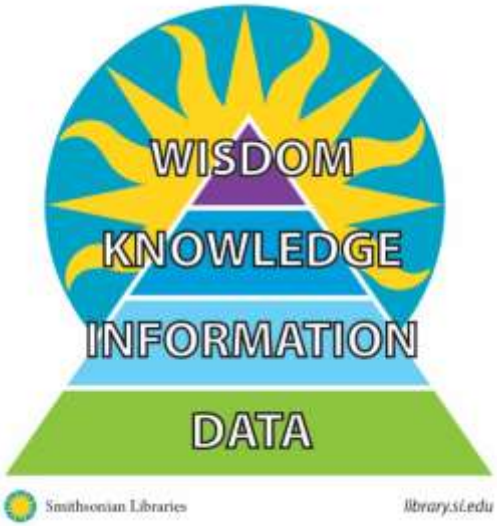
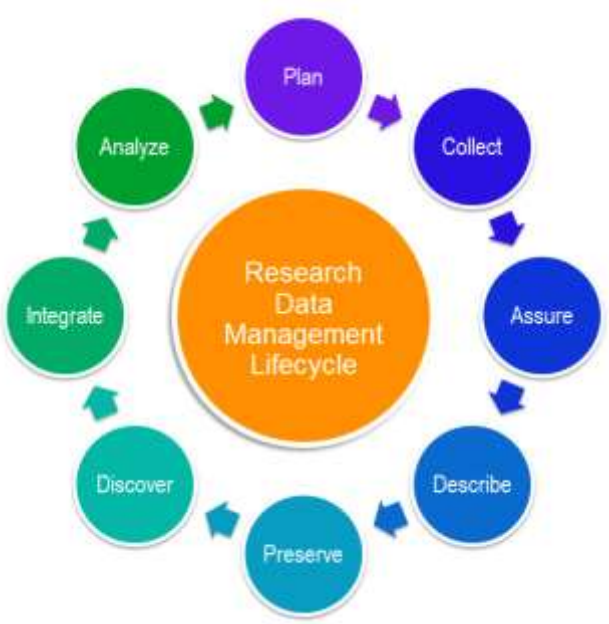
Research Data Management Best Practices



INTRODUCTION

The following best practices are intended for use by Smithsonian researchers and affiliated staff who plan for, create, and/or work with digital research data.
Additional information about available tools, policies, and resources for managing research data can be found on https://library.si.edu/research/data-management .



There are many phases in the research data lifecycle and they do not always occur in the tidy order pictured in the diagram (left). These best practices are designed to improve overall management of data at each point in the lifecycle, resulting in published data that are not only easy to care for long after the project is complete, but that are also findable, accessible, interoperable, and reusable.

The Smithsonian has other resources that can contribute to effective data management including software and high performance computing provided by OCIO, and training and planning consultation services from the Libraries. Contact AskaLibrarian@si.edu for more information.

SI also has two locally managed repositories that accept research data for publication and/or archiving:

- **SIdora** – is best for larger, or more complicated datasets, including actively updated datasets.
  →To deposit data in SIdora contact Beth Stern or email si-sidora@si.edu

- **Smithsonian Research Online (SRO)** – is best for smaller (<50GB), fixed (inactive) datasets that accompany or support publications deposited in SRO.
  →To deposit data and publications in SRO, you can self-deposit using the forms found here http://staff.research.si.edu/input_forms.cfm or contact research-online@si.edu

Actively managing your data throughout the research process enables reproducibility, reusability, and discovery, and can help maximize the impact of your research into the future.

PLANNING & DATA MANAGEMENT PLANS

Many granting agencies, such as NSF and the Alfred P. Sloan Foundation, require a formal **data management plan (DMP)** as part of a grant proposal.

Even if a granting agency does not require a DMP, SI strongly recommends that PIs create a planning document before starting any project that will create **digital research data**. DMPs are valuable tools for addressing issues that affect not only collection and use, but also the long-term viability of your data.

A written data management plan can:

- provide **continuity** on projects if staff join or leave
- allow for future **validation** or reproduction of results
- enable **reuse** of your data in potentially novel ways

SI Libraries staff are available for consultation on creating DMPs and are happy to review draft DMPs before submission with a proposal. Contact AskaLibrarian@si.edu for more information.

Proposals

The Smithsonian Office of Sponsored Projects (OSP) provides administrative and financial services for externally funded grants and contracts, and is available to assist PIs with technical and procedural questions related to managing grants and awards.

OSP also provides training in proposal development, writing and editing, and compliance oversight for areas such as Institutional Animal Care & Use, Export Control, Human Subjects in Research, and Responsible Conduct of Research. Their list of online and in person learning opportunities are available on their PRISM site.

Planning checklist

Any plan should *at a minimum* answer the following questions in **bold** for each stage in the data management lifecycle. More specific guidance for questions in the data collection, publishing, and archiving stages is available at https://library.si.edu/research/data-management

Proposal/planning Stage

- ☐ **What type of data is being collected/generated?**
- ☐ **Who is involved in data collection?**
- ☐ **Who "owns" the rights to the data?**
- ☐ **Are there restrictions on sharing and reuse?**
- ☐ **Are there applicable institutional policies on how the data is handled, shared, or archived?**
- ☐ Who will be using the data?
- ☐ If a collaborative project, are there MOUs that define roles and responsibilities?
- ☐ How do the outcomes need to be reported, e.g., to a sponsor or publisher?

Data Collection Stage

- ☐ **How will data be acquired/collected?**
- ☐ **What metadata standards and schema will be used?**
- ☐ **What are the file and data field naming conventions?**
- ☐ **What are the temporary storage requirements (size,cost,media)?**
- ☐ **How, where, and how frequently will data be backed up?**
- ☐ Are there existing standards for data structure and vocabularies, or will they be developed?

- ☐ Are there existing workflows for collecting, processing, describing, and storing the data, or will they need to be developed?
- ☐ Is there a data model for the project?
- ☐ Will your data be versioned, and if so, how will versioning be handled?
- ☐ What is your quality assurance/quality control process?

