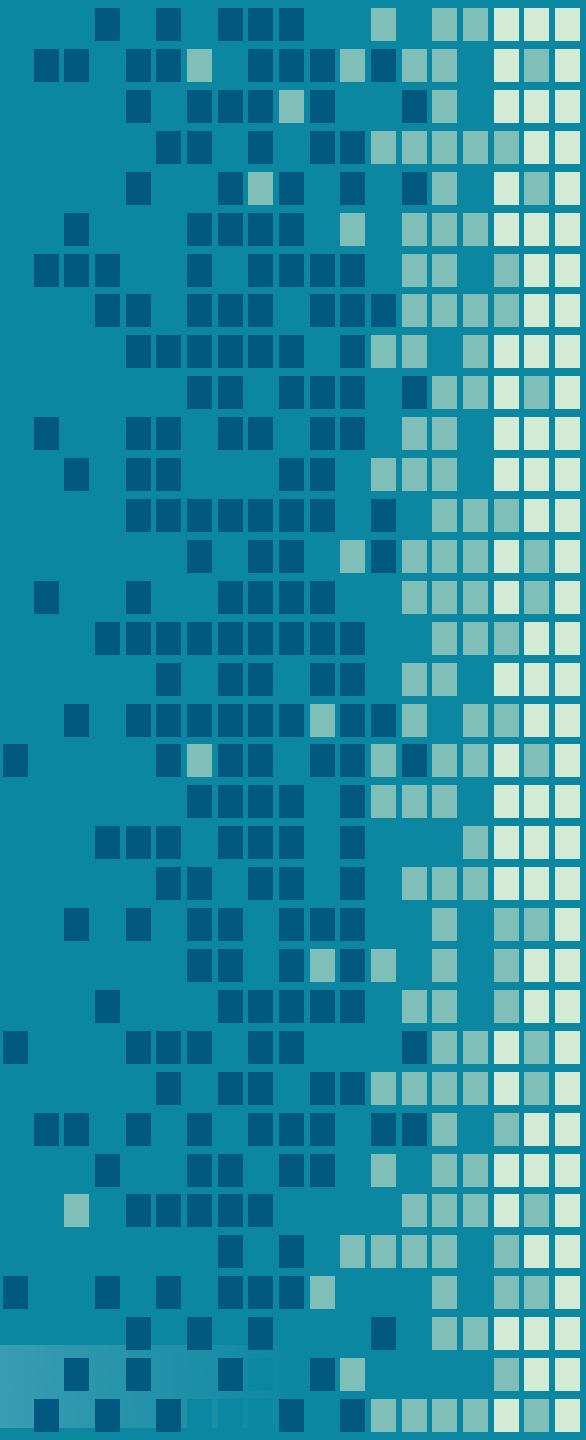


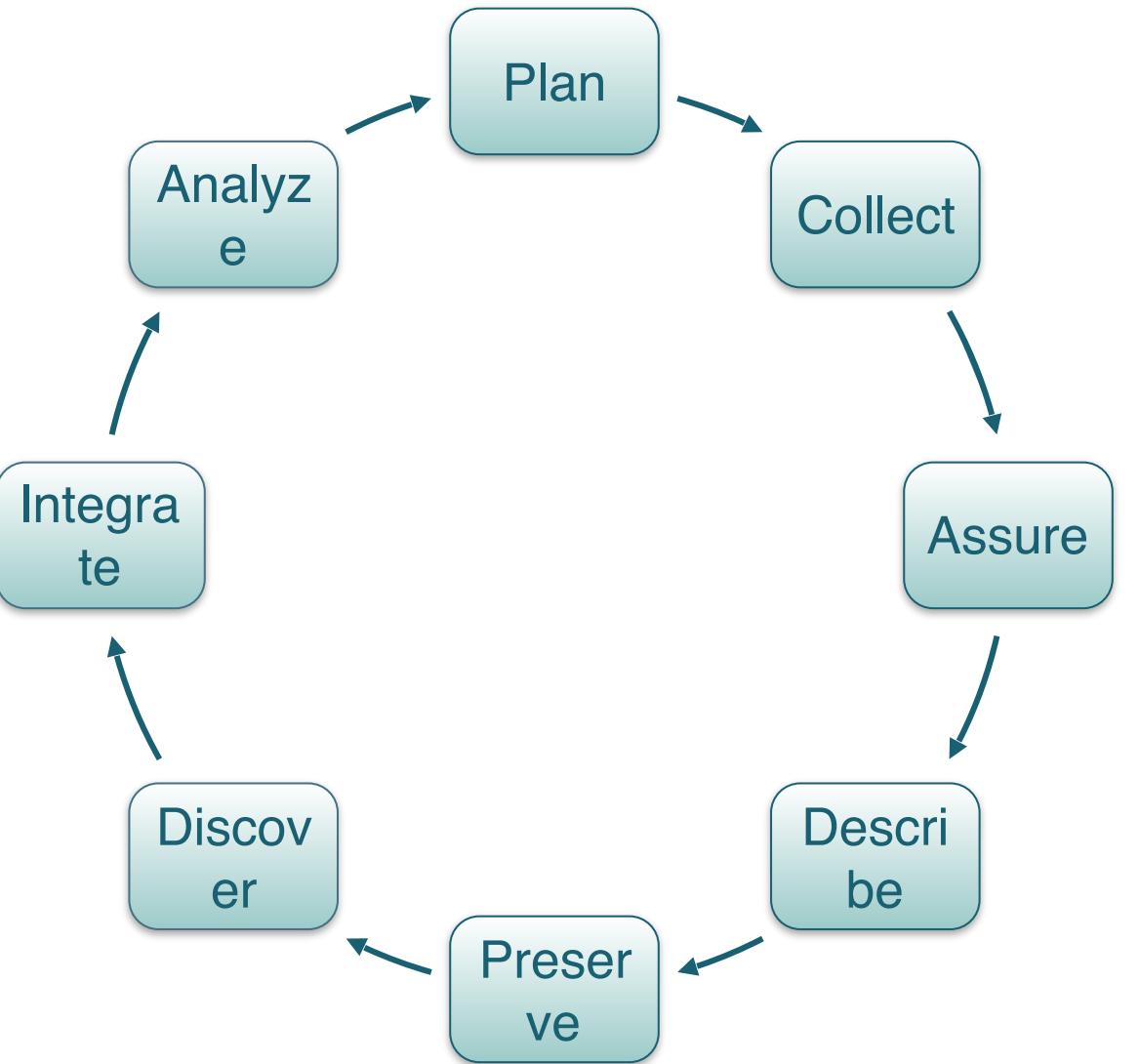
Authoring High Quality Metadata

Jesse Goldstein and Jeanette Clark
