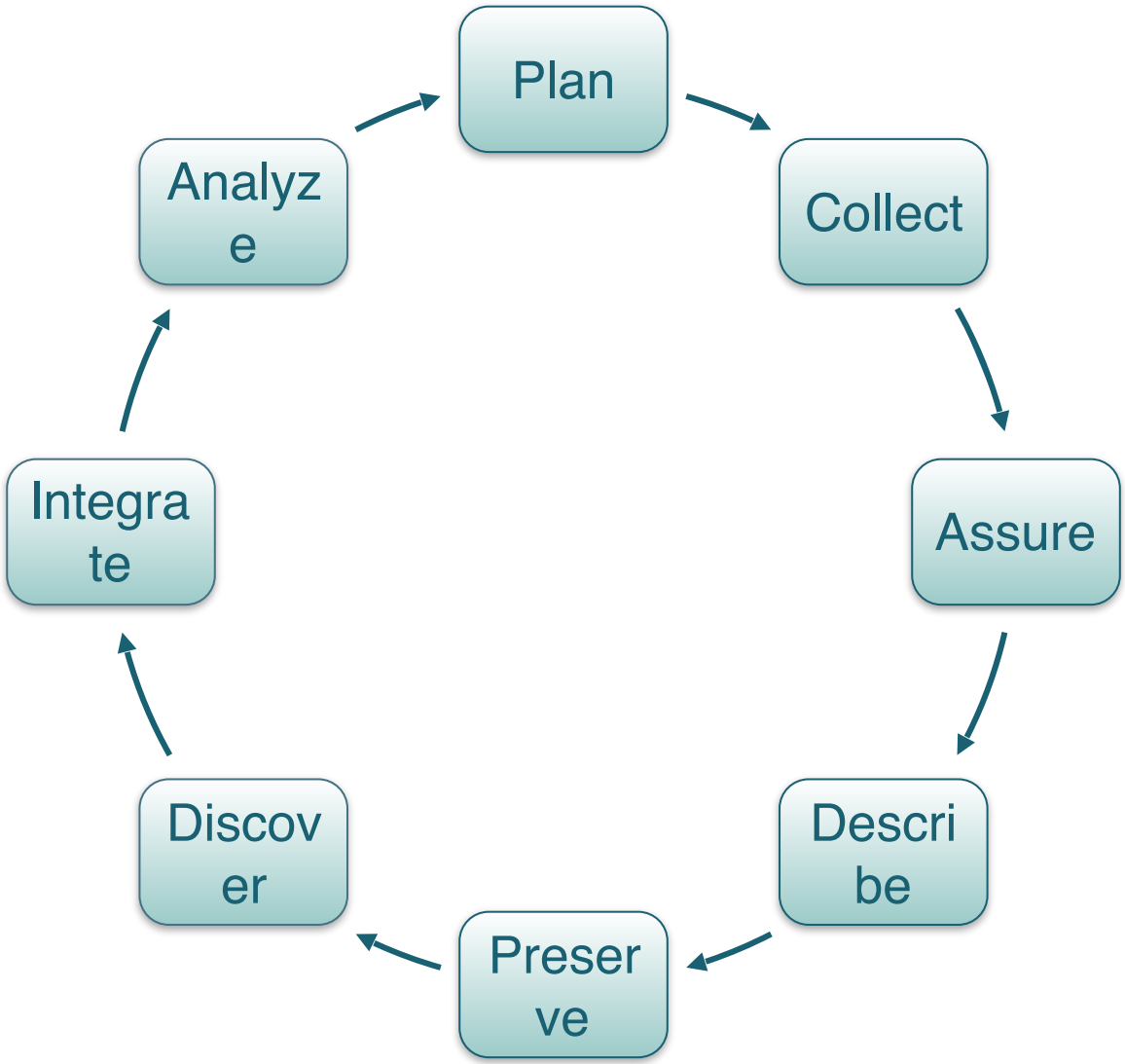
Authoring High Quality Metadata

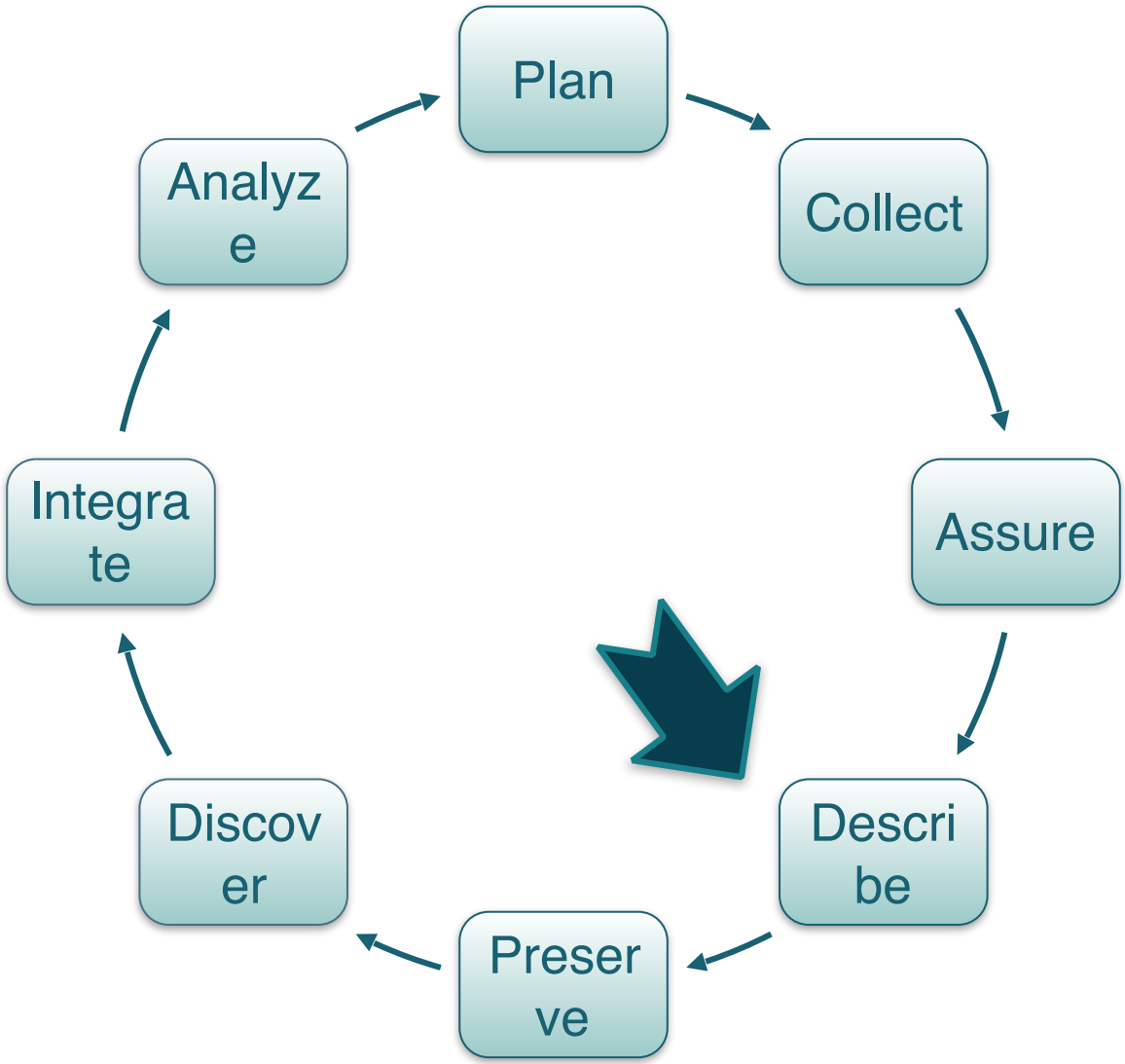
Jesse Goldstein and Jeanette Clark
UC Santa Barbara

