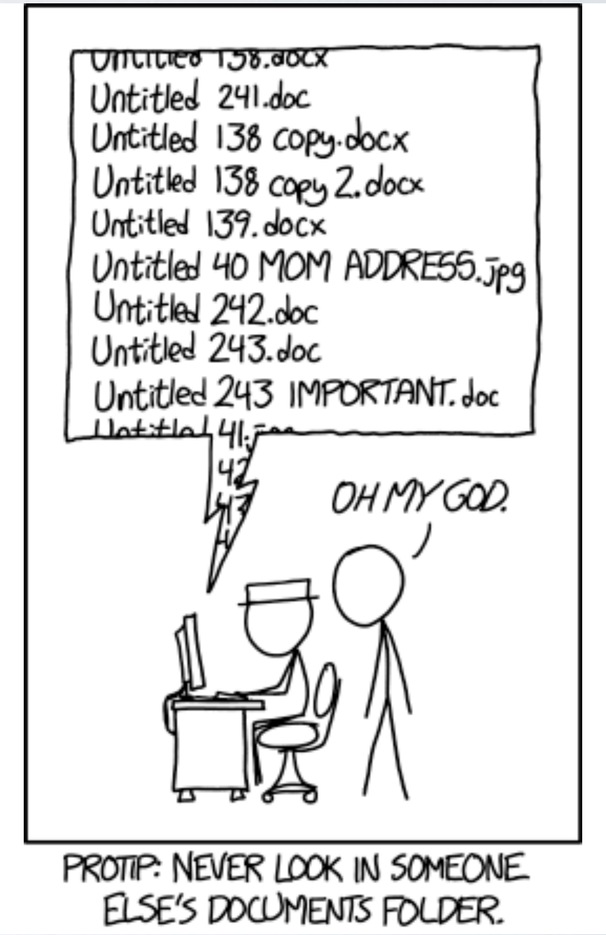
The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 license. DOI: 10.5281/zenodo.3332807.

File Names

Leverage filenames to help you manage complex projects

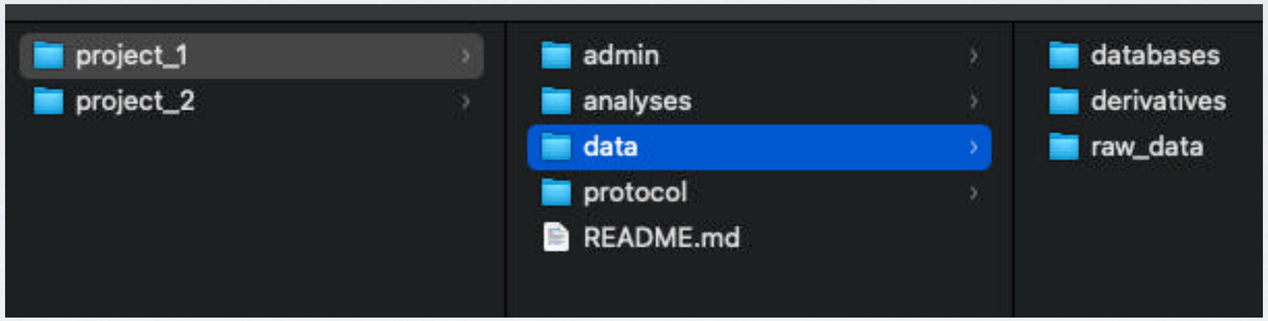
Deep Dive:

- What to consider in file names
- Best practices
- Examples of field standards (e.g., BIDS, MIxS)



Directory Structure

Directory Structures: organization of files into a hierarchical structure



Deep Dive:

