**Introduction**

The purpose of this document is to create a baseline of how data collected as part of the Lone Cabbage Reef (LCR) restoration project should be organized and managed. The recommendations of this document should be considered and included in design efforts prior to data collection. It is common to see data in formats that are not consistent, not organized, and not uniform. This makes data analyses difficult. Biological data has the extra disadvantage of only being interpreted by the context and units of which it is collected. If these units change, or are unknown, additional errors can be introduced into the analyses. This document will provide guidelines on how to set up a “packet” that will manage data entry and organization and identify the purpose and a meaning behind every observation collected into the field and entered electronically for analyses.

**Methodology**

The methodology behind creating a data packet will ensure a smooth transition from data entry to data analysis. Many functions in R, or other programming languages, require consistent data format layouts. The methodology promotes data integrity standards to ensure data reliability and consistency. It is imperative to discuss what data will be collected prior to entering data. Creating a data packet will give a better insight into how and what is being collected in the field and what data are necessary for the project to answer the biological question motivating the research. For these reasons, it is imperative to create the data packet prior to any data collection.

Overall, this packet is designed to describe data collected as part of the oyster monitoring aspects of the Lone Cabbage Reef project. However, it is important to note these types of data packets can be applied to broader ecological projects.

**Long Format**

Long format data are defined as a data frame with each row containing ONE observation, and with multiple columns describing the observation, such as date, time, and location. Long format data also specifies that each of the biological observations are fully defined within the row. It is recommended to use this format for all data entry regardless as to when the data analyses will occur. Taking time and concentrating on the best way to enter data for future use is crucial in what we do as

<https://www.theanalysisfactor.com/wide-and-long-data/>

<https://en.wikipedia.org/wiki/Wide_and_narrow_data>

**Terms and definitions**

Packet- An Excel workbook containing sheets/tabs that contain specific functions and data

Sheet/tabs- A function in Excel to create a view to separate data

<https://www.wikihow.com/Add-a-New-Tab-in-Excel>

* Sheets can be read into R using the `read\_excel` package

Column- A vertical data group starting with a header

* Headers should be all lowercase, with no spaces, and words separated by a underscore “\_”
* Columns are required to have a predetermined data type such as date, time, numerical, character

Row- A horizontal data containing only one biological observation

Cell- An Excel cell that contains only ONE piece of information, guided by rows and columns to determine what that information is

GitHub repository- Online repository at github.com that will be the version control software required for this project. View the GitHub workflow documentation for more information.

**LCR Project Expectations**

Project expectations include knowing when and where data have been collected, and who has entered these data. Data collected for the project will be used for combined (i.e, multiple years) and independent (i.e., individual year) analyses. Data must be structured in the same way, so that they can be analyzed with ease. Data integrity is necessary in this project as part of reproducibility standards required by funding agencies and best scientific practices (<https://datacarpentry.org/rr-intro/aio.html>). These points cannot be emphasized enough and following the packet structure will help meet these expectations.

**LCR Project and Double Entry**

LCR project standards include a double entry system for field collected data in Excel when entering biological observations. A double entry system refers to entering biological data twice, by different users, to ensure data integrity. Once the data is entered by the two users, these entries are compared to each other, differences reconciled, and the final data approved by a third party (Research Coordinator or PI). This is standard practice in many data collection efforts and follows USGS guidelines. Currently, the LCR project uses double entry practices in transect, spat, and oyster height observations through an Excel worksheet packet.

**Packet Requirements**

The packet requirements are not optional. All the sheets with the Excel workbook are needed for the packet to be a success. The formation and completion of the packet will keep data organized. Additional sheets can be created in the packet, but the packet must have a minimum of the required sheets, which are defined in this document.