Publishing Stage

- ☐ **What repository or platform will be used to share the data?**
- ☐ **Who will be responsible for deposit and archiving after the project ends?**
- ☐ **If the data is to be shared publicly, what license should be applied? Are there any use restrictions?**
- ☐ If the data is embargoed, what is the embargo period, and who will manage it?
- ☐ If the data is not public, how will access be restricted?
- ☐ What costs are associated with publishing?
- ☐ What unique identifier will be assigned to the data (DOI, etc.)?

Archiving Stage

- ☐ **Who is responsible for preserving the datasets in the future?**
- ☐ **What data should be retained?**
- ☐ **Where will the data be archived?**
- ☐ **How much storage will be needed?**
- ☐ **How long should the data be maintained, and why?**
- ☐ **What are the risks for future access to the data, i.e., proprietary file formats, specialty software needed to interpret, password-protected systems?**
- ☐ How should the data be maintained in the future?
- ☐ Is there a cost associated with archiving the data?
- ☐ How will the data be found?

Funder-specific DMP Requirements

Some funding agencies require that plans submitted with grant proposals include specific elements or specific formatting. Below is a list of links to those requirements, alphabetical by funder, for selected granting organizations.
* = sample plans available on their site

- Alfred P. Sloan Foundation
- BCO-DMO NSF OCE: Biological and Chemical Oceanography
- Department of Energy – DOE: Generic
- * Gordon and Betty Moore Foundation (pdf)
- Institute for Museum and Library Services IMLS : guidelines for datasets (Word doc)
- * National Aeronautics and Space Administration NASA
- * National Endowment for the Humanities NEH-ODH (pdf)
- National Oceanic and Atmospheric Administration NOAA
- * National Science Foundation NSF-Generic DMP
    - NSF-Atmospheric and Geo
    - NSF-Astronomy (pdf)
    - NSF-Biology (pdf)
    - NSF-Earth Sciences
    - NSF-Education and Human Resources (pdf)

- United States Geological Survey USGS

Tools and templates

Smithsonian Specific Templates

The Data Management Team has developed boilerplate (temporarily located on an internal Confluence site) that can be used when applying for an **NSF** grant if you plan to deposit data either in **SRO** or **SIDora.** The boilerplate address the specifics of data archiving, dissemination, policies, and roles and responsibilities within the SI data management ecosystem.

DMPTool

One of the major tools for creating data management plans is the DMPTool, hosted by the University of California Curation Center (UC3 The Smithsonian was one of the original partner institutions involved in creating the DMPTool.

The DMPTool website includes templates and requirements for a large number of granting bodies, including NSF, DOE and NIH.
Any researcher at the Smithsonian can create an account and login to the DMPTool by selecting "Smithsonian Institution" from the list of institutions and then using their **SI network** username and password.

Digital Curation Centre Checklist

The Data Curation Centre's (DCC) Checklist can help you craft a custom DMP. The Checklist covers the main elements of a good plan, most of which are listed above, with suggested content for each element.

NAMING AND ORGANIZING YOUR FILES

Name and organize your files in a way that indicates their **contents** and specifies any **relationships** to other files.

The five precepts of file naming and organization:

- Have a **distinctive**, **human-readable** name that gives an indication of the content.
- Follow a **consistent pattern** that is machine-friendly.
- Organize files into **directories** (when necessary) that follow a consistent pattern.
- **Avoid repetition** of semantic elements among file and directory names.
- Have a **file extension** that matches the file format (no changing extensions!)

File naming

A file name should enable **disambiguation** among similar files and, for large numbers of files that make up a dataset, facilitate sorting and reviewing. Ideally, file names should be **unique**.

Keep in mind that files can be moved and, without the inherited folder structure, important descriptive information about the contents could be lost. Consider whether a filename would be meaningful outside of your chosen directory structure, and if not, how important the loss of that context would be, e.g., if the date a file was created is important, include it in the filename rather than just the directory name.

To provide a description of the file contents in the name itself, you should include elements such as:

- a **date**, or at least the year, the contents of the file were created, in the **YYYYMMDD** format (four digit year, two digit month, two digit day.)
    - start the filename with the date if it is important to store or sort files in chronological order.
- the **project name**, or documented abbreviation for the project.
- an **accession or other standard record number** if data is based on or includes SD-600 collections.
- your **organization's name** or abbreviation (if files are to be shared among collaborators.)
- the **location** related to the contents of the file, such as city, research site, etc.
- a **version number**, prefaced by "v", or another indicator of the file content's status such as "_draft" "_final" or similar.
    - Avoid *starting* the filename with version number, "draft" or "final"
- an **ordinal number padded with zeros** (particularly if the file needs to be sequenced/sorted with many other files).
    - use a minimum of two zeros for padding, with as many as necessary to accommodate the quantity of files you expect, e.g., if you expect 1,200 data files from one instrument, pad the filenames with three zeros, starting with _0001

Filenames for any given project or program should follow a **consistent pattern**. Adopt a pattern that will enable you to make filenames unique within each project, and are machine-friendly.

- Omit spaces and punctuation other than **hyphen** and **underscore.**
- Use **underscore** or "**camelCase**" between file name elements, e.g., my_data_file.txt or myDataFile.txt . Neither approach is better - just choose *one* and stick to it!
- Do not use spaces, tabs, semicolons or periods to separate elements of a filename.
- Try to use only **ASCII-encoded alphanumeric characters**, e.g., letters found in the Latin alphabet, and numbers between 1 and 10.
- Limit the name to **25 characters** in length if possible. **Short but meaningful is best.**

Examples of well-formed file names

1. For an image of a specimen in the Fishes collection, NMNH, collected in Mindoro, Philippines in 2000 with the catalog number USNM 379221 (3 options):

    a. 2000_USNM_379221_01.tiff
    b. USNM_379221_01.tiff
    c. PHL2000USNM379221.tiff

2. A versioned file of tabular data and the accompanying data dictionary for a project in 2018 called "Multi-site cross-cutting longitudinal study" (two potential abbreviations for the project are given):

    a. 2018MSCCLSv1.txt ; 2018MSCCLSReadMe.txt
    b. MultisXxLong2018v1.txt ; MultisXxLongAbout.txt
    c. 2018_MSCCLS_v1.txt ; 2018_MSCCLS_readme.txt

Tip: You can bulk **rename** and manipulate files by scripting in the programming language of your choice, using PowerShell (Windows) or the Finder (Mac), or you can use an application like:

- Adobe Bridge
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/Main_Intro.php
- Renamer 5 for macOS: https://renamer.com/
- PSRenamer (requires JRE/JVM): http://www.powersurgepub.com/products/psrenamer/index.html

File Organization

Like file naming, **consistency** is key. Organize files in a way that makes sense within the context of your project, but would also make sense to someone who was not intimately familiar with your project.