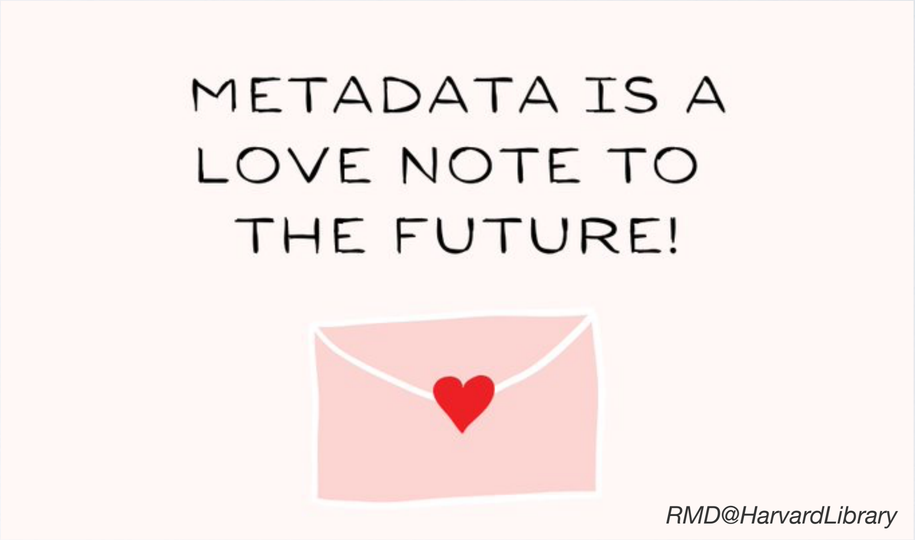
- Why create a hierarchy?
- How to create a hierarchy based on data into categories or attributes
- Examples of field standards (e.g., BIDS, MIxS)

Metadata

Metadata: data that provides information about other data

Deep Dive:

- Types of metadata
- How to provide metadata
- Examples of README information



Version Control

Deep Dive:

- ‘Good enough’ approaches
- Code-based approaches

File Names

Goal:

- Identify file/contents in a clear way
- Have a consistent approach across projects and collaborators
- Should be meaningful but brief

File Names

Goal:

- Identify file/contents in a clear way
- Have a consistent approach across projects and collaborators
- Should be meaningful but brief

Do Not Use

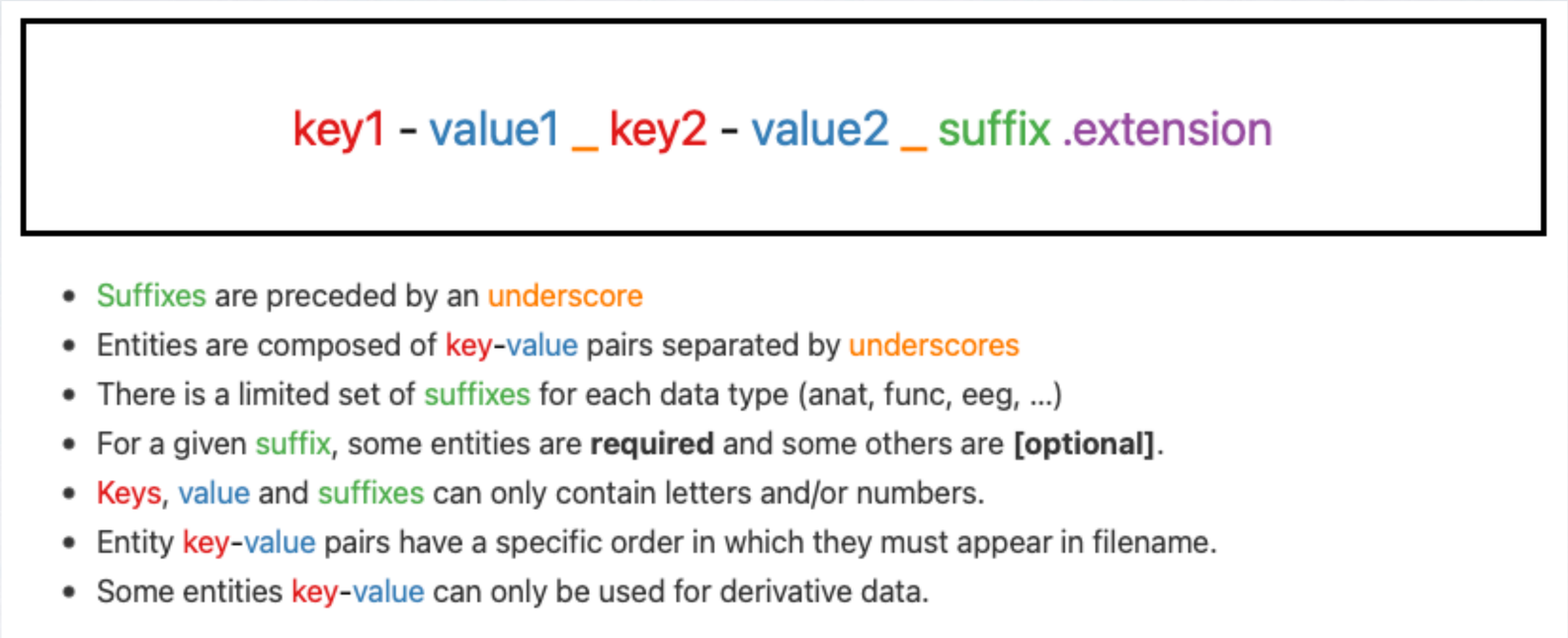
- Spaces
- Periods (except for file extensions)
- Other special characters (&, *, ^, etc)

