

DataONE

Data Observational Network for Earth

Rebecca Koskela

William Michener

Dave Viegla

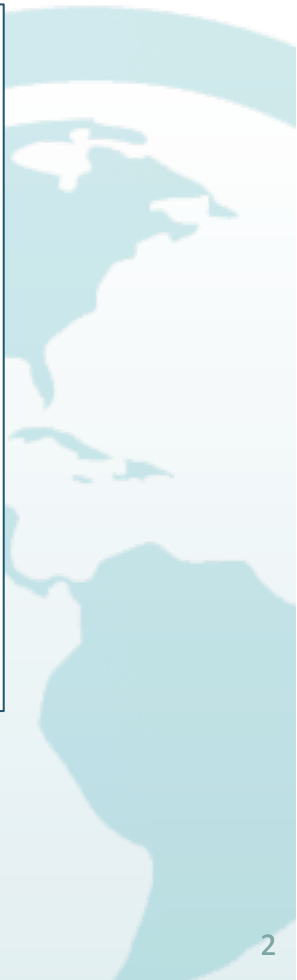
Amber Budden

OSTP/NITRD Data Sharing and Metadata Curation:
Obstacles and Strategies

May 29, 2013



The metadata problem



OPEN ACCESS Freely available online

PLOS one

Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir^{1*}, Suzie Allard¹, Kimberly Douglass¹, Arsev Umur Aydinoglu¹, Lei Wu¹, Eleanor Read², Maribeth Manoff², Mike Frame³

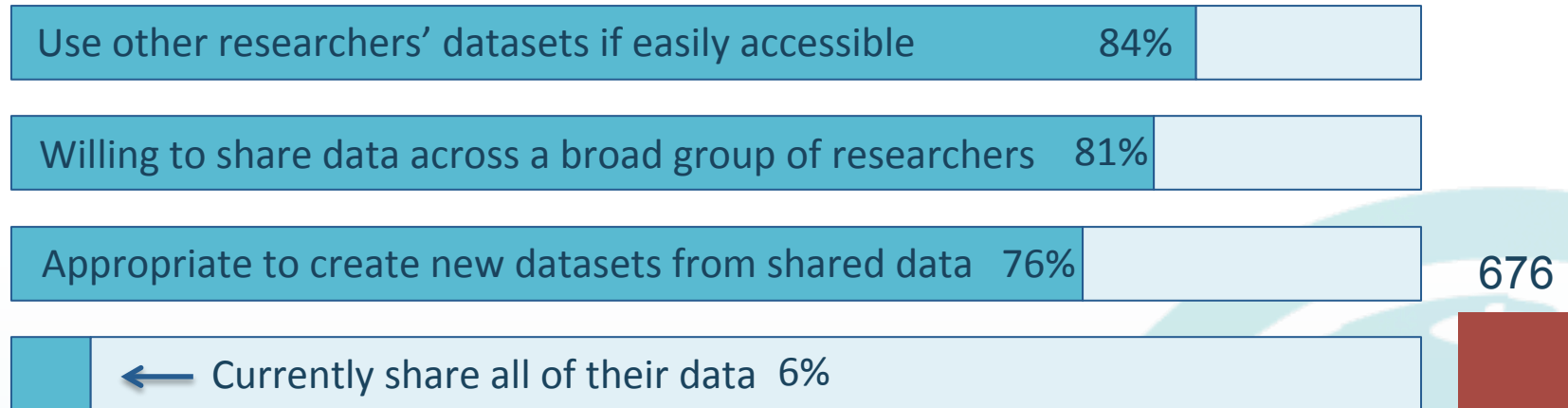
¹ School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, ² University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, ³ Center for Biological Informatics, United States Geological Survey, Oak Ridge, Tennessee, United States of America

Abstract

Background: Scientific research in the 21st century is more data intensive and collaborative than in the past. It is important to study the data practices of researchers – data accessibility, discovery, re-use, preservation and, particularly, data sharing. Data sharing is a valuable part of the scientific method allowing for verification of results and extending research from prior results.

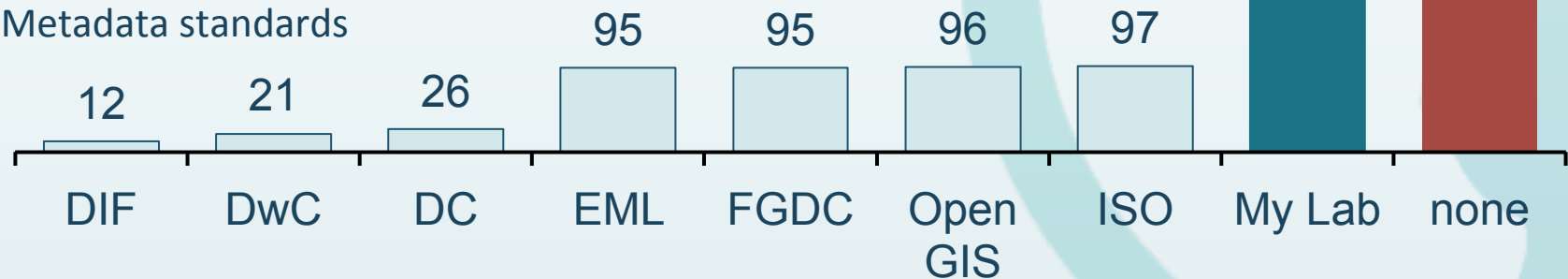
Methodology/Principal Findings: A total of 1329 scientists participated in this survey exploring current data sharing practices and perceptions of the barriers and enablers of data sharing. Scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding. Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle (collecting their research data; searching for, describing or cataloging, analyzing, and short-term storage of their data) but are not satisfied with long-term data preservation. Many organizations do not provide support to their researchers for data management both in the short-

Scientists want to share data



but don't know how to and, if they do, want to get proper credit for doing so.

Metadata standards



Some solutions

- Make it easy to describe data
- Provide credit to the data/metadata author
 - Citation
 - Promote discoverability
- Mandates (ideally, funded!)



Best Practices and Software Tools



Data Observation Network for Earth

Search For

Connect 

About Participate Resources Education Data

Home » Resources » [Software Tools Catalog](#)

Software Tools Catalog

Resources

- Investigator Toolkit
- Data Management Planning
- Best Practices
- Software Tools Catalog
- Publications

Tags

- statistics graphics web 2.0
- geospatial metadata edit
- visualization mode
- metadata map
- analyze database G



Data Observation Network for Earth

Search For

Connect 

About Participate Resources Education Data

Home » Resources » [Best Practices](#)

Best Practices

Search Best Practices

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle (shown below). Users can access best practices within the database by either clicking on a stage of the lifecycle, selecting keywords (under advanced search) or using free search.

