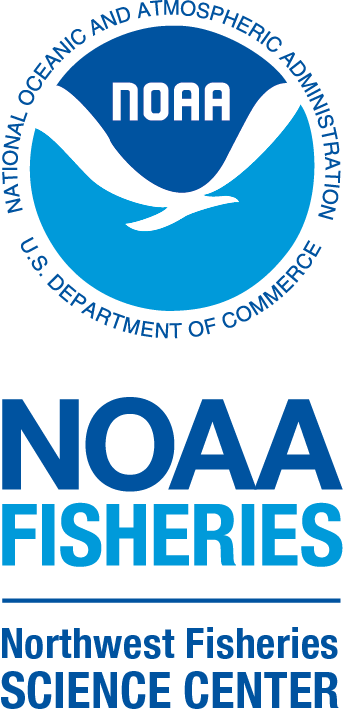
|  |  |
| --- | --- |
| FROM | Danielle Perez, Kate Rovinski |
| TO | Paul McElhany, OA Lab |
| DATE | (Wednesday) 2021.01.13 |
| SUBJ | *Draft*  Data Management Policies  Understanding types of data, handling, and storage  Protocols for OA Lab Personnel |



**Table of Contents**

1 Overview

Application of This Document

Consistency Rules -Naming Protocols

Consistency Rules -Date Formats

Consistency Rules -Colors & Highlighting

2 Definitions of Data

Raw Data

Processed Data

Data Dictionary

Meta Data

Version Control

Code, types of Code

Received Data

CSV files

Dynamic Excel files

Google Drive Documents

3 Data Organization, format, and Storage (including Repositories)

meta requirements/format

file locations - Before implementing anything (e.g. renaming and moving files )

Back ups

4 OAP & NWFSC Regulation

5. Review of All Goals and Rules Mentioned

1. **Overview** 
   1. **Goals of This Document & OA Lab Produced Data**

**Goal 01:** Format it once

**Goal 02:** Produce robust, re-usable data

Reusable Data has additional information for complete user understanding about collection and any manipulation/ filtering/ formatting

**Goal 03:** Provide lab members a clear idea about preparing, disseminating, and archiving data

**1.1a Consistently Rules – Naming Convention**

**1.1a.1 Rule 01: Be consistent especially with categorical variable names**

**Use consistent codes for categorical variables**. For a categorical variable like sex, use a single common value for males (e.g. “male”) and a single common value for females (e.g. “female”).

* Pick one method of referring to something and stick to it.
* Keep it short but keep it obvious
* Use all lowercase letters unless referring to an acronym
* Four- five digits is the ideal but, be sure that the shortened expression can’t be confused with another value. This is to help identify variables- especially categorical values
* Example: Describing the first Mobile Ocean Acidification Treatment System replicate 01 could be written a number of ways. The character "1" or "01" could work however, "M01" or "MOATS01" can't be mistaken for scalar values

Typical Categorical Variables Examples

|  |  |  |  |
| --- | --- | --- | --- |
| **Long Descriptive Name** | **Best** | **Better** | **Good** |
| Mobile Ocean Acidification Treatment System replicate 01 | MOATS01 | M01 | 1 – but could be mistaken as a scalar |
| Treatment | treatment | NA | NA |
| Temperature | temperature | tempC | | Temp – could be confused with as temporary value | tmp – could again could be confused with as temporary value |
| High CO2 | hiCO2 (or) HiCO2 | HighCO2 | (not good) High CO2 **NO SPACES!!!** |
| 1hr.treatment | onehr -Can’t start variable names with scalars | One.hr | (not good) **can’t start with numbers!!!** |

|  |  |
| --- | --- |
| Using Digits at the End of Variable/database Names | Logo  Description automatically generated |

**Use consistent variable names.** The repose metric should appear somewhere in the variable name with numerical descriptors separated with an underscore or period. Never leave spaces in variable names or in file names. If the response metric isn’t conducive use the test species/animal/invertebrate

Examples: crab | juvenile | zoea | megalopae | survival | slope | krill | coho | etc.