Use

- CamelCase
- Underscores (_)
- Consistent date format - YYYYMMDD recommended
- Pad with zeros when using numbers

File Names

Example: Brain Imaging Data Structure



sub-035_task-flanker_events.txt
sub-035_ses-2_task-flanker_events.txt

https://bids-standard.github.io/bids-starter-kit/folders_and_files/files.html

Directory Structures

Hierarchical Structures - makes it easier to find what you are looking for

- Names
- Structures
- Relationships

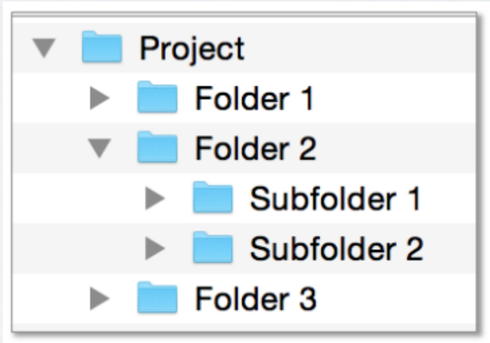
Goals

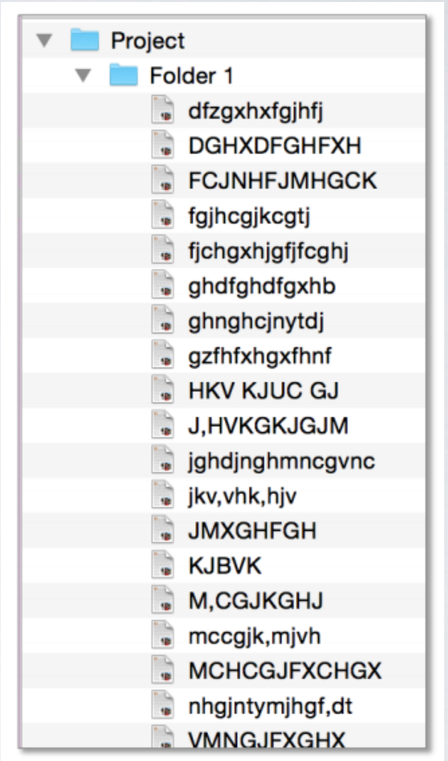
- Transparent
- Consistent — allows for continuity in the future and across projects
- Built BEFORE data collection (if possible)

Directory Structures

General Best Practices

- Structure logically based on project
- Keep subfolder categories narrow to limit number of files in each one
- Define abbreviations in README
- Follow file naming best practices