How files are nested in directories can be dependent on the **number of files** you are working with, and what aspect of those files is **most important for analyzing** or re-using the information in them.
For instance, if you have hundreds of thousands of image files collected over many years from many different locations, you may want to organize first by year, then month, then location. You could also organize them entirely by date, and include the location in the filename. Alternatively, organize by location, and only include the date in the filename.

If you are working on a collaborative project, make sure all collaborators are using the same principles to organize and name files!

Examples of Directory Organization

Example 1: SI and UCSD are both contributing to a five year project that involves taking measurements over time on two sample materials, A and B. Submitted files are for analyzed rather than raw data, and UCSD is employing two methodologies for analysis, submitted as versions.
Because the date of the measurement is important, files are first named by date, then sample. Each are organized into directories by contributor, and further grouped at the top level by year.

- 2017
    - UCSD
        - 20171001_B_v1.csv
        - 20171001_B_v2.csv
        - 20170930_B_v1.csv
        - 20170930_B_v2.csv
    - SI
        - 20170930_B_SI.csv
        - 20170925_A_SI.csv

Example 2: Images and corresponding description of those images from various sites in Pennsylvania, taken over the course of several years. The researcher expects to have between 150-300 images per site per year. In this example, a text file with descriptive metadata for all the images taken on one day is stored in a separate directory. This metadata file could also be co-located with the images.

- 2017_Images
    - Philadelphia
        - phil_20171028_001.tiff
        - phil_20171028_002.tiff
        - phil_20171028_003.tiff
        - phil_20171029_001.tiff
        - phil_20171029_002.tiff
    - Pittsburgh
        - pitt_20170922_001.tiff
        - pitt_20170922_002.tiff
        - pitt_20170922_003.tiff
- 2017_Metadata
    - Philadelphia
        - phil_20171028.txt
        - phil_20171029.txt

References

Briney, Kristin. 2015. Data management for researchers: organize, maintain and share your data for research success.

NIST. 2016. *Electronic File Organization Tips* https://www.nist.gov/sites/default/files/documents/pml/wmd/labmetrology/ElectronicFileOrganizationTips-2016-03.pdf

Purdue Library. 2017. *Data Management for Undergraduate Researchers: File Naming Conventions*. http://guides.lib.purdue.edu/c.php?g=353013&p=2378293

Stanford Libraries. (viewed 2018). *Best practices for file naming*. http://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

University of Edinburgh. 2007. Records Management: Naming Conventions. https://www.ed.ac.uk/records-management/records-management/staff-guidance/electronic-records/naming-conventions

CHOOSING FILE FORMATS

Support for a particular file format now does not guarantee readability in the future. If your data is in a proprietary format, or is dependent on specialty software or hardware to read it, consider converting it to one of the more sustainable file types below and deposit your data in both its original format plus an equivalent sustainable format.

In general, save data for archiving in a non-proprietary, platform-independent, unencrypted, lossless, uncompressed, commonly used file format.

If you need to store multiple files in one container, choose an **uncompressed TAR, GZIP, or ZIP**.

File Type	Prefer	Avoid
Text	PDF/A-1 or PDF/A-2, OpenDocument Text Format (.odt), Plain Text (.txt - using ASCII or Unicode encoding)	WordPerfect, Microsoft Word
Structural markup	SGML with DTD, XML with DTD	SGML or XML without DTD
Tabular data	comma (.csv) or tab-delimited (.txt or .tab) file, OpenDocument Format Spreadsheet (.ods)	
Database	comma or tab delimited flat file(s) (.csv or .txt), XML with DTD, JSON, OpenDocument Format Spreadsheet (.ods)	MSAccess, FileMaker Pro
Geospatial*	GeoTIFF, Geographic Markup Language (GML) or GML in JP2, Keyhole Markup Language (KML)	
Audio	Broadcast Wave Format (BWF – an extension of WAVE), Free lossless codec of the Ogg project (FLAC), Audio Interchange File Format (AIFF), MP3 (uncompressed), Microsoft Wave (WAV)	
Video	AVI, Quicktime (MOV), Windows Media Video (WMV), MPEG4 or MPEG-2 Video, Material Exchange Format – lossless (MXF)	Raw, Flash
Still image	TIFF, JPEG2000, PNG, PDF/A	Camera RAW
Presentation	PDF/A, PDF	

*for Geospatial files, ESRI Shapefiles and ESRI ARC files are acceptable, but not preferred according to NARA.

References

(accessed 2018-02-23) Wikipedia. List of file formats. https://en.wikipedia.org/wiki/List_of_file_formats

2014. NARA Records Management Regulations, Policy, and Guidance Appendix A: Tables of File Formats https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

(accessed 2018). Library of Congress. Recommended Formats Statement. http://www.loc.gov/preservation/resources/rfs/data.html

2016. Recommended Preservation Formats for Electronic Records.https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records

WORKING WITH TABULAR DATA

Putting data into simple tables is one of the most common ways to store and then work with data. Below are some basic principles for organizing data into tables so that both humans and machines can use that data.

If your project or group already follows a convention for putting data in tabular form, always follow that convention. However, if that convention is significantly different from one of the guidelines below, *consider retaining a copy of the data in its original form as well as a normalized version* that conforms to SI guidelines.

