DATA MANAGEMENT

Presenters: Nell Aronoff & Rachel Starry
November 2, 2018
Learning objectives

1. Recognize what research data is and what data management entails
2. Recognize why managing data is important for your research career
3. Identify common data management issues
4. Learn best practices and resources for managing these issues

Objectives are from NECDMC Module 1 Overview of Research Data Management
https://library.umassmed.edu/resources/necdmc/modules
What are research data?
Three Types of Data

• Research Data **Objects**
  - primary data that can take many different forms

• Research Data **Products**
  - built on primary data and important for secondary analyses

• Research Data **Documentation**
  - metadata, codebooks, and information about methodologies
Research Data Objects

- Documents and spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artifacts, specimens, samples

- Project files
- Digital objects created during research
- Statistical or other data files
- Database contents
- Models, algorithms, scripts
- Software

Adapted from: MANTRA https://mantra.edina.ac.uk/researchdataexplained/ Slide 12
Research Data Products

- Grant applications
- Ethics applications
- Technical reports
- Technical appendices
- Research reports
- Research publications
- Models, algorithms, scripts
- Data files

Adapted from: MANTRA https://mantra.edina.ac.uk/researchdataexplained/ Slide 13
Research Data Documentation

- Correspondence (electronic and paper)
- Social media communications (blogs, wikis, tweets etc.)
- Signed consent forms
- Methodologies and workflows
- Standard operating procedures and protocols
- Questionnaires, transcripts, codebooks

Adapted from: MANTRA https://mantra.edina.ac.uk/researchdataexplained/ Slide 13
Activity #1: What Does Your Data Look Like?

Take a few minutes to write down your responses to the following questions.

Questions:
1. What types of research data objects do you deal with in your own research?

2. How might thinking about research products as data change how you manage your own research outputs?

3. What are some of the ways you currently document your data? How do you document your research processes and methodologies?
What is research data management?
Definitions

• “Data management refers to all aspects of creating, housing, delivering, maintaining, and archiving and preserving data. It is one of the essential areas of responsible conduct of research.” (MANTRA [https://mantra.edina.ac.uk/datamanagementplans/](https://mantra.edina.ac.uk/datamanagementplans/) Slide 5)

• Data Management “consists of two major activities conducted in coordination: data management services and data stewardship. They constitute a comprehensive end-to-end process including movement of data and information from the observing system sensors to the data user. This process includes the acquisition, quality control, metadata cataloging, validation, reprocessing, storage, retrieval, dissemination, and archival of data.” (NOAA's Administrative Order 212-15)
Lynn Jamieson, Professor of Sociology
The University of Edinburgh

MANTRA - Professor Lynn Jamieson - Importance of data management

https://www.youtube.com/watch?v=YQNadL5t8hg
Benefits of Good Data Management

- Meeting funder **requirements**
- Ensuring research **integrity** and **reproducibility**
- Increasing your research **efficiency**
- Ensuring data and records are **accurate, complete, authentic**, and **reliable**
- Saving time and **resources**
- Enhancing data **security** and minimizing risk of data loss
- Preventing duplication of effort by enabling others to **re-use** your data
- **Complying with practices** in industry and commerce
Why Manage Data: The Researcher’s Perspective

- Keep yourself **organized** – be able to find your files
- Track research process for **reproducibility**
- Better control **versions** of data
- Better control the **quality** of data
- Avoid **data loss**
- Format your data for **re-use**
- To **document** your data for accountability and verification
- Gain **credibility** and **recognition** by sharing data

DataONE [https://www.dataone.org/education-modules](https://www.dataone.org/education-modules)
Why Data Management, Slides 15-16
Data Management and Sharing Snafu

https://www.youtube.com/watch?v=N2zK3sAtr-4
“...because we do not have access to the original data, we cannot assure you that the results of the studies are valid.”

https://jamanetwork.com/journals/jama/fullarticle/2703449
Activity #2: What Would You Do?

Read the scenario below and answer the following questions. What do you think you might do in such a case?