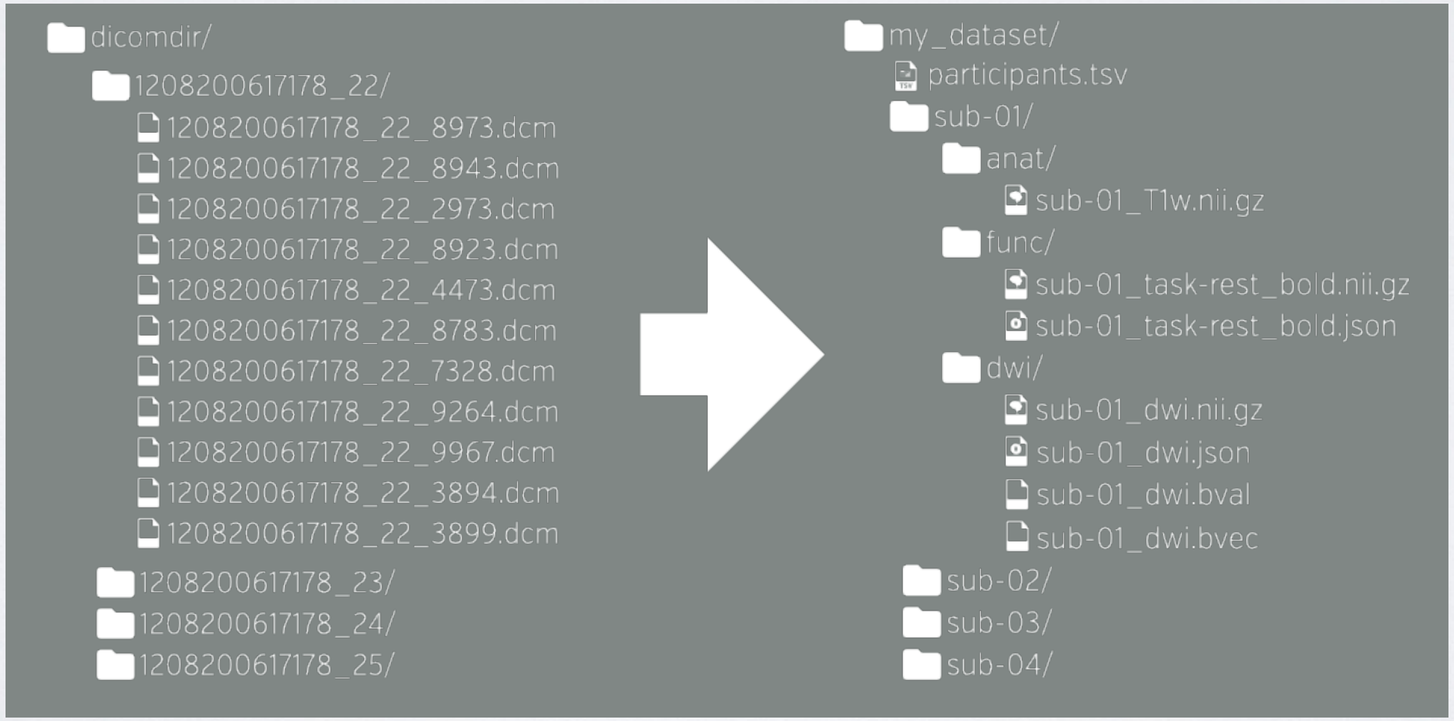


Directory Structures

Tricky Choices

Organize by data type vs sample/participant?

- Brain Imaging Data Structure

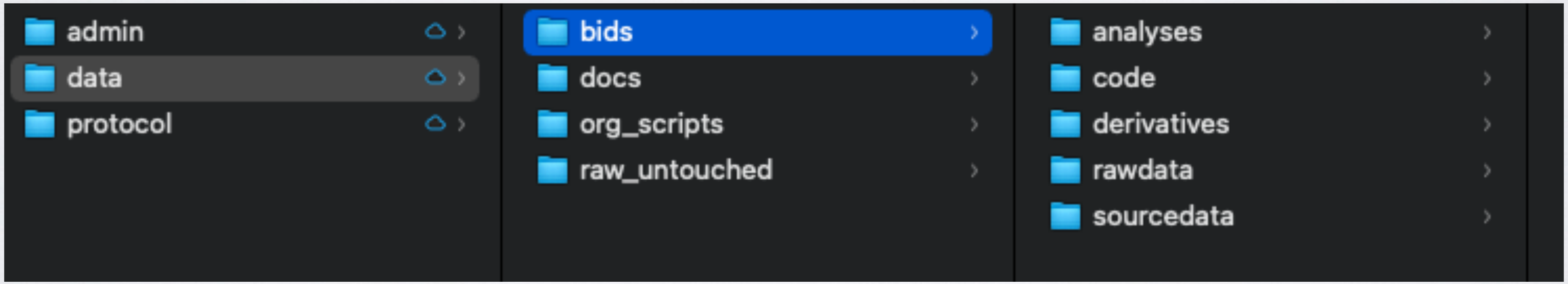


Directory Structures

Tricky Choices

Store RAW data in data directory vs in separate folder?

- Separate folder - will copy from raw_data directory to data directory
- Same folder - risk individuals touching/using only copy of raw data
- Duplicate??



