NSF Arctic Data Center

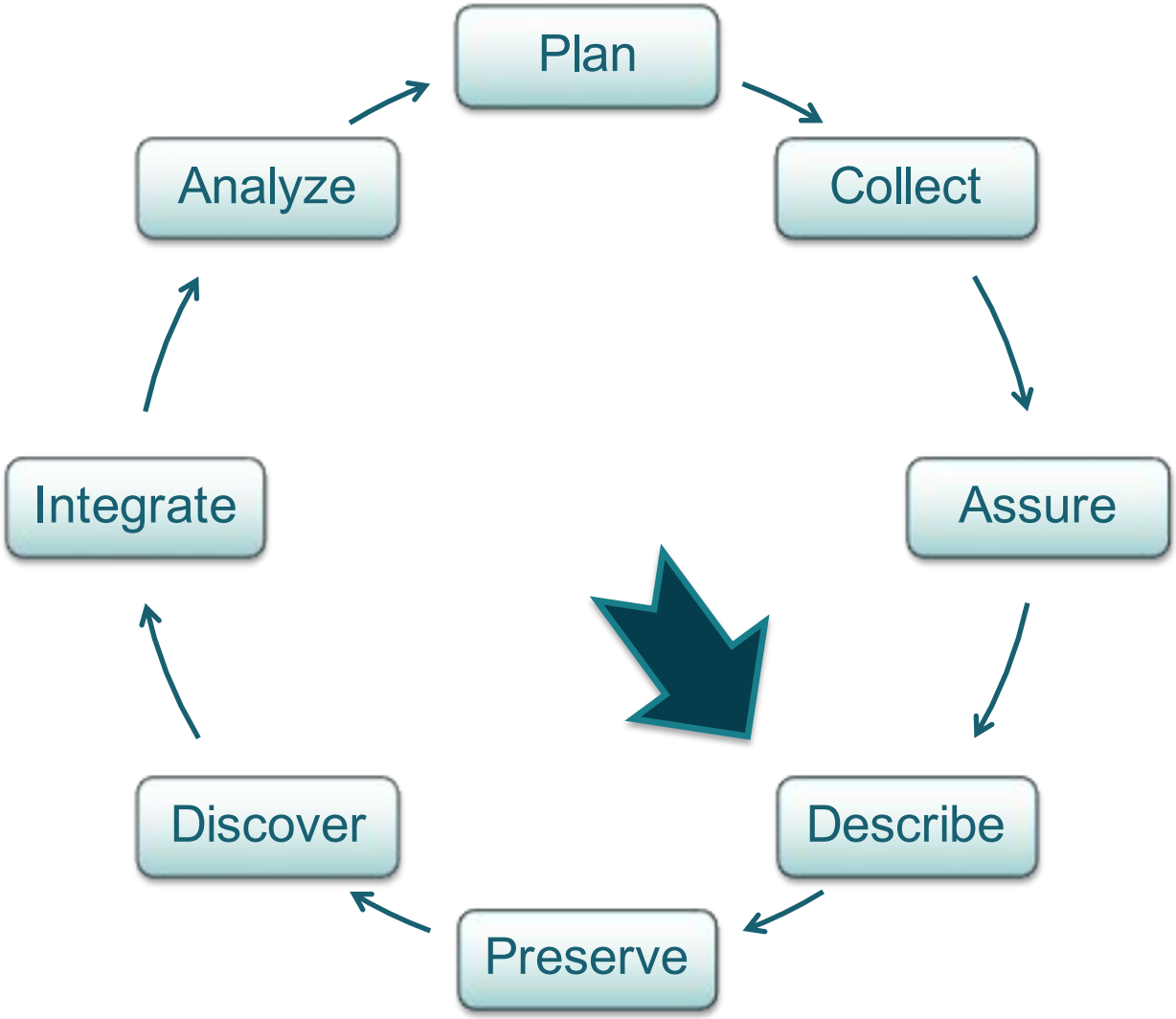
Authoring High Quality Metadata

Bryce Mecum
Scientific Software Engineer
NCEAS
orcid.org/0000-0002-0381-3766

https://arcticdata.io
NSF Award #: 1546024

Photo credit: duncan c (http://flic.kr/p/bBtZhy)

The Data Life Cycle



What is metadata?

Think of metadata as “data reporting”

- **Who** created the data?
- **What** is the content of the data?
- **When** were the data created?
- **Where** are the data from?
- **How** were the data developed?
- **Why** were the data developed?

