Research Data Management: Strategies for Data Sharing and Storage

Research Data Management Services @ MIT Libraries

- Workshops
- Web guide: http://libraries.mit.edu/data-management
- Individual assistance/consultations
  - includes assistance with creating data management plans

Why Share and Archive Your Data?

- Funder requirements
- Publication requirements
- Research credit
- Reproducibility, transparency, and credibility
- Increasing collaborations, enabling future discoveries

Research Data: Common Types by Discipline

General	Social Sciences	Hard Sciences
• images • video • mapping/GIS data • numerical measurements	• survey responses • focus group and individual interviews • economic indicators • demographics • opinion polling	• measurements generated by sensors/laboratory instruments • computer modeling • simulations • observations and/or field studies • specimen

Digital data

+

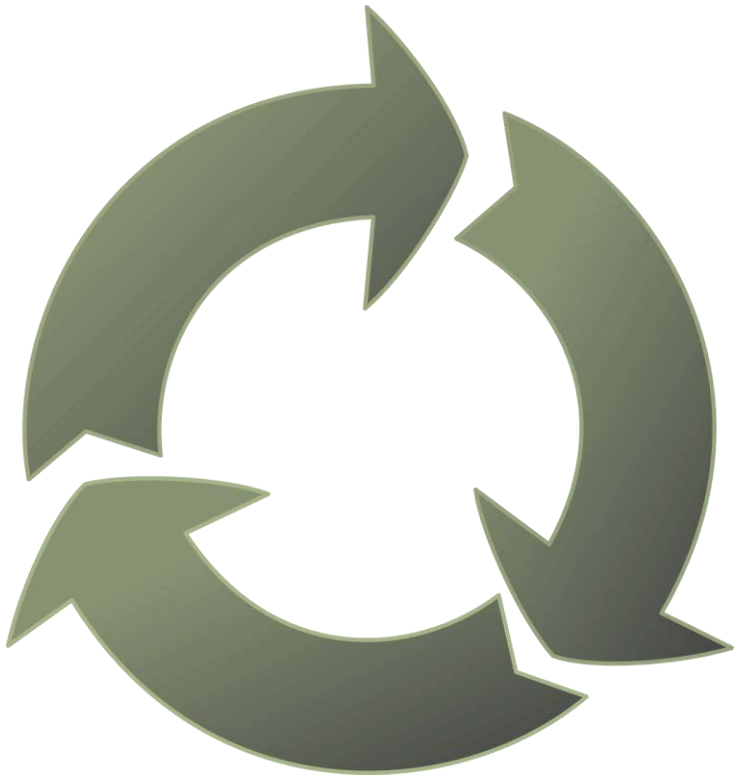
Complex workflows

Research Data: Stages

Raw Data	raw txt file produced by an instrument
Processed Data	data with Z-scores calculated
Analyzed Data	rendered computational analysis
Finalized/Published Data	polished figures appear in *Cell*

Setting Up for Reuse:

- **Formats**
- Versioning
- Metadata



Formats: Considerations for Long-term Access to Data

In the best case, your data formats are both:

- Non-proprietary (also known as *open*), **and**
- Unencrypted and uncompressed

Formats: Considerations for Long-term Access to Data

In the best case, your data files are both:

- **Non-proprietary (also known as *open*)**, and
- Unencrypted and uncompressed

Formats: Preferred Examples

Proprietary Format	Alternative/Preferred Format
Excel (.xls, .xlsx)	Comma Separated Values (.csv) ASCII
Word (.doc, .docx)	plain text (.txt), or if formatting is needed, PDF/A (.pdf)
PowerPoint (.ppt, .pptx)	PDF/A (.pdf)
Photoshop (.psd)	TIFF (.tif, .tiff)
Quicktime (.mov)	MPEG-4 (.mp4)

Formats: Considerations for Long-term Access to Data

In the best case, your data files are both:

- Non-proprietary (also known as *open*), and
- **Unencrypted and uncompressed**

Formats: Preferred Examples

Type of Data	Preferred Formats
Text	TXT, XML, PDF/A, HTML, ASCII, UTF-8
Still images	TIFF, JPEG 2000, PDF, PNG
Moving images	MOV, MPEG, AVI, MXF
Sounds	WAVE, AIFF
Statistics	ASCII
Databases	XML, CSV
Containers	TAR, GZIP, ZIP

Formats: Converting



Photos courtesy of Christine Malinowski, used with permission.

