LA-COE Data and File Management

BEST PRACTICES

October 2021

1. Introduction

According to the LA-COE Standard Operating Procedure Version 3 (SOP V3) 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 the funded project. The following table (Table 1) from SOP V3 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 V3).

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 "backup" storage strategies	Identify possible data "backup" storage strategies and tools
Long-term data storage/archival	Identify possible long-term data storage options

2. Backup Strategies

Research products from projects funded by the RESTORE Act Center of Excellence for Louisiana (LA-COE) need to be securely stored during the development stage and upon project 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 'backup', 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 'backup' process.

Backup frequency

The frequency of backups depends on how often the data changes. Backing up requires effort to efficiently copy and track material versions (a cost); however, it mitigates the need to recreate material if corrupted or lost (a benefit). Backup software can help automate the backup process and help save storage space (Table 2). Recommended frequencies include:

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

When you try to decide how often data should be backed up, think about the effort required to recreate all your work and decide what is the right frequency for your project.

Different kinds of backup

It is not necessary to back up all of your data every time you initiate the backup process. As opposed to performing a complete backup of all files (which, because of the file sizes and copy time involved, is practical to be performed at discrete time intervals, i.e., daily or weekly, and allows for few versions to be archived at a time; this type of backup is called a 'differential backup'), backup software may permit backing-up of the changes at predetermined (prescribed time intervals, i.e., every 10 minutes) or manually-initiated instances ('incremental backup'). Incremental backup allows for a relatively large number of versions to be saved. As suggested, both of these types of backups have their advantages and disadvantages.

While the incremental backup can help save space, the restore time will be longer, and if one of the incremental backups 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 backups can be deleted without loss of data for the newer backups. Along with doing frequent backups, periodic restore-tests should be done to ensure that the backup process successfully completed without errors and that the restore functionality is working properly.

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 (quality and quantity of backup devices), consequence of losing each item, and the effort required to backup 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 tweak its configuration, you might consider backing up the whole workspace which may include all 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 common stand-alone programs (e.g., Microsoft Office, Matlab, ArcGIS) with few external dependencies (e.g., runtime libraries) it will be sufficient to only backup 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 locally or ask your IT department to host it for you remotely. These are tools that come in handy for tracking 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.

Table 2. List of various software options to help securely store your data.

Software	Website	Cost	Free version available?
Acronis	http://www.acronis.com/en-us/business/backup	From $5.38/month	Trial
Macrium	http://www.macrium.com/product/2/a-nameworkstationamacrium-reflect-v6-workstation.aspx	$75 one time purchase	Trial
Novabackup	http://novabackup.novastor.com/data-backup-products/pc-backup-software/	$50/year	Trial
EaseUS Backup	http://www.todo-backup.com/products/home/free-backup-software.htm	$40 for business version	Yes
SyncBack	http://www.2brightsparks.com/syncback/syncback-hub.html	Up to $54.95	Yes

3. 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 (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 DDMMYYYY-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 an underscore rather than creating whitespace through the use of spaces.

4. Data Archives

LA-COE requires that you archive your data with an established data repository and report on the location within **one year after project completion**. This requirement is part of the data accessibility criteria recommended by Treasury to meet federally-recognized data accessibility standards. To facilitate ease of data discovery, the LA-COE requires data to be archived in a digital format whose structure is well-

organized and follows the International Organization for Standards (ISO) format for metadata documentation describing data and the location (link to established data repository). In addition, data resulting from LA-COE must be readable without using proprietary software.

LA-COE highly recommends the use of repositories that automatically generate digital object identifier (DOI), which is in wide use to identify academic information, such as journal articles, and data sets. See Table 3 for more information about the repositories that support DOI. It is also beneficial to select the repository that is most relevant to the associated data sets, 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 the LA-COE's mission, NOAA's NCEI is the recommended repository. Other repositories that may be beneficial to LA-COE's research grantees are listed in Table 3.

Table 3. List of repositories for data archive recommended by the LA-COE.

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