

# Metadata and data documentation

Oregon Data Science Collaborative

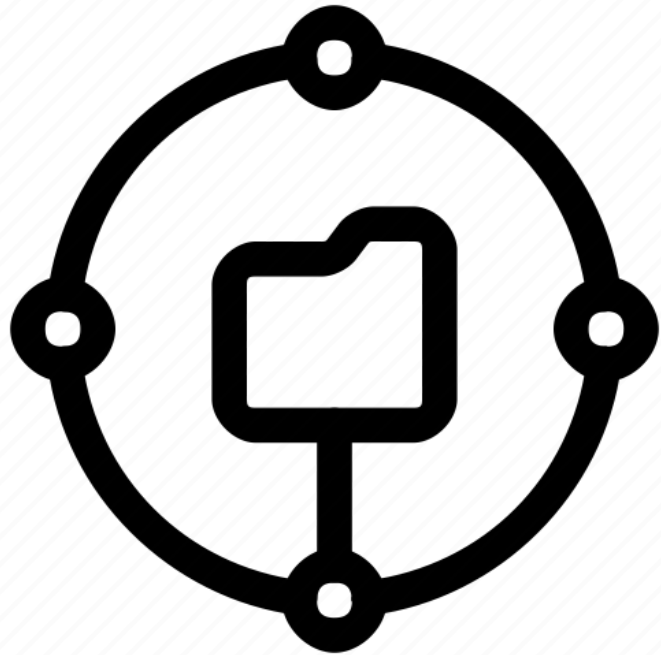
Spring 2022



# Exercises

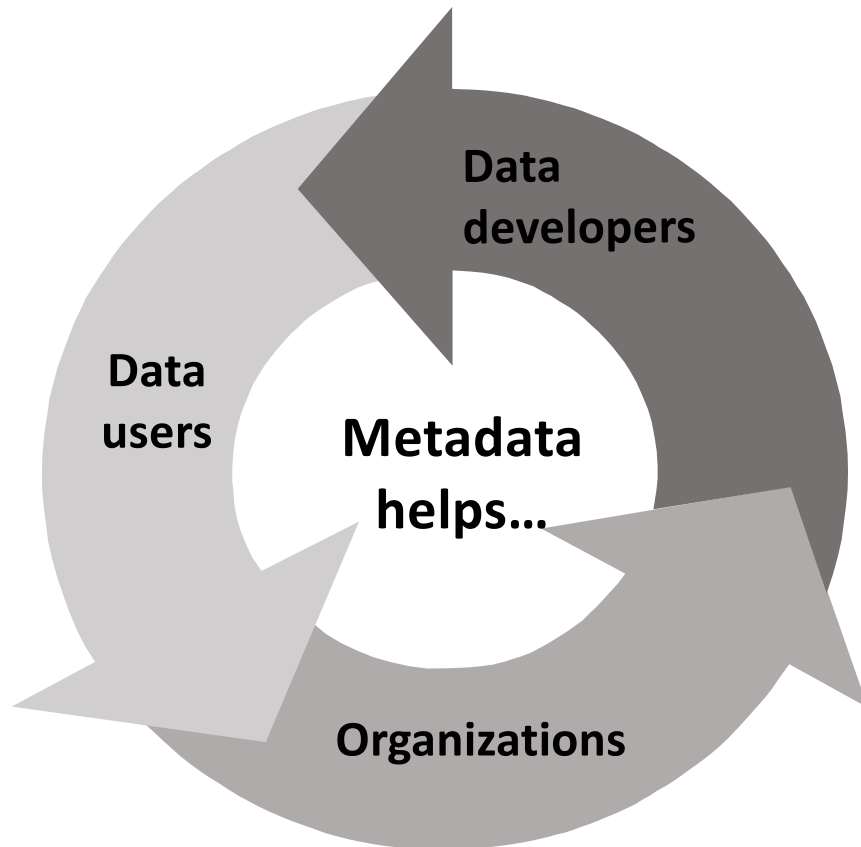
- For today's exercises, you will create metadata for a research project
- You will need information about the project and data related to the project
- If you have your own project to use, make sure you have tabular data available
- If you do not have your own project, visit this link to access the demo data: [\*Compiled annual statewide Alaskan salmon escapement counts, 1921-2017.\*](#)
- You will need to download the file 'CompiledEsc.csv'
- Keep the repository page open as well

# Metadata



**WHO** created the data?  
**WHAT** is the content of the data?  
**WHEN** were the data created?  
**WHERE** are the data from?  
**HOW** were the data developed?  
**WHY** were the data developed?

# Who uses metadata?



- Consider the audience
- Who might use your data in the future?
  - Your future self!
  - Your group – including future members
  - Current collaborators
  - Researchers in your field – for meta-analysis, synthesis
  - Researchers outside your field – for interdisciplinary applications

# Metadata occurs in many forms

Site Name and Code FJ / 4 <sup>th</sup> of July	Region WA	Date 07/20/21	People OG KT
Tide stage ebb / low / <u>flood</u> / high	Weather (rain, wind) light breeze	Swell height (m) 0-0.5 0.5-1 1-2 2-3	

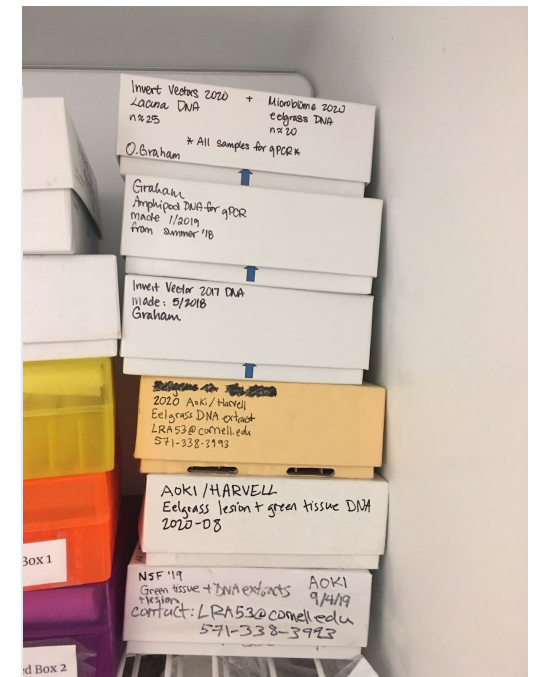
Transect	Replicate	Bottle Number	Collection time	Seagrass present on transect?	Filtration time	Total volume filtered (mL)	Notes
U1	1	1	11:46	Y	3:36	1111 500	filter marked counter
U1	2	2	11:44	N? 10m	3:05	1111 450	
U1	3	3	11:43	N	4:09	1111 500	
U2	1	4	11:41	N	2:45	1111 500	
U2	2	5	11:41	N	3:53	1111 500	

## MATERIALS AND METHODS

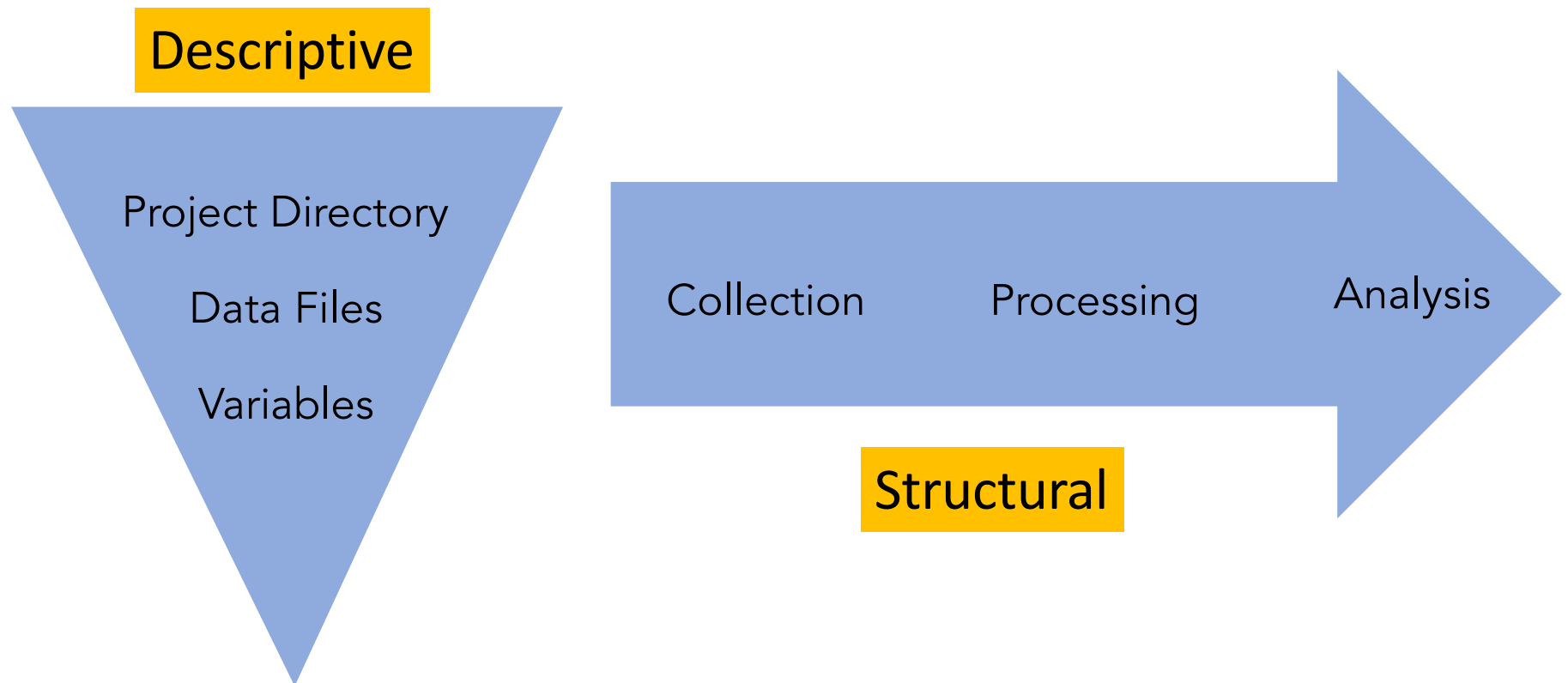
### Field Study 1: Eelgrass Growth

To determine temporal differences and interactions between eelgrass growth, productivity, and SWD, we conducted field trials in June and July 2019 at Fourth of July Beach, San Juan Island, WA (48°28'01.4" N, 123°00'00.4" W). We selected this

OR.A.L4.11.tiff  
OR.A.L5.8.tiff  
OR.A.L5.20.tiff  
OR.A.L6.1.tif  
OR.A.L6.4\_5.tif  
OR.A.L6.15.tif  
OR.A.L6.18.tiff  
OR.A.U1.3.tiff  
OR.A.U2.9.tif  
OR.A.U2.11.tif  
OR.A.U2.17.tif  
OR.A.U2.19.tif  
OR.A.U3.6\_7.tif  
OR.A.U3.15.tif

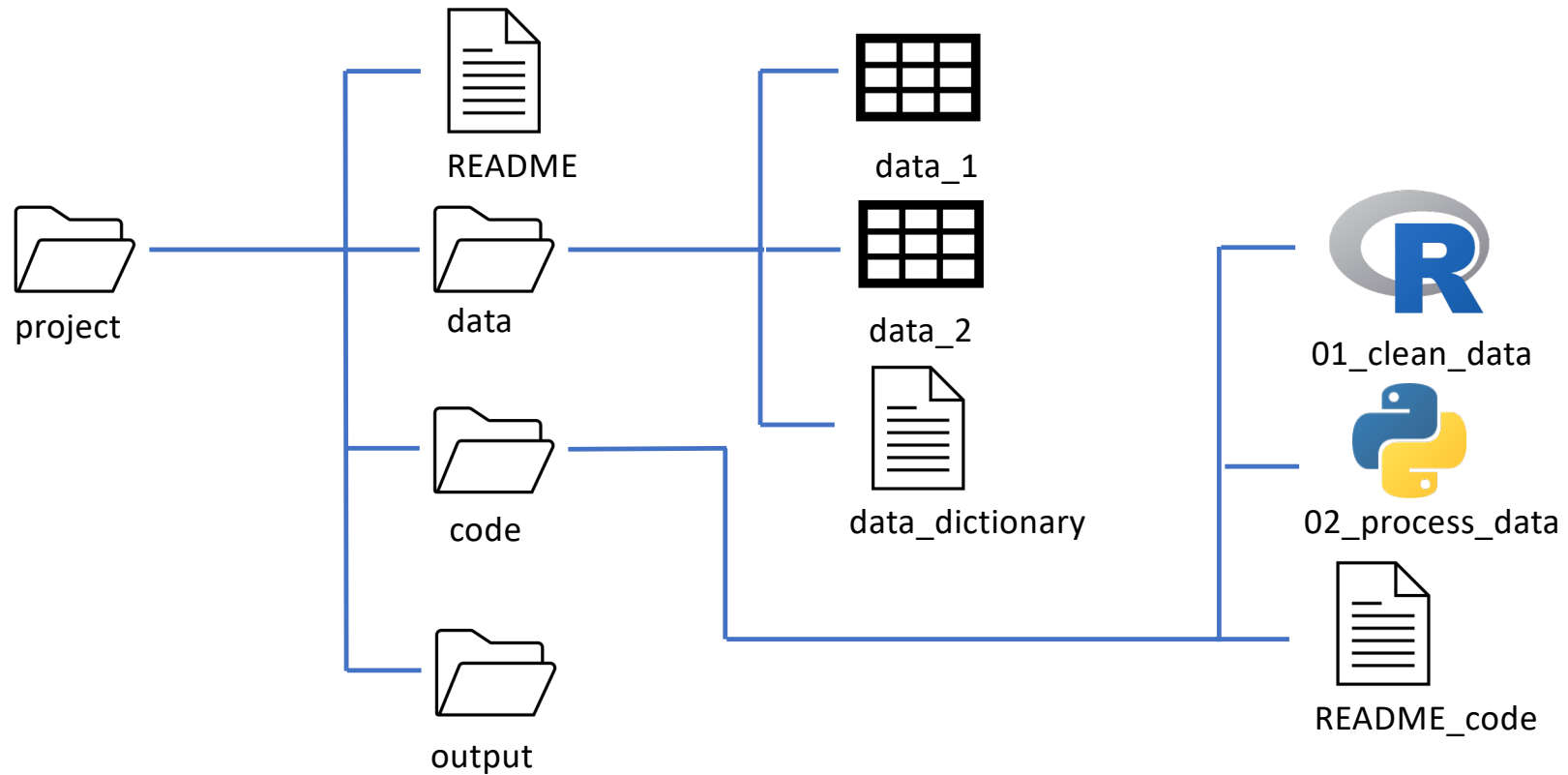


# Metadata occurs at multiple levels



*Adapted from UC Davis DataLab*

# Start with ReadMe and data dictionary



# ReadMe

1. Project name
2. Dates (when project began, last updated)
3. Author names and contact info
4. Origin of data
5. Description of the goal of the project
6. Dependencies, installation
7. Summary of methods
8. ...

- Gives users important context for a project or dataset
- Should be in plain text format, either .txt or .md
- Contains both descriptive and structural metadata

## ReadMe for subdirectory

1. Contents and purpose of subdirectory
2. Source of raw data
3. Exact tools and steps taken to modify data



# Sample README from a published dataset

Paper: A continuous morphological approach to study the evolution of pollen in a phylogenetic context: an example with the order Myrtales

Authors: Ricardo Kriebel, Mohammad Khabbazian, Kenneth J. Sytsma

Description:

This README file describes the data package accompanying the above publication.

Files:

1. Myrtales\_pollen\_evolution.Rmd: This R markdown file reproduces the analyses of the paper cited above.
2. Data folder: Includes three spreadsheets with the pollen measurement data, the names of the data checked for there current accepted taxonomic names, and the latitudinal data.
3. Functions folder: Includes three R functions that are sourced during the markdown analyses.
4. Outlines folder: Includes two subfolders, one with the outline files of the polar view of Myrtales pollen grains, and one with the outlines of the equatorial view of the pollen grains. In addition, within each of the subfolders is a .csv file that is used to assign the groups.
5. Phylogenies: This folder includes the three phylogenies described in the paper.
6. Results: This folder includes 18 rds files with the results of the shift detection analyses and their corresponding bootstrap analyses.

[Open in a browser from this link](#)

# ReadMe exercise one (full group)

- Review sample READMEs
  - [Example 1 – University of Arizona](#)
  - [Example 2 – Hack for California](#)
- What is effective about these READMEs? What would you change?

# ReadMe exercise two (breakout groups)

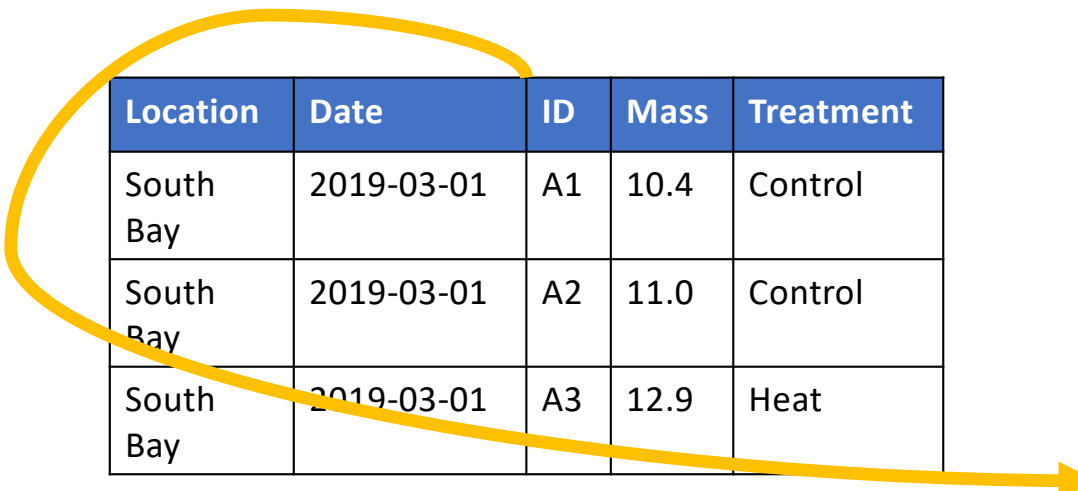
## If you have your own project/data

- Draft a README for your own project – past, present, or future
- Use [this template](#) to get started
- Are there any roadblocks to drafting the ReadMe? What information is harder to gather?

## If you are using the demo project

- Visit the [project repository](#)
- Draft a README document using the available metadata
- Use [this template](#) to get started
- Is any information missing?

# Data dictionary defines the data!



Location	Date	ID	Mass	Treatment
South Bay	2019-03-01	A1	10.4	Control
South Bay	2019-03-01	A2	11.0	Control
South Bay	2019-03-01	A3	12.9	Heat

Other useful information to include:

- Accepted values
- Null values
- Notes – especially to explain unexpected values

Variable name	Data type	Units	Description
Location	text	NA	Location where experiment occurred
Date	Date-time	NA	Date of measurement
ID	text	NA	Sample identifier
Mass	numeric	grams	Mass of sample
Treatment	text	NA	Experimental temperature treatment

# Codebooks

- For categorical data, especially data collected in surveys, metadata needs to define each response code
- Note that the terms README, data dictionary, and codebook all overlap – exact use depends on the field!

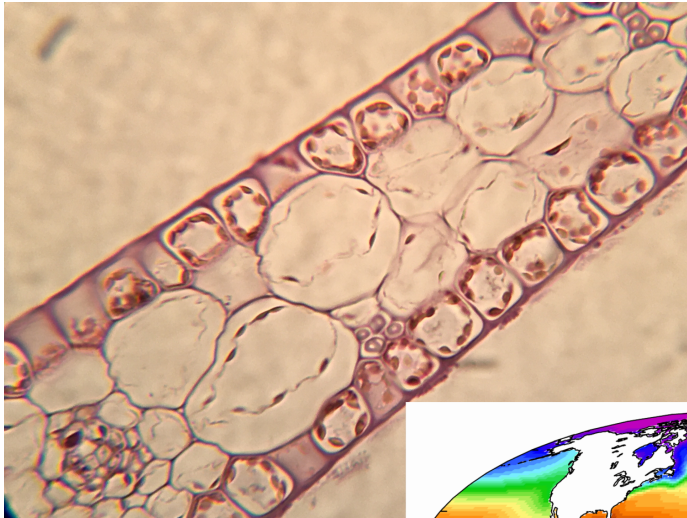
Question	Answer	Code
RespondentID		
Q1_Group_Learning	Never	1
	Once	2
	2–3 times	3
	4–5 times	4
	More than 5 times	5
Q2_Rate_Learning	Excellent	1
	Very Good	2
	Good	3
	Fair	4
	Poor	5
Q3_Leadership_Position	Yes	1
	No	2
	Don't Know	3
Q4_Communicate_Thoughts	Very easy	1
	Easy	2
	Unsure	3
	Difficult	4
	Very difficult	5
Q5_Team_Size	2 person	1
	3 person	2
	4 person	3
	5 person	4
	6 or more person	5
	I do not like to work in teams	6
Q6_Like_Most	Verbatim	
Q7_Like_Least	Verbatim	
Q8_Gender	Male	1
	Female	2
Q9_DOB	Verbatim	

*Sue and Ritter 2012*

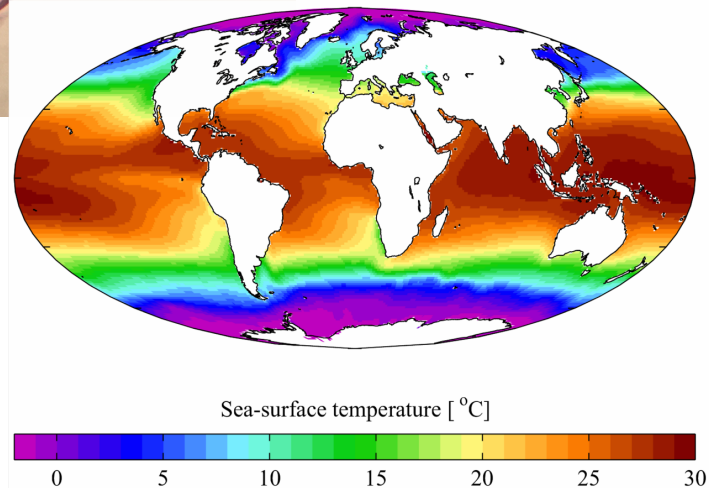
# Data dictionary best practices

- Within the data dictionary
  - List all the variables in the dataset
  - Include a definition, data type, and units
  - Include range or list of accepted values
  - Include missing value codes
- Across a project
  - Practice consistency in variable naming and definitions
  - Store data dictionary with the dataset
  - Use the data dictionary to make metadata explicit
  - Don't rely on your memory!

# Non-tabular data need metadata too



- Imagery
- Spatial data
- Spectral data
- Sensor data



- ReadMe and descriptive data dictionary can add value
- Goal is still to make metadata explicit and consistent

# Data dictionary exercise

## If you have your own project/data

- Start drafting a data dictionary for the tabular data in your project
- Consider what information to include, such as flags for quality control
- Do you encounter any unexpected properties of your data?

## If you are using the demo project

- Visit the [project repository](#)
- Scroll down to the “Data Table” details and click through the different attributes
- Are any of the variables unclear? Is any information missing?



# Metadata standards

- Many fields have developed metadata standards to describe data or resources
- Metadata standards require specific information and are validated
- You may need to create metadata according to a standard when you publish datasets in a repository

## Ecological Metadata Language (EML)

- EML is built on XML
- Extremely adaptable

## Biodiversity Information Standards (TDWG)

- Standards for biodiversity data, e.g. Darwin Core

## Examples of EML record

Knowledge  
Network for  
Biocomplexity



[Compiled annual statewide  
Alaskan salmon escapement  
counts, 1921-2017](#)

Environmental  
Data Initiative



[United States Pacific Northwest  
surveys of coastal foredune  
topography and vegetation  
abundance, 2012-2014](#)

# Standardizing terminology

- Various organizations publish lists of controlled vocabulary – including recommended and related terms
  - [Medical Subject Headings](#) from NIH for biomedical and life sciences
  - [NASA Thesaurus](#) for earth and space sciences
  - [USGS Thesaurus](#) for earth and environmental sciences
- These can be helpful when adding keywords to metadata
  - Keywords make your dataset more findable
  - Use specific keywords that are different from each other
  - [Some best practices for keywords from EDI](#)

## ezEML practice

- Visit <https://ezeml.edirepository.org/eml/>
- Log in with your GitHub, ORCID, or Google account
- Click 'EML Documents' and select 'New'
- Enter a name for your EML document – 'Demo EML document' or name with your project data
- Upload a data table – either your own tabular data or the demo CSV
- Work through the ezEML wizard to enter information and validate the metadata

# Working in reality

- Data documentation is additional to "real" work and can be hard to prioritize
  - What are some strategies to incorporate data documentation into workflow?
- Collaborators use different tools and have different workflows
  - How do you align metadata creation with collaborators?