For students and others new to data management, we provide a **Best Practices Primer** as an introduction to the DataONE Best Practices database and data management in general.

The development of the DataONE Best Practices database was a collaborative effort across many individuals ([credits](#)).

Resources

- Investigator Toolkit
- Data Management Planning
- Best Practices
- Software Tools Catalog
- Publications

Tags

- metadata analyze
- preserve data archives
- assure documentation
- describe format provenance

Making it easy to describe data



Intercept researchers where **they already work**







An open source tool helping researchers document, manage, and archive their tabular data. DataUp operates within the scientist's workflow and integrates with Microsoft® Excel.

Looking for the **DataUp** blog?

[Go to DataUp](#)

DataUp Features	Start Using DataUp	Customize DataUp	Contact Us
 See what DataUp can do for you - Check for Best Practices - Create Metadata - Get Credit for Data - Archive Data			

About the Project

Funders, partners, history and evolution of DataUp

Resources

CDL resources, data management help, presentations about DataUp

News & Events

Stay informed about DataUp, including upcoming presentations and demos

Discover UC3

Connect to the UC Curation Center's services for data

Add your email to the DataUp listserve for updates

Email Address:

[Submit](#)

Tweets

[@DataUp at CDL](#) 17 Oct
[@cdljournal](#) New post on [#DataUpBlog](#): Provenance and [#DataUp](#) at the [#MSEScience](#) workshop

[@DataUp at CDL](#) 16 Oct
[@cdljournal](#) I'm finally willing to tweet my [#DataUp](#) interview w/ [@clarkefor](#) Okios blog (previous photo = REALLY embarrassing)

[@cdljournal.wordpress.com](#) [Show Summary](#)

[@DataUp at CDL](#) 16 Oct
 Tweet to [@DataUpCDL](#)


 Powered by The California Digital Library Terms and Conditions Privacy Policy Accessibility Policy Contact Us © 2012 The Regents of the University of California

[illegible]

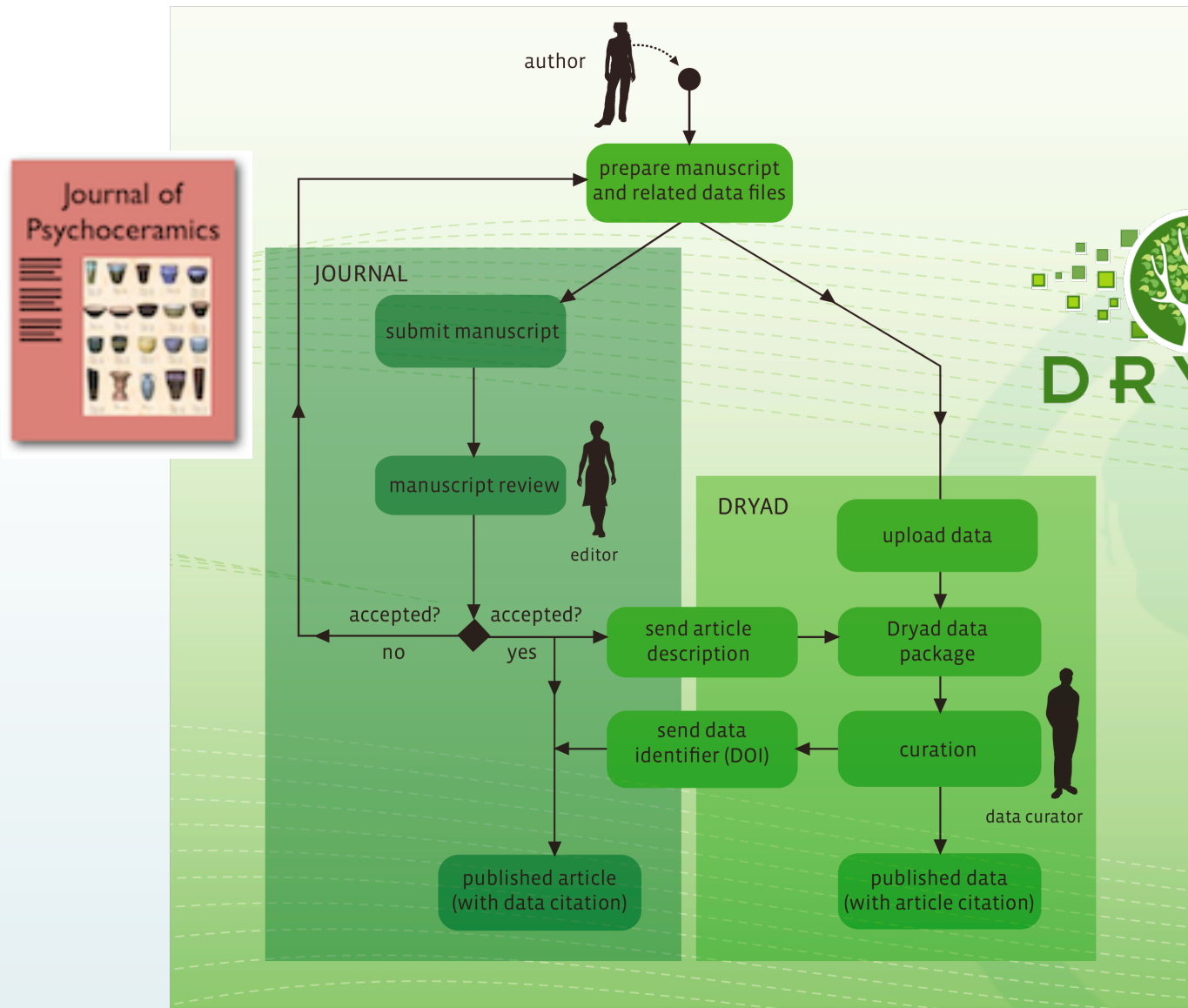
GENERAL METADATA

Creator-First name: Carly
 Creator-Last name: Stromer
 Creator-Address: 413 250th St
 Creator-State/province: Oakland
 Creator-Postal code: CA
 Creator-Country: USA
 Creator-Email: carly.stromer@ucop.edu
 Creator-Phone: 510-987-0379
 Creator-Organization: California Digital Library
 Title of dataset*: Geopanelog_2012-06
 Today's date*: June 23 2012
 Abstract
 Repository name and contact info...: ODSpace
 url for data:

Keyword(s)
 Keyword thesaurus used
 Geographic coverage: Description
 Geographic coverage: West bounding coordinates
 Geographic coverage: East bounding coordinates
 Geographic coverage: North bounding coordinates
 Geographic coverage: South bounding coordinates
 Temporal coverage: Beginning date
 Temporal coverage: Ending date
 Project title
 Project description
 Funding
 Intellectual rights

CASSI
 CC-BY

Credit: Dryad repository for journal data & metadata



Promoting data citations via Dryad

Article

Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, Venter JC, Eisen JA (2011) Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in phylogenetic trees of phylogenetic marker genes. PLoS ONE 6(3): e18011.

