Metadata - Making Data Understandable

Week 3
Data Science Workshop for
NGA LTER REU Students

Review of Last Week

- Good data organization is the foundation of any research project.

- Assume you are going to mess up.
  - Never modify raw data
  - Keep records of every step

Goals for Today

- What is Metadata?
- Why is Metadata useful?
- How might you create and use Metadata?
  - Emphasis on machine readable formats
- Strategies

What is Metadata?

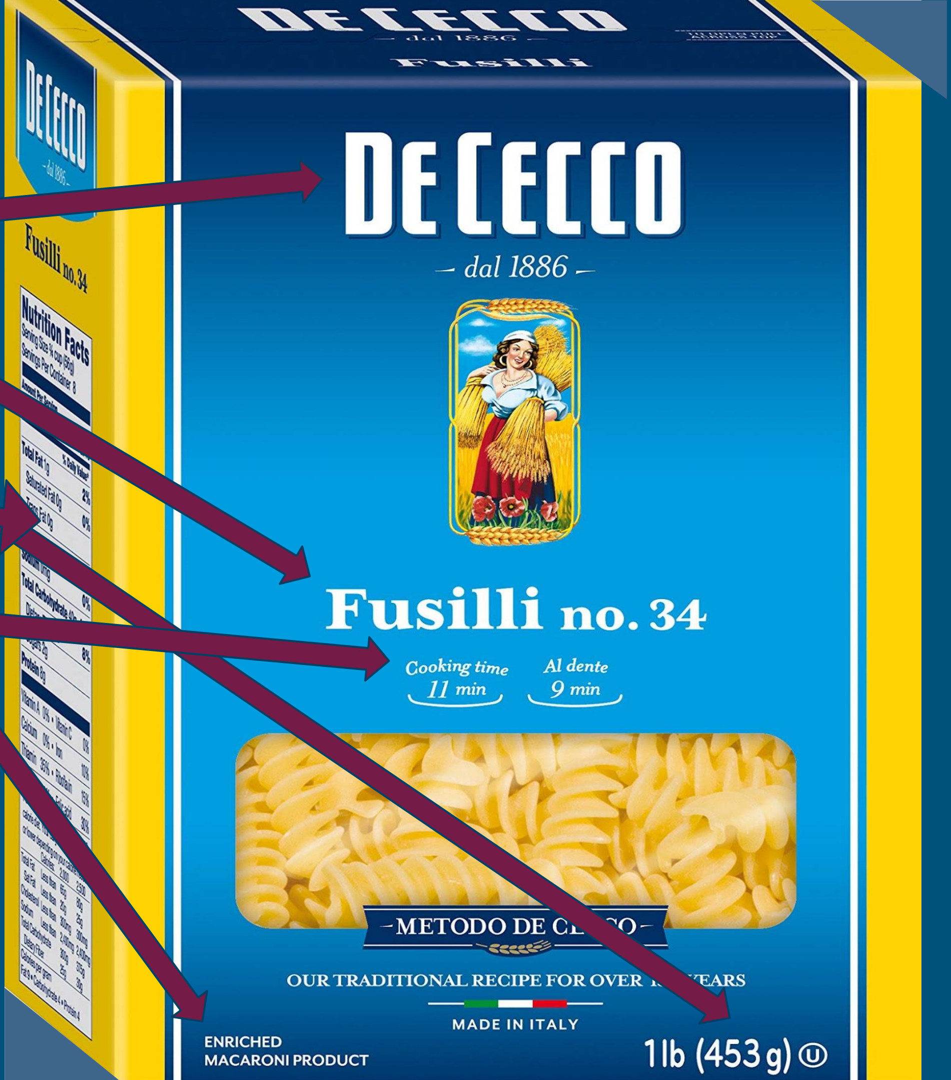
Metadata

“Data about data”

Information used to describe other resources for purposes of re-use, but also for discovery, identification, and management.

The (data) product

- Name of thing
- Creator
- Amount
- Description
- Contents
- Usage



Scientific Metadata

Who: is responsible for the dataset?

What: is the content of the dataset?

Where: does the dataset describe?

