

Data Management Workshop for Library Staff

David Durden, Data Services Librarian

Outcomes

1. Recognize the stages of the data lifecycle
2. Understand the 6 main parts of a data management plan (DMP)
3. Know how to ask the right questions about data
4. Apply data management practice to your daily work and projects
5. Learn how to document and describe data products

What is Data Management and Why Does it Matter?

Data Management is Information Management

"...university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management – a discipline that encompasses the entire life cycle of data, from how they are acquired and stored to how they are organized, retrieved and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge."

What is Research Data Management?

1. Umbrella term for **describing, planning, organizing, documenting, storing, and sharing** data
2. Emphasis on data **protection, security, and confidentiality** issues
3. **Framework** for supporting researchers and their data during and beyond a research project

Levels of Data Management

1. Administrative -- policy

- Data Use Agreements
- Data Management Plans
- Data Safety and Monitoring Plans

2. Applied -- procedural

- Workflows
- data management plans
- Documentation

Why Does Data Management Matter?

1. Fulfills **funder, university, or industry** requirements and standards
2. Ensures data are **accurate, complete, authentic, and reliable**; good research practice
3. Data security and **minimized risk of loss**
4. Increases efficiency; saves time and resources
5. Data remain accessible for future use; re-use/aggregation, defense against claims of data alteration, flawed methodology, etc.

