• add menti meter set up slide
• add explanation of all lifecycle steps - what do they mean, have menti meter poll - after slide 5
• do you have a naming schema
• how do you
First Question:
What does data management mean to you?
1. Principles of data management
2. Avoiding mishaps and headaches
3. Purposeful documentation
4. Publishing data and code
5. Closing thoughts
6. Questions
The (Research Data) Future is Open

**Office of Science & Technology Policy**

- Issues directive to federal funders with >$100m annual R&D budgets: Data should be stored and publicly accessible to search, retrieve, and analyze.

**Office of Science & Technology Policy**

- Issues directive to *all* federal funders: Data underlying publication should be immediately accessible; all data shared; should use repositories for sharing data.

**National Institutes of Health**

- Implement Data Sharing Requirement for certain kinds of research grants.

**National Institutes of Health**

- Begins requiring Data Management Plans for research grants.

**National Science Foundation**

- Begins requiring Data Management Plans for research grants.

**United Nations**

- Issues recommendations for Open Science “transformative potential of open science for reducing the existing inequalities ... and accelerating progress towards ... Sustainable Development Goals”.

**National Institutes of Health**

- Begins requiring Data Management and Sharing Plans for *all* research grants.
Principles of Data Management
What Does “Managing Data” Mean?

- Happens throughout the research data lifecycle
- Making intentional decisions on how you collect, organize, store, preserve, and share your data
- Documenting your decisions
### The Research Data Lifecycle

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan</td>
<td>Apply for funding, write proposals, start new collaborations</td>
</tr>
<tr>
<td>Acquire</td>
<td>Find data, create data, understand permissions to use data</td>
</tr>
<tr>
<td>Process</td>
<td>Limit data to that which is relevant to your research question</td>
</tr>
<tr>
<td>Analyze</td>
<td>Tailor data for use with programs or processes that help explore or explain the data</td>
</tr>
<tr>
<td>Preserve</td>
<td>Put files into format(s) and place(s) that reduce the likelihood of data degradation/loss</td>
</tr>
<tr>
<td>Share Results</td>
<td>Place data in a location where other researchers can find it</td>
</tr>
<tr>
<td>Discover &amp; Re-use</td>
<td>Make your data more valuable by enriching its metadata as much as possible (link it to Orcid/journal articles/etc)</td>
</tr>
</tbody>
</table>
Go to Menti.com code: 9863 5615

For one project right now, which step are you currently at in the research data lifecycle?
Research Data Stewardship

Taking responsibility for research data
- Both as a producer and as a secondary user
- Requires vigilance throughout the lifecycle

Practicing academic citizenship
- Provide for others what you expect them to provide for you
- Unlock a multiplier effect for resources spent

Developing better habits
- Not difficult conceptually
- Just requires practice (easier said than done)
Data Management Planning

“Data Management Plan”

- Formal document following specific requirements
- Submitted as part of an application for research funding

VS

A plan for managing data

- Living document to plan and document the processes around data
How do you plan to manage your data?

Who? Roles and responsibilities

What? Tasks and products

Where? Storage and access

When? Timetable
Avoiding Mishaps and Headaches
The 3-2-1 Rule for Backups

3 different copies of each file

2 different physical locations

1 offsite location
Scanning Physical Copies

BookEye Scanner
an overhead, large format scanner available on the C-floor (C-6K) of Firestone Library.
<table>
<thead>
<tr>
<th>Order in Place</th>
<th>Order in Perspective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Folder structure, file naming system, etc.</td>
<td>Stems from order in place</td>
</tr>
<tr>
<td>Always a good idea</td>
<td>Can remedy some disorder in place</td>
</tr>
<tr>
<td>Not always a complete solution</td>
<td>Fills out the solution</td>
</tr>
<tr>
<td>• Different places</td>
<td>• Centralized view</td>
</tr>
<tr>
<td>• Different systems</td>
<td>• Comprehensive sorting</td>
</tr>
<tr>
<td>• Different people</td>
<td>• Manage tasks and people</td>
</tr>
</tbody>
</table>
File Organization

**Nominal** – A descriptive and consistent file naming scheme

**Structural** – A coherent and easy-to-navigate folder hierarchy

**Relational** – An arrangement that highlights interdependencies

Be pragmatic
1. Pick a system that works
2. Write it down
3. Make it a habit
Effective Naming Schemes

Think in terms of metadata fields
- Namespace for sorting, sifting, and grouping
  - E.g., content type, location, purpose, date/time, source, etc.

