Data management best practices

Properly managing data is of the upmost importance to ensure data quality and integrity, while also improving data access and discovery. To be successful, proper data management must be incorporated in all stages of data processing including collection, analysis, and long-term storage. The Washington University in St. Louis Libraries, has outlined step by step procedures for researchers to undertake to improve their data management techniques.

Recommendations:

1. Conduct a data asset inventory.
    a. Assess and inventory data collected
    b. Create a standard data intake form to document details on data collection entities and methods for current and future data acquisitions
2. Adopt best practices for managing data
    a. Create an intuitive folder structure and hierarchy that reflects the project goals
    b. Adopt and apply consistent file naming conventions for all files and folders
    c. Actively manage digital objects
    d. Establish a robust backup strategy (preferably 3 copies – 2 locally – 1 offsite)

1. Inventory data

An incremental, stepped approach to undertaking a data inventory is described below that can be applied flexibly according to context and needs.

1. Plan and define the purpose and scope of the inventory (determine which projects, whether or not to include off-site collaborators, what variables to include in the inventory, etc.).
2. Identify:
    a. what data assets exist,
    b. the values of the variables,
    c. where management could be improved,
    d. and any classifications of the data.

Free, open source software can help automate the file inventory and identification process.

FITS: http://projects.iq.harvard.edu/fits
BitCurator: http://wiki.bitcurator.net/?title=Software

Record the outputs of your inventory on a spreadsheet. Variables to track may include:

- Dataset location (folder path)
- Dataset File Name
- Date of data collection
- Project Name
- Data Type (interview recording, transcript, etc.)
- Data status (active, raw, analyzed, closed)
- Collection Method (instrument, survey, interview, etc.)
- Name of data collector

2. Data Management Best Practices

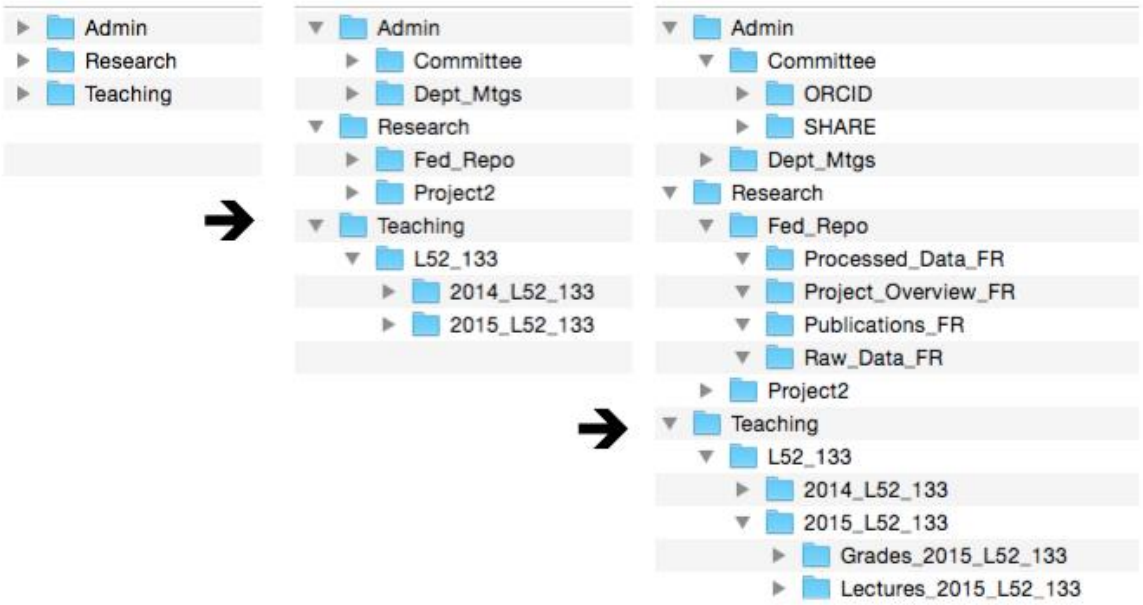
a. Folder organization

Use the information gathered in step 1 to develop a logical folder organization structure/hierarchy. This structure should reflect the manner in which data is accessed and collected.

In general, there are three ways to organize folders:

1. Object type – e.g. Interview Data, Survey Data
2. Organization structure – e.g. Location, Department, Individual, Project
3. Combine – structure directory by organization then object type

In most cases, the combined structure is the preferred method to organize files, but it is ultimately up the needs of the project.



