Research Data Management for the Health Sciences

Summer 2019

Sawyer Newman
Data Librarian for the Health Sciences

Yale *Harvey Cushing / John Hay Whitney Medical Library*

Objectives

- Define data and research data management
- Highlight the importance of research data management
- Identify best practices, strategies, and software for data management
- Identify research data support providers at Yale

What is research data?

Recorded factual material commonly accepted in the scientific community as necessary to document and support research findings.

This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based.

NIH Data Sharing Policy and Implementation Guidance

Research data includes

Raw data and derived variables

- Lab notebook content
- Imaging outputs
- Mined text
- Computerized datasets

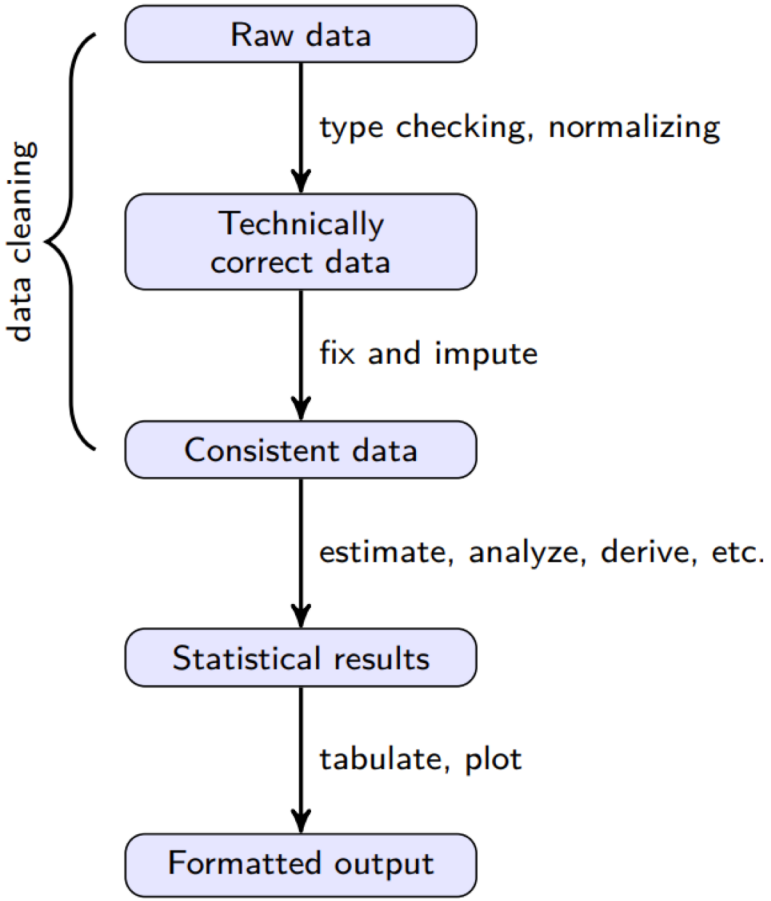
Code

- Data cleaning
- Data transformation
- Data analysis/statistics

Research data does not include

- Statistics
- Summary tables
- Graphs or charts
- Physical objects
- Books
- Plans for future research

Stages of research data



Raw vs summary data

uid	date_coll	species	color	type
1	2018-12-01	Canis lupus familiaris	Brown	Terrestrial
2	2019-12-20	Betta Splendens	Red	Aquatic
3	2019-12-20	Neritina natalensis	Brown	Aquatic
4	2019-01-08	Felis catus	Orange	Terrestrial
5	2019-01-09	Sciurus carolinesis	Gray	Terrestrial
6	2019-12-25	Larus argentatus	Brown	Terrestrial

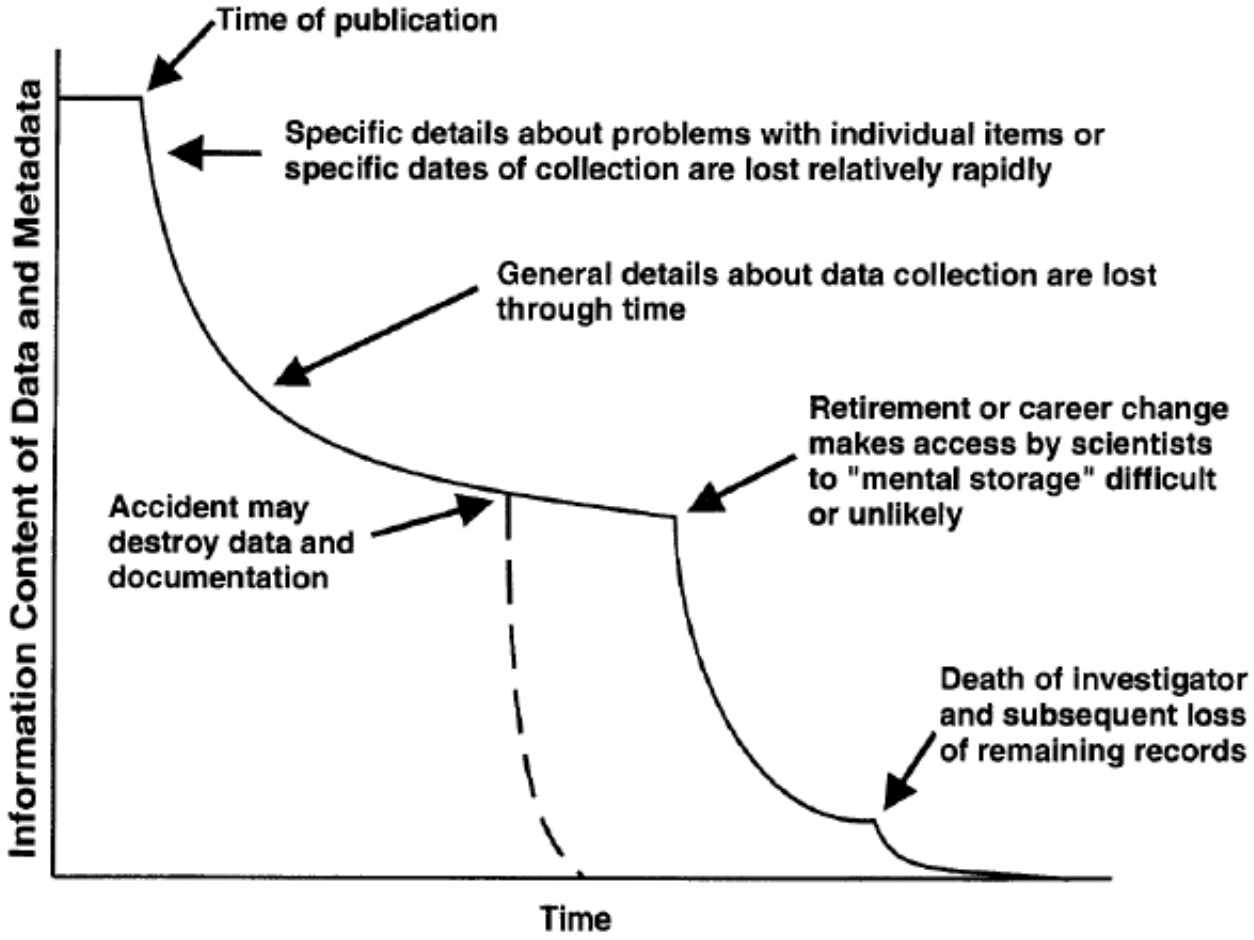
color	count
Brown	3
Red	1
Orange	1
Gray	1