Play nice with the machines
- No spaces or special characters
- No more than 255 characters in a full path
- Dates in the form YYYY-MM-DD (ISO 8601)
- Leading zeros for numbers

Semantic versioning
- Meaningful numbers, not ambiguous words (e.g., not “draft” or “final”)
  - major.minor.patch (e.g., 1.2.1)
Go to Menti.com code: **9863 5615**

Do you use a naming schema for your files?

Which metadata field do you think is most helpful to have in the naming schema?
Centralized overview

• Files, folders, locations, etc.
• Status and dates
• Content description
• Relations and dependencies

Fulfill multiple purposes at once

• Navigation
• Prioritization
• Progress tracking
• Data security assurance
# Data File Inventory Example

<table>
<thead>
<tr>
<th>FileName</th>
<th>FileLocation</th>
<th>CreatedBy</th>
<th>ManagedBy</th>
<th>DateCreated</th>
<th>DateUpdated</th>
<th>UpdateComment</th>
<th>Status</th>
<th>StatusComment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-01-08_NIH_SurveyW1_1-raw.csv</td>
<td>H:\Active\Data\NIHISurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-08</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-01-10_NIH_SurveyW1_2-cleaned.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-01-10_NIH_SurveyW1_3-labeled.dta</td>
<td>H:\Active\Data\NIHISurveys_3-Labeled</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-03-15_NIH_SurveyW2_1-raw.csv</td>
<td>H:\Active\Data\NIHISurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-15</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-03-16_NIH_SurveyW2_2-cleaned.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-16</td>
<td></td>
<td></td>
<td>ContainsErrors</td>
<td>Several variables miscoded</td>
</tr>
<tr>
<td>2019-03-18_NIH_SurveyW2_2-cleaned_v2.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-18</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td>Double-checked by Fran</td>
</tr>
<tr>
<td>2019-03-18_NIH_SurveyW2_3-labeled.dta</td>
<td>H:\Active\Data\NIHISurveys_3-Labeled</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-03-18</td>
<td>2019-03-19</td>
<td>Quick fix of script typo</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-08-28_NIH_SurveyW3_1-raw.csv</td>
<td>H:\Active\Data\NIHISurveys_1-Raw</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-08-28</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-08-30_NIH_SurveyW3_2-cleaned.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Sydney</td>
<td>Fran</td>
<td>2019-08-30</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-07-07_NIH_Surveys1to3_Long.dta</td>
<td>H:\Active\Data\NIHISurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2018-07-07</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-07-07_NIH_Surveys1to3_Wide.dta</td>
<td>H:\Active\Data\NIHISurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2019-07-07</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2019-09-10_NIH_SurveyW4_1-raw.csv</td>
<td>H:\Active\Data\NIHISurveys_1-Raw</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-09-10</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2019-09-15_NIH_SurveyW4_2-cleaned.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-09-15</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2019-09-16_NIH_SurveyW4_3-labeled.dta</td>
<td>H:\Active\Data\NIHISurveys_3-Labeled</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-09-16</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-06_NIH_SurveyW5_1-raw.csv</td>
<td>H:\Active\Data\NIHISurveys_1-Raw</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-06</td>
<td></td>
<td></td>
<td>Frozen</td>
<td></td>
</tr>
<tr>
<td>2020-01-10_NIH_SurveyW5_2-cleaned.csv</td>
<td>H:\Active\Data\NIHISurveys_2-Cleaned</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-10</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-12_NIH_SurveyW5_3-labeled.dta</td>
<td>H:\Active\Data\NIHISurveys_3-Labeled</td>
<td>Fran</td>
<td>Pat</td>
<td>2019-01-12</td>
<td></td>
<td></td>
<td>ErrorFree</td>
<td></td>
</tr>
<tr>
<td>2020-01-20_NIH_SurveysAll_Long.dta</td>
<td>H:\Active\Data\NIHISurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-20</td>
<td>2020-01-22</td>
<td>Regenerated in Shareable folder</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2020-01-20_NIH_SurveysAll_Wide.dta</td>
<td>H:\Active\Data\NIHISurveys_4-Shareable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-20</td>
<td>2020-01-22</td>
<td>Regenerated in Shareable folder</td>
<td>ReadyForReview</td>
<td></td>
</tr>
<tr>
<td>2020-01-22_YouthHealthAttitudesSurvey.csv</td>
<td>H:\Active\Data\NIHISurveys_5-Publishable</td>
<td>Pat</td>
<td>Pat</td>
<td>2020-01-22</td>
<td></td>
<td></td>
<td>ReadyForReview</td>
<td></td>
</tr>
</tbody>
</table>
Purposeful Documentation
Being a Good Citizen with Your Documentation