JG orcid.org/0000-0002-1006-9496
JC orcid.org/0000-0003-4703-1974

The Data Life Cycle



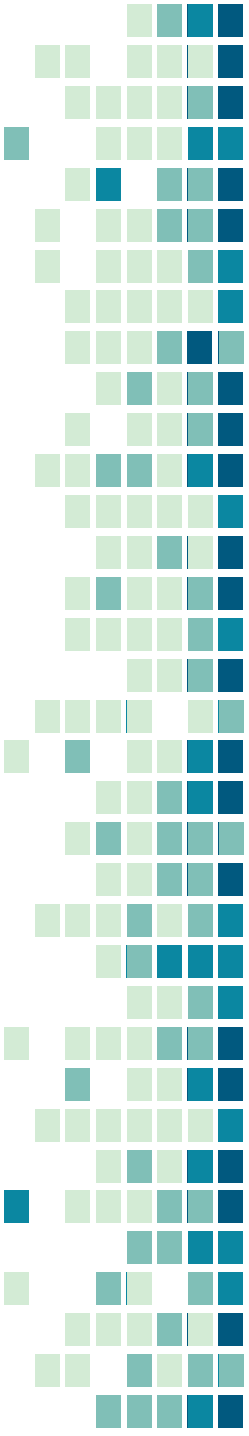
The Data Life Cycle



What are metadata?

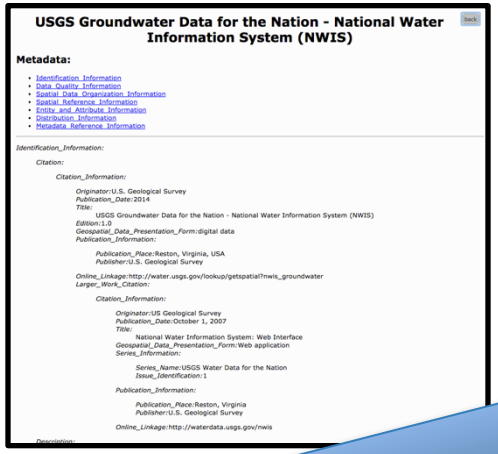
Think of metadata as “data reporting”

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

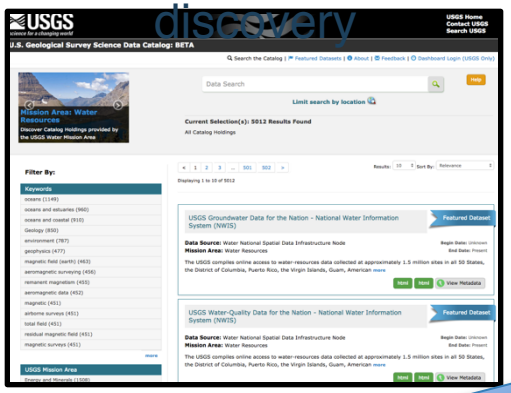


Why are metadata important?

