**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, organized, and uniform. This makes data analyses difficult because the data must be extensively cleaned and standardized. 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 the data entry “packets” are set up and used for entry of oyster data for the Lone Cabbage Reef restoration project. The goals of this packet system are to promote efficient data collection (data collected with specific purpose) and minimize data entry errors to provide an accurate and reliable data entry system to inform restoration actions.

**Methodology**

Creating a data packet will promote accurate and efficient data workflow from data entry to data analysis. By standardizing data entry, this will make analyses more efficient by reducing data cleaning and standardization efforts by having consistent data layouts. The methodology promotes data integrity standards to ensure data reliability and consistency. Creating a data packet also motivates a critical discussion of understanding what data will be collected prior to actually collecting the data and help meet project objectives. 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.

**Data format definitions**

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. This type of data structure is generally recommended for data entry regardless of data analyses approaches that may be considered (reference). This is because XYZ.

Wide format data by contrast contain each data observation as a separate column.

**Terms and definitions for the Lone Cabbage Project data packet**

**Packet**: Microsoft Excel workbook consisting of sheets/tabs that contain specific functions and data

**Sheet/tabs**: Tabs are separate virtual pages of information as a spreadsheet that are nested within MS Excel to create a workbook of spreadsheets. These individual spreadsheets are organized tabs on the bottom of the MS Excel main screen.. Using R, these can be read into R using the `*read\_excel*` or `*readxl*` packages.

**Column**: vertical data group beginning with a header (all headers are in row 1)

Column guidelines include (1) Headers should be all lowercase, with no spaces, and words separated by an underscore “\_” (e.g., date, total\_length, fork\_length, dry\_weight, season). (2) Columns are required to have a predetermined data type such as date, time, numerical, character

**Row**: horizontal data containing only one biological observation (row 1 always contains the headers of the columns)

**Cell**: an individual box in the MS Excel workbook that contains only ONE piece of information, guided by rows and columns to determine what that information is

**GitHub** **Repository**: online repository at <https://github.com/> that will be the version control software required for this project (view the GitHub workflow documentation [*github\_workflow.docx*].

**LCR Project Expectations of Collected Data**

Project expectations define essential data elements including where data have been collected and who has entered these data. Data collected for the LCR project will be used for combined (e.g., through years, over seasons) and independent (e.g., single year, single season) analyses. Data must be structured in the same way through time so analyses can be simplified and data will not have to be standardized each year separately. This also reduces risks of errors being introduced through the data standardization process and promotes best reproducibility practices (<https://datacarpentry.org/rr-intro/aio.html>) as required by funding agencies.

Double Data Entry Standards

The LCR project standards include a double entry system entering field data in a spreadsheet for storage and access for analyses. This double data entry system requires entering biological data twice in separate sheets/tabs, 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. This is standard practice in many data collection efforts and follows USGS guidelines (reference).

**Packet Requirements**

Packet requirements listed below are required for all data entry. All the sheets/tabs that create the MS Excel workbook are needed for the packet to successfully standardize and check data. Additional sheets can be created in the packet, but the packet must have a minimum of these required sheets (sheets/tabs should be labeled with what is in the parentheses next to each title):

**Sheet/Tab 1 ­– Physical Field Datasheet (*field\_datasheet*)**

This is a copy of the datasheet that is printed onto rite-in-the-rain paper and used in the field for data collection. By including this datasheet as part of the packet, the same datasheet can be used for each sampling effort and will never be lost. The field datasheet includes all parameters needed for the data collection. Examples of data collected, and data standards are:

date – YYYMMDD in UTC

location – GPS coordinates (UTMs in decimal degrees)

Decimal degrees is a coordinate type is easily read into programs such as ArcMap, R, and QGIS. The GPS units used in the field are pre-set to record in decimal degrees.

A screenshot of a computer

Description automatically generated

Figure 1. Screenshot of MS Excel “Packet” used for data entry and management. This image is of the first tab, which is a copy of the data sheet used in the field. This data sheet is the data sheet used when one person is counting the oysters along the transect.

**Sheet/Tab 2 – Data Entry 1 (*raw\_data\_1*)**

This is the first data entry sheet. Columns must have no spaces in-between separate words, using “\_” instead and are in all lowercase. For example:

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

This sheet has built in data validation parameters that are linked to the data validation pick list (sheet/tab 6). These data validation links to the picklist impose rules on data entry to minimize data entry errors in various ways. These include the use of dropdown menus that require the user to select names (such as site locations) from a predefined list or implements rules that keep entries from being made that are above or below defined values (such as oyster size).

units are not needed in the column names; units will be specified in the meta data sheet/tab (sheet/tab 7)

note that you should `*freeze*` the top row of this sheet so the user entering the data can see which data are needed for specific columns (<https://edu.gcfglobal.org/en/excel2013/freezing-panes-and-view-options/1/>)

**Sheet/Tab 3 – Data Entry 2 (*raw\_data\_2*)**

This is the second data entry sheet, which has the exact same columns in the same order as the first data entry sheet. Sheet/Tab 2 and sheet/tab 3 will look and be exactly the same.