Provide for others what you would want them to provide for you

What would they need to:
- Understand the data?
- Replicate the study?
- Re-use the data?
What really needs to be written down?

Orienting information

- **What** kind of thing is this? What is it comprised of?
- **Who** made it? Based on whose prior work?
- **Why** was it made? What can it be used for?

The story behind the final product

- **How** you got from origin to endpoint
- Steps required to do it all over again
- Decision points, where others might fork off
**Study-Level Documentation**

- Research design
- Context for data collection
- Scope of samples and populations
- Data collection methods
- Quality assurance procedures
- Data transformation protocols

**Data-Level Documentation**

- Contents of data files
- Variables, labels, and types
- Values, measures, and units
- Encoding/interpretation methods
- Treatment of missing data
- Handling of private information
Standard Forms of Documentation

- **Metadata**
  - Context: Findability

- **README**
  - Context: Understanding

- **Codebook**
  - Usage: Metrics & Meanings

*aka “Data Dictionary”*
Elements of Good Metadata

- Both human- and machine-readable

- Adheres to a standard schema
  - Standards Directory from the Research Data Alliance

- Includes descriptions and relations
  - Title (distinct from corresponding paper)
  - Creators and contributors (with ORCID)
  - Abstract and keywords
  - Notes on collection and references to sources
  - Links to papers using the data (with DOIs)
Elements of a Good README

- **Basic, easily-accessible file type**
  - Plain text (.txt) in UTF-8 is preferred
  - Markdown (.md) is increasingly common

- **Distills standard metadata**
  - Give users the contextual information to retain after download
  - Especially: attribution guidance, permissions, and persistent identifiers

- **Supplements standard metadata** (if no codebook/data dictionary)
  - Background on research design and methodology
  - Outline files, contents, and relationships
  - Document variables and values with labels, types, and units
Elements of a Good Codebook/Data Dictionary

- **Outlines the scope and features of the data**
  - Conceptual and operational definitions of variables
  - Observed values within the space of possible values
  - Representative population and sampling techniques

- **Provides all definitive mappings**
  - Variable names to labels and data types
  - Raw variables to constructed variables
  - Values to labels and units
  - Labels to descriptions

- **Human- and machine-actionable**
  - Organized for interpretation and quick use
  - Includes descriptions along with codes and labels
  - Provides mappings and groupings in a computable structure
Image that two researchers had to leave their positions is a hurry. Based on their files, which work of the two would you want to continue with?

**Researcher 1**
- figure1.pdf
- figure2.pdf
- file1.py
- file2.py
- file3.py
- main.tex
- out_a_10
- out_a_20
- out_a_30
- out_b_1
- out_b_2
- out_b_3
- refs.bib
- output1.log
- output2.log

**Researcher 2**
- code
  - analysis.py
  - main.py
  - README
  - tests
    - test_analysis.py
- data
  - effect_of_length
    - length.log
    - system_length_10.csv
    - system_length_20.csv
    - system_length_30.csv
  - effect_of_width
    - width.log
    - system_width_1.csv
    - system_width_2.csv
    - system_width_3.csv
- manuscript
  - figures
    - length.pdf
    - width.pdf
  - text
    - main.tex
    - refs.bib
- README

Code was written by Alan Turing (aturing@gmail.com)

January 15, 2023
We investigate the effects of changes in length and width. It was shown previously that height is not important.
Hands On Activity: Directory README

Imagine that you are about to start a new line work on your research project.

1. Create a new directory for this new work using a descriptive name.
2. Inside this directory, create a README file with some practical notes such as
   (a) what is the purpose of the subproject
   (b) what will the directory structure be
   (c) include the date.

