

# Our Data Legacy: Metadata and Data Publication Best Practices

Trinh Nguyen, Senior Environmental Scientist Specialist

Interagency Ecological Program

California Department of Fish and Wildlife

April 18, 2024

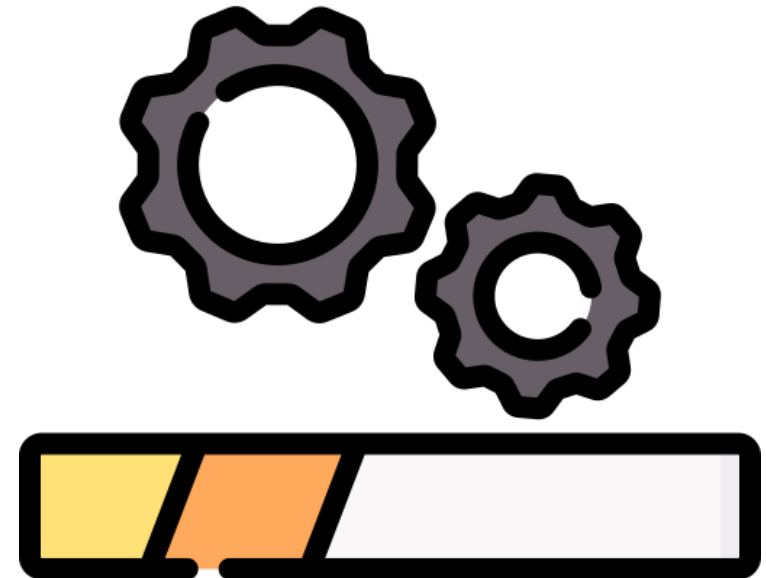


# What to Expect in This Course

- A detailed look at:
  - Metadata
  - Data publication best practices
- Practical guidance to these concepts
- Opportunities for collaboration

# More to Come!

- Full training later in the year, day-long
- More in-depth discussion on these concepts
- Workshop component to:
  - Showcase practical guidance
  - Application to your own data package

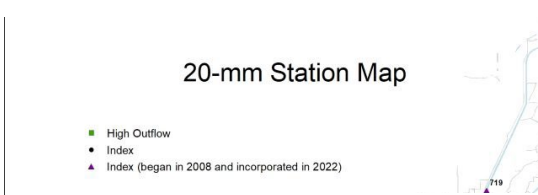


# Table of Contents

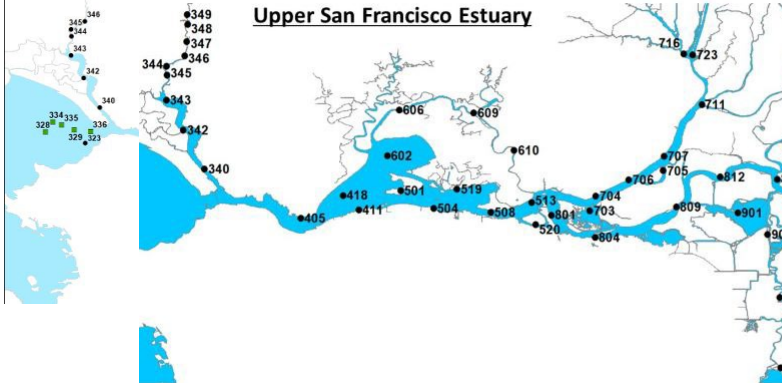
- Vision
- Defining metadata
  - What are the benefits?
  - How do we do this?
- Defining data publication
  - What are the benefits?
  - How do we do this?
- Culture

# Vision

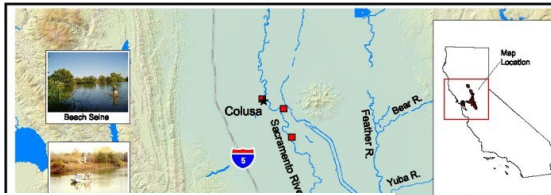
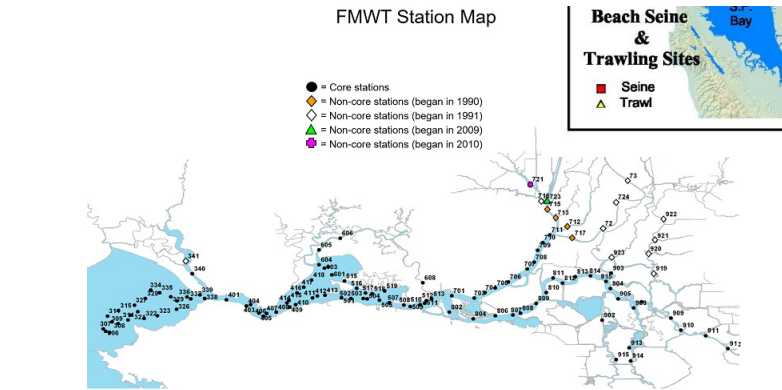
20-mm Station Map



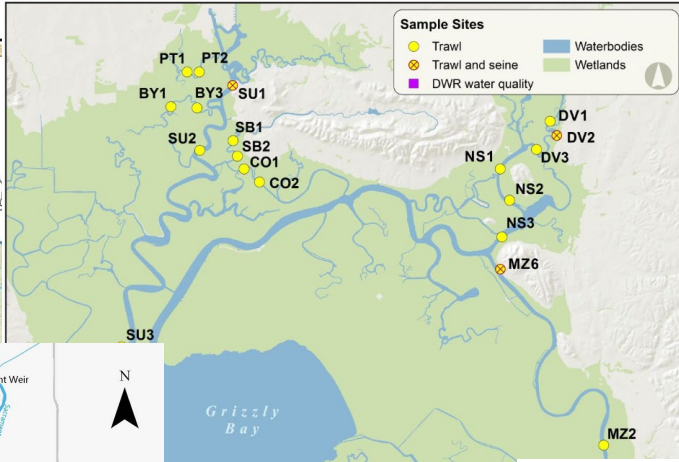
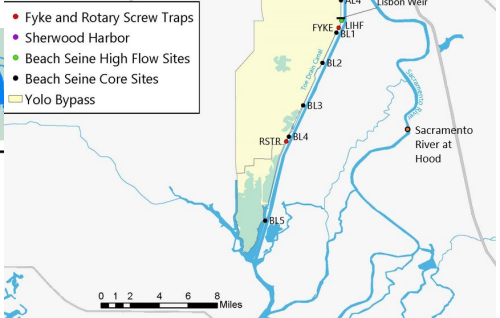
Upper San Francisco Estuary