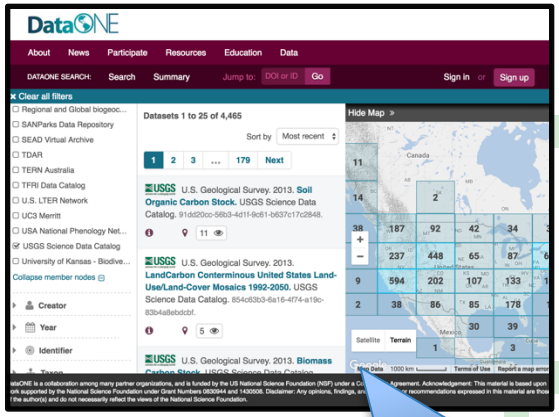
Metadata capture information



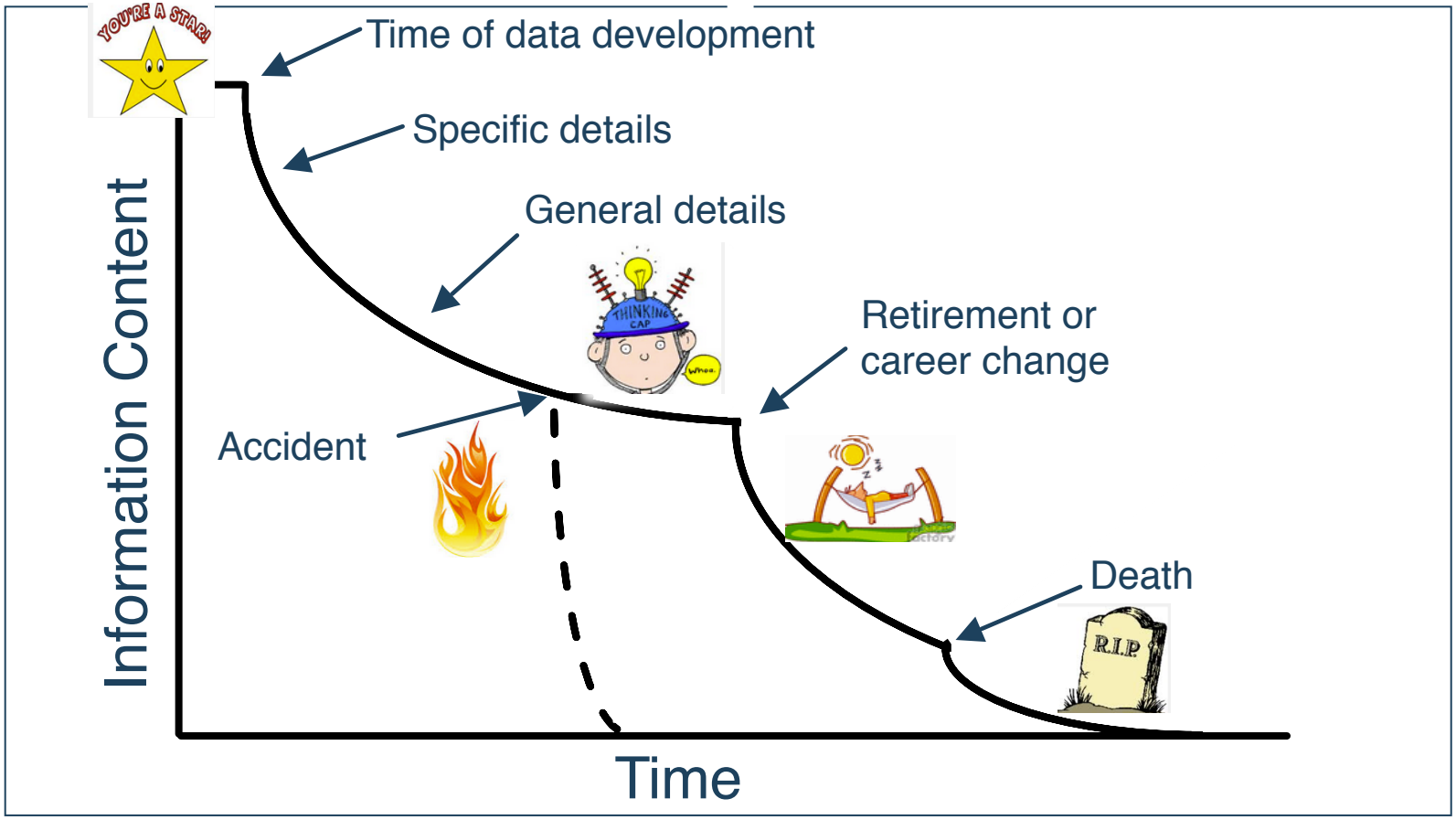
USGS Science Data Catalog: enables discovery



DataONE: enables



Why are metadata important?



(modified from Michener et al. 1997)

Why are metadata important?

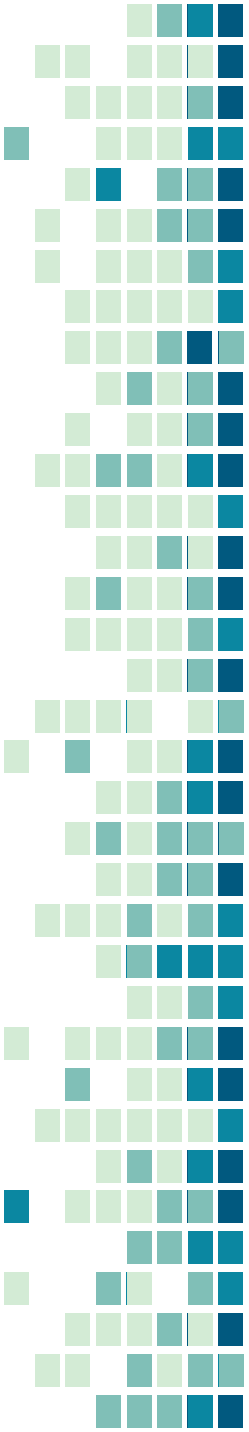
Metadata are important for the short and long-term utility of data

Why are metadata important?

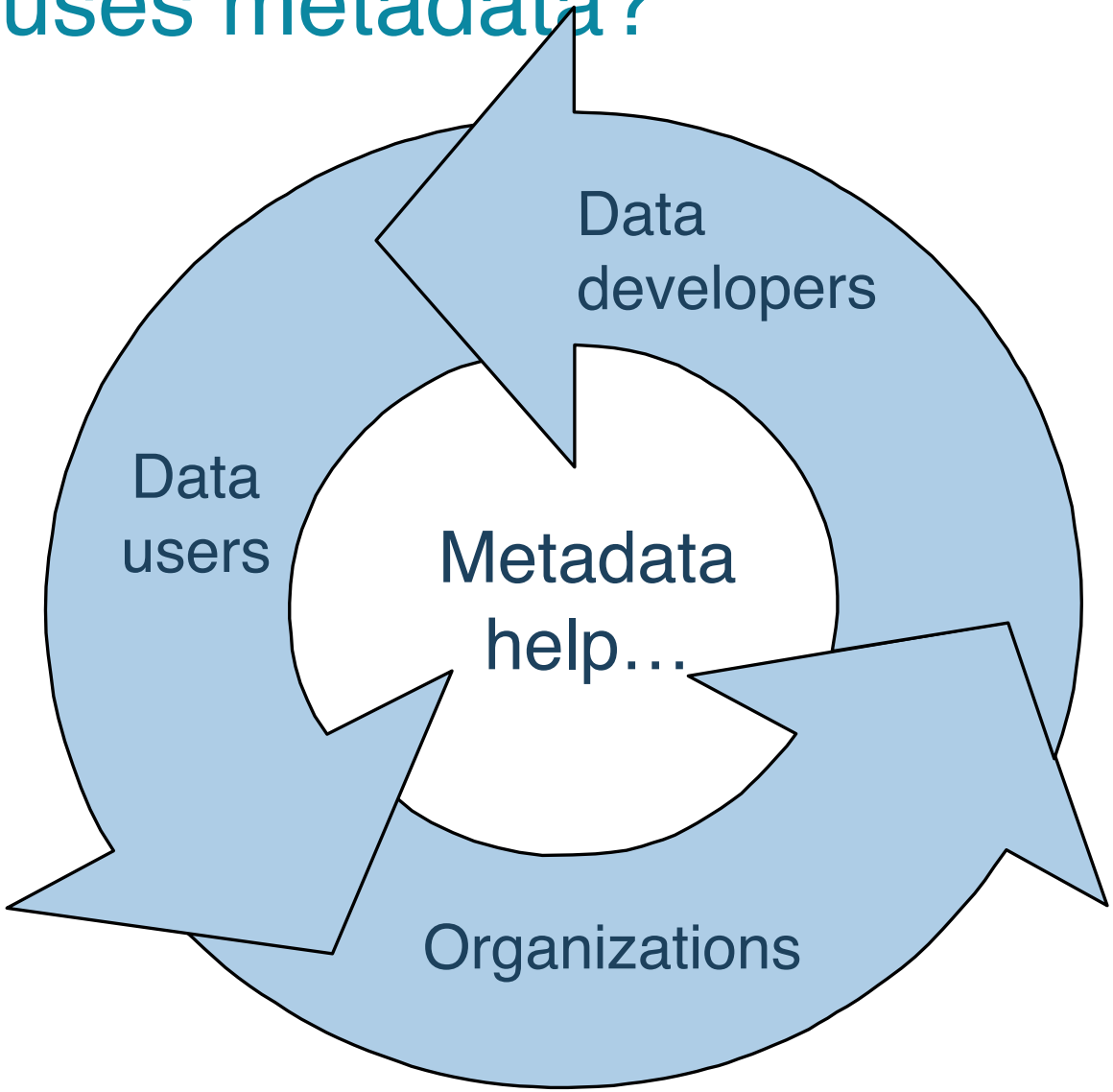
Why are metadata important?

