Research Data Management: Best Practices
Hilda Teresa Ayala González
March 5th, 2018
Graduate Research and Innovation Center (GRIC)
University of Puerto Rico, Mayagüez Campus

Pre-test

• What is data management

• What is the research data lifecycle

• Mention three (3) benefits of proper data management

• Select all the examples of good data management practice

Goals

• Understand what is data management

• Learn about the benefits of proper data management

• Find out some of the best practices for data management

• Improve your data management skills

Research Data

Data means different things to different people in different contexts. Different disciplines have and use discipline-specific language around the subject research data.

• Data may be viewed as the lowest level of abstraction from which information and knowledge are derived.

• Research data are collected, observed, or created, for the purposes of analysis to produce and validate original research results.

• Examples of research data may include a spreadsheet of statistics, a series of email messages, a sound recording of an interview, a descriptive record of a rock specimen, or a collection of digital images from a SEM (scanning electron microscope)
**Data Management**

*Data management* refers to the everyday handling and workflow of research data during the active phase of a project as well as the practices that support long-term preservation, access, and use after the project has been completed.

Essential part of the responsible conduct of research. Before project starts. Throughout the entire research data lifecycle.

---

**Data Management: Benefits**

- Easier to find, use, and analyze
  - Better understanding of your data
    - Unnecessary duplication
  - Decreased risk of loss, theft or inappropriate use of data
  - Proper and thorough documentation
  - At the end of the project
    - Faster disposition or sharing
    - Increase citations

*Many organizations find out the hard way that the costs of maintaining data properly are insignificant compared to the expenses incurred in trying to replace or recreate lost data."

---

**Research Data Lifecycle**

- What data do I need? What for?
- Where can I find trustworthy data?
- Where will I publish findings?
- Plan
- Discover & Route
- Create
- Analyze variables to answer research question
- Collect or edit data
- Publish
- Appraise
- Use
- Select what will be kept and published

*MANTRA and the Research Data Lifecycle*

University of Edinburgh, 2012

---

**Data Management: Benefits**

- Easier for your collaborators (or others) to understand the data
- Allows for project continuity if team changes
- Promotes reproducibility and validation of research results
- Increase discoverability, interpretation and reuse by others
- Diversifies research collaborations
- Avoid duplication of research
- Accelerate scientific progress

*An effective data management program would enable a user 20 years or longer in the future to discover, access, understand, and use particular data."

---

*FOR YOU*

*FOR OTHERS*
Research Data Management
Best Practices
Basics for research integrity

Exercise #1: Current state of your data

- What type of data are you collecting
  - In what format
- Do you have a favorite folder structure
  - Can you draw an example
- How do you name your files
  - Give an example
- How do you differentiate versions
  - Give an example
- How do you keep track of changes
  - What method you use
- Where do you store it
  - Name all the places
- What security measures do you use so the data won’t be touch by unauthorized people
- Who owns the copyrights of your data

The Basics

- Formats
- Folder structure & file names
- Version control
- Essential documentation
- Storage options
- Security
- Rights & ownership
Formats

- Multiples
- Proprietary
- Discipline oriented
- Compatibility
- Obsolescence

**Tool:**
LC Sustainability of Digital Formats Initiative

Formats: Best Practices

- Open
- Documentation
- Non-proprietary
- Non-compression
- Un-encrypted
- Metadata friendly
- Widely accepted by research community

Structure & File Naming

**Electronic Files**

- Chaotic quantities
- Duplicity
- They are everywhere
- Junk
- Investment of time

- Electronic Files
  - Chaotic quantities
  - Duplicity
  - They are everywhere
  - Junk
  - Investment of time
Structure & File Naming

**General Rules**

- Avoid special characters
  - !@#$%^&&*?\|
- Instead of spaces, use an underscore (_) or dash (-)
- You can also use a capital letter, but it is not always reliable
- Use shortcuts to avoid file duplication and version control
- Screenshots or a **content file**
  - Know what is what
  - Unnecessary looking through every folder

Structure & File Naming: Best Practices

**General Rules**

- Organize yourself
  - Think about the types of documents and data you will be generating/using
- Keep folders and files out of your desktop
  - Shortcuts are ok, for short periods of time
  - This is also a security issue!
- Name folders according to project, function, activity, type
  - Thesis / Finances / Administrative / Images-2017
- Use descriptive file naming
- Create something simple that works for you and your team
- Be consistent with format

**Example (ok deep)**

- **Courses** / Procedures / UBC
- [Spring 2015 / Summer 2015 / Winter2014](#)
- HistoryRecordsKeeping / ProjectManagement
Structure & File Naming: Best Practices

- As possible incorporate the date YYYYMMDD in the name of files
  - Don’t trust the file date
  - Helps organize files in chronological order
- Include elements that facilitates the recognition of the file without opening it
  - Project number, description of content, date of creation or modification, name of creator or team, version number
  - Reference article: Johan_2017_DigitalPreservation

Example (ok short)

<table>
<thead>
<tr>
<th>Date</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/23/2016 6:30 PM</td>
<td>DisclosureStatement</td>
</tr>
<tr>
<td>6/23/2016 6:30 PM</td>
<td>DisclosureStatement-1</td>
</tr>
<tr>
<td>4/15/2015 2:31 PM</td>
<td>NEWSUI_loan_A4_formGradMarch2015</td>
</tr>
</tbody>
</table>

Author       Year  Short title
---          ---    ---

<table>
<thead>
<tr>
<th>Project</th>
<th>MMDD</th>
<th>Time of sample</th>
<th>Initials of Researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>0823</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Version Control: Best Practices

- Avoid:
  - Draft, Revision, Final, Approved
  - Thesis_Final_Draft.doc

- Significant changes
  - Date and version
    - V01, V02, V03
    - POP_20120312_V03

- Minor revisions
  - V1-01
  - POP_20120312_V03-01

- Keep older versions in a separate Folder
  - Erase when the older versions are no longer needed

- Use track changes!

