View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DigitalCommons@University of Nebraska

University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Library Conference Presentations and Speeches | Libraries at University of Nebraska-Lincoln

8-2014

Data Management Quick Guides

Jennifer L. Thoegersen
University of Nebraska-Lincoln, Jennifer.Thoegersen@oslomet.no

Follow this and additional works at: https://digitalcommons.unl.edu/library_talks

Part of the Library and Information Science Commons

Thoegersen, Jennifer L., "Data Management Quick Guides" (2014). *Library Conference Presentations and Speeches*. 103.
https://digitalcommons.unl.edu/library_talks/103

This Article is brought to you for free and open access by the Libraries at University of Nebraska-Lincoln at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Library Conference Presentations and Speeches by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

Data Management Quick Guide: Data & Metadata

July 1, 2014

Data Lifecycle

Consider data management at (and between) each stage of the life cycle
UK Data Archive Research Data Life Cycle Model: http://www.data-archive.ac.uk/create-manage/life-cycle

File Naming Conventions

General rules to follow:

- Avoid special characters ("/ \ : * ? " < > [ ] & $)
- Use underscores, not spaces
- Avoid names longer than 25 characters (supercalifragilisticexpialidocious)
- Use the ISO 6801 standards for date formats (YYYY-MM-DD)
- Use consistent versioning identification (DM_Guide_v03)
- Use names that describe the content

Define conventions and BE CONSISTENT

File Organization

Organize files into folders/subfolders that describe the general category of files
Can organize by project, date, researcher, location, etc.
Choose a structure that makes sense for the project and BE CONSISTENT

File Formats

Select formats that ensure the best change for long-term access to data
Favor commonly used and non-proprietary formats
Consider longevity, popularity, and potential for migration
Consider requirements of selected data repository
UK National Archives' **PRONOM:** http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx

- Provides detailed technical information about data file formats

File Format Recommendations/Preferences from:

- UK Data Archive: http://data-archive.ac.uk/create-manage/format/formats-table
- Library of Congress: http://www.digitalpreservation.gov/formats/content/content_categories.shtml
- Purdue University Research Repository: https://purr.purdue.edu/legal/file-format-recommendations

Metadata Standards

Select standards based on discipline

- Researcher might know standards
- If not, a place to start: http://www.dcc.ac.uk/resources/metadata-standards

Consider standards used by selected data repository

Controlled Vocabulary

Select based on discipline

- Researcher might know standards
- If not, a place to start: http://www.jiscdigitalmedia.ac.uk/guide/controlling-your-language-links-to-metadata-vocabularies/

Consider standards used by selected metadata standard and data repository

Data Dictionary

Provides a detailed description for each element or variable in a dataset and data model
Ensures consistent data entry and allows for future interpretation of data
Example: For a column in a spreadsheet, document meaning of column, allowable values, format of values, etc.

README.txt

Provides introductory documentation for a dataset

Data Management Quick Guide: Storage, Backup, & Security

July 8, 2014

Storage

Storage Option	The Good	The Bad
PC/Laptop	Convenient for active data	Lost/stolen; fail; manual backup
Network	Automatic backup & security	Access/capacity limitations
External devices	Low cost; portable; easy use	Lost/stolen; fail
Remote/Cloud	Global access; collaboration	Security/privacy limitations
Physical	Convenient; tangible	Manual backup

Backup

Allows for the restoration of data in the event that it is lost or compromised due to disaster, theft, hardware/software malfunctions, or unauthorized access. Vital for data that are unique or difficult/expensive to reproduce. Remember to create digital surrogates to backup analog materials

What?
Everything that would be required to restore data in event of loss (data/software/scripts/documentation)

How many?
Follow the Rule of 3: Original copy, second local copy, remote copy

How often?
Backup frequency is dependent on the project and the data. Consider how much data you would be willing to lose.

What type?
Full: Backup all files
Incremental: Backup only files that have changed since last backup (either full or incremental)
Differential: Backup only files that have changed since last full backup
For more details: http://support.microsoft.com/kb/136621

Test your system: Go through the exercise of accessing backup to see that procedure works & you can fully restore your data

Security

Access security ensures only authorized users can access data.
Utilize unique, role-based user IDs & passwords
Password tips:

- ✓ Consider length, complexity, variation, and uniqueness
- ✓ Include no personal information, sequences, or repetition
- ✓ Don't reuse passwords
- ✓ Balance difficulty to guess with difficulty to remember

Systems security is the protection of hardware and software.
Update anti-virus software, applications, and operating system and utilize firewall & intrusion detection
Control access to hardware (e.g. keep doors to office/server room locked)

Data Integrity ensures data has not been manipulated in an unauthorized way.
Encryption: Coding information that cannot be read/deciphered unless someone has the decoding key
Electronic signature: Coded message that is unique to both the document and the signer
Watermarking: Embeds digital marker for authorship verification & can alert someone of alterations

Data Management Quick Guide: Legal & Ethical Considerations

July 15, 2014

Intellectual Property

- "Intellectual property (IP) refers to creations of the mind: inventions, literary and artistic works, and symbols, names, images, and designs used in commerce." (World Intellectual Property Organization, wipo.int/about-ip/en/)
- Who owns the products of research is a complex issue that is dependent on many factors, including the funder.
- Generally speaking, for federally funded research, the university owns the data but allows the PI on the grant to be the steward of the data.
- The University of Nebraska Board of Regents Policy Section RP-4.4 (nebraska.edu/docs/board/RegentPolicies.pdf) outlines the institution's intellectual property policy.
- Contact the Office of Research Responsibility with questions concerning intellectual property (research.unl.edu/researchresponsibility/)

Retention

Retention concerns how long a researcher should retain data and is dependent on the agencies involved and whether there is personally identifiable data. The University of Wyoming Office of Research and Economic Development outlines the following steps (uwyo.edu/research/_files/docs/investigator%20requirements%20for%20retaining%20research%20data.pdf):

1. Determine which regulation(s) apply to your research (e.g. HIPAA, NIH, NSF Engineering Directorate)
2. Determine the time requirement (longest required; minimum of 3 years)
3. Determine what information to keep (Appraise and Select: dcc.ac.uk/resources/how-guides/appraise-select-data#5)

Ethics

- The Institutional Review Board (IRB) oversees research involving humans and ensures all applicable regulations are being followed (mediahub.unl.edu/media/3762).
- DMPs should include "…policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security…" (NSF Grant Proposal Guide, www.nsf.gov/pubs/policydocs/pappguide/nsf14001/gpgprint.pdf)
- **De-identified information** not only does not directly identify the individual (name, address, social security number), but should not either alone or in combination with other available data/knowledge allow an individual to be uniquely identified.
    - **Suppress:** Remove all identifiers that could be used to uniquely ID an individual
    - **Code:** Replace identifiers with a code that is kept separate from the data
    - **Generalize:** Make potentially identifiable data less specific (e.g. birth year instead of birth date, state instead of zip code)
    - **Aggregate:** Summarize data into related categories so individual records are
- The Office of Research Responsibility (http://research.unl.edu/researchresponsibility/) can help researchers ensure that they are meeting all federal, state and university guidelines.

Data Management Quick Guide: Sharing and Preservation

July 22, 2014

Data Sharing

Why share	Who benefits	Ways to share	When to share	What to share
Public investment	Researcher & team	Informal	During project	Raw data
Required by publishers/funders	Scientific communities	Supplemental	Immediately after project	Processed data
Inform new research	Students	Repository	Given time after project	Software/scripts used
Maximize transparency	Public			Documentation
Increase impact	Funding agencies			
Reduce duplication effort				
Provide credit to researcher				

Researchers should consider the legal and ethical issues involved in sharing (e.g. do they have consent to share participant data?). They should also consider the potential for reusability of their data, as well as whether outsiders will be able to understand the data. There are some potential drawbacks to sharing. Ensuring data is fit to share may be time-intensive. Others could misuse or misrepresent a dataset. Data released in the middle of a project may not have undergone sufficient quality assurance. There may be an overlap of publications if data are released during or immediately following a research project.

Data Reuse

When considering whether to reuse other researchers' data, determine whether the data is suitable for your purposes and, if so, determine the terms for reuse of the data. Properly cite the dataset in order to:

- Provide credit to data creators
- Assist in measuring impact of data
- Enable others to access the data
- Help researchers know *how* their data is being used

A **data citation** should include:

- Authors/Creators
- Publication data
- Title of dataset
- Publisher/Archive
- Version information
- Identifier/Locator (DOI/URL)

For more information on citing datasets, visit the Digital Curation Centre website: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

Repositories

For long-term preservation, datasets should be deposited in a data repository or archive. There are a wide array of **domain repositories** available, which accept data from specific subjects or domains. The following websites provide directories of repositories and are a great starting point for considering a domain repository:

- DataBibRepository List (http://www.databib.org)
- DataCite Repository List (www.datacite.org/repolist)
- Re3data (http://www.re3data.org)
- Open Access Directory (OAD) Data Repositories (http://oad.simmons.edu/oadwiki/Data_repositories)

If no suitable domain repository can be located, UNL Libraries hosts the **UNL Data Repository** (UNLDR), which provides researchers with a secure site for storage and long-term preservation of datasets that are no longer actively in use. UNL researchers can preserve up to 50 GB of data in UNLDR for free. Above that, there is a one-time fee (see https://dataregistry.unl.edu/ for details).

DataONE (https://www.dataone.org/best-practices/preserve) provides best practice guides on things like deciding what data to preserve, identifying data sensitivity and what data has long-term value.