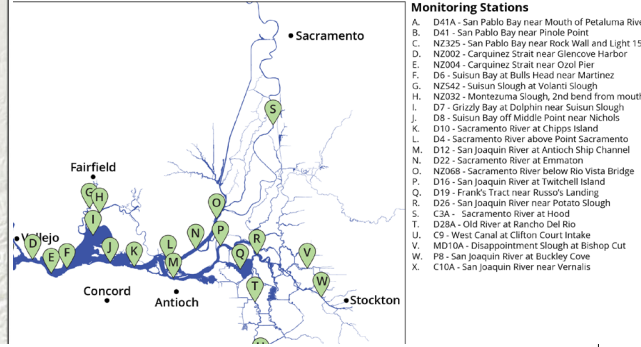
FMWT Station Map



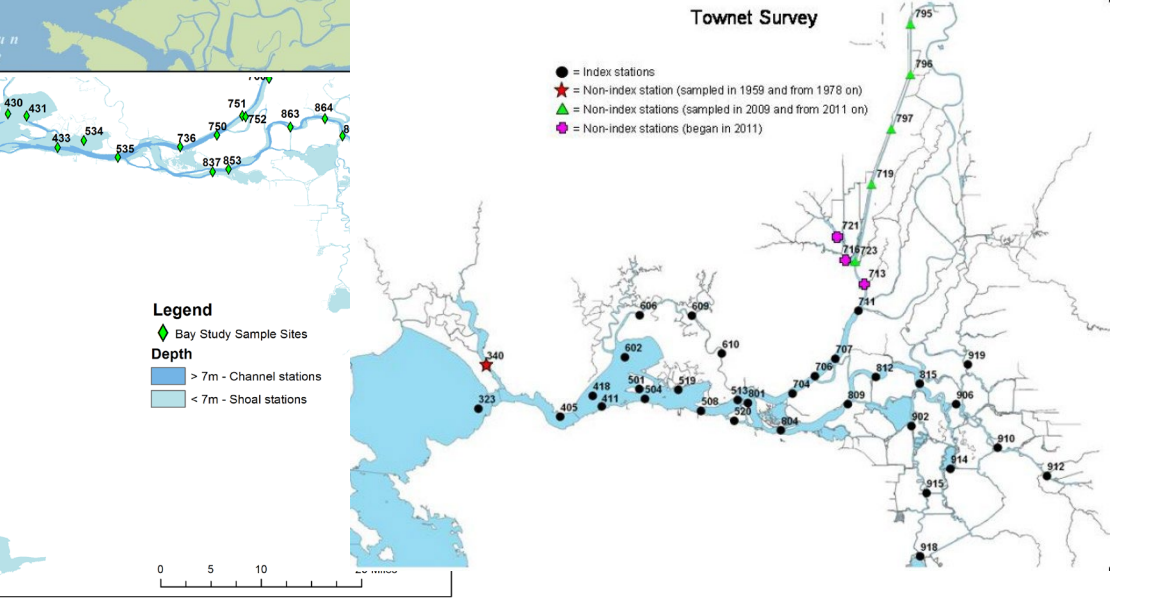
Beach Seine & Trawling Sites



Environmental Monitoring Program - Phytoplankton

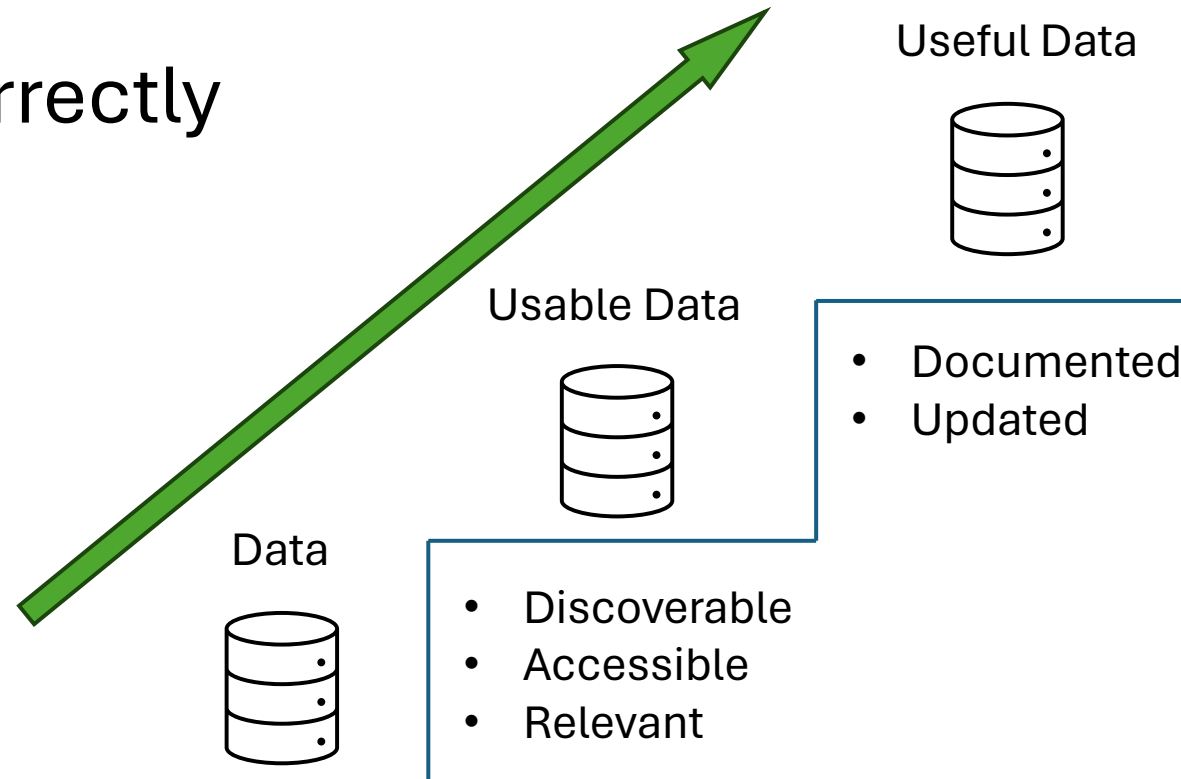


Townnet Survey



# Vision: Useful Data

- Data is valuable and impactful if:
  - It is used
  - It is used correctly

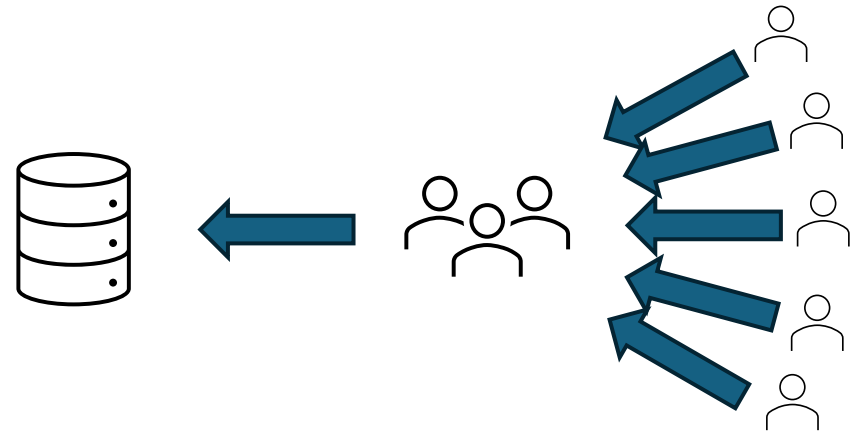
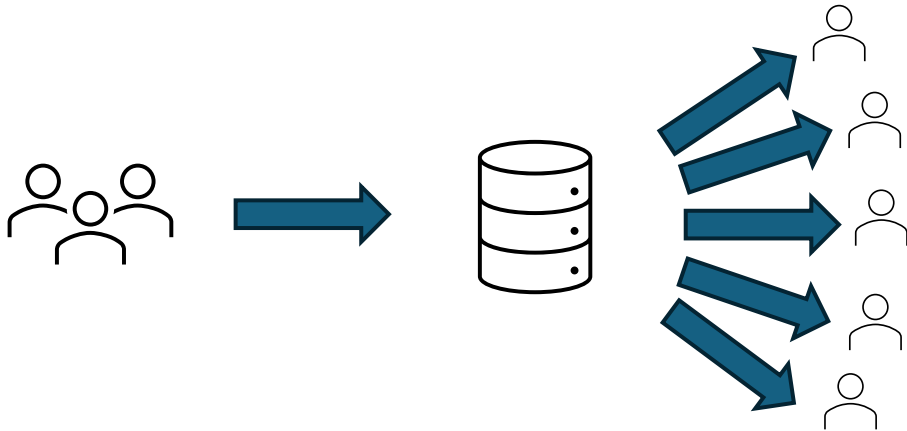