Metadata powers our world

https://arstechnica.com/gadgets/2011/07/mac-os-x-lion-a-visual-introduction/

Author(s)	Boullosa, Carmen.
Title(s)	They're cows, we're pigs / by Carmen Boullosa
Place	New York : Grove Press, 1997.
Physical Descr	viii, 180 p ; 22 cm.
Subject(s)	Pirates Caribbean Area Fiction.
Format	Fiction

Nutrition Facts

Serving Size 4 OZ. SERVING (112g)
Servings Per Container VARIED

Amount Per Serving

Calories 170 — Calories from Fat 70

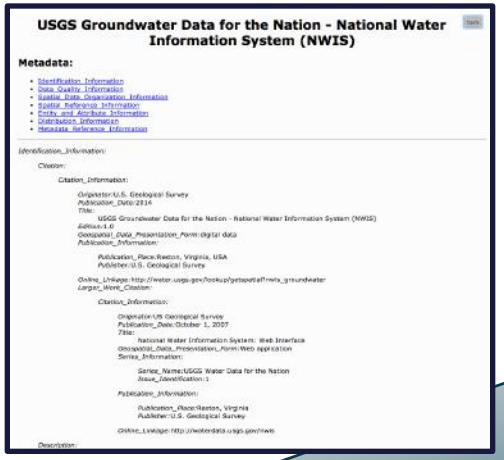
	% Daily Value*
Total Fat 8g	12%
Saturated Fat 3g	15%
Cholesterol 65mg	22%
Sodium 70mg	3%
Total Carbohydrate 0g	0%
Dietary Fiber 0g	0%
Sugars 0g	
Protein 23g	

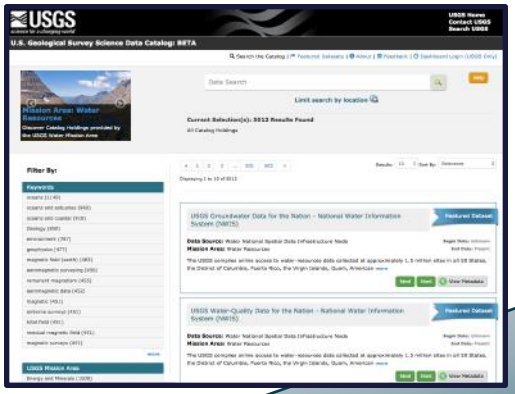
Vitamin A 0% • Vitamin C 0%
Calcium 0% • Iron 15%

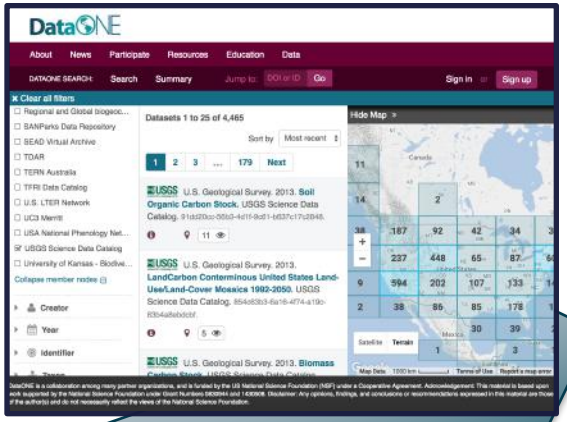
*Percent Daily Values are based on a 2,000 calorie diet.

What are metadata good for?

