

Intro to Data Management with ARCC

Introduction: Contemporary research is data intensive that requires developing plans and practicing good management strategies. While there are not clear cut answers to every data management situation, there are some best practices to keep in mind throughout your research project. On this page you will find a list of topics relating to data management for research using ARCC resources. These topics are modular and do not need to be followed in order, so please feel free to jump to the topic you are most interested in.

Goals:

- ❑ Understanding the Research Data Management Life-cycle
 - ❑ The importance of metadata
 - ❑ Data Storage Options
 - ❑ Organizing and naming files
-

Sections:

1. [What is data lifecycle and how ARCC can play a role in it?](#)
 2. [What is metadata and the importance of a README file?](#)
 3. [Data value and storage](#)
 4. [Organizing data logically](#)
 5. [file naming conventions](#)
-

What is the data lifecycle and how ARCC can play a role in it?

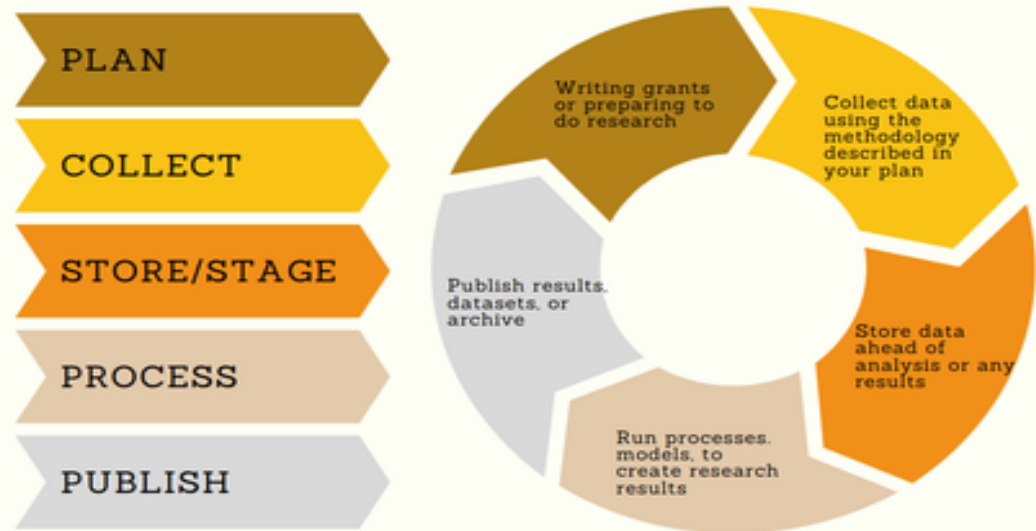
In this section of the workshop we will discuss data management for research workflows, why it's important, and introduce how you can use ARCC resources to manage your data. This page gives background information for future topics, if you are looking for specific examples, please head back to the main Data Management page to navigate to other pages.

It is also important to note that the content of these pages are general suggestions and guidelines to assist in your research workflows. The content is NOT rules or requirements for using ARCC resources during your research project.

- [Research Data Life-cycle](#)
 - [The Planning Phase](#)
 - [The Collection Phase](#)
 - [The Storage Phase](#)
 - [The Analysis Phase](#)
 - [The Publishing Phase](#)
 - [Next Steps](#)
-

Research Data Life-cycle

The Life-cycle of data in a research project can be broken down into multiple phases. These phases can be thought of as distinct phases, but often blend into each other with little to distinguish the differences between them. Below we provide details and guidelines for each phase.



The Planning Phase

Data Management Planning is an often overlooked, but critical phase of the Research Data Management Life-cycle. Not only will it be useful for the execution of your research project, a formalized plan is often required by funding agencies such as the National Science Foundation (NSF) and National Institutes of Health (NIH) among many others. The planning phase of the research data life-cycle usually comes after a research project has been conceptualized but before a the project is underway (or even funded), but can always be re-visited in an informal manner. It is important to consider a variety of things during this phase as well as establish goals for your data:

- What kind of data is required to answer our research question?
- What file formats will be collected?
- Is there a particular software needed in the other phases that requires the data to be formatted in a particular way?
- Are there any federal compliance requirements?
- How will the data be stored and protected prior to analysis?