**Scenario**: You have completed your postgraduate study with flying colors and published a couple of papers to disseminate your research results. Your papers have been cited widely in the research literature by others who have built upon your findings. However, three years later a researcher has accused you of having falsified the data.

**Questions**:
1. What evidence would you need to prove that you didn’t falsify the data?
2. What could/should you have done throughout your study in order to document your research methods and show the integrity of your data?

MANTRA [https://mantra.edina.ac.uk/datamanagementplans/](https://mantra.edina.ac.uk/datamanagementplans/) Slide 2
The Research Data Lifecycle

Create → Document → Use → Share → Preserve → Create

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
The Research Data Lifecycle: Create

Create → Document → Use → Share → Preserve → Create

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Data Creation

- Decide **what** data will be created and **how**
- Develop **procedures** for consistency and data quality
- Choose **appropriate software and formats**
- If necessary, ensure consent forms, licenses, and partnership agreements **don’t limit options** to share data

DCC [http://www.dcc.ac.uk/training/rdm-librarians](http://www.dcc.ac.uk/training/rdm-librarians)
Why develop a Data Management Plan (DMP)?

DMPs can help researchers to:
- Make informed decisions to anticipate & avoid problems
- Avoid duplication, data loss and security breaches
- Develop procedures early on for consistency
- Ensure data are accurate, complete, reliable and secure

DCC [http://www.dcc.ac.uk/training/rdm-librarians](http://www.dcc.ac.uk/training/rdm-librarians)
Components of a DMP

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Roles and responsibilities
6. Budget
DCC Checklist for a DMP

- What data will you collect or create?
- How will the data be collected or created?
- What documentation and metadata will accompany the data?
- How will you manage any ethical issues?
- How will you manage copyright and intellectual property rights issues?
- How will the data be stored and backed up during research?
- How will you manage access and security?
- Which data should be retained, shared, and/or preserved?
- What is the long-term preservation plan for the dataset?
- How will you share the data?
- Are any restrictions on data sharing required?
- Who will be responsible for data management?
- What resources will you require to implement your plan?

What is Metadata?

Metadata is “Data Reporting”

WHO created the data?
WHAT is the content of the data?
WHEN were the data created?
WHERE is it geographically?
HOW were the data developed?
WHY were the data developed?
Metadata: Why Does it Matter?

Data is not self-describing. Metadata, or “data about data” explains your dataset and allows you to document important information for:

- **Finding** the data later
- **Understanding** what the data is later
- **Sharing** the data (both with collaborators and future secondary data users)

Consider it an **investment** of time that will save you trouble later several-fold

MIT Libraries [https://www.dropbox.com/s/7k4qt43phcsohhg/rdm101_Slides_MIT.pdf?dl=0](https://www.dropbox.com/s/7k4qt43phcsohhg/rdm101_Slides_MIT.pdf?dl=0)

Research Data Management 101, Slide 16
## Concerns About Creating Metadata

<table>
<thead>
<tr>
<th>Concern</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>workload required to capture accurate robust metadata</td>
<td>incorporate metadata creation into data development process (distribute the effort)</td>
</tr>
<tr>
<td>time and resources to create, manage, and maintain metadata</td>
<td>include in grant budget and schedule</td>
</tr>
<tr>
<td>readability / usability of metadata</td>
<td>use a standardized metadata format</td>
</tr>
<tr>
<td>discipline-specific information and ontologies</td>
<td>use a standard ‘profile’ that supports discipline-specific information</td>
</tr>
</tbody>
</table>

DataONE [https://www.dataone.org/education-modules](https://www.dataone.org/education-modules) Metadata, Slide 21
The Research Data Lifecycle: Document

- Create
- Preserve
- Share
- Store
- Use
- Document

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Data Documentation: Recommendations

- Collect all the information users would need to understand and reuse the data
- Create metadata at the time - it’s harder to do later
- Use standards where possible
- Name, structure, and version files clearly

DCC [http://www.dcc.ac.uk/training/rdm-librarians](http://www.dcc.ac.uk/training/rdm-librarians)
Revisiting Data Documentation

- laboratory notebooks & experimental protocols
- questionnaires, codebooks, data dictionaries
- software syntax and output files
- information about equipment settings & instrument calibration
- database schema
- methodology reports
- provenance information about sources of derived or digitized data
Levels of Data Documentation

• **Project level**: what the study set out to do, how it contributes new knowledge to the field, what the research questions/hypotheses were, what methodologies were used, what sampling frames were used, what instruments and measures were used, etc.

