Publishing Data

NCEAS Learning Hub

for

Delta Science Program

October 2023

Learning Objectives

- Overview best practices for organizing data for publication

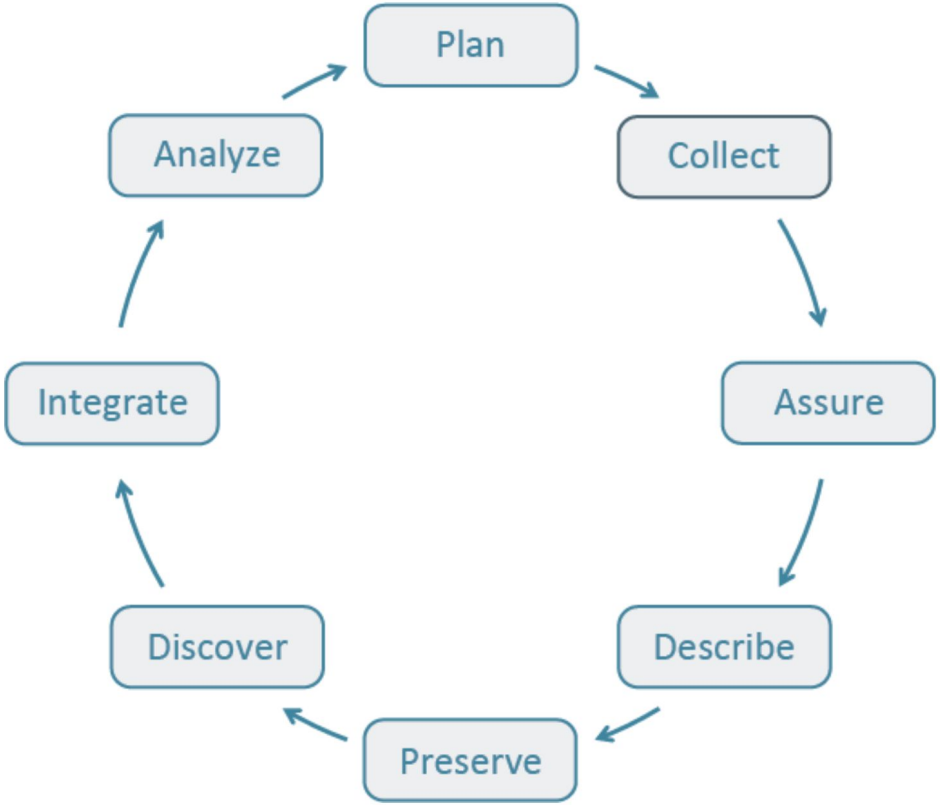
Learning Objectives

- Overview best practices for organizing data for publication
- Review what science metadata is and how it can be used