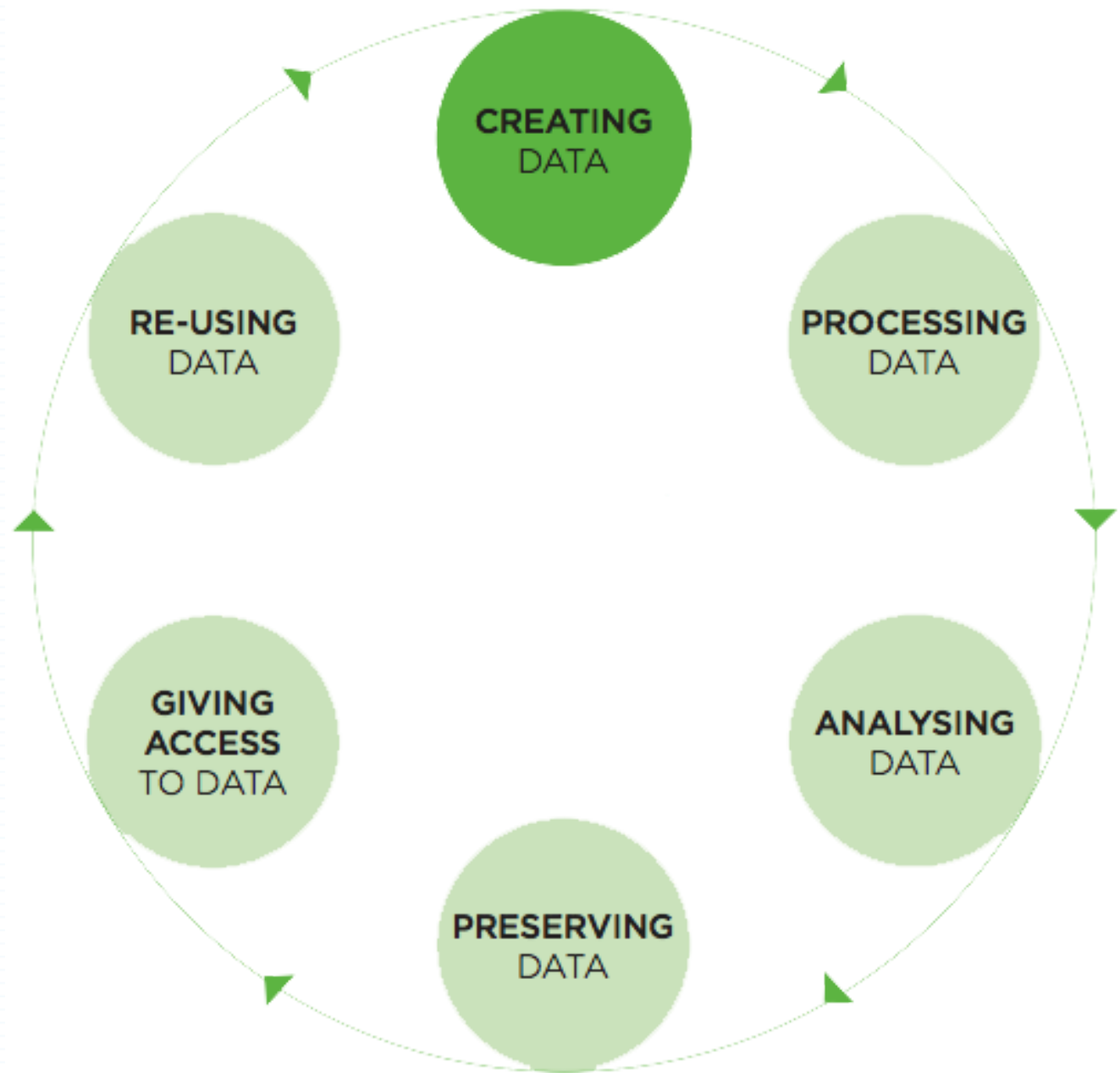
The Funding Agency Perspective

1. Research = Time and Money
2. Data are considered a public good
3. Review for standards and best practices
4. Articulate levels of discoverability (metadata)
5. Stipulate the levels of access (embargo, limited data set, etc.)
6. Explicit inclusion of data management activities in the funding request

The Data Lifecycle

? What Counts as Research Data?

Data that are used as *primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity*, and that are used as *evidence in the research process* and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. *Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data.*



1. **Creating data:** research design, data management planning, data location/collection, create basic metadata
2. **Processing data:** digitize, enter, transcribe, code, translate, etc., anonymize, describe data, active management
3. **Analyze data:** interpret data, production research outputs/publications
4. **Preserving data:** format standardization/normalization, create robust metadata and documentation, archive data
5. **Access to data:** Distribute/share, establish access control, copyright/license
6. **Re-use of data:** follow-up research, new research, scrutinize/review research, teach/learn

Group Discussion

See handout

Group Discussion

Q1: How would this lifecycle apply to your work?

Q2: What stages do not make sense to your work?

The Data Management Plan

"When things go wrong, they do so in the manner that yields the most difficulty."
- The Principle of Maximum Inconvenience



Institutions That Often Require DMPs

- The Sloan Foundation
- Institute of Museum and Library Services (IMLS)
- Institute of Education Sciences (IES)
- DoD, DoE
- U.S. Geological Survey
- NASA, NOAA
- USDA
- NSF, NIH



DMPs Often Require The Following Descriptions:

1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. the standards to be used for data and metadata format and content;
3. the policies for access and sharing, provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for re-use, redistribution, and the production of derivatives; and
5. plans for archiving data, samples, and other research products, and for preservation of access to them.

The Basics of a DMP

A good DMP answers the questions:

- who,
- what,
- where,
- when,
- why,
- and how

Data Management Plans: Section by Section

A General Description of Your Data

- how big will the data be?
- how fast will the data grow?
- what are the likely file formats for the data?
- how unique is the data?
- what is the source of the data?
- who owns the data?

Organization and Documentation

- how will you document your data?
- which metadata schema will you use?
- how are your data organized
- how are individual files named?

Storage and Backup

- how and where are the data stored?
- what is your backup schedule?
- who manages storage and backups?
- cost?

Security

- what security measures will you use?
- who has access to data and data stores?
- who manages the security systems?
- where is PII stored?

Security and Retention

What security policies apply to your data?

- UMD (e.g., data retention)
- IRB (human subjects and ethics)
- HIPAA (digital health records)
- FERPA (student and educational records)
- ITAR (DoD and other federal organizations)
- other ethical/legal concerns (at-risk populations, protected species, trade secrets, etc.)

Post-project Data Management

- how will you store and backup your older data?
- how long will you maintain data after the project is completed (e.g., institutional data retention policies, state or federal policies)?
- will you migrate your storage media over time?
- how will you prepare data for long-term storage/preservation?
- will you use a third party to preserve your data? e.g., a data repository
- what happens to data if you leave your current institution?

Data Sharing and Access

One of the most important sections for publicly funded research.

