Chapter 1- Data management workflow

Introduction

* 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 how data are collected.
* Technological changes including new sensor technology and data collection platforms has caused rapid changes in the spatial and temporal scale of data collected. As an example, improvements 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 single locations or for some variables large spatial areas from satellite or other remotely sensed data collection platforms.
* OK a good transition then would be to do something similar to the Lowndes et al. 2017 paper and reference that paper, Yanni and the Barone et al. paper and point out how biologist are not trained in how to effectively manage these data
* You could point out how analytical methods for doing analyses have proliferated and increasingly biologist work in large teams with complex large datasets.
* But biologists are not trained in basic data management, curation, and workflow
* This creates problems because no matter how good the analyses methods are, if the data are not well managed then the data may not be accessible, that could be lost or not recoverable, or there could be mistakes in analyses and interpretation. These mistakes can cause problems because policy and management decisions that can impact people’s livelihoods (such as harvest regulations) may be incorrect because they are based on erroneous analyses from bad data.
* Related to this, expenditures of public money for research projects or large restoration efforts are dependent on good decisions being made based on the data collected
* Increasingly granting and governmental agencies require data management and analytical plans to promote best data management practices. See USGS Fundamental Science Practices as example
* <https://www2.usgs.gov/usgs-manual/500/502-1.html>
* Biological data have historically been management ineffectively, i.e storing on loose hard drives, not digitizing or entering data from physical data sheets.
* Biological data analysis are not always reproducible for further analysis, especially by other organizations.
* Difficult to provide data checks and quality control manually, for large sets of data.

Objectives

* Creating and maintaining a workflow specifically for the Oyster Restoration Project, and for general data management workflow use.
* Highlighting the best ways to manage different types of data for continuous and spatial data.

Implementing a modern data workflow

* Increasing need for computer and analysis techniques to be a core skill for biologists.
* Need for flexibility in data management.

Data management tools

* Excel - computing
* R - computing
* Python - computing
* Git / Github – version control
* ArcMap- mapping and geodatabases
* MySQL - database

QA in data entry

* Water quality QA in MySQL checking for double entries.
* Field technician check on physical data sheets.
* Additional code to find outliers in data, i.e values over 40 ppt for salinity

Adding data to databases on Github

* Version control with Github, when adding new data. Able to add and remove data through “committing” and “pushing/pulling” changes.

Automated updating of supplemental data

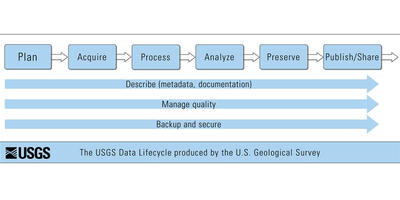
* Creating .rmd for figures or analysis on newly imported data.

Discussion/ Synthesis

* Many new journals are increasing their requirements for data management and sharing.
  + This is a key point, it is not just journals but granting programs require this, including the grant that funded the Lone Cabbage project. Agencies also require this as part of various laws and guidelines. I think this should really be in the introduction and likely the first paragraph. Take a look at this
  + <https://www.usgs.gov/products/data-and-tools/data-management>
  + This site was the model for the data sharing plan I developed for the Lone Cabbage project.

Another example is here: <https://www.nsf.gov/bio/pubs/BIODMP_Guidance.pdf>

* Discuss limitations.
* Discuss how other data types can benefit from the data management.
* Discuss advantages of the approach
  + can be used by many people not necessarily IT
  + Maybe create a test repository



**Plan**

Data Management Plan

* + Creating and continuing to
  + Data management checklist for each sample data

Data Stewardship Roles and Responsibilities

* Having rules for each user on what data they can edit

**Acquire**

Methods of acquiring data, to this project

* Field measurements from sampling
* Data from other cooperators

Data and File format standards

* Standardization for all data collected, date, coordinates, measurements
* Creating a data dictionary

Organize Files/ Data

* Determine the format of each data set collected/entered
* Determine the storage requirement of each type of data set
* Use logical naming conventions and readmes for clarification

**Process**

Workflow capture

* Describe technology and software needed for data processing
* List all inputs and output
  + If an R code will output a .csv
* Documenting the process with steps
* Creating best practices, a possible checklist
* Some mentions in USGS about version control, publishing data and analysis to be public

**Analyze**

Data Analysis

* Statistical analysis, i.e avg daily mean for salinity
* Visualizations for cleaned and analyzed data
* Spatial analysis of cooperator data and collected salinity data, and shoreline changes
* Analysis Documentation
  + Notes in scripts
  + Readmes
  + Documenting steps in analysis

**Preserve**

Repositories

* Storing final data and analysis within repositories, Github

Archiving

* Long-term storage, T Drive

**Publish/Share**

Data Release

* Email releases for major Github repository publishes