Additionally, you can organize your folders by:

- Project
- Researcher
- Date
- Research Notebook Number
- Sample Number
- Experiment type/instrument
- Data Type

Whatever structure is chosen, the most important part is to be consistent and apply it. This structure should also be documented in a readme.txt file.

b. Naming conventions

Folder and file naming conventions should be descriptive, unique and reflect the file content/sample. The file and folder names should also be portable, which means that if the folders or files within happen to be moved outside of a parent file/folder – they are sufficiently understandable by name alone. When deciding on project naming conventions, consider future dependencies, file formats, and order of analysis.

The same naming conventions apply to files. For example:

7509228.txt is not a helpful file name.

Instead:

Deomurari_Transript_Interview_OralHistory_20150101_ST_LOUIS_MO.txt

This descriptive file name tells the file user immediately that this is a text transcription of an interview with an individual name Demourari for the Oral History project that took place on January 1, 2015 in St. Louis, MO. The file user understood this all without opening the file.

Good names convey context about what the file contains by stating information such as:

- Experiment type
- Experiment number
- Researcher name or initials
- Sample type
- Sample number
- Analysis type
- Date
- Site name
- Version number

Additional tips:

- List versions alphanumerically, e.g. v1,v2,v3 rather than last, final, finalfinal, useTHISone
- Use numerical dates, e.g. YYYYMMDD rather than Dec15
- Some computers and statistical programs will not understand file names with uppercase letters, weird characters (/#?,) or spaces between words

Good, consistent naming conventions significantly improve the access of data both in time and effort. They are essential for proper data management, but they also must not be so cumbersome that they take extremely long to implement. The researcher must find a solution that works for them.

Whatever naming conventions the researcher decides upon, document the file naming convention format in a metadata readme.txt file and adhere to it.

Tools for automating the bulk changes of names of documents are available freely online, and the Libraries are happy to recommend a program based upon your needs and operating system.

c. Manage Digital Objects

Spreadsheet best practices

1. Keep the raw data (unprocessed and unanalyzed) on a separate tab or spreadsheet. All manipulations should be carried out on the other tabs or in other files to ensure that the original data are not lost.
2. Only put one type of a data in any given cell. For example, rather than putting a location in a cell as "St. Louis, MO", this should be broken into two cells: one for city (St. Louis) and one for state (MO).
3. Consider breaking spreadsheets into discrete tables to mimic a relational database. Rather than having all the contact information for a location (site name, site number, latitude, longitude, village, district, state) directly in the data table, move this information to a separate table designed for describing location. Then you can use the site number to refer to the location in the data table.

Data dictionary

In simplest terms a data dictionary is a text file that records the definitions (semantics) for all the variables used in a database or spreadsheet. A data dictionary may also include detailed documentation about the relationships among metadata elements, as well as syntax and schema application rules. The term data dictionary comes from the relational database community and may be viewed as a type of metadata specification[1].

A sample data dictionary readme.txt file is available in Appendix B. Generally, a data dictionary includes an overall description of the data along with more detailed descriptions of each variable, such as[2]:

- Variable name
- Variable meaning
- Variable units
- Variable format
- Variable coding values and meanings
- Known issues with the data (systematic errors, missing values, etc.)
- Relationship to other variables
- Null value indicator
- Anything else someone needs to know to better understand the data

Metadata

Documentation should also be developed for the collection at various levels of granularity. At the minimum, basic elements of the project should be recorded at the collection level answering the questions of who, what, where, when and why.

A collection level metadata record should include:

[1] Drake, M. A. (2003). Metadata in the World Wide Web in Encyclopedia of library and information science. 2nd ed. / New York: Marcel Dekker.

[2] "Data Dictionaries," a blog post from Kristin Briney (dataabinitio.com)

- Author/Creator
- Project title
- Author contact information
- Date created
- Rights for reuse
- Preferred citation
- Project abstract
- Related publications/research outputs
- File inventory (naming conventions)
- Folder structure

Additionally, it is highly recommend to create additional metadata at the individual object level. This read-me file should include:

- Author/Creator
- Project title
- Date recorded/Date created
- Rights for reuse
- Project abstract
- Name of individual
- Other contextual information

d. Backup Strategy

And once you've gotten all this data managed properly, last thing you want to have happen is for it to all disappear. Following a robust storage and backup plan is the first step to ensuring that your data sticks around for as long as you need it to, and then into the future.

Store copies of all files and folders in three distinct locations.

- Network drive
- External storage device
- Geographically disparate location

Make sure that you have regularly scheduled backups in place, AND DO THEM. Again back up to an external device and/or a hosted service. Check to make sure your backups are indeed running.

Create checksums for all of your files and run checks against those checksums to ensure FES data/files do not encounter any digital bit rot or file corruption. A great free MD5 & SHA checksum generator and checking tool may be found at: http://download.cnet.com/MD5-SHA-Checksum-Utility/3000-2092_4-10911445.html.

Check the files every few years to ensure that they are there, accessible, you can open and/or execute them, and that they are uncorrupted. If you do notice corruption, data loss, or any other problems take action immediately to transfer and refresh the data.

Additional resources:

- *Data Asset Framework Implementation Guide* (October 2009)
- *Research Data Management* A Primer Publication of the National Information Standards Organization (July 2015).

Appendix A: Sample Collection ReadMe File

Please complete this template ReadMe file for each data package you have, and save as a TEXT file named "README_{Metadata/Dataset title}_{Todays date}.txt" (ex. "ReadMe_FedRepo_20080131").

The outline below should be completed with information relevant to the data package or data file:

Required information:

1. Author/Creator
2. Project title
3. Author contact information
4. Date created
5. Rights for reuse
6. File names, directory structure (for zipped files) and brief description of each file or file type
7. Definitions of acronyms, site abbreviations, or other project-specific designations used in the data file names or documentation files, if applicable
8. Method(s) for processing data, if data other than raw data are being contributed
9. Specialized software (including version number) used to prepare, to analyze and/or needed to read the dataset, if applicable
10. Quality assurance and quality control that have been applied, if applicable
11. Known problems that limit the data's use or other caveats (e.g., uncertainty, sampling problems, blanks, QC samples)
12. Date dataset was last modified
13. Relationships with any ancillary datasets outside of this dataset, if applicable

Appendix B: Sample Data Management Plan

Data format –

Whenever possible, open data formats or formats that do not use closed proprietary specifications will be adopted as standards for the project. For example, all text/data will be encoded using Unicode to prevent data loss. Uncompressed TIF (or comparable) will be used for all images. Well-documented industry standard video formats, such as MPEG-2 will be used to archiving. Digital geographic data may be stored in the defacto standard ESRI Shapefiles, or emerging specifications such as GeoJSON.

The following file naming conventions will be applied:

Datatype_DataStage_Location_Date_ProjectAcronym

The following folder organization will be applied:

>Project Title
>>Year1
>>>Location1
>>>>>DataType1
>>>>>DataType2
>>>Location2
>>>>>DataType1
>>>>>DataType2
>>Year2
>>>Location1
>>>>>DataType1
>>>>>DataType2
>>>Location2
>>>>>DataType1
>>>>>DataType2

Archival copies and originals of the data will be maintained according to the WU Libraries archiving policies outlined below.

Metadata – Metadata will be created and saved throughout the lifecycle of the research project and will be in line with the commonly accepted scientific community standards. All collections will have a DublinCore metadata record created. This record includes elements the following elements:

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source Language
- Relation
- Coverage
- Rights

As appropriate, the Data Documentation Initiative (DDI) metadata standard will be used to describe data granules (individual files rather than collections) and ISO19115 compliant metadata will be generated to describe spatial and temporal data elements. For continued preservation of the data and materials, metadata elements consistent with the PREMIS data dictionary and data model will be implemented.

A "read-me" file will also be created that includes: a dataset inventory, the DublinCore metadata information, contact information, specialized software used for analyzing/viewing the data, definitions of acronyms, data limits, methods for processing the data, a recommended citation, and an abstract of the project.

Access to data and data sharing practices & policies – Resulting unique research data and supporting documentation will be made accessible to the public via the Washington University data repository. Additionally, metadata records are propagated to metadata harvesters, such as SHARE (share-research.org) and DataCite through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

The dissemination information package (DIP) will include an access copy of the dataset, any required analysis code or software, a data dictionary/code book, the DublinCore metadata record(s), the discipline metadata record(s) and the "read-me" file. A digital object identifier (DOI) will also be assigned to the collection or at necessary granularities of the data.

Archiving of data – Data archiving is supported through an approach that starts with adequate documentation of data using metadata and other formats appropriate for long-term preservation. Unique digital research data from this project will be deposited in the WUSTL Libraries digital preservation repository and archived according to the Open Archival Information System (OAIS) framework.

The archival information package (AIP) will contain all the materials previously mentioned in the DIP, a copy of the untreated, original submission information package, the PREMIS metadata, a checksum manifest, and a curation treatment actions file.

Selection and Retention – The PI, and/or the core research team members, will select and prepare the research data for deposit. WUSTL Libraries follows a model of forward migration and will archive the dataset and its documentation for the long term, migrating the data through changing technologies, new media, and data formats for a minimum of 10 years.

Storage and Backup – The AIP will be backed up regularly, according to WUSTL Libraries practices and policies. Copies of the archival versions of the data will be stored at two locations one locally on a WU Library server and one off-site at the WU IT data center.