

RESEARCH DATA MANAGEMENT FUNDAMENTALS

Data Documentation

File Organization

Storage & Backup

**Data Publishing,
Sharing & Reuse**

**Protecting Data &
Responsible Reuse**

FUNDAMENTAL BEST PRACTICES IN RESEARCH DATA MANAGEMENT

The data underlying published research are increasingly viewed as an important resource with the potential for publishing, reuse, and the promotion of new research. Recent grant funder, federal, and MSU policies mandate data management planning, and many research communities actively support the sharing of data. This increased interest in data points to the need for guidelines to fundamental best practices in research data management.

NEW POLICIES

Grant funders such as the National Science Foundation now require the submission of a “Data Management Plan” outlining how projects will conform to its policy on the dissemination and sharing of research results. Government directives such as the February 2013 White House policy memo, which will require public access to federally funded scientific research, reinforce the importance of research data management planning. At MSU, the University Research Council has endorsed best practices for research data management, control, and access (http://rio.msu.edu/research_data.htm).

THE CHANGING RESEARCH LANDSCAPE

More and more, the research community itself supports data sharing and open access. This changing research landscape is exemplified in the rise of data citations and other metrics for tracking data reuse, publishing supplementary data with journal articles, datasets as standalone publications, data journals, and data repositories and archives.

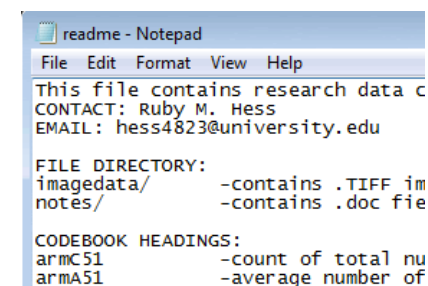
MANAGING YOUR DATA

Data management is the process of planning for and implementing a system of care before, during, and after a research project in order to ensure a usable resource. Caring properly for data will allow you and other researchers to access and understand your data during and after the life of a project. Following best practices in research data management can help you secure grant funding and create a data output that becomes part of the scholarly record. This document walks you through the key areas of fundamental data management best practices.

Without proper documentation, researchers may find even their own datasets difficult to decipher or reuse. Preserving the *who*, *what*, *where*, and *when* of data collection and analysis along with the data itself helps make datasets more accessible and intelligible to all users.

CREATING A “README” FILE

One simple way to store data documentation is by creating a file called `readme.txt` to accompany each dataset. The `readme` file houses all of the significant documentation about the dataset, providing users with a starting point that tells them what the data consists of, how it was collected, whether any restrictions apply to its distribution or use, along with a range of other descriptive information.



```

readme - Notepad
File Edit Format View Help
This file contains research data c
CONTACT: Ruby M. Hess
EMAIL: hess4823@university.edu

FILE DIRECTORY:
imagedata/      -contains .TIFF im
notes/          -contains .doc file

CODEBOOK HEADINGS:
armC51          -count of total nu
armA51          -average number of
  
```

USING METADATA STANDARDS

Another word for the descriptive documentation that accompanies a dataset is “metadata,” or, data about data. Using a particular metadata standard means recording your documentation in certain formalized ways. This standardization means that data can more easily be found as well as compared to other datasets.

A FEW EXAMPLES

- **Dublin Core:** Commonly-used descriptive metadata format to facilitate discovery of datasets across the Web.
- **Data Documentation Initiative (DDI):** Standard that defines metadata content, presentation, transport, and preservation for the social and behavioral sciences.
- **ISO 19115:** Describes geographic data such as maps and charts.

Metadata fields could be: Title, Language, Dates, File Structure or Formats, Creator, Location, Methodology, Code Lists, Rights, Versions, Funders, Access Information, List of Files Names, Variables, and much more.

FILE ORGANIZATION

It is important to understand the risks of using certain file formats, and to make informed decisions regarding the appropriate “container” for the content of your files. When in doubt, use standard file formats that are widely supported by your community of practice.

UNDERSTAND FILE FORMATS

It is important to choose platform and vendor-independent file formats to ensure the best chance for future compatibility. “Open” formats are often supported broadly by a community rather than individually by a company or vendor.

GOOD CHOICES FOR FILE FORMATS

- Non-proprietary
- Open, documented standard
- Common usage by research community
- Standard representation (ASCII, Unicode)
- Unencrypted
- Uncompressed

FORMAT GENRE	OPTIMAL STANDARDS
TEXT	.txt; .odt; .xml; .html
AUDIO	.flac; .wav
VIDEO	.mp2; .mp4; .mkv
IMAGE	.tif; .png; .svg; .jpg
DATA	.sql; .csv

DESIGN A FILE PLAN

A file structure is the framework of your file plan. Think of this as a classification system to make it easier to locate folders and files.

