Hands-on data management kickstarter: a practical, hands-on overview of data management
Slide deck: https://osf.io/eguk4?view_only=2b5ed7c0339746718581ef9da77c150c

Pre-workshop survey:
https://docs.google.com/forms/d/e/1FAIpQLSeTpIup6UBdBQw_yIXXE0cRt1n5Z6v1Uk6MC4mfGizrUyvFKQ/viewform?usp=sf_link

OSF Project: https://osf.io/j7b8q/?view_only=2b5ed7c0339746718581ef9da77c150c

Agenda, notes, & questions

Introduction & Background (10 mins)

- Workshop target skills
- Reproducibility learning paths
- Terms we will use
- **POLL: What is the hardest part of onboarding someone to your research?** *(+1 the best answer for you)*
    - Pointing them to all the materials they need +1
    - Ensuring they can understand and use your materials +1 +1 +1 +1
    - Keeping track of changes to the research and to the materials: +1,
    - Finding materials and data from past contributors
    - Other
- Questions / comments? (*Write below and +1 your favorites)*
- Organization of materials (sometimes internal folders vs external)
- sharing terminology and getting people familiar with the processes specific to our institution
- Providing the appropriate amount of time to onboard (sometimes new employees takes 2-3 weeks to get onboarded but they're also expected to learn the project during that time). Hope this made sense!

+1 Having my new lab team member be able to access all of the data files and STATA codes from prior analyses and understand what has all been done before.

Why should we think about data management? (5 mins)

- Onboarding
- Preserving access
- Data sharing requirements
- Tracking of changes
- Questions / comments? (*Write below and +1 your favorites)*
-
-

What to think about when creating a data management plan? (2 mins)

- What types of data will be produced as part of the project? (image files, csv files, etc.)
- Where will the data be organized, documented, stored
- What are the data sharing requirements, when will these outputs be shared
- What metadata will you need to create
- Questions / comments? (*Write below and +1 your favorites)*
-
-

What does data management look like in the wild? (10 mins)

- There are three different DMPs in our shared notes, each related to cardiology but for different projects with different data approaches. Some are real, some are samples
- We are going to split into 3 breakout rooms with random distribution
- Find your room number at the top of your Zoom window
- Find your room number in the shared notes for a link to a DMP
- Discuss as a group; some things to consider
    - Are the authors including the required elements that you would expect?
    - Does their data management and sharing plan effectively consider reproducibility of their study?
    - Are they considering how their data can be found, accessed, and maintained?
    - What do you really like about their approach?
    - What would you do differently?
    -
- Group 1:
    - Example from DMPonline: https://dmponline.dcc.ac.uk/plans/53671/export.pdf?export%5Bquestion_headings%5D=true
    - Observations/questions
- Discuss as a group; some things to consider
    - Are the authors including the required elements that you would expect? There is a lack of detail in the DSP, including actual instruments, validation of instruments, collection and protection of privacy, and outcome measures.
    - Does their data management and sharing plan effectively consider reproducibility of their study?
    - Are they considering how their data can be found, accessed, and maintained? No references to storage tools and links to data. Not sure how much it costs/budget
    - What do you really like about their approach? Clearly written and important topic.
    - What would you do differently? Very vague; need specifics in each category, including who did the interviews, how they will identify the specific interview, and others. Typos throughout (minor). Concerns about protection of participants

- Group 2:
    - Example at DMPTool: https://dmphub.cdlib.org/dmps/doi:10.48321/D1J31B
    - Observations/questions
        - Alejandra notes: https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan#sample-plans
        - Links to repository were available but difficult to navigate
        - Big gaps in describing the data to be collected, and where it will be stored, preserved, or shared

- Group 3:
    - Example at DMPonline: https://dmponline.dcc.ac.uk/plans/61337/export.pdf?export%5Bquestion_headings%5D=true
    - Observations/questions
        - Hitting main points of management and curation
        - Backup and security
            - Clearly identify research team and particular roles in data management, HIPAA compliance
        - Very specific de-identification of data sets specific to MRI brain scans (very sensitive information) in order to put them into a repository
            - Can be difficult especially with large data sets
            - Some of the de-identified data might even be re-identifiable within some contexts, even from general locations, age range, and gender
            - De-identification of patients with rare diseases
            - Where is the line for security boundaries?
        - Concerns from researchers do arise around ethics
            - It's an area that may need more consideration and training before allowing wide sharing of data

- Other DMP examples:
    - https://dmponline.dcc.ac.uk/public_plans
    - https://abcdstudy.org/scientists/data-sharing/
    - https://dataverse.harvard.edu/dataverse/cardiovasculardiseasesintehran

- Questions / comments? (Write below and +1 your favourites)

- **ACTIVITY: Let's discuss a data management plan that you wrote.** Does it have?
    - Good README file
    - Good data documentation
    - Well organized materials

    - Metadata
    - Questions / comments? (*Write below and +1 your favorites)*
    -

Questions to remember when creating a data management plan? (5 mins)

- What
- How
- Who
- When
- Metadata
- Questions / comments? (*Write below and +1 your favorites)*
-
-
-

Plan a home for your research in advance? (5 mins)

- Use an AHA approved data repository
- Use a generalis/specific repository : https://journals.plos.org/plosone/s/recommended-repositories
- Persistent identifiers
- Persistent access
- Preservation
- Backup
- Management of access
- Versioning
- Licensing
- Questions / comments? (*Write below and +1 your favorites)*
-
-
-

Create a central workspace for your project (5 mins)

- Project management
- Bundle your dependencies
- Collaboration
- Questions / comments? *(Write below and +1 your favorites)*
-
-
-