UC Santa Barbara

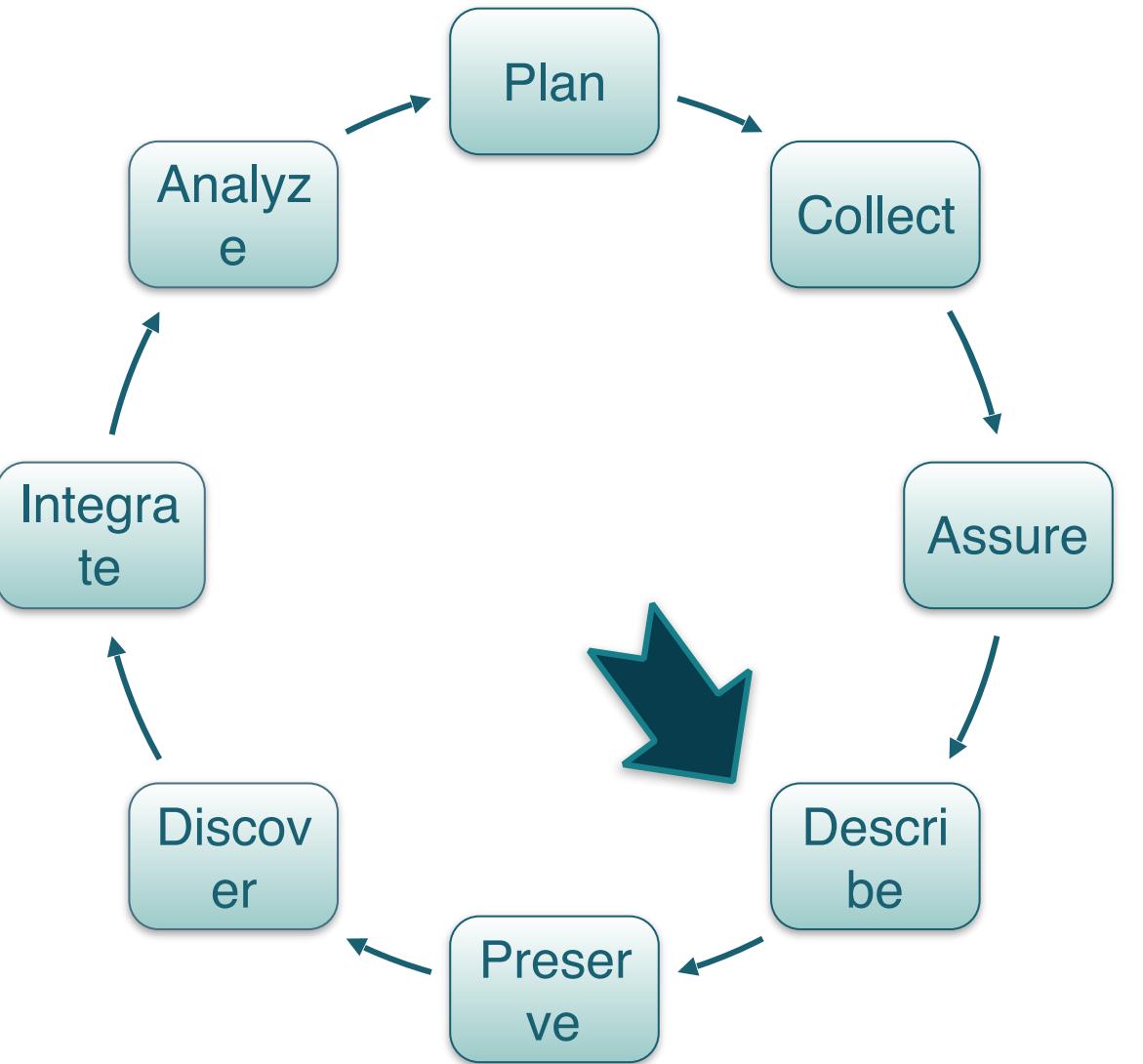
JG orcid.org/0000-0002-1006-9496
JC orcid.org/0000-0003-4703-1974



