**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 ecologists. When interpreting and analyzing these data, they need to be in an organized and easily readable format for others (i.e., if you disappeared today, someone tomorrow should be able to pick up where you left off) and coding languages (e.g., R).

Below are links to examples of long (i.e., narrow) and wide data formats:

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

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

**Terms and definitions**

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

**Sheet/tabs**: function of MS Excel to create a view to separate types of data or lists and are found at the bottom of the workbook (see <https://www.wikihow.com/Add-a-New-Tab-in-Excel>); 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)

* headers should be all lowercase, with no spaces, and words separated by an underscore “\_” (e.g., date, total\_length, fork\_length, dry\_weight, season)
* 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*] in the LCRoysterproject `*repo\_structure*` repository for more information); GitHub is a great version control software for all ecological studies

**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 (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 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 this packet structure will allow you to meet these expectations.

**LCR Project and Double Entry**

LCR project standards include a double entry system for collected field data in MS Excel when entering biological observations. A double entry system refers to 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.

**Packet Requirements**

The packet requirements listed below **are NOT optional**. All the sheets/tabs that create the MS Excel workbook are needed for the packet to successfully standardize and monitor data. The formation of this packet was strategically created to keep data organized and easily accessible. 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 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)
  + reason for decimal degrees is that this coordinate type is easily read into programs such as ArcMap, R, and QGIS
* clearly marked areas for observational counts/measurements
* formatted in a way to include units with all fields and format type
* example below:

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 are represented in the entry sheets/tabs.

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

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

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

* **this sheet will have data validation parameters that will be set up from the data validation pick list (sheet/tab 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 possible**
* 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/>)
* test feasibility of entering data prior to data collection; this step can help double check the information being entered is what corresponds to the physical data sheet (sheet/tab 1) prior to data collection; can be time consuming, yet worth it

**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.

* **must have data validation parameters that will be set up from the data validation pick list (sheet/tab 6)**
* same parameters in sheet/tab 2 will be applied here to sheet/tab 3; again, units are not needed in the column headers, they will be defined in in the meta data sheet (sheet/tab 7).
* 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 come back saying that all entries are non-matching, removing the purpose of double entry
* example below:

A close up of a computer

Description automatically generated

Again, sheet/tab 3 is identically made as sheet/tab 2.

**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 include individual cells that will need to be “checked” if sheets/tabs 2 and 3 do not have matching data. You need to use the equation below for each individual cell in this sheet/tab:

* 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

**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** |

* this sheet is for internal purposes and not usually used for analyses
* 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)
* example below: A screen shot of a computer

  Description automatically generated

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

This pick list 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 pick list 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 **HAVE** to be in the same order and use the same names as the data entry sheets/tabs 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)
* 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>

* example below:

A screenshot of a computer

Description automatically generated

**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. The compilation of this sheet is also very important and highly advised. It 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/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

**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.