- Will the data be preserved or discarded after the project is complete?
-

How ARCC Can Help With Planning

ARCC is a good resource for many of the phases in the Research Data Management Life-cycle, but in the planning phase, is a bit more limited in scope. That said, there are some things researchers can interact with ARCC on in forming this plan:

- The [ARCC documentation](#) and policies can provide researchers with much of the background information required about resources available
 - ARCC resources are described in a [Facilities Statement](#)
 - ARCC is always willing to meet with researchers to discuss any Data Management issues by scheduling through our [ticketing system](#)
 - We work closely with [UWyo Libraries](#), who are well versed in Data Management and can refer to them for more nuanced questions
 - They also administer the UWyo instance of a Data Management Planning Tool called [DMPTool](#), which can be very useful for writing data management plans
 - They also have resources available for publishing research data, which will be discussed later in this module
-

The Collection Phase

There are multiple types of data and collection of these data vary greatly depending on the kind of research being done. Below is a table of some types of data that could apply top any management scenario. Please note this is not a comprehensive list and many more types of data that exist.

Classic	Simulated/automated	Social
----------------	----------------------------	---------------

<ul style="list-style-type: none"> <input type="checkbox"/> Text files <input type="checkbox"/> Tabular (Spreadsheets, Databases, etc.) <input type="checkbox"/> Matrices <input type="checkbox"/> Observations/field notes 	<ul style="list-style-type: none"> <input type="checkbox"/> Computer Models <input type="checkbox"/> Instruments (Microscopes, Weather Stations, Satellite Imagery, etc.) <input type="checkbox"/> Audio/video recordings 	<ul style="list-style-type: none"> <input type="checkbox"/> Surveys <input type="checkbox"/> Interviews <input type="checkbox"/> Focus groups <input type="checkbox"/> Exit Polling
---	--	---

During this phase, it is important to keep data that are being collected organized and named with appropriate conventions to assist with the next phases, and examples will be discussed in other modules.

How ARCC Can Help With the Collection Phase

Since ARCC does not advise on how research should be done, how data are collected is not usually an area of expertise we provide. However, we can provide advice on how the data maybe used in later phases of the Research Data Management Life-cycle, that you may want to be mindful of while you are in the collection phase. If you are unsure about anything that you may run into, please remember that ARCC provides the following that may assist you:

- The [ARCC documentation](#) and policies can provide researchers with much of the background information required about resources available
 - ARCC resources are described in a [Facilities Statement](#)
 - ARCC is always willing to meet with researchers to discuss any Data Management issues by scheduling through our [ticketing system](#)
-

The Storage Phase

Once your research data are collected, you will need a place to keep them before moving onto the next phases. This phase is often the longest of the phases and sometimes overlaps many of the others. While seemingly trivial, the storage phase is vital to the Data

Management Life-cycle. Here are a few nuances to be aware of before we discuss the systems and services ARCC provides that can assist in this phase, and it is important to ask yourself a few questions before making a decision on where your data will be stored:

- Does the data fall under any federal compliance or other security restrictions?
- How are the data to be accessed and how frequently?
- Do the data need to be backed up or version controlled?
- Do other collaborators require access and are they local to your institution or not?

How ARCC Can Help With the Storage Phase

[Research data storage](#) is a core service that ARCC provides and we have several storage options available for you that will be discussed in subsequent modules, but to state it briefly there are three core storage systems that ARCC provides that fit different phases of the Research Data Life-cycle each filling different roles detailed in the table below:

The ARCC Data Portal (Storage)	MedicineBow HPC system (Analysis)	Pathfinder (Storage)
<ul style="list-style-type: none"><input type="checkbox"/> Free for UWyo researchers up to a default limit<input type="checkbox"/> Accessible via the UWyo network or VPN<input type="checkbox"/> Includes backups and snapshots	<ul style="list-style-type: none"><input type="checkbox"/> Home (for configuration and profiles)<input type="checkbox"/> Project (for shared data during analysis)<input type="checkbox"/> gscratch (for actively read/write during analysis)<ul style="list-style-type: none"><input type="checkbox"/> MedicineBow is NOT backed up, but includes snapshots	<ul style="list-style-type: none"><input type="checkbox"/> Cloud-like backend<input type="checkbox"/> Web-enabled S3 buckets for data storage, data transfer, etc.<input type="checkbox"/> Is NOT backed up

Transferring data to and from these systems is discussed in another workshop. Please also be aware that none of these systems meet any federal compliance requirements.