Metadata

Types of Metadata:

- Directory structure and definitions (README)
- Project information (authors, funding - README)
- Data dictionaries
- Pre-processing steps/information (e.g., software versions, processing steps)
- Data Manuals
  - Larger - often combines multiple sources of metadata
  - More verbose protocol descriptions
  - Can include some science/rational/citations

Metadata

Metadata Standards:

- https://www.dcc.ac.uk/guidance/standards/metadata
- https://rdamsc.bath.ac.uk

Metadata

README:

- Should go in the top folder of your directory hierarchy
- Standard file format is Markdown
- What to include?
  - Project title and description
  - Authors and funding
  - License information
  - Directory and file organization
  - How to use the project/data

Metadata

README Example: Harvard Research Library

Dataset Title: Raw Images for Experiment A, Smith Lab

Principal Investigator: John Smith, PI, 555-555-5555, jsmith@hms.harvard.edu

File Naming Convention:
ExperimentName_InstrumentID_CaptureDateTime_ImageID.tif
The base file name is composed of the name of the experiment, the ID number of the instrument used, the date and time that the image was captured, and the unique identifier of the image.

Attributes: Also see the Codes section for a list of instruments and their ID numbers

- ExperimentName = Name of the experiment
- Instrument ID = Five-digit code assigned to the lab instrument
- CaptureDateTime = Date and time at which the image was captured, in YYYYMMDD format
- Image ID = Three-digit unique identifier for image, such as 001, 002, 003
-

Codes:

- [List of instruments and IDs]
-

Examples:

- File formats: daf2-age1_14052_20150412T0515_005.tif
- Versioning: All changes to this dataset will be documented in a changelog in this README file

Metadata

Project Metadata Example: Brain Imaging Data Structure


```
{“Name": "The mother of all experiments",
 "BIDSVersion": "1.6.0",
 "DatasetType": "raw",
 "License": "CC0",
 "Authors": [
   "Paul Broca",
   "Carl Wernicke"
 ],
 "Acknowledgements": "Special thanks to Korbinian Brodmann for help in formatting this
dataset in BIDS.",
 "HowToAcknowledge": "Please cite this paper: https://www.ncbi.nlm.nih.gov/pubmed/
001012092119281",
 "Funding": [
   "National Institute of Neuroscience Grant F378236MFH1",
 ],
 "EthicsApprovals": [
   "Army Human Research Protections Office (Protocol ARL-20098-10051, ARL 12-040,
and ARL 12-041)"
 ],
 "ReferencesAndLinks": [
   "https://www.ncbi.nlm.nih.gov/pubmed/001012092119281",
   "Alzheimer A., & Kraepelin, E. (2015). Neural correlates of presenile dementia in
humans. Journal of Neuroscientific Data, 2, 234001. doi:1920.8/jndata.2015.7"
 ],
 "DatasetDOI": "doi:10.0.2.3/dfjj.10",
```

Metadata

Data Dictionary:

- Where to put it?
  - Specific to a single database: save with the database
  - Generalizes to many files (e.g., for each sample/participant): Save at highest directory structure that contains all files (e.g., data directory)
- What to include?
  - Variable names
  - Full variable definitions
  - Optional:
    - Number of observations
    - Ranges
    - Type of data
    - License information (if different from README)

Metadata

Data Dictionary Example

column	variable	label	value_labels	type	n_na	range
1	id	ID	NULL	double	0	c(1, 133)
2	v1_date	date from participant contacts databases ('verified_visit_da	NULL	character	0	c("2018-01-31", "2022-05-07")
3	bmi_screenout	Child BMI Percentile Screen Out	c(`YES, child is overweight, sc	double	0	c(0, 1)
4	parent_respondent	Parent Reported: Parent relationship to child re-leveled in R	c(Mother = 0, Father = 1, Oth	double	0	c(0, 1)
5	parent_respondent_o	Parent Reported: Parent specify relationship to child if other	NULL	character	0	c("", "")
6	hw_measured	Parent attending Visit 1 had measured height and weight	c(No = 0, Yes = 1)	double	0	c(1, 1)
7	measured_parent	Parent with measured BMI at Visit 1	c(mom = 0, dad = 1)	double	0	c(0, 1)
8	risk_status_mom	Child risk categor: Low risk: Mom BMI < 26, High Risk: Mom	c(`Low Risk` = 0, `High Risk` =	double	0	c(0, 1)
9	risk_status_both	Child risk category: Low Risk: Mom and Dad BMI < 25, High	c(`Low Risk` = 0, `High Risk` =	double	0	c(0, 2)
10	sex	Child Sex re-leveled in R to start with 0	c(Male = 0, Female = 1)	double	0	c(0, 1)
11	dob	date of birth converted to format yyyy-mm-dd in R	NULL	double	0	c(14333, 16391)
12	age_yr	Age in years calculated from dob and start_date	NULL	double	0	c(7, 8.99)
13	age_mo	Age in months calculated from dob and start_date	NULL	double	0	c(84, 107.9)
14	ethnicity	Parent Reported: Child ethnicity	c(`NOT Hispanic or Latino` = 0	double	0	c(0, 0)
15	race	Parent Reported: Child race -- Note: prefer not to answer (p	c(`White/Caucasian` = 0, `Am	double	0	c(0, 2)
16	income	Parent Reported: Yearly household income -- Note: prefer n	c(`Less than $20,000` = 0, `$2(	double	3	c(0, 5)
17	parent_ed	Parent Reported: Parent education re-leveled in R to start w	c(`High School or GED (12 yea	double	0	c(0, 5)

