LA-COE Data and File Management

BEST PRACTICES

July 2024

Introduction

According to the RESTORE Act Center of Excellence for Louisiana (LA-COE) Standard Operating Procedure (V4) that can be found here, "**All data, collected data products, and metadata must be made publicly available within one year after submission of the final report**." Therefore, to assist the researchers with managing their data, producing metadata, and supporting the inclusion of their data into a public repository, the following best practices are provided. The LA-COE team will be made available to researchers via email (la-coe@thewaterinstitute.org) or webinar, when requested, to help answer specific project-related questions and help expand the data management plan from their proposal to a fully functioning plan for funded project. The following table (Table 1) from SOP V4 lists the responsibilities of research subrecipients and the related support by the LA-COE team.

Table 1. LA-COE team support and the research subrecipient responsibility (SOP V4)

Research Subrecipient Responsibility	LA-COE team Support
Plan to manage data	Support research subrecipient in answering questions in the data management plan checklist
Collect, generate, acquire, and organize data	• Ensure that researchers collect, record, and organize information required to complete metadata records • Assist researchers in implementing data management best practices for their data and projects
Create metadata	Provide information and links about metadata creation tools, specifically related to proper formats and information that should be included
Plan data "back-up" storage strategies	Identify possible data "back-up" storage strategies and tools
Long-term data storage/archival	Identify possible long-term data storage options

The LA-COE team will review all semi-annual Performance Progress Reports (PPRs) for updates on the data management. Key tasks that need to be reported in the PPRs include:

1. Choosing "Back-up" Strategies

Research products awarded by the LA-COE need to be securely stored during the development stage and upon completion for archival purposes. Secure storage requires that the products, such as data, metadata, and interpretative documents, must remain accessible and suffer no loss of fidelity over time. The main objective of secure storage is to prevent data loss through user behavior or equipment failure. The most

effective method of secure storage is data "back-up", i.e., maintaining redundant copies of all material at multiple locations. Retaining multiple copies of working datasets and material requires efficient "versioning", i.e., tracking alternative versions of material created during the "back-up" process.

Back-up frequency

The frequency of back-ups depends on how often the data changes. Back-ups require effort to efficiently copy and track material versions (a cost); however, it mitigates the need to recreate material if corrupted or lost (a benefit). Back-up software can help automate the back-up process and help save space when making these. Recommended frequencies include:

- If you do a lot of changes to your data every day, daily back-ups are recommended.
- If the data changes less frequently back-ups every few days are sufficient.
- If the data is being worked on only sporadically, weekly back-ups are recommended.

When you try to decide how often you need data back-ups, think about the effort required to recreate all your work and decide what is the right frequency for your project.

Different kinds of back-ups

There are several strategies for back-ups. A complete back-up of all files includes all files during every back-up. Because of the file sizes and copy time involved, this kind of back-up is practical to be performed at discrete time intervals, i.e., daily or weekly. One of the disadvantages of complete back-ups is it only allows for a few versions to be archived at a time. It is not necessary to back-up all data every time you initiate the back-up process. An alternative type of back-up is called a "differential back-up", where back-up software may permit backing-up of the changes made at predetermined (prescribed time intervals, i.e., every 10 minutes). A third type of back-up is determined by manually-initiated instances, and is called "incremental back-up". Incremental back-up allows for a relatively large number of versions to be saved. As suggested, these types of back-ups have advantages and disadvantages.

While the incremental back-up can save space, its restore time will be longer, and if one of the incremental back-ups gets lost or is corrupt, recovery won't be possible. The differential back-up will take up more space, but the restore process will be faster, and older differential back-ups can be deleted without loss of data for the newer back-ups. Along with doing frequent back-ups, periodic restore-tests should be done periodically to ensure the backup process went without errors, as well verify the restore functionality is working.

What should be backed-up?

The decision of what to back-up is up to you. This decision can be assessed in terms of acceptable risk and will depend on the probability of equipment failure (equality and quantity of back-up devices), consequence of losing each item, and the effort required to back-up each item (e.g., copy time, maintaining adequate storage space). It depends on the kind of work that you are doing and the kind of system that you are using.