The Analysis Phase

The analysis phase can include a variety of methodologies and tools to complete. This phase also often includes different stages and versions of data. Here are a few questions to ask yourself before entering this phase of the Research Data Life-cycle:

- How large are the data that I am working with?
 - Will I need a powerful system such as a High Performance Computing system to complete this work?
 - What software will I need to perform the analysis?
 - Will there be new data generated as a result of this work (simulated data for model training, summarized subset of raw data etc.)
 - Will this work change my raw data and do I need to keep a copy of either the raw data or results?
 - How will I manage the changes that will happen during this phase and maintain a record of them?
-

How ARCC Can Help With the Analysis Phase

[High Performance Computing](#) is another core ARCC service and we offer an assortment of support for this type of work. Along with the [MedicineBow HPC system](#), we provide documentation, troubleshooting consultations, software management, and workshops among the system administration of the system. Additionally, we provide facilitation of and technical support for [NCAR Wyoming Supercomputing Center's Derecho system](#).

If neither of these systems meets your needs for the analysis phase and you still require assistance, please reach out to us via our [service portal](#) to discuss what your requirements are and potential options.

Another service that may be of use during this phase that ARCC provides is [GitLab](#) for collaborative code development and version control. We do also recommend maintaining a README file that is associated with your work to record additional metadata that will be useful for the publishing phase of the Research Data Life-cycle. Metadata and README files are discussed in the next module.

The Publishing Phase

This phase of the Research Data Life-cycle usually occurs after the work has been completed but before other work (such as a manuscript) is published. What exactly it involves depends on the requirements of the various funding agencies and/or scientific journals that you are working with. For example, if your work was funded by the NSF the resulting data of your work must be made publicly available, and if you are wanting to publish in the Journal of Science, your data has to be available before your manuscript will be published itself. Good scholarly metadata (described in the next section) will be key to completing this phase. Other key concepts in this phase are:

- ❑ Discipline specific data repositories
- ❑ General or institutional data repositories
- ❑ Digital identifiers, such as a [Digital Object Identifiers \(DOIs\)](#)
- ❑ Personal scholarly identifiers, such as an [ORCID](#)

How ARCC Can Help With the Publishing Phase

ARCC supports some of the systems used in publishing research data along with the Data Librarians at The University of Wyoming Libraries. The Data Librarians will be the primary points of contact during this phase and can seek ARCC's assistance if needed. Additionally, some larger datasets will require ARCC to host or move for the researcher. Lastly, if the data to be published are already stored on one of ARCC's systems, ARCC can assist in getting it moved to the appropriate place for publishing.

Metadata and README files

Documentation of research data, also called, metadata is an often overlooked, but critical, aspect of research data management. It's important to note that this type of metadata is not as simple as basic system metadata familiar to computer scientists such as file sizes,

ownership, etc. When we are talking in a data management context, metadata provides several descriptive elements that inform viewers of the data of who collected it, how it was collected, where and when it was collected, processing methods, and much more. Not only is this a requirement for data publishing, it is very useful for collaboration between other researchers and can serve as a tool to ensure consistency throughout the other steps taken in the Research Data Management Life-cycle.

- ❑ [Common Metadata Standards](#)
 - ❑ [Common Metadata Fields](#)
 - ❑ [Metadata File Types / README](#)
 - ❑ [README Example](#)
 - ❑ [How to Download the Libraries' README on ARCC HPC](#)
 - ❑ [Next Steps](#)
-

Common Metadata Standards

There are several standards available depending on discipline that provide the advantages of ensuring you have a complete, standard set of information about each part of your data and enable your dataset to be organized with other datasets, a few examples are:

- ❑ [FGDC \(Federal Geographic Data Committee\)](#)
- ❑ [DDI \(Data Documentation Initiative\)](#)
- ❑ [Dublin Core](#)
- ❑ [Darwin Core](#)
- ❑ [ABCD \(Access to Biological Collections Data\)](#)
- ❑ [AVMS \(Astronomy Visualization Metadata Standard\)](#)
- ❑ [CSDGM \(Content Standard for Digital Geospatial Metadata\)](#)

While these standards exist and can help aid researchers in recording complete metadata, they are not universally required to be used. A best practice is to record the information that works for you and your collaborators. It's better to have something than nothing at all.

Common Metadata Fields

While each of the standards listed above are each unique and recommend recording different information, there are several commonalities. Listed below are a few of them that each metadata document could contain for a directory of research data.