- Generated in R using labeled data (sjlabelled) and labelled::generate_dictionary
- SPSS and REDCap also have built-in tools

Metadata

Data Manual Best Practices:

- Include table of contents
- Go from broad to detailed
- Goals: define data and provide instructions for use
  - Should be able to pull from it to start writing outlines for paper methods sections
- Document all steps in your pipeline

Version Control

Goal: distinguish between individual file versions and maintain previous versions

'Good enough' practices

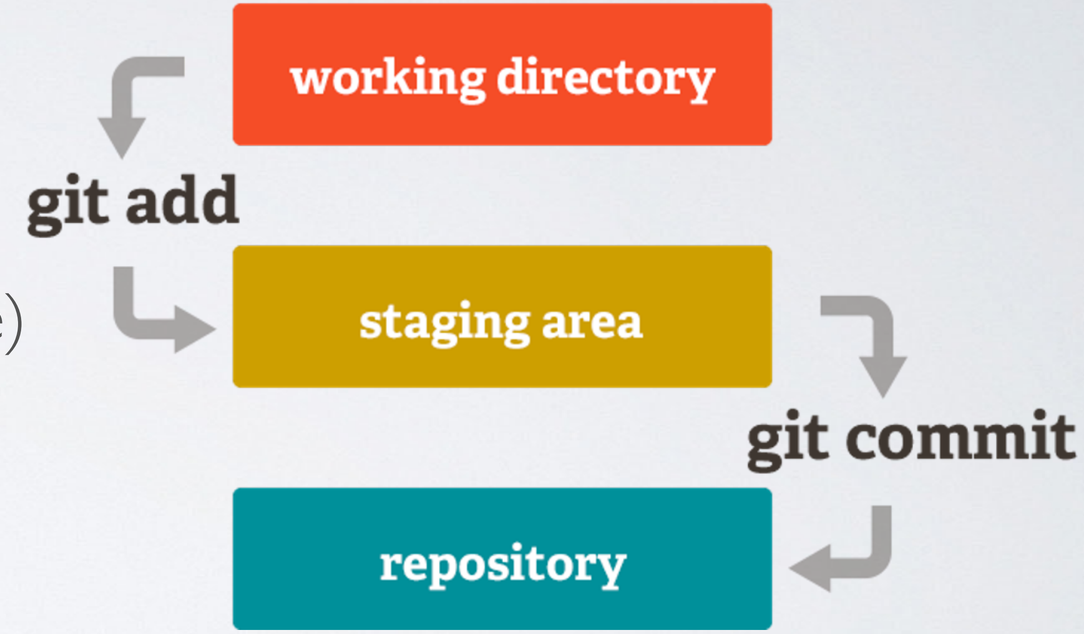
- Maintain a master file with original content
- Make new versions when meaningful changes/updates have been made
- Use clear file naming conventions with version numbers (_v2) or dates (YYYYMMDD)
- Keep file history in CHANGES or README

Version Control

Goal: distinguish between individual file versions and maintain previous versions

git

- Keeps snapshots of entire projects
- Documentation integrated in commits
- Command line based system
- Integrated with R (other GUIs also available)