What else could you add that would be useful to someone taking over the work at a later time?
Key Points on README Files

• You should create README files liberally

• Don't worry about writing polished text inside the files
  • Be practical

• It is perfectly fine to create a README file in every directory
  • For most projects this is unnecessary

• Think of these files as notes to yourself at a later time
Publishing Data and Code
<table>
<thead>
<tr>
<th>Data Sharing</th>
<th>Data Publishing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definitions of “the data” vary</strong></td>
<td><strong>Explicit, citable digital object</strong></td>
</tr>
<tr>
<td>• Complete dataset?</td>
<td>• Clearly defined products</td>
</tr>
<tr>
<td>• Particular file? Version?</td>
<td>• Title, author, date, etc.</td>
</tr>
<tr>
<td>• Part of supplementary materials?</td>
<td>• Given a DOI</td>
</tr>
<tr>
<td><strong>Hosting conventions differ</strong></td>
<td><strong>Responsible publisher</strong></td>
</tr>
<tr>
<td>• A journal’s online version?</td>
<td>• Durable institution</td>
</tr>
<tr>
<td>• Project website?</td>
<td>• Professional curators</td>
</tr>
<tr>
<td>• GitHub? Whose account?</td>
<td>• Connected systems</td>
</tr>
<tr>
<td><strong>Life spans typically unclear</strong></td>
<td><strong>Preservation plan</strong></td>
</tr>
<tr>
<td>• Preserved by the journal?</td>
<td>• File integrity</td>
</tr>
<tr>
<td>• Dependent on the institution?</td>
<td>• Backups</td>
</tr>
<tr>
<td>• Websites maintained?</td>
<td>• Long-term hosting with stable URLs</td>
</tr>
</tbody>
</table>
What really needs to be published?

**You don’t have to stress about everything**

- Final products: cleaned, documented, and ready for re-use
- Not every raw instrument output or lab notebook

**Align with the incentives**

- Carrots: career advancement via open science
  - Evidence that papers with open data are getting cited at higher rates (Colavizza et al. 2020; Piwowar & Vision 2013; Piwowar et al. 2007)
- Sticks: funder/journal requirements
  - For example, the NSF expects that publicly-funded research data and software are shared along with the significant findings (see “Dissemination and Sharing of Research Results”)

(Colavizza et al. 2020; Piwowar & Vision 2013; Piwowar et al. 2007)
The FAIR Principles

Findable

Accessible

Interoperable

Reusable

https://www.go-fair.org/fair-principles/
## The FAIR Principles

<table>
<thead>
<tr>
<th>Findable</th>
<th>Globally unique and persistent identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Registered or indexed, with rich metadata</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accessible</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Interoperable</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Reusable</th>
<th></th>
</tr>
</thead>
</table>
The FAIR Principles

Findable
- Globally unique and persistent identifier
- Registered or indexed, with rich metadata

Accessible
- Retrievable by open, free, universal protocols
- Metadata persist even if data become unavailable

Interoperable

Reusable
The FAIR Principles

**Findable**
- Globally unique and persistent identifier
- Registered or indexed, with rich metadata

**Accessible**
- retrievable by open, free, universal protocols
- Metadata persist even if data become unavailable

**Interoperable**
- Metadata given in a standard format used across systems
- Includes explicit references to other digital objects

**Reusable**
The FAIR Principles

Findable
- Globally unique and persistent identifier
- Registered or indexed, with rich metadata

Accessible
- retrievable by open, free, universal protocols
- Metadata persist even if data become unavailable

Interoperable
- Metadata given in a standard format used across systems
- Includes explicit references to other digital objects

Reusable
- Clear and accessible data usage license
- Rich metadata to foster replication and combination

https://www.go-fair.org/fair-principles/
Preservation Formats for Data

• **Key characteristics**
  • Cross-platform, widely accessible, easy to validate
  • Provides complete data, with full detail/precision
  • Character-based ([UTF-8](https://en.wikipedia.org/wiki/UTF-8)), rather than binary (if practical, given size)

• **Preferred formats:**
  • Tabular data: [CSV](https://en.wikipedia.org/wiki/Comma-separated_values), [TSV](https://en.wikipedia.org/wiki/Tab-separated_values), [XLS/XLSX](https://en.wikipedia.org/wiki/Excel_format)
  • Text and documents: [TXT](https://en.wikipedia.org/wiki/Text_file), [PDF/UA](https://en.wikipedia.org/wiki/PDF_ua)
  • Audio: [WAV](https://en.wikipedia.org/wiki/Wave_format), [DSD](https://en.wikipedia.org/wiki/DSD_format)
  • Video: [MXF](https://en.wikipedia.org/wiki/QuickTime_Media_Example_File), [MPG](https://en.wikipedia.org/wiki/MPEG_video_format)
Preservation Formats for Data (cont.)