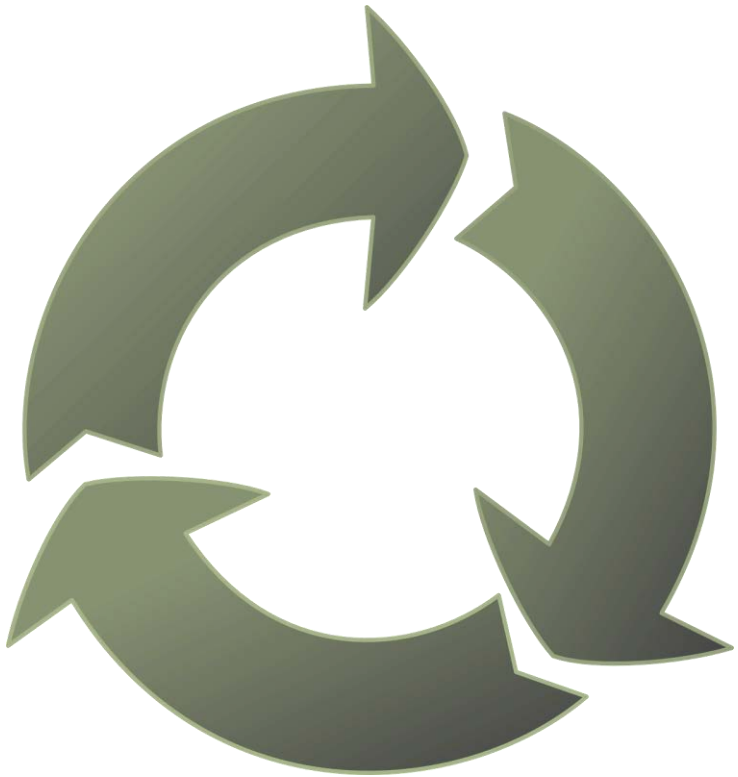
Information can be lost when converting file formats.

To mitigate the risk of lost information when converting:

- Note the conversion steps you take
- If possible, keep the original file as well as the converted ones

Setting Up for Reuse:

- Formats
- **Versioning**
- Metadata

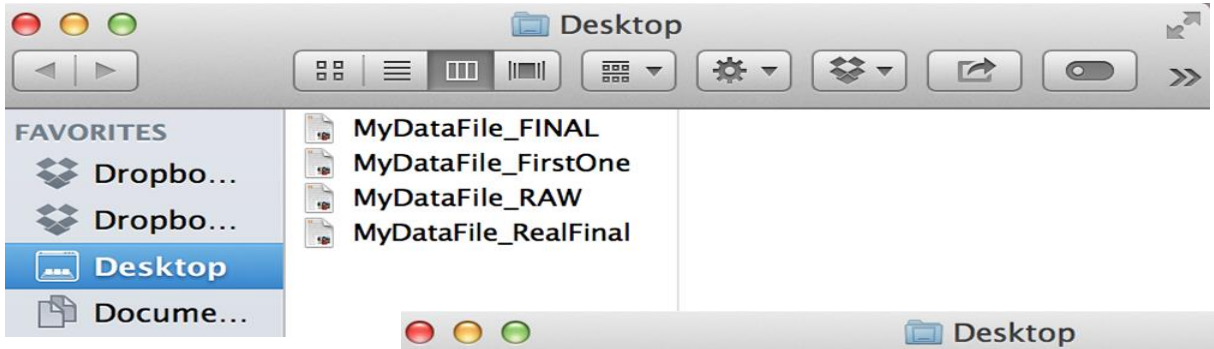


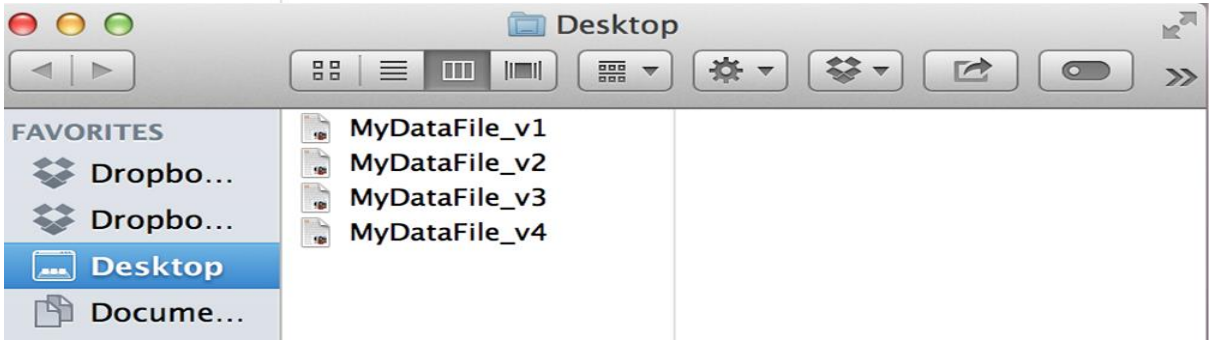
Versioning: Why do I need to worry about that?

- Have you ever had to leave the lab for a few days and have someone else pick up your project?
- Or picked up someone else's project?
- Will you leave your lab before a project is complete?
- Have you ever had to revisit a project after a break (to publish or pick it up again)?

Versioning: Basic Practices

Keep the original version of the data file the same and save iterative versions of the analysis/program/scripts files



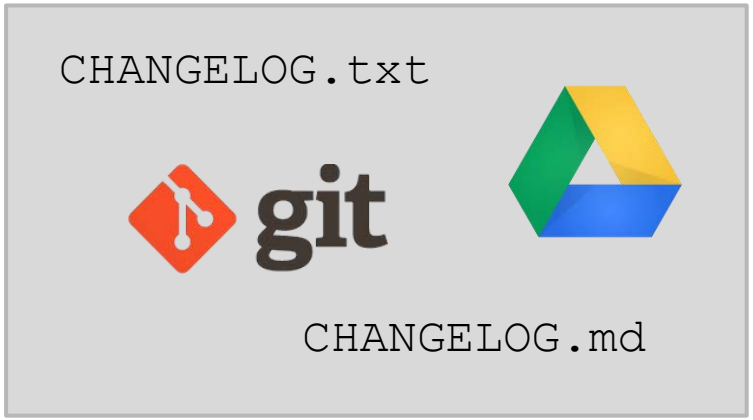


Versioning: Basic Practices

In some cases, it may make sense to log the changes so that you can quickly assess and access the versions.

It’s good to document:

- What was changed?
- Who is responsible?
- When did it happen?
- Why?



Versioning: File Naming Conventions

Naming conventions make life easier!

Naming conventions should be:

- **Descriptive**
- Consistent

Consider including:

- Unique identifier (ie. Project Name or Grant # in folder name)
- Project or research data name
- Conditions (Lab instrument, Solvent, Temperature, etc.)
- Run of experiment (sequential)
- Date (in file properties too)
- Version #

Versioning: File Naming Conventions

Naming conventions make life easier!