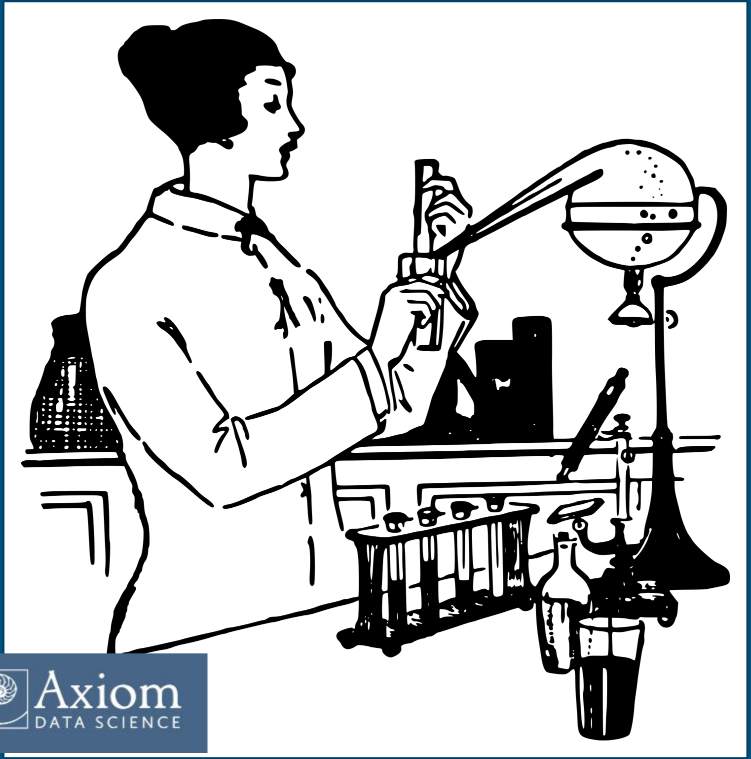
When: does the dataset describe?

How: was the dataset created and evaluated?

Why: this dataset and these methods?

Why Bother?

Value for Data Creators



- Publicize work, get credit
- Communicate quality and limitations of data
- Promote collaboration and synthesis
- Preserve your memory of the data and processes

https://www.goodfreephotos.com/vector-images/lady-scientist-doing-experiment-vector-clipart.png.php

Benefits for Data Users

- Avoid duplication of effort
- Save time deciphering dataset structure and content
- Understand quality and limitations
- Find collaborators

https://commons.wikimedia.org/wiki/File:201705_Scientist_desk_F.svg

Axiom DATA SCIENCE

Metadata Snafu: Act 3

The story so far:

Dr. Judy Benign is trying to use a colleague's data. Three times, she's returned to get information on how to use the data. They are both getting exasperated.

Hanson, Karen; Surkis, Alisa; Yacobucci, Karen: Data Sharing and Management Snafu in 3 Short Acts. https://doi.org/10.5446/31036

https://www.youtube.com/watch?time_continue=5&v=-MIH8PkuUo4

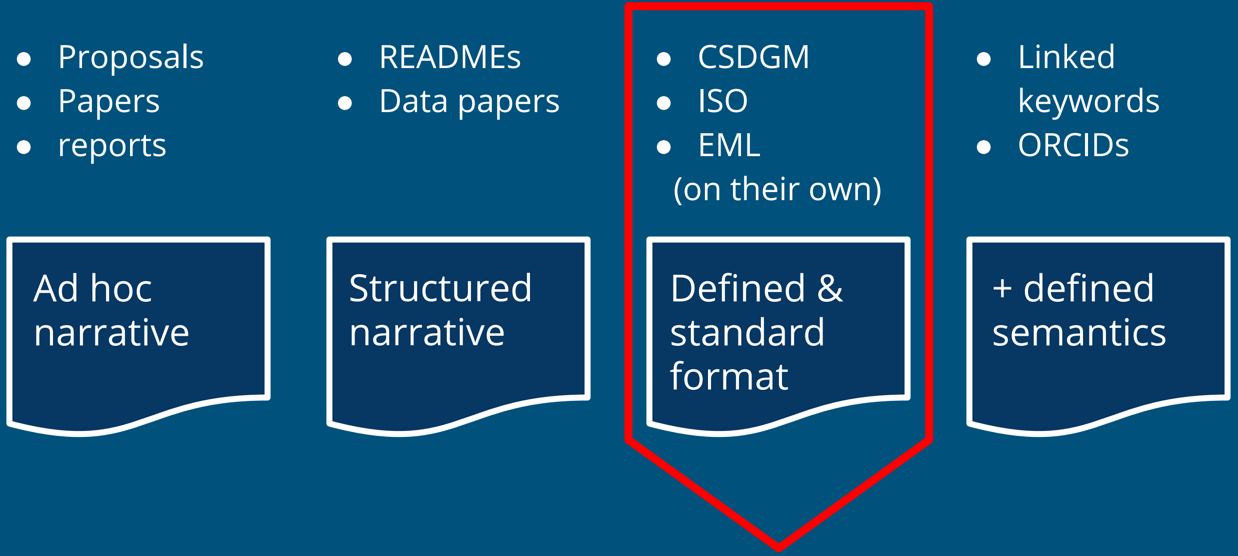
Metadata
is a love note
to the future

What Actually Goes into Metadata?

Jamboard!

https://jamboard.google.com/d/144zev4wCsL9iiGNker0fKyjgMMbk4iNgMrWotb3yS6U/edit?usp=sharing

How to Organize All Those Elements?



