**Introduction**

The purpose of this document is to create a baseline of how sampled data should be organized and arranged. The recommendations of this document should be made prior to data collection. It is common to see data in formats that are not consistent, not organized, and not uniform in nature. Biological data has the extra disadvantage of only being interpreted by the context and units of which it is collected. In this case this document will provide guidelines on how to set up a “packet” that should give your data entry and organization a purpose and a meaning behind every observation entered.

**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 similar data format layouts. The methodology creates 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 is necessary for the project/ the biological question. It is also imperative to create the data packet in full prior to any data collection.

**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 refers that each of the biological observations are fully defined within the row. It is recommended to use this format regardless if data analysis will occur shortly after data collection or sometime in the future.

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

**Terms and definitions**

Packet- A 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 “\_”
* 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 has been collected, and who has entered them. Data collected for the project will be used for combined and independent analysis. Data must be structured in the same way, so that they can be analyzed with ease. Data integrity is necessary in this project. 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 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, it is finally checked by a third party. This is standard practice in many data collection efforts. Currently, the project using double entery practices in transect, spat, and oyster height observations through an Excel worksheet packet.

**Packet Requirements**

The packet requirements are not optional. All the sheets 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**

Field data sheet including all parameters needed for the data collection. Some examples include:

- 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

Your 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.

- Test the easibility 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).**

**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.

- 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>

A screenshot of a computer

Description automatically generated

**Sheet 7- Meta data Sheet**

This meta data worksheet include the data entry (Sheet 2 and 3) column names and their parameters 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

**Common column names**

obs

date

year

month

date

start\_time

end\_time

locality

site

bar

station

counter

treatment

strata

period

**Data Type Guidelines Overview**

GPS Coordinates- Decimal Degrees in UTM

Time zone: UTC

Date- Normally in YYYY/MM/DD, but can be DD/MM/YYYY, 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 as N\_A\_ or 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.