Proposal

Mel Moreno

February 19, 2019

# Chapter 1- Data management workflow

## Introduction

#### Background and context

Traditional field biology programs designed to assess animal populations, their habitats, and how people use and modify these populations and associated environments have experienced large changes in data collection, management, and storage technology in recent years. Changes including new sensor technology, data collection methods, and observing platforms (i.e. Southeast Coastal Ocean Observing Regional Association [SECOORA] and National Ecological Observing Network [ NEON]) have caused rapid changes in the spatial and temporal scale of data collected. As an example, advancements in sensor technology have allowed for changes in water quality monitoring to transition from single samples at specific locations in space and time collected in the field and then processed in a lab, to real-time observations at multiple locations for multiple variables in large spatial areas (see https://portal.secoora.org/). Many of these programs are conceived, planned, and used by biologists, but these users often have training in ecology and biology with limited experience in basic data management, curation, and workflow of data generated from these platforms. Lowndes et al. (2017) highlights the results of a recent survey of program needs of NSF funded principal investigators in biological sciences. Of the 704 scientists who participated in the survey, these respondents identified data skills as the largest unmet need (Barone, Williams, and Micklos (2017); Lowndes et al. (2017)).

#### Continuous data, management and analyses

In the US Gulf of Mexico, large restoration efforts are currently underway to reverse observed declines in key ecosystem components including sea grass, fish communities, and oyster reefs using funding from the consolidated Deepwater Horizon settlements (see <https://www.nfwf.org/gulf/Pages/home.aspx> as example). These restoration projects vary in spatial scale and funding, but like other restoration actions, these projects often have data collection and evaluation efforts that occur frequently over the duration of the project.

#### Adaptive management

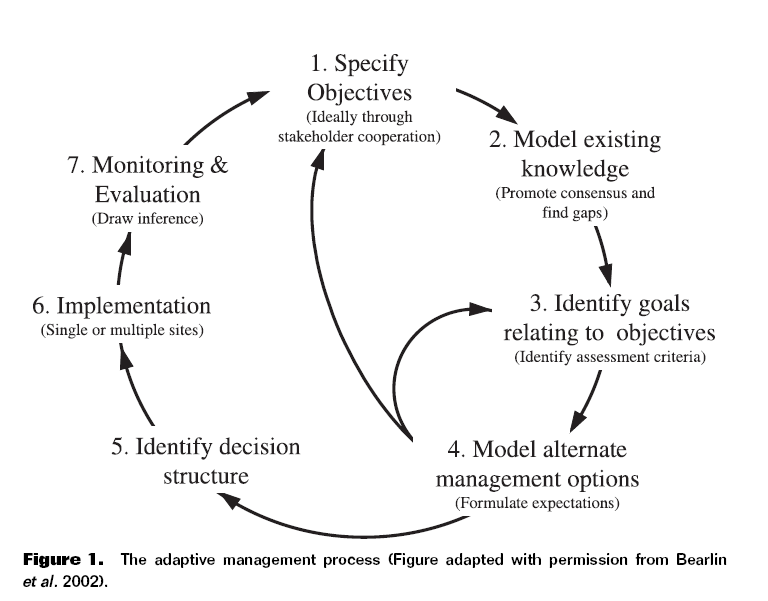
In some cases, data collected from these programs adaptively inform ongoing restoration actions to help improve their outcome. For this type of application, data collection systems and related workflow have to be flexible over time as data collection efforts expand or contract, projects evolve (i.e., construction aspects of restoration begin or end), management actions are implemented, or ecosystems change over short and long-time periods. At the same time, as the systems being monitored change, the people conducting the monitoring and technology used in the field may also change, potentially introducing unanticipated variability in data. These changes must also be considered along with ecosystem response to restoration and other management actions.

Data that are continuously collected or “living data” (Yenni et al. 2018) are critical to this type of adaptive learning to inform restoration and management actions. These adaptations that result from learning from new data can be small such as shifting the location of a sensor, to large changes including restoration practices or revamping of sampling programs because of low statistical power. Many living-data management plans commonly fail due to poor planning and lack of focus (Lindenmayer and Likens (2009)). Living data are challenging to work with from a data management perspective because the data (by design) change as new data are collected. In a restoration or management context as these data are collected they must be processed, and analyses of these data completed to help draw inferences on how the system of interest is responding to the restoration action. This idea of iteratively integrating new data, analyses, and comparing these outcomes with previously stated objectives is not new, and is a central aspect of the “adaptive management” process for natural resources first described in the 1970’s (Holling 1978; Walters 1986). The adaptive management framework is widely discussed and considered in a variety of management and restoration projects (<https://www2.usgs.gov/sdc/doc/DOI-%20Adaptive%20ManagementTechGuide.pdf> ) but success of these programs is highly variable for a variety of reasons (Walters 2007).

By design, an adaptive management program requires rapid feedback between data collection, analyses, and interpretation to drive the process of updating knowledge, examining management and restoration options, making decisions, and implementing actions that are then monitored and evaluated to improve management actions. These programs create challenges from a data management perspective because these continuous efforts must ensure that data meet quality assurance/quality control protocols to identify and correct inconsistencies and errors in field or sensor observations on a continuous basis. Errors in these data, or delays in producing the data in a usable framework, can quickly lead to a breakdown in the adaptive learning process either in terms of slowing the analyses limiting their utility for timely decision making, or worse, erroneously informing the decision making process because of errors in data management or analyses.

An example program that requires an adaptive management approach to restoration are projects funded by the National Fish and Wildlife Federation as part of the Gulf Environmental Benefit Fund (NFWF-GEBF). These projects explicitly require an adaptive management plan to guide the restoration process of the funded project. These projects also require extensive data management plans to capture data collected and analyzed as part of these programs with the overall purpose of creating opportunities to improve future restoration actions by maximizing learning from previous and ongoing restoration efforts.

The Lone Cabbage Reef (LCR) restoration project is a large restoration effort in the eastern Gulf of Mexico funded by NFWF-GEBF. The main goal of this effort is restore historical oyster reefs so that they may be plastic to sea level rise, and fluctuations in river discharge. This project generates data from multiple sources including continuous autonomous water quality data from sensors and observations of oyster populations from field biologists. These data are generated at different time frequencies with sensor data obtained at hourly time intervals from multiple spatial locations and biological data collected at discreet time intervals from multiple spatial locations. For both cases, there is a need to prepare data, meet data quality standards, and complete routine analyses of data to ensure data collected are useful for project objectives. Because this is a long-term restoration project with numerous uncertainties in how the ecosystem will respond to restoration actions, developing a data management and workflow system that automates as many aspects of the workflow including QA/QC, measurement errors, and inconsistency in naming conventions is essential to allow for rapid analyses of data to inform decision making related to sensor deployment or modifying the reef restoration process through additional construction efforts.



### Objective

In this chapter, I will document how the basic elements of the LCR restoration project water quality and biological data associated with oyster populations are managed. The objective is to develop and implement a data management workflow, which starts at the data collection point (i.e physical data sheet if required) and ends at the visualization/ interpretation of collected data from different data streams. These data streams include water quality data from a network of sensors that record observations hourly to counts of oyster populations on reefs that occur seasonally and are recorded on paper data sheets. I document how these data are recorded, and then the data quality assurance/quality control procedures, data checking (anomalous values), data visualization, and data releases for analyses using multiple software tools. This chapter provides an example of a living data project can function to inform an ongoing, long-term restoration project and serve as an example for other projects with data collection efforts.

### Implementing a modern data workflow

Creating and automating a data management workflow for living data is an emerging skill for natural resource professionals. Increasingly data management is recognized as a core skill for biologists and ecologists (Hampton et al. 2017). Even though the design of my workflow will be specific to the Lone Cabbage Reef restoration project, the steps outlined can be broadly used for many conservation efforts. The tools used to implement the data management workflow, are also readily available online and most tools offer tutorials and workshops for a more in-depth training. The approach for this workflow requires basic knowledge of computer coding and version control structure (to track changes in data and computer code). I will use freely available open source tools including program R (<https://www.rstudio.com/>) and GitHub (<https://github.com/>) for version control.

## Methods

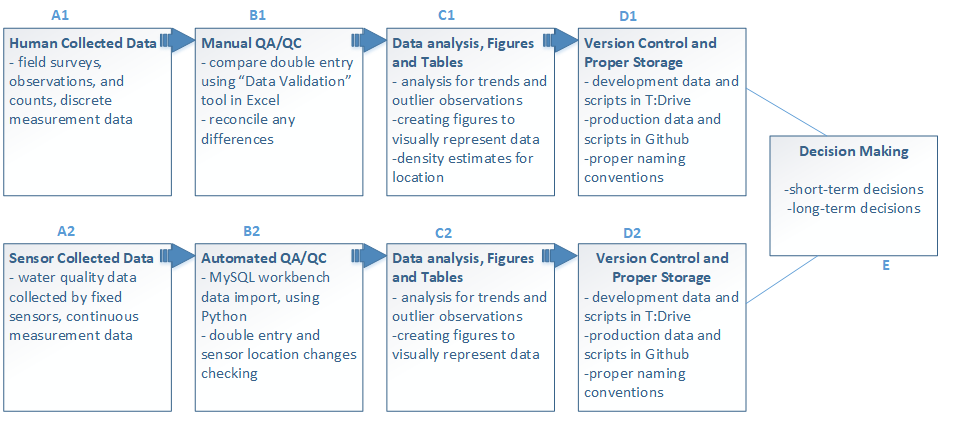


Figure 2- Data management workflow for the Lone Cabbage Oyster Restoration Project.

### Field Collections

One of the goals of a data management plan is to minimize errors in data collected. Often, the first step in the data collection process is transcribing an observation in the field to paper or electronic datasheets for analyses back in the lab. This simple effort of recording the data in the field is the first opportunity to introduce errors in the data collection process. These errors can come from a variety of sources such as the wrong date or site name may be written on a sheet or the person recording the data may be unfamiliar with terminology or protocols. To minimize these types of mistakes best practices for data management such as those adopted by USGS recommend development of a standard set of data guidelines before field collections begin (Figure 3).

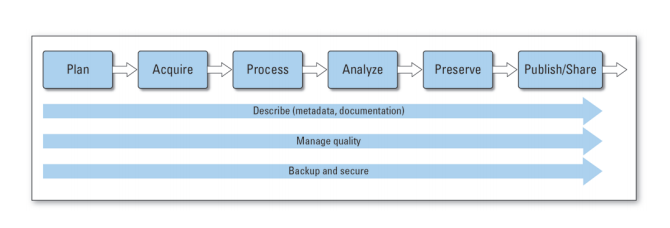


Figure 3- USGS Science Data Lifecycle Model (<https://pubs.usgs.gov/of/2013/1265/pdf/of2013-1265.pdf>)

These guidelines define how basic data are recorded such as date and time formats, site naming conventions, and units of measure for specific observations. This type of predetermined information is a key first step in reducing the risk of one type of data in the field. As an example, simple differences in how dates are recorded by different people such as YYMMDD or MMDDYY formats can create confusion as to when a sample was actually collected. Errors in site names can place the data observations in the wrong location spatially. To minimize this risk, when possible data sheets can be pre-populated with as much information as possible before going into the field.

#### Human collected data

For the Lone Cabbage Reef project (Figure 2, box A1), observational data collected in the field primarily consists of counts and size measurements of oysters from line transects among randomly selected oyster reefs delineated into strata based on specific research questions. These data are recorded on waterproof paper data sheets. To reduce chance of field errors and save time while in the field, I will work to improve data workflow by providing guidance on pre-populating datasheets when possible with basic information including, date and location following data naming standards and field protocols that I am helping to develop have been developed for the project (Figure 4).

#### Sensor collected data

Sensor collected data differs from human collected data, in that sensor data are observations recorded by an instrument automatically. These types of data are a common component of many large-scale observation platforms that may record environmental or biological data continuously, and then make these observations available for use at set time intervals or through “live” feeds. Examples of these types of data include river discharge information provided by USGS or wind observations from a NOAA weather buoy.

The LCR project has a small array of sensors that track temperature and conductivity of water near the oyster reef restoration site. To retrieve the data from these sensors, the sensor must be physically removed from the water and a data file downloaded from the receiver (Figure 2, box A2). To ensure that sensors are functioning, the sensors are serviced bi-weekly to retrieve data.

Each data file contains about X lines of data and a total of about Z observations. While the observations are collected automatically, there are still opportunities to introduce errors when these data are collected. This includes incorrect naming of files once downloaded to a laptop in the field, copying over files on the laptop erroneously, or failing to “start” the sensor once redeployed. to ensure that it will continuous to record a stream of interrupted measurements. The LCR project has developed existing protocols to minimize these errors. I will review these protocols and revise as necessary as part of my data workflow development.

### QA/QC during data entry

Paper data sheets to electronic records

The process of transferring data from paper datasheets to electronic form that can be understood by a computer and use for analyses is a source of potential errors. I will work to minimize this risk of errors as part of my workflow. For electronic data, I will use automated data checks built into the system used to upload data to check for errors. For data entered by hand, I will first use a system that reduces the likelihood of an error being introduced into the data entry to start with. This will be done by using a data entry template that mirrors how the data are collected in the field on paper data sheets, with how the data are entered in the computer. This follows USGS Data Management guidelines which suggests that the most effective way to ensure data quality, is to prevent the creation of defective data. I will use a Data template structure based on USGS Data Management Standards (<https://www.usgs.gov/products/data-and-tools/data-management/quality-design-recommended-practices?qt-science_support_page_related_con=0#qt-science_support_page_related_con>).

Once data have been collected and returned to the lab the data must be converted from paper based observations to electronic records for storage and use. This conversion to electronic records also can be mandated by some granting agencies to meet a 2013 US Government Executive Order mandating publicly funded data to be available in “machine readable” format ; “data easy to find, accessible, and usable”" (see <https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->). These transfer from field observations recorded on paper to electronic records introduces a second source of potential error in the data work flow.

For the LCR project, I will design an an Excel workbook designed as a Data Template for easy data entry (Figure 5). This workbook will be been modified for data entry using “Data Validation” features in Excel that restrict the types of data that can be entered into each predefined column (Figure 2, box B2). These restrictions include the use of “drop down” style menus that require the person entering data to choose a value for entry based on a pre-populated list of values. These pre-populated list of values, such as site name abbreviations, are based on the terms defined by the data abbreviations guide for the project. Other types of restrictions include specific formatting for date or time values, as well as “limits” on observational data entered into each cell. By limiting the choice the researcher when selecting locations, dates, times, units, and measurement ranges this limits the potential for data entry errors such as capitalization or use of zeros instead of the letter O. To simplify entry, each data column matches an entry on the physical data sheet used in the field.

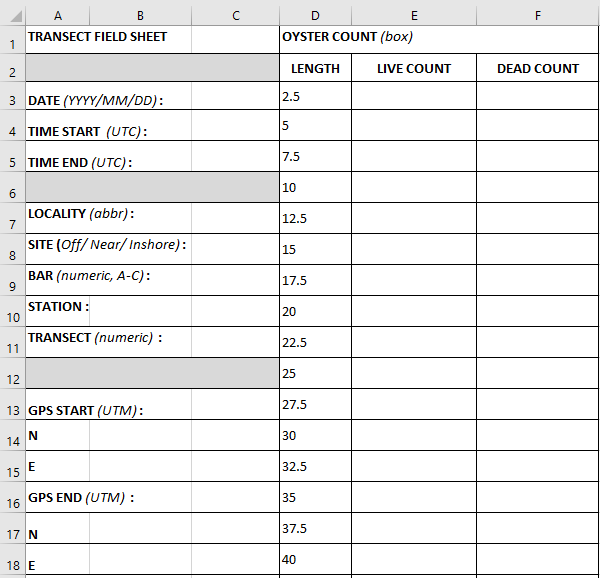


Figure 4- Physical Data sheet created and managed in Excel

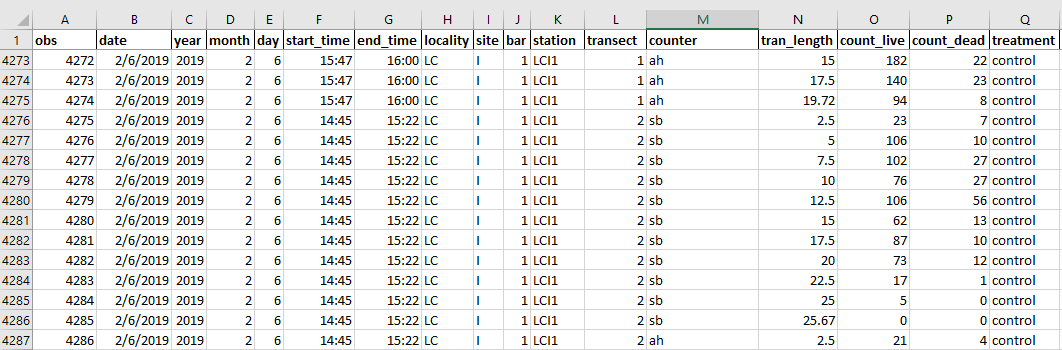


Figure 5- Data entry workbook in Excel

As an example, oyster length measurements will be restricted from being entered at a size greater than 125-mm. While oysters greater than this size could be found, to enter a value above this level requires manual override from someone with supervisory control. This data entry system will also require a “double entry” system where each line of data will be be entered into the computer twice, and then these data compared electronically. If the entered data do not match exactly, the original data sheets are examined to determine why discrepancies exist. Using different people for each round of data entry is preferred because different people may interpret the handwriting on the field data sheets in different ways. A third tab will then be used in Excel to compare the two data entry tabs for discrepancies. Any identified errors are then reconciled against the field data sheets and by a project supervisor.

Transfer electronic records from sensor to database

Once sensor collected data are transported back to the lab as individual data files on a laptop, these files must be checked for errors and the data amended to an existing database to provide a continuous record of the water quality observations of interest. I will develop a three-step process where:

Step 1. Working with UF Library team, I will develop Python code that will distinguish files from each of the two types of sensors that make up the water quality sensor array (Star-Oddi or Diver), based on proper file naming convention.

Step 2. Python code will then check for errors in these data including duplicate observations or data from a sensor that does not have an “identity” in our database.. As an example, all active and functioning sensors that are deployed in the field are stored as a data table in our MySQL database where the start day, time, and location are recorded. If the data file list of sensors does not match the list of active sensors known in the database, then an error message will be generated.

Step 3. MySQL will import all checked and correct observations in their appropriate tables.

Once imported, a second set of QA/QC protocols will be performed within the MySQL database such as examining observations for non-sense values based on expected temperature and conductivity values for the array location. While I will not develop the MySQL database as part of my thesis, I will work closely with UF Library staff to define database relationships, error checking routines, and workflow within the MySQL database. I will also develop basic Python skills to allow me to conduct routine maintenance on the database such as error checking and adding additional water quality stations as needed..

### Data analysis, figures and tables

Once data are standardized and available for use in the computer, basic visualization of the data via graphs and figures is a key next step for data checking and the beginning of the analyses. I will develop a group of data visualization products to be used both to check data from field collections and water quality sensors. These figures will be integrated with the living data such that as data are entered into the database and after they pass initial QA/QC the figures will be automatically updated to allow visual assessments of the recorded data. I will focus my efforts on creating these visualization products for the water quality data collected by the LCR project.

I will also develop a set of summary tables as part of the data workflow to provide basic information on water quality variables at different time intervals. These summary tables and figures will follow data reproducibility guidelines from USGS where the tables will be created from the living data using standard code that reproduces the same table, just with updated data, when needed. By developing code for both tables, figures, and any analyses that are reproducible as the data are updated, this will reduce analyses

Using thoroughly cleaned data, it is important to use analytical methods to determine patterns, conceive generalizations, notice biological trends, and estimate data uncertainty. Data types can be evaluated by statistical, visual, spatial, and image analysis. These evaluation types are the basis for all biological data collected. Having a reproducible way to analyze these data every time is collected, will ensure the quick feedback loop from collected data to the decision making (Figure 2). For funding agencies, having a clear data analyis workflow, allows the stakeholders to know that the data were structured the same way, every time. Having realiable data interpretation, could lead to higher rates for conservation projects to be funded in the future, and create a standard of reliability for agencies conducting research. “Interpretation is the act of using data and analytic output to evaluate hypotheses and methods, extrapolate from observations to predictions, detect patterns, and explore the consequences of assumptions”, (<https://www.usgs.gov/products/data-and-tools/data-management/analyze>).

For the LCR project, summary data tables are populated after every survey. Having updated and reproducible tables and figures is highly desired because it will save time and money on management decisions. These tables will help spot any erroneous collected observations, as well. It is necessary to check, statistically, newly collected data to ensure their integrity and biological plausability. Data analysis are conducted on ongoing newly updated data to update current biological analysis and keep tables up to date. Using R, almost all analyses can be reproduced immediately. These R scripts also check for potentially questionable data by checking for out of bound data ranges. This can be especially important in case a Data Validation tool does not spot an outlier error, or the the ranges of the data have to be adjusted.

### Version Control

Version control is defined as a software that allows for the saving and management of changes in content, documents, and other developmental information. The main focus of version control is to confirm that changes in content are intended and planned. Version control is “a tool for managing changes to a set of files” (Huang and Gonzalez 2016,<http://swcarpentry.github.io/git-novice/>). Version control can be incorporated into a data workflow using software such as Github, (Figure 2, boxes D1 and D2). The USGS Data Management Guidelines encourage the use of version control software and respositories of data and code used for projects, to allow the research data to be accessible and reproducible (<https://www.usgs.gov/products/data-and-tools/data-management/repositories>).

With living data that is accessible and used by multiple people, version control can be critical to ensuring that data are not duplicated, lost, or time wasted by not working with the proper files. The Data Carpentries provide detailed reasons for using version control (Huang and Gonzalez 2016) that can be generalized as (1) a version control system saves all versions of a file. This allows you to go back in time to old versions if needed and see what person made changes or used a particular file, (2) version control records who made what changes to specific files and (3) allows these changes to be undone if needed, (4) version control software notifies each user when there is a conflict between different people’s work such as code, (5) allows users to see when and how different files have been added or modified.

The LCR project will use a GitHub structure for version control. I will manage the Excel workbooks used for data entry and initial QA/QC in Git to allow each user to see when new data are available. I will also use GitHub to track changes in routine R files used for data summaries that are pushed to the web and included in standard reports to funding agencies. By design, the version control software tracks every possible change in the data file and allows reversion back to previous versions at any time. Once field data are reconciled they are exported as a standard CSV file and posted as a “production” copy of the data and any updates are tracked by git to document changes in individual files.

#### Proper Storage

I propose that the data workflow for both data and code scripts be separated into two modes. The first mode is “development” mode, meaning that the data are currently being processed and going through the QA/QC process. The second mode is “production” mode, where the data has been thoroughly reviewed and it is ready to conduct data analysis on. Github repositories will only have publicly available production data and scripts (Figure 2, boxes D1 and D2). These repositories will not contain individual raw sensor data files, but rather the processed and cleaned CVS file of combined processed data.

For the data and scripts that are in development mode, the proper storage for these documents will be in our projects internal server, commonly referred to as the T:Drive. This server is only available to members of the LCR project, and cannot be viewed by the public. All raw sensor data files are also located in this server. These raw files are considered to be in development mode because they have not been processed by the import and/or QA/QC scripts. It is necessary to store these files with the same naming structure, and in the same file location.

