Hands-on Activity 7: Metadata

Associated DataONE Lecture: Lesson 7: *Metadata*

Objectives: Students consider the level of detail that is necessary for metadata to adequately describe data sets, and work with a metadata record.

Outcomes: (1) Students can explain why detailed metadata are valuable. (2) Students can provide suggestions for improving metadata descriptions.

Time Needed: 45 minutes in class.

URLs: Morpho (https://knb.ecoinformatics.org/#tools/morpho), DataUp (http://dataup.cdlib.org/)

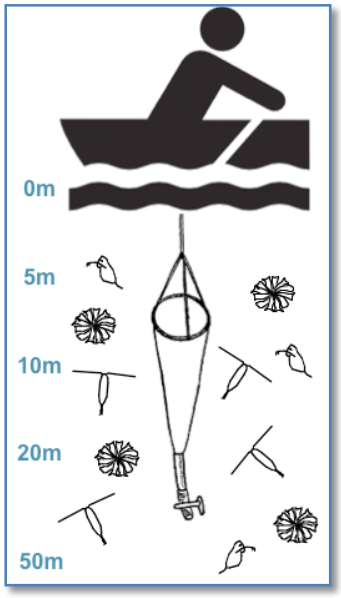
Additional Files Needed: xlsx, zoop- temp-main.xlsx; zoop-temp.xlsx

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.

White, E. P., E. Baldridge, Z. T. Brym, K. J. Locey, D. J. McGlinn, and S. R. Supp. 2013. Nine simple ways to make it easier to (re) use your data. PeerJ PrePrints.

Notes and Instructions for Instructors:



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.

The files have some problems in how they are organized, which you have already discussed in a previous exercise. Now let's think about writing some good metadata that describes the data set. Note that Activities 1-4 refer to the gray areas in the metadata record, which is found later in this document.

Activity 1

As individuals or in small groups, look through the files and locate all the information that describes these data – the metadata. Some of this information is found in this handout, and some of it is within the 3 data sheets provided. Describe where you found the information that is needed to populate the metadata record.

Example answer:

> Look at the column headers in all the sheets, a brief table on zoop + temp.xlsx, and a second worksheet on zoop + temp-main.xlsx.
>
> Some trainees may also suggest that information is online or elsewhere – e.g. the geographical coordinates may be used to locate lake names, and information about the organisms may be published.

Activity 2

Now let's focus on a metadata description just for pond2010.xlsx. Look at the table contained in the file. Write an appropriate title for this data set.

Example answer:

> There are many good answers here but we are looking for very descriptive titles, and consider that keywords can be used to complement the titles so that they don't get too long!
>
> Here's one suggestion:
> Summer population density and colony size of Conochilus hippocrepis and Conochilus unicornis at multiple pond depths in Littlevick Pond Natural Reserve, Surrey, UK in 2010

Activity 3

"Time Period of Content" represents the time period the data was collected. What dates would you enter?

Example answer:

Look to the columns for dates.

5 June 2010 – 18 June 2010 is the time period covered by pond2010.xlsx. In the metadata record the dates would be represented as YYYYMMDD: 20100605 and 20100618.

Activity 4

What would be some appropriate theme keywords for this dataset? Where can you find help for developing keywords?

Example answer:

Again, there are many good answers here. You may find that some of the same terms appear in both the title and keywords section.

Words might be taxonomic like: rotifers, zooplankton, plankton. They may describe the process that the researchers are studying such as: diel vertical migration.

Taxonomic references may include Cowardin Wetland Classification System and other discipline specific taxonomies. Place Keyword thesauri could include Geographic Names Index Service (GNIS). Discuss relevant taxonomies with participants.

Activity 5

Take a look at the metadata record in this exercise. Note that there are a variety of domain types, and some are noted as "unrepresentable." What that might mean?

Example answer:

Attributes such as temperature, diameter, and density are listed as "unrepresentable" instead of listing a range of values (ie, 10-30 cm) because there is no absolute max and min value for the attribute noted anywhere.

A "percent" attribute is a good example of a range domain because the values must be greater than or equal to 0 and less than or equal to 100.

Pond2010 Metadata

This is some (fictional) information about the (fictional) data set called pond2010.xlsx. The data set can be used to fill in metadata fields in a formal record, such as the one below, but note that there may also be additional important metadata within the pond2010 file and its related files, zoop-temp-main.xlsx and zoop-temp.xlsx.

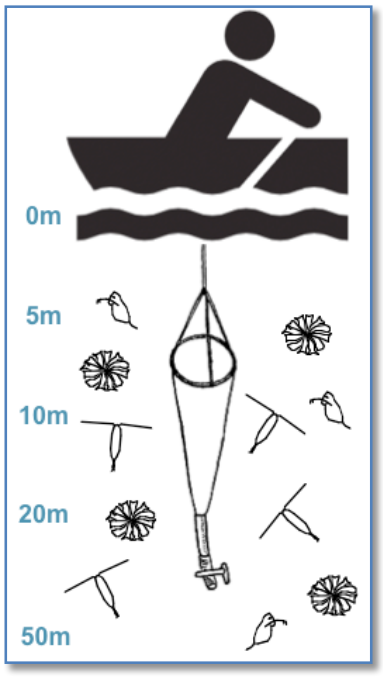
Title of the Data set	
Originator/Dataset Author	Anna Sassin Dan D. Lyons
Abstract	This dataset is one of a collection of four population survey datasets documenting colony growth, reproduction, and survival of two rotifer species (*Conochilus unicornis* and *Conochilus hippocrepis*) at four time periods of the year. This dataset describes population data for the summer season. Samples of both species were taken at Littlevick pond, Surrey, UK. Measurements taken include depth, temperature, colony density and colony diameter.
Purpose	Data were collected to evaluate how temperature and depth affect the survival of rotifer colonies in ponds within the UK.
Publication	*Publisher:* International Rotifer Recovery Science Center *Place:* Surrey, UK *Publication_Date:* 12/08/2012 *Series Name:* Four Season Rotifer Survey *Name of Issue*: Summer Survey
Larger_Work_Citation	*Originator:* Sassin, Anna and Lyons, Dan .D. *Publication_Date:* 12/08/2012 *Title:* Relationships between population and temperature: Tracking rotifers over the course of four seasons in the United Kingdom. *Publisher:* Rotifer Conservation *Place:* UK *Volume;Issue;Pages:* 4(2): 325-340
Time Period of Content	Begin Date: End Date:
CurrentnessReference	Ground Condition
Progress/status:	Complete
Maintenance_and_Update_Frequency	None planned

Geographic coverage	Littlevick Pond Natural Reserve, Surrey, UK.
Bounding_Coordinates:	*West_Bounding_Coordinate:* -0.92456818028327 *East_Bounding_Coordinate:* 0.371818538415 *North_Bounding_Coordinate:* 51.511581803063 *South_Bounding_Coordinate:* 50.808817656094
Keywords (theme)	
Keywords (place)	Surrey UK International Littlevick Pond Natural Reserve
Keywords (temporal)	summer, June
Data Access_Constraints	No legal or policy restriction for accessing this dataset.
Data Use_Constraints:	Must properly cite originator if used in publications, reports, presentations, etc. Please cite data set according to DataCite.org standards
Contact_Person_Primary:	*Contact_Person:* Tad Pohl (Data steward) *Contact_Organization:* International Rotifer Recovery Science Center *Address:* 5638 Independence Way *City:* Guildford *State_or_Province:* Surrey, UK *Contact_Telephone:* +44 (0) 888-8888
Data_Set_Credit	Funding was provided by International Rotifer Foundation
Analytical_Tools	SAS, R, MatLab
Data_Quality_Information	
Attribute_Accuracy_Report	Temperature instrument was tested and calibrated for accuracy before each sampling. Density and colony counts were conducted according to the Standard Plate Count procedure. Counts were conducted by two data counters. Each technicians count was verified by the second technician. Counting accuracy was found to be 95% accurate.
Completeness_Report	The data set is generally complete although the temperature for one sample depth could not be recorded due to instrument malfunction. Colony and density counts are also mostly complete except for two instances where the data is missing and is therefore unknown. Statistical summary (boxplot) of the data was performed and no outstanding outliers or potentially erroneous values were found.
Positional_Accuracy:	Positional Accuracy was not assessed
Process_Step:	Data was collected by 2 people the first week and by the same 2

Process_Description:	people the following week. Water samples and temperature were taken at five different depths. In order to account for variability in sample measurements, 6 water samples were taken at each depth. These 6 samples were later randomly divided into two even groups of three. The two groups were randomly assigned a rotifer species name whereby data counters would perform the density and colony counts for the particular species.
Entity and Attribute Information	
Detailed_Description *Entity_Type*	*Entity_Type_Label:* pond2010.xlsx *Entity_Type_Definition:* Rotifer population survey at various depths and temperature
Attribute	*Attribute_Label:* z *Attribute_Definition:* Depth in centimeters from the surface *Attribute_Domain_Values:* *Enumerated_Domain:* *Enumerated_Domain_Value:* 0.5 *Enumerated_Domain_Value_Definition:* 0.5 cm below surface *Enumerated_Domain_Value:* 5 *Enumerated_Domain_Value_Definition:* 5 cm below surface *Enumerated_Domain_Value:* 10 *Enumerated_Domain_Value_Definition:* 10 cm below surface *Enumerated_Domain_Value:* 25 *Enumerated_Domain_Value_Definition:* 25 cm below surface *Enumerated_Domain_Value:* 50 *Enumerated_Domain_Value_Definition:* 50 cm below surface
Attribute	*Attribute_Label:* Temperature *Attribute_Definition:* Temperature of water in Celsius *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Density *Attribute_Definition:* Number of individuals per colony *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Colony Diameter *Attribute_Definition:* Length of longest colony diameter in millimeters *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Species *Attribute_Definition:* Rotifer species *Attribute_Domain_Values:* *Enumerated_Domain_Value:* cuni

	Enumerated_Domain_Value_Definition: Conochilus unicornis *Enumerated_Domain_Value:* chippo *Enumerated_Domain_Value_Definition:* Conochilus hippocrepis
Distribution Information	
Distributor *Contact_Information* *Contact_Organization_Primary*	*Contact_Organization:* Rotifer Network for Biocomplexity (RNB) *Contact_Person:* Metadata Coordinator *Address:* 6534 Biodata Way *City:* Novel Jersey *State_or_Province:* New Jersey *Postal_Code:* 97564 *Contact_Voice_Telephone:* 555-555-1034 *Contact_Email:* info@rnb.net
Distribution_Liability	The Rotifer Network for Biocomplexity (RNB) shall not be held liable for improper or incorrect use of the data described and/or contained herein. It is the responsibility of the data user to use the data appropriately and consistent within the limitations of the data.

Student Instructions:



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. 2am) 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.

The files have some problems in how they are organized, which you have already discussed in a previous exercise. Now let's think about writing some good metadata that describes the data set. Note that Activities 1-4 refer to the gray areas in the metadata record, which is found later on in this document.

Activity 1

As individuals or in small groups, look through the files and locate all the information that describes these data – the metadata. Some of this information is found in this handout, and some of it is within the 3 data sheets provided. Describe where you found the information that is needed to populate the metadata record.

Activity 2

Now let's focus on a metadata description just for pond2010.xlsx. Look at the table contained in the file. Write an appropriate title for this data set.

Activity 3

"Time Period of Content" represents the time period the data was collected. What dates would you enter?

Activity 4

What would be some appropriate theme keywords for this dataset? Where can you find help for developing keywords?

Activity 5

Take a look at the metadata record in this exercise. Note that there are a variety of domain types, and some are noted as "unrepresentable." What that might mean?

Pond2010 Metadata

This is some (fictional) information about the (fictional) data set called pond2010.xlsx. The data set can be used to fill in metadata fields in a formal record, such as the one below, but note that there may also be additional important metadata within the pond2010 file and its related files, zoop-temp-main.xlsx and zoop-temp.xlsx.

Title of the Data set	
Originator/Dataset Author	Anna Sassin Dan D. Lyons
Abstract	This dataset is one of a collection of four population survey datasets documenting colony growth, reproduction, and survival of two rotifer species (*Conochilus unicornis* and *Conochilus hippocrepis*) at four time periods of the year. This dataset describes population data for the summer season. Samples of both species were taken at Littlevick pond, Surrey, UK. Measurements taken include depth, temperature, colony density and colony diameter.
Purpose	Data were collected to evaluate how temperature and depth affect the survival of rotifer colonies in ponds within the UK.
Publication	*Publisher:* International Rotifer Recovery Science Center *Place:* Surrey, UK *Publication_Date:* 12/08/2012 *Series Name:* Four Season Rotifer Survey *Name of Issue*: Summer Survey
Larger_Work_Citation	*Originator:* Sassin, Anna and Lyons, Dan .D. *Publication_Date:* 12/08/2012 *Title:* Relationships between population and temperature: Tracking rotifers over the course of four seasons in the United Kingdom. *Publisher:* Rotifer Conservation *Place:* UK *Volume;Issue;Pages:* 4(2): 325-340
Time Period of Content	Begin Date: End Date:
CurrentnessReference	Ground Condition
Progress/status:	Complete
Maintenance_and_Update_Frequency	None planned

Geographic coverage	Littlevick Pond Natural Reserve, Surrey, UK.
Bounding_Coordinates:	*West_Bounding_Coordinate:* -0.92456818028327 *East_Bounding_Coordinate:* 0.371818538415 *North_Bounding_Coordinate:* 51.511581803063 *South_Bounding_Coordinate:* 50.808817656094
Keywords (theme)	
Keywords (place)	Surrey UK International Littlevick Pond Natural Reserve
Keywords (temporal)	summer, June
Data Access_Constraints	No legal or policy restriction for accessing this dataset.
Data Use_Constraints:	Must properly cite originator if used in publications, reports, presentations, etc. Please cite data set according to DataCite.org standards
Contact_Person_Primary:	*Contact_Person:* Tad Pohl (Data steward) *Contact_Organization:* International Rotifer Recovery Science Center *Address:* 5638 Independence Way *City:* Guildford *State_or_Province:* Surrey, UK *Contact_Telephone:* +44 (0) 888-8888
Data_Set_Credit	Funding was provided by International Rotifer Foundation
Analytical_Tools	SAS, R, MatLab
Data_Quality_Information	
Attribute_Accuracy_Report	Temperature instrument was tested and calibrated for accuracy before each sampling. Density and colony counts were conducted according to the Standard Plate Count procedure. Counts were conducted by two data counters. Each technicians count was verified by the second technician. Counting accuracy was found to be 95% accurate.
Completeness_Report	The data set is generally complete although the temperature for one sample depth could not be recorded due to instrument malfunction. Colony and density counts are also mostly complete except for two instances where the data is missing and is therefore unknown. Statistical summary (boxplot) of the data was performed and no outstanding outliers or potentially erroneous values were found.
Positional_Accuracy:	Positional Accuracy was not assessed
Process_Step:	Data was collected by 2 people the first week and by the same 2

Process_Description:	people the following week. Water samples and temperature were taken at five different depths. In order to account for variability in sample measurements, 6 water samples were taken at each depth. These 6 samples were later randomly divided into two even groups of three. The two groups were randomly assigned a rotifer species name whereby data counters would perform the density and colony counts for the particular species.
Entity and Attribute Information	
Detailed_Description *Entity_Type*	*Entity_Type_Label:* pond2010.xlsx *Entity_Type_Definition:* Rotifer population survey at various depths and temperature
Attribute	*Attribute_Label:* z *Attribute_Definition:* Depth in centimeters from the surface *Attribute_Domain_Values:* *Enumerated_Domain:* *Enumerated_Domain_Value:* 0.5 *Enumerated_Domain_Value_Definition:* 0.5 cm below surface *Enumerated_Domain_Value:* 5 *Enumerated_Domain_Value_Definition:* 5 cm below surface *Enumerated_Domain_Value:* 10 *Enumerated_Domain_Value_Definition:* 10 cm below surface *Enumerated_Domain_Value:* 25 *Enumerated_Domain_Value_Definition:* 25 cm below surface *Enumerated_Domain_Value:* 50 *Enumerated_Domain_Value_Definition:* 50 cm below surface
Attribute	*Attribute_Label:* Temperature *Attribute_Definition:* Temperature of water in Celsius *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Density *Attribute_Definition:* Number of individuals per colony *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Colony Diameter *Attribute_Definition:* Length of longest colony diameter in millimeters *Attribute_Domain_Values:* *Unrepresentable_Domain*
Attribute	*Attribute_Label:* Species *Attribute_Definition:* Rotifer species *Attribute_Domain_Values:* *Enumerated_Domain_Value:* cuni

	Enumerated_Domain_Value_Definition: Conochilus unicornis *Enumerated_Domain_Value:* chippo *Enumerated_Domain_Value_Definition:* Conochilus hippocrepis
Distribution Information	
Distributor *Contact_Information* *Contact_Organization_Primary*	*Contact_Organization:* Rotifer Network for Biocomplexity (RNB) *Contact_Person:* Metadata Coordinator *Address:* 6534 Biodata Way *City:* Novel Jersey *State_or_Province:* New Jersey *Postal_Code:* 97564 *Contact_Voice_Telephone:* 555-555-1034 *Contact_Email:* info@rnb.net
Distribution_Liability	The Rotifer Network for Biocomplexity (RNB) shall not be held liable for improper or incorrect use of the data described and/or contained herein. It is the responsibility of the data user to use the data appropriately and consistent within the limitations of the data.