General guidelines

- If possible, store tabular data in a non-proprietary file format such as comma-delimited .csv or tab-delimited .txt files.
- Do not rely on special formatting such as cell colors, text bolding, or other visual cues to provide meaning.
- Do not include figures, analyses, or charts.
- If your tables contain formulas or macros, create a second copy at the end of the project that contains only the results of those formulas.
- If possible, use only Latin (English) alphanumeric characters (a-z and 1-10) in data and headers. Avoid the use of commas in data if possible.
    - If your data includes non-alphanumeric characters, e.g., letters with diacritics (accents), always check your data to see that it has been correctly interpreted when you open or reuse the file in different software applications.
    - If your data must include commas, and you are saving the file as comma or tab delimited, make sure to qualify or "escape" the data between the columns by adding double quotes around the data values.

Rules for Rows

- A few files with many rows is preferable to many files with few rows. However, you may want to consider splitting files with more than 1,000,000 rows or 15,000 columns depending on what programs are typically used to read the data.
- Each row in your file should represent a single record or data point, e.g., the measurements of one sample or the response of one individual.
- The first row in the table should be reserved for column headers, aka field names.
    - Each column header should be concise but meaningful, contain only alphanumeric characters (with the addition of hyphens or underscores if necessary) and should never be duplicated in the same table.
    - If possible (and relevant), include units of measurement in the column header.

Data Standardization

- Standardize the format of data within each column, e.g., calculate numerals to a set decimal place.
- Use international, e.g., ISO; national, e.g., FGDC; or field-specific, e.g., LCSH; standards when collecting common types of data.
    - Use the ISO standard for recording dates – four digit year first, then two digit month, then day, e.g., 2018-01-31
- Decide on a consistent way to indicate missing data, and stick with that convention! Document that convention in your data dictionary (see **Describing Your Data: Data Dictionaries**)
    - Common ways to indicate missing data is to use a code such as -999 or -9999, or use text like "missing"
    - Always check to make sure that your missing data is interpreted correctly in any software you use to analyze or process it.
- Provide a data dictionary that explains the contents of your tabular files and gives additional context, including what any abbreviations mean, the units used, and any standards followed.
    - The data dictionary should be named similarly to the data file (see file naming best practices). If you use Excel and want to keep the data dictionary as a separate "tab" in the file, that is acceptable, but be

aware that other software applications may not be able to correctly interpret the relationship between the contents of the two tabs.

Examples

Fig. 1 Well-formatted tabular data

Site	Ecosystem	Plot	Depth_cm	Section_length_cm	Total_core_length_cm	Percent_LOI	Percent_TC
Al Aryam	salt marsh	6	30-50	20	-9999	5.818	10.26
Al Aryam	salt marsh	6	50-81	31	-9999	3.813	10.58
Eastern Mangrove	salt marsh	1	0-15	15	85	4.861	11.14

Fig. 2 Poorly formatted tabular data

						Data collated by Dr. R.E. Searcher 1/10	
Soil carbon data							
	Ecosystem	Plot #	Depth	Sect length	core	LOI	%TC
Al Aryam	salt marsh	6	30-50	20		5.82	10.26
Al Aryam	sm	6	50-81	31		3.813	10.58
Eastern Mangrove	salt marsh	1	0-15	15	85	4.861	11.14

References

2007. Best Practices for Preparing Environmental Data Sets to Share and Archive. Hook, L.A., Beaty, T.W., Santhana-Vannan, S., Baskaran, L., & Cook, R.B. http://daac.ornl.gov/PI/bestprac.html

2009. Some Simple Guidelines for Effective Data Management. Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. Bull. Ecol. Soc. Am. 90(2)205-214. http://www.nceas.ucsb.edu/files/news/ESAdatamng09.pdf

Preparing tabular data for description and archiving. Cornell University Research Data Management Service Group. https://data.research.cornell.edu/content/tabular-data

Expressing intentional blanks (null values) in a tabular dataset. DataOne. http://www.dataone.org/best-practices/identify-missing-values-and-define-missing-value-codes

2017. Ecology Tutorial: Data Organization in Spreadsheets. Data Carpentry. http://www.datacarpentry.org/spreadsheet-ecology-lesson/

DESCRIBING YOUR DATA: DATA DICTIONARIES

A "**data dictionary**" or a "**readme**" file includes crucial information about your data that ensures it can be correctly interpreted and re-used by yourself, possible collaborators, and other researchers in the future. Depending on the nature of your datasets, it may include collection methods or any processing/calculations that were applied to the dataset as a whole or to specific data elements.

> *" The increased use of data processing and electronic data interchange heavily relies on accurate, reliable, controllable, and verifiable data recorded in databases. One of the prerequisites for a correct and proper use and interpretation of data is that both users and owners of data have a common understanding of the meaning and descriptive characteristics (e.g., representation) of that data. To guarantee this shared view, a number of basic attributes has to be defined."*
> *-International Standards Organization (ISO) Information Technology Parts 1-6 (2nd Edition),2004.*

If the data you are describing is primarily tabular, the description could be in a tabular form as well. In most cases, you should create your dictionary as a **plain text file** with an introduction giving basic information about the dataset, followed by detailed definitions for each element in the dataset.

- Create one descriptive file for each dataset.
- Name the dataset, data dictionary, and any other supporting files similarly. (see **File naming**)
- See **Working with Tabular Data** for more details

Follow the conventions of your discipline when choosing standardized terms or when structuring your data, e.g., use USGS Thesauri terms for Earth science data, or Darwin Core for Natural History collections.

You should also provide sufficient metadata to cite your dataset (see **Describing Your Project: Citation Metadata**). This information may be included in the data dictionary, or be stored separately in another file, or as a metadata record in a repository.

A basic overall definition of the data should be at the beginning of your data dictionary.

Basic dataset introduction *must* include:

- Who collected or aggregated that data, or in the case of many contributors, who is the principal investigator or contact.
- When the data was collected.
- What the data elements are measuring or describing.
- Why the data was collected.
- Methodologies used or assumptions made while collecting the data.