- Proposals
- Papers
- reports

Ad hoc narrative

- READMEs
- Data papers

Structured narrative

- CSDGM
- ISO
- EML

(on their own)

Defined & standard format

- Linked keywords
- ORCIDs

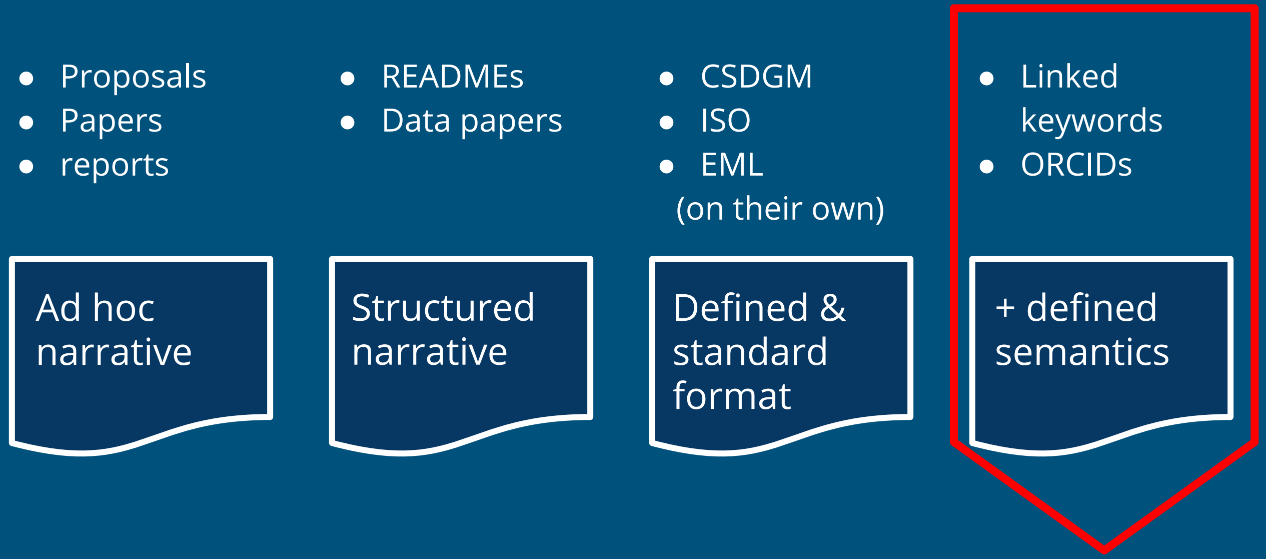
+ defined semantics

Both Human and Machine Readable

Example of XML: INSDC

```
<INSDSet>
 <INSDSeq>
  <INSDSeq_locus>KU141605</INSDSeq_locus>
  <INSDSeq_length>535</INSDSeq_length>
  <INSDSeq_strandedness>double</INSDSeq_strandedness>
  <INSDSeq_moltype>DNA</INSDSeq_moltype>
  <INSDSeq_topology>linear</INSDSeq_topology>
  <INSDSeq_division>INV</INSDSeq_division>
  <INSDSeq_update-date>24-MAR-2020</INSDSeq_update-date>
  <INSDSeq_create-date>09-MAY-2016</INSDSeq_create-date>
  <INSDSeq_definition>
   Pseudocalanus minutus isolate IBPS120 cytochrome c oxidase subunit I (COI) gene,
   partial cds; mitochondrial
  </INSDSeq_definition>
  ...
```

How to Organize All Those Elements?



- Proposals
- Papers
- reports

Ad hoc narrative

- READMEs
- Data papers

Structured narrative

- CSDGM
- ISO
- EML

(on their own)

Defined & standard format

- Linked keywords
- ORCIDs

+ defined semantics

Semantics = meanings of words

Another Metadata Example: ISO 19115

```
<?xml version="1.0" encoding="UTF-8"?>
<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd"
xmlns:gco="http://www.isotc211.org/2005/gco"  ... >
...
  <gmd:identificationInfo>
    <gmd:MD_DataIdentification>
      <gmd:citation>
        <gmd:CI_Citation>
          <gmd:title>
            <gco:CharacterString>CT
Northern Gulf of Alaska, 1970-20
          </gmd:title>
```

```
<xs:complexType name="CI_Citation_Type">
  <xs:annotation>
    <xs:documentation>Standardized resource
    reference</xs:documentation>
  </xs:annotation>
  <xs:complexContent>
    <xs:extension base="gco:AbstractObject_Type">
      <xs:sequence>
        <xs:element name="title"
        type="gco:CharacterString_PropertyType"/>
        ...
<xs:element name="CI_Citation" type="gmd:CI_Citation_Type"/>
```

Machine Readable?

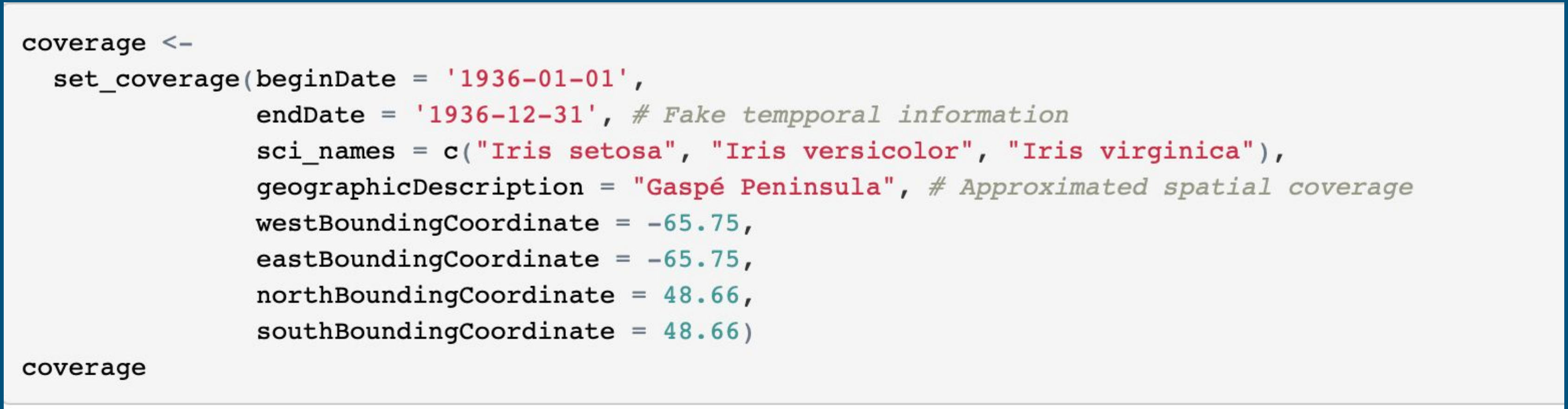
Advantages to Machine Readable

- Tools to create
  - Fillable web forms
  - Error checking
  - Code that adds tags automatically
- Tools to display
  - Data Catalogs
  - Highlight organization
- Tools for discovery
  - Identifiable
  - Searchable

Tools to Edit: Code that Reads and Writes

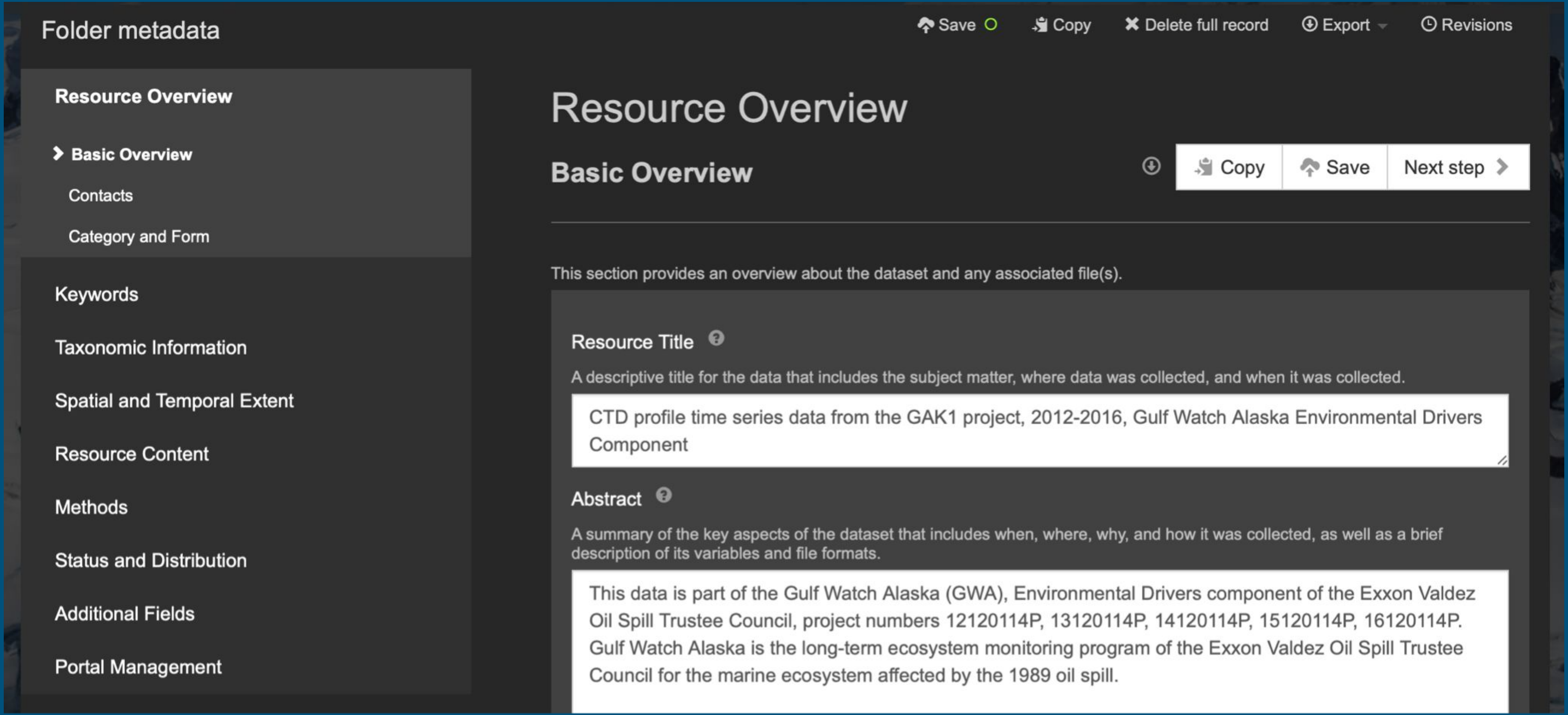
R EML package:

https://www.rdocumentation.org/packages/EML/versions/2.0.5



```
coverage <-
  set_coverage(beginDate = '1936-01-01',
               endDate = '1936-12-31', # Fake tempporal information
               sci_names = c("Iris setosa", "Iris versicolor", "Iris virginica"),
               geographicDescription = "Gaspé Peninsula", # Approximated spatial coverage
               westBoundingCoordinate = -65.75,
               eastBoundingCoordinate = -65.75,
               northBoundingCoordinate = 48.66,
               southBoundingCoordinate = 48.66)
coverage
```

Tools to Edit: Fillable Web Forms



Resource Overview

Basic Overview

This section provides an overview about the dataset and any associated file(s).

Resource Title

A descriptive title for the data that includes the subject matter, where data was collected, and when it was collected.

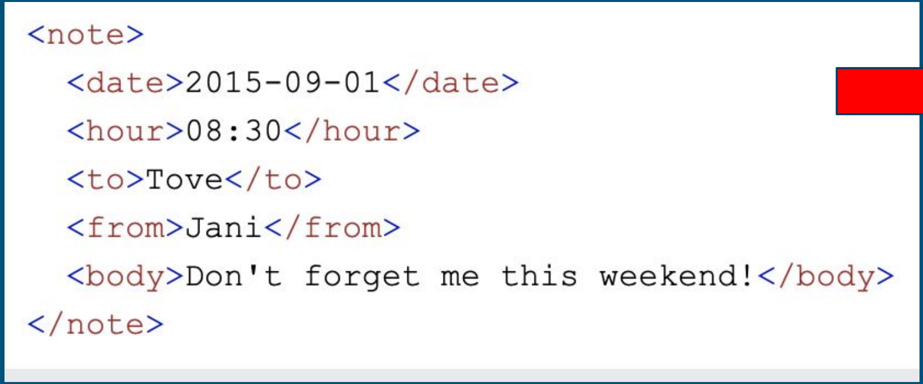
CTD profile time series data from the GAK1 project, 2012-2016, Gulf Watch Alaska Environmental Drivers Component

Abstract

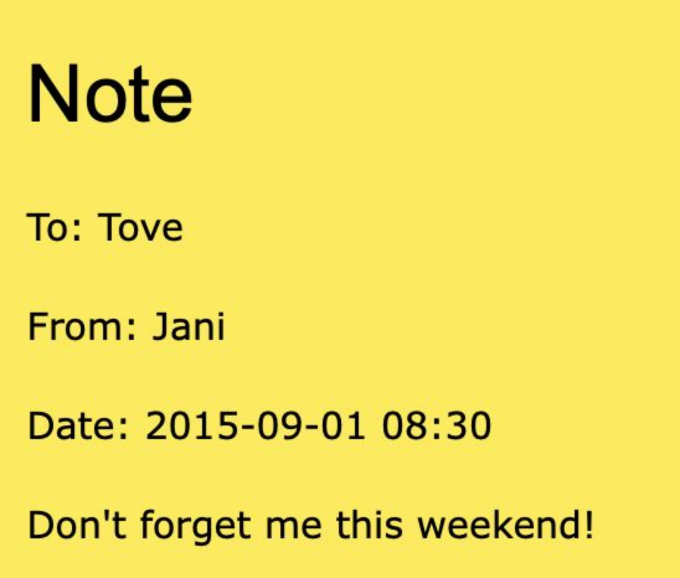
A summary of the key aspects of the dataset that includes when, where, why, and how it was collected, as well as a brief description of its variables and file formats.

This data is part of the Gulf Watch Alaska (GWA), Environmental Drivers component of the Exxon Valdez Oil Spill Trustee Council, project numbers 12120114P, 13120114P, 14120114P, 15120114P, 16120114P. Gulf Watch Alaska is the long-term ecosystem monitoring program of the Exxon Valdez Oil Spill Trustee Council for the marine ecosystem affected by the 1989 oil spill.

Tools to Display: XML to Formatted



```
<note>
  <date>2015-09-01</date>
  <hour>08:30</hour>
  <to>Tove</to>
  <from>Jani</from>
  <body>Don't forget me this weekend!</body>
</note>
```



Note

To: Tove

From: Jani

Date: 2015-09-01 08:30

Don't forget me this weekend!

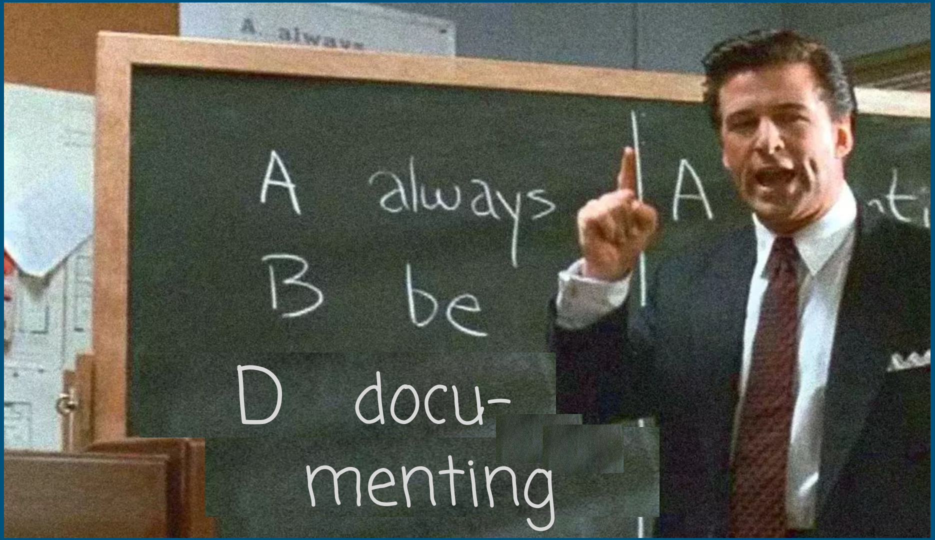
Tools to Display: NGA LTER Data Catalog

https://search.dataone.org/view/10.24431%2Frw1k595

Strategies

Strategies

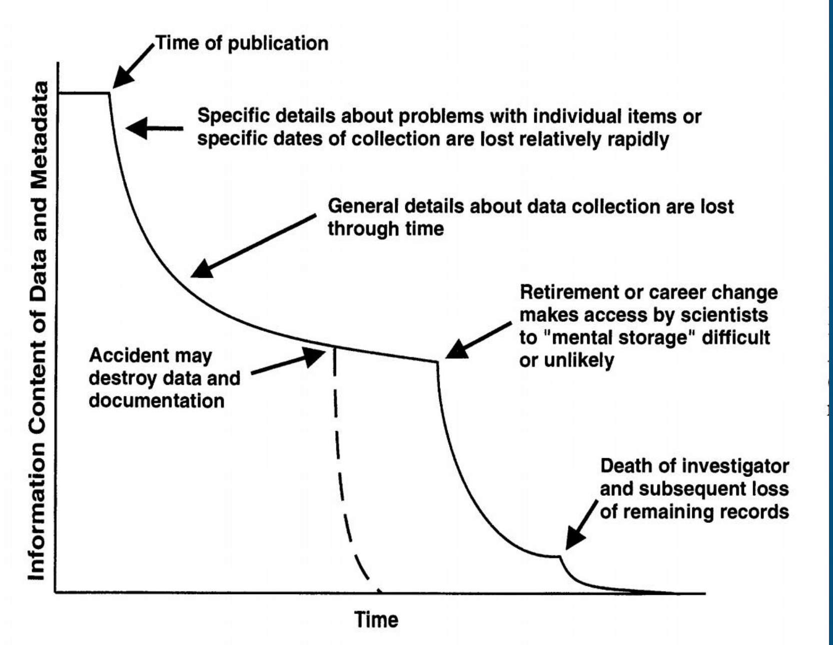
1. **Get organized and make a plan.**
   a. Gather docs from planning, collection & processing.
   b. Be clear about what the data is/are.



