Hands-on Activity 4: Data Entry & Manipulation

Associated DataONE Lecture: Lesson 4: *Data Entry and Manipulation*

Objectives: Students recognize the importance of organizing data so that others can understand them, by experiencing the challenges of reviewing data sheets that are organized in a common but inappropriate way.

Outcomes: (1) Students can explain why standardizing descriptions and organization of data is important. (2) Students can strategize about the best ways to organize data.

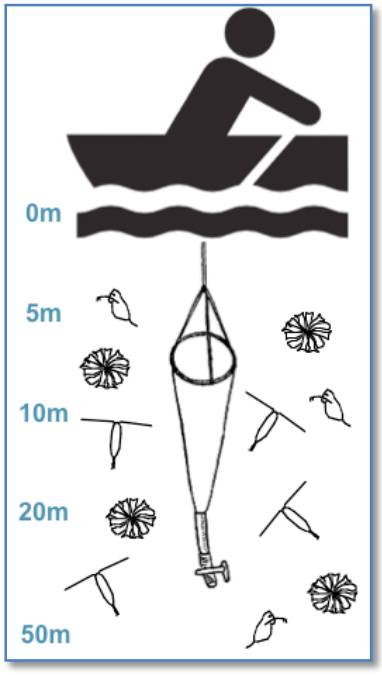
Time Needed: One hour in class.

URLs: None.

Additional Files Needed: pond2010.xlsx, zoop- temp-main.xlsx; zoop-temp.xlsx

Key Reading: Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M., 2009. Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America 90, 205–214.

Notes and Instructions for Instructors:



Data Organization

Background: Plankton are microscopic organisms that form the base of many aquatic food webs – fueling the growth of fish and other larger organisms. It's common to sample them using a net or another container that can be controlled to collect water just from certain depths; so you can see how plankton collected at the surface (0 meters) might be different from plankton at another depth (e.g. 10 meters below the surface).

(For more information: http://en.wikipedia.org/wiki/Phytoplankton and http://en.wikipedia.org/wiki/Zooplankton.)

They are identified and counted under a microscope, and usually their numbers are reported as individuals per liter or milliliter.

Frequently, aquatic scientists collect plankton samples during both day (e.g. noon) and night (e.g. 2 am) because plankton change their distributions from day to night, and not all species alter their distributions in the same way. (For more information, search "diel vertical migration" on the web.)

You should have 3 (fictional) data files: pond2010.xlsx, zoop-temp-main.xlsx; zoop-temp.xlsx.

These 3 files were all intended to be part of the same study – the investigators wanted to examine the day-night distribution of 2 species of zooplankton across multiple years. The type of zooplankton they studied is called rotifers generally, and specifically the genus *Conochilus,* in which groups of individual rotifers stick together in colonies (see http://eol.org/pages/43393/overview). The investigators plan to repeat this study for several more years.

Activity 1

As individuals or in small groups, open the 3 files and inspect them. Based on what you have learned so far about data management, what are some problems in the way the data are currently organized?

Some example answers:

- The 3 separate files are organized inconsistently - pond2010.xlsx appears to have data from only one station in 2010, while zoop-temp-main.xlsx and zoop-temp.xlsx represent Station A and Station B separately, both in 2011.
- The 3 files use different names for the columns (e.g. "Density" vs "#/L"), and the columns differ in their ordering.
- The metadata are represented differently across the sheets – in one there is a header describing the Station details, while in another there is a second spreadsheet with metadata.
- There are graphs embedded in some of the spreadsheets.
- There are multiple tables in some of the spreadsheets, e.g. random calculations in zoop-temp.xlsx.
- There are no units associated with some of the headers in the spreadsheets (temperature, colony diameter, depth).
- Color is being used as metadata, and it is uncertain how it is used – it could mean missing values that were interpolated, or values for which equipment seemed faulty, or that these values were particularly interesting!
- The file names are not descriptive or consistent.

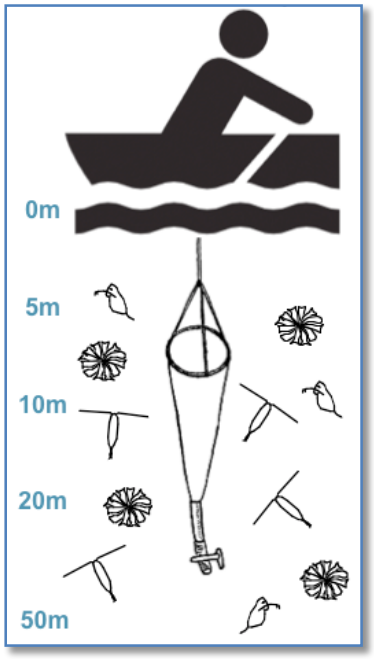
Activity 2

Suggest a new system for organization. Create a new spreadsheet that can be used as a template for later years of data collection.

Some example answers:

- There are many correct answers here. Some people may suggest separate data sheets for each year (a program like R or SAS could then just call each of the sheets to aggregate them for analyses), others may suggest moving them all into the same data sheet and adding new data each year. In the latter case, recognize that any "copy and paste" that is done to aggregate the old files may introduce new errors, so some data checking will be necessary to ensure data quality in this migration.
- Some may suggest that the metadata about the lake itself could be recorded in a separate table, to create a relational database, e.g. with depth and other characteristics of the lake, rather than just having that information embedded in a metadata file.
- At the very least, we'd expect to see a data sheet with consistent naming conventions for the variables and consistent flagging conventions for missing data or interpolations, with a full metadata record kept separately from the data table.

Student Instructions:



Data Organization

Background: Plankton are microscopic organisms that form the base of many aquatic food webs – fueling the growth of fish and other larger organisms. It's common to sample them using a net or another container that can be controlled to collect water just from certain depths; so you can see how plankton collected at the surface (0 meters) might be different from plankton at another depth (e.g. 10 meters below the surface).

(For more information: http://en.wikipedia.org/wiki/Phytoplankton and http://en.wikipedia.org/wiki/Zooplankton.)

They are identified and counted under a microscope, and usually their numbers are reported as individuals per liter or milliliter.

Frequently, aquatic scientists collect plankton samples during both day (e.g. noon) and night (e.g. 2 am) because plankton change their distributions from day to night, and not all species alter their distributions in the same way. (For more information, search "diel vertical migration" on the web.)

You should have 3 (fictional) data files: pond2010.xlsx, zoop-temp-main.xlsx; zoop-temp.xlsx.

These 3 files were all intended to be part of the same study – the investigators wanted to examine the day-night distribution of 2 species of zooplankton across multiple years. The type of zooplankton they studied is called rotifers generally, and specifically the genus *Conochilus,* in which groups of individual rotifers stick together in colonies (see http://eol.org/pages/43393/overview). The investigators plan to repeat this study for several more years.

Activity 1

As individuals or in small groups, open the 3 files and inspect them. Based on what you have learned so far about data management, what are some problems in the way the data are currently organized?

Activity 2

Suggest a new system for organization. Create a new spreadsheet that can be used as a template for later years of data collection.