Metadata are essential for policy work

- Data discovery to help answer policy questions
- Metadata support scrutiny of data
  - Motivations
  - Methodologies
  - Conflicts of interest

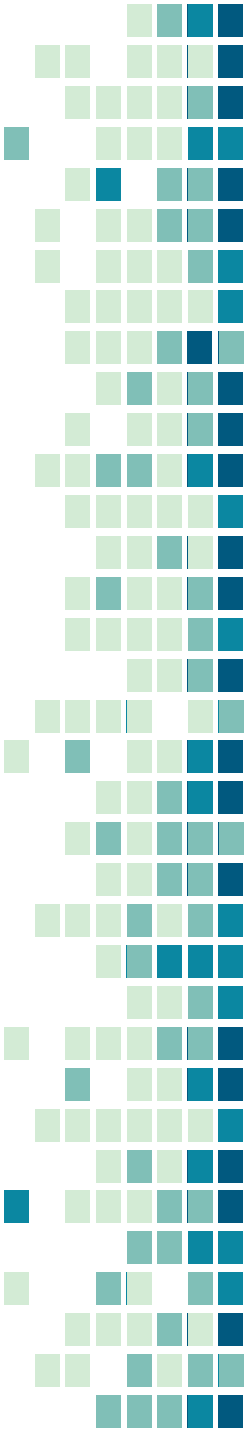


Who uses metadata?



Metadata for data developers

- Avoid data duplication
  - What data have already been collected?
  - Save time the next time
  - “Hey, I’ve already done this!”
- Share reliable information
  - What methods were used?
  - What methods are in common use in my field?
- Publicize your work
  - “Hey, I made this!”



Metadata for data users

- Find relevant data
- Evaluate what is suitable for use in your work
- Retrieve the data you've found
- Understand if and how to actually use the data

Metadata for organizations

- Help ensure the organization's investment in the data
  - Documentation for sampling and data processing methods are recorded
  - Ability to use data after initial intended purpose
  - Track data re-use and citation
- Transcend people and time
  - Data are not lost when researchers or labs leave
  - Avoid duplication in new work
- Advertise organization's research
  - What data has our organization produced?

Concerns about creating metadata

Even if the value of data documentation is recognized, researchers are often concerned about the effort required to create metadata that effectively describe their data.

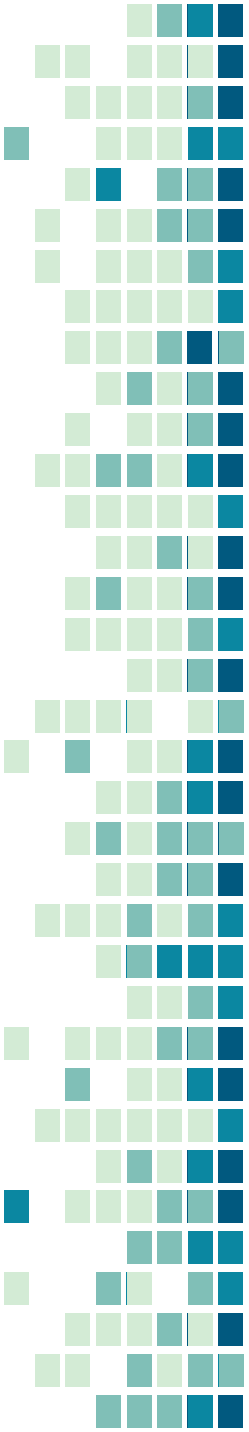
Concerns about creating metadata

Concern	Solution
Workload required to capture accurate robust metadata	Incorporate metadata creation into data development process – distribute the effort
Time and resources to create, manage, and maintain metadata	Include in grant budget and schedule
Readability / usability of metadata	Use a standardized metadata format
Discipline specific information and ontologies	Use a standard ‘profile’ that supports discipline specific information

Metadata standards

A metadata standard provides a uniform structure to describe data:

- Machine readable (usually XML)
- Common terminology
- Common structure



Metadata standards

Example standards:

- Dublin Core (emphasis on publications)
- Darwin Core (emphasis on collections)
- FGDC (emphasis on spatial data)
- ISO19115 (emphasis on spatial data and services)
- Ecological Metadata Language (general, but emphasis on filesystem artifacts, attributes, taxonomy)

Metadata standards

```
<?xml version="1.0" encoding="UTF-8"?>

<gmi:MI_Metadata xmlns:gmi="http://www.isotc211.org/2005/gmi" xmlns:gco="http://www.isotc211
  <gmd:fileIdentifier gco:nilReason="missing"/>
  <gmd:language>
    <gco:CharacterString>eng;USA</gco:CharacterString>
  </gmd:language>
  <gmd:characterSet>
    <gmd:MD_CharacterSetCode codeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schem
  </gmd:characterSet>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>Axiom Data Science</gco:CharacterString>
      </gmd:organisationName>
      <gmd:positionName>
        <gco:CharacterString>Metadata Specialist</gco:CharacterString>
      </gmd:positionName>
      <gmd:contactInfo>
        <gmd:CI_Contact>
          <gmd:address>
            <gmd:CI_Address>
              <gmd:deliveryPoint>
                <gco:CharacterString>1016 W 6th Ave, Ste 105</gco:CharacterString>
              </gmd:deliveryPoint>
              <gmd:city>
                <gco:CharacterString>Anchorage</gco:CharacterString>
              </gmd:city>
              <gmd:administrativeArea>
```