Sheet/tab 3 format should be exactly the same as sheet/tab 2 (columns need to be in the same order with same names)

To keep accurate data entry, the second user must enter the data in the same order as the first data entry user. If the data are not entered in the same way, sheet/tab 4 will return an error message identifying the cells between sheets 2 and 3 that do not match. The cells must then be manually reconciled.

* example below:

A close up of a computer

Description automatically generated

Figure 2. Screen shot of Sheet/Tabs 2 and 3.

**Sheet/Tab 4 – Data Validation (*raw\_data\_check*)**

This sheet/tab is 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 flag with the word “check” any cells that do not match exactly between the two data entry tabs. The data checking equation in Excel is::

Make sure to apply this equation to **all cells** that will correspond to the double entry sheets/tabs (i.e., sheets/tabs 2 and 3)

This worksheet needs to include all column names, in the same order, as the double entry sheets

As new data are entered on sheets/tabs 2 and 3 the “checking” equation above needs to be expanded to ensure all entries on sheets/tabs 2 and 3 are checked

If a “check” appears on the cell, it is up to the third-party individual, different from the two users that entered the data, to check and correct the discrepancy

Data validation “checks” will need to be reconciled prior to the packet being accepted into the LCRoysterproject `*master\_data*` GitHub repository

Example below:

A screenshot of a computer

Description automatically generated

Figure 3- Screen shot of Sheet/Tab 4 and the corresponding equation for this sheet.

**Sheet/Tab 5 – Sampling Progress (*progress*)**

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

Include similar columns as previous sheets/tabs; however, not all will be applicable:

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

Additional information, such as GPS coordinates, can also be added for each sampling trip

Add any additional information that is pertinent to data management

Include information about the data that will be useful in the future (e.g., gear type, specific sampling information)

Include all information that describes the entered data; however, this should not repeat the metadata sheet/tab (sheet/tab 7; see below)

A screen shot of a computer

Description automatically generated

Figure 4. Screen shot of Sheet/Tab 5 displaying the sampling progress.

**Sheet/Tab 6 – Pick List used in Data Validation (*pick\_list*)**

This picklist will govern and validate sheets/tabs 2 and 3. Data validation ensures that individual cells will only have specific options that can be selected and not entered by the user. These are determined by the picklist options per column. When the user clicks on a cell in sheets/tabs 2 and 3, they will be prompted to select one of the variables mentioned in this pick list sheet. A range (i.e., minimum and maximum values) can also be set on individual columns.

columns of this sheet must match data entry sheets 2 and 3

each column needs to have listed all of the possible variables that can be selected by the user (e.g., month only has the options 1–12 because there are only 12 months in a year, and without this data validation it could be possible for the user to enter 13). For sites, as new sites are added then this list will have to be updated.

having a pick list with data validation steps, will ensure that no selection outside of the allowed possibilities can be entered by the user

More information on how to set up a data validation pick list can be found at the links below: <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

Figure 5. Screen shot of Sheet/Tab 6 including all columns and their allowable selections.

**Sheet/Tab 7 – Metadata (*metadata*)**

This metadata sheet/tab includes the data entry (sheets/tabs 2 and 3) column headers, their parameters, and units explained. All columns in sheets/tabs 2 and 3 need to be represented in the metadata worksheet. This sheet must include all information that describes the column data (i.e., list what each column header means and how those data were collected). This sheet is different than Sheet/Tab 5 because it provides a detailed definition of each column. This sheet/tab must include:

all columns and their applicable unit descriptions

easy to understand language

as much information describing the columns as needed

example below:

A screenshot of a computer

Description automatically generated

Figure 6. Screen shot of Sheet/Tab 7 including the definitions of all columns.

**Standardized Column Names Commonly Used:**

* **obs** – observation number, normally numerical and in ascending order
* **date** – date needs to be consistent, normally in YYYY-MM-DD or YYYY/MM/DD format
* **year** – numerical value of the year of the observation, four digits
* **month** – numerical value of the month of the observation, two digits
* **day** – numerical value of the day of the observation, two digits
* **site, locality, station, bar** – location of the sample (see below for examples)
* **start\_time** – time value in UTC
* **end\_time** – time value in UTC
* **treatment** – whether the bar was built with rocks or wild
* **strata** – size of rocks and if harvested or not
* **locality** – standardized locality names (could expand as study continues)
* 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 (could expand as study continues)

|  |
| --- |
| 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 repetitive with the locality, site, and bar columns to ensure the correct spatial names are used (could expand as study continues)

|  |
| --- |
| 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 (could expand as study continues)

|  |
| --- |
| 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 (could expand as study continues)
* 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; also separated into columns for year, month, day
* **Capitalization** – keep capitalization in columns completely consistent and the same throughout the packet; **all** **lowercase is preferred!!**
* **Missing Numbers** – missing numerical values should entered as -999
* **Missing Characters** – missing character values should be entered N\_A
* **\*\*\*** missing numbers and characters can be all “na” and removed in R easily, just keep it consistent \*\*\*

All cells should be completed and filled per observation, and if some piece of information is missing, add a missing number or missing character selection in the data validation sheet/tab.