Additional definition *should* include when relevant:

- Description of any transformations or calculations applied to the raw data (if the data being described is not the raw data) or to specific data elements, including references to any scripts used.
- Version.
- Any validation or quality control process that has been applied.

After the overall definition, describe each component or element of your data. If your data is tabular, describe each column (field) and what it should contain. If your data includes images, describe how they are organized, and where detailed metadata can be found.

Data element description may include:

- Element name as found in the dataset, i.e., the data label or column header.

- A full “human readable” name of the element if the dataset uses codes or abbreviations.
- A definition of the data element.
- Any units of measure and precision (if applicable), e.g., “measured in meters, rounded up to the nearest .01 meter”.
- The format of the data element (if applicable), such as integer, text, date-time, etc.
- All valid/allowed values.
- Any codes, symbols or abbreviations used in the values themselves.
- If the element is required or not.
- The source of the controlled vocabulary or thesaurus used (if applicable).
- The source of the data element, e.g., sensor, observation, etc.
- If the element is “null”(for a non-required element) the convention for how that is represented, e.g., “unk”, "-999999"

Example

File 1: Amendment seed packets and fungi_all.txt

This datafile includes the numbers of protocorms recovered from seedpackets exposed to amendment with different organic amendments, compared to no amendment. Data were collected 2010-04-02 and 2010-04-08 with results published in the paper “title of paper.” Missing data are indicated by a “.”. Data were collected by M------- and R-------. Questions should be directed to M--------.

Column headings:

Species: The orchid species of seeds added to the plot in seedpacket. Goodyera=Goodyera pubescens; Liparis=Liparis liliifolia; Tipularia=Tipularia discolor

Site: Designated numerically 1-6. All sites are forest stands at the Smithsonian Environmental Research Center, Edgewater, Maryland, USA. Sites 1-3 are old stands and 4-6 are young stands (see Siteage, below).

Subplot: Designates the subplot location within each site. Thirty-six subplots were arranged in a square with columns labeled A-F and rows labeled 1-6.

Siteage: Old=120-150year old forest. Young-50-70year old forest.

Treatment: The amendment added to a subplot (Leaves=tulip poplar leaf litter; Wood=chipped fresh tulip poplar wood). Subplots with no amendment added are designated Control.

Inoculated?: Designates whether mycorrhizal host fungi were inoculated into the subplot.

fungusyn: Indicates whether appropriate host fungi were detected (1) or not (0) using PCR amplification of the soil in the subplot.

fungusInt: A semi-quantitative measure of the abundance of appropriate host fungi. The intensity of fluorescence by a post-PCR gel band 0=no band visible to 3=intensely bright fluorescence.

fung2YN: For Tipularia discolor, indicates whether an appropriate host fungus was detected (1) or not (0) using PCR amplification of the soil in the subplot using a second primer set (TipC2F/TipR) that detects an appropriate host fungus not detected by the first primer set (TipC1F/TipR).

References

2006. Northwest Environmental Data Network. Best Practices for Data Dictionary Definitions and Usage. http://www.pnamp.org/sites/default/files/best_practices_for_data_dictionary_definitions_and_usage_version_1.1_2006-11-14.pdf

Retrieved 2017. Open Science Framework. How to make a data dictionary. http://help.osf.io/m/bestpractices/l/618767-how-to-make-a-data-dictionary

2017. USGS. Data Management: Data Dictionaries and Thesauri. https://www2.usgs.gov/datamanagement/describe/dictionaries.php

DESCRIBING YOUR PROJECT: CITATION METADATA

The overall description for your project could be referred to as project metadata, citation metadata, a data record, a metadata record, or a dataset record. The information supplied in the project description should be sufficient to enable finding and properly citing your data.

An easy way to ensure you have supplied enough information in the citation record is to ask yourself if you have answered the "Who" "What" "Where" "When" and "How" of your project, and that you have included a persistent identifier. Avoid abbreviations and short hand. Remember that the description of your project may be read by someone outside your field of study or even yourself years later.

Always include:

- **Creator**/author(s) -- including complete names, institutional affiliations (including SI unit if depositing into an SI repository) and any ORCIDs
- **Title** -- a meaningful and descriptive title, prefaced with "Dataset:". Title can include a facility, or title of a larger dataset if the one you are describing is a derivative or subset of that dataset.
- **Publication Date** -- year (and if relevant, month and day) the data was made public, or if under embargo, the date the embargo expires. If data is restricted and not publicly available, use the date it was deposited.
- **Persistent Identifier/Location** -- a DOI is preferred, but a URN, Handle, EzID or ARK are acceptable. If no persistent identifier is available, a working URL/URN for the data is mandatory.

Include when possible:

- **Resource type** –the general format of your data, e.g., tabular data, database, audio files, sensor data, images, etc.
- **Publisher** – usually this will be the hosting location or organization with which you have deposited your data. You can use the institution or project name, or a URL or URN for the repository.
- **Grant** – either the name of the grant, e.g., "CLIR Hidden Collections 2017" or the grant number associated with the dataset
- **Abstract/Description** – an abstract for the dataset that covers who, what, where, when, why in a narrative format.
- **Preferred citation format –** MLA, APA, Chicago, etc.
- **Related publications** – this could be a published article, or related datasets, referenced with a URL or a DOI
- **Rights** – any licenses, intellectual property rights, and/or restrictions that should be applied to the data
- **Version** – a number increased when the data changes, e.g., through addition of data or re-running an analysis or derivation process.

Example:

[Dataset:] Templates for Statistical Resample Methods Maximize Accuracy and Efficiency of Colorimetric Data Collection for Monitoring Biocolonization on Stone. Perets, Ethan A.; Charola, A. Elena; Liu, Yun; Grissom, Carol; DePriest, Paula T.; Koestler, Robert J. 2016. Repository.si.edu. DOI: https://doi.org/10.5479/data.mci.2016.0629

Abstract:
Non-parametric and semi-parametric statistical approaches were developed to maximize accuracy of colorimetric data for monitoring biocolonization on stone surfaces, while simultaneously optimizing efficiency of data collection in the field. These approaches were applied to colorimetric data sets collected on three Kasota limestone capstones located at the National Museum of the American Indian in Washington, DC. Data was randomly resampled without replacement (the statistical "jackknife"), producing data subsets of diminishing resample sizes.