- Creator/Author: <Researcher name/ORCID>
- Subject/title: <Name of the data>
- Description: <Short paragraph describing the data and how they got to this state e.g., image taken, data processed, etc.>
- Contributor(s)/Collaborator(s): <Names of people associated with the project>
- Date: <use a format that is standardized across all the data e.g., YYYYMMDD>
- Original Format/File types: <.txt .csv .png .sql>
- Relation: <list any relating files/folders>
- Location: <e.g., Latitude & Longitude in decimal degrees>
- Rights: <funder grant number, or open source>

Metadata File Types / README

Metadata can be recorded in multiple ways including in a filename, in a spreadsheet, in an XML file, or into a database. However, a very common type is a simple text file called a README file. A README provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data. In general a good README should include several things in addition to what is listed above:

General Items	Optional Items	Other Recommendations
---------------	----------------	-----------------------

<ul style="list-style-type: none"> □ File naming system (with examples) □ Folder structure □ Relationships and dependencies between files □ Other documentation files of interest within dataset (notes, companion files) □ For each major file, short description of contents □ Date of creation of each major file 	<ul style="list-style-type: none"> □ Experimental & environmental conditions of collection (if appropriate) □ Standards and calibration for data collection (if applicable) □ Uncertainty, precision and accuracy of measurements (if appropriate) 	<ul style="list-style-type: none"> □ Methods used for data processing □ Software used in data collection and processing, including version numbers □ File formats used in the dataset & recommended software □ Quality control procedure applied □ Description of file versioning system if appropriate □ Dataset changelog
--	---	---

README Example

Luckily for UWYO researchers the University Data Librarians provide an extensive sample [README](#) that can be [downloaded directly from their website](#). Below is an example of what the beginning of the file looks like:

```
This DATSETNAMEREADME.txt file was generated on YYYY-MM-DD by NAME
GENERAL INFORMATION
1. Title of Dataset:
2. Description or abstract of dataset:
3. Author Information
   A. Principal Investigator Contact Information
      Name:
      Institution:
      Address:
      Email:
4. Date of data collection (single date, range, approximate date) <suggested format YYYY-MM-DD>:
5. Geographic location of data collection <latitude, longitude, or city/region, State, Country, as appropriate>:
6. Information about funding sources that supported the collection of the data:
7. Keywords for dataset:
8. Discipline of dataset:
```

9. License for dataset:

How to Download the Libraries' README on ARCC HPC

ARCC recommends researchers to download the Libraries README file into the project directory for High Performance Computing (HPC) projects when they initially get the project setup. That way the file can be maintained throughout the process, so that when it comes to the publishing phase of the Research Data Life-cycle, the metadata is already recorded and can be shared quickly.

Here is an example of a Linux command to run to download the Libraries README file:

```
#To Download
wget https://uwyo.libguides.com/ld.php?content_id=61572044

#To rename as README.txt
mv 'ld.php?content_id=61572044' README.txt
```

Data Value and Storage

Another critical aspect of Data Management is consideration of which datasets are the most valuable and what protections need to be in place for them. Discussed in this section of the workshop, are the different stages data could be in, what to consider before choosing a storage option, a comparison of the storage offerings from ARCC, and long-term planning for the data.

- [Stages of Data](#)
- [Assessing the Needs](#)
- [Comparing Storage Options](#)
- [Considering Other Requirements](#)
- [How to Decide](#)
- [Next Steps](#)

Stages of Data

During a research project, data takes on different stages of use each with a different storage requirement. Some research projects will use all of these stages, some will only use a few.

1. **Potential Data** - Data that are not yet collected, but there is a plan to store them.
2. **Raw Data** - This transitional stage includes everything that is collected from the potential stage into a place for processing or pre-processing.
3. **Prepared Data** - This stage describes the pre-processing of the raw data that prepares for a model or other processes.
4. **Intermediate Data** - This stage is the most temporary of all data, it could be a step in a process that creates these data before processing into final data or simulated data that helps train a model.
5. **Final Data** - The resulting data of a process. These data tell the story of the research and indicate the results.
6. **Published Data** - These data are the same as the final data, but are in a format optimized for sharing
7. **Archived Data** - This stage of data are no longer needed for ongoing research projects but are not deleted.