Documentation

Data documentation explains how data were created or digitized, what data mean, what their content and structure are, and any manipulations that may have taken place.

It ensures that data can be understood during research projects, that researchers continue to understand data in the longer term and that re-users of data are able to interpret the data.

Good documentation is also vital for successful data preservation.
Documentation

Where to look and what to keep
- Laboratory Notebook
- Laboratory procedures, user manuals
- Recordings
- Photographs
- Software and lab equipment documentation and user manual
- Data Management Plan
- Digital files (emails)

Documentation: Best Practices

- Document from the beginning
- Do it constantly
- Keep your documentation with your data
  - Text “Read Me” File
  - Separate tab on spreadsheet
- The more “born-digital” the better
- Digitize hard data (e.g. laboratory notebook)

Documentation

Describe the digital context
- Name of the data set
- Example of the data file(s) in the data set
- Date the data for each data type file
- Pertinent companion files
- List of related or ancillary data sets
- Software used to prepare/read the data set
- Data processing that was performed

Describe the personnel and stakeholders
- Who collected the data
- Who should be contacted with questions
- Sponsors

Describe the scientific context
- Scientific reason why the data were collected
- What data were collected
- What instruments (model and serial) were used
- Environmental conditions during collection
- Where and when collected and spatial resolution
- Standards or calibrations used

Information about parameters
- How each was measured or produced
- Units of measure
- Format used in the data set
- Precision, accuracy, and uncertainty
- Definitions of codes used
- Quality assurance and activities
- Known problems that limit data use

Documentation: Best Practices

- Spreadsheets: Create a template
  - Content: parameter, format, units, and codes for missing values
  - Atomize: one piece of data per entry (e.g. database format)
  - Consistency
  - Completeness

- Lab Notebooks
  - Cross reference with data location
  - Always date & sign entries
  - Bound book (not spiral), number pages, don’t tear
  - Write clearly in permanent Ink
  - Document changes, even if they are minor
  - Keep in safe cabinet
Documentation: Best Practices

Retention

- 3 to 7 years after project completion
- Or the period established by funding agency
- Should remain in the research institution, in a safe place and be readable
- If future value, might need to be kept longer

Nevertheless, it is important to have a clear retention policy that balances the best interests of society with those of the research institution and the individual researcher.

Storage & Security: Issues

- Control access to data
- Tracking
- Obsolescence
- Corruption of files
- Loss
- Backups
- Third parties
Storage & Security: Best practices

- Control access to data
  - Access privileges
  - Limit access to raw data folder
  - Keep your stuff secure & safe

- Tracking
  - Keep a log
    - Medium + content + access privileges
  - Screenshot or content file
    - Know where is what
    - Extra control to who gets to actually open something

Storage & Security: Best practices

- Third parties: Cloud storage
  - It means
    - Someone else can access your data
    - Bankrupt, disappear, sold
  - Read their terms and services
    - Data encryption (transfer or storage)
    - Servers location
    - Data ownership
    - Data deletion
  - Don’t have it ALL up there!

Copyrights and Ownership

- Data are facts, which means that they are not subject to restrictions or protections associated with copyright

- According to the US Copyright Law of the United States (Title 17)

  “compilation” is a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship.
Copyrights and Ownership: Best Practices

Ownership and rights will determine how the data can be managed into the future, so these should be documented early in a project through data management planning.

- Clarify who has primary ownership of the data
- Whose rights should be considered when making decisions about the management and dissemination of the data
  - Funders
  - Your institution
  - Research subjects
  - Collaborators
  - Publishers
  - General public

Exercise #2: Apply best practices to your data

- What type of data are you collecting
  - What format you will use for accessibility
- Do you have a favorite folder structure
  - Draw a better folder structure
- How do you name your files
  - Give a proper example
- How do you differentiate versions
  - Give a proper example
- How do you keep track of changes
  - What other methods and documentation you will keep
- Where do you store it
  - How can you make it better
- What security measures do you use so the data won’t be touch by unauthorized people
  - How can you make it better
- Who owns the copyrights of your data
  - Who else is involved and what type of rights they own.

Issues: Preservation & Long-term access

- Up to a few years ago, preserving data was not considered as important as the research product
- Agencies and institutions around the World have learn about the importance and advancements of data sharing
- This interest in research data as an asset, is pushing creators to care about preservation and access since the early stages of their research, as they may be required to share it when they publish an article or as a requirement of the grant that funded the project
Issues: Preservation & Long-term access

- Ensures:
  - Integrity
  - Authenticity
  - Reliability
  - Usability
  - Interoperability
  - Discoverability

- Applying these best practices will not only make your research process easier, but will assure your data can be preserved and access in the future by you and others

Post-test

- What is data management
- What is the research data lifecycle
- Mention three (3) benefits of proper data management
- Select all the examples of a good data management practice

Questions and comments

References