…. Factors affecting the necessary minimum sample size for achieving pre-selected confidence levels and acceptable measurement error – including the impacts of a biocide treatment and heterogeneity of surface textures – were also investigated. Comparison of results for textured capstones suggests that rougher stones require greater numbers of measurements at identical d and confidence. Corresponding author: Paula DePriest.

References

CrossRef. Required, Recommended and Optional Elements. https://support.crossref.org/hc/en-us/articles/213077846-Required-Recommended-and-Optional-Elements

DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. 10.5438/0014.

Ball, A. & Duke, M. (2015). 'How to Cite Datasets and Link to Publications'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

PREPARING FOR STORAGE AND PRESERVATION

Storage and Archiving

Research data and related files require reliable and trustworthy storage at all phases of the research process. Best practices include documenting the information below either in a **Data Management Plan**, or in project protocol documentation.

To provide that trustworthy storage during the **planning and active** phases of a research project be sure you can document:

- Data **ownership** and responsibility.
- Who has **access** to the original or raw data, and how access is restricted (password protected, networked server with limited user accounts, etc.) Restricting access to the original data reduces risk of inadvertently (or intentionally!) altering or deleting data.
    - When doing analysis, transformations, or other work, always **use a copy** of the original.
- Estimate of **storage space** needed, including for backup copies, and storage **media**.
    - Preferred storage media include "spinning disc" hard drives, solid state hard drives (SSD), magnetic tape (often found in large data centers), and thumb/jump drives.
    - Optical media, i.e., CDs and DVDs, are not good long-term storage options as they can degrade quickly and fewer machines are able to read them.
- Location and methodology for **backups**, including **schedule** of periodic backups.
    - There should be **two**, preferably three, backup copies.
    - Copies may be physically stored with the researcher, on a networked server, or in the cloud. Each backup copy should be in a **different location**, and/or on different media.
    - Periodically **verify** backup integrity (can you access and read the files).
- Estimate of storage **costs** over the course of the project. Is there a cost to get data out of storage, e.g., Amazon Glacier or similar cloud storage.
    - Media represents only a small part of total cost for storage in the long term. Maintaining content for preservation involves human resources as well.

All storage mechanisms are subject to failure and do not last forever. If managing your own data during a long-term project, plan to migrate data from one storage platform to another at least once to ensure that data remains accessible, and to prevent data loss due to media failure. If a third party is storing the data, migration to newer storage should be part of its service.

Before the **conclusion** of a project you should decide and document:

- **Where** the data will be archived, and who will be responsible for data stewardship and curation.
    - The archive or repository must have a **disaster recovery plan**.
    - The organization responsible for long-term archiving must have sufficient resources (staff and funding) to maintain the data for the requisite period of time.
- **Length of time** data must be maintained.
    - Data that cannot easily be reproduced, such as climate sensor data or data from destructive sampling, should probably be retained indefinitely.
    - Data that has legal value should be retained indefinitely or for a length of time stipulated by law or policy.
    - Data that has significant historical or research value should probably be retained for at least 20-50 years.
    - Data that could be re-created with better technologies in the future could be retained only until it can be re-created and replaced.
- **Accessibility** of the data. Is it available immediately or stored in a dark archive offline that might require some retrieval time, or incur costs for retrieval?

- If data is the result of a multi-institution collaboration, with archiving occurring in multiple places, that all data is accounted for. If you only are responsible for partial data, indicate where the other data is stored

Preservation

Preservation of data requires active management of that data over time to maintain accessibility, authenticity, and usability. Future data managers' ability to maintain your data is influenced by the format of the data, the extent and quality of description (metadata), and how the data was stored and managed during the project.

When thinking about preservation, be aware that *not all data should be retained forever*. Drafts of papers, duplicate copies, superseded versions of datasets, beta versions of software, and other working files are transitory in nature and should probably not be preserved indefinitely. Preservation is a resource-intensive activity allocating resources toward storage and preservation.

To ensure your data is viable in the future:

- **Choose open formats when possible.** Store data, or a copy of the original, in a non-proprietary, commonly used, open format. (see Choosing File Formats)
    - If any of your files are in a proprietary format, consider creating and depositing a second copy in an open format with it when archiving, e.g., if your metadata file is in MS Word, save a copy and convert to PDF/A and archive both with the data.
- **Always include sufficient metadata** with your data files. (see Describing Your Data: Data Dictionaries) This should include:
    - Any requirements from funders or publishers that mandate sharing the data for a specified period.
    - Documentation of what software (and its version) and hardware were used to create the data, and what quality assurance processes were done on the data. Keep in mind that in the future, data created with special software might require the use of an emulator to view or use it.
- **Follow good storage and data management practices during the project.** Ensure a responsible party is maintaining the data according to best practices, and has created sufficient documentation to provide continuity should they leave the project.

References

(accessed 2018-02-28). Smithsonian Institution Archives. Records Management at the Smithsonian Institution Archives - https://siarchives.si.edu/what-we-do/records-management
(accessed 2018-02-28). Digital Curation Centre. Why Preserve Digital Data? - http://www.dcc.ac.uk/digital-curation/why-preserve-digital-data
2017-10. University of Edinburgh. MANTRA – Research Data Management Training https://mantra.edina.ac.uk/

CHOOSING A REPOSITORY

Given the large number of specialty repositories that exist or are being built for specific data types, specific organisms, and large grant-funded collaborative projects, it is impractical to list all the data repositories that could be used by SI researchers to conform to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles.

Before depositing data in a repository not included in one of the lists below, the data owner *should at a minimum* ensure that the repository:

- Has a plan and sufficient funding to ensure its long-term viability.
- Allows export of data and data descriptions in a standards-compliant format, preferably identical to the format you deposited.