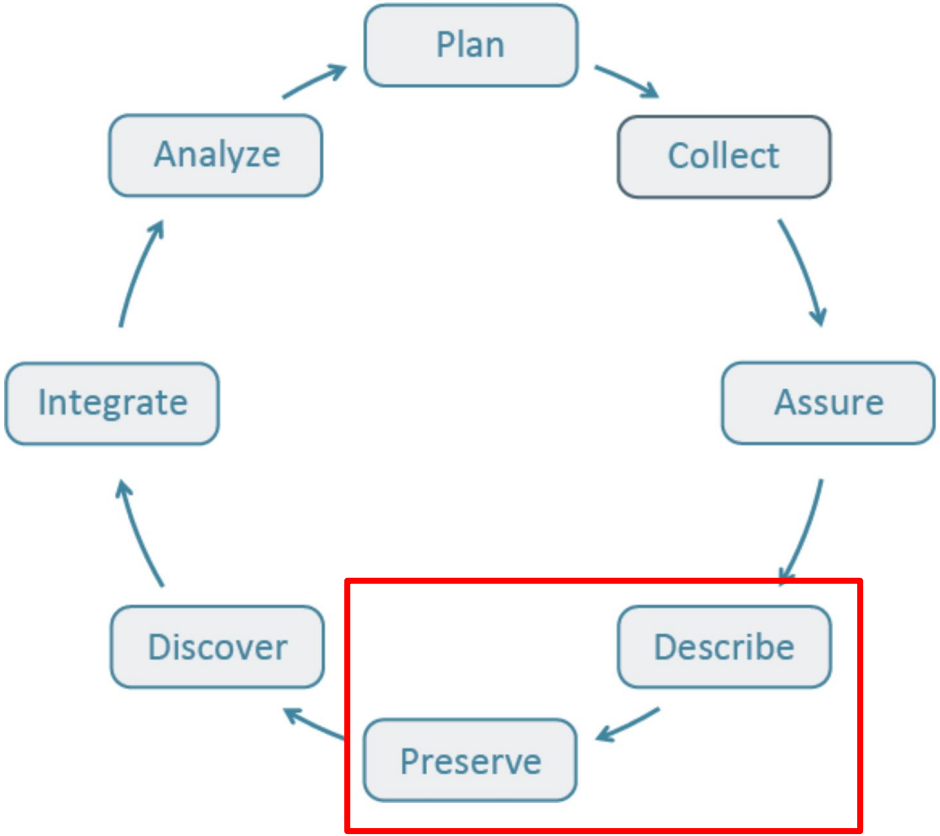
Learning Objectives

- Overview best practices for organizing data for publication
- Review what science metadata is and how it can be used
- Demonstrate how data and code can be documented and published in open data archives

Data Life Cycle

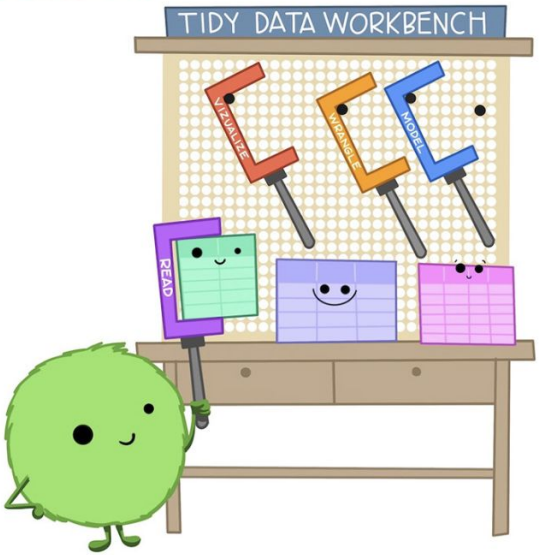


Data Life Cycle



Organizing Data ~Tidy Data

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.

Organizing Data - Best Practices

Clean data programmatically

Organizing Data - Best Practices

Clean data programmatically

Design your tables to add rows (not columns)

Organizing Data - Best Practices

Clean data programmatically

Design your tables to add rows (not columns)

Include header lines in tables

Organizing Data - Best Practices

Clean data programmatically

Use non-proprietary formats

Design your tables to add rows (not columns)

Include header lines in tables

Organizing Data - Best Practices

Clean data programmatically

Use non-proprietary formats

Design your tables to add rows (not columns)

Have descriptive files names (with no spaces)

Include header lines in tables

Organizing Data - Best Practices

- Clean data programmatically
- Use non-proprietary formats
- Design your tables to add rows (not columns)
- Have descriptive files names (with no spaces)
- Include header lines in tables
- Make sure your file paths are reproducible

Organizing Data - Best Practices

Large Data Packages

When you have or are going to generate large data packages (in the terabytes or larger), it's important to establish a relationship with the data center early on.

The data center can help come up with a strategy to tile data structures by subset, such as by spatial region, by temporal window, or by measured variable. They can also help with choosing an efficient tool to store the data (ie NetCDF or HDF), which is a compact data format that helps parallel read and write libraries of data.

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

What was measured?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

Who measured it?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

When it was measured?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

Where was it measured?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

How was it measured?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

How is the data structured?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

Why was the data collected?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

Who should get credit?

Metadata

- The goal is to have enough information for researcher to understand the data, interpret the data, and then reuse the data in another study

Under what license this data can be reused?

Metadata Standards: EML

- How will computers organize and integrate this information?
- Ecological Metadata Language also known as EML is commonly use in the earth and environmental sciences.

> “The Ecological Metadata Language (EML) defines a comprehensive vocabulary and a readable XML markup syntax for documenting research data” (https://eml.ecoinformatics.org/)

EML & XML

- EML or Ecological Metadata Language is the name of the metadata standard.
- EML are stored in an XML file.
- XML (Extensible Markup Language), is a markup language that provides rules to define any data.
- XML file extension is `.xml`. So an EML file will be something like `metadata.xml`.

EML & XML

```
<?xml version="1.0" encoding="UTF-8"?>
<eml:eml packageId="df35d.442.6" system="knb"
    xmlns:eml="eml://ecoinformatics.org/eml-2.1.1">
    <dataset>
        <title>Improving Preseason Forecasts of Sockeye Salmon Runs through
            Salmon Smolt Monitoring in Kenai River, Alaska: 2005 - 2007</titl
        <creator id="1385594069457">
            <individualName>
                <givenName>Mark</givenName>
                <surName>Willette</surName>
            </individualName>
            <organizationName>Alaska Department of Fish and Game</organizatio
            <positionName>Fishery Biologist</positionName>
            <address>
                <city>Soldotna</city>
                <administrativeArea>Alaska</administrativeArea>
                <country>USA</country>
            </address>
            <phone phonetype="voice">(907)260-2911</phone>
            <electronicMailAddress>mark.willette@alaska.gov</electronicMailAd
        </creator>
        ...
    </dataset>
</eml:eml>
```

Data Identifier & Citation

- Many journals require a **DOI - a digital object identifier** - be assigned to the published data before the paper can be accepted for publication.

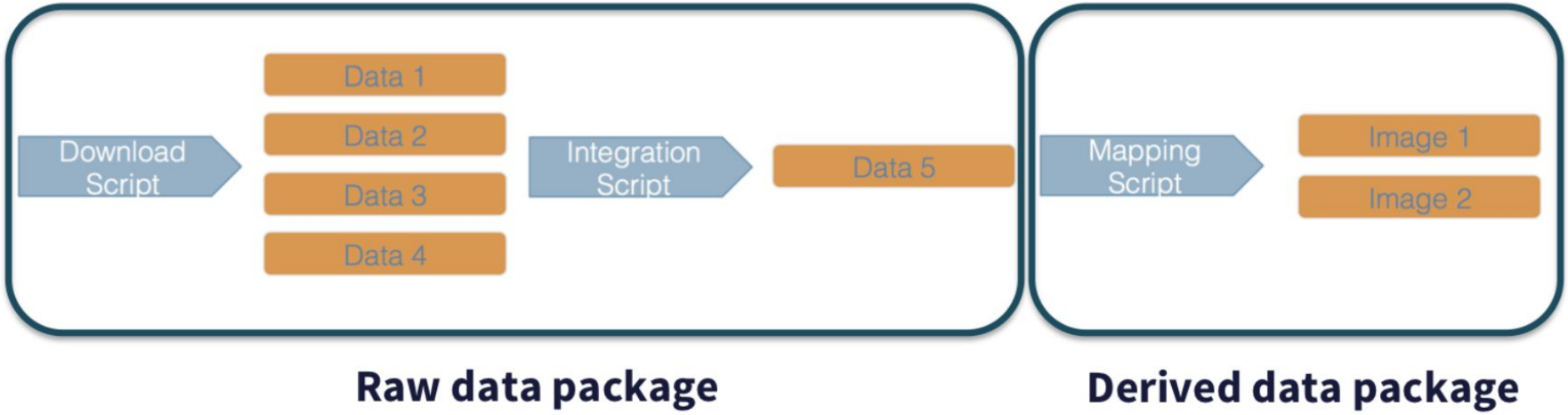
Data Identifier & Citation

- Many journals require a **DOI - a digital object identifier** - be assigned to the published data before the paper can be accepted for publication.
- Keep in mind that generally, if the data package needs to be updated (which happens in many cases), **each version of the package will get its own identifier.**

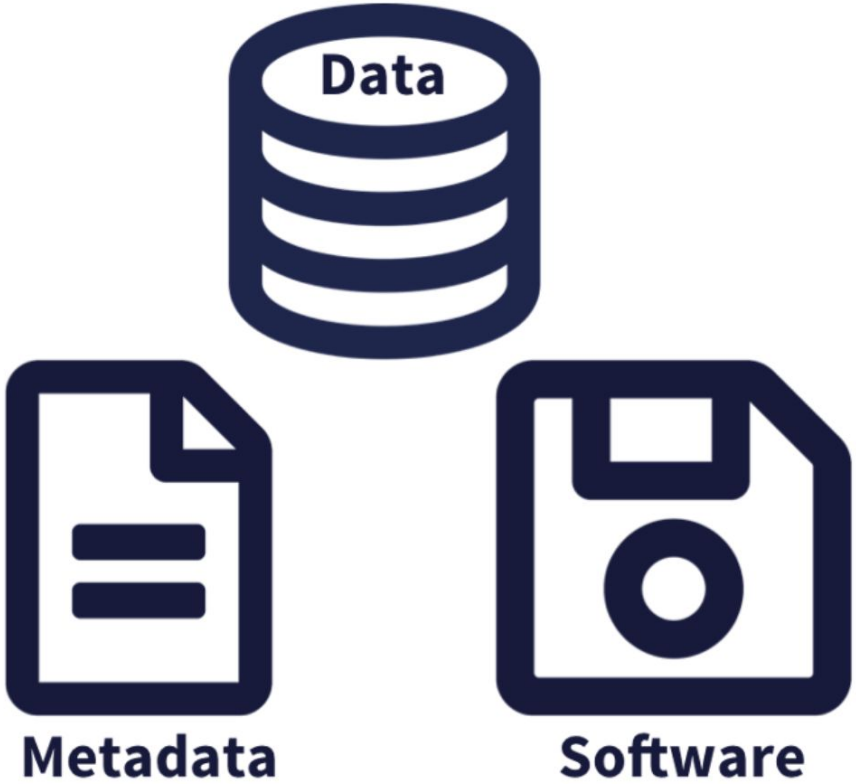
Data Identifier & Citation

- Many journals require a **DOI - a digital object identifier** - be assigned to the published data before the paper can be accepted for publication.
- Keep in mind that generally, if the data package needs to be updated (which happens in many cases), **each version of the package will get its own identifier.**
- Researchers should get in the habit of **citing the data that they use** (even if it's their own data!) in each publication that uses that data.