Assessing the Needs

Stage	Storage Need
Potential	None yet, but a plan for raw is in place
Raw	Could be stored in a temporary place or in a more permanent place if keeping raw is determined to be valuable
Prepared	Should be stored or transferred to storage that will optimize the next process
Intermediate	Should be stored in a highly performant in read and write operations and backups are not necessarily a requirement
Final	Should be stored in a safe place if the process is difficult to re-do

Published	May be in a different format for data sharing, possibly a compressed file stored on a repository
Archived	Could be stored somewhere in “cold” storage in the most cost effective way possible

Comparing Storage Options

Storage Type	Advantages	Disadvantages
External Storage i.e., portable hard-drive or Laptop	Fully user controlled, can be encrypted, portable, and not accessible without physical access	Easily lost, vulnerable to damage, no extra copy, only as safe as the circumstances
Cloud backed service i.e., Google Drive or Dropbox	User friendly, accessible from anywhere, interactive use of native files, shareable, syncable	Possibly costly and subject to unexpected terms of service changes, potentially unauthorized access
Cloud storage services i.e., AWS, GCP, or Azure	Robust, scalable storage with customizable access and interoperability within the cloud environment	Potentially costly egress fees, terms of service changes
Institutional Research storage service i.e., ARCC Data Portal	Free up to default limits, support for UWyo researchers, included backups and snapshots	Requires a UWyo based PI, does not include an offsite back up, non-compliant data only
Institutional HPC Storage i.e., ARCC MedicineBow	Access to compute power, specialized directories for performance and collaboration, snapshots	Linux only permissions, not backed up, non-compliant data only
Specialized Institutional storage i.e., ARCC Pathfinder	Cloud-like backend and functionality with S3 protocol for sharing	Not backed up, requires specialized software clients to interact with, non-compliant data only

Considering Other Requirements

Before determining a storage solution for a research project, researchers should take a moment and consider all requirements they may have and what sort of compromises they can live with. Here are a few additional questions to consider prior to making a choice:

- How frequently will I need access to my data and how do I want to access?
 - Will I have collaborators that need access?
 - Do I require backups?
 - Will I need to compute on these data?
 - Are there any federal compliance requirements such as HIPAA or NIST 800-171?
 - Is this production-like data that need to have a systems with near 100% uptime?
 - Do I require proprietary software to access the data?
-

How to Decide

It may seem like a daunting task to choose where to put research data, but the reality is that data can be transferred to different systems when needed. There will always be nuances to migrating data from one platform to another as well as potential costs. If you are unsure, you can always request for a consultation on what ARCC can provide to get clarification on if that will meet your research needs or not.

Organizing Data Logically

Organizing data can help make research efforts more efficient and logically separate. Many research programs are composed of multiple projects, investigators, and organizations working collaboratively to collect, share, analyze, or disseminate scientific results. Projects are analogous to a file-storage directory but are more flexible and can hold their own metadata. Within a Program, most projects contain a collection of files that can all be described with similar data collection methods, and which typically come from the

same funded effort. Smaller, focused projects may contain only a single dataset, while larger projects that collect or produce multiple types of data may contain several datasets. Identifying the specific dataset(s) that will be produced by a project is a central aim of project data management planning, and is necessary for planning an organizational structure within a project. Projects can also be used to share information that doesn't require metadata, such as administration or outreach materials.

- ❑ [Project Naming](#)
 - ❑ [Project Naming on ARCC Systems](#)
 - ❑ [Folder Organization Example](#)
 - ❑ [Level of Granularity](#)
 - ❑ [Folder Naming](#)
 - ❑ [Next Steps](#)
-

Project Naming

Project titles should be as concise as possible while still containing key information about the dataset. The title is often the most important piece of metadata describing a resource. It is the first thing seen by people when browsing or searching for a resource, and may be the only information used to evaluate the content of the resource.

At a minimum, project titles should contain the following information:

- ❑ Location
- ❑ Data type
- ❑ Year (or other time unit) range
- ❑ Program or institution name, if your dataset is part of a large effort

Naming Examples

Poor project titles:

- Data Management
- Workshop for Mike

Better project titles:

- Data Management Workshop, University of Wyoming ARCC, Fall 2024
 - Conductivity, temperature and depth data for 12 northwestern Gulf of Mexico locations, May to July 2012
 - SAFARI 2000 Upper Water Column Profiles, Gulf of Alaska, 2011-2012
-

Project Naming on ARCC Systems

While the project names above are very descriptive and something to record in a README file, ARCC systems have restrictions on how the type of characters and how many can be used in a project name. This is due to how permissions work on the system and very long project names with spaces or other special characters can cause problems with the administration of the system.