Acceptable formats:

- Publicly documented, adopted by communities, open/portable
- Hierarchical scientific data: H5, CDF, MAT
- Text and documents: HTML, DOCX, PDF, RTF
- Images: PSD (Photoshop), Camera Raw Formats, GIF
- Video: MP4, MKV

Many research-oriented formats are not ideal for preservation!

- Statistical software: RDATA (R), DTA (Stata), XPT (SAS), POR (SPSS)
Licensing Considerations

**Accessibility** – Open up data/code as much as possible

**Reusability** – Make it clear how others may make use of your work

**Attribution** – You may choose to oblige others to give you credit

→ **Creative Commons**: CC-BY is good; CCO is better

→ **Open Data Commons**: ODC-By is good; PDDL is better

→ **Code is different**: MIT is a good default; but lots of options
Types of Research Data Repositories

**Domain-Specific**
- Stronger communities and higher standards
- More inter-connections and richer metadata
- Examples: ICPSR; GenBank (see ACS Research Data Guidelines)

**Generalist**
- Loose communities and boilerplate standards
- Often unmediated (fast, but no one reviews submissions)
- Examples: Zenodo; DRYAD; Harvard Dataverse

**Institutional**
- Open to all disciplines, but restricted to affiliates
- Typically offer curation services and accept large files (100 GB+)
- Example: Princeton’s Data Repository
Go to Menti.com code: **9863 5615**

Where do you share your data?
Closing Thoughts
Cultivate Good Data Management

**Routinize** – Set a schedule and stick to it

**Habituate** – Lighten the cognitive load of decision making

**Automate** – Minimize the effort required for mundane tasks
Additional Resources
More Tools for File Inventory

- **DROID**
  - A free, cross-platform GUI app designed for digital archivists
  - Determines format by file signature, not just extension
  - Does checksums automatically

- **DataLad**
  - An open-source CLI tool catered to academic researchers
  - Uses Git to manage metadata, regardless of storage location
  - Great for managing dynamic datasets among collaborators

- **Pachyderm**
  - A freemium platform for data-driven pipelines
  - Tracks data versions and lineage, with containerization
  - Geared toward commercial data engineers, but applicable to academic researchers, too

Image credit: https://bit.ly/3mJvsCE
More Resources on Documentation

- PRDS Guide on Data Documentation
- README Guide from Cornell’s Research Data Management Service
- Best Practices for Data Description from DRYAD
- Open Science Framework How-To for Data Dictionaries
- USGS Guide to Data Dictionaries
Frictionless Data

• An open-source framework to reduce friction in data workflows

• Multiple standards developing for data scientists and researchers

• The key standard for us is the Data Package
  • Systematic ways to describe the structure and contents of datasets
  • Includes metadata specifications (e.g., Dublin Core)
  • Validated with JSON Schema for machine-readable variable-level descriptions
  • A basic package is a CSV with data and a JSON with robust metadata
  • Implementations in R and Python
Welcome to the Princeton Research Data Service

Established in 2019, we provide Princeton’s diverse research community with expert services and infrastructure to store, manage, retain, and curate digital research data, and to make their digital research data available to the broader network of academic researchers, as well as the general public. We provide consultations, training, and data curation services to researchers throughout the life cycle of research projects, working with them to make the process of data management and storage as seamless as possible with their current research practices.

Make an Appointment
Thank You!

Neggin Keshavarzian
Research Data Management Specialist
neggink@princeton.edu

Sarah Reiff Conell, PhD
Research Data Management Specialist
reiff.conell@princeton.edu
Attribution for these slides

Copyright (2023) The Trustees of Princeton University.

License:
CC BY 4.0 (https://creativecommons.org/licenses/by/4.0): everyone is free to copy, adapt, and reuse this document for any purpose, so long as they give attribution following the copyright notice and suggested citation.

Suggested Citation: