**Chapter 1**

**Data Management Workflow**

#### 

#### Background and Context

Traditional field biology programs, many of which are designed to monitor animal populations and their environments, have experienced substantial changes in data collection, management, and storage technology in recent years. Changes include new sensor technology, data collection methods, and data observing platforms that are being used in large-scale monitoring programs including SECOORA (Southeast Coastal Ocean Observing Regional Association) and NEON (National Ecological Observing Network). As an example, advancements in sensor technology have allowed for significant changes in water quality monitoring such as transitioning from discrete single location and single point in time sample collections to real-time continuous observations at multiple locations. While the scale and technological capacity of many monitoring programs has increased these monitoring programs are still most often conceived, planned, and used by personnel trained as biologists and not data scientists. The lack of training in basic data management, curation, and workflow of data generated from these types of data collection platforms was demonstrated in a recent NSF (National Science Foundation) survey (Lowndes et al., 2017) which highlighted that of the 704 scientists who participated in the survey, “data skills” (e.g. multi-step workflows, ability to store, share and publish data) was identified as the largest unmet need (Barone, Williams, and Micklos, 2017; Lowndes et al., 2017).

**Continuous Data, Management and Analyses**

The US Gulf of Mexico region is undergoing a large restoration effort to reverse observed declines in key ecosystem components including seagrass, fish communities, and oyster reefs using funding from the consolidated Deepwater Horizon settlements (see https://www.nfwf.org/gulf/Pages/home.aspx as an example). These restoration projects vary in spatial scale and funding, but, like other restoration efforts, these projects will have data collection and evaluation efforts that occur frequently throughout the project.

Several of the ongoing restoration programs in the Gulf of Mexico have similar objectives such as restoring oyster or seagrass habitats. For programs with similar objectives, restoration actions may be improved by adaptively informing ongoing restoration while the restoration is taking place, or informing restoration based on results from a similar study elsewhere, to help improve restoration outcomes. For this type of system to work efficiently, a workflow must be developed that captures data as it is collected, and guides this information from collection, to analyses, and data storage. Developing this type of data system is essential to improve data quality by reducing the likelihood of data collection errors or reducing the availability of data for use by other researchers conducting similar restoration efforts.

**Adaptive Management**

“Living data” are defined as data which are continuously collected and updated. These types of data are critical to this type of adaptive learning to inform restoration and management actions (Yenni et al., 2018). These informed adaptations during a restoration project can be small such as shifting the location of an autonomous sensor, to larger changes including restoration practices or revamping of sampling programs because of low statistical power. Living data are challenging to work with from a data management perspective because the data (by design) change as new data are collected and usually frequently. In a restoration or management context as these data are being collected, they must be processed, and analyses of these data to be 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).

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. This process is repeated (Figure 1) to improve management actions such as identifying the best restoration approach. To carry out a restoration project adaptively, data used in these continuous efforts meet quality assurance/quality control (QA/QC) protocols to identify and correct inconsistencies and errors in field or sensor observations before these data are used in an analysis. 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.

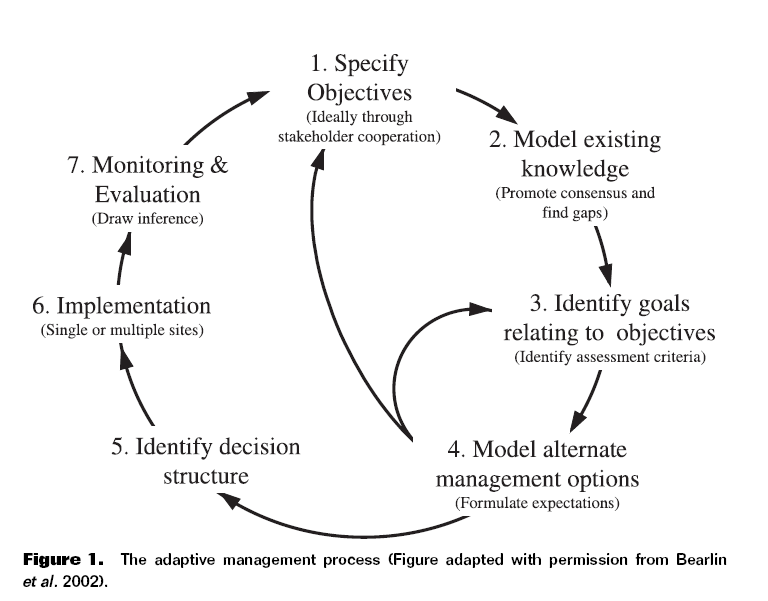
An example of programs that require an adaptive management approach to restoration are projects funded by the National Fish and Wildlife Federation as part of the Gulf Environmental Benefit Fund (NFWF-GEBF). These projects explicitly require an adaptive management plan to guide the restoration process. Extensive data management plans are also mandated to capture data collected and analyzed 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 part of the Gulf of Mexico funded by NFWF-GEBF. The project’s primary goal is to restore specific historical oyster reefs so that they may be resilient to changing sea level and river discharge. This project generates data from multiple sources including continuous autonomous water quality data via sensors and observations of oyster populations by 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 discrete 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 and questions. Because this is a long-term restoration project with numerous uncertainties in how ecosystems 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. Creating this workflow allows for rapid analyses of data to inform decision making related to sensor deployment or modifying the reef restoration process through additional construction efforts.

Figure 1- The adaptive management process (Schreiber et al., 2004, adapted from Bearlin et al., 2002).

### Implementing a modern data workflow

Creating and automating a data management workflow for living data is an emerging skill for natural resource professionals. More than ever, data management is recognized as a core skill for biologists and ecologists (Hampton et al., 2017). Even though the design of workflow is specific to the LCR 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 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 used freely available open source tools including program R (<https://www.rstudio.com/>) for data QA/QC, analysis, and visualizations, and GitHub (<https://github.com/>) for version control. Steps 1, 3, 4, and 7, in Figure 1, are directly used in my proposed workflow, with applicable modifications.

## Workflow methodology

The workflow methodology described can be used in a broader sense for many biologically data driven efforts. Though the LCR restoration project is the case study for these workflow methods, the QA/QC, analyses, version control and adaptative management decisions steps are flexible in nature.

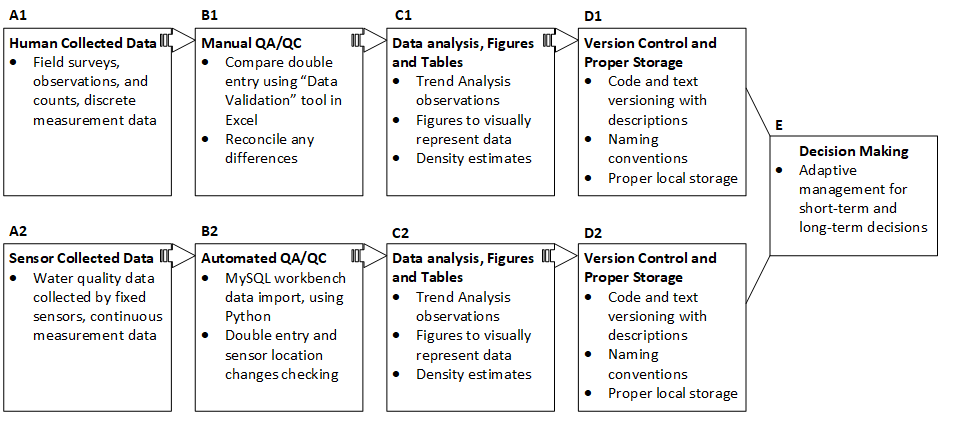


Figure 2- Data management workflow designed for the LCR restoration project.

### Reducing errors

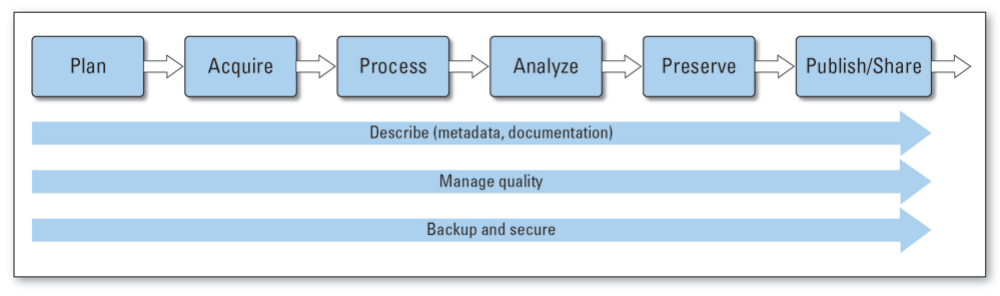
The goal of a successful data management plan is to minimize errors in collected data. Often, the first step in the data collection process is transcribing an observation in the field to paper data sheets for analyses back in the lab. This simple effort of recording the data in the field is the first 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 on a sheet or the person recording the data may be unfamiliar with terminology or protocols. To minimize these types of mistakes it is best to follow proven practices for data management such as those from the United States Geologic Survey (USGS) who 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 types of data guidelines define the basic types of data that are to be collected including date and time, site naming conventions, and units of measurements for specific observations. This type of predetermined information is a key first step in reducing the risk of this type of data error 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 physically collected. Errors in site names can place the data observations in the wrong location spatially. To minimize this risk, whenever possible data sheets can be pre-populated with as much information as possible before going into the field.

#### Field Data Sheet Recommendations

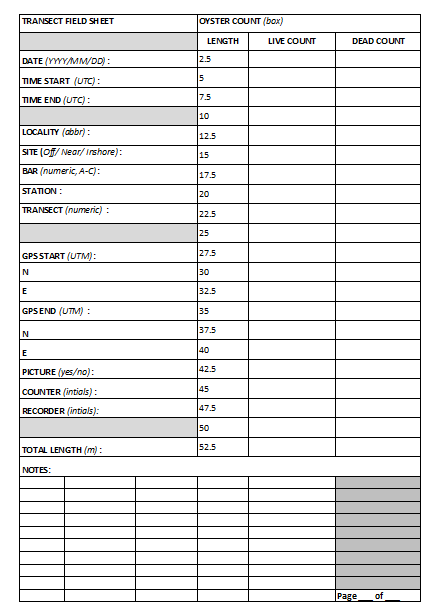
For the LCR restoration project, observational data collected in the field primarily consists of oyster counts and size measurements from line transects among randomly selected oyster reefs delineated into strata based on specific research questions, which are then recorded on waterproof paper data sheets (Figure 2, Box A1). To reduce the chance of field errors and generally save time, we pre-populate fields in data sheets whenever possible with basic information including, date and location following data naming standards and field protocols (Figure 4). Clear and concise fields with their applicable units ensures that the recorder of this datasheet knows what they are writing and how they should write it.

Figure 4- Field data sheet template created and managed in Excel with labels and units for each field.

#### Sensor Collected Data Recommendations

Sensor collected data differs from human collected data, in that sensor data are measurements recorded by an instrument automatically. These types of data are a common component of many large-scale observational 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 (https://github.com/USGS-R/waterData) or wind observations from a National Oceanic and Atmospheric Administration (NOAAA) weather buoy (https://github.com/ropensci/rnoaa), which can be accessed online by APIs or software such as R.

The LCR restoration project has a small array of sensors (N=10) that track the temperature (°C) and conductivity (μS/m) 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 the associated data files are downloaded from the receiver (Figure 2, Box A2). Sensors are serviced bi-weekly to ensure functionality.

An individual sensor data file with 14-days’ worth of observations contains about 900 lines of data and a total of about 400 observations. While the observations are collected automatically, there are still opportunities to introduce errors when these data are collected. This can include 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. Reducing these error opportunities ensures a continuous sensor stream of interrupted measurements.

#### Paper data sheets to electronic records recommendations

The process of transferring data from paper data sheets to electronic form, which makes it compatible with a computer for data analyses, is the most common source of potential errors. Minimizing the risk of errors is the main characteristic of the workflow design. For data entered by hand a data entry system that reduces the likelihood of introducing errors via data entry is used. This is done by using a standardized template, so data sheets and digital spreadsheets are input similar ways. 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. The LCR restoration project uses 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>).

For the LCR restoration project, a designed Excel workbook is used and is inteded as a Data Template for easy and efficient data entry (Figure 5). This workbook is modified for data entry using “Data Validation” features in Excel that restrict the types of data that can be typed into each predefined column (Figure 2, Box B1). These restrictions include the use of “drop down” style menus that require the user entering data to choose a value for entry based on a pre-populated list of values. These pre-populated lists 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 in each cell. By restricting the choice of the user when selecting locations, dates, 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.

As an example of the capability of “Data Validation” features, oyster length measurements are restricted from being entered at a size greater than 125-mm and give an error message to the user if done so. While oysters greater than this size are observable in nature, to enter a value above this level requires manual override from someone with supervisory control. This data entry system also requires a “double entry” system where each line of data are entered into the workbook twice, typically by separate users, and then these data are compared electronically. If the double 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 differently. A third tab is be used in Excel to compare the two user data entry tabs for discrepancies. Any identified errors are then reconciled against the field data sheets and by a project supervisor.

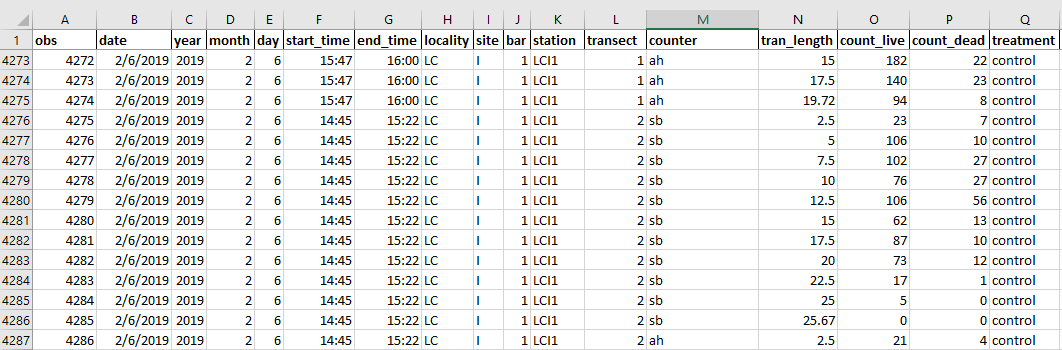


Figure 5- Data entry workbook in Excel to enter in data collected from the field. Each column is restricted on what information can be entered into it.

#### Transfer of electronic records from sensor to database

When individually collected sensor data files are transported back to the lab, 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 (Figure 2, Box B2). A three-step process has been developed where:

Step 1. Working with University of Florida Library team, Python code has been developed that distinguishes 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 conventions (i.e 20200219\_wq6\_star.dat).

Step 2. Python code checks data for errors including duplicate observations or checks that the sensor is identified properly in our database. As an example, all active and functioning sensors, which are deployed in the field, are stored in 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 is reported.

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

Once imported, a second set of QA/QC protocols is performed within the MySQL database, which examines observations for non-sense values based on expected temperature and conductivity values for the array location. While I have not directly developed the MySQL database, I have worked closely with University of Florida Library staff to define database relationships, error checking routines, and workflow within the MySQL database.

Developing a database to store the water quality datasets, prior to data collection, has given the project a security in knowing that every possible instance of a mistake or error has be thoroughly thought through. Some mistakes that have been taken into consideration in the database are the possible locations of the sensors. Sensors can move from one location to another if need be, and the database has a “check-in” and “check-out” procedure to ensure that only the data from active sensors can be imported. Thinking through the possibilities of how data collection can possibly be entered or imported incorrectly, is a major advantage in creating a workflow design.

### Data Analysis, Figures and Tables

Once data are standardized and available computational use, basic visualization of the data via graphs and figures the next step for data checking and the beginning of the analyses (Figure 2, Boxes C1 and C2). A group of data visualizations has been developed to produce products to be used both to check data from field collections and water quality sensor data. These figures are integrated with the living data such that as data are entered into the database, and after they pass initial QA/QC, the figures are automatically updated to allow visual assessments of the recorded data.

A set of summary tables has been developed as part of the data workflow to provide basic information on water quality variables at different time intervals. These summary tables and figures follow data reproducibility guidelines from USGS where the tables are created from the living data using standard code that reproduces the same table and adding newly updated data when needed. Many functions from the R package `tidyverse` (<https://www.tidyverse.org/>) are maintained and updated regularly to ensure compatibility with other existing functions, are be used to create these tables and figures. By developing code for tables, figures, and any other reproducible analyses as the data are updated, total time for data feedback loop is reduced.

### 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 over time. The key purpose of using version control software is to document and confirm that changes in content are intended and planned. Version control can be incorporated into a data workflow using software such as Github (https://github.com), (Figure 2, Boxes D1 and D2). The USGS Data Management Guidelines encourage the use of version control software and repositories for data and code used for projects, which allows the project data analysis to be accessible and reproducible (https://www.usgs.gov/products/data-and-tools/data-management/repositories).

Version control can be critical to ensuring that data are not duplicated, lost, or time is not wasted by not working with the proper files. The Data Carpentries (https://datacarpentry.org/) provide detailed reasons for using version control (<http://swcarpentry.github.io/git-novice/>) that can be generalized as (1) a version control system saves all versions of a file, (2) version control records who made what changes to specific files and makes the user write detailed notes about what they changed (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.

The LCR restoration project uses use a GitHub structure for version control (https://github.com/LCRoysterproject). 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.

##### Naming conventions for files

I will develop a naming structure which will require files start with the date of creation, in the format YYYYMMDD. Each file will have additional information which will usually have a prior set of approved abbreviations, after the date. One advantage of this naming structure is that all files will be ordered chronologically when sorted by name, so there is very little confusion on when the files were created (Table 1). Following guidelines from USGS Data Standards suggest that file names should be in 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).

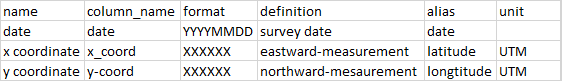


Table 2- Data Standard examples of creating a definition list for naming conventions.

#### Data 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 data that are currently undergoing a QA/QC process. The second mode is “production” mode, where the processed data are ready to be analyzed. Github repositories will only have publicly available production data and scripts (Figure 2, Boxes D1 and D2). Raw sensor data files will not be found in these repositories.

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 (Figure 2, Box D2). This server is only available to members of the LCR restoration project is not publicly available. Raw sensor data files would be stored and archived in this server. However, for other projects it would be advised to look into a protected and secure server to store raw data files.

### Discussion

Using the Lone Cabbage Reef restoration project as a case history, I will develop a data management workflow that is adaptable for multiple types of data and meets best practices of data validation and reproducibility (Table 3). This data workflow will integrate living data from observations recorded from paper data sheets and autonomous sensors that monitor water quality.

|  |  |
| --- | --- |
| Organization/ Person | Deliverables |
| ARCS (Academic Research Consulting and Services) | 1. Data management to UF faculty through the lifecycle of their data 2. Providing researchers, a competitive edge to secure funding 3. Creating plans that follow funding agencies compliance |
| LCR Project | 1. Executing Data Management and Access Plan:  * Observational data including autonomous stations and oyster reef sampling * Management of specific data products * Quality Assurance/Quality Control * Access to Data and Data Sharing Practices and Policies |
| Proposed research | 1. Investigate data management needs for LCR project 2. Design and create data management workflows as it pertains to the LCR project, for each data sampling type 3. Continually updating workflow for maximum efficiency 4. Ensure data management plan is compliant with USGS Data Standards |

Table 3- Deliverables comparison of ARCS, LCR Project and my proposed research.

The main goal of my proposed workflow is to make data available for rapid analyses to adaptively assess the LCR restoration project and ongoing 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 restoration project. The LCR restoration 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 data will then be available and used to make decisions related to future conservation and restoration efforts like the LCR restoration project. Having precise knowledge of biological data interpretations, will ensure both time and money are 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 biofouling of water quality sensors. (3) Long-term decisions as part of the adaptive management process of this project can also be informed by this data workflow. Overall 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.