Axiom DATA SCIENCE

Strategies

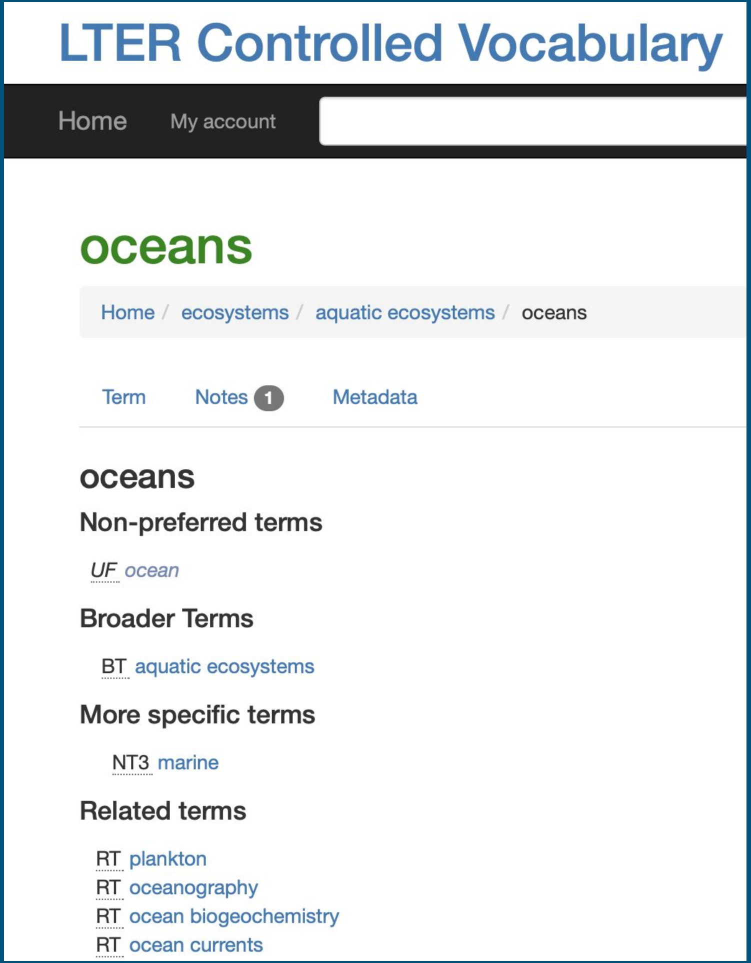
1. Get organized and make a plan.
2. **Get started as soon as possible.**
   a. Write metadata early and often.
   b. Write for humans and machines.



Michener, et al. (1997). Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications* 7(1):330-342
http://dx.doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Strategies

1. Get organized and make a plan.
2. Get started as soon as possible.
3. **Use controlled vocabularies.**



LTER Controlled Vocabulary

oceans

Home / ecosystems / aquatic ecosystems / oceans

oceans

Non-preferred terms

UF ocean

Broader Terms

BT aquatic ecosystems

More specific terms

NT3 marine

Related terms

RT plankton
RT oceanography
RT ocean biogeochemistry
RT ocean currents

Axiom DATA SCIENCE

Strategies

1. Get organized and make a plan.
2. Get started as soon as possible.
3. Use controlled vocabularies.
4. **Treat it like an important part of your science.**
   1. Plan to revise & review before you publish.
   2. Have someone else read your record.

Conclusions

1. Metadata is the backbone of data discovery

2. Think about Metadata from the beginning of your research

3. Take advantage of:

   - Tools
   - Code
   - People

4. Conform to standards

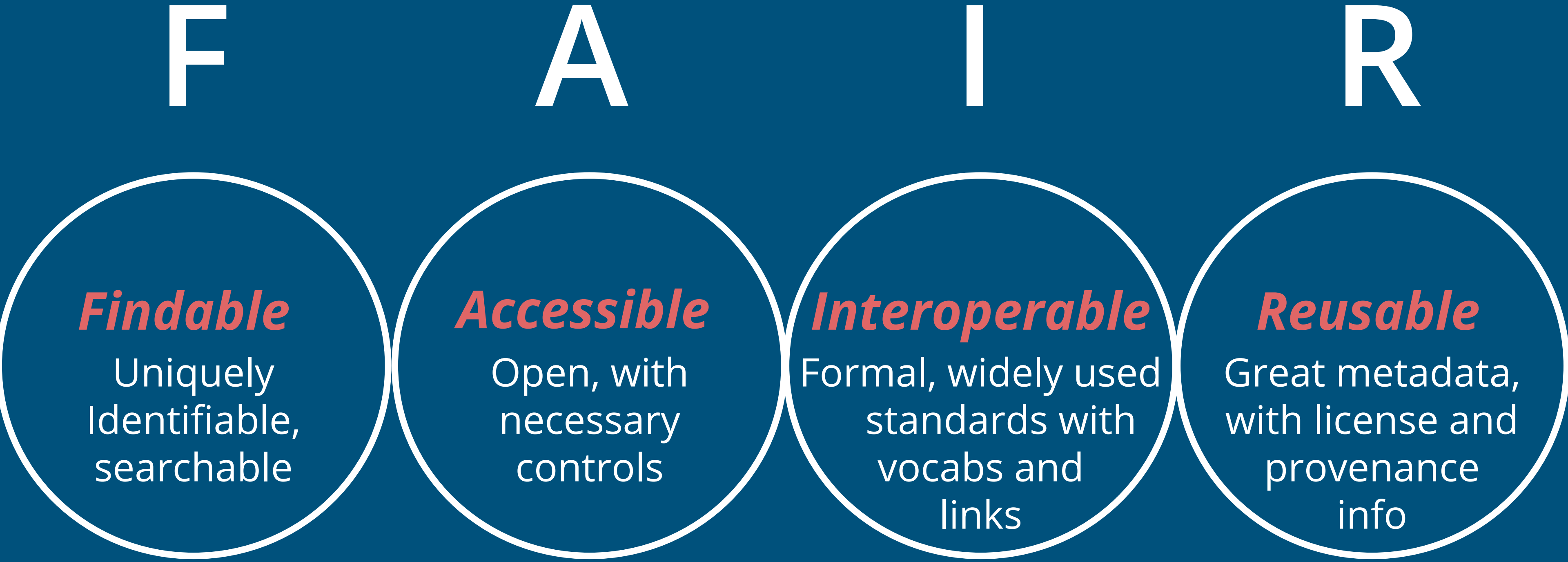
Finish-up

- Assignment #3 on GitHub - Presentations on Data Archives
- Next week we will be at a this Zoom Link again

Resources

- Hanson, Karen; Surkis, Alisa; Yacobucci, Karen: Data Sharing and Management Snafu in 3 Short Acts. https://doi.org/10.5446/31036
- RDA Metadata Standards Directory Working Group
- Understanding Metadata, National Information Standards Organization (NISO)
- FAIRsharing Searchable Table of Standards
- Marine Metadata Interoperability Project Semantic Web Services
- LTER Controlled Vocabulary
- NGA LTER DataONE Data Catalog

Reproducibility Requires FAIR (meta)data



F

Findable

Uniquely Identifiable, searchable

A

Accessible

Open, with necessary controls

I

Interoperable

Formal, widely used standards with vocabs and links

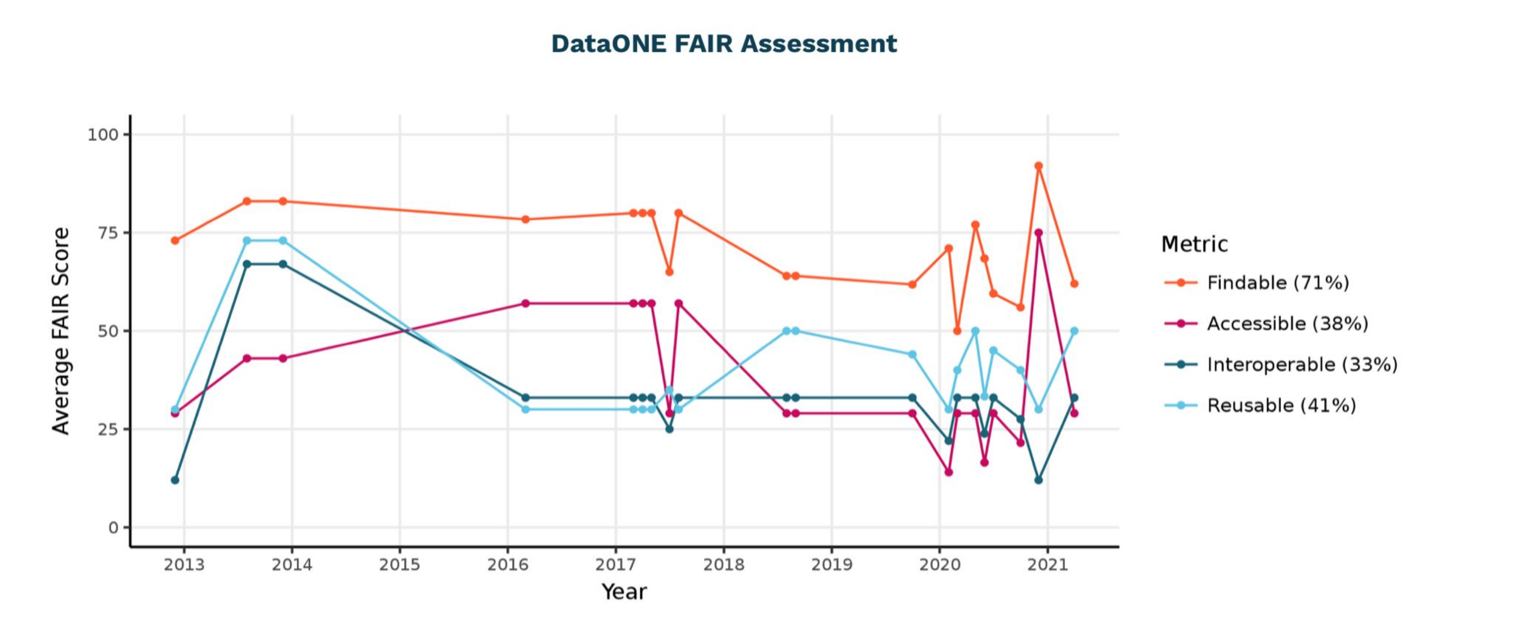
R

Reusable

Great metadata, with license and provenance info

https://www.go-fair.org/fair-principles/

Tools to Edit: Compliance Checkers



Discovery

https://search.dataone.org/portals/NGALTER/Data