Ideally, the repository should also:

- Enable easy citation of your data, including supporting DOIs (either minted by the repository or by SIL).
- Be searchable, and indexed in a service such as DataCite or Elsevier's DataSearch.
- Support application of an appropriate license, and embargo of data if necessary.
- Support metadata standards for your data, e.g., ISO 19115 for Geographic data.

In general, any repository managed by a U.S. **Federal agency or national laboratory**, e.g., NIH's GenBank, NASA's National Space Science Data Center, or ORNL's DAAC should be considered a **preferred repository** for any SI research data that meets their criteria for deposit. In addition, **national data services** such as Australia's ANDS, or the Netherlands DANS-EASY are also acceptable, as are many data repositories run by established U.S. institutions such as Harvard's Dataverse.

General-purpose

SI has two repositories, SRO and SIdora, that accept Smithsonian-produced data.

Both SRO and SIdora support multiple file types, and are discipline agnostic. Both accommodate use of DOIs for citation. Both have actively managed, backed-up, secure storage in the Herndon Data Center. Both support having open (accessible) and closed (private) data, though SRO additionally supports embargoes and other restrictions.

- **SRO** – is best for smaller (<50GB), fixed (inactive) datasets that accompany or support publications deposited in SRO.
  →To deposit data and publications in SRO, you can self-deposit using the forms found here http://staff.research.si.edu/input_forms.cfm or contact research-online@si.edu

- **SIdora** – is best for larger, or more complicated datasets, including actively updated datasets
  →To deposit data in SIdora contact Beth Stern or email si-sidora@si.edu

If you or your publisher prefer to deposit with a non-SI repository, there are four recommended general-purpose repositories. Their features are compared below.

feature	Dryad	Figshare	Open Science Framework	Zenodo
	http://datadryad.org/pages/faq	https://support.figshare.com/	http://help.osf.io/	http://help.zenodo.org/

Fees/terms	$120/deposit	free, premium service for a fee	free	free, limited to 50GB/dataset
dataset licensing	CC-0 (Creative Commons 0 only)	CC-By, CC-0, MIT, GNU GPLv3,Apache 2.0	CC, Apache, MIT, GNU, other	CC, other
access options	open; embargoed (only for certain publishers)	open; restricted (unpublished)	open; restricted; closed	open; restricted; closed; embargoed
versioning available	yes	yes	yes	yes
formats accepted	office documents, scientific & statistical data, plain text, structured text, software, source code, other	office documents, images, structured graphics, audiovisual data, raw data, plain text, archived data	any	any
provides usage stats	yes	yes	yes	yes
persistent identifiers	will assign DOI	supports ORCID; will assign DOI	supports ORCID; will assign ARK and DOIwill assign ARK and DOI	supports ORCID; will assign DOI or use provided DOI

Discipline-specific

The following non-comprehensive list does *not include* repositories for model organisms, those managed by U.S. government agencies, or institutional data repositories such as Harvard's Dataverse or UIUC's Illinois Data Bank. *Omission from this list does not imply that it is not an acceptable repository for your data*. If you would like to see a repository listed here, contact Keri Thompson thompsonk@si.edu.

- Astronomy
    - SIMBAD http://simbad.u-strasbg.fr/simbad/
- Biodiversity/taxonomy
    - GBIF Global Biodiversity Information Facility https://www.gbif.org/
    - KNB Knowledge Network for Biocomplexity https://knb.ecoinformatics.org/
- Omics/sequencing
    - ArrayExpress https://www.ebi.ac.uk/arrayexpress/
    - BioGRID https://thebiogrid.org/
    - DGVA Database of Genomic Variants Archive https://www.ebi.ac.uk/dgva
    - NURSA Nuclear Receptor Signaling Atlas https://www.nursa.org/nursa/index.jsf
    - ProteomeXchange http://www.proteomexchange.org/
    - UniProt https://www.ebi.ac.uk/uniprot
- Environment/Climate
    - NERC Environmental Information Data Centre (multiple repositories) http://www.nerc.ac.uk/research/sites/data/
    - Environmental Data Initiative https://portal.edirepository.org/nis/home.jsp
    - World Data Center for Climate (at DRKZ) https://www.dkrz.de/up/systems/wdcc

    - PANGAEA https://www.pangaea.de/
- Humanities
    - ICPSR https://www.icpsr.umich.edu/icpsrweb/landing.jsp
    - Qualitative Data Repository (Political Science) https://qdr.syr.edu/
- Marine Sciences
    - Seanoe http://www.seanoe.org/
- Science (General)
    - Mendeley https://data.mendeley.com/faq

References

Fairsharing.org : PLOS recommended repositories. https://fairsharing.org/recommendation/PLOS . Accessed 2018-02-01

Re3Data : Registry of Research Data Repositories. http://re3data.org Accessed 2017-11-02

GLOSSARY

Archive - A managed system and service that organizes and preserves information for future retrieval and use, see also **Repository**.

ARK – Archival Resource Key. See **Persistent Identifier**

Camel case – aka camel-cased or camelCase. The practice of writing compound words using capital letters at the beginning of each word. In programming languages the first letter of the compound word is frequently not capitalized, e.g., camelCaseFileName.

Citation metadata - The overall description of your dataset (or project), sufficient to find and properly cite your dataset (or project).
Essential components of a dataset citation include: **Creator**/owner (with ORCID); **Title**; **Date** made public; **DOI** (or other persistent id); and, preferably, an **abstract** which provides a brief narrative describing the data (not identical to the paper abstract, if data is supporting a paper.)

Data - The primary source materials that support research, scholarly enquiry, or artistic or technical innovation. Research data are used as evidence in in the research process and/or are necessary to validate research findings and results.

Dataset - a logical grouping of files or data points that support, or are used as evidence in, a specific research enquiry. Datasets may include one tabular file, a set of 10,000 images, R code, or combinations of multiple types of files.