• **File or database level**: how all the files (or tables in a database) that make up the dataset relate to each other; what format they are in; whether they supercede or are superceded by previous files

• **Variable or item level**: name and description of each variable in a dataset, what each item/row in a data table represents
Why Document your Data?

The Three Rs of Reproducibility:

1. **Replication** verifies that the observed findings are sufficiently stable to be discovered more than once using the same methods.

2. **Reproduction** verifies that the findings are not solely attributable to the experimental method. For this reason, a new experiment is run using different experimental methods to test the same hypotheses as the baseline experiment.

3. For **reanalysis**, the data of a previously conducted experiment are used to verify the results rather than re-running the experiment.
File/Folder Naming: Benefits

- Data files are distinguishable from each other within their containing folder
- Data file naming prevents confusion when multiple people are working on shared files
- Data files are easier to locate and browse
- Data files can be retrieved not only by the creator but by other users
- Data files can be sorted in logical sequence
- Data files are not accidentally overwritten or deleted
- Different versions of data files can be identified
- If data files are moved to another storage platform their names will retain useful context
File/Folder Naming: Best Practices

- **Organization** - important for future access and retrieval, and needs to take into account the file naming constraints of the system where the file is located

- **Context** - this could include content specific or descriptive information, independent of where the data are stored

- **Consistency** - choose a naming convention and ensure that the rules are followed systematically by always including the same information (such as date and time) in the same order (e.g. YYYYMMDD)
File Organization: Naming Conventions

<table>
<thead>
<tr>
<th>Best Practice</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Limit the file name to 32 characters</strong> (preferably less!)</td>
<td>32CharactersLooksExactlyLikeThis.csv</td>
</tr>
<tr>
<td><strong>When using sequential numbering, use leading zeros to allow for multi-digit versions</strong></td>
<td>NO ProjID_1.csv</td>
</tr>
<tr>
<td></td>
<td>YES ProjID_01.csv</td>
</tr>
<tr>
<td></td>
<td>NO ProjID_12.csv</td>
</tr>
<tr>
<td></td>
<td>YES ProjID_12.csv</td>
</tr>
<tr>
<td><strong>Don’t use special characters</strong></td>
<td>NO name&amp;<a href="mailto:date@location.doc">date@location.doc</a></td>
</tr>
<tr>
<td>&amp; , * % # ; * ( ) ! @$ ^ ~ ' { } [] ? &lt; &gt; -</td>
<td></td>
</tr>
<tr>
<td><strong>Use only one period</strong> and use it before the file extension</td>
<td>NO name.date.doc</td>
</tr>
<tr>
<td></td>
<td>NO name_date..doc</td>
</tr>
<tr>
<td></td>
<td>YES name_date.doc</td>
</tr>
<tr>
<td><strong>Avoid using generic data file names</strong> that may conflict when moved from one location to another</td>
<td>NO MyData.csv</td>
</tr>
<tr>
<td></td>
<td>YES ProjID_date.csv</td>
</tr>
</tbody>
</table>
File Versioning

- File versioning ensures that a **clear audit trail** exists for tracking the development of a data file and identifying earlier versions when needed.
- Common to use ordinal numbers (1, 2, 3, …) for major version changes and dashes for minor changes (1-01, 1-02, 1-03, …)
- Beware of using confusing labels: revision, final, final2, definitive_copy

MANTRA [https://mantra.edina.ac.uk/organisingdata/](https://mantra.edina.ac.uk/organisingdata/) Slide 17
Activity #3: Practicing Good File Naming

**Scenario:** You have been asked by your colleague, Dr. Psi, to help reorganize their team’s data files and folders for their “Monorail Project.” Look over the various data files and sub-folders in Dr. Psi’s current project folder. Can you create a file naming convention so all the data files for the project could live in a single directory called “Monorail Project”?