If you have a complicated software setup, keep adding software as you go, or regularly adjust its configuration, you might consider backing-up the whole workspace, which may include all or elements of your operation system configuration and research data. This way, in case of hardware failure for example, the system could be copied to a new hard drive without starting over with the installation of all tools

needed to do the work. If you only use a few commonly stand-alone programs (e.g., Microsoft Office, MATLAB, ArcGIS) with few external dependencies (e.g., runtime libraries) it will be sufficient to only back-up your research data.

If you do any kind of programming, you might want to consider using versioning tools designed for ascii files, e.g. GitHub, subversion, CVS or others. You could install the repository either locally or ask your IT department to host it for you remotely. These are tools that come in handy for keeping track of changes to code, to perform cross-platform programming, and to easily try out new programming strategies without risking loss of the changes or even a functioning version of your code.

The following table lists the various software to help securely store your data:

Software	Website	Cost	Free version available?
Acronis	Fast & Secure MSP Cloud Backup Solution & Service – Acronis	From $5.38/month	Trial
Macrium	Macrium Standalone Business Products	$65 one time purchase	Trial
Novabackup	http://novabackup.novastor.com/data-backup-products/pc-backup-software/	$99.95/year	Trial
EaseUS Backup	EaseUS Todo Backup - Home Backup Solutions for Data/Files/OS/Devices.	$40 for business version	Trial
SyncBack	http://www.2brightsparks.com/syncback/syncback-hub.html	Up to $59.95	Yes

2. Setting Directory and File Naming Guidelines

The purpose of these guidelines is to improve naming of files when adding new files to your project.

- Avoid long folder names and complex hierarchical structures, but use information rich file names; examples are: Data\Data\Data\Infrastructure\ → Data\GISData\Infrastructure\
- Use capital letters or underscores to delimit words in folder names, not spaces → salinityData.csv
- Rely on commonly understood abbreviations when applicable.
- When using dates in file names, be consistent (i.e., always use one DD MM YYYY-type format)
- Name files from general information to specific (e.g., instrument - location – time – sample number)
- Avoid unnecessary repetition and redundancy in file names and file paths (e.g. file names in a folder called “Field Work” should not include “fieldwork” in the filename)
- Avoid using non-alphanumeric characters; however, in many cases it will be beneficial to use underscore (“_”) rather than create whitespace/ use spacebar.

3. Selecting Data Archives

As recommended in “Centers of Excellence Data Accessibility and Management Best Practices” released by Treasury’s Office of Gulf Coast Restoration (OGCR) the “data resulting from COE awards should be publicly available in an **open-standard, machine-readable** format and include **appropriate metadata** documentation”.

To meet these criteria for data accessibility required by Treasury, the LA-COE requires that you archive your data with an established data repository and report on the location within **one year after project completion**. To facilitate ease of data discovery, the LA-COE requires data to be archived in a digital format whose structure is well-organized and follow International Organization for Standards (ISO) format with metadata documentation brief describing data and the location (the repository and the DOI). In addition, data resulting from LA-COE requires to be read without proprietary software

Ideally, it will be beneficial to select the repository most relevant to the associated datasets, queried by researchers with similar interests, and with a high number of dedicated users. At the current time, because of its relative maturity, free public access, and relevance to the mission of most Center's research foci, NOAA's NCEI is the recommended repository. Other repositories that may be beneficial to LA-COE's research grantees include:

Repository	Metadata Standard	Cost	Focus	Persistent Identifier	Relationship required?
NCEI	ISO 19115	None	Environmental	DOI	Helpful, but not necessary
Dryad	Dublin Core	Yes	Data from publications	DOI	…
ICPSR	DDI/Dublin Core	None	Social, behavioral	DOI	Helpful, but not necessary
GRIIDC	ISO 19115	Yes	Gulf of Mexico	DOI	Yes
DataOne	Variety	None	None	DOI	No
IEDA	…	None	Geosciences	DOI	No
OBIS	Dublin Core 2	None	Biogeographic	None	Likely helpful
KNB	EML default, ISO, FGDC, MML	None	Ecological	Varied, DOIs supported	No
NCBI	Custom	None	Biotechnology	None	No
Figshare	Compatible with: DataCite, Dublin Core, RDF, CERIF XML and other Schemas (It is customizable/flexible)	Yes	Generalist repository	DOI	No