Adopt a file naming convention (5 mins)

- The rules don't matter; that you have rules matters
- Create names that are machine readable

- Create names that are human readable
- Questions / comments? *(Write below and +1 your favorites)*
-
-
-

Create an informative directory structure (5 mins)

- What it is
- Why it exists
- How it relates to other files
- Questions / comments? *(Write below and +1 your favorites)*
-
-
-

Activity: Look at example directories

- Group 1: https://osf.io/mpyvx
    - Observations/questions
        -
- Group 2: https://osf.io/wu37k/
    - Observations/questions
        -
- Group 3: https://osf.io/squy7/
    - Observations/questions
        - Metadata not descriptive, could be richer
        - Well organized

-

Hands-on research sharing kickstarter
Slide deck:
Shared notes:
Pre-workshop survey:
Feedback survey:

Agenda, notes, & questions

Introduction & Background (10 mins)

- Workshop target skills
- Reproducibility learning paths
- Terms we will use
- How are researchers finding materials
- What does it look like to share materials

POLL: How do you share materials with collaborators? *(+1 the best answer for you)*

    - Server
    - Hard-drive
    - Dropbox+1 +1 +1
    - Google Drive+1+1+1+1
    - Github
    - Email attachments+1 +1+1
    - Evernote
- Questions / comments? (*Write below and +1 your favorites)*
-
-

Research sharing happens on a spectrum (5 mins)

- As open as possible, as closed as necessary
- What does sharing mean?
- Questions / comments? (*Write below and +1 your favorites)*
-
-
-

Research sharing 101 (5 mins)

- What to share?
- Why Share
- How to share?
- Questions / comments? (*Write below and +1 your favorites)*

-
-

Examples of common sharing scenarios (5 mins)

- As open as possible, as closed as necessary
    - Proprietary data or methods
    - Data reuse restrictions
    - Large datasets
    - Complex analyses
    - Embargoes
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Sharing code, data, and materials (5 mins)

- Example of the American Journal of Political Science: https://dataverse.harvard.edu/dataverse/ajps
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Sharing selected data, code, and materials (5 mins)

- Sharing de-identified data
- Sharing synthetic data
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Brokered, embargoed, or other limited sharing (5 mins)

- Analysis portals
- Data brokers
- Limited access
- Sharing rich metadata
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Sharing Metadata

- Closed, restricted and embargoed deposits can still be discoverable through repositories
-
-

Benefits to sharing (5 mins)

- Prevent loss of access
- Improve onboarding
- Increase the impact
- Gain credit
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Preserve access to your research with repositories (5 mins)
https://journals.plos.org/plosone/s/recommended-repositories

- What are the advantages of repositories
- Data licenses
- Code licenses
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Specialized repositories (5 mins)

- Repositories make sharing easier through design
- Protocol repositories
- Reagent repositories
- Discipline specific repositories
- Data type repositories
- Large data repositories
- Brokered repositories
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Compare repositories using re3data (5 mins)

- Funder specified repository
- Institutionally specified data repository
- Domain or discipline-specific data repository

- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Protocol repositories share methods (5 mins)

- Protocol Exchange
- Protocols.io
- PLOS ONE
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

Code sharing repositories (5 mins)

- GitHub
- Zenodo

General purpose repositories work too (5 mins)

- DataDryad
- Figshare
- Zenodo
- Open Science Framework
- Questions / comments? (*Write below and +1 your favourites)*
-
-
-

FAIR Metadata (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable)

- Metadata Quality
- Documentation Quality
- Reuse indicators
-
-

Persistent identifiers (PIDs)

- Long lasting reference to a resource
- Goal is to solve lost/broken links to important information through technical and human failures
- Enables access to a resource even if it is moved to different location or changes ownership
- Creates opportunities for interoperability with various different infrastructure systems
- Key element in making data FAIR (**F**indable, **A**ccessible, **I**nteroperable and **R**eusable)

Persistent identifiers (PIDs)

- PIDs for people (researchers) ISNIs and ORCIDs
- PIDs for places (research organizations) include ROR and funder IDs
- PIDs for things (research outputs/inputs like grants, papers, projects, etc.) include Crossref and DataCite DOIs (**D**igital **O**bject **I**dentifiers), IGSNs (**I**nternational **G**eneric **S**ample **N**umber) and more

Activity:
Thinking about your research projects
Considering metadata and sharing practices we've discussed is your project leveraging what's available?
Are there missed opportunities?
Are there other interesting ways that they are enabling discoverability?

- Group 1: https://osf.io/mpyvx
    - Observations/questions
        -
- Group 2: https://osf.io/wu37k/
    - Observations/questions
        -
- Group 3: https://osf.io/squy7/
    - Observations/questions
        -

Where to start:

- Upload all your materials in one place
    - Utilize a file tree to organize your files
    - Maintain version control
- Customize your privacy settings
- License your research
    - Communicates who and how your project is reused
    - License picker for common licenses
    - CC-BY ensures people are prompted to give attribution
    - Licenses controlled at the component level (within the OSF)

Customize what and how you share (20 mins)

- **EXERCISES:**
    - **Get an ORCID (if you don't have one) https://orcid.org**
    - **Include relevant PIDs in you metadata (even if it's optional)**
    - **Use repositories that provide PIDs**
    - **License your research**
- Questions / comments? (*Write below and +1 your favourites)*
-

-
-

Summary and next steps (5 mins)

- Key points
- Post-workshop survey
- Next steps
- Questions / comments? *(Write below and +1 your favourites)*
-
-
-