**Issues:**
1. What file formats are included in the project data files?
2. What pieces of information should each file name include?
3. How might you indicate whether a file contains raw or processed data?
4. How might you standardize file versions?

University of Minnesota’s Data Management Workshop Series, Session 2
https://tinyurl.com/y9zlglmw
The Research Data Lifecycle: Use

Create → Document → Use → Share → Preserve → Create

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Data Access and Use

- Restrict access to those who need to read/edit data
- Follow good practices for usernames and passwords
- Consider the data security implications for where you store data and from which devices you access files
- If storing sensitive data on USB drives, use drives with encryption software, which protects the data if the drive is lost
- Choose appropriate methods to transfer and share data - institutional file storage (UB Box) and encrypted media are better than email or cloud storage (Google Drive, Dropbox, etc.)
The Research Data Lifecycle: Store

Create → Document → Use → Share → Preserve → Store → Create

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Data Storage and Backup

- Use managed services where possible - e.g. institutional file storage (UB Box) rather than local or external hard drives
- Ask the local IT team for advice
- The 3-2-1 rule for data storage:
  - at least 3 copies of a file
  - on at least 2 different media
  - with at least 1 offsite

Would anyone like to share a data loss horror story?

DCC [http://www.dcc.ac.uk/training/rdm-librarians](http://www.dcc.ac.uk/training/rdm-librarians)
Why Perform Backups?

- Limit or negate loss of data, some of which may not be reproducible
- Save time, money, productivity
- Help prepare for disasters
  - Accidental deletions
  - Fires, natural disasters
  - Software bugs, hardware failures
- Reproduce results of past procedures (if they were based on older files)
- Respond to data requests
- Limit liability
Backup Strategy

• How will you back up your data?
• How regularly will backups be made?
• Will all data be backed up or only those that have been changed? (incremental vs. full backups)
• How often will full and incremental backups be made?
• How long will backups be stored?
• How much storage space is required to maintain backups?
• How will you keep track of different data versions, across multiple devices?
• What backup services or tools will you use?
Data Security

Planning for data security can help prevent:

- Accidental or malicious damage/modification to data.
- Theft of valuable data.
- Breach of confidentiality agreements and privacy laws.
- Premature release of data, which can void intellectual property claims.
- Release before data have been checked for accuracy and authenticity.
The Research Data Lifecycle: Share

Create → Document → Use → Store → Share → Preserve → Create

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Benefits of Sharing Data

1. It reinforces open scientific inquiry.
2. It supports the verification and replication of original results.
3. It promotes new research and allows for the testing of new or alternative methods.
4. It encourages collaboration and multiple perspectives.
5. It provides important teaching resources.
6. It reduces costs by avoiding duplicate data collection effort.
7. It protects against faulty or fraudulent data.
8. It enhances the visibility and overall impact of research projects.
9. It preserves data for future use.
10. Helps the broader community and individual researchers “do better research.”
Module 5: Data Sharing, Citation and Re-Use

https://www.youtube.com/watch?v=z94ZWwJ5GZ4
Challenges to Sharing Data

1. It takes time and effort to prepare data for sharing.
2. There are perceived risks from loss of control of the data.
3. Data can contain confidential or sensitive information.
4. Ownership of data can be unclear or problematic.
5. Lack of incentives for sharing data.
6. Lack of knowledge and experience with data management.
File Formats

• Open vs. Proprietary Formats
  • Open formats can be opened by a variety of programs; open standards are preserved throughout time
  • Proprietary formats can only be read by proprietary software

• Value of saving research data as text files: .txt, .csv, .asc, .html, .xml
  • Both human-readable and readable by any text editor
  • Very unlikely to become obsolete anytime soon

• Internationally standard file formats enable data sharing
Three Types of Repositories

- Institutional
- Cross Disciplinary
- Discipline Specific

NYU Teaching Toolkit
The University at Buffalo Institutional Repository (UBIR) collects, preserves, and distributes UB's research and scholarship. UB faculty and staff interested in contributing a collection or an item should contact Karlen Chase (klchase@buffalo.edu).