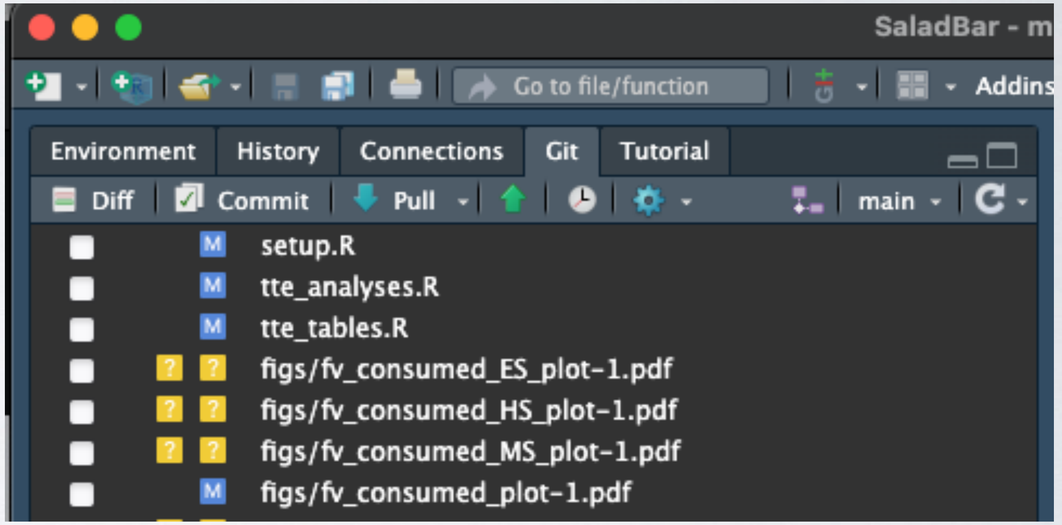


Version Control

Goal: distinguish between individual file versions and maintain previous versions

git

- Keeps snapshots of entire projects
- Documentation integrated in commits
- Command line based system
- Integrated with R (other GUIs also available)

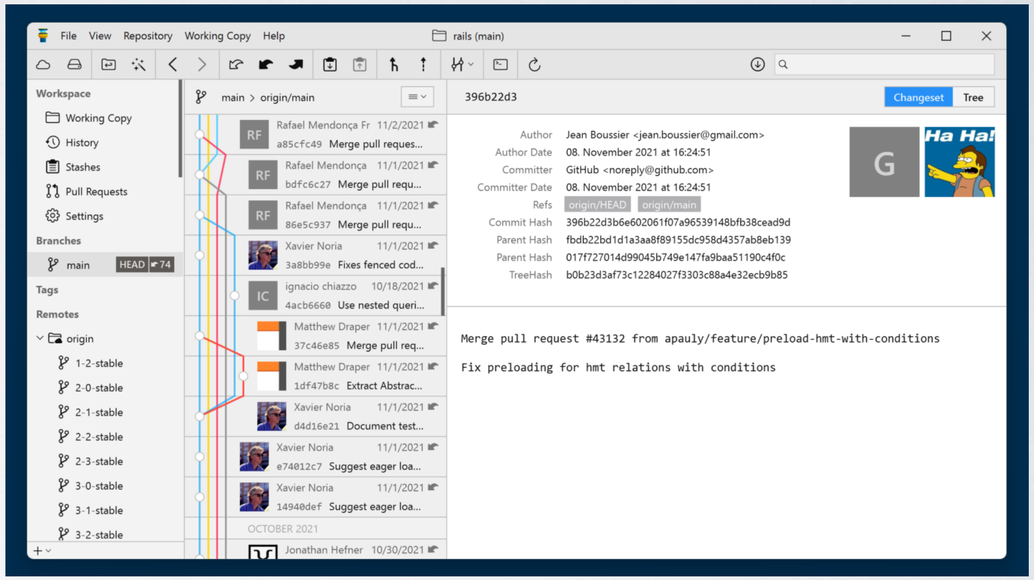


Version Control

Goal: distinguish between individual file versions and maintain previous versions

git

- Keeps snapshots of entire projects
- Documentation integrated in commits
- Command line based system
- Integrated with R (other GUIs also available)
  - Git-tower (https://www.git-tower.com/mac)