The recommended limitation for project names on ARCC are to use acronyms when possible or shorten words in a logical way. The restrictions are as follows:

- Lowercase letters and numbers only
- Hyphens are allowed, but no other special characters such as underscores
- No longer than sixteen (16) characters

For Example, if this tutorial were to be a project on ARCC systems, we would take the long title of “Data Management Workshop, University of Wyoming ARCC, Fall 2024” and change it to read as one of these suggestions:

- data-mgmt-arcc24
 - arcc-datmgt-uw24
-

Organizing Folders within a Project

Folders are an important way to organize your project files into smaller, easier-to-manage, and identifiable units. Create a logical folder structure to help you stay organized and easily find and retrieve your stored files, and initiate it at the beginning of your project to save time and frustration.

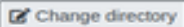
Avoid complex, deeply-hierarchical folder structures, which require extra browsing for file storage and retrieval. Try to keep the folder levels to **no more than three deep**. Folder structures can be simplified by including all the essential information concisely in the file name.






The following best practices are recommended for creating an effective project folder structure:

- ❑ Organize folders by major project components.
- ❑ Create a hierarchical system with nested subfolders (high-level folders for broad topics with more specific folders within).
Examples of high-level folder topics include:
 - Input data files by discreet location/source/type
 - Metadata
 - Code or scripts
 - Results or output data
- ❑ Organize the data by data type and then by research activity.
- ❑ Separate preliminary and final data into different folder structures.
- ❑ Be consistent with your folder organization throughout the life of your project and/or Research Campaign.

Folder Organization Example

On ARCC systems, folder/directory names do not have to comply with the project name restrictions, but it helpful to keep them as short as possible while being descriptive, without using too many special characters and no spaces.

 / project / arcc / dperkin6 / data-mgmt-arcc24 / 

<input type="checkbox"/>	Type ▲	Name
<input type="checkbox"/>		code
<input type="checkbox"/>		metadata
<input type="checkbox"/>		ModelResult_07012024
<input type="checkbox"/>		ResearchData_Instrument1_06012024
<input type="checkbox"/>		ResearchData_Location1_05012024

 / project / arcc / dperkin6 / data-mgmt-arcc24 / ResearchData_Location1_05012024 /

<input type="checkbox"/>	Type ▲	Name	
<input type="checkbox"/>		conditions	
<input type="checkbox"/>		images	
<input type="checkbox"/>		samples	

Level of Granularity

It may be unrealistic to anticipate and pre-create every folder that will be needed for a project. Instead, consider the level of folder hierarchy that will provide sufficient structure for users and collaborators on your project to create their own subfolders.

A good approach is to establish the first one or two levels in the hierarchy, then let your collaborators create subfolders for lower levels as needed.

Granularity Examples

- Project: data-mgmt-arcc24
 - Parent folder: ResearchData_Location1_05012024
 - Child folder: images
 - *Users can create subfolders within as needed*
 - Child folder: samples
 - *Users can create subfolders within as needed*

Folder Naming

How you name folders will have an impact on you and your collaborator's ability to find and understand the folder contents. Naming folders consistently and descriptively will help users identify records at a glance, and will help to facilitate the storage and retrieval of data.

Folder names should adhere to the following best practices:

- Rename default folder names generated by the Research Workspace with descriptive titles.
- Name folders according to the areas of work to which they relate, and not after individuals. Classify file types with broad folder names.
- Use folder names that are unambiguous and meaningfully describe the folder contents to you and your collaborators.
- Be consistent when developing a naming scheme. Ideally, a scheme is created at the start of a project and used consistently throughout.
- Avoid extra long folder names, but use information-rich file names instead (refer to File Naming).

- Try to avoid duplicate folder names or paths. For example, if a folder is named “Photos” in one directory, don’t create a subfolders elsewhere named “Images”.

Examples of folder names

Poor folder names:

- My Data
- Data From Ben

Better folder names:

- GPS-locations-sagebrush-study-2021
- Raw-songbird-acoustic-data2012-2016

File Naming Conventions

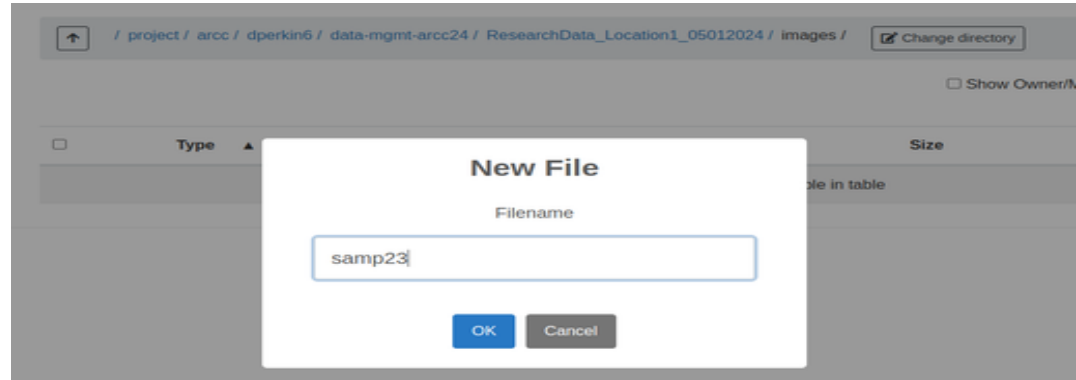
This page walks researchers through the process of creating a file naming convention for a group of files. This process includes: choosing metadata, encoding and ordering the metadata, adding version information, and properly formatting the file names.

-
- [Unique File Names](#)
 - [Abbreviations](#)
 - [Ordering](#)
 - [Separating Characters](#)
 - [Versioning](#)
 - [Patterns](#)
 - [Next Steps](#)
-

Unique File Names

The first thing to consider when naming files is to determine what information (metadata) is important about these files and makes each file distinct?

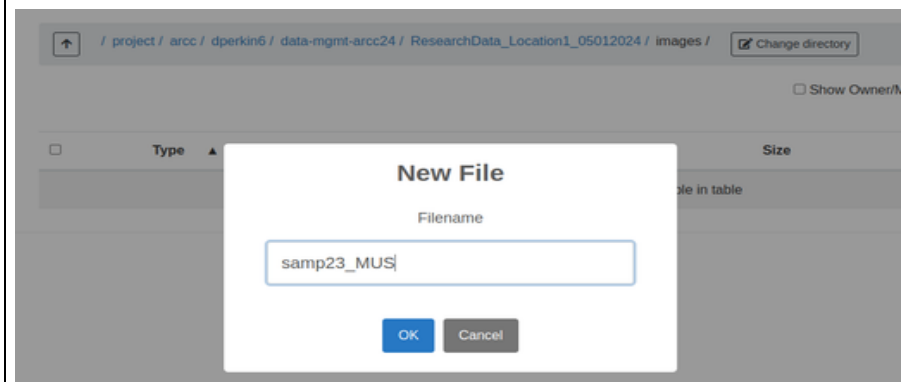
Ideally, pick three pieces of metadata; use no more than five. This metadata should be enough for you to visually scan the file names and easily understand what's in each one. Example: For my images, I want to know date, sample ID, and image number for that sample on that date.



Abbreviations

Do you need to abbreviate any of the metadata or encode it?

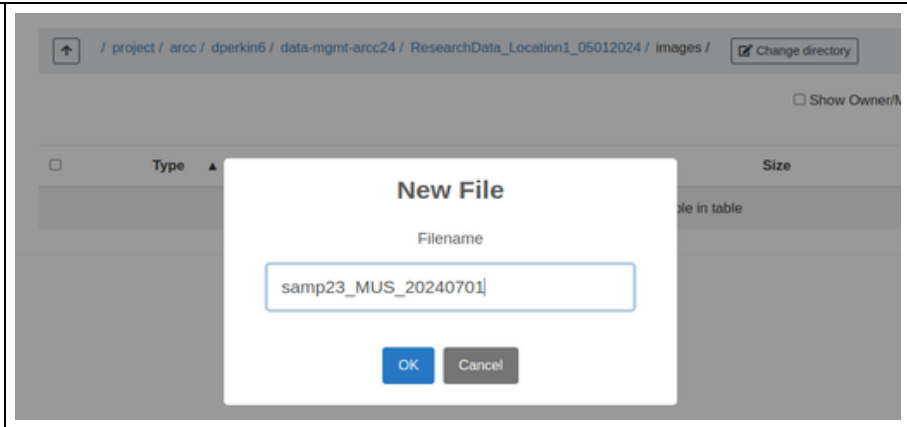
If any of the metadata from step 2 is described by lots of text, decide what shortened information to keep. If any of the metadata from step 2 has regular categories, standardize the categories and/or replace them with 2- or 3-letter codes; be sure to document these codes. Example: Sample ID will use a code made up of: a 2-letter project abbreviation (project 1 = P1, project 2 = P2); a 3-letter species abbreviation (mouse = "MUS", fruit fly = "DRS"); and 3-digit sample ID (assigned in my notebook).