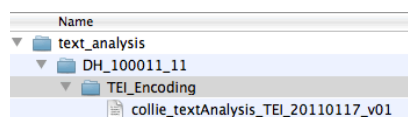
BENEFITS

- Simple organization is intuitive to team members and colleagues
- Reduces duplicate copies in personal drives and email attachments

GOOD PRACTICES

- Choose a sortable directory hierarchy
 - Investigator, Process, Date
 - Instrument, Date, Sample

EXAMPLES



README.txt Documentation:

Here is an example of a method you might use to document a file directory structure in your plain text README file.

If your desired directory structure is:

/text_analysis/DH_100011_11/TEI_Encoding

You can specify a generic template:

/[project]/[grant number]/[event]/

This way you will know how to name future folders as your project grows!

USE A FILE NAMING CONVENTION

A consistent file naming convention will enable better access to your files, create logical sequences for file sorting, and make it easier to search for information.

GOOD PRACTICES

- Meaningful but short (255 character limit)
- Use alphanumeric characters (e.g. abc123)
- Capital letters or underscores differentiate between words
- Surname first followed by initials of first name
- Use the year-month-day format for dates, with or without hyphens (e.g., 2006-03-13 or 20060313)
- Decide on a simple “versioning” method (e.g. file_v001)

BENEFITS

- Create logical sequences for sorting through many files and versions
- Identify what you’re searching for by filename

FILE NAME EXAMPLES

☐ sharpeW_krillMicrograph_20110117.tif

☐ borgesJ_collocation_20080414_d001.xml



README.txt Documentation:

Here is one way you might create a template for file names by using your plain text README.

To create consistent file names like:

sharpeW_krillMicrograph_20110117.tif

borgesJ_collocation_20080414_d001.xml

You should specify a template such as:

[investigator]_[descriptor]_[YYYYMMDD].[ext]

STORAGE & BACKUP

Without a storage and backup plan data are at significant risk of loss. Hardware failures, network failures, bit rot, and human errors are only a few of the risks to data longevity. Commercial grade hard drives such as those found in your laptop and desktop cannot be trusted singularly for storage of data. An effective data storage plan will make provisions for a primary authoritative copy of data, a secondary local backup, and a tertiary remote backup.

AVOID SINGLE POINTS OF FAILURE

A single point of failure occurs when it would only take one event to destroy all data on a device. Imagine if that happened right now. Do you have a backup of your data to restore from? Is it current?

Good practices for avoiding single points of error:

- Use managed networked storage whenever possible
- Move data off of portable media
- Never rely on one copy of data
- Do not rely on CD or DVD copies to be readable
- Be wary of software lifespans (e.g. ANGEL)

ENSURE DATA REDUNDANCY AND REPLICATION

Proper data storage uses a tiered approach, and does not rely solely on hardware or software redundancy. Data must be replicated to a reliable backup system.

- Make 3 copies
 - E.g. original + external/local + external/remote
 - E.g. original + 2 formats on 2 drives in 2 locations
- Geographically distribute and secure
 - Local vs. remote, depending on needed recovery time
 - Personal computer, external hard drives, departmental, or university servers may be used



README.txt Documentation:

Create an inventory of the locations of your important data. This might include descriptions of users, computers, server addresses, backup locations, and third-party storage details.

COMMON TYPES OF STORAGE							
	DESCRIPTION	PORTABLE DATA TRANSFER	SHORT TERM STORAGE	PROJECT TERM STORAGE	NETWORKED DATA TRANSFER	LONG TERM STORAGE	RELIABLE BACKUP OPTION
Optical Media	CDs, DVDs, Blu-Ray	✓	✗	✗	✗	✗	✗
Portable Flash Media	SD cards, USB sticks, internal device memory	✓	✓	✗	✗	✗	✗
Commercial Hard Drives	Commercial grade spinning disk drives	✓	✓	✓	✗	✗	✗
Commercial NAS	Commercial network attached storage	✗	✓	✓	✓	✗	✗
Cloud Storage	Cloud service storage with web interface	✗	✓	✓	✓	✗	✗
Enterprise Network Storage	Rack-based high capacity storage with customer service	✗	✓	✓	✓	✓	✓
Trusted Archival Storage	Enterprise storage with a sustainable business model for preservation	✗	✗	✗	✓	✓	✓