- what data will be shared and in what forms?
- who is the audience for your data?
- when and where will you share?
- how will the data be prepared for sharing?
- who is responsible for making the data available and/or answering question about access?
- how much will this cost?



DMP Resources

1. UMD Libraries Guide to Writing DMPs (<https://www.lib.umd.edu/data/dmp-guide>)
2. Univ. of California Curation Center's DMPTool (<https://dmptool.org>)
3. UK's Digital Curation Centre's DMPonline (<https://dmponline.dcc.ac.uk/>)

Thinking Data Management

Data Organization

- plan storage/organization before you start generating data/creating files
- organize systematically so that hierarchies are easily identifiable and understandable
- establish a directory structure and file naming convention
- use file versioning or version control

Folder Structures

Levels/Hierarchies

- project
- researcher/collaborator
- date
- sample number
- experiment type/instrument
- data type

File Naming Conventions

The most granular level of metadata.

Use a combination of:

- experiment type/number
- researcher name
- sample type/number
- analysis type
- date
- site/lab/location
- version number

File Naming Conventions

*There is **no** absolute way to name your files. However...*

Names should be:

- descriptive
- unique
- consistent
- relatively short



Some File Naming Tips

Avoid:

- whitespace
- special characters / \ : () * ? ' < > [] & \$ @ ! " % ^
- localized date formats
- using commas or periods in file names

Do:

- use dashes or underscores (CamelCase or pot_hole_case)
- use the ISO 8601 date convention (YYYYMMDD or YYYY-MM-DD)

Group Discussion

See handout

Group Discussion

Q1: How would you improve these names to become more specific?

Q2: What are some example naming conventions make sense to you?

Backup and Storage Strategies

The level of redundancy in your storage media should reflect the value of your data.

but...

"Lots of copies keep stuff safe."

Backup and Storage Strategies

The "3-2-1" backup rule:

1. At least **three** copies,
2. in **two** different formats,
3. with **one** of those copies off-site.

REMEMBER: *the cloud is just someone else's computer...*

Hardware	Rating	Notes
Personal computer	Good	Good when used with other storage
External hard drive	Good	Good when used with other storage
Local server/drive	Good	Good when used with other storage
Magnetic tape	Good	Good when used with other storage
CD/DVD	Acceptable	Cumbersome to use/becoming obsolete
Cloud storage	Depends on product	Read the Terms of Service
USB flash drive	Do not use	Use only for file transfer
Obsolete media	Do not use	Remove data as soon as possible

Backup and Storage Strategies

- establish a backup strategy for your data and project materials
- test your backup strategy
 - test that your backups work
 - prove that you can recover lost work
- scan and/or copy analogue data (notebooks, research notes, surveys, etc.)
- for institutional resources (DIT, Library, Departmental), determine who manages data recovery and understand the timeline for restoration

Group Discussion

See handout

Documenting Data



Reasons for Documenting Data

- Remember details of what you did
- Aid others in understanding your work
- Verify and/or replicate your work
- Capture metadata as you work

Examples of Data Documentation

- Lab notebooks
- Field notes
- Questionnaires
- Standard Operating Procedures (SOPs)
- Methodologies used
- Code books/Data dictionaries

Levels of Documentation

Project

- study background, methodologies, instruments, hypothesis

File/Database

- formats, relationships between files

Variable/Item

- variable generation, label descriptions, measure descriptions

Metadata vs. Documentation

- Metadata is created for computers to read
- Documentation is created for people to read

Levels of Metadata


- *Descriptive*: title, author, keywords, etc. -- used for searching/browsing
- *Administrative*: preservation activities, rights management, technical/format description
- *Structural*: the relationship between related resources -- raw data & scripts, database ERD, load order, etc.

This Workshop was Developed Using the Following Sources:

1. EDINA and Data Library, University of Edinburgh. DIY research data MANTRA training kit for librarians. <http://datalib.edina.ac/mantra/libtraining.html>.
2. Briney, K. (2015). Data management for researchers: Maintain and share your data for research success. Exeter: Pelagic Publishing, UK.

 **Thanks!**

 durden@umd.edu

 lib-research-data@umd.edu

 <https://www.lib.umd.edu/data>



This work is licensed under a [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/).