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 Groundwater Data for the Nation - National Water Information System (NWIS)	
Metadata:	
<ul style="list-style-type: none">• Identification_Information• Data_Organization_Information• Source_Data_Organization_Information• Spatial_Reference_Information• Temporal_Reference_Information• Distribution_Information• Metadata_Reference_Information	
Identification_Information:	
Citation:	
Citation_Information:	
Originator: U.S. Geological Survey	
Publication_Date: 2014	
Title:	
USGS Groundwater Data for the Nation - National Water Information System (NWIS)	
Edition: 1	
Geospatial_Data_Presentation_Form: digital data	
Publication_Information:	
Publisher: Peace River, Virginia, USA	
Publisher_U.S._Geological_Survey:	
Online_Linkage: http://water.usgs.gov/lookup/getspatial/nwis_groundwater	
Larger_Works_Citation:	
Citation_Information:	
Originator: U.S. Geological Survey	
Publication_Date: October 1, 2009	
Title:	
National Water Information System: Web Interface	
Geospatial_Data_Presentation_Form: Web application	
Series_Information:	
Series_Name: USGS Water Data for the Nation	
Issue_Number: 1	
Publication_Information:	
Publisher: Peace River, Virginia	
Publisher_U.S._Geological_Survey:	
Online_Linkage: http://waterdata.usgs.gov/nwis	

JSGS Science Data Catalog: enables

The screenshot shows the USGS Water Resources Data Catalog homepage. The top navigation bar includes links for "USGS Home", "Contact USGS", and "Search USGS". Below the navigation is a search bar with the placeholder "Search the Catalog" and a dropdown menu showing "Featured Datasets" and "About". To the right of the search bar are links for "Feedback" and "Download Log (USGS long)". The main content area features a large search result summary for "Groundwater" with a thumbnail image of a water drop. The summary includes the number of datasets (5012), a "View Details" button, and a "View All" link. Below this is a "Data Search" section with a search bar, a "Limit search by location" button, and a "Search" button. A "Current Selection(s): 5012 Results Found" message is displayed above the catalog holdings table. The catalog holdings table lists various datasets with columns for title, source, date, and type. At the bottom of the page, there are "USGS Water Resources Data Catalog" links and a "View Previous" button.

DataONE: enables

The screenshot shows the DataONE search interface. At the top, there's a navigation bar with links for About, News, Participate, Resources, Education, and Data. Below that is a search bar with dropdowns for 'DATACITE SEARCH:' and 'Search' (with 'Summary' as the current selection), and a 'Jump to: DOI or ID' input field with a 'Go' button. To the right are 'Sign In' and 'Sign up' buttons. A 'Clear all filters' link is also present.