Metadata standards

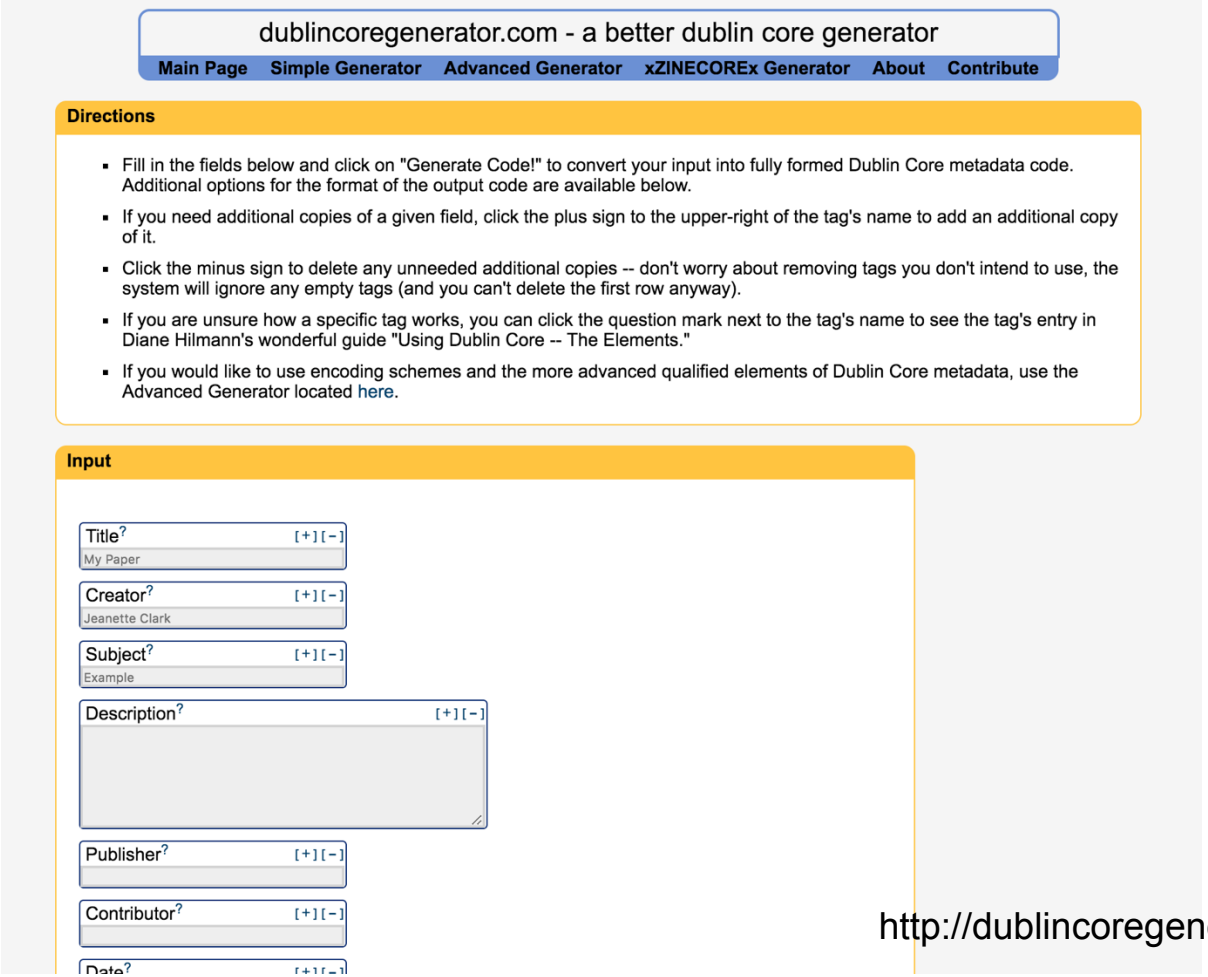
```
<?xml version="1.0" encoding="UTF-8"?>

<gmi:MI_Metadata xmlns:gmi="http://www.isotc211.org/2005/gmi" xmlns:gco="http://www.isotc211
  <gmd:fileIdentifier gco:nilReason="missing"/>
  <gmd:language>
    <gco:CharacterString>eng;USA</gco:CharacterString>
  </gmd:language>
  <gmd:characterSet>
    <gmd:MD_CharacterSetCode codeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schem
  </gmd:characterSet>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>Axiom Data Science</gco:CharacterString>
      </gmd:organisationName>
      <gmd:positionName>
        <gco:CharacterString>Metadata Specialist</gco:CharacterString>
      </gmd:positionName>
      <gmd:contactInfo>
        <gmd:CI_Contact>
          <gmd:address>
            <gmd:CI_Address>
              <gmd:deliveryPoint>
                <gco:CharacterString
              </gmd:deliveryPoint>
              <gmd:city>
                <gco:CharacterString
              </gmd:city>
              <gmd:administrativeAr
```

…is a person that creates and manages metadata for resources and services. This person generally has expertise in documentation standards and has enough experience and understanding of the resource to document it in partnership with the originator or resource contact.

