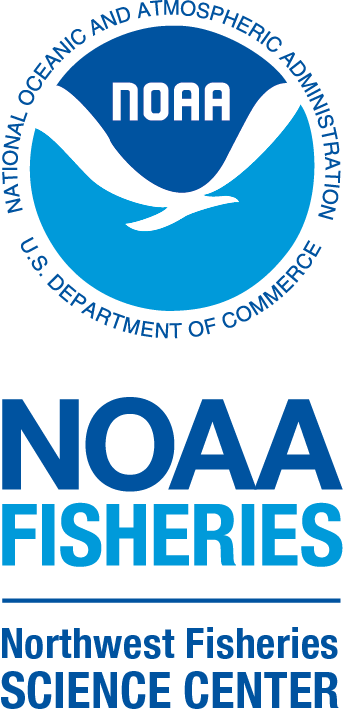
|  |  |
| --- | --- |
| FROM | Danielle Perez, Kate Rovinski |
| TO | Paul McElhany, OA Lab |
| DATE | (Wednesday) 2021.01.13 |
| SUBJ | *Draft*  Data Management Policies  Understanding types of data, handling, and storage  Protocols for OA Lab Personnel |



**Table of Contents**

1 Overview

Consistently Rules (Including Naming Protocols)

2 Users- Who should read and implement this?

3 Definitions of Data

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

meta requirements/format

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

5 OAP & NWFSC Regulation

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: The first rule of data organization is 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 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!!!** |

**1.1b Consistently Rules – Date Formating**

# 2 Users - Should this Apply to Me, What Projects does this Apply

# 3 Types of data Definition

## 3.1 Raw data

## 3.2 Derived data

## 3.3 Other types of data & things that are borderline data

## 3.4 Received data

## 3.5 Code

# 4 Data organization and formats 3

## 4.1 File formats

## 4.2 Data formatting

## 4.3 Metadata

## 4.4 Cleaning data

# 5 Where to store data 5

## 5.1 Data storage in the lab

## 5.2 Preferred repository for data archiving and publication

Creating Folders - Format of First Four Folders

1. Raw Data

1a. Processed Data

2. Scripts and R documents

Put a time stamp with the commit (giuthub gives time from now rather than exact dateTime delivered onto the repository)

A commit needs and update- be clear about the sections and preferably the lines adjusted

example T/S, Troubleshooting plotting error on line 933

3. Outputs (objects & plots)

4. Presentations

## 5.3 READme files- Governing Large Workflows. Solving the difficulty to reproduce executions

# x Redistribution policies & timelines 6

## x.1 Length of time after collection to be made available to rest of lab

## x.2 Use of lab data

MOATS : it is an acronym, always all CAPs including the S for System

PSU : prefered salinity/conductivty units

Celsius : prefered temperature units

unless acroynm all lowercase

#### Use of lab data - Consistency

```{r 1.0) Use of lab Data - Consistency}

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# 1.0) Use of lab Data - Consistency

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# examples of how things can break\

```

\*Consistency Rules\*

\*Rule 01\*: The first rule of data organization is be consistent.

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 refering to something and stick to it.

Keep it short but keep it obvious

four digits to help identify variables- especially categorical values

to describe MOATs 01 the character "1" or "01" could work

However, "M01" or "MOATS01" can't be mistaken for scalar values

\*Rule 02\*: Be consistent about null or missing values; Use "NA"

Use a single fixed code for any missing values.

Every cell needs filled in (more discussion here), so that one can distinguish between truly missing values and unintentionally missing values. R users prefer “NA”. You could also use a hyphen. But stick with a single value throughout your data. Definitely don’t use a numeric value like -999 or 999; it’s easy to miss that it’s intended to be missing. Also, don’t insert a note in place of the data, explaining why it’s missing. Rather, make a separate column with such notes.

Use consistent variable names. If in one file (say the first batch of subjects), you have a variable called “Glucose\_10wk”, then call it exactly that in other files (say for other batches of subjects). If it’s variably called “Glucose\_10wk”, “gluc\_10weeks”, and “10 week glucose”, then downstream the data analyst will have to work out that these are all really the same thing. (More on naming variables here.)

Use consistent subject IDs. If sometimes it’s “153” and sometimes “mouse153” and sometimes “mouse-153F” and sometimes “Mouse153”, there’s going to be extra work to figure out who’s who.

Use a common data layout in multiple files. If your data are in multiple files, use the same layout in all files. (More on layout here.)

Use consistent file names. Have some system for naming files. If one file is called “Serum\_batch1\_2015-01-30.csv”, then don’t call the file for the next batch “batch2\_serum\_52915.csv” but rather use “Serum\_batch2\_2015-05-29.csv”. (More on naming files here.)

Use a single common format for all dates, preferably YYYY-MM-DD, like 2015-08-01. If sometimes you write 8/1/2015 and sometimes 8-1-15, you’re asking for trouble. (More on dates next.)

Use consistent phrases in your notes. If you have a separate column of notes (for example, “dead” or “lo off curve”), be consistent in what you write. Don’t sometimes write “dead” and sometimes “Dead”, or sometimes “lo off curve” and sometimes “off curve lo”.

Be careful about extra spaces within cells. A blank cell is different than a cell that contains a single space. And “male” is different from “ male ” (that is, with spaces at the beginning and end). These can be a headache later on.

mg/L : particular units for dissolved oxygen measurements

what is our pH measurement

## x.3 Length of time after collection for data to be made public

## x.4 Data licenses

# x UBC Policies

# 9 NOAA/NWFSC Policies... OAP policies