Ordering

What is the order for the metadata in the file name?

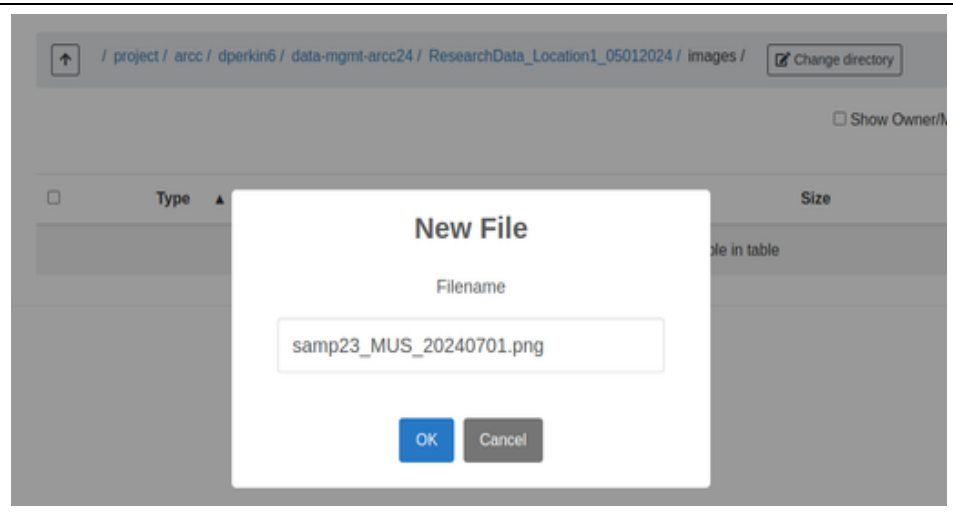
Think about how you want to sort and search for your files to decide what metadata should appear at the beginning of the file name. If date is important, use ISO 8601-formatted dates (YYYYMMDD or YYYY-MM-DD). Example 1: My sample ID is most important so I will list it first, followed by project abbreviation, then date.



Separating Characters

What characters will you use to separate each piece of metadata in the file name?

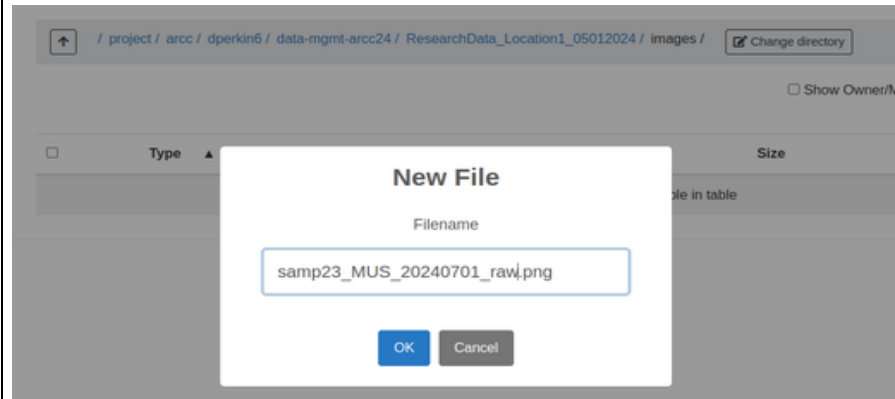
Many computer systems cannot handle spaces in file names. To make file names both computer- and human-readable, use dashes (-), underscores (_), and/or capitalize the first letter of each word in the file names. Example: I will use underscores to separate metadata.



Versioning

Will you need to track different versions of each file?

You can track versions of a file by appending version information to end of the file name. Consider using a version number (e.g. “v01”) or the version date (use ISO 8601 format: YYYYMMDD or YYYY-MM-DD). Example: As each image goes through my analysis workflow, I will append the version type to the end of the file name (e.g. “_raw”, “_processed”, and “_composite”)



Patterns

It is a good idea to write down your naming convention pattern in a README file or another place that you can look up.

Make sure the convention only uses alphanumeric characters, dashes, and underscores. Ideally, file names will be 32 characters or less. Example: My file naming convention is “SAMPLEID_AB_YYYYMMDD_status.png”
Examples are “samp23_MUS_20240701_raw.png” and “samp25_MUS_20240701_composite.png”.



Next Steps

Previous

[Organizing Data Logically](#)

Workshop Home

[Intro to Data Management with ARCC](#)