Creating Standardized Metadata

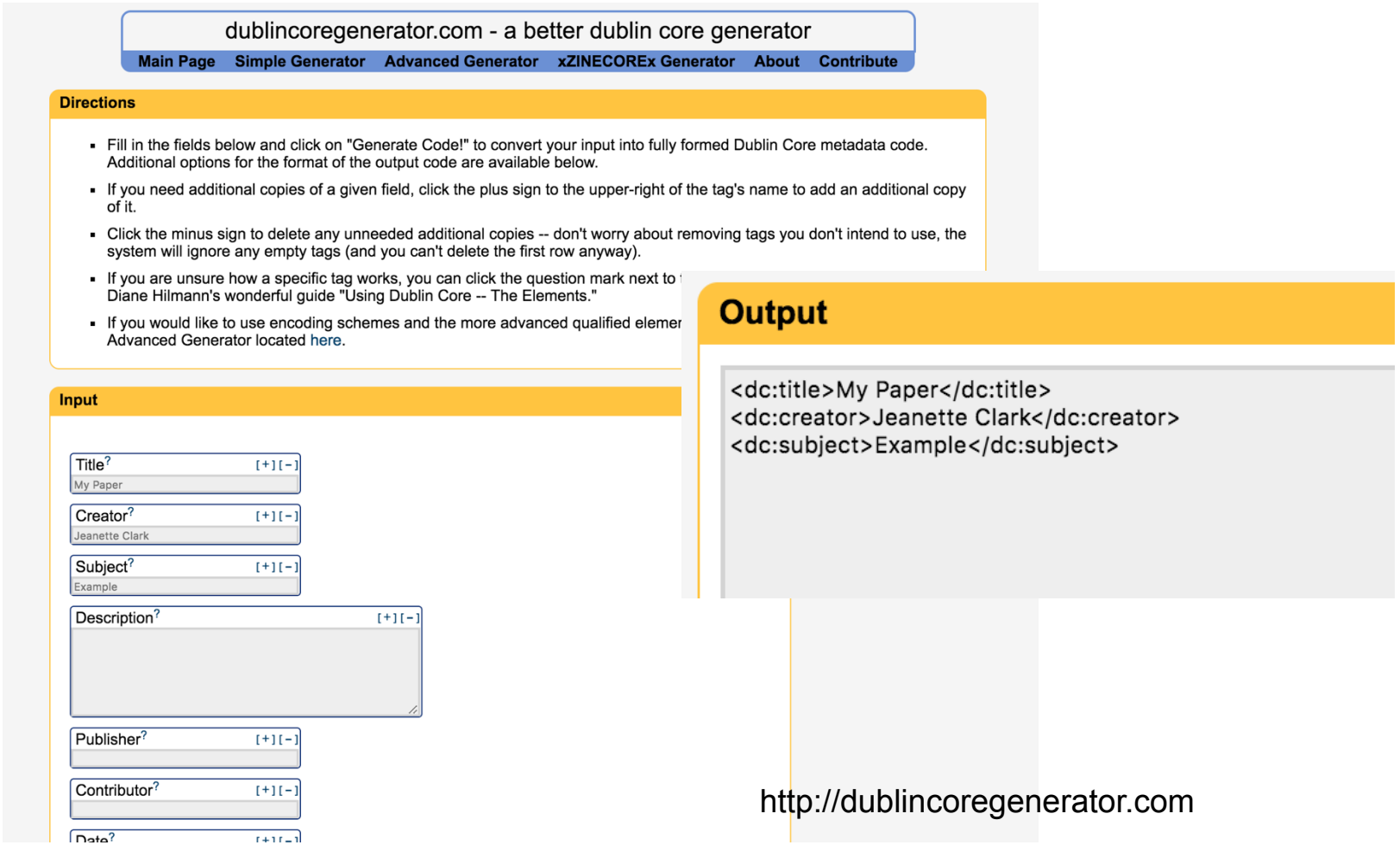
- Specialized tools are your friend!



http://dublincoregenerator.com

Creating Standardized Metadata

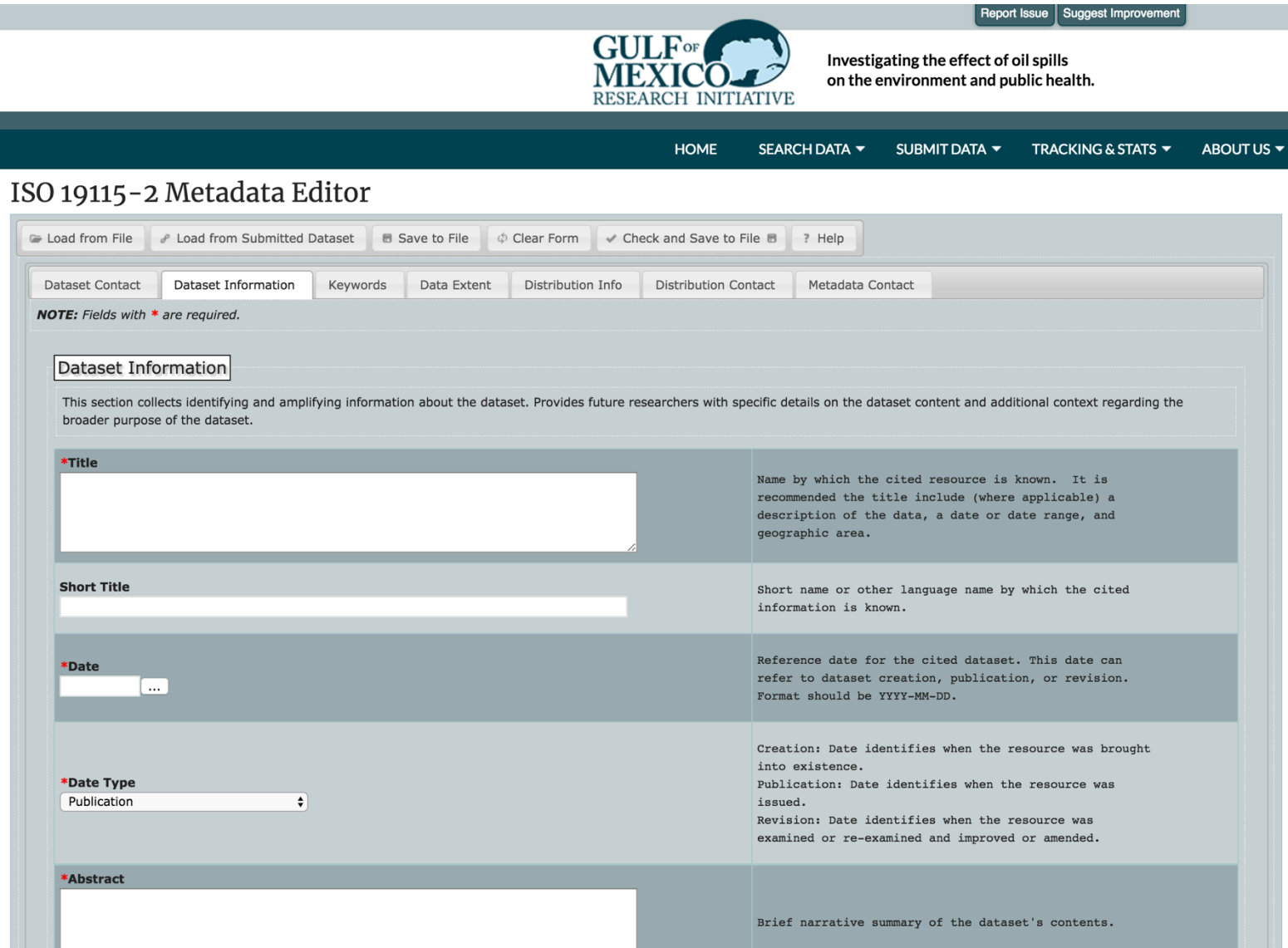
- Specialized tools are your friend!