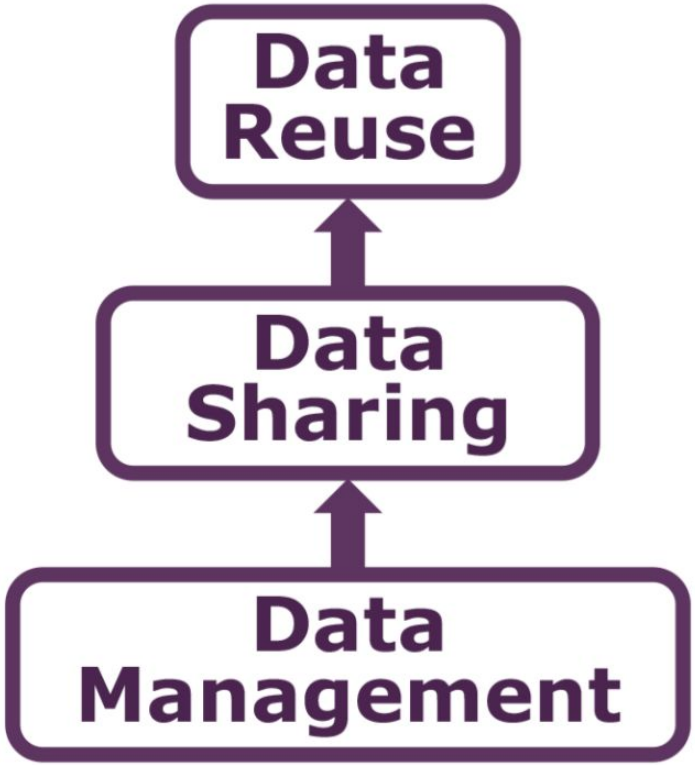
Provenance & Computational Workflows



Provenance & Computational Workflows



Preserving Your Data

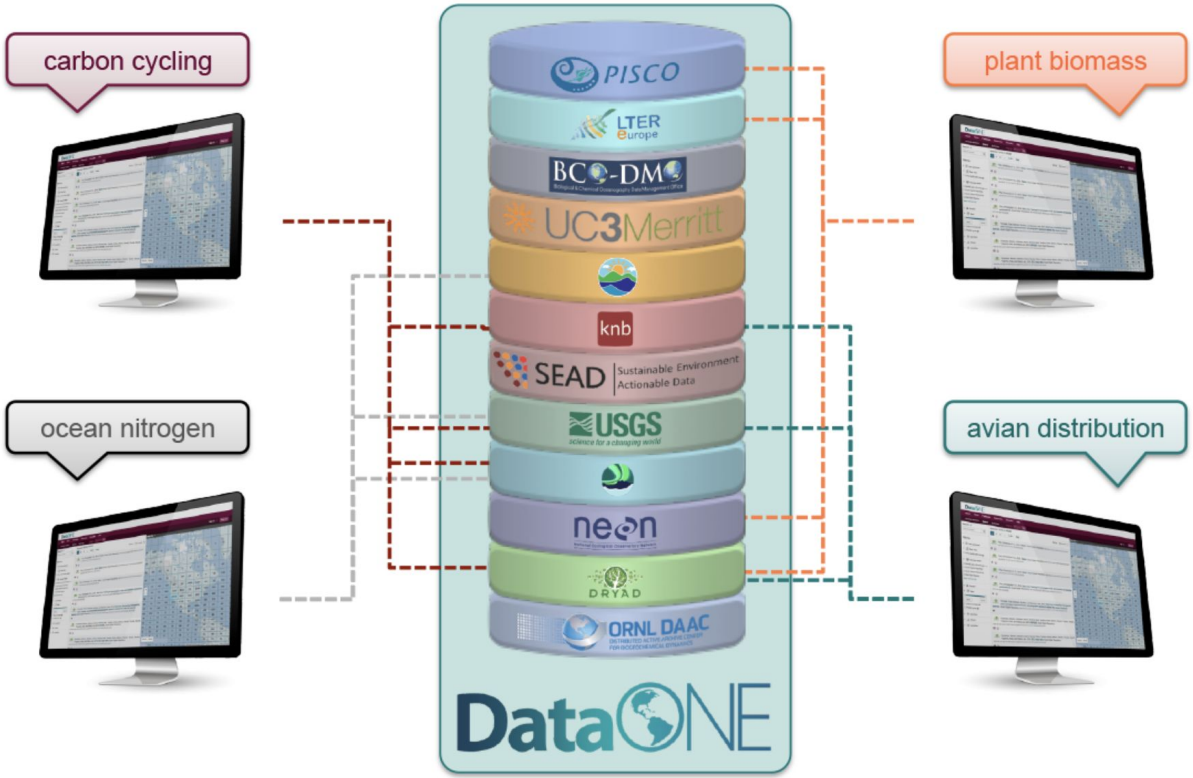


Data Repositories: Build for data and code

- *GitHub is not an archival location*
- Dedicated data repositories: KNB, Arctic Data Center, Zenodo, FigShare
  a. Rich metadata
  b. Archival in their mission
- Data papers, e.g., Scientific Data

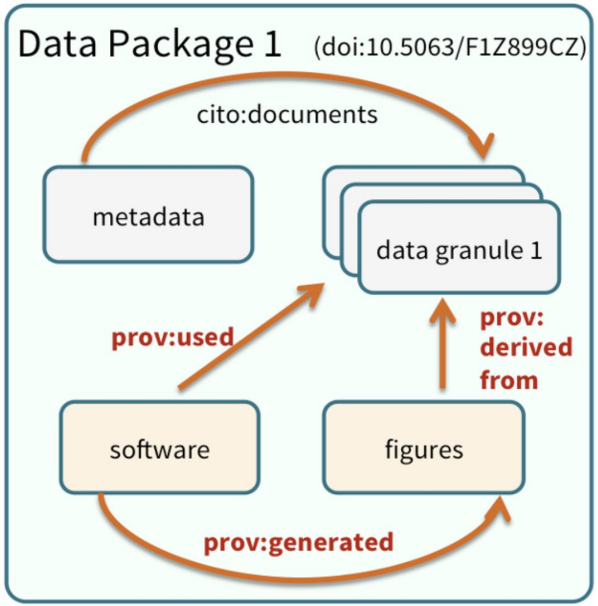


DataONE


carbon cycling
ocean nitrogen
plant biomass
avian distribution
PISCO
LTER Europe
BCO-DMO
UC3Merritt
knb
SEAD
Sustainable Environment Actionable Data
USGS
neon
DRYAD
ORNL DAAC
DataONE

Data Package

- *Data package* as a scientifically useful collection of data and metadata that a researcher wants to preserve.



Publishing Data from the Web

- *Go to book.*