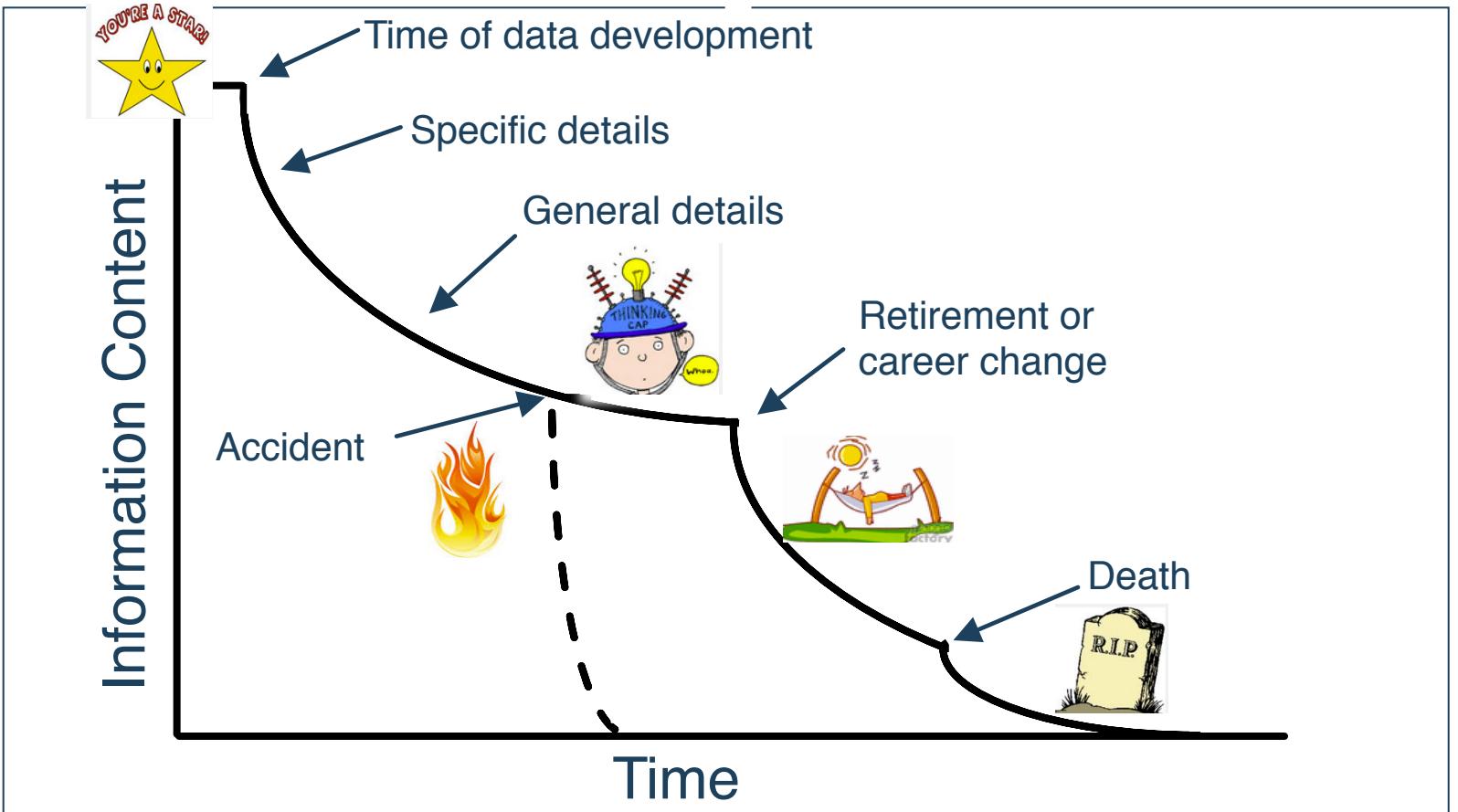
The main content area has a header 'Datasets 1 to 25 of 4,465'. It includes a 'Sort by' dropdown set to 'Most recent' and a 'Next' button. Below this, a grid of dataset cards is shown:

	11	14	2	34
	187	92	42	34
	237	448	65	87
	594	202	107	133
	38	86	85	178

Each card contains a thumbnail, the dataset title, and a 'View Details' button. The first card is for 'U.S. Geological Survey, 2013. Soil Organic Carbon Stock USGS Science Data Catalog. 9162220-2953-4617-9819-1937017c2948.' The second card is for 'U.S. Geological Survey, 2013. Land-Contaminant Cen...'. The third card is for 'U.S. Geological Survey, 2013. Land-Cover/Mosaic 1992-2050. USGS Science Data Catalog. 6540530-6a16-4f74-a18c-83d0a6d60d0'. The fourth card is for 'U.S. Geological Survey, 2013. Biomass Carbon Stock USGS Science Data Catalog.'

To the right of the dataset grid is a 'Hide Map >' button followed by a map of North America with state/province boundaries. A sidebar on the left lists 'Regional and Global biogeoc... SANData Repository SEAD Virtual Archive TDAR TERIN Australia TERRI Data Catalog U.S. TERI Network JCU3 Merritt USA National Phenology Net... USGS Science Data Catalog University of Kansas - Biogeoc... (la)use member nodes'. On the far left, there are 'Creator', 'Year', 'Identifier', and 'Type' filters.

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?