##### File and folder organization

For the LCR project, we have organized folders into categories based on mandatory tasks needed to be completed for this restoration effort (Figure 6). We have structured folders to be in this format in our internal server, to be able to locate files easily. It is beneficial to map out how folders will be arranged prior to the start of the project, since this will ensure that files will be stored correctly and that users will know where files are located, saving team member time. Usually meta-data, such as “readme” file are created to keep a written record of the folder structure.

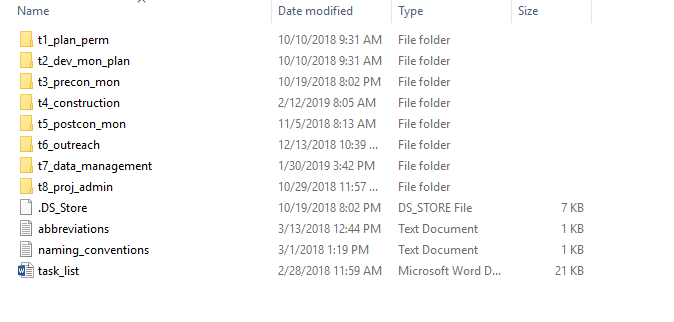


Figure 6- LCR current working folder organization

##### Naming conventions for files

It is imperative to establish naming conventions for file types and folders. I will develop a naming structure that requires that all files start with the date of creation, in the format YYYYMMDD. Each file will have additional information, that will usually have a prior set of approved abbreviations, after the date. These naming structures automatically set all files chronologically, so there is very little confusion on when the files were created (Table 1). It is also recommended that all files should be either all uppercase or all lowercase letters, instead of a combination of both. For the LCR project, I will propose as part of the naming convention standards that all files are lowercase and the context of the file names are separated with an underscore. If files are not named correctly, they will be renamed to follow our guidelines. Files that are not named correctly, also have the risk of being overlooked, or re-organized in an incorrect folder. Correct naming conventions are critical to create the correct interface between the field collected water quality sensor data and the Python code that reads and stores these data.



Table 1- Example of file naming structure

As per USGS Data Standards, naming conventions are necessary to make data easier to use, to integrate and to share. This is especially true because data that are represented will be in a format that has already been established and planned (<https://www.usgs.gov/products/data-and-tools/data-management/data-standards#examples>). Creating a table beforehand, on how each data type will be named, formatted, and defined will provide data integrity and accuracy (Table 2).

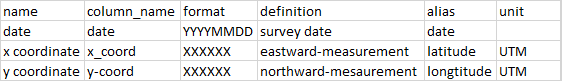


Table 2- Data Standard examples

### Discusssion

Using the Lone Cabbage Reef Restoration as an example, I will develop a data workflow that is adaptable to multiple types of data and meets best practices for data validation and reproducibility. I will use the LCR restoration project as a case history to develop this data workflow. This data workflow will integrate living data from observations recorded on paper data sheets and electronically recorded from sensors that monitor water quality.

The main goal of this workflow is make these data available for rapid analyses to adaptively assess the LCR restoration project to inform water quality and oyster monitoring efforts. This will help to meet the adaptive management requirements for this project by providing the data in a structure that allows rapid assessment and evaluation to inform decision making related to the ongoing monitoring efforts (Figure 2, box E). To do this, these data must be properly processed and managed to support reproducible analyses. My project will ensure that best practices are established and followed for data input, management, and basic summaries and visualization. This information will be useful for (1) increasing efficiency in the LCR project. The LCR project involves a large restoration project as well as integration of historical data from two other sampling epochs. Because a single data management workflow was not used across these epochs, significant effort has been required to standardize existing data. By establishing a data workflow at the beginning of the LCR restoration epoch, the data will be managed in a common structure over the life of the project. These productive data are used to make decisions in future conservation efforts. Having precise knowledge of biological data interpretations, will ensure both time and money are being used efficiently. (2) This data workflow will inform a variety of short-term decisions that must be made to adaptively improve the ongoing LCR monitoring efforts. As an example, sampling frequency, sampling locations, and sampling times of both the oyster populations and water quality can be informed by rapidly processing existing data. This can prevent data gaps from occurring from events such as fouling of water quality sensors. By knowing the rate that the water quality sensors biofoul, we can increase our sensor servicing interval to clean the sensors before their precision is compromised. Because of quick data interpretation turn around time, we were able to determine that fixed sensors in warmer summer months experience more fouling, and must be serviced more frequent than 14 days, usually about 9-10 days. Being able to keep sensors free of fouling, we are able to ensure proper and continuous data measurements of the sensor. We can also determine, whether or not a fixed sensor is in an appropriate location to measure changes in the parameters of interest such as changes in salinity following reef construction. (3) Long-term decisions as part of the adaptive management process of this project can also be informed by this data workflow. For example, this project is one of the first large oyster restoration projects funded in Florida by GEBF. Oyster reef restoration is a common topic for other possible projects and the LCR project can provide information on how funds could be allocated for sampling trips, surveys and equipment. Overally well designed data workflow programs are critical to meeting basic requirements of an adaptive management plan. When combined this approach can be highly effective in maximizing the effectiveness of conservation actions such as the LCR restoration in a cost effective manner.

This foThe main goal of designing an adaptive data management plan is to maximize funds, man power, and data feedback loops. The LCR project is funded by NFWF, and due to the Deepwater Horizon oil spill, a total of $2.544 billion dollars has been allocated to benefiting natural resources along the Gulf Coast. It is ideal to maximize the impact of these funds as much as possible. These funds essentially are to provide support for informed decision making when it comes to managing ecological restoration efforts. Good data management decisions, that are designed early, are predicated on following a streamlined and reproducible process. USGS Data Science guidelines conclude with, “The [USGS’s] most valuable resource (its people) cannot operate without good, solid, accurate, reliable, useful, and timely data. Furthermore, taxpayers and [USGS’s] customers have paid for and are entitled to know the factual basis for [USGS] decision making (in other words, the data relied upon to make those decisions)”, (<https://www.usgs.gov/products/data-and-tools/data-management/why-manage-data>). Collecting and trying to manage poorly handled data will slow learning, waste money, and ultimately contribute to incorrect decision making.

### Decision Making

## References

Barone, Lindsay, Jason Williams, and David Micklos. 2017. “Unmet Needs for Analyzing Biological Big Data: A Survey of 704 Nsf Principal Investigators.” *PLoS Computational Biology* 13 (10). Public Library of Science: e1005755.

Holling, Crawford S. 1978. *Adaptive Environmental Assessment and Management*. John Wiley & Sons.

Lindenmayer, David B, and Gene E Likens. 2009. “Adaptive Monitoring: A New Paradigm for Long-Term Research and Monitoring.” *Trends in Ecology & Evolution* 24 (9). Elsevier: 482–86.

Lowndes, Julia S Stewart, Benjamin D Best, Courtney Scarborough, Jamie C Afflerbach, Melanie R Frazier, Casey C O’Hara, Ning Jiang, and Benjamin S Halpern. 2017. “Our Path to Better Science in Less Time Using Open Data Science Tools.” *Nature Ecology & Evolution* 1 (6). Nature Publishing Group: 0160.

Walters, Carl J. 1986. *Adaptive Management of Renewable Resources*. Macmillan Publishers Ltd.

Walters, Carl J. 2007. “Is Adaptive Management Helping to Solve Fisheries Problems?” *AMBIO: A Journal of the Human Environment* 36: 36–36–4.

Yenni, Glenda M, Erica M Christensen, Ellen K Bledsoe, Sarah R Supp, Renata M Diaz, Ethan P White, and SK Morgan Ernest. 2018. “Developing a Modern Data Workflow for Living Data.” *BioRxiv*. Cold Spring Harbor Laboratory, 344804.