Naming conventions should be:

- Descriptive
- **Consistent**

YYYYMMDD
MMDDYYYY
YYMMDD
MMDDYY
MMDD
DDMM

Maintain order

TimeDate
DateProjectID
TimeProjectID

Sample001234
Sample01234
Sample1234

Include the same information

Versioning: File Naming Conventions

Best Practice	Example
Limit the file name to 32 characters (preferably less!)	32CharactersLooksExactlyLikeThis.csv
When using sequential numbering, **use leading zeros** to allow for multi-digit versions For a sequence of 1-10: 01-10 For a sequence of 1-100: 001-010-100	**NO** ProjID_1.csv ProjID_12.csv **YES** ProjID_01.csv ProjID_12.csv
Don't use special characters & , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? < > -	**NO** name&date@location.doc
Use only one period and use it before the file extension	**NO** name.date.doc **NO** name_date..doc **YES** name_date.doc
Avoid using generic data file names that may conflict when moved from one location to another	**NO** MyData.csv **YES** ProjID_date.csv

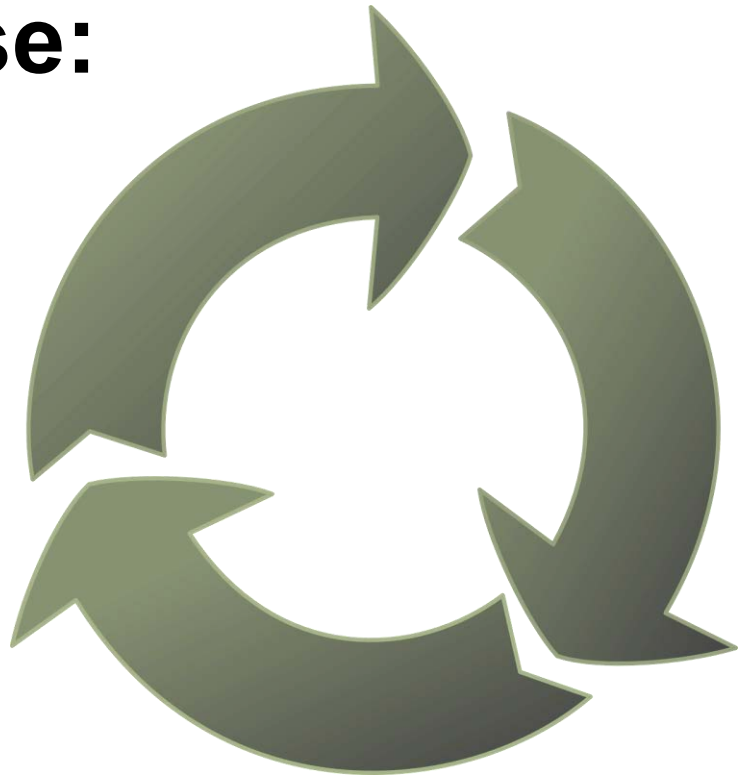
Versioning: File Naming Conventions

Resources:

- Check for Established File Naming Conventions in your discipline
  - DOE's Atmospheric Radiation Measurement (ARM) program
  - GIS datasets from Massachusetts
  - The Open Biological and Biomedical Ontologies
- File Renaming Tools
  - Bulk Rename Utility
  - Renamer
  - PSRenamer
  - WildRename

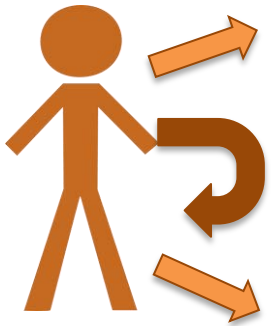
Setting Up for Reuse:

- Formats
- Versioning
- **Metadata**



Metadata should tell you…

- **What** do the data consist of?
- **Why** were the data created?
- What **limitations**, if any, do the data have?
- What does the data **mean**?
- How should the data be **cited**?