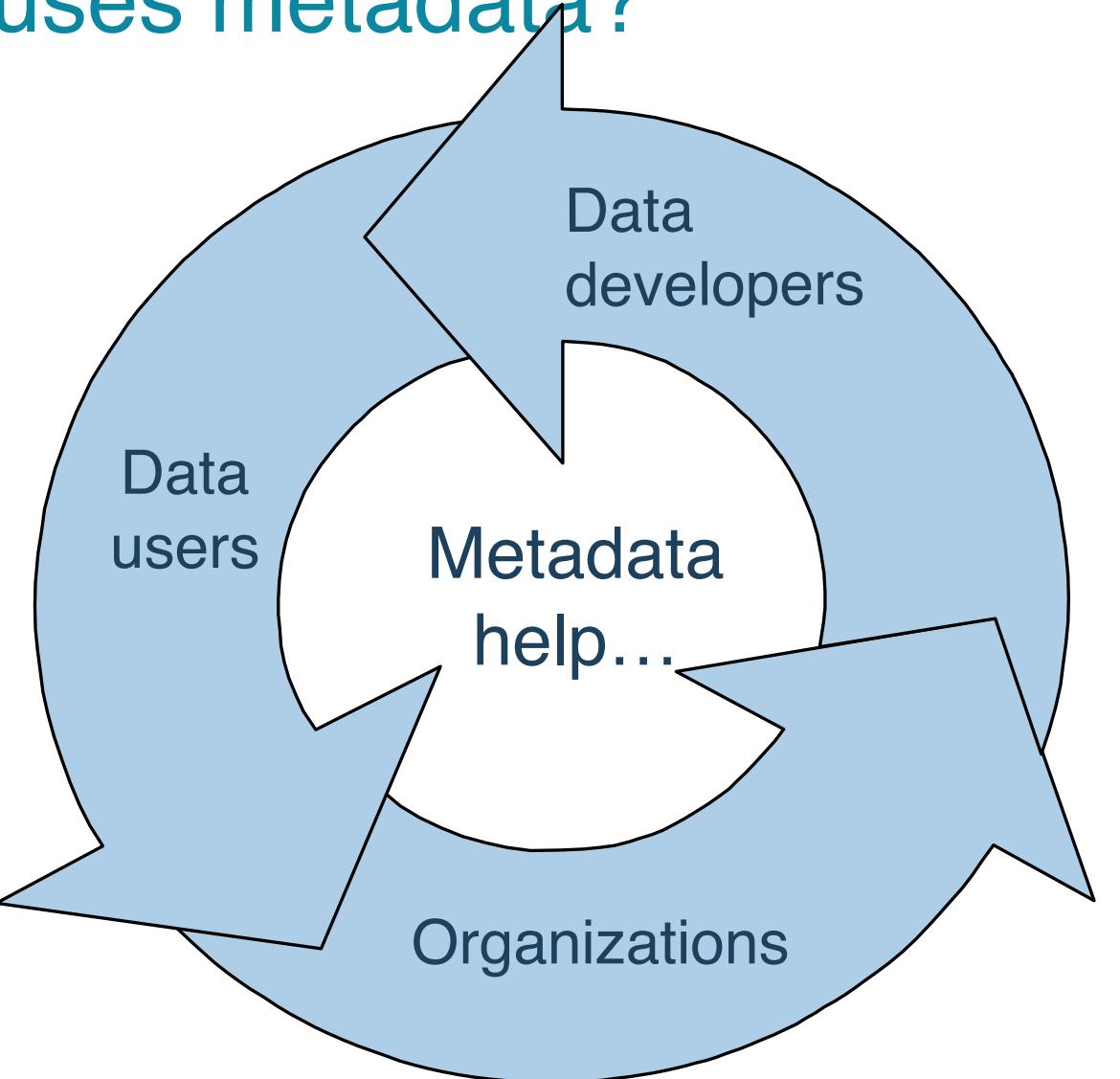


Why are metadata important?

