An Introduction to Data Management

David Durden, Data Services Librarian

Universiy of Maryland Libraries

February 25, 2019

👍 **Get the Slides and Follow Along!**

https://go.umd.edu/190225-dm

🔍 The Data Problem

🔍 Where Does All the Data Go?

> ...for every yearly increase in article age, the odds of the data set being extant decreased by 17%.

4

❓ What is Data Management

Can be defined as: *The **process** of documenting, organzing, and maintaining the processes used in the information/data lifecycle.*

- Storage & backup procedures

- Script/code documentation

- Provenance & citation information

- Quality assurance & security protocols

- Licensing agreements

🏛️ Motivations for Data Management

"*When things go wrong, they do so in the manner that yields the most difficulty.*"

   – The Principle of Maximum Inconvenience

🏛 Motivations for Data Management

1. Funding & money

2. Ethics & transparency

3. Reproducibility & accountability

4. Re-use

5. Good organizational practice

6. For the public good!

🔀 Approaches to Data Management

1. Administrative (policy)

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

2. Applied (procedural)

- Workflows
- data management plans
- Documentation

📊 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.*

♻ DataONE Lifecycle Model
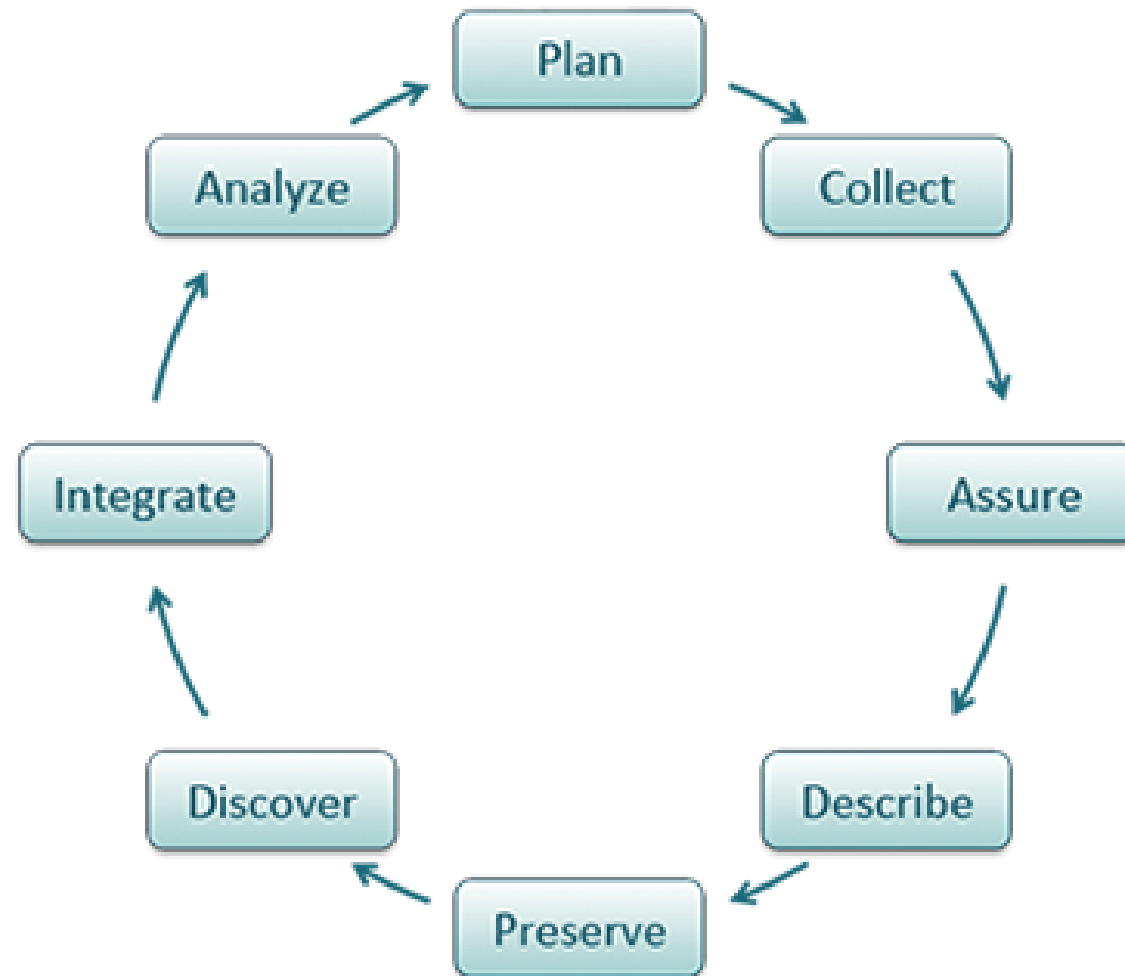
♻️ DataONE Lifecycle Model

1. *Plan:* description of the data that will be compiled, and how the data will be managed and made accessible throughout its lifetime

2. *Collect:* observations are made either by hand or with sensors or other instruments and the data are placed a into digital form

3. *Assure:* the quality of the data are assured through checks and inspections

4. *Describe:* data are accurately and thoroughly described using the appropriate metadata standards

