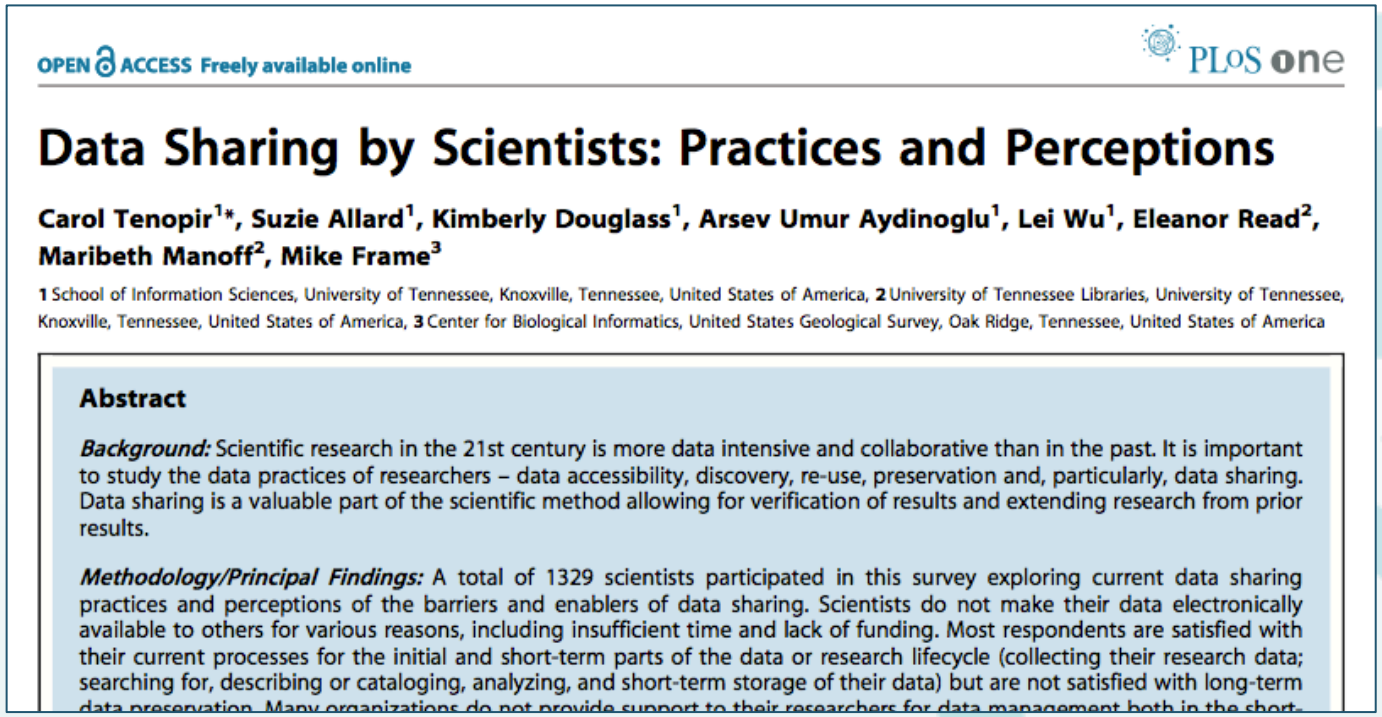
# DataONE
# Data Observational Network for Earth

Rebecca Koskela
William Michener
Dave Vieglais
Amber Budden

OSTP/NITRD Data Sharing and Metadata Curation:
Obstacles and Strategies
May 29, 2013

DataONE

# The metadata problem



OPEN ACCESS Freely available online

PLoS one

## Data Sharing by Scientists: Practices and Perceptions

**Carol Tenopir[1]*, Suzie Allard[1], Kimberly Douglass[1], Arsev Umur Aydinoglu[1], Lei Wu[1], Eleanor Read[2], Maribeth Manoff[2], Mike Frame[3]**

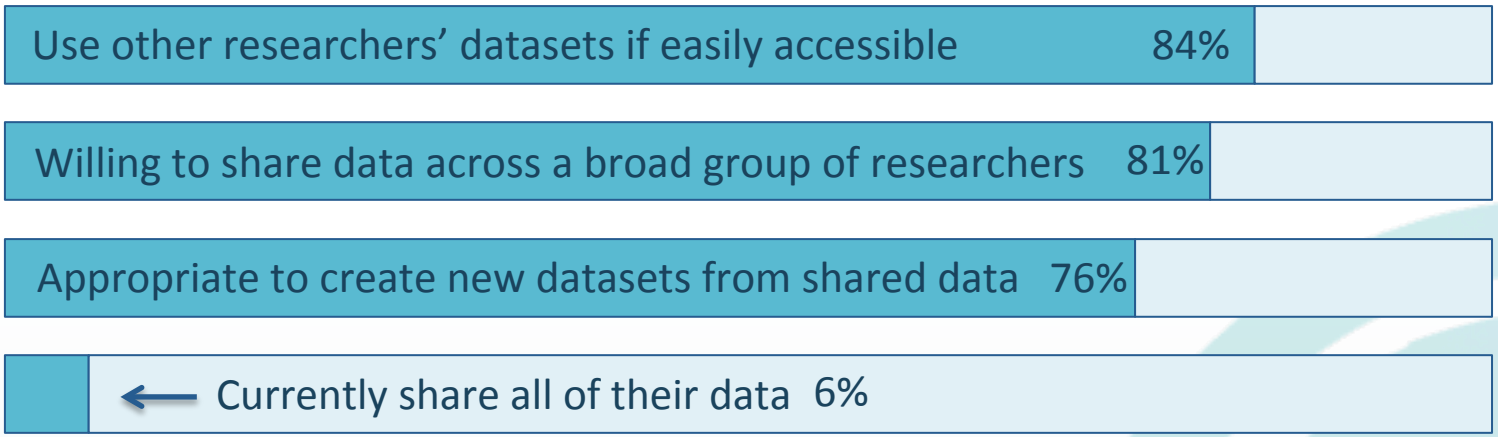
**1** School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, **2** University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, **3** Center for Biological Informatics, United States Geological Survey, Oak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