type	percentage
Terrestrial	0.6666667
Aquatic	0.3333333

What is data management?

- Measures taken to ensure data is findable, accessible, interoperable and reusable
- Takes place throughout the course of the research data lifecycle
- Is performed by one or more people involved in a research project

RDM & Research Data Entropy



Benefits of managing research data

- Verify the integrity of your data
- Make your data findable and reusable
- Help others understand your data
- Encourage other researchers to reuse and cite your data
- It is required by funding agencies such as NIH and NSF

What does data management entail?

- File backups
- File names
- File and folder organization
- Documentation and metadata
- File format considerations
- Data security
- Long term data storage

Data management starts when your research project starts

How can you plan ahead?

- Data management roles
- Data types, formats and quantities
- Data and metadata standards
- Data access and sharing policies
- Hardware required for data storage
- Data sharing through repository deposits

Data management/sharing plans

- Written statements provided to an agency during the application process for research funding
- DMPs can be "living documents" that are updated as a project changes

File and folder naming practices

- Create unique and simple files names
- Use only alpha-numeric characters. Avoid using special characters such as: ? / $ % & ^ # . \ : < >
- Use underscores (_) and dashes (-) to represent spaces
- Use leading zeros with the numbers 0-9 to facilitate proper sorting
- Dates should follow the ISO 8601 standard of YYYY_MM_DD or YYYYMMDD

Data management software options

Electronic lab notebooks

- Rspace
- LabArchives

Survey capture and management tools

- REDCap
- Qualtrics

Database management systems

- Relational database management systems (SQL databases)
- NoSQL databases (Not Only SQL)
- Microsoft access

Data management software options

Data Analysis

- R, Python, SPSS, STATA, SAS
- NVivo
- Bioinformatics support through CWML

Data Storage

- Box Secure
- Dropbox
- Storage @ Yale
- Spinup server

Data Repositories

- NIH repositories
- Disciplinary repositories
- Generalist repositories

Documentation

What information should you capture about your data?

- The data creator
- Data file continents
- Data creation times
- Data creation locations
- Reasons why the data were created
- Methods used to generate the data
- Units of measurement
- Instruments used

Version control

- Version control allows you to see what changes you have made to a file over time and allows you to restore old versions of a file
- Document:
  - What changed?
  - Who changed it?
  - Why was the change made?
  - When was the change made
- File name example: 2019-01-09_workingData_v01.doc

Data backups

LOCKS - lots of copies, keep stuff safe

- Store in multiple physical locations
  - Local machine
  - Cloud
  - External hard drives
- Maintain version control
- Yale Systems
  - Storage @ Yale
  - CrashPlan → Desktop Backup (free to the user)

Data security

How can I securely share working data?

- Yale's Secure File Transfer service
- Secure Box at Yale

Which software are compliant with high risk data?

- https://your.yale.edu/technology/data-security/protect-your-data#approved

End of project data management

- Choose a long term storage location for data
- Maintain compliance with data sharing agreements
- Ensure your data is findable through your research publication (DOI)

Other data support groups at Yale

- Core Research Facilities
- Joint Data Analytics Team
- StatLab
- Yale Center for Analytical Sciences
- Yale Center for Clinical Investigation
- Yale Center for Genome Analysis
- Yale Center for Research Computing

How can the CWML help?

Consultations, custom workshops, instruction sessions

- Check our calendar for classes
  - Intro to R with Swirl
  - Intro to Git and GitHub
  - Data Management and the Unix Shell
- View our online resources

Research Data Services @ CWML

Data Management Plans

Data Tools & Software

Data Policy Guidance

Find Datasets

Data Storage

Best Practices & Definitions

Data Support Groups at Yale

Consultations & Drop-Ins

medicaldata@yale.edu