http://dublincoregenerator.com

Creating Standardized Metadata

https://data.gulfresearchinitiative.org/metadata-editor/



Creating Standardized Metadata

https://github.com/ropensci/EML/

```
attributes2 <- data.frame(attributeName = c('Time', 'Wind_Speed'),
                          attributeDefinition = c('Date and time of wind speed reading', 'Measured
                          measurementScale = c('dateTime', 'ratio'),
                          domain = c('dateTimeDomain', 'numericDomain'),
                          formatString = c('YYYY-MM-DD hh:mm:ss', NA),
                          definition = c(NA, NA),
                          unit = c(NA, 'metersPerSecond'),
                          numberType = c(NA, 'real'),
                          missingValueCode = c(NA, NA),
                          codeExplanation = c(NA, NA),
                          stringsAsFactors = FALSE)
attributeList2 <- set_attributes(attributes2)
id2 <- 'PID2'
physical2 <- pid_to_eml_physical(mn, id2)
dataTable2 <- new('dataTable',
                  entityName = 'EagleMtnWindData.csv',
                  entityDescription = 'Wind data from Eagle Mountain',
                  physical = physical2,
                  attributeList = attributeList2)
```

Creating Standardized Metadata

https://github.com/ropensci/EML/

```
attributes2 <- data.frame(attributeName = c('Time', 'Wind_Speed'),
                          attributeDefinition = c('Date and time of wind speed reading', 'Measured
                          measurementScale = c('dateTime', 'ratio'),
                          domain = c
                          formatStri
                          definition
                          unit = c(NA
                          numberType
                          missingVal
                          codeExplan
                          stringsAsF
attributeList2 <- set_attributes(att
id2 <- 'PID2'
physical2 <- pid_to_eml_physical(mn,
dataTable2 <- new('dataTable',
                  entityName = 'Eagl
                  entityDescription
                  physical = physica
                  attributeList = at
```

```
> attributeList2
<attributeList>
  <attribute>
    <attributeName>Time</attributeName>
    <attributeDefinition>Date and time of wind speed reading</attributeDefinition>
    <measurementScale>
      <dateTime>
        <formatString>YYYY-MM-DD hh:mm:ss</formatString>
      </dateTime>
    </measurementScale>
  </attribute>
  <attribute>
    <attributeName>Wind_Speed</attributeName>
    <attributeDefinition>Measured wind speed</attributeDefinition>
    <measurementScale>
      <ratio>
        <unit>
          <standardUnit>metersPerSecond</standardUnit>
        </unit>
        <numericDomain>
          <numberType>real</numberType>
        </numericDomain>
      </ratio>
    </measurementScale>
  </attribute>
</attributeList>
```

What makes a good metadata record?

Overall goal: Could a reasonable scientist make sense of your data in 10, 20, 30+ years without contacting you?

When in doubt, be more specific:

- Spell out acronyms
- Use full names, emails, addresses, etc.

Include as much information as possible directly in the metadata record

What makes a good metadata record?

Target multiple user groups:

- Someone looking directly for your data
- Someone who does not know about your work but should
- Someone looking to scrutinize your work
- Someone trying to reproduce your work
- Someone looking to give you credit for your work

What makes a good metadata record?

Good titles include:

- What
- When
- Where

The title is often the first way a user will evaluate your data set

What makes a good metadata record?

Title:

“ITP37”

What makes a good metadata record?

Title:

“ITP37” 😕 ⚠️

What makes a good metadata record?

Title:

"Ocean water property observations reported from ice-tethered profiler #37, Transpolar Drift, 2009"

What makes a good metadata record?

Title:

"Ocean water property observations reported from ice-tethered profiler #37, Transpolar Drift, 2009"

😃 ✅

What makes a good metadata record?

Begin: 2003-04-14
End: 2003-04-13

Sag River

😕 ⚠️

What makes a good metadata record?

What makes a good metadata record?

"Begin: 2002-04-14
End: 2003-04-13"

"Sagavanirktok River, North Slope, Alaska"

What makes a good metadata record?

“Begin: 2002-04-14
End: 2003-04-13”

“Sagavanirktok River, North Slope, Alaska”

😃 ✅

What makes a good metadata record?

“ ”

What makes a good metadata record?

“ ”

What makes a good metadata record?

Transect

We established three 100-m transects at the Airport Site to quantify differences in micro-topography, soil temperatures, thaw depth, soils, vegetation, permafrost and snow in relationship to distance from the road. Pin flags were placed at 1-m intervals along Transects 3 and 4, and vertical 150-cm PVC posts were placed at 0, 5, 10, 25, 50 and 100 m. The poles have blue stripes at 50, 100 and 150 cm height to help locate the transects in winter. No poles or pin flags were placed along T5, but the plots are permanently marked by wooden corner stakes and an aluminum-capped piece of rebar at the center bearing the plot number.

Vegetation Plots

We established permanent vegetation plots with photo points in polygon centers and troughs at 5, 10, 25, 50 and 100 m from the road along T3 and T4, and at 25, 50 and 100 m from the road on T5. Voucher collections of all vascular plants, mosses and lichens were collected from each plot and are stored at the Alaska Geobotany Center. Species cover was measured using 100 points from a 1 x 1 m2 point-quadrat. Cover of all species was estimated using Braun-Blanquet cover abundance scores. The species at the top of the plant canopy were recorded at 100 grid points within each plot. Leaf Area Index (LAI) was measured using an AccuPAR LP-80 PAR/LAI Ceptometer. Soil temperature loggers were installed at all permanent plots on T3 and T4. Air temperature loggers were installed along the T3 and T4 flag transects.

Topographic Surveys

The location and elevation of all boreholes, transects, vegetation plots and other reference points were

What makes a good metadata record?