An example…





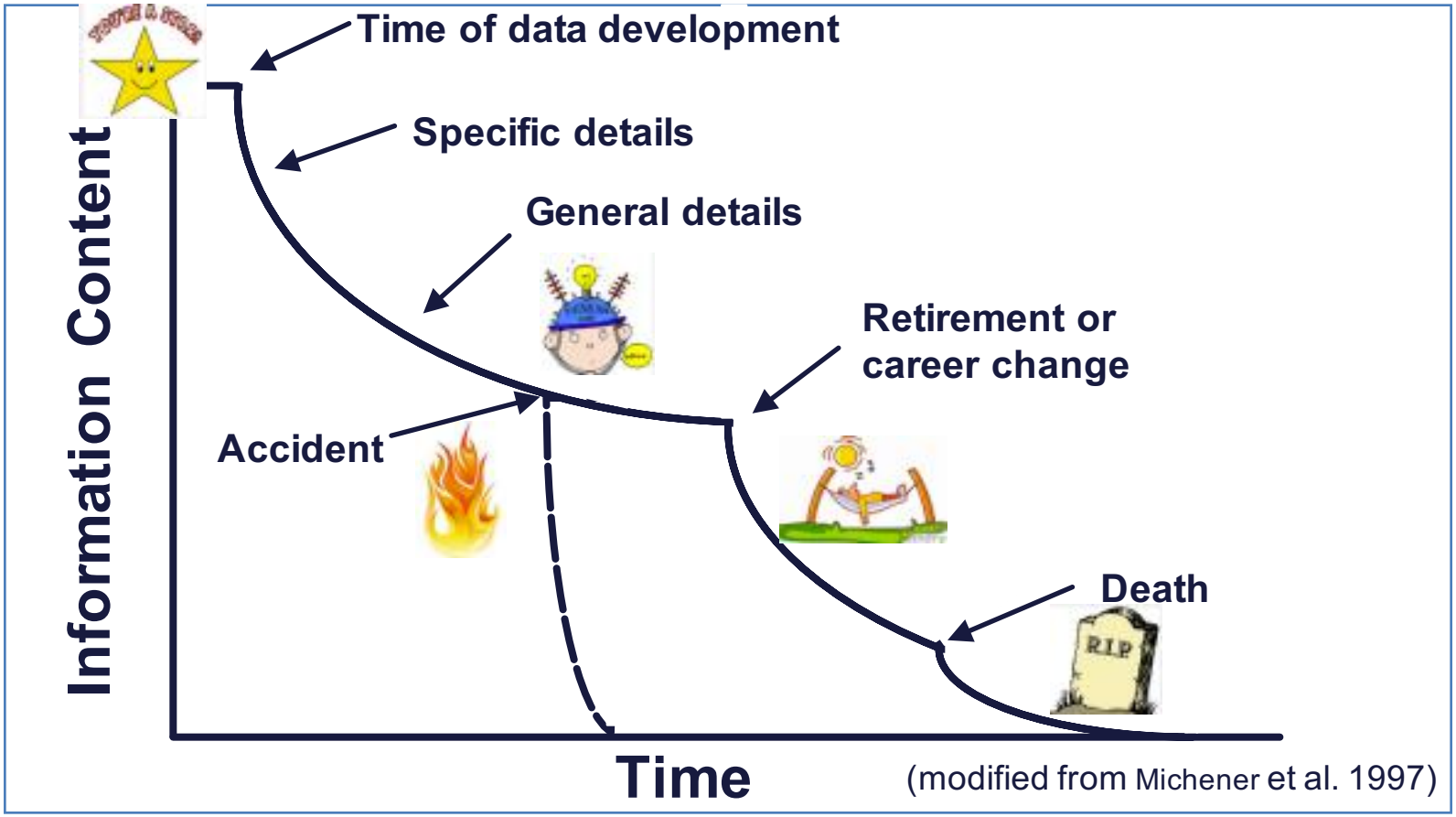


Metadata: captures information

USGS Science Data Catalog: enabling discovery

DataONE: enables exchange

The importance of metadata



(modified from Michener et al. 1997)