**Sheet 1- Physical Data Sheet**

This is a copy of the data sheet as used for field data collections. By creating this datasheet as part of the packet, the same data sheet will be used for each sampling effort. The field data sheet includes all parameters needed for the data collection. Examples of data collected and data standards are:

- Date- YYYMMDD in UTC

- Location (GPS) in UTMs and decimal degrees

- The reason for decimal degrees is that this coordinate type is easily read into programs such as ArcMap, R and GQIS

Data sheet should also be:

- Clearly marked areas for observational counts/ measurements

- Formatted in a way to include units with all fields as well as format type

A screenshot of a computer

Description automatically generated

- It is advised to keep the physical data sheet in the packet, to ensure that all data needs are represented in the sheet.

**Sheet 2- Data Entry Sheet 1**

This is the first data entry sheet, which includes concise names for the columns. Columns must have no spaces in-between, preferably using “\_”. Columns should be in all one letter case, preferably lowercase.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **obs** | **date** | **year** | **month** | **day** | **start\_time** | **end\_time** | **locality** | **site** | **bar** |

-Units are not needed in the column names, because the units will be specified in the meta data sheet (Sheet 7).

- It can also be noted to `freeze` the top row of this sheet so the user entering the data can see which data are need for the specific columns (https://edu.gcfglobal.org/en/excel2013/freezing-panes-and-view-options/1/).

- Test the feasibility of entering data prior to data collection. This step can help the project coordinator/student to double check the information being entered is what corresponds to the physical data sheet (Sheet 1) prior to data collection.

**- This sheet will have data validation parameters that will be set up from the data validation pick list (Sheet 6). This pick list is critical as it defines the naming convention of all sites and also includes built in data checks such as minimum/maximum size of oysters possible.**

**Sheet 3- Data Entry Sheet 2**

This is the second attempt at data entry, which should have the exact same columns in the same order, as the first data entry sheet. Sheet 2 and Sheet 3 will look and be exactly the same.

- All of the same parameters in Sheet 2 will be applied to Sheet 3. Again, units are not needed in the column headers, they will be defined in in the meta data sheet (Sheet 7).

- Sheet 3 format should be exactly the same as Sheet 2. The columns need to be in the same order, and have the same names.

- To keep accurate data entry, the second user must enter the same data in the same order as the first data entry user. If the data are not entered in the same way, Sheet 4 will come back saying that all entries are non-matching.

**- This sheet will have data validation parameters that will be set up from the data validation pick list, (Sheet 6).**

A close up of a computer

Description automatically generated

**Sheet 4- Data Validation Sheet (check)**

This sheet should be solely for checking data integrity. There is no data entry in this sheet. Do not type any data information in this sheet. This sheet will include individual cells that will need to be “checked” if Sheets 2 and 3 do not have matching data. You can use the equation for each cell in the data checking tab:

*=IF(raw\_data\_1!A2=raw\_data\_2!A2,"","check")*

-Make sure to apply this equation to all cells that will correspond to the double entry sheets (Sheets 2 and 3). This worksheet needs to include all column names, in the same order, as the double entry sheets. This is important as new data are entered on Sheets 2 and 3 the “checking” equation needs to be expanded to make sure the new entries on sheets 2 and 3 are checked.

- If a “check” appears on the cell, it is up to the data manager, or third-party individual from the two users that entered the data, to check the discrepancy. The data validation “checks” will need to be reconciled prior to the packet being accepted into its GitHub repository.

A screenshot of a computer

Description automatically generated

**Sheet 5- Sampling Progress Sheet**

Progress of the data collection, which is basically a summary of sampling events.

Include similar columns (not all will be applicable):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **QUESTION** | **STATION** | **TRANSECT** | **SAMPLED** | **SCANNED** | **ENTRY1** | **ENTRY2** | **DATABASE** |

- This sheet will be for internal purposes and not usually used for analysis.

- Additional information such GPS coordinates can also be added in these sampling trips.

- Add any additional information that is pertinent to data management.

- Include information about the data that will be useful in the future, such as gear types, or specific sampling information.

- Include any and all information that describes the entered data.