UBIR Repository

The University at Buffalo Institutional Repository collects, preserves, and distributes UB's research and scholarship. This Institutional Repository is a group of collections of full-text articles, full-text reports, audio mp3 files, images, and video clips produced by UB researchers and students. Other types and formats of digital scholarship will be added as they become available. UB faculty and staff interested in contributing a collection or an item should contact us at klchase@buffalo.edu.

Communities in UBIR

Select a community to browse its collections.

Anderson Gallery

CAS - Biological Sciences

Center for High Performance Computing (CHPC)
Storage ≠ Preservation ≠ Access

the “dark archive” zone
The Research Data Lifecycle: Preserve

Image from: Digital Curation Conference
http://www.dcc.ac.uk/training/rdm-librarians
Preserving Your Data

Store your data in a place where it will be:

• Backed up
• Discoverable
• Accessible for the future

Preservation requires long-term storage and access.
FAIR Data Principles for Data Management & Curation

Your data should be...
- Findable
- Accessible
- Interoperable
- Reusable


Challenges to Preservation

- **Benign neglect**: not having active long-term preservation strategy
- **Bit rot**: digital data degradation over time without refreshing
- **Obsolescence**: software and hardware become obsolete over time
- **Insufficient documentation**: even if data are saved, they’re useless if we don’t know what they are
Preservation Best Practices

• To ensure **authenticity**: keep single master file of data; assign responsibility for master files to single person; regulate write-access to master files; record all changes to master files; archive copies of master files at regular intervals; develop formal procedure for destruction of master files

• To ensure **integrity**: create automated backup process for critical or frequently used files; store master files in open-source formats; verify backup copies against originals using checksums; store copies of files on two different media; refresh data by copying to new media every 2-5 years
WRAP-UP & RESOURCES

Major take-away: data management is most effective when it is performed throughout the entire research data lifecycle!
Resources Cited

- UMass Medical School, New England Collaborative Data Management Curriculum, Modules for Managing Research Data: [https://library.umassmed.edu/resources/necdmc/modules](https://library.umassmed.edu/resources/necdmc/modules)
- University of Edinburgh, MANTRA Research Data Management Training, Learning Units: [https://mantra.edina.ac.uk/](https://mantra.edina.ac.uk/)
- DataONE Education Modules: [https://www.dataone.org/education-modules](https://www.dataone.org/education-modules)
- Data Curation Center, Training and Reference Materials, RDM for Librarians: [http://www.dcc.ac.uk/training/rdm-librarians](http://www.dcc.ac.uk/training/rdm-librarians)
- Data Curation Center, Checklist for a Data Management Plan: [http://www.dcc.ac.uk/resources/data-management-plans/checklist](http://www.dcc.ac.uk/resources/data-management-plans/checklist)
- MIT Libraries, Research Data Management 101: The Lifecycle of a Dataset: [https://www.dropbox.com/s/7k4qt43phcsohhr/rdm101_Slides_MIT.pdf?dl=0](https://www.dropbox.com/s/7k4qt43phcsohhr/rdm101_Slides_MIT.pdf?dl=0)
- University of Minnesota Libraries, Data Management Workshop Series: [https://www.lib.umn.edu/datamanagement/workshops](https://www.lib.umn.edu/datamanagement/workshops)
- UNC Chapel Hill and the University of Edinburgh, Research Data Management and Sharing, Coursera Course: [https://www.coursera.org/learn/data-management](https://www.coursera.org/learn/data-management)