Intro to Research Data Management

Christiana Dobrzynski and Hannah Calkins

What is Research Data Management?

Children's Hospital of Philadelphia®

Research Data Management (RDM) 101

- Research data: Information collected during the course of research processes used for analysis
- RDM: Process of organizing, annotating, preserving, and sometimes sharing research data in all its forms
- Applies to entire lifecycle of research data
- Requires ongoing iteration rather than one-time application
- Best practices/recommendations rather than prescriptions
  - Universal
  - Domain-specific

Children's Hospital of Philadelphia

Why do we care about RDM?

Children's Hospital of Philadelphia®

Regulatory Requirements

- Many funding agencies (CDC, EPA, DoD & more) require data sharing & pre-written data management plans
- NIH requires a data sharing plan for all grants over 500k
- NIH will require data sharing & management plan for **all** grants starting Jan 2023
- Many sub groups such as data consortiums or centers of interest require data sharing and management
- Many publishers require data sharing along with publication

Benefits to the Researcher

- Well managed data:
  - Is easier to quality check & validate during collection and analysis
  - Reduces knowledge drain with staff turnover
  - Reduces bottlenecks by shifting implicit knowledge to explicit knowledge
  - Maximizes ease of reuse for future analysis & publications
  - Reduces administrative burden for combining with additional datasets for expanded analysis
  - Reduces administrative burden for sharing data & collaborating with new researchers
  - Makes responding to questions or challenges arising from publication and review simpler
- Research Data Management is about getting the most out of the hard work researchers are doing and the valuable data being collected

Children's Hospital of Philadelphia

Basic Best Practices

File Organization (Project template)

- File directory structure for any CHOP system with nested folders
- Organized but flexible
- Project-based instead of person-based
- Research data + access tools + contextual files + metadata
- Use for new, in-progress, on-going, and completed work
- Recommended components
  - Standardized project name
  - Parent folders (data/, manifests/, src/), no orphan files
  - Version control
  - Naming conventions
  - READMEs
- Arcus resource: Project Template Overview; Project Template GitHub repo

Children's Hospital of Philadelphia

Project Documentation (READMEs)

- Non-proprietary “homepage” with context and guidance around files and processes in your research
- Document how team members manage, organize, and use files, as well as make transitions easier between team members and over time
- Recommended for each directory and subdirectory
  - People: roles and responsibilities
  - Overview: Description of data/files and how they fit into overall research effort
  - Dates: README creation/maintenance, research milestones, methods changes
  - Related files: Navigation tips for finding important context, information, documentation
  - Process information: Details of process, methods, settings, parameters for easy reference
  - Change log: Version control to track major changes of README over time (who, what, when)
  - Naming conventions: Record naming conventions of files and data
- Arcus resource: README Best Practices

File Naming

- Standard file naming conventions increase usability of files and data
- Spend time thinking about what aspects of your files are important *in advance* and you will have names that are useful and meaningful to your process
- Standardizing the format of file names means better ability to have
  - Informative sort order
  - Easy scripting and file manipulation
  - Validation checks for missing data files
- Arcus resource: File Naming Activity & File Naming Tip Sheet

Data Dictionaries

- Data Dictionaries document the data we are collecting/creating by describing various important facets of that data & collection process. They can help with:
  - Consistency over time
  - Onboarding
  - Protect against knowledge loss
  - Troubleshooting
- Data Dictionaries should be flexible and should capture what is important about *your* data. This may include:
  - Formulas for calculated fields
  - References to papers or other publications that provide measure definitions
  - Dates on which data began or ceased to be collected
  - Question text or field text
  - Description
  - More! Less! Whatever fits your needs
- Arcus Resource: Data Dictionaries Best Practices

Children's Hospital of Philadelphia

Process Documentation

- Documenting multi-step processes increases transparency and allows for easier sharing of responsibilities
- Processes you can document:
  - **Data Collection**
  - **Analysis**
  - **Data Transformation**
- Sometimes just undergoing the act of writing out the process documentation uncovers inconsistencies within the project team!
- Arcus Resource: Process Documentation Template

Others

- Intro to research data management
- Ontologies
- Extract, transform, load (ETL)
- REDCap survey methodology
- NIH's data management plan requirements
- Narrative descriptions of research recommendations (scope note)

Custom resource development, implementation recommendations, consultations:

Email arcus-support@email.chop.edu

Children's Hospital of Philadelphia

Questions?