A screen shot of a computer

Description automatically generated

**Sheet 6- Pick List used for Data Validation**

This pick list will govern and validate Sheets 2 and 3. Data validation ensures that individual cells will only have specific options that can be selected, and not entered by the user. They are determined by the pick list options per column. When the user clicks on a cell in Sheets 2 and 3, they will be prompted to select one of the variables mentioned in this pick list sheet.

- The columns of this sheet need to be in the same order and same names as the double entry sheets, Sheets 2 and 3.

- Each column needs to have listed all of the possible variables that can be selected by the user. For example, for month we only have the options 1 through 12, because there are only 12 months in a year. Without this data validation it could be possible for the user to enter 13, but having a pick list with data validation steps, will ensure that no selection outside of the allowed possibilities can be chosen by the user.

More information set up a data validation pick list can be found here: <https://www.officetooltips.com/excel_2016/tips/check_data_entry_for_invalid_entries.html>

https://support.office.com/en-us/article/apply-data-validation-to-cells-29fecbcc-d1b9-42c1-9d76-eff3ce5f7249

A screenshot of a computer

Description automatically generated

**Sheet 7- Meta data Sheet**

This meta data worksheet includes the data entry (Sheet 2 and 3) column names and their parameters and units explained. All columns in Sheets 2 and 3 need to be represented in the meta data worksheet. The compilation of this sheet is also very important and highly advised.

* Include all columns and their applicable unit descriptions
* Include easy to understand language
* Include as much information describing the columns as needed

A screenshot of a computer

Description automatically generated

**Standardized column names**

obs -observation number, normally numerical and in ascending order

date – date needs to be consistent, normally in a YYYY-MM-DD or YYYY/MM/DD format

year – numerical value that is only the year of the observation, four digits

month – numerical value that is only the month of the observation, two digits

start\_time – time value in UTC

end\_time - time value in UTC

locality – standardized locality names

* LC – Lone Cabbage
* BT – Big Trout
* LT – Little Trout
* NN - No Name
* CK – Cedar Key
* CR - Corrigans
* HB – Horseshoe Beach

site – standardized site names

* I - inshore
* O - offshore
* N – nearshore

bar – numerical value of reef bars

|  |
| --- |
| 1 |
| 2 |
| 3 |
| 4 |
| 5 |
| 6 |
| 7 |
| 8A |
| 8B |
| 8 |
| 9A |
| 9B |
| 9C |
| 9 |
| 10A |
| 10B |
| 10 |
| 11A |
| 11B |
| 11 |
| 12 |
| 13 |
| 14 |
| 15 |
| 16 |
| 17 |
| 18 |
| 19 |
| 20 |
| 21 |
| 22 |
| 23 |
| 24 |
| 25 |
| 26 |
| 27 |
| 28 |
| 29 |
| 30 |
| 31 |
| 32 |
| 33 |
| 34 |
| 35 |

station – standardized station names, which are a combination of locality + site + bar. This is by design duplicative with the locality, site, bar columns to ensure the correct spatial names are used.