**Abstract**

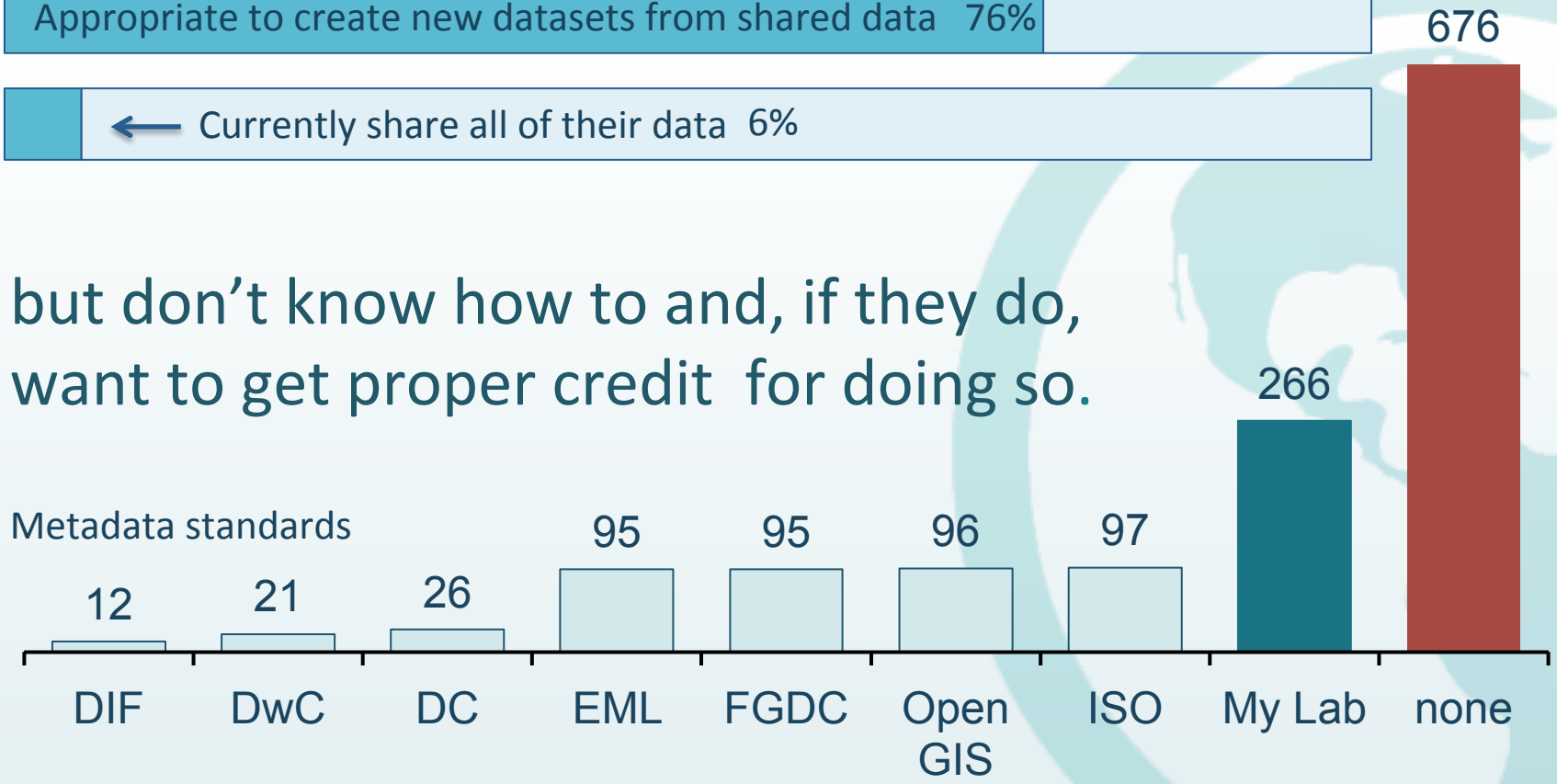
***Background:*** Scientific research in the 21st century is more data intensive and collaborative than in the past. It is important to study the data practices of researchers – data accessibility, discovery, re-use, preservation and, particularly, data sharing. Data sharing is a valuable part of the 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