RECOMMENDED STORAGE @ MSU: CLOUD STORAGE						
	DESCRIPTION	SHORT TERM STORAGE	PROJECT TERM STORAGE	NETWORKED DATA TRANSFER	LONG TERM STORAGE	RELIABLE BACKUP OPTION
ANGEL	Free. Ideal for collaboration; not intended as storage space. Phase out date of 2015. http://angel.msu.edu/	✓	✓	✓	✗	✗
Desire2Learn	Free. Ideal for collaboration; not intended as storage space. https://d2l.msu.edu/	✓	✓	✓	✗	✗
Google Apps	Free. Ideal for collaboration; not intended as storage space. https://googleapps.msu.edu	✓	✓	✓	✗	✗
RECOMMENDED STORAGE @ MSU: ENTERPRISE STORAGE						
	DESCRIPTION	SHORT TERM STORAGE	PROJECT TERM STORAGE	NETWORKED DATA TRANSFER	LONG TERM STORAGE	RELIABLE BACKUP OPTION
AFS Storage	Free up to 1GB, additional space can be purchased with department account. http://afs.msu.edu/	✓	✓	✓	✗	✓
Individual Storage	Free based. For more information visit: http://tech.msu.edu/storage/	✓	✓	✓	✗	✓
Mid-Tier Storage	Free based. For more information visit: http://tech.msu.edu/storage/	✓	✓	✓	✗	✓
Enterprise Storage	Free based. For more information visit: http://tech.msu.edu/storage/	✓	✓	✓	✗	✓
HPCC Home Directory	Free up to 1TB. Fee based additions available. For more information visit: https://wiki.hpcc.msu.edu	✓	✓	✓	✗	✓
HPCC Research Directory	Free up to 1TB. Fee based additions available. For more information visit: https://wiki.hpcc.msu.edu	✓	✓	✓	✗	✓

Research datasets are on the way to becoming first-class scholarly contributions on par with peer-reviewed journal articles. There are a number of considerations when deciding whether or not your research project is best served by restricting access or including a data publication to more broadly share the results of your research.

REASONS TO SHARE AND PUBLISH YOUR DATA

Preparing data in a format suitable for sharing and publication is a time-intensive process. The return on investment has the potential to be well worth the effort. Consider these positive outcomes:

- Increased research impact and citations
- Enable additional scientific inquiry
- Provide opportunities for co-authorship and collaboration
- Enhance your grant proposal's competitiveness

DATA PUBLICATION VENUES

There are multiple ways to publish research data, each offering varying levels of support for indexing, access controls, and long-term curation.

- Faculty or project website
- Journal supplementary materials
- Repository (data archive)

ARCHIVE IT!

Disciplinary data repositories, also known as data archives, provide a secure way to share your data and ensure long-term access. They are usually the most visible place to publish and often offer persistent citations. The availability of disciplinary repositories varies across domains as data sharing norms continue to evolve. Consult *Databib.org* for a directory of research data repositories.

PROTECTING DATA & RESPONSIBLE REUSE

When sharing your data, remember to consider how you plan on protecting the data and any intellectual property rights, while also encouraging the reuse of your data by other researchers.

INTELLECTUAL PROPERTY

Intellectual property (IP) refers to the exclusive rights creators of works have. While individual data cannot be protected by copyright, the organization of the data, such as in a database, any creative work produced with the data, and research instruments used may be protected by copyright, patents, trademarks, or trade secrets. Consider the following when publishing your data:

CITE IT!

Providing an example of how you'd like your work to be cited encourages proper attribution when it is reused.

- Ownership of intellectual property rights
 - The principal investigator's institution usually holds any IP rights
- Provide a clearly stated license for producing derivatives, reusing, and redistributing data sets
 - License your work under Creative Commons (creativecommons.org) or provide an explicit statement of use on your work
 - State if there are any restrictions or delays in using the work

ETHICS & DATA SHARING

Keep in mind the following ethical concerns when sharing your data:

- Privacy
- Confidentiality
- Security and integrity of the data

For data involving human subjects, obtaining written permission or consent stating how the data may be reused can help with some ethical issues.



README.txt Documentation:

If you are aware of IP or copyright issues with your data, document these concerns as free text in your README file. Note the team member's roles and permissions, and if you are utilizing a software license be sure to include the license text!

FOLLOWING BEST PRACTICES = HIGH IMPACT DATA

As recognition for the value of sharing data continues to grow, so will the need for fundamental best practices in research data management.

- File organization ensures easier access and retrieval of your data during and after your project.
- Documentation makes your datasets accessible and intelligible to users.
- Storage and backup safeguards your data against technical failure, human error, and natural catastrophe.
- Data publishing and sharing encourages the most widespread reuse of your data.
- Data protection ensures responsible reuse in light of intellectual property and ethical concerns.

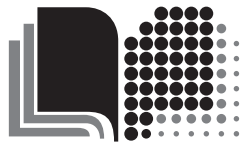
Following best practices in research data management accomplishes more than compliance with policy mandates. In the spirit of the new research landscape of accessible and open data, enabling reuse of data will increase the impact of your research and promote new research opportunities.

FOR MORE INFORMATION

To learn more about how to manage your research data throughout its lifecycle, visit the Research Data Management Guidance website at <http://lib.msu.edu/rdmg/>. Contact us at researchdata@mail.lib.msu.edu.

**RESEARCH DATA
MANAGEMENT GUIDANCE**

<http://lib.msu.edu/rdmg>
researchdata@mail.lib.msu.edu



Michigan State University

LIBRARIES

366 W. Circle Drive
East Lansing, MI 48824

University Archives
& Historical Collections

Conrad Hall, 888 Wilson Rd., Room 101
East Lansing, MI 48824

MICHIGAN STATE
UNIVERSITY

MSU is an affirmative-action, equal-opportunity employer.