Examples: krillsurv\_10wks

Don’t then allow substitutions: krillsurvival\_10weeks or surv\_10wks Be kind to downstream data

**Using a dropdown selection to keep items specific**

|  |  |
| --- | --- |
| Data Validation Rule, filter by a list, use a selection  use a separate spreadsheet to clearly designate those selections | Graphical user interface, application, icon  Description automatically generated |

**Use Data Validation.** Show briefly above, select a column. In the menu bar choose data validation. For validation criteria select list. That list should have all the categorical variable names needed. Whole numbers inside a range can also be used, a range for a decimal number or a limited string of text.

**1.1a.2 Rule 02 : Recognize a dataframe’s unique qualities; don’t name everything “d”**

**1.1a.3 Rule 03 : Be consistent, unique, and informative with subject IDs**

Examples: Krill476 | Krill\_476 | krill\_r4\_76 | krill\_476 are all going to be determined to be unique observations unless they are cleaned up. Be kind to your future self.

To make sure that the subject ID speaks is unique and tells a story don’t be afraid to combine variables.

Use the paste function along with some character string **“\_”** to create descriptive subject IDs

|  |  |
| --- | --- |
| A picture containing chart  Description automatically generated  Using the paste function to bring together Krill ID with SensorName for Descriptive subject ID | Logo  Description automatically generated |

**1.1a.4 Rule 04 : Make the data square**

The best layout for data that is imported into rStudio is a big rectangle without holes. Rows should correspond to subjects and columns should correspond to variables.

|  |  |
| --- | --- |
| Example of a Rectangle Salinity Log – where measurements were not made daily | Graphical user interface, application, icon  Description automatically generated |

**1.1a.5 Rule 05 : Communicate “NA”, leave no missing values, no gap**

Use “NA”

Do not use a numeric value like 999

Don’t leave a note inside the data, be uniform and use “NA”

Notes belong in a new notes column where observations that don’t need notes would get “NA”

[Folder Name Conventions Needed ]

**1.1a.6 Rule 06 : Use Consistent File Names, dates help**

Having some system for naming all files in a project. Dates are handy at ensuring you have the most recent file to hand. Don’t get pulled into the Master\_file, new\_Master\_file, new\_new\_Master\_file black hole.

|  |  |
| --- | --- |
| Text  Description automatically generated  New files are created with the date first  Naming it by will also always have the | A picture containing text  Description automatically generated |
| Graphical user interface, application  Description automatically generated  There is almost a pattern here with animal, round number of RESP trials, and then date | Logo, company name  Description automatically generated |

**1.1a.7 Rule 07 : Colorings and highlighting is no substitute for assigning a value**

Colors on a spreadsheet aren’t data. Highlighting doesn’t live outside its present program/application. Why use a color when you can assign a value?

**1.1b Consistently Rules – Date Formatting**

**1.1b.1 Rule 08 : YYYY-MM-DD or bust**

Excel will not always be your friend. It displays dates but stores them internally as integers counting up along a number line from 1900-01-01. Don’t forget about time zone when you bring in data.

|  |  |
| --- | --- |
| Figure detailing how you may have to hunt for the source format of dateTimes    Previous dateString holds data in the PAX Roma 0019 whereas the ObservationDate values shows the date in 2019 | Logo  Description automatically generated |

1. **Definitions of Data**
   1. **Raw Data-** just the data and nothing else: no calculations, no graphs.

**2.1a.1 Rule 09 : Ensure raw data lives in its own folder. Create a copy, label processed data and continue work**

**[Meta Data File]**

It’s too easy to open the designated raw data file and inadvertently entering in new data and selecting a field. Designate raw data and figure out a way to only import it and find a way to leave it in its whole and raw form. Data in its original form before you begin transforming in rStudio.

**2.2 Processed Data** – data they you imported into a project have begun to manipulate/transform in rStudio. Examples would be changing column names, snipping headers, and changing date formats.

|  |  |
| --- | --- |
| Read the CSV files    read.csv function most common | Logo  Description automatically generated |
| Host of other functions that can be used to bring in data  Note that headers, row names, column names are fairly key to the dataset being read in | Logo  Description automatically generated |

**2.3 Data Dictionary** - helpful to have a separate file explaining all variables are. The best datasets are rectangular in form:

* The exact variable name as in the data file.
* A version of the variable name that might be used in data visualizations.
* A longer explanation of what the variable means.
* The measurement units
* Expected minimum and maximum values.

|  |
| --- |
| Under Construction DRAFT DRAFT DRAFT under construction |

2.4 Types of Code and R Documents

Types of R documents

Scripts

Rule [No.?]