The importance of metadata

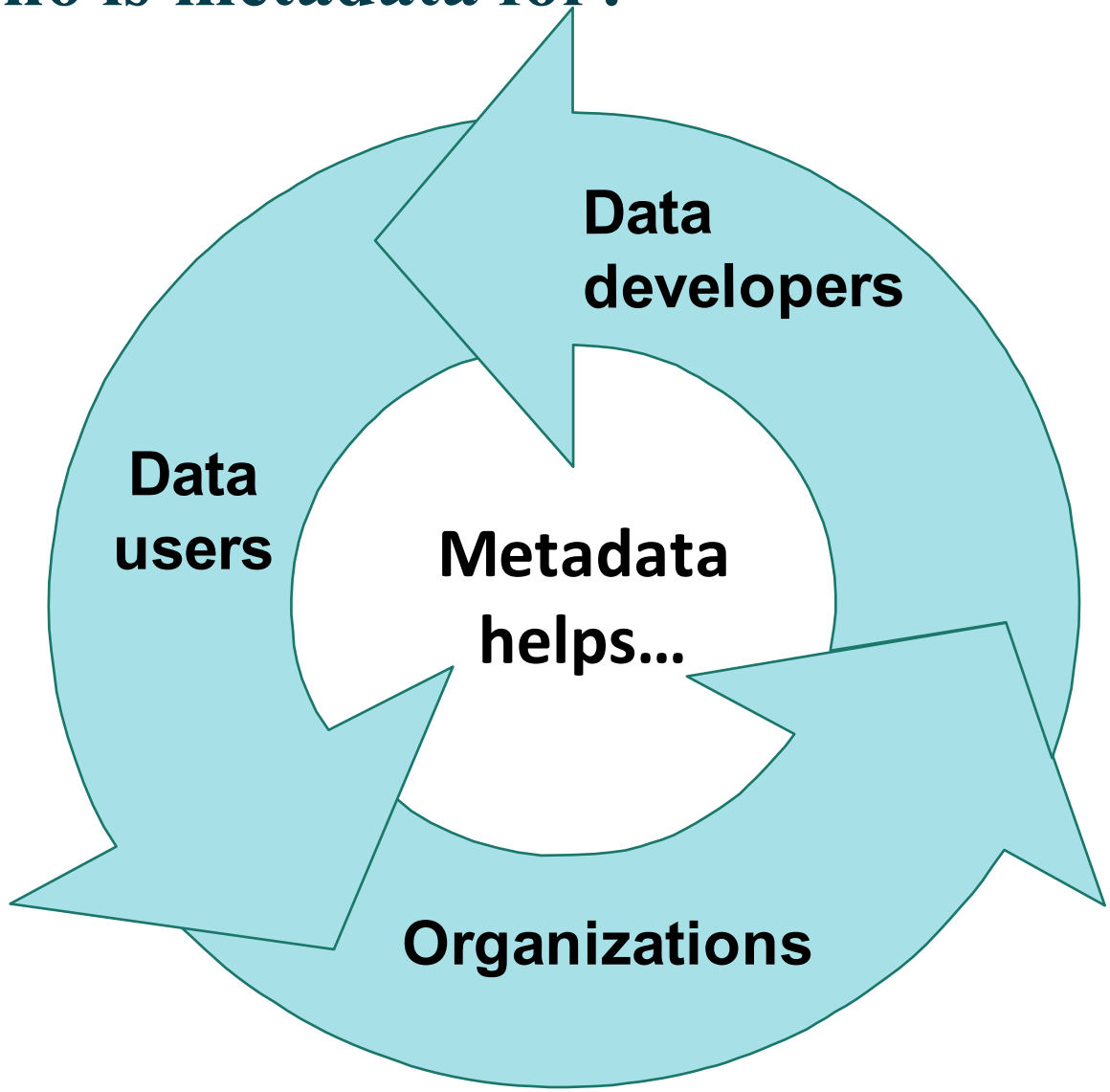
Metadata are important for the short and long-term utility of data

The importance of metadata

The importance of metadata

- Metadata are essential for policy work
  - Discovering data relevant to policy questions in the first place
  - Metadata support scrutiny of our data
    - Motivations
    - Methodologies
    - Conflicts of interest

Who is metadata for?



Metadata for data developers

- Avoid data duplication
  - What has been collected already?
- Share reliable information
  - What method was used?
  - What methods are in common use in my field?
- Publicize our work
  - Hey, I made this!
- Save time the next time
  - Hey, I've already done this!

Metadata for data users

- Find relevant data
- Evaluate what you find for suitable 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 & data processing methods get recorded
  - Ability to use data after initial intended purpose
  - Track data re-use and citation

Metadata for organizations

- Transcend people and time
  - Don’t lose our data when researchers/labs leave
  - Avoid duplication in new work

Metadata for organizations

- Advertise organization's research
  - What data has our organization produced?

Metadata for organizations

Metadata is all about scale

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 (emph. publications)
- Darwin Core (emph. collections)
- FGDC (emph. spatial)
- ISO19115 (emph. spatial & services)
- Ecological Metadata Language (general, but emphasis on filesystem artifacts, attributes, taxonomy)

Metadata standards


```xml
<?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>
                <gco:CharacterString>AK</gco:CharacterString>
              </gmd:administrativeArea>
              <gmd:postalCode>
                <gco:CharacterString>99501</gco:CharacterString>
```