Metadata fields

- Title
- Creator
- Identifier
- Funders
- Dates
- Rights
- Processing
- Location

- Instruments used
  - Standards/calibrations used, environmental conditions
  - Units of measure
  - Formats used in the data set
  - Precision/accuracy
  - Software, data processing
  - Date last modified
  - …

Metadata: Things to Document

- Title……..datasetName
- Creator…….Malinowski, Christine
- Identifier……..dataID
- Funders……...NIH
- Dates…….20140123-20150114
- Rights……...We own this data.
- Processing……...Normalized
- Location…..…….This file is located in this directory MyProject_NSF_2014

Metadata Standards

- Provide common terms, definitions, structures.
- Ensure you have a complete, standard set of information
- Enable your dataset to be organized with other datasets

Examples:

- DDI (Data Documentation Initiative)
- Dublin Core
- FGDC (Federal Geographic Data Committee)

Capturing Metadata

- In a readme file
- In a spreadsheet
- In an XML file
- Into a database (when I share the data)

Document your workflow

- Workflow: how you get from raw data to the final product of research
- Documentation could be a flowchart or document
- Comment your code and scripts
- Well-commented code is easier
  - –to review
  - – share
  - –and use for repeat analysis

Time types

Time in the database is stored with one of two suffixes: `TIME` and `DATE`. If a column has `TIME` as the suffix, e.g. `CHARTTIME`, then the data resolution is down to the minute. If the column has `DATE` as the suffix, e.g. `CHARTDATE`, then the data resolution is down to the day. That means that measurements in a `CHARTDATE` column will always have 00:00:00 has the hour, minute, and second values. This does *not* mean it was recorded at midnight: it indicates that we do not have the exact time, only the date.

Date shifting

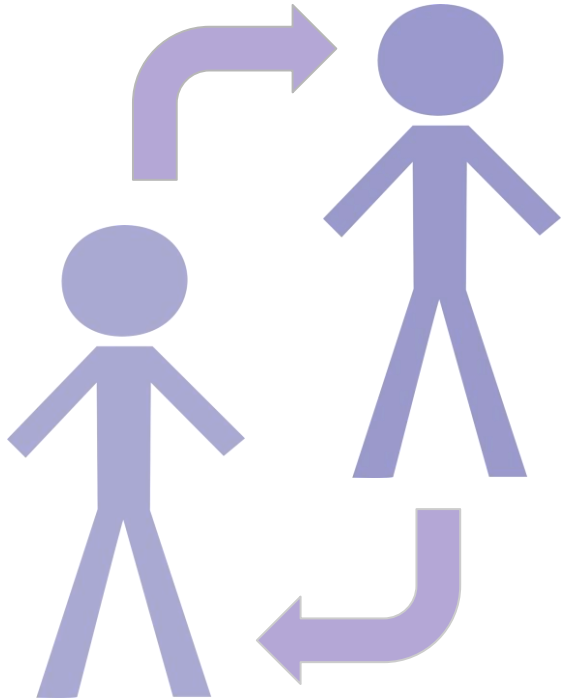
All dates in the database have been shifted to protect patient confidentiality. Dates will be internally consistent for the same patient, but randomly distributed in the future. Dates of birth which occur in the present time are *not* true dates of birth. Furthermore, dates of birth which occur before the year 1900 occur if the patient is older than 89. In these

Metadata: Best Practices

- Consistent data entry is important
  - Avoid extraneous punctuation & most abbreviations
  - Use templates, macros & existing standards when possible
  - Keep a data dictionary
- Extract pre-existing metadata
- Document production and analysis steps
- Consult a metadata librarian!

Setting Up for Sharing:

- Publishing
- Copyright / Licensing
- Citations
- Persistent IDs
- Private / Confidential data



Data Sharing: Options

- Individual request
- Personal website
- Publish as supplementary material
- Deposit in a repository
- Publish a data paper

Data Sharing: Publication

On your own:

- Pros:
  - Little up-front work
  - Allows for careful control of private/confidential data
- Cons:
  - Hard to find and/or access
  - Ongoing management burden
  - High risk for data loss

Data Sharing: Publication

Data as Supplementary Material:

- Pros:
  - Associates data with published articles
  - Provides a citable source
- Cons:
  - Limits to number and sizes of files
  - Possible format limitations
  - Reduced metadata

Data Sharing: Publication

Data repositories:

- Pros:
  - Allows addition of metadata to provide context
  - Subject-specific repositories collocate related data sets
  - Often provide archiving/long-term preservation services
- Cons:
  - Up-front work to submit data
  - Limitations on what can be submitted

More on repositories later...

Data Sharing: Publication

Data journals:

- Publish “data papers”
- Help make data sets discoverable and citable
- Peer-reviewed

Data Sharing: Publication

Data journal examples:

- *Scientific Data* http://www.nature.com/sdata/about
- *Journal of Chemical and Engineering Data* http://pubs.acs.org/journal/jceaax
- *Open Health Data* http://openhealthdata.metajnl.com/
- *Earth System Science Data* http://www.earth-system-science-data.net/
- And more...

Data Sharing: Copyright / Licensing

Type of Information	Copyrightable?
Raw data	No
Processed/cleaned data	No
Data in a creative visual representation (chart, graph)	Yes
Database	Maybe

Data Sharing: Citation

- Facilitates discovery of data
- Gives credit to the researcher
- Recognizes data as substantial output of the research process
- Allows for citation/impact analysis, as with article publications

Data Sharing: Citation

Important components:

- Creator/author
- Title
- Publisher
- Publication date
- Version
- Persistent ID

Data Sharing: Citation

Persistent identifier:

> "A unique web-compatible alphanumeric code that points to a resource (e.g., data set) that will be preserved for the long term (i.e., over several hardware and software generations)."[2]

[2] Hakala, J. Persistent identifiers – an overview. http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/

Data Sharing: Persistent IDs

- DOI - Digital Object Identifier
- ARK - Archival Resource Key
- Researcher identifier
  - ORCID - Open Researcher and Contributor ID

Data Sharing: Persistent IDs

- ORCID - Open Researcher and Contributor ID
  - Registry of researchers with unique identifiers
  - Name disambiguation helps with attribution
  - Supported by many publishers and repositories
  - Free to register at http://orcid.org/

Data Sharing: Citation

- Cite others' data properly
- Ensure that your data has sufficient information to be cited properly:
  - Creator, title, publisher, publication year, version
  - Persistent ID

Data Sharing: Managing Private / Confidential Data

Things to consider:

- de-identification / anonymization
- segregation of sensitive information
- adherence to relevant laws & policies

http://informatics.mit.edu/classes/managing-confidential-data

Long-term Storage:

- Definition
- Active Management
- Management Strategies
- Repositories

Long-term Storage

- What does “long-term” mean?
  - Two years?
  - Ten years?
  - Fifty years?

Long-term Storage

- Preservation = active management
  - Backup
  - Fixity checks
  - Format migration
  - Security/permissioning

Long-term Storage

- Backup
  - Multiple types of storage (spinning disk, tape, cloud servers)
  - Distributed across geographic locations
  - At least three copies

Long-term Storage

- Fixity checking
  - Generate and store checksums / cryptographic hash values for all files
  - MD5 and SHA-1 are common
  - Verify checksums regularly

Long-term Storage

- Format Migration
  - Obsolescence due to evolution of software
  - Reiterate: open, uncompressed formats!
  - Requires monitoring of formats over time

Long-term Storage

- Security
  - Physical space - access to storage hardware
  - Virtual space - permission controls
    - Access to read/use vs.
    - Write/edit

Long-term Storage

Management strategies

- Institutional resources
  - Backup services
  - Storage
- Grant/project funding
- Repositories: a great solution for many challenges!

Long-term Storage

Discipline-specific repositories

- Inter-university Consortium for Political and Social Research (ICPSR)
  http://www.icpsr.umich.edu
- Dryad - Scientific and medical data
  http://datadryad.org/

Long-term Storage

Find a repository:

- Registry of Research Data Repositories (re3data) http://www.re3data.org/

- MIT Libraries Data Management Services http://libraries.mit.edu/data-management/

Long-term Storage

Repositories: What to look for

- Open access
- Generates persistent IDs
- Good archival practices (Trusted Digital Repository certification)
- Flexible metadata
- Additional services (data cleanup, format migration/normalization, metadata assistance, etc.)

RES.STR-002 Data Management
Spring 2016