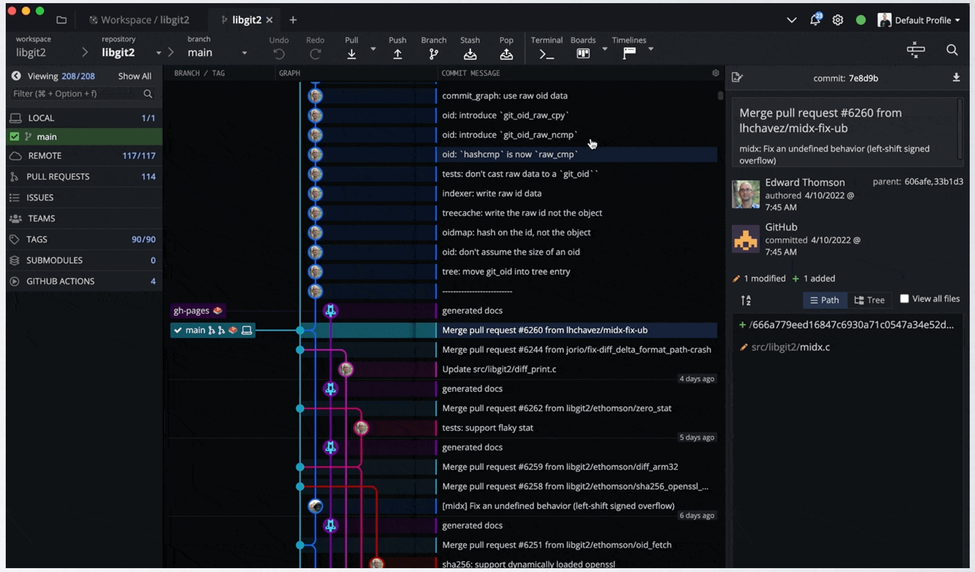


Version Control

Goal: distinguish between individual file versions and maintain previous versions

git

- Keeps snapshots of entire projects
- Documentation integrated in commits
- Command line based system
- Integrated with R (other GUIs also available)
  - Git-tower (https://www.git-tower.com/mac)
  - Git-Kraken (https://www.gitkraken.com/)

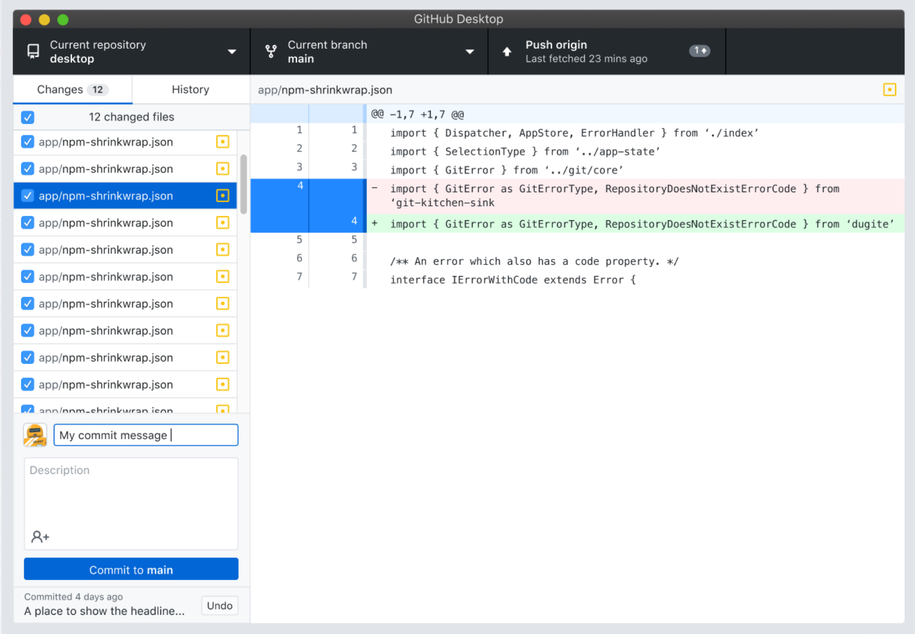


Version Control

Goal: distinguish between individual file versions and maintain previous versions

git

- Keeps snapshots of entire projects
- Documentation integrated in commits
- Command line based system
- Integrated with R (other GUIs also available)
  - Git-tower (https://www.git-tower.com/mac)
  - Git-Kraken (https://www.gitkraken.com/)
  - GitHub Desktop (https://desktop.github.com/)



Helpful Resources

- https://www.dropbox.com/s/0lysIli2wk1al6o/Template_README_fileOrg.txt?dl=0
- https://www.dropbox.com/sh/3lwnnab54o55hqg/AAD0_6mwZW3vH4xkILFwkcDza/Sample_README_fileOrg.docx?dl=0
- https://datamanagement.hms.harvard.edu/collect-analyze/documentation-metadata/readme-files
- https://data.research.cornell.edu/content/readme
- https://www.dropbox.com/s/ttv3boomxlfgiz5/Handout_fileNaming.pdf?dl=0
- https://www.employedforgood.com/how-to-create-a-user-manual-for-your-database-7-steps/