Rmarkdown

Rule [No.?]

Rnotebook

Rule [No.?] Primary use of detailing investigations/studies

Rule [No.?] Name all chunks

Types of code

Applications

Interfaces

Pipelines (aspirational)

Model Simulations (aspirational)

1. **Data Organization** 
   1. Format. Data should be organized in both their respective GitHUB repositories per project and the Google Drive. GitHUB has a 5MB limit on files which requires a Google drive back up.
      1. **Rule 10 : Find the Experiment Folder, The subject invertebrate, and organize folders by experiment year**

3.1.1a Create four folders – create these folder below in every project to make clear what is raw data, where to find code, and where to find the presentational objects. Following folders can be specific to one aspect of the project e.g. protocols, meeting minutes, etc. Respirometry or survival may want to be called out in another folder but those four main folders need to be the neatest aggregate

|  |  |  |
| --- | --- | --- |
| Folder Name and Ordinal Number | Notes | Logo, company name  Description automatically generated |
| 1.0 Raw Data |  |
| 2.0 Processed Data |  |
| 3.0 Scripts, rNotebooks, rMarkdowns |  |
| 1. Outputs (tables and plots) |  |

3.2 Start with GITHUB repository.

* 1. Storage

Commits to the Repository (once daily- time stamp and call out lines adjusted)

Storing on the OA drive

Things need to live in 1 of 3 (maybe 5 folders)

Google drive – map document

1. Experiments
2. OA Systems
3. OA Outreach (all media files- even pictures of us doing stuff)

Being hard pressed to store anything outside these folders

Then nest folders by project and year

1. \_
2. \_
3. Purchases….market research

3.2.1 Repositories (Storage)

3.2.2 Google Drive (Storage)

3.2.2a File locations

3.2.2b Renaming and Moving files

3.2.3 Schedule of Back Ups

4 OAP & NWFSC Regulation

|  |  |
| --- | --- |
| Titles |  |
| Investigators |  |
| Abstract |  |
| Citation |  |
| **IDENTIFICATION INFORMATION FOR THIS DATA PACKAGE** | |
| NCEI accession number |  |
| NCEI DOI: | DOI Web address |
| Types of Study | Example: Laboratory Experiment |
| Temporal Coverage- Start Date | YYYY-MM-DD |
| Temporal Coverage- End Date | YYYY-MM-DD |
| Spatial Coverage- Latitude |  |
| Spatial Coverage- Longitude |  |
| Geographic Names | Examples: Possession Sound, Puget Sound, Washington; West Coast of the U.S |
| Platforms | None if no vessel or platform was used |
| Research Project(s) |  |
| **VARIABLES / PARAMETERS**  Dissolved Inorganic Carbon Examples in Gray | |
| Abbreviation | DIC |
| Unit: | umol/kh |
| Observation Type: | Laboratory Experiment |
| In-Situ / Manipulations / Response Variable: | Manipulation Condition |
| Measured or Calculated | Measured |
| Sampling instrument: | Acid-washed borosilicate glass bottles |
| CRM manufacturer: | Dickson Laboratory, Scripps Institution of Oceanography |
| Poison name: | HgCl2 |
| Poison volume: | 20 uL HgCl2 is added to 500 mL samples. |
| Poison correction: | No correction |
| **VARIABLES / PARAMETERS**  Total alkalinity Examples in Gray | |
|  |  |
|  |  |
|  |  |
|  |  |

**Goals and Rules**

**Goal 01:** Format it once

**Goal 02:** Produce robust, re-usable data. Reusable Data has additional information for complete user understanding about collection and any manipulation/ filtering/ formatting

**Goal 03:** Provide lab members a clear idea about preparing, disseminating, and archiving data

**Rule 01:** Be consistent especially with categorical variable names

**Rule 02 :** Recognize a dataframe’s unique qualities; don’t name everything “d”

**Rule 03 :** Be consistent, unique, and informative with subject IDs

**Rule 04 : Make the data square**

**Rule 05 : Communicate “NA”, leave no missing values, no gap**

**Rule 06 : Use Consistent File Names, dates help**

**Rule 07 : Colorings and highlighting is no substitute for assigning a value**

**Rule 08 : YYYY-MM-DD or bust**

**Rule 09 :** Ensure raw data lives in its own folder. Create a copy, label processed data and continue work