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:administrativeAre
                <gco:CharacterString>AK</gco:CharacterString>
              </gmd:administrativeArea>
              <gmd:postalCode>
                <gco:CharacterString>99501</gco:CharacterString>
```


…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.

What makes a good metadata record?

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

- When in doubt, be more specific:
  - Spell out acronyms
  - Use full names, emails, addresses, etc.

- Include as much info 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 doesn't 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:
  - Who
  - What
  - When
  - Where
  - Why

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

What makes a good metadata record?

- Titles: Which is preferable?

River Data

or…

Greater Yellowstone Rivers from 1:126,700
U.S. Forest Service Visitor Maps (1961-1983)

What makes a good metadata record?

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

What makes a good metadata record?

- Documented filesystem artifacts
  - File formats
  - Checksums (Do I have the same file?)
  - Where to download (web address)
  - Attributes used (variables)

What makes a good metadata record?

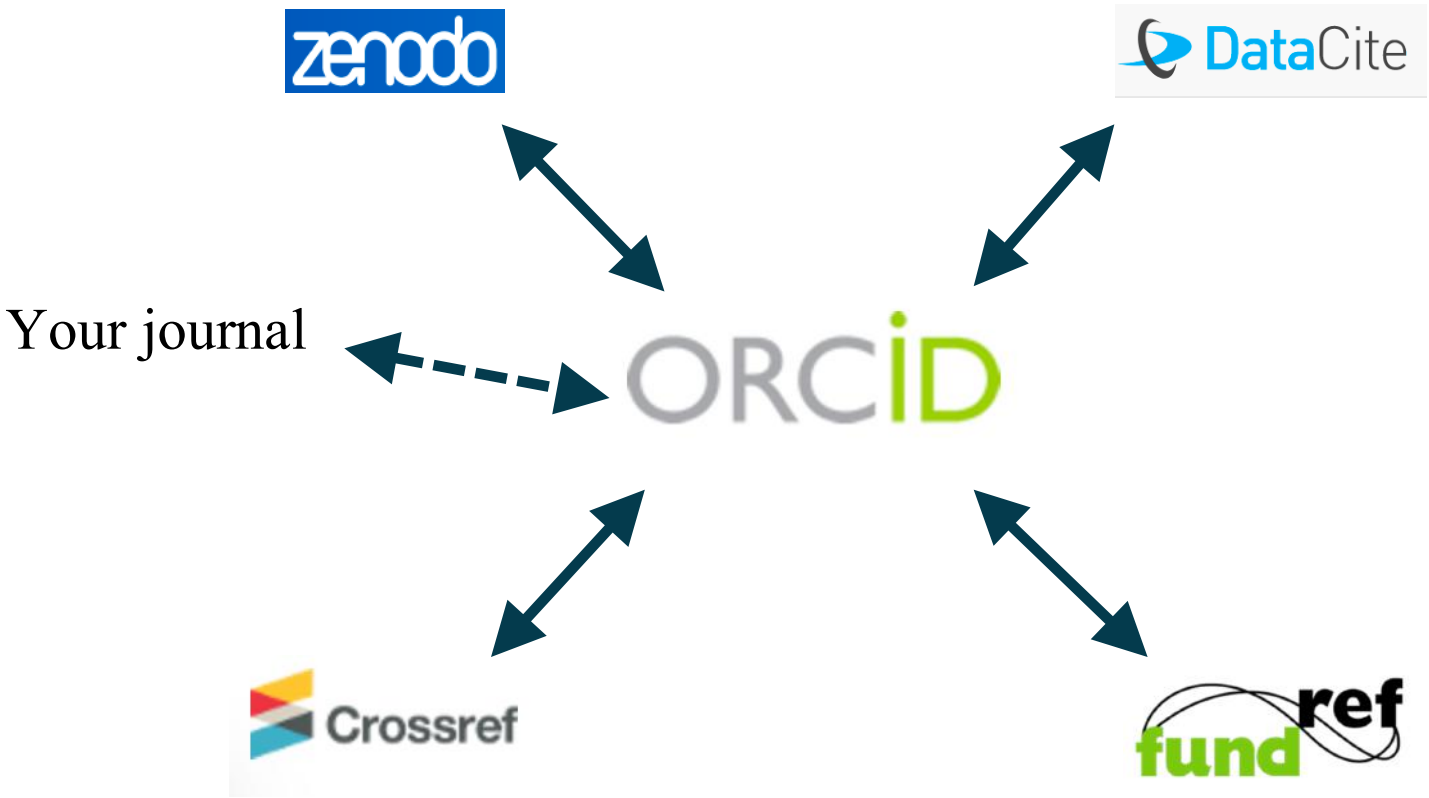
- Involved parties
  - Name alone is not enough
    - To assign credit
    - To disambiguate across datasets
  - Email helps
  - ORCiD (w/ above) is best

ORCiDs: Wait, what is an ORCiD?

- Like an ISBN for people
  - e.g. mine: 0000-0002-0381-3766
- Enables unambiguous reference to humans
- Free
- Becoming a community norm
- Inherently connected…

ORCiDs

- Inherently connected



Activity

- Get an ORCiD:
  - https://orcid.org/
- Sign in to https://dev.nceas.ucsb.edu/