- Datasets can be raw, e.g., measurements reported by a sensor and output to a file or un-edited responses to a questionnaire; or analyzed, e.g., questionnaire responses that have been tagged with a controlled vocabulary, or measurement data that has had algorithms applied to it to calculate accuracy.
- Figures in a publication are generally considered *not* to be datasets, as they are usually an alternate visual representation of the data on which the publication is based.
- Datasets are often *actionable.* That is, they can be re-used with minimal manipulation by a machine or other researchers to reproduce research results or support novel enquiry.

Data dictionary - a file that describes each element of your dataset. If your dataset includes tabular data, R code, and images, the data dictionary would include a list of the fields in the table and what they mean, including units and precision; a brief overview of the purpose of the code (if not already contained in comments); and information about the images and how they relate to the dataset (more detailed metadata for the images should be embedded). For a brief guide see http://datadryad.org/pages/readme

Data Management Plan (DMP) - A data management plan is a formal document that outlines what you will do with your data during and after a research project. DMPs help articulate who has responsibility for your data throughout its lifecycle, what can and cannot be done with the data, as well as how to find, re-use, and interpret your data. Many potential data management disasters can be handled easily or avoided entirely by planning ahead.

Data model - A data model defines how the different elements (measurements, data points, files, etc.) of your data are connected, and how they interact. Creating a data model can be very helpful when planning how to collect and later display large amounts of information.

Digital preservation - The process that maintains the integrity and readability of digital data for a specific period of time. This process may include data integrity checks, migration, transformation of the data into new formats, or environment emulation.

Disaster Recovery Plan – A documented process that describes the procedures an organization will follow to recover data and restore IT infrastructure functionality after a disaster.

DOI – A digital object identifier (DOI) is a very specific kind of persistent identifier. DOIs are assigned to a digital object like a publication or dataset to facilitate finding and citing that object. DOIs are "resolvable" – that is, if the object moves to a new online location, the registry that tracks that object can be updated to make sure any links to that object do not break.
Many repositories will issue DOIs for deposited data. The Libraries can also issue a DOI for a dataset or publication that is created by SI staff, and stored in an SI managed location such as in SRO, on an si.edu website. All digital objects with DOIs issued by the Libraries are included in metrics and impact reports submitted to the Castle.

EZID – see **Persistent Identifier**

FAIR data principles – FAIR is an acronym for Findable, Accessible, Interoperable, and Re-usable. The principles are meant to guide data producers and publishers as they strive to practice good data management and enable leveraging high quality formal digital publications to facilitate future reuse, discovery, and evaluation.

Read all the principles on the FORCE11 website: https://www.force11.org/group/fairgroup/fairprinciples

Handle – see **Persistent Identifier**

Metadata – Descriptive "data about data." This term can be used in many ways—in the data dictionary, metadata provides descriptions of and context for the contents of your dataset. In a dataset citation record, the metadata provides a high-level description of the overall dataset.

Metadata record - see Citation metadata

Metadata schema - a standard framework or set of rules that are used to provide a consistent, structured description of an object, e.g., DarwinCore, schema.org.

Ontology - often used interchangeably with taxonomy. The specification of names, definitions, and relationships among entities in a given system.

ORCID – An ORCID (Open Researcher and Contributor Identifier) is a persistent digital identifier for academic authors and researchers. When used in manuscript and grant submission activities, an ORCID can automate linking between you and your digital publications. ORCID is also the name of the non-profit organization that maintains the registration of researcher ids called ORCIDs.
SI Libraries is an ORCID member, and leverages their services to create standardized researcher bibliographies and track impact and publication activities of SI scholars.

Persistent identifier (PID) - A Persistent identifier is a stable alphanumeric or numeric descriptor attached to a digital object that is used to uniquely identify it. A GUID, or Globally Unique ID, is a type of persistent identifier that is unique within a large identifier ecosystem. Common types of persistent identifiers used in data management include DOIs, handles, ORCiDs (for authors), ARKs (Archival Resource Keys), and EZIDs.

Quality Assurance/Quality Control – Though often used interchangeably, quality assurance (QA) and quality control (QC) are activities invoked at different stages of a project.
QA establishes processes and standards to ensure data is collected or created in a way that meets project requirements, i.e., error prevention. QA may involve using data collection forms with drop-down menus instead of free-text fields, or writing protocols that specify use of particular measuring tools. QC establishes processes to check for compliance with project requirements, and specifies how to correct or normalize data that does not meet those requirements, i.e., error detection and correction.

ReadMe file - see Data dictionary

Repository - an actively managed system used to store scholarly output for preservation. A repository may or may not be publicly accessible. A web server, external hard drive, or cloud storage system such as Google Drive, are not repositories.

Research outcome - the result of a particular line of research. This could be a peer-reviewed paper, presentation, dataset, algorithm, computer application, process, or object.

SIDora – the Smithsonian data repository. Managed out of the Office of Research Computing, OCIO, SIDora is built on a Fedora Commons 3 platform, with a Drupal front end. It can accommodate data projects that are large, have custom interface needs, or are structurally complex. More information can be found at https://oris.si.edu/sidora or contact Beth Stern sternb@si.edu

Smithsonian Research Online (SRO) – is managed by Smithsonian Libraries and has two primary components: a bibliography of publication citations and a repository of full-text online editions built on the DSpace platform. While the repository is primarily used for publications, it can also accommodate datasets and other supplementary material. More information can be found at https://research.si.edu/about/ or contact Alvin Hutchinson hutchinsona@si.edu

URN – a Uniform Resource Name is an identifier that uniquely identifies a resource on the web, and that uses the URN scheme for naming. Both URNs and URLs (Uniform Resource Locators, aka web addresses) are types of Uniform Resource Identifiers (URIs.) Unlike URLs, a URN does not necessarily enable location of the resource that it identifies.

Versioning – the process of saving, and then assigning unique names or numbers, to a file whenever significant changes are made. Versioning files enables tracking changes over time, as well as supporting reverting to an older version of a file. Several software tools and file systems support automatic versioning including: DropBox, Git, ArcGIS, and most wikis.