|  |
| --- |
| LCI1 |
| LCI2 |
| LCI3 |
| LCI4 |
| LCI5 |
| LCI6 |
| LCI7 |
| LCI8 |
| LCI9 |
| LCI10 |
| LCI11 |
| LCI12 |
| LCI13 |
| LCI14 |
| LCI15 |
| LCI16 |
| LCI17 |
| LCI18 |
| LCI19 |
| LCI20 |
| LCI21 |
| LCI22 |
| LCI23 |
| LCI24 |
| LCI25 |
| LCI26 |
| LCI27 |
| LCI28 |
| LCI29 |
| LCI30 |
| LCI31 |
| LCI32 |
| LCI33 |
| LCN1 |
| LCN2 |
| LCN3 |
| LCN4 |
| LCN5 |
| LCN6 |
| LCN7 |
| LCN8 |
| LCN9 |
| LCN10A |
| LCN10 |
| LCO2 |
| LCO3 |
| LCO4 |
| LCO8A |
| LCO8B |
| LCO9A |
| LCO9B |
| LCO9C |
| LCO10A |
| LCO10B |
| LCO11A |
| LCO11B |
| LCO12 |
| LCO13 |
| LCO14 |
| LCO15 |
| LCO16 |
| LCO17 |
| LCO18 |
| LCO19 |
| LCO20 |
| LCO21 |
| BTI1 |
| BTI2 |
| BTI3 |
| BTI4 |
| BTI5 |
| BTI6 |
| LTI1 |
| LTI2 |
| LTI3 |
| LTI4 |
| LTI5 |
| LTI6 |
| LTI7 |
| LTI8 |
| LTI9 |
| LTI10 |
| LTI11 |
| NNI1 |
| NNI2 |
| NNI3 |
| NNI4 |
| NNI5 |
| LTI1 |
| LTI2 |
| LTI3 |
| LTI4 |
| LTI5 |
| LTI6 |
| NNI1 |
| NNI2 |
| NNI3 |
| NNI4 |
| NNI5 |
| LTI1 |
| LTI2 |
| LTI3 |
| LTI4 |
| LTI5 |
| LTI6 |
| LTI7 |
| LTI8 |
| LTI9 |
| LTI10 |
| LTI11 |
| LTI12 |
| CRI1 |
| CRI2 |
| CRI3 |
| CRN1 |
| CRN2 |
| CRN3 |
| CRO1 |
| CRO2 |
| CRO3 |
| CRO4 |
| CKI1 |
| CKI2 |
| CKI3 |
| CKN1 |
| CKN2 |
| CKN3 |
| CKO1 |
| CKO2 |
| CKO3 |
| HBI1 |
| HBI2 |
| HBI3 |
| HBI4 |
| HBN1 |
| HBN2 |
| HBN3 |
| HBN4 |
| HBN5 |
| HBN6 |
| HBO1 |
| HBO2 |
| HBO3 |

counter – initials of oyster counter

|  |
| --- |
| at |
| bp |
| cw |
| kk |
| lg |
| mm |
| NA |
| pf |
| rb |
| rh |
| sb |
| sl |
| jc |
| ds |
| ah |
| dc |
| ar |
| jb |
| ja |
| la |
| ah |
| tr |
| pfat |
| tc |
| attc |
| jh |
| ec |
| sw |
| jv |

strata – the rock and harvest status of an oyster bar

* N\_NA – no harvest, no rocks
* N\_LG – no harvest, large rocks
* N\_SM – no harvest, small rocks
* Y\_NA – yes harvest, no rocks
* Y\_SM – yes harvest, small rocks

period – sampling time definition

1 - Summer 2010  
2 - Winter 2010-2011  
3 - Summer 2011  
4 - Winter 2011-2012  
5 - Summer 2012  
6 - Winter 2012-2013  
7 - Summer 2013  
8 - Winter 2013-2014  
9 - Summer 2014  
10 - Winter 2014-2015  
11 - Summer 2015  
12 - Winter 2015-2016  
13 - Summer 2016  
14 - Winter 2016-2017  
15 - Summer 2017  
16 - Winter 2017-2018  
17 - Summer 2018  
18 - Winter 2018-2019  
19 - Summer 2019  
20 - Winter 2019-2020

**Data Type Guidelines Overview**

GPS Coordinates- Decimal Degrees in UTM

Time zone: UTC

Date- Required to be in YYYY/MM/DD, keep it consistent. It can also be advised to have the year, month, and day in separate columns as well to tease apart in scripts at a later time.

Capitalization: Keep capitalization in columns completely consistent and the same throughout the packet. Normally lowercase is preferred, for coding easibility.

Missing numbers- Missing numerical values should entered as -999.

Missing characters- Missing character values should be entered N\_A.

All fields should be completed and filled per observation, and if some piece of information is missing add a missing number or missing character selection.