[doi:10.1371/journal.pone.0018011](https://doi.org/10.1371/journal.pone.0018011)

Dryad data package

Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, Venter JC, Eisen JA (2011) Data from: Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in phylogenetic trees of phylogenetic marker genes. Dryad Digital Repository.

[doi:10.5061/dryad.8384](https://doi.org/10.5061/dryad.8384)

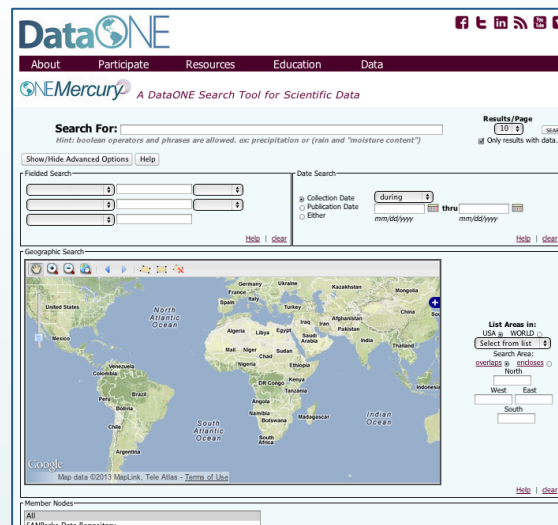
Promoting data discovery

Provide universal access to data about life on earth and the environment

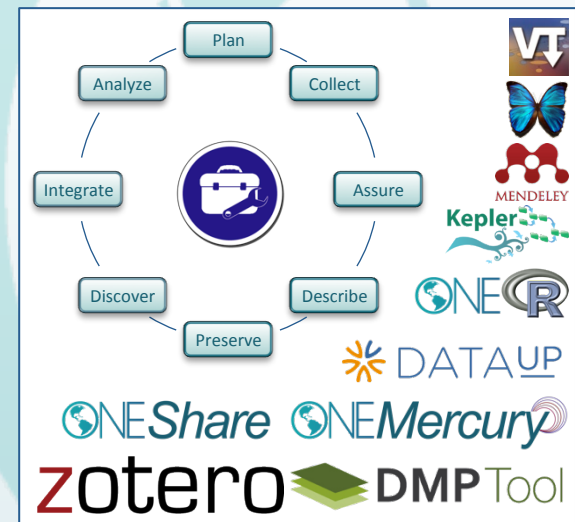
1. Building community



2. Developing sustainable data discovery and interoperability solutions



3. Enabling science through tools and services

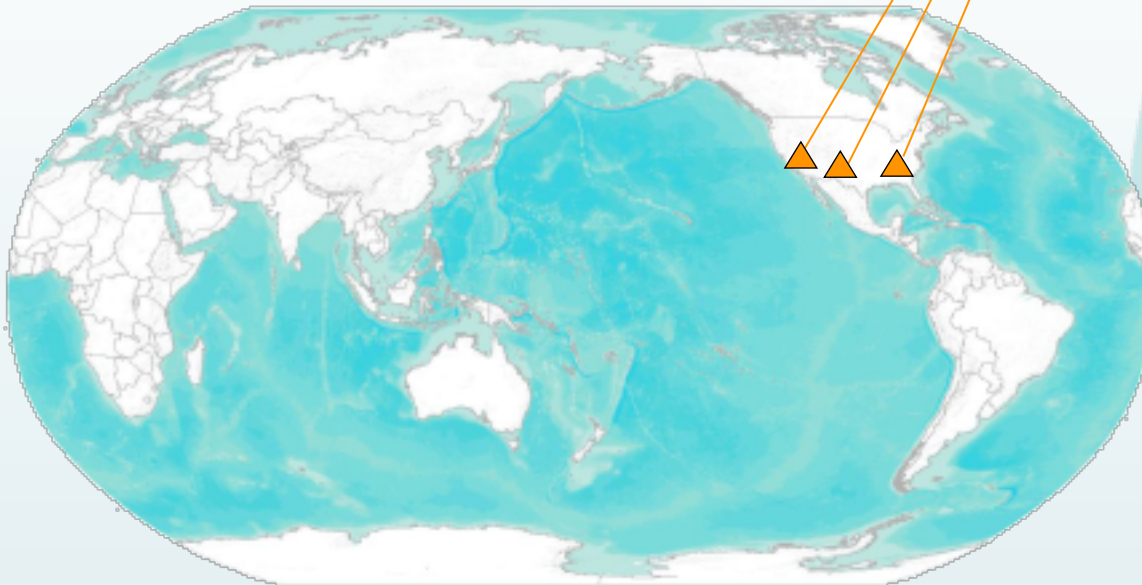


DataONE

Three major components for a flexible, scalable, sustainable network

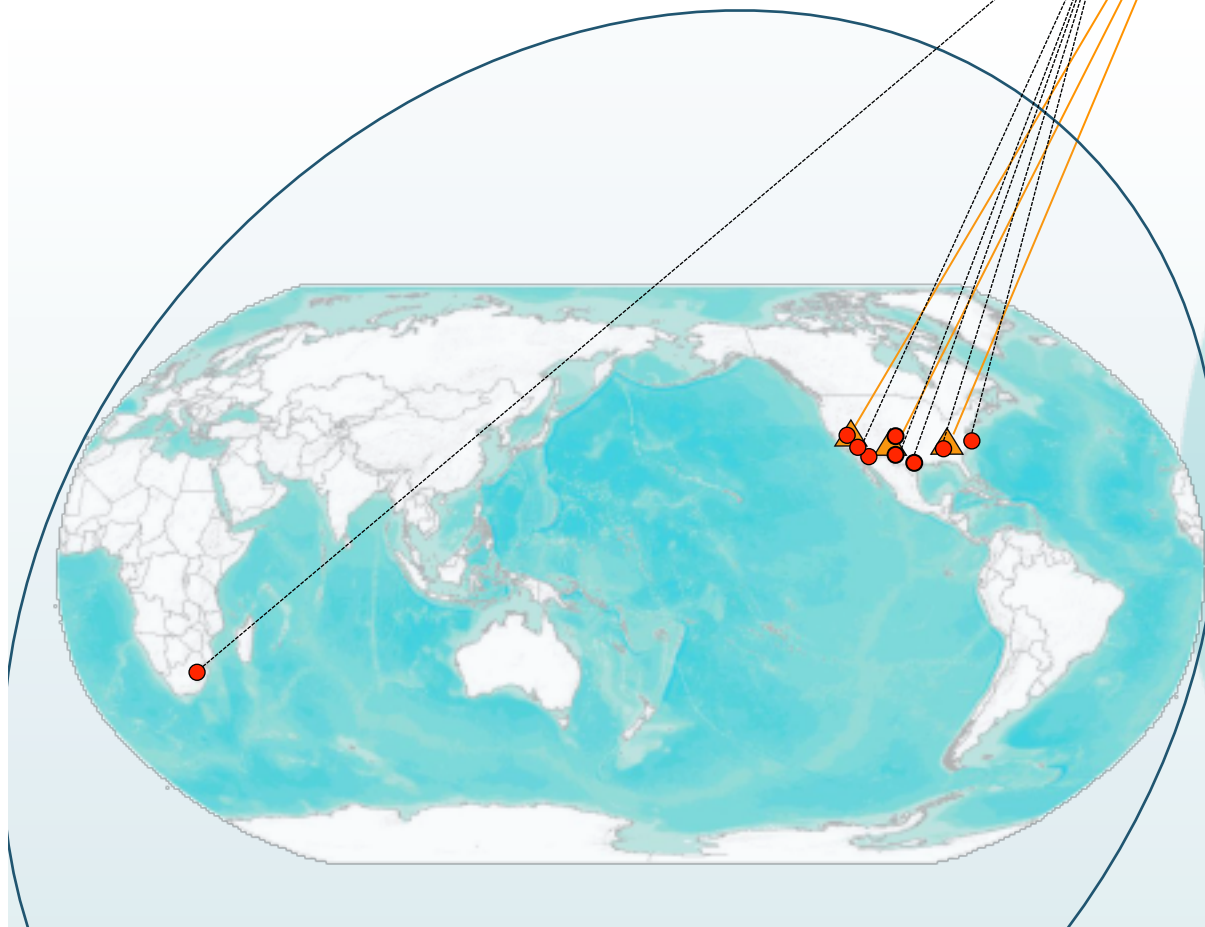
Coordinating Nodes

- retain complete metadata catalog
- indexing for search
- network-wide services
- ensure content availability (preservation)
- replication services



DataONE

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data



The **Cornell** Lab 



UC3 Merritt



PISCO



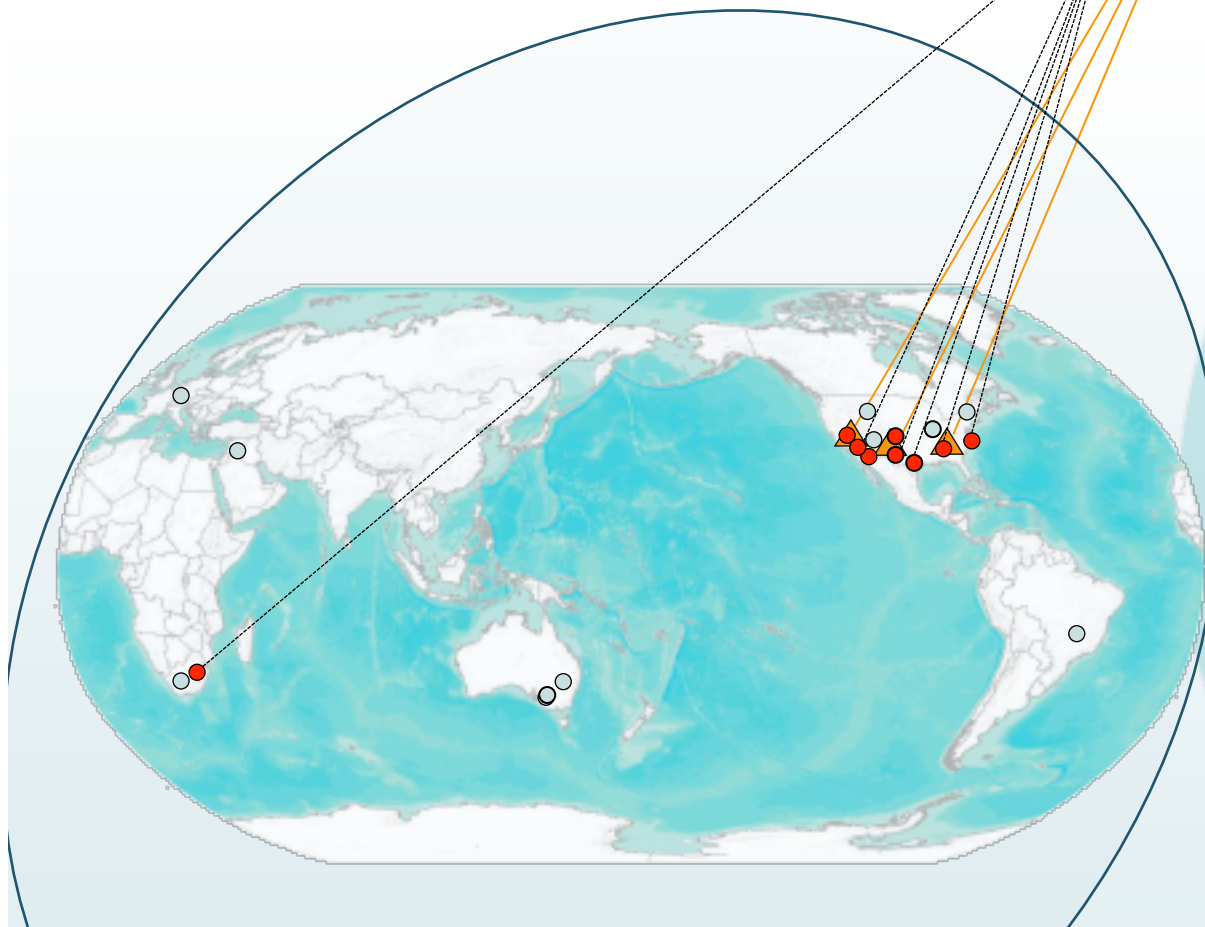
ECOLOGICAL SOCIETY OF AMERICA
esa



ONE Share

DataONE

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data



The **Cornell** Lab



UC3 **Merritt**



PISCO



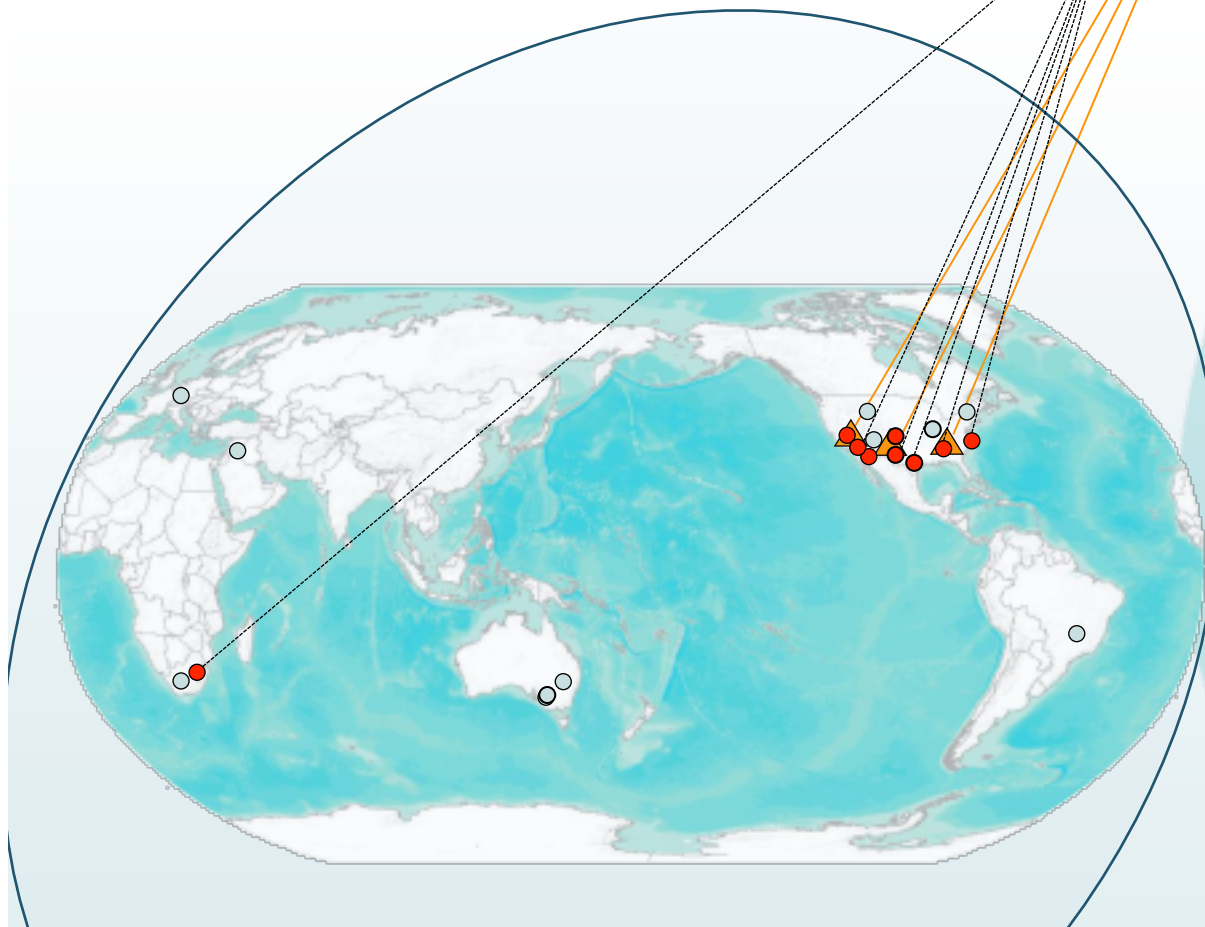
ECOLOGICAL SOCIETY OF AMERICA
esa



ONE **Share**

DataONE

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

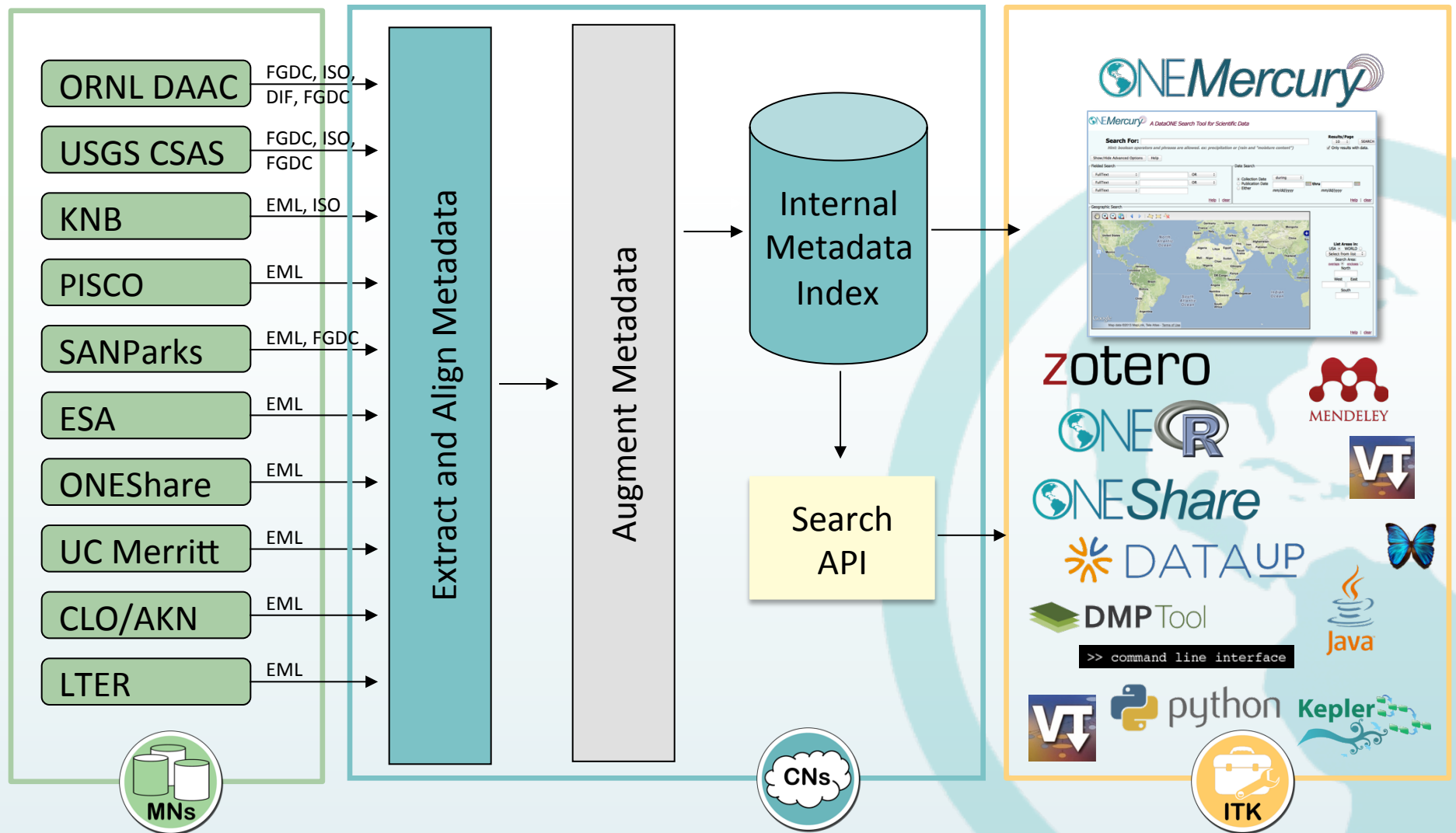
Member Nodes

Investigator Toolkit

>> command line interface



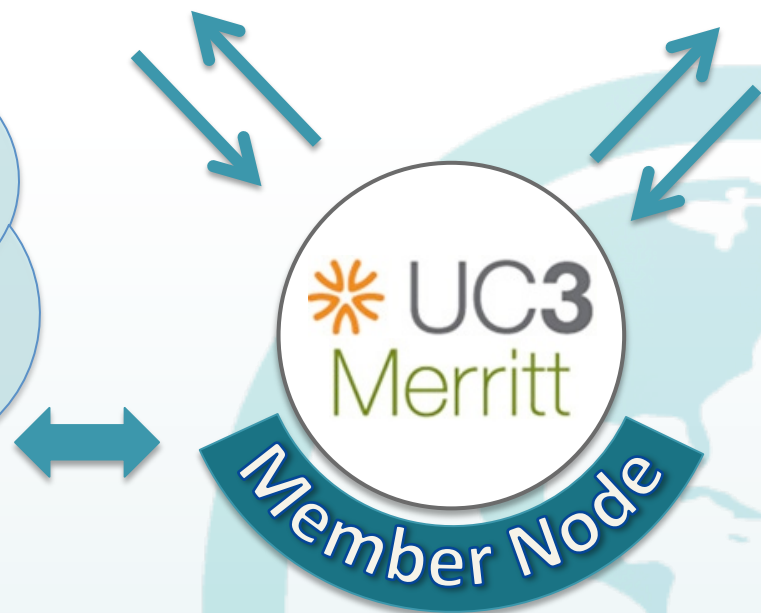
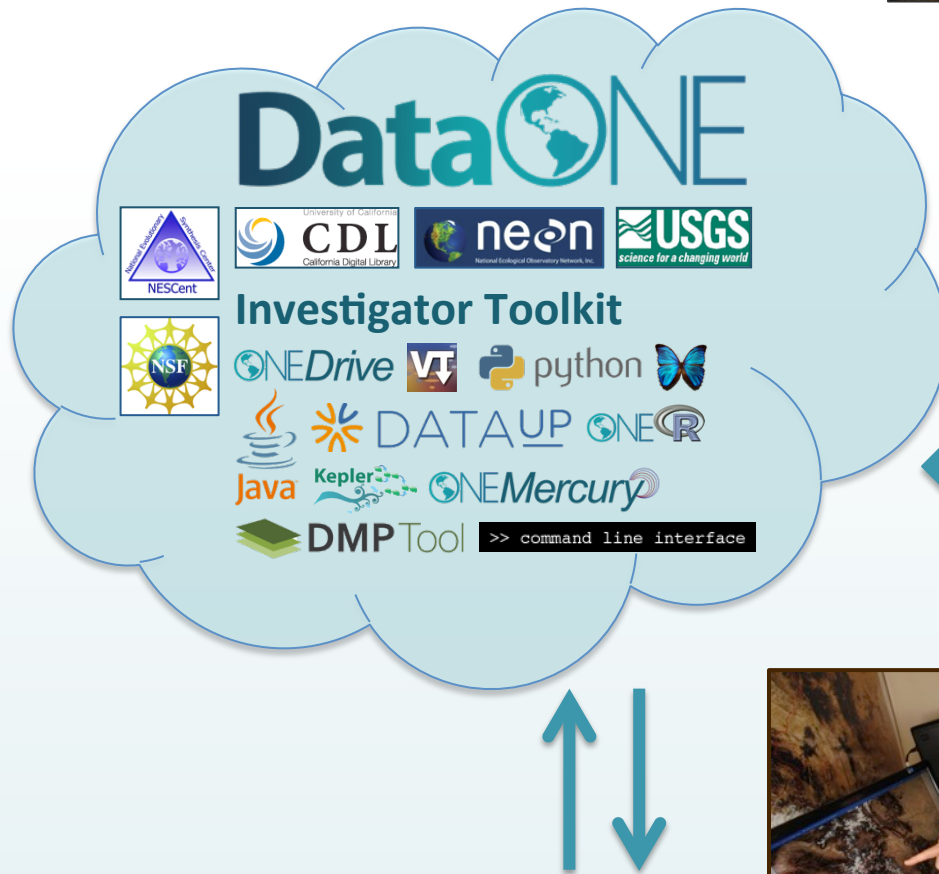
DataONE: Enabling data discovery



 Information
Center for the
Environment
(ICE) UC Davis



agencies
citizens
faculty
ICE Users



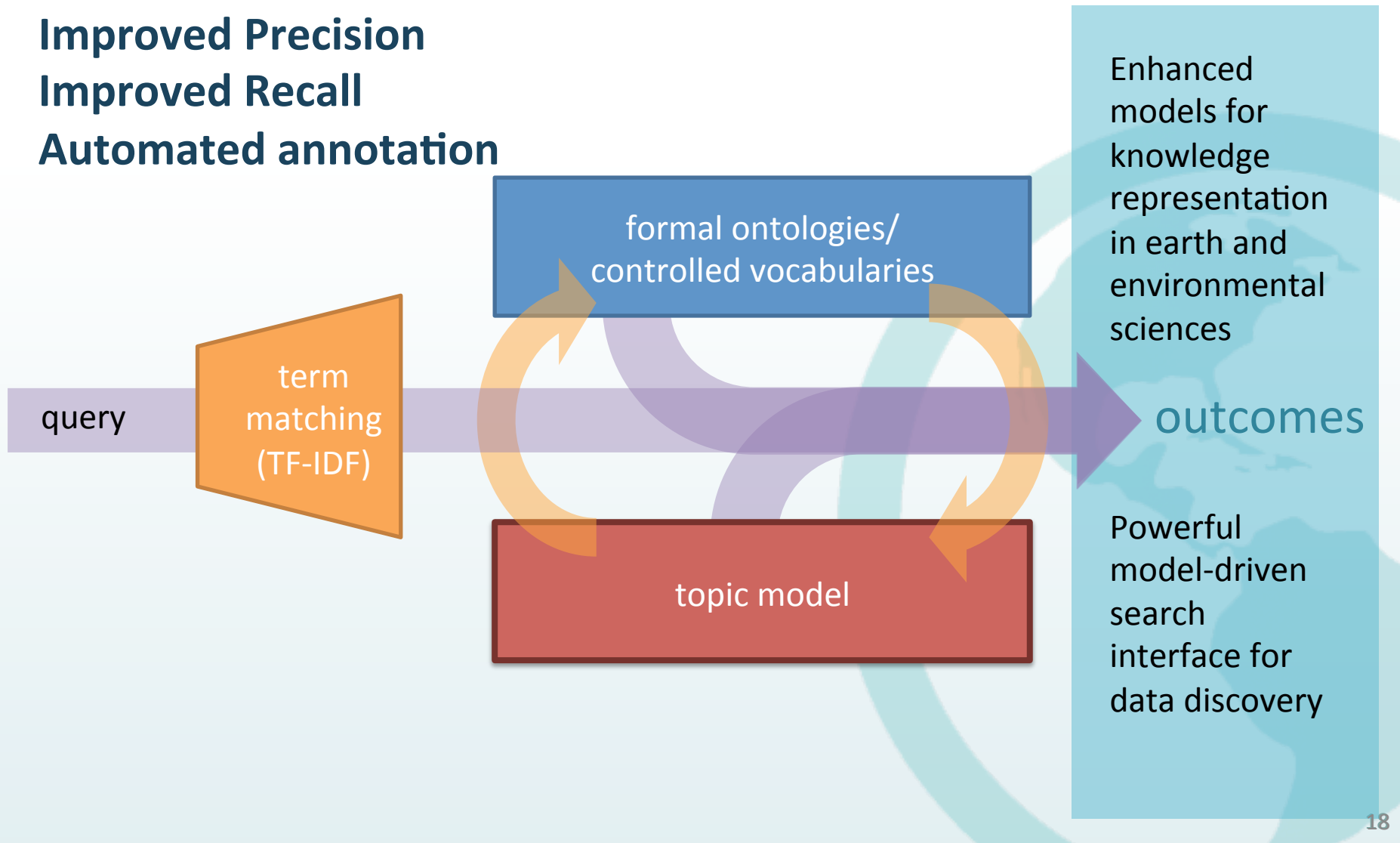
Some remaining challenges

- Semantic mediation
- Provenance
- Improving metadata quality over time



Powerful Data Discovery via Semantics

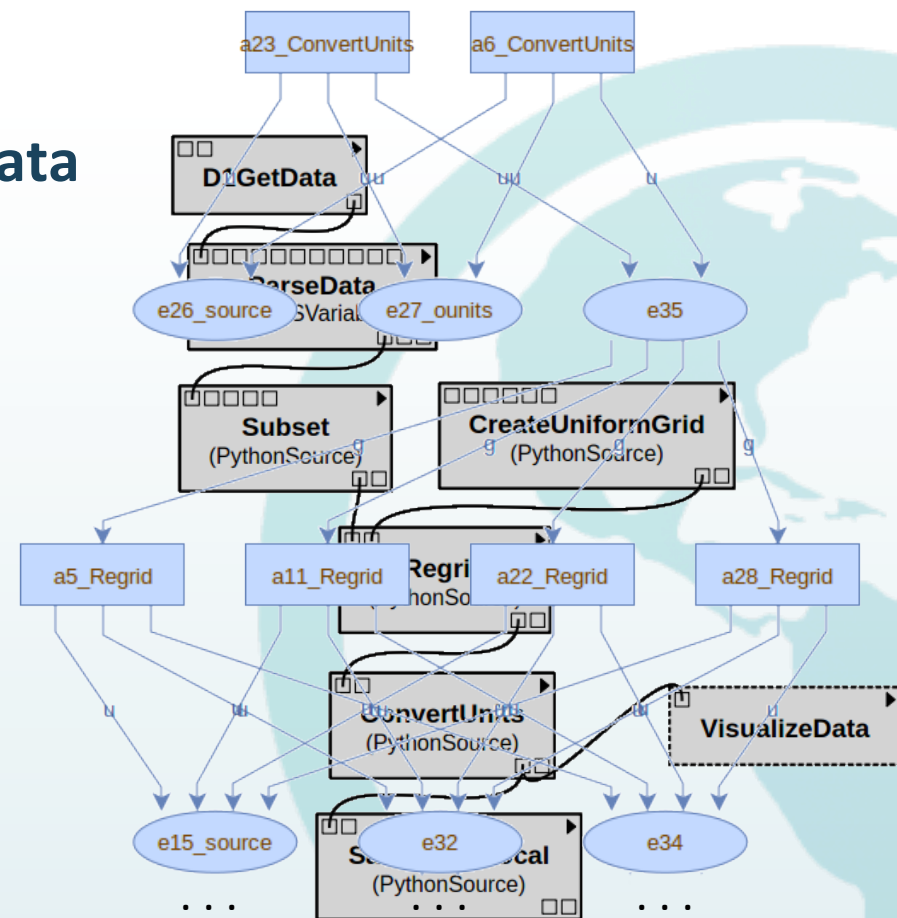
Improved Precision
Improved Recall
Automated annotation



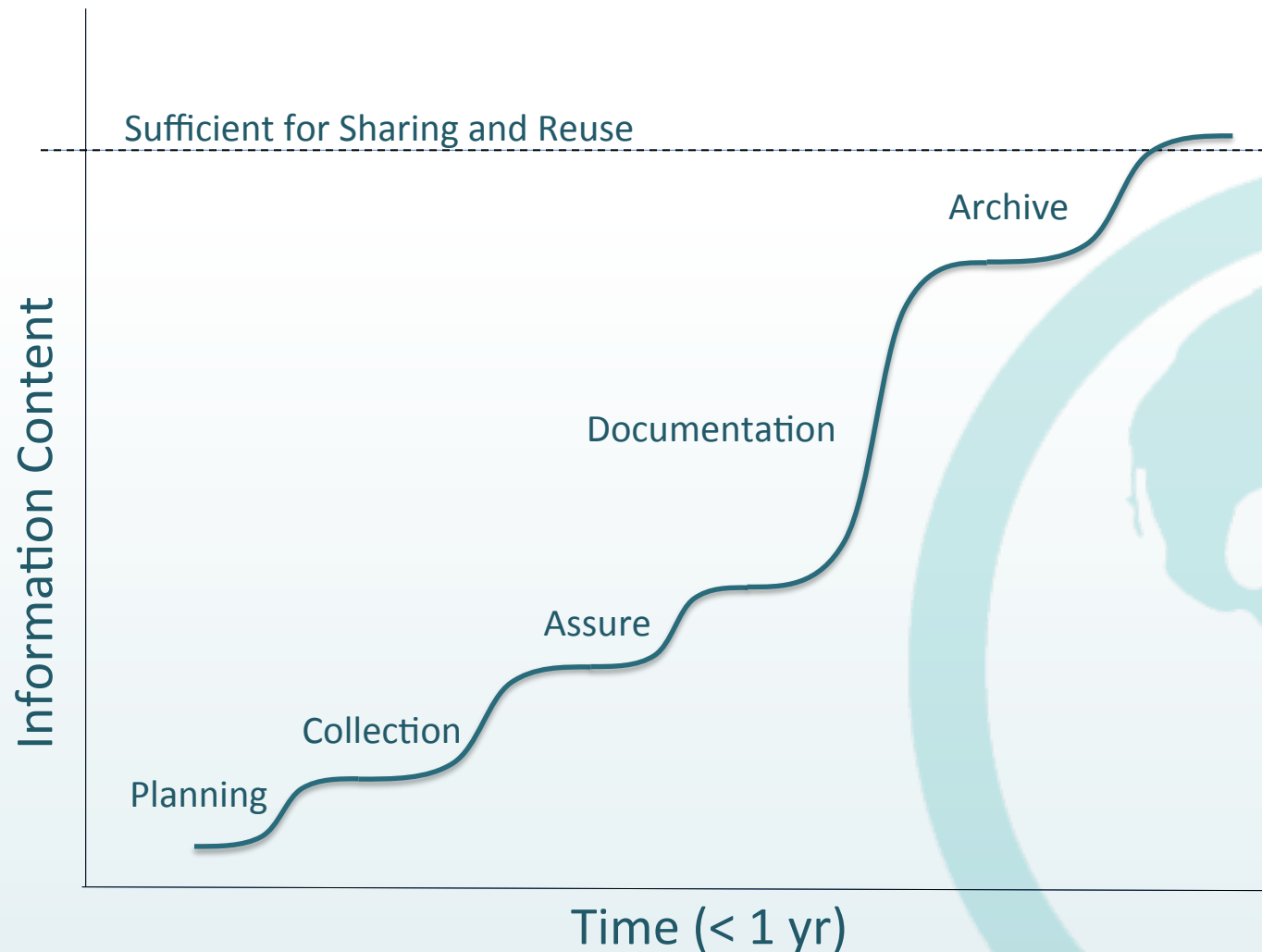
Provenance

Origin, context, derivation, ownership, history of (data) artifacts

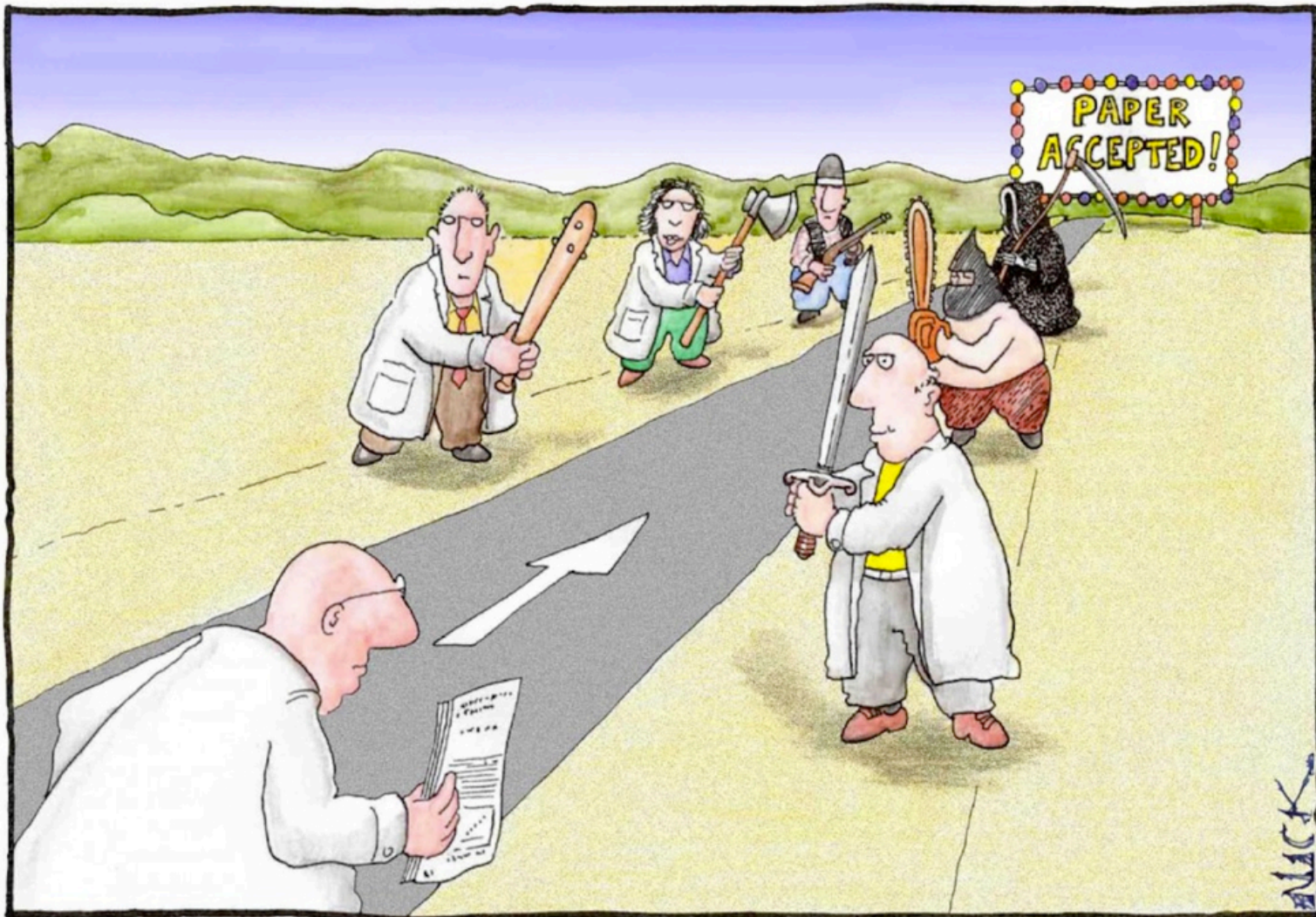
- Record **processing history, data lineage**
 - *dependency graph*
- W3C standard: PROV
- DataONE Extension: **D-PROV**
 - Workflow provenance
 - System agnostic!



Improving metadata quality for data reuse



Mandates (ideally, funded!)



Most scientists regarded the new streamlined peer-review process as "quite an improvement."

DataONE: Supporting scientific data preservation, discovery, and innovation



www.dataone.org