Transect

We established three 100-m transects at the Airport Site to quantify differences in micro-topography, soil temperatures, thaw depth, soils, vegetation, permafrost and snow in relationship to distance from the road. Pin flags were placed at 1-m intervals along Transects 3 and 4, and vertical 150-cm PVC posts were placed at 0, 5, 10, 25, 50 and 100 m. The poles have blue stripes at 50, 100 and 150 cm height to help locate the transects in winter. No poles or pin flags were placed along T5, but the plots are permanently marked by wooden corner stakes and an aluminum-capped piece of rebar at the center bearing the plot number.

Vegetation Plots

We established permanent vegetation plots with photo points in polygon centers and troughs at 5, 10, 25, 50 and 100 m from the road along T3 and T4, and at 25, 50 and 100 m from the road on T5. Voucher collections of all vascular plants, mosses and lichens were collected from each plot and are stored at the Alaska Geobotany Center. Species cover was measured using 100 points from a 1 x 1 m2 point-quadrat. Cover of all species was estimated using Braun-Blanquet cover abundance scores. The species at the top of the plant canopy were recorded at 100 grid points within each plot. Leaf Area Index (LAI) was measured using an AccuPAR LP-80 PAR/LAI Ceptometer. Soil temperature loggers were installed at all permanent plots on T3 and T4. Air temperature loggers were installed along the T3 and T4 flag transects.

Topographic Surveys

The location and elevation of all boreholes, transects, vegetation plots and other reference points were

What makes a good metadata record?

Abstract

- Distinct from publication abstract
- Should provide more context for the title
- Should give a high-level summary of methodologies, data formats, coverages, etc.

What makes a good metadata record?

Abstract

- Distinct from publication abstract
- Should provide more context for the title
- Should give a high-level summary of methodologies, data formats, coverages, etc.

These data are ocean water property observations reported from ice-tethered profiler #37.

What makes a good metadata record?

Abstract

- Distinct from publication abstract
- Should provide more context for the title
- Should give a high-level summary of methodologies, data formats, coverages, etc.

These data are ocean water property observations reported from ice-tethered profiler #37. Profiler #37 was deployed near Barrow, Alaska from a research vessel. These data are used to characterize upper ocean dynamics to better understand the underlying conditions for sea ice formation. Included in this dataset are …

😀 ✅

What makes a good metadata record?

Documented filesystem artifacts

- File formats
- File sizes
- Checksums (“Do I have the same file?”)
- Where to download (web address)
- Attributes used (variables)

What makes a good metadata record?

Involved parties

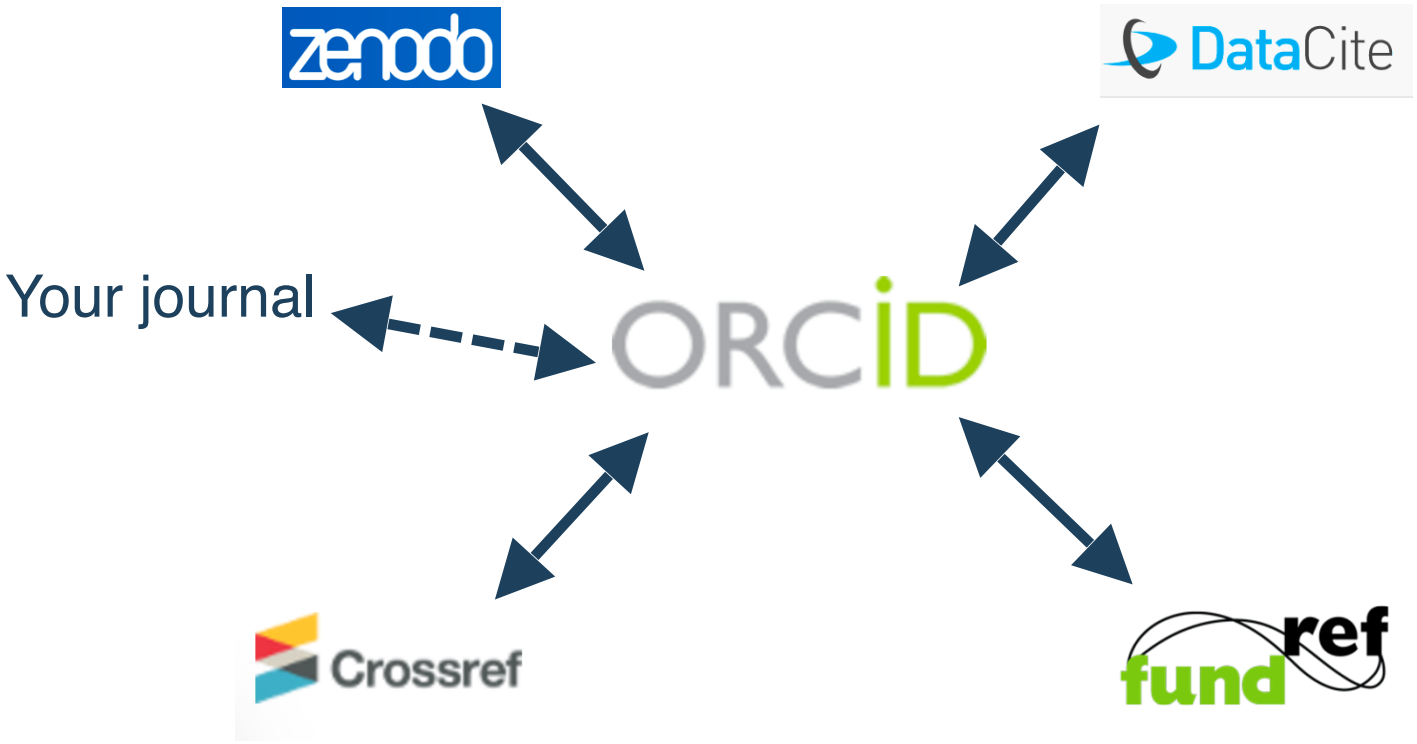
- Name alone is not enough…
  - to assign credit, nor
  - to disambiguate across data sets
- Email addresses help
- Including ORCID iD is best

ORCID iDs: ”Wait, what is an ORCID iD?”

- Like an ISBN for people
  - e.g. mine: orcid.org/0000-0002-1006-9496
- Enables unambiguous reference to humans
- Free
- Becoming a community norm
- Inherently connected…

ORCID iDs

Inherently connected



Activity

Register an ORCID iD:

- orcid.org/register

Sign in to dev.nceas.ucsb.edu/#share