♻️ DataONE Lifecycle Model

5. *Preserve:* data are submitted to an appropriate long-term archive (i.e. data center)

6. *Discover:* potentially useful data are located and obtained, along with the relevant information about the data (metadata)

7. *Integrate:* data from disparate sources are combined to form one homogeneous set of data that can be readily analyzed

8. *Analyze:* data are analyzed

📄 The Data Management Plan

💡 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
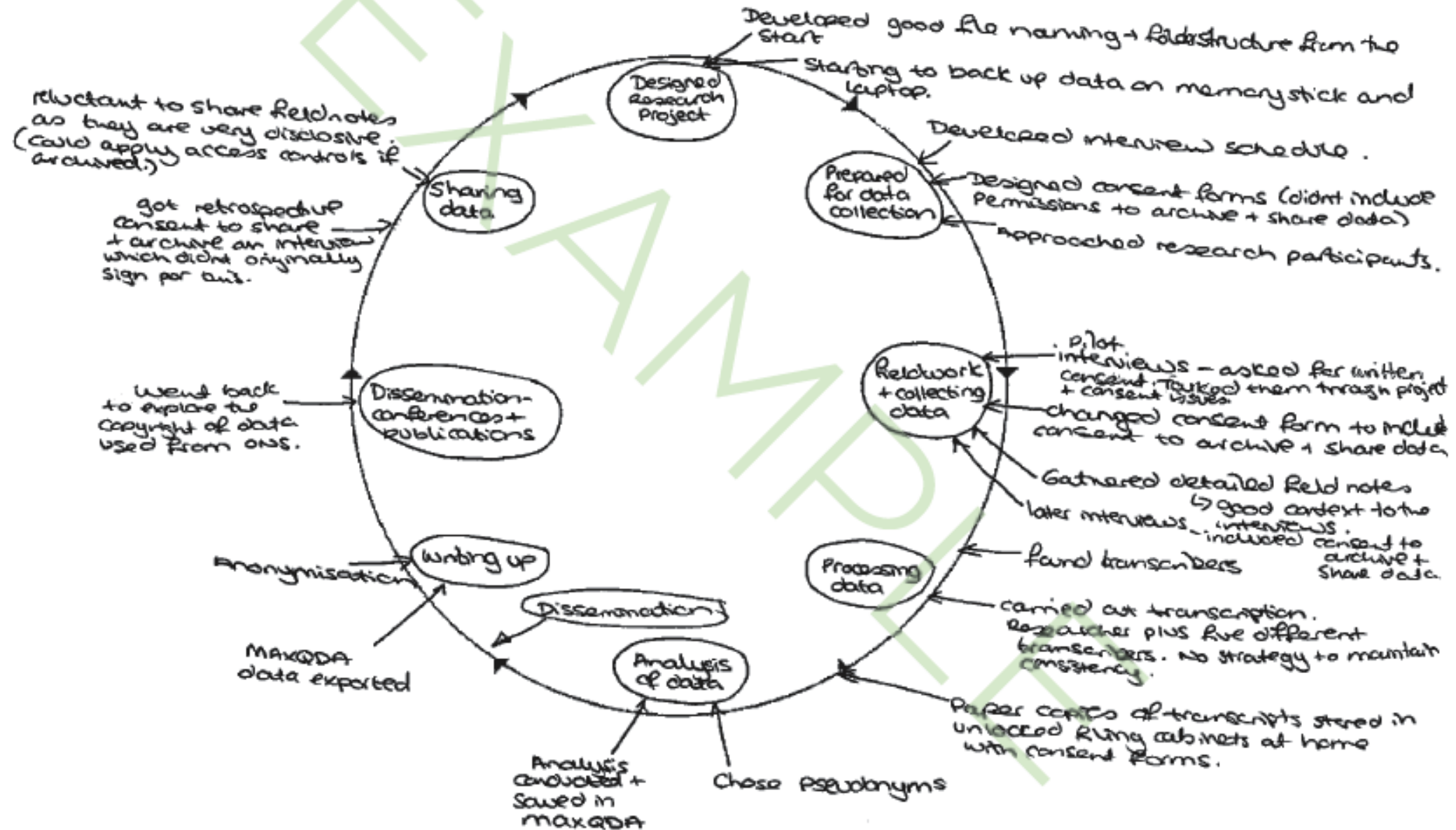
- NASA, NOAA

- USDA

- NSF, NIH

⚙️ The Basics of a DMP

A good DMP answers these questions:

- Who

- What

- Where

- When

- Why

- How

📝 **Exercise: Planning your Research Using a Data Management Lifecycle**

Developed good file naming + folder structure from the start

Starting to back up data on memory stick and laptop.

Developed interview schedule.

Designed Research Project

Prepared for data collection

Designed consent forms (didn't include permissions to archive + share data)

Approached research participants.

reluctant to share fieldnotes as they are very disclosive. (could apply access controls if archived)

Sharing data

got retrospective consent to share + archive an interview which didn't originally sign for this.

Dissemination - conferences + publications

went back to explore the copyright of data used from ONS.

Fieldwork + collecting data

· pilot interviews – asked for written consent. Tracked them through project + consent issues.

changed consent form to include consent to archive + share data

Gathered detailed field notes ↳ good context to the later interviews. interviews included consent to archive + share data.

found transcribers

Processing data

carried out transcription. Researcher plus five different transcribers. No strategy to maintain consistency.

Paper copies of transcripts stored in unlocked filing cabinets at home with consent forms.

anonymisation

Writing up

MAXQDA data exported

Dissemination

Analysis of data

Analysis conducted + Saved in maxQDA

Chose pseudonyms

🔧 Data Management Resources

Data Management for Researchers
is a step-by-step guide designed for academics who need to take care of their data.
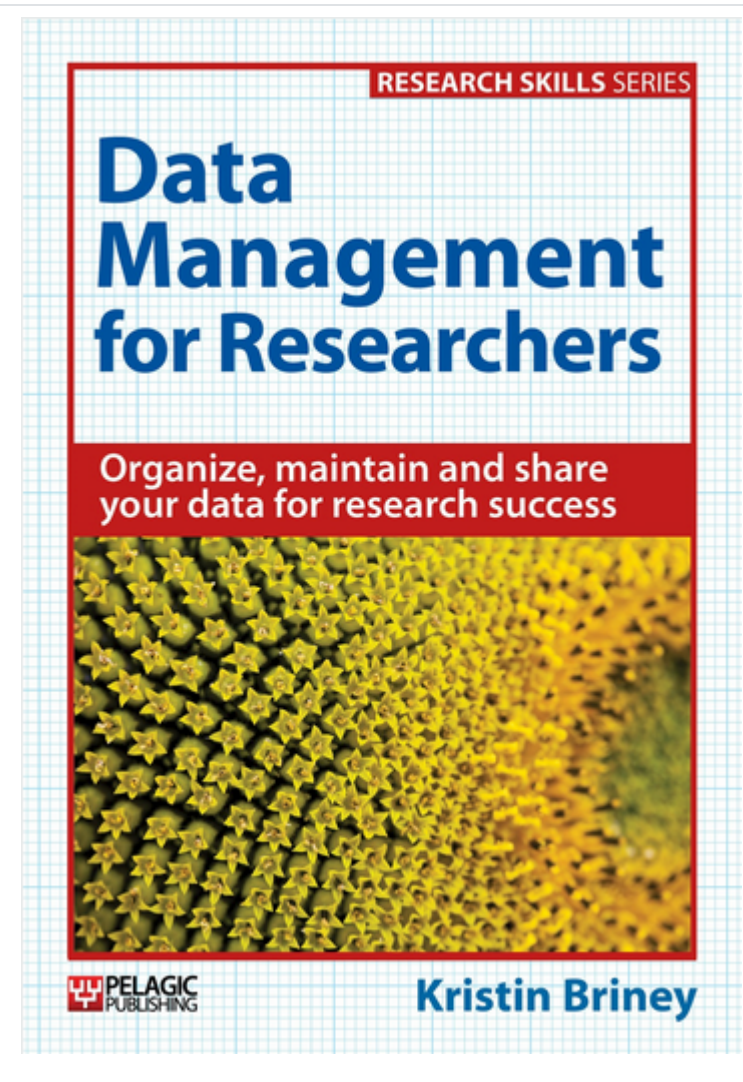
📝 Write your DMP with DMPTool

Build your Data Management Plan

Login at https://dmptool.org

💻 Use Version Control (e.g. Git)

- See Fred Hutch's blog post on Git and GitHub for Computational Research

- Use the Git Tutorials from Software Carpentry to practice using Git

💿 Backup and Storage Strategies

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

- 💡 *Lots of copies keep stuff safe.*

- *The cloud is just someone else's computer...* (Always read the terms of service for any cloud storage solution.)

- The 3-2-1 Rule:

  i. At least **three** copies

  ii. in **two** different formats

  iii. with **one** of those copies off-site

📄 Think Security

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.)

📄 File Naming Conventions

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

Names should be:	Avoid:
Descriptive	Spaces (use **CamelCase** or **pot_hole_case** instead)
Unique	Special characters (/ \:*?'< >: &$@!%^)
Consistent	Localized data formats – Use ISO 8601 (YYYYMMDD or YYYY-MM-DD)
Relatively short	Using commas or periods in file names

📗 Take an Online Data Management Training

- https://dataone.org/education

- https://mantra.edina.ac.uk/libtraining.html

💡 **Data Management Wrap-up**

💡 Data Management Wrap-up

- Analyze your workflows

- Create systems that work for you, and stick to them
    - When they don't, change them

- Organize and describe your data and processes in way that is understandable and repeatable

- Don't be afraid to change or make improvements

💡 Data Management Wrap-up

- Use a DMP or data lifecylce as part of project planning

- Omit DMP sections that won't apply

- Don't reinvent the wheel – use existing policies and/or DMPs

- Always look out for '*future you*'

💾 Thanks!

💻 https://lib.umd.edu/data

✉ lib-research-data@umd.edu

✉ durden@umd.edu