# Vision: Data Stewards

Data producers



Data stewards



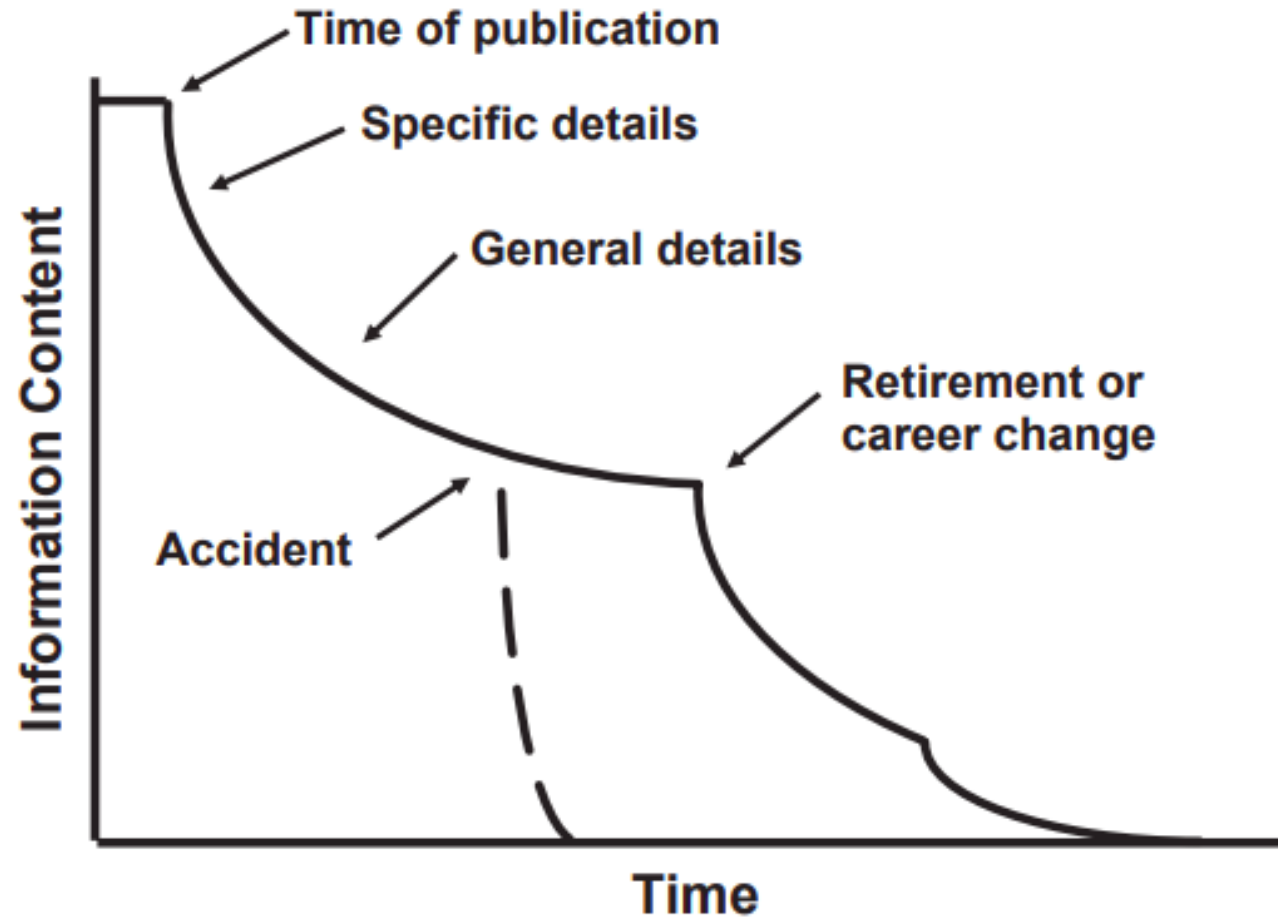
# Vision: Legacy

- Management actions based on best available science
- Best available science depends on best available data



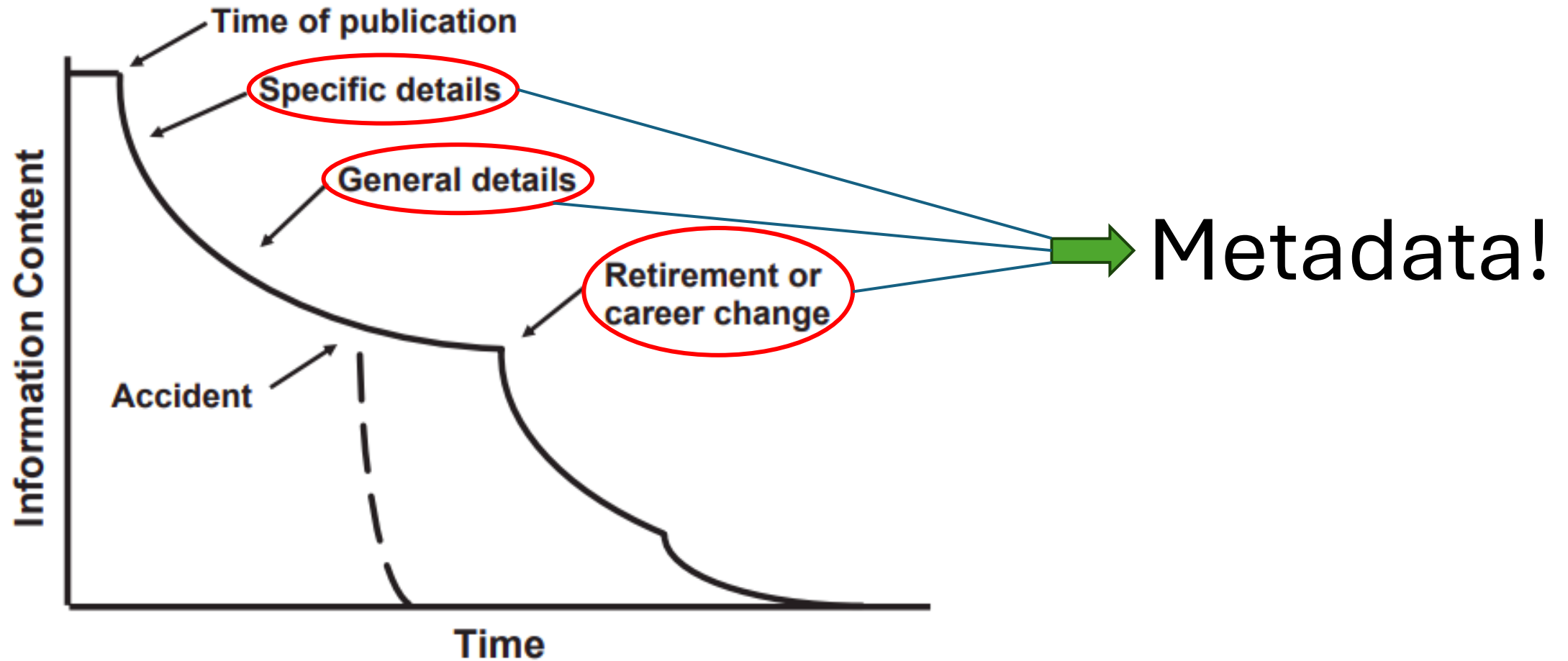


# Vision: Legacy



Adapted from Michener et al., 1987

# Fighting Information Loss



Adapted from Michener et al., 1987

# How comfortable are you with the concept of metadata?

- A) Very comfortable, I work with metadata regularly
- B) Somewhat comfortable, I have a basic understanding
- C) Not very comfortable, I need more information
- D) I am unfamiliar with the term “metadata”

# Metadata fundamentals

- Metadata is “data about data”
  - Many types, based on field and needs
    - Historical metadata
- Provides critical context to help us:
  - Understand
  - Manage
  - Find, Access, and Use

# How Useful can Metadata be: Only Data

1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

# How Useful can Metadata be: Column Names

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

# How Useful can Metadata be: Keys

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1 <sup>st</sup> , 2 = 2 <sup>nd</sup> , 3 = 3 <sup>rd</sup>
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard	
parch	# of parents / children aboard	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

# How Useful can Metadata be: Notes

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1 <sup>st</sup> , 2 = 2 <sup>nd</sup> , 3 = 3 <sup>rd</sup>
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard	
parch	# of parents / children aboard	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

## Variable Notes

**pclass:** A proxy for socio-economic status (SES)

1<sup>st</sup> = Upper, 2<sup>nd</sup> = Middle, 3<sup>rd</sup> = Lower

**age:** Age is fractional if less than 1. If the age is estimated, it is in the form of xx.5

**sibsp:** The dataset defines family relations in this way...

Siblings = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

**parch:** The dataset defines family relations in this way...

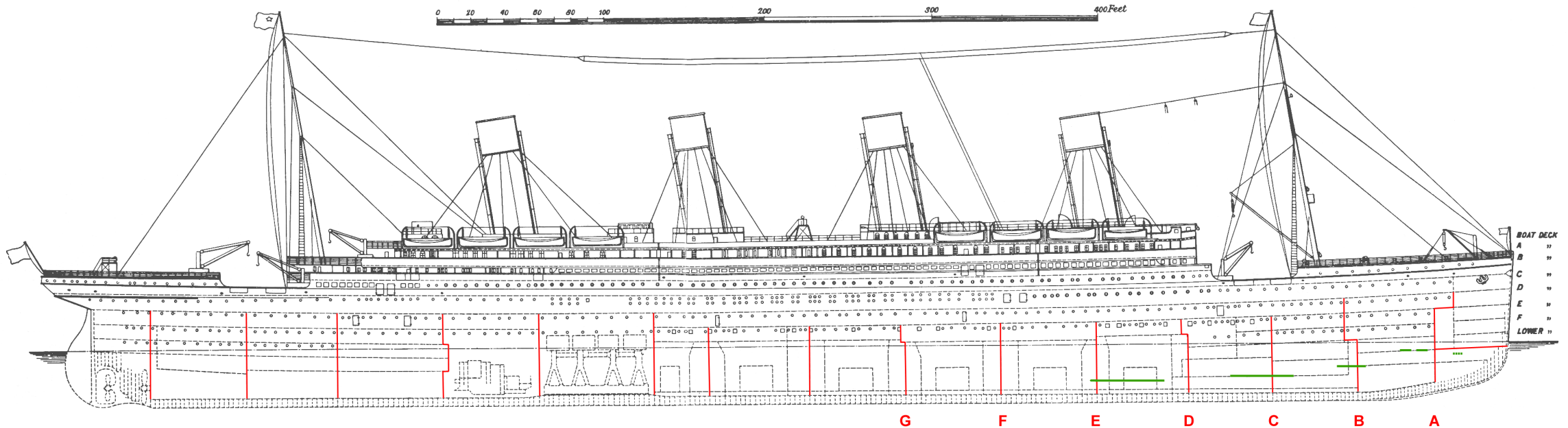
Parents = mother, father

Child = daughter, son, stepdaughter, stepson

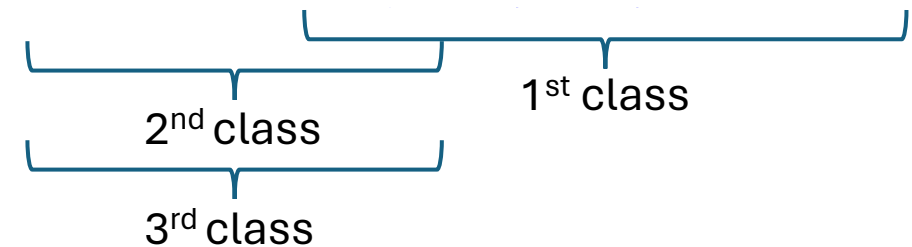
Some children travelled only with a nanny, therefore parch=0 for them.



# How Useful can Metadata be: Notes

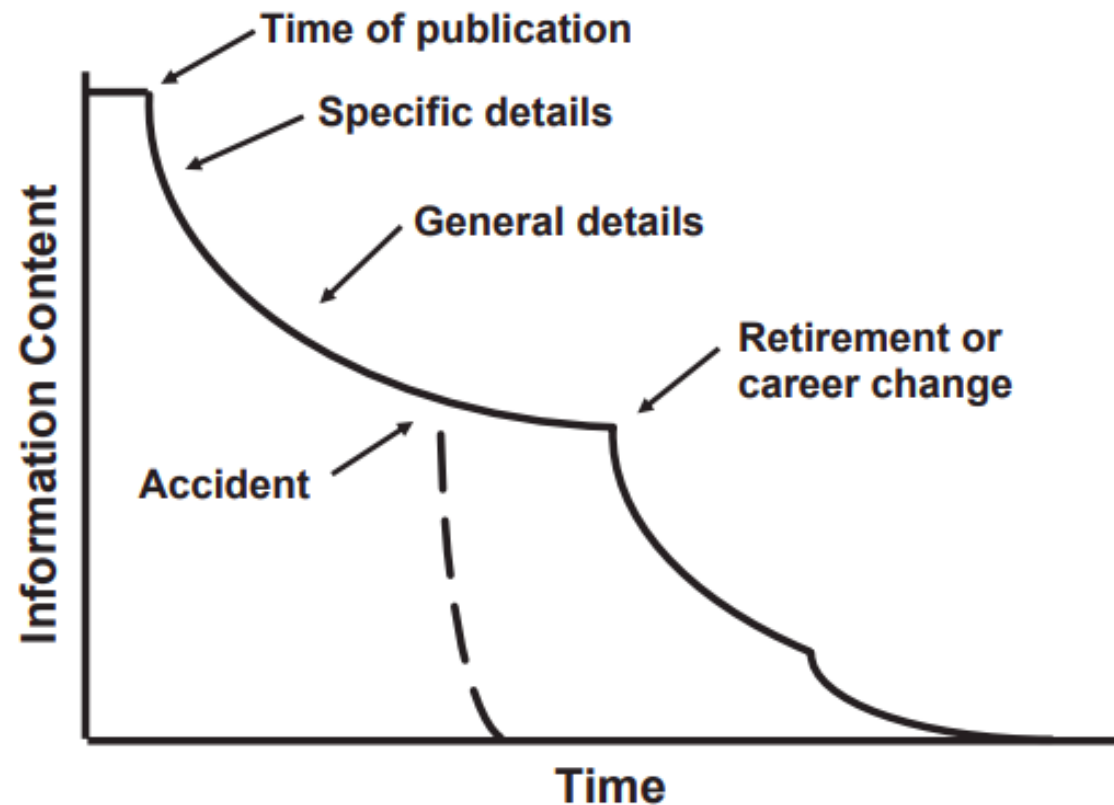


Name	Cabin
Braund, Mr. Owen Harris	
Cumings, Mrs. John Bradley (Florence Briggs Thayer)	C85
Heikkinen, Miss. Laina	
Futrelle, Mrs. Jacques Heath (Lily May Peel)	C123
Allen, Mr. William Henry	
Moran, Mr. James	
McCarthy, Mr. Timothy J	E46
Palsson, Master. Gosta Leonard	
Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	
Nasser, Mrs. Nicholas (Adele Achem)	
Sandstrom, Miss. Marguerite Rut	G6
Bonnell, Miss. Elizabeth	C103



# Why metadata matters

- A foundation to build the legacy of the dataset on



Adapted from Michener et al., 1987

# How important do you think comprehensive metadata documentation is for IEP datasets?

- A) Extremely important, essential for data usability
- B) Important, but not always necessary
- C) Somewhat important, mainly to know what the columns are
- D) Not very important, the data speaks for itself

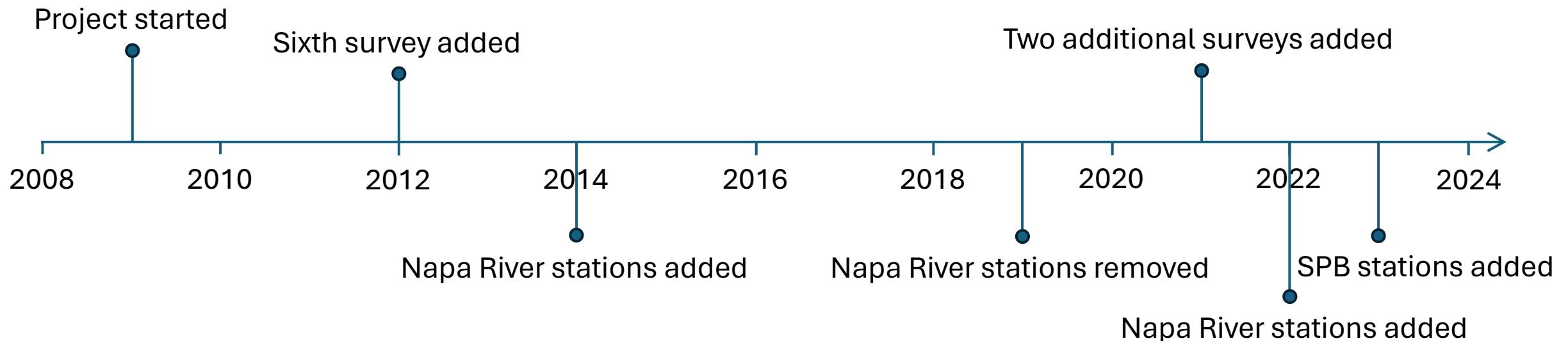
# IEP Datasets

- Long term-datasets
- Ecological data is diverse: IEP data spans many types

Dataset	Start Year	End Year	Target
Environmental Monitoring Program	1971	Present	Phyto- and Zooplankton, WQ and Nutrients, Benthic
Bay Study	1980	Present	Fish, Shrimp, Crab, WQ
Summer Tow Net	1959	Present	Fish, WQ
Fall Midwater Trawl	1967	Present	Fish, WQ, Zooplankton
Yolo Bypass	1998	Present	Fish, WQ, Zooplankton
Delta Juvenile Fish Monitoring Program	1970s	Present	Fish, WQ, Zooplankton
20mm	1995	Present	Larval Fish, Zooplankton

# Importance of Historical Metadata

- Refers to documentation of the origin, modifications, and movements of data over time.
- Extremely important for IEP surveys:
  - Long-term surveys
  - Tendency to focus only on current metadata

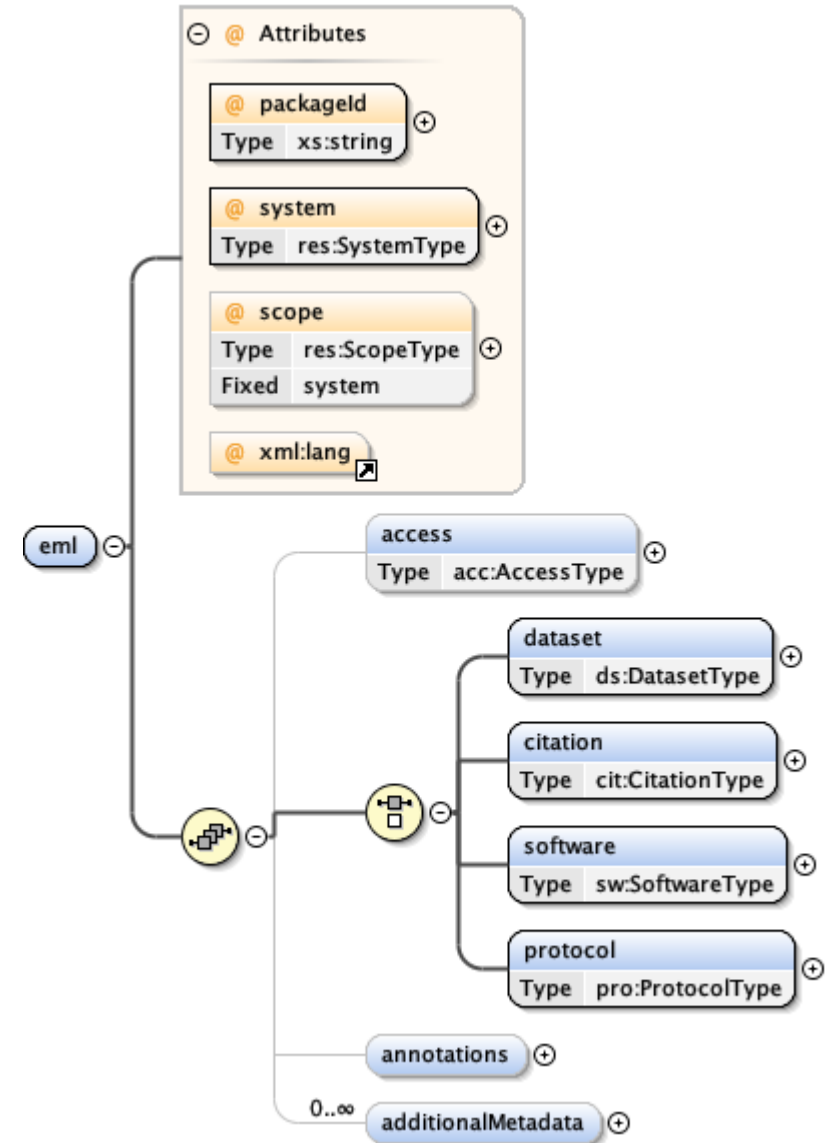


# Where do we Start: Metadata Standards

- List of standardized elements to facilitate consistency
- Structured approach enables:
  - Discovery
  - Accessibility
  - Automation
  - Reuse
  - Interoperability
- Ease of use, for all
  - ↳ Metadata
    - ↳ Element – Title
    - ↳ Element – PIs
    - ↳ Element – Publication Date
    - ↳ Element – Abstract
    - ↳ Element – Package ID
    - ↳ Element – Data Table(s)
    - ↳ Element – Intellectual Rights

# Metadata Standards: EML

- Ecological metadata language
- Curated specifically for ecological data
- Nested structure
  - XML
    - ↳ Element
      - ↳ (Elements)
      - ↳ (Attributes)
      - ↳ Attributes
- Ideal for complex ecological data



# EML Example

- SLS
  - Creator
    - Organization Name: Interagency Ecological Program (IEP)
  - Data Entities
    - Data Table
      - Name: Catch.csv
      - Description: Fish catch data from the Smelt Larval Survey
      - Table Structure
        - Size: 297714 byte
      - Table Column Descriptions
        - Attribute
          - Column Name: Date
          - Definition: Date when sampling occurred
        - Attribute
          - Column Name: Station
          - Definition: Project station number



# Metadata Standards: Adopting EML

- Made with ecological data in mind
- Comprehensive
- Extendable
  - Important to serve IEP datasets, e.g., historical metadata
- Officially supported standard of the Environmental Data Initiative (EDI)
  - Infrastructure in place to help guide users
    - ezEML
    - EMLassemblyline (EAL)

**Title**

Data Tables

Creators

Contacts

Associated Parties

Metadata Providers

Abstract

Keywords

Intellectual Rights

Geographic Coverage

Temporal Coverage

Taxonomic Coverage

Maintenance

Publisher

Publication Info

Methods

Project

Other Entities

Data Package ID

Check Metadata ●

Check Data Tables ●

Explore Data Tables

Submit/Share Package

## Title

Enter a

Title \*

Save

## Welcome to ezEML



Welcome to ezEML! Here is some initial guidance on its use.

You can think of using ezEML as filling out one big form that's organized as a number of separate pages. As you move from page to page, your changes are automatically retained (as you'd expect when filling out a form), except in the cases where a page has a **Cancel** button, which lets you exit that page without saving.

**Save and Continue** buttons save any changes you've made on the current page and move you ahead to the next section of the metadata. You can step through the sections sequentially using **Save and Continue**, or you can click any of the links under **Contents** in the left margin to save changes and jump directly to a particular section of the metadata.

**Reset Changes** buttons clear the changes you've made on the current page and leave you on that page.

The **ezEML** link at the top left of each page serves as a "Home" button, taking you back to the top of the Contents list.

It's usually best to refrain from using your browser's forward and back buttons. Using them may cause you to lose unsaved edits.

Notice that there are lots of Help buttons in ezEML (see the circular question mark icons). You may find them helpful.

If you click the checkbox below we won't show you this welcome dialog again, but you can always review how to navigate in ezEML either by clicking the help button next to **Reset Changes** on this **Title** page or by going to [Navigating in ezEML](#) in the [User Guide](#).

You can also watch a short demo video here: **YouTube**

We hope you find ezEML useful, and we welcome your feedback at [support@edirepository.org](mailto:support@edirepository.org).

☐ Got it. Don't show this dialog again.



## Contents

### Title

Data Tables

Creators

Contacts

Associated Parties

Metadata Providers

Abstract

Keywords

Intellectual Rights

Geographic Coverage

Temporal Coverage

Taxonomic Coverage

Maintenance

Publisher

Publication Info

Methods

Project

Other Entities

Data Package ID

Check Metadata 

Check Data Tables 

Explore Data Tables

Submit/Share Package

# Title

Enter a title for the data package:

Title \*



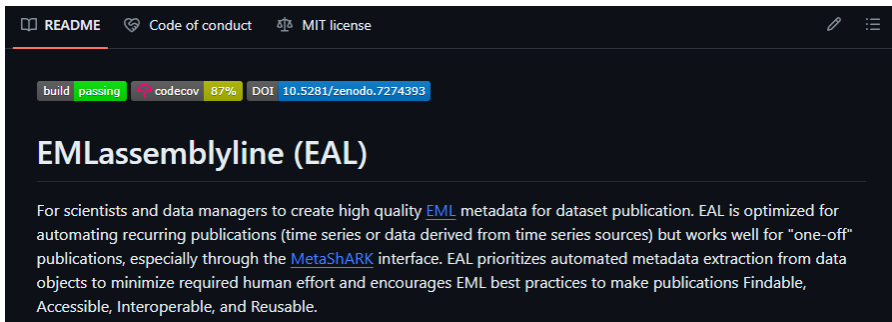
Save and Continue

Reset Changes



# EMLassemblyline

- Programmatically create EML metadata, R
- Upload data package to a repository
- Automate data publication process
- <https://github.com/EDIdorg/EMLassemblyline?tab=readme-ov-file>



## Instructions for the EML assembly line

Original Author: Colin Smith (EDI)

### IEP test-case comments Brittany Davis

Expect to take about a day to become familiar with the [EML assembly line](#) R package.

Edits by Rosemary Hartman, CDFW 7 Dec 2018

Edits by Ryan Mckenzie, USFWS 21 April 2020

Edits by Rosemary Hartman DWR 23 June 2021

### Overview

The EML assembly line will help you create high quality metadata for your dataset. Below is a set of step-by-step instructions for making EML metadata for tabular data and other data, including spatial vector, spatial raster, and images.

### Installation (periodic reinstallation is recommended)

The assembly line is under constant revision and improvement. Please reinstall the assembly line periodically to ensure a successful experience. Installation from GitHub requires the devtools package.

```
# Install devtools
install.packages("devtools")

# Load devtools
library(devtools)

# Install and load EMLassemblyline
install_github("EDIdorg/EMLassemblyline")
library(EMLassemblyline)
```

# Template guides

- IEP EDI Metadata Template 2020.docx

## EDI/IEP Metadata Template (2020)<sup>1</sup>

This template follows the Ecological Metadata Language schema for generation of standardized metadata. Additional notes or examples for Interagency Ecological datasets can be seen in **green** color.

*Data should be in csv text file. If starting with an Excel spreadsheet, please make sure it does not contain any formulas and comments on cells. If you need comments put them in their own column. If data were used in a database and major table linking is necessary to analyze, please de-normalize into a flat file, not just database table exports. This will often be several flat files, especially for large, complex data sets. **Replace any empty cells with "NA".***

### Dataset Title

*Be descriptive, more than 5 words. This should include IEP in the title, for example "Summer Townet Survey" should become "Interagency Ecological Program Sacramento-San Joaquin Delta Summer Townet Survey for Juvenile Fish 1969-2018", or something similar:*

Interagency Ecological Program:

### Short name or nickname you use to refer to this dataset:

*Start all IEP data sets with "IEP-". For example the Yolo Bypass Fish Monitoring Program, could be IEP-YBFMP.*

### Abstract

*Include what, why, where, when, and how. This may be copied directly from the IEP workplan checklist, but will likely need edits to make sure it is readable for a wider audience.*

<b>Title:</b> Metadata Template	<b>Status:</b> Effective
	<b>Approval Date:</b> 6/9/2021
<b>Document Number:</b> DWR-1-TEM-002	<b>Version:</b> 1.0

*This template is based on the Interagency Ecological Program (IEP) Metadata template and the Environmental Data Initiative (EDI) Metadata template. The template follows the Ecological Metadata Language schema for generation of standardized metadata. Notes from the IEP template are in **green**. Additional notes are in **red**. Notes and instructions are italicized.*

*Data should be in csv text file. If starting with an Excel spreadsheet, please make sure it does not contain any formulas and comments on cells. If you need comments put them in their own column. If data were used in a database and major table linking is necessary to analyze, please de-normalize into a flat file, not just database table exports.*

### Dataset Title

*(be descriptive, more than 5 words):*

*If this is an IEP dataset, include Interagency Ecological Program in the title.*

### Abstract

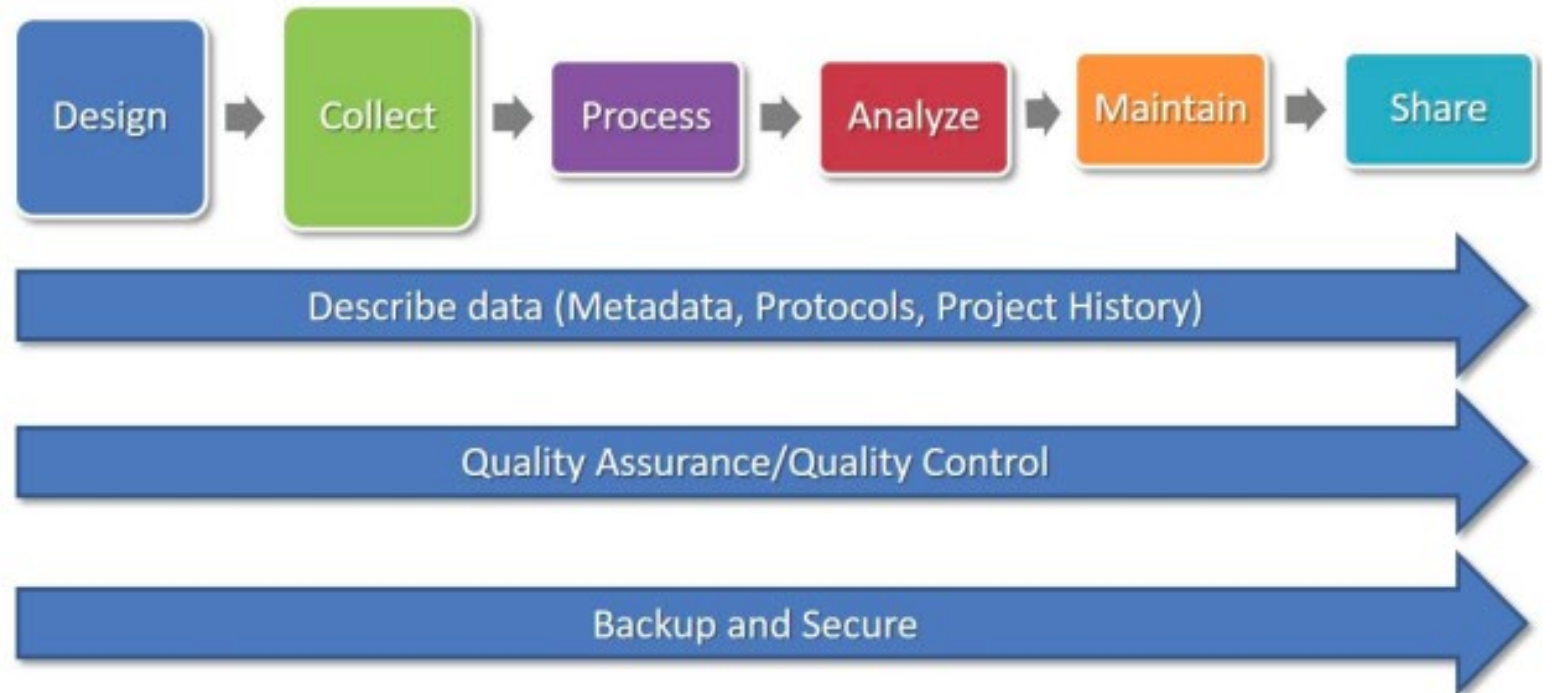
*(include what, why, where, when, and how)*

# Which tool can you see yourself using to document rich metadata?

- A) ezEML, the wizard seems straightforward enough
- B) EAL, I am comfortable using R and automation is critical
- C) IEP Word Document, a pre-formatted template is easy to follow
- D) None, I prefer to use other methods

# Data Publication: Goal

- Making our research data publicly available to facilitate:
  - Data discovery
  - Access and use
  - Reuse



# Data Publication: Benefits

Stewards	Users	Community
Contractual obligations	Ease of accessibility	Increased collaboration
Enhances transparency	Increased data quality and trustworthiness	Accelerates scientific progress
Increased visibility	Increased use and reuse	Promotes open science
Increased recognition	Increased efficiency	Supports evidence-based decision making
Increased impact		
Improved data management		
Promotes data preservation		



# Where do we start: FAIR

- FAIR principles: meant to enhance
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
- Developed by an international team (Wilkinson et al., 2016)
  - Gold standard for the industry



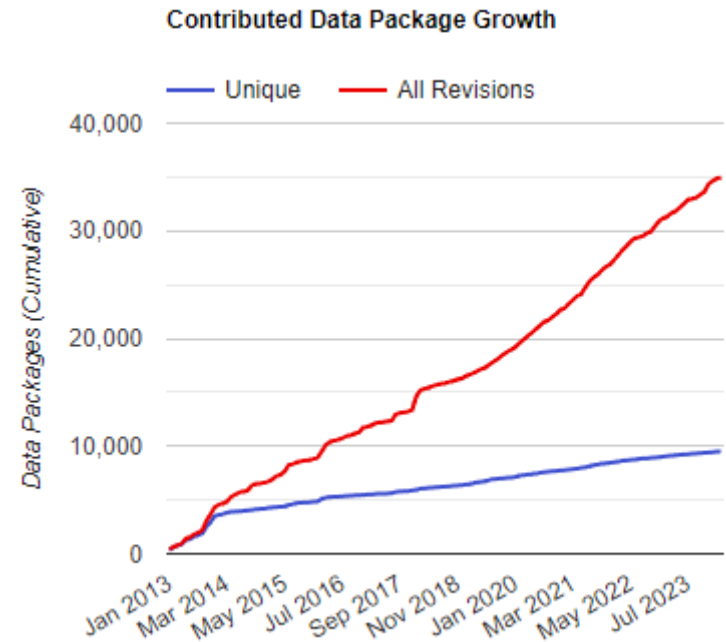
# FAIR: Summary

- <https://www.go-fair.org/fair-principles/>
- Revolves around:
  - Document metadata thoroughly
  - Adopt a metadata standard
  - Lean on accepted infrastructure



# FAIR: EDI Repository

- Environmental Data Initiative (EDI)
  - A data repository
- Existing infrastructure to support FAIR data
  - Unique and persistent DOIs per data package
  - Utilizes EML to produce rich metadata
  - Part of a data federation, DataONE
  - Access with any browser or through an API
  - Guidance and support
    - ezEML, EMLassemblyline, assessment reports
    - Dedicated curation team
  - Version control and data preservation



# FAIR and EML: Great Initial Steps

- Ecological data is diverse and unique to each system
- Things to consider:
  - System-specific data types
    - Long-term datasets
  - System-specific reporting requirements
    - Required data portals, reporting frequencies
  - Existing infrastructure
    - Excel or Access based
    - Scripted QAQC queries



# Data Publication Working Group






- A DUWG subgroup focused on providing guidance more specific to our system
- Two main functions:
  - Develop guidance for the community:
    - Guidelines, e.g., data publication best practices
    - Templates, e.g., historical metadata template
    - Guides, e.g., publishing to EDI









# DUGW and DPWG Products

- <https://iep.ca.gov/Data/Data-Utilization-Working-Group>
- Homepage > Data > Data Utilization Working Group

## Data Publication Materials

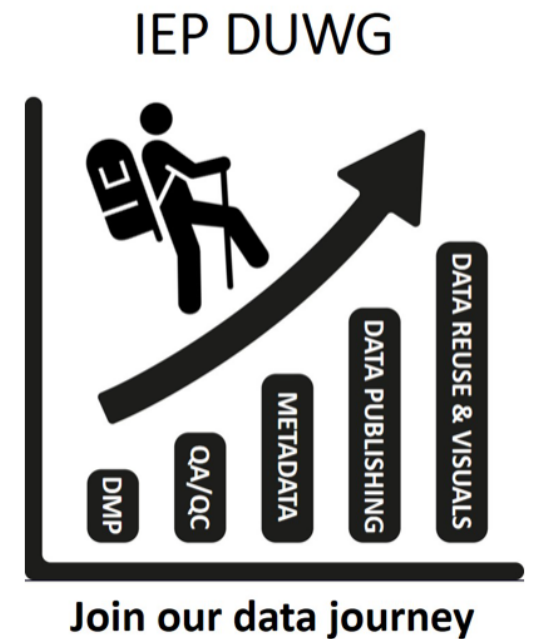
- [Guidelines for Data Publishing and Formatting \(PDF\)](#) 
- Code for formatting metadata in Ecological Metadata Language format and publishing to the Environmental Data Initiative (EDI) website is available on the [DUWG GitHub page](#) .
- [Publishing IEP datasets to EDI \(PDF\)](#) 
- [Connecting EDI datasets to the California Natural Resources Agency Open Data Platform \(PDF\)](#) 
- [EML assembly line instructions for creating machine-readable metadata \(PDF\)](#) 

## Templates and Guidelines

- [IEP Historical Metadata Template \(PDF\)](#)  **NEW!**
- [IEP Digital Datasheet Best Practices \(PDF\)](#)  **NEW!**
- [IEP Fish Data Quality Control Best Practices \(PDF\)](#)  **NEW!**
- [IEP Data Management Plan template \(Word\)](#)
- [IEP Data Management Plan instructions \(PDF\)](#) 
- [IEP Metadata template \(PDF\)](#) 
- [Synthesis Data Management and QAQC guidelines \(PDF\)](#) 
- [Vocabulary Crosswalk and Standards \(Excel\)](#)
- [IEP Standard Operating Procedures template \(Word\)](#)

# Data Publication Working Group

- A DUWG subgroup focused on providing guidance more specific to our system
- Two main functions:
  - Develop guidance for the community:
    - Guidelines, e.g., data publication best practices
    - Templates, e.g., historical metadata template
    - Guides, e.g., publishing to EDI
  - Serve as an open forum to help IEP surveys through the publication process
    - Provide relevant guidance
    - Review drafted publication packages



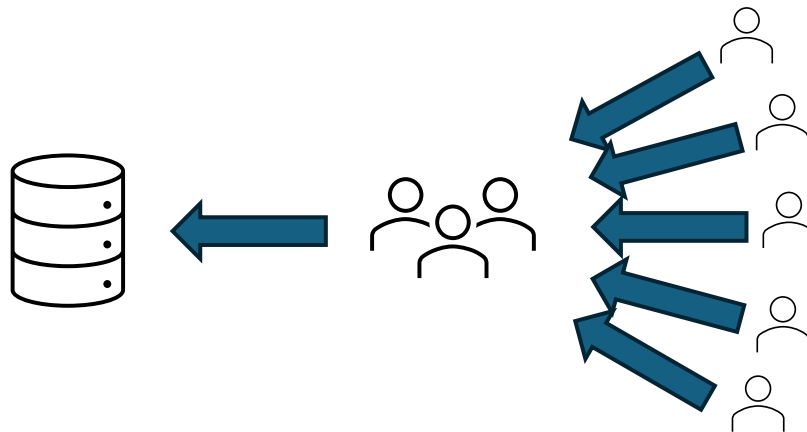
# Which data repositories do you prefer to use for data publication?

- A) EDI Repository
- B) California Natural Resources Agency Open Data Platform
- C) In-house network, e.g., FTP
- D) None, reach out to me personally for data



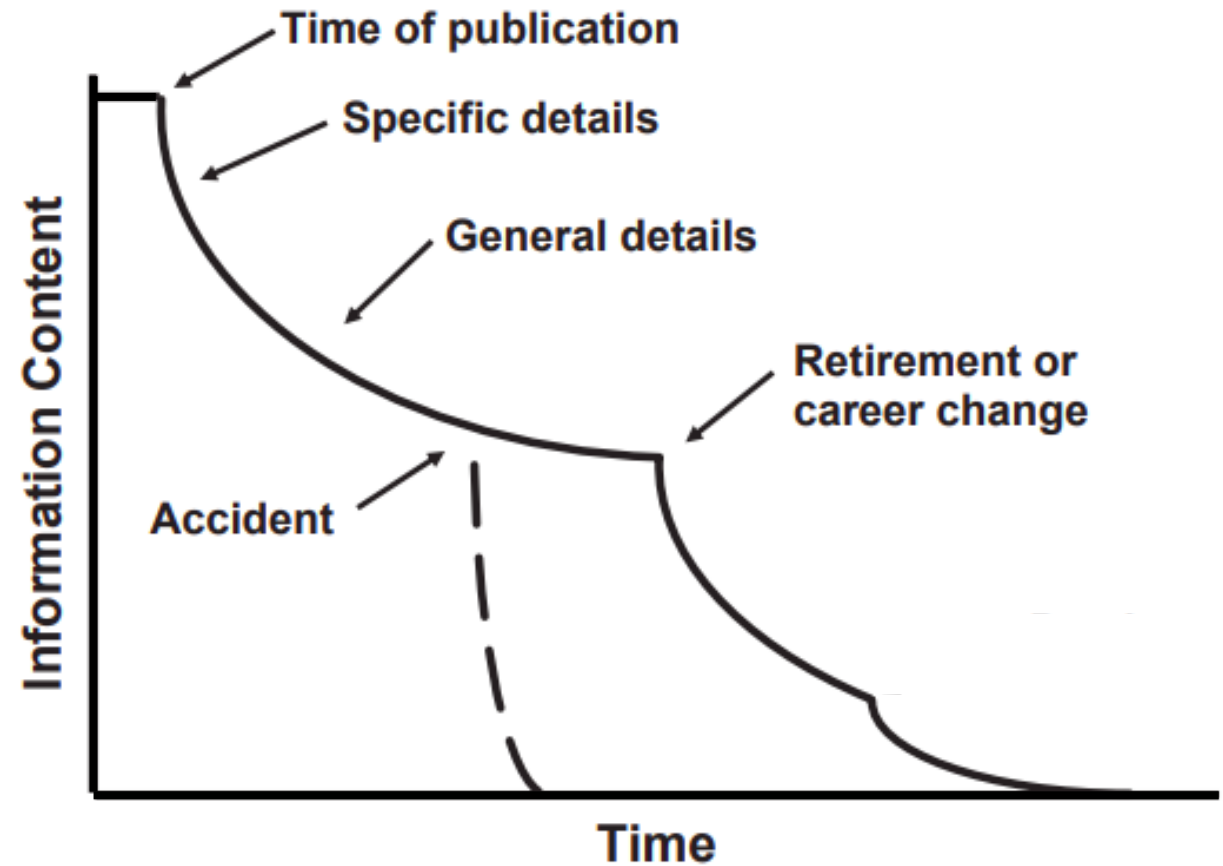
# A Barrier to Consider: Culture

- Transition from data producers to data stewards
- Data stewards manage and oversee our datasets to help users discover and use our datasets correctly
- Data stewardship: two parts
  - Technical guidelines
  - Attitudes, behavior, and practices



# Why should we care?

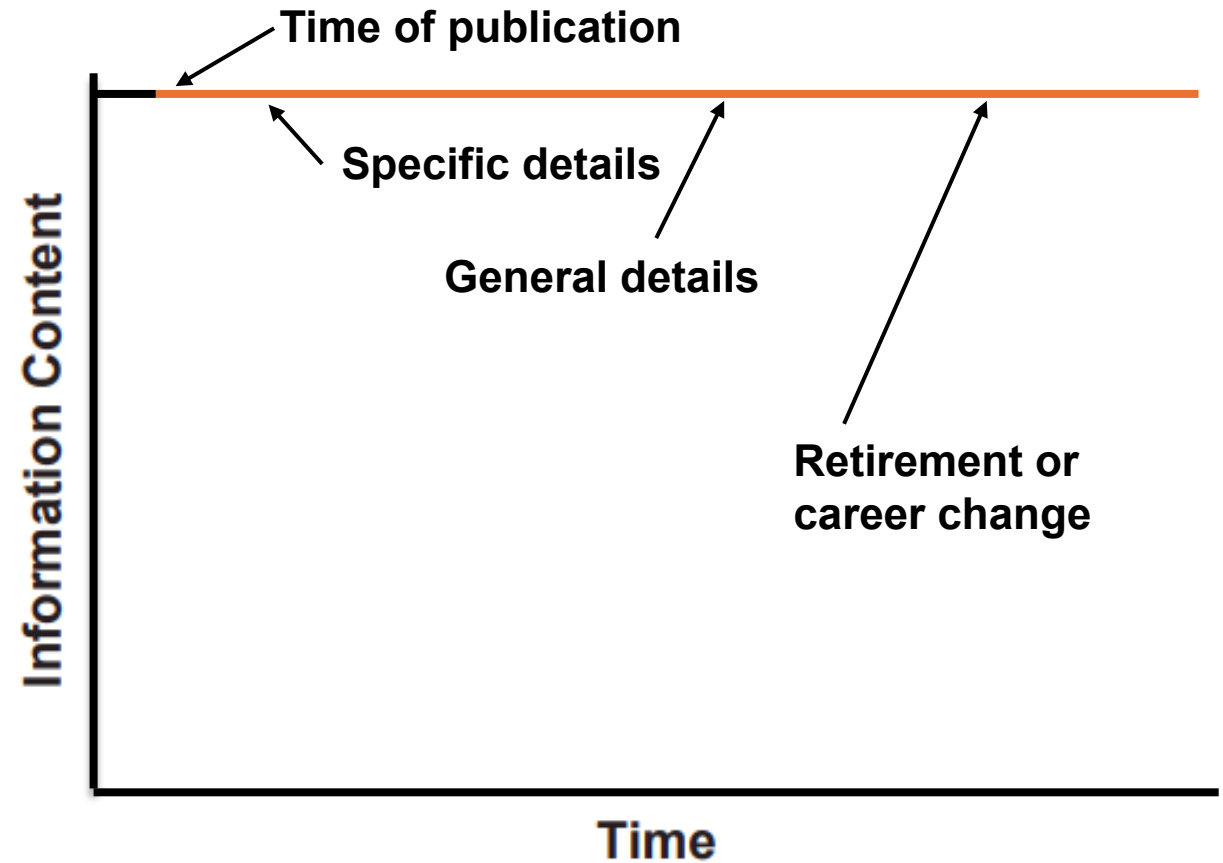
- Best available science depends on best available data
- Helps drive meaningful changes to our system and field
- The future is infinite, and so is the potential use of our data
- Each line of useful data that we produce is a part of our legacy we leave behind



Adapted from Michener et al., 1987

# Why should we care?

- Best available science depends on best available data
- Helps drive meaningful changes to our system and field
- The future is infinite, and so is the potential use of our data
- Each line of useful data that we produce is a part of our legacy we leave behind



Adapted from Michener et al., 1987

# How can we start?

- Prioritize documenting comprehensive metadata for all IEP datasets
- Explore publication through the EDI repository
- Engage with the DPWG to follow and develop best practices
- Develop and embrace a culture of data stewardship to leave a lasting legacy through our research data

Thank you!