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"?>

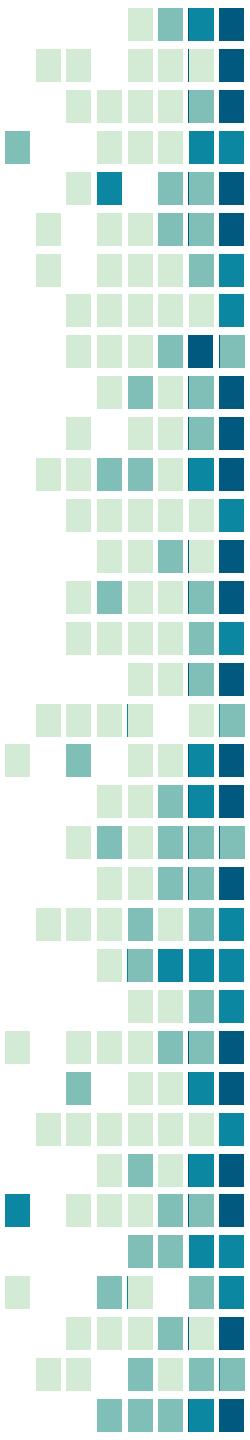
<gmi:MI_Metadata xmlns:gmi="http://www.isotc211.org/2005/gmi" xmlns:gco="http://www.isotc211.org/2005/gco">
  <gmd:fileIdentifier gco:nilReason="missing"/>
  <gmd:language>
    <gco:CharacterString>eng;USA</gco:CharacterString>
  </gmd:language>
  <gmd:characterSet>
    <gmd:MD_CharacterSetCode codeList="http://www.ngdc.noaa.gov/metadata/published/xsd/schem>
  </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.org/2005/gco"
  <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/schemacodeList.xml" codeListValue="UTF-8"/>
  </gmd:characterSet>
  <gmd:contact>
    <gmd:CI_ResponsponsibleParty>
      <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>1234 Main Street</gco:CharacterString>
              </gmd:deliveryPoint>
              <gmd:city>
                <gco:CharacterString>Anytown USA</gco:CharacterString>
              </gmd:city>
              <gmd:administrativeArea>
                <gco:CharacterString>USA</gco:CharacterString>
              </gmd:administrativeArea>
            </gmd:CI_Address>
          </gmd:address>
        </gmd:CI_Contact>
      </gmd:contactInfo>
    </gmd:CI_ResponsponsibleParty>
  </gmd:contact>
</gmi:MI_Metadata>
```

...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!

dublincoregenerator.com - a better dublin core generator

Main Page Simple Generator Advanced Generator xZINECOREx Generator About Contribute

Directions

- Fill in the fields below and click on "Generate Code!" to convert your input into fully formed Dublin Core metadata code. Additional options for the format of the output code are available below.
- If you need additional copies of a given field, click the plus sign to the upper-right of the tag's name to add an additional copy of it.
- Click the minus sign to delete any unneeded additional copies -- don't worry about removing tags you don't intend to use, the system will ignore any empty tags (and you can't delete the first row anyway).
- If you are unsure how a specific tag works, you can click the question mark next to the tag's name to see the tag's entry in Diane Hilmann's wonderful guide "Using Dublin Core -- The Elements."
- If you would like to use encoding schemes and the more advanced qualified elements of Dublin Core metadata, use the Advanced Generator located [here](#).

Input

Title?	[+] [-]
My Paper	
Creator?	[+] [-]
Jeanette Clark	
Subject?	[+] [-]
Example	
Description?	[+] [-]
Publisher?	[+] [-]
Contributor?	[+] [-]
Data?	[+] [-]

<http://dublincoregenerator.com>

Creating Standardized Metadata

- Specialized tools are your friend!

dublincoregenerator.com - a better dublin core generator

Main Page Simple Generator Advanced Generator xZINECOREx Generator About Contribute

Directions

- Fill in the fields below and click on "Generate Code!" to convert your input into fully formed Dublin Core metadata code. Additional options for the format of the output code are available below.
- If you need additional copies of a given field, click the plus sign to the upper-right of the tag's name to add an additional copy of it.
- Click the minus sign to delete any unneeded additional copies -- don't worry about removing tags you don't intend to use, the system will ignore any empty tags (and you can't delete the first row anyway).
- If you are unsure how a specific tag works, you can click the question mark next to it to learn more. See Diane Hilmann's wonderful guide "Using Dublin Core -- The Elements."
- If you would like to use encoding schemes and the more advanced qualified elements, click the link to the Advanced Generator located [here](#).

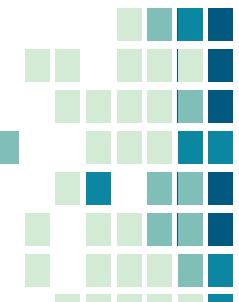
Input

Title?	[+] [-]
My Paper	
Creator?	[+] [-]
Jeanette Clark	
Subject?	[+] [-]
Example	
Description?	[+] [-]
Publisher?	[+] [-]
Contributor?	[+] [-]
Data?	[+] [-]

Output

