

# Managing Data From Your Research

David Durden

Data Services Librarian

# The Data Problem

*"If my calculations are correct... you're gunna see some serious s\*\*\*."*

- Dr. Emmett Brown

## Where Does All the Data Go?

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

## Why Data Management?

1. *money.*
2. ethics & transparency
3. reproducibility & accountability
4. re-use
5. good organizational practice
6. for the public good



What Qualifies as "Research Data"?



## Research Data Are:

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.*



## According to NIH:

The definition of digital scientific data includes data that are used to support a scientific publication as well as data from completed studies that might never be published.



## According to NSF:

... may include, but is not limited to: data, publications, samples, physical collections, software and models.



## Exercise 1

*What do you think qualifies as research data?*

On a piece of paper, write down 5-10 items that you think should/could be classified as data.



## What is "Research Data"?

- drafts
- bibliographies
- oral histories & A/V materials
- code/scripts
- survey instruments
- "data"
- description of methodologies
- etc.

# Data Management Terminology

*"And then the world is your mollusc!"*

*"My mollusc! What's a mollusc?"*

- Terry Pratchett

## Terminology - Preservation

- *archive* - a facility that appraises, preserves, and maintains access to materials on a long-term or permanent basis; also used to mean the transfer of materials to such a facility
- *institutional repository* - a service for storing and providing access to digital materials; *Digital Repository at the University of Maryland (DRUM)* is your local insitutional repository
- *preserve* - activities that ensures access, completeness, and integrity of information; in the case of data, this can include emulation, format migration, and creating interoperable metadata

## Terminology - Legal

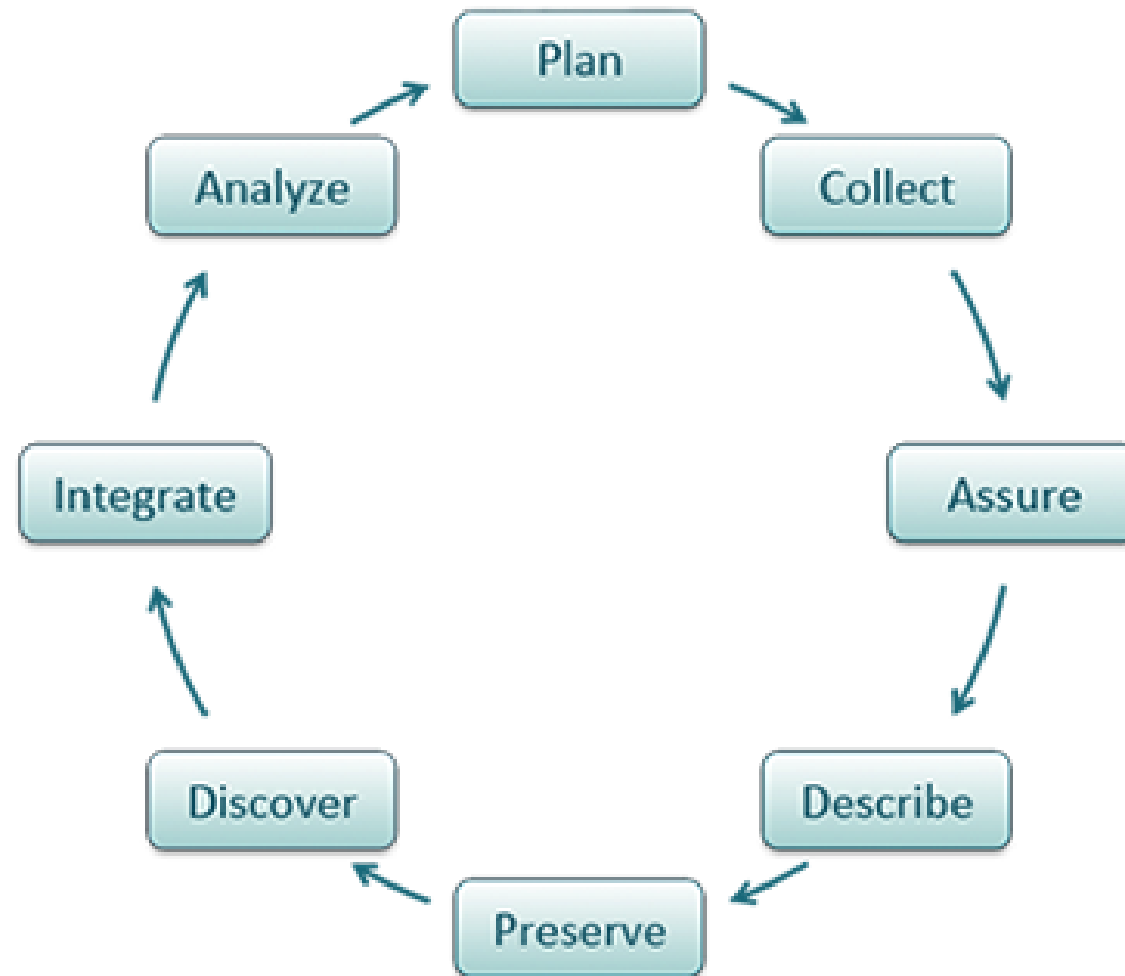
- *intellectual property* - rights applied to creative works; data are often considered *statements of fact* and are not copyrightable; derivatives, such as databases, analytic scripts, aggregate data, code, papers, etc. are copyrightable
- *open access* - most commonly applied to publications, materials are accessible free of license agreements or copyright restrictions, *or subscription fees...*
- *use statement* - the conditions under which materials may be used, often a license or contract; for sensitive data, a data use agreement often used

## Terminology - Technical

- *file format* - data encoding scheme used to make digital information readable; .txt, .csv, .tif, .mp3, etc.
- *identifier* - a reference that uniquely and permanently points to a single object; conceptually follows levels of abstraction - within a single project, identifiers can be file names, folders, etc.; in a broader sense, DOIs, Handles, URIs, URLs, etc.
- *metadata* - (literally "data about data") descriptive elements arranged in a structured way, often using a standard (e.g., ISO, NIST, etc.); example schema include Dublin Core, Darwin Core, DDI, EML, MIBBI, etc.)

# The Data Lifecycle

## DataONE Model





## DataONE 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 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 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

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

- The Principle of Maximum Inconvenience

## NSF DMP Requirements:

Proposals submitted... must include a supplementary document of no more than two pages labeled "Data Management Plan." This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.

*Note:* NSF DMP requirements vary by directorate. In fact, DMP requirements are hardly standardized across institutions and agencies.

## The Basics of a DMP

Describe:

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

Remember:

*A good DMP answers the who, what, where, when, why, and how about your data.*

## DMPs - 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?
- what metadata schema will you use?
- how are your data organized and 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

*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?

## Post-project Data Management

- 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?

## Data Sharing and Access

- 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?





## How Much Will All This Cost?

- always budget for data management
- time is money; not all costs are measured in dollars
- be realistic
- seek out free or low cost resources

## Organization

## Folder Structures

### Levels/Hierarchies

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

## Folder Structures - Example

Project\_1

    Data

        AnalyzedData

        RawData

            20171201

            20171215

                Mterrapin-observ-1

    DataManagementPlan

    Documentation

    Draft

    LitReview

    Scripts

## 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

## Exercise 2

Read through the following file names and try to figure out what each file contains. How could you improve these file names?

 Doc. 1	 My data
 IMPORTANT	 My Passwords
 Thesis Final final	 Thesis version 12
 My study	 Data chart for interviews
 Interview with Jane	 Int 1 (2)
 Interview with Janet	 My thesis (copy)
 Int. 1	 New doc.

## File Naming Conventions

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

Names should be:

- descriptive
- unique
- consistent
- relatively short

## Some File Naming Standards

Avoid:

- spaces - use dashes or underscores (use CamelCase or pot\_hole\_case)
- special characters / \ : ( ) \* ? ' < > [ ] & \$ @ ! " % ^
- localized date formats - use the ISO 8601 date convention (YYYYMMDD or YYYY-MM-DD)
- using commas or periods in file names



## 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.



## Exercise 3

Using the 3-2-1 rule and the following list of potential storage media, design an appropriate storage and backup strategy. Assume you have access to all of them.

- personal computer
- external hard drive
- local server or network drive
- cloud services
- CD/DVD or flash drive
- LTO/tape

## Backup and Storage Strategies

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

Always read the terms of service for any cloud storage solution.

## Backup and Storage Strategies

- automate your backups so that you don't forget to do them
- determine a backup frequency that matches the time and value of your data
  - daily
  - weekly
  - bi-weekly
- incremental vs. full backups

## Backup and Storage Strategies

- 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, etc.)
- for institutional resources (DivIT, Library, Department), determine who manages data recovery and understand the timeline for restoration

## Data Management Wrap-up

*"You have lost everything you value."*

- Lt. Cmdr. Data

## Data Management Wrap-up

*Conceptually easy, realistically... something else.*

- analyze your workflows
- create systems that work for you, and stick to 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

- create a data management plan even when you are not required
- omit DMP sections that won't apply
- don't reinvent the wheel - use existing policies and DMPs
- look out for 'future you'

 Thanks!

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

 [lib-research-data@umd.edu](mailto:lib-research-data@umd.edu)



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