```
<dc:title>My Paper</dc:title>
<dc:creator>Jeanette Clark</dc:creator>
<dc:subject>Example</dc:subject>
```

<http://dublincoregenerator.com>



Creating Standardized Metadata

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

[Report Issue](#) | [Suggest Improvement](#)



Investigating the effect of oil spills
on the environment and public health.



HOME SEARCH DATA ▾ SUBMIT DATA ▾ TRACKING & STATS ▾ ABOUT US ▾ HELP ▾ RESEARCH

ISO 19115-2 Metadata Editor

Load from File Load from Submitted Dataset Save to File Clear Form Check and Save to File Help

Dataset Contact Dataset Information Keywords Data Extent Distribution Info Distribution Contact Metadata Contact

NOTE: Fields with * are required.

Dataset Information

This section collects identifying and amplifying information about the dataset. Provides future researchers with specific details on the dataset content and additional context regarding the broader purpose of the dataset.

*Title <input type="text"/>	Name by which the cited resource is known. It is recommended the title include (where applicable) a description of the data, a date or date range, and geographic area.
Short Title <input type="text"/>	Short name or other language name by which the cited information is known.
*Date <input type="text"/> ...	Reference date for the cited dataset. This date can refer to dataset creation, publication, or revision. Format should be YYYY-MM-DD.
*Date Type <input type="button" value="Publication"/>	Creation: Date identifies when the resource was brought into existence. Publication: Date identifies when the resource was issued. Revision: Date identifies when the resource was examined or re-examined and improved or amended.
*Abstract <input type="text"/>	Brief narrative summary of the dataset's contents.

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', 'unit', 'formatString', 'numberType', 'missingValue', 'codeExplanation',
                           stringsAsFactors = FALSE)
                           measurementScale > attributeList2
                           domain = c(<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



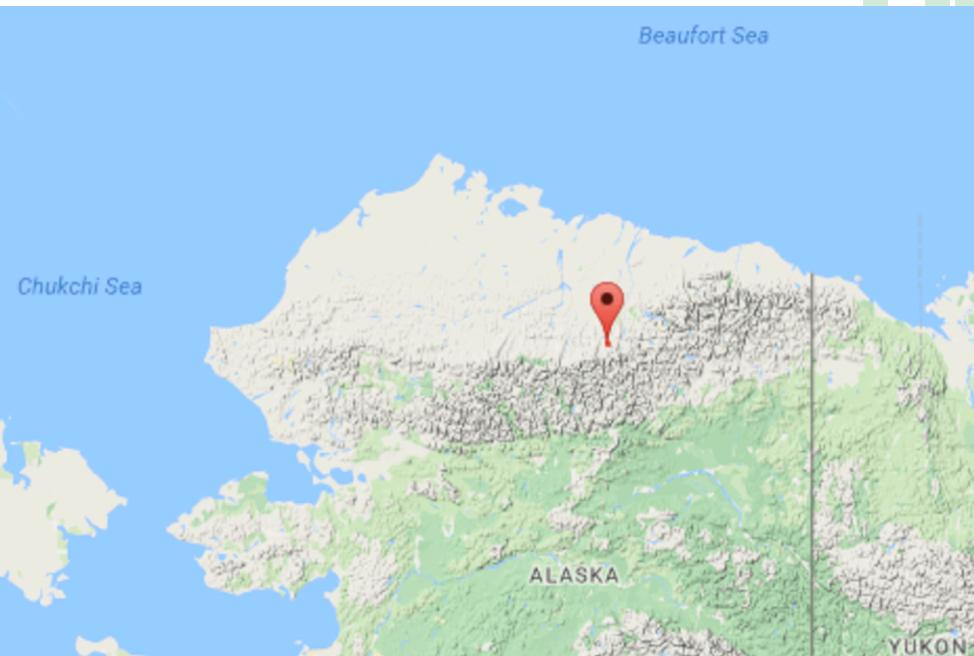
What makes a good metadata record?



What makes a good metadata record?

“Begin: 2002-04-14
End: 2003-04-13”

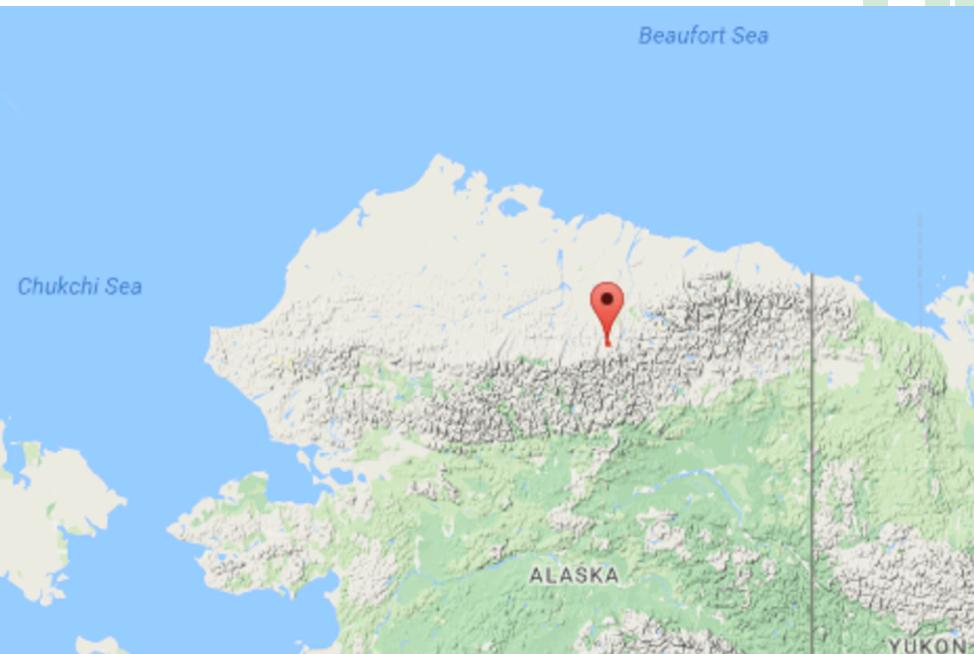
“Sagavanirktok River,
North Slope, Alaska”



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

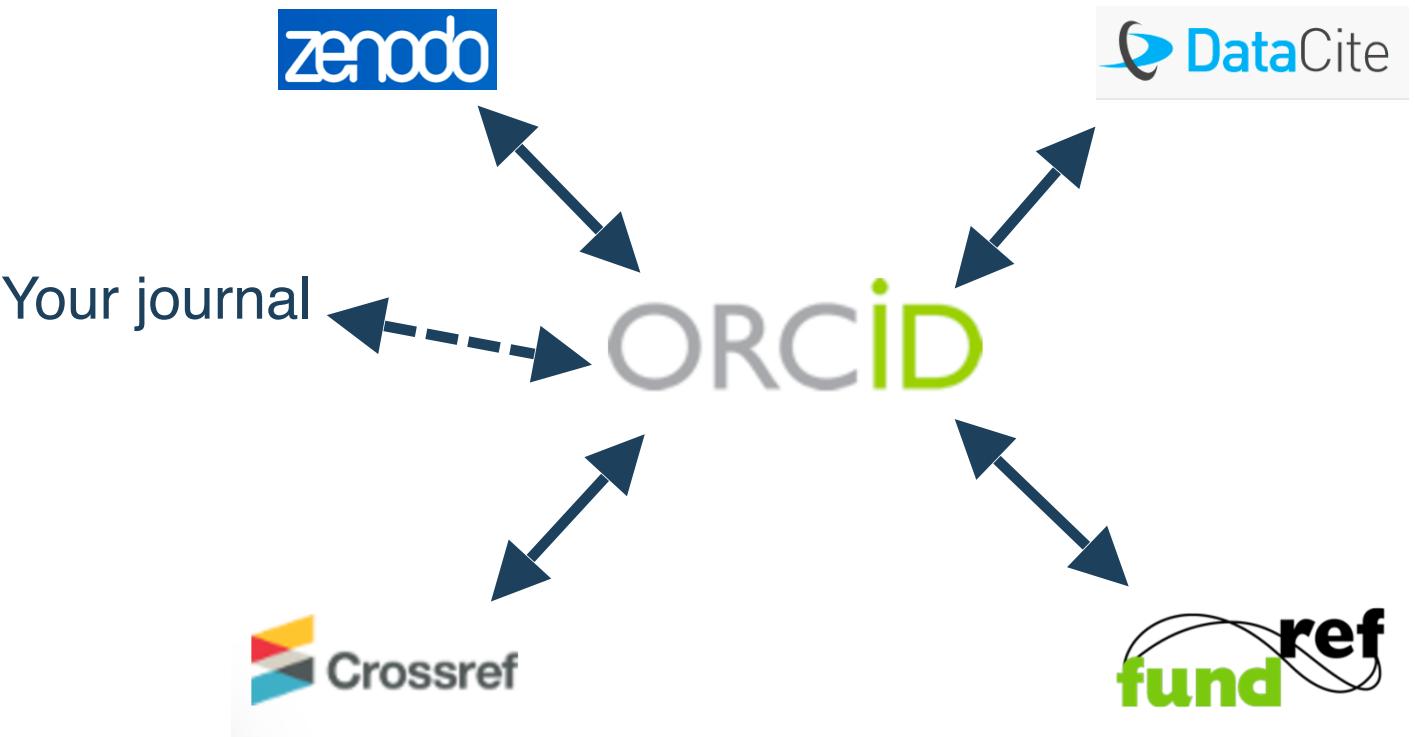
- Name alone is not enough...
 - to assign credit, nor
 - to disambiguate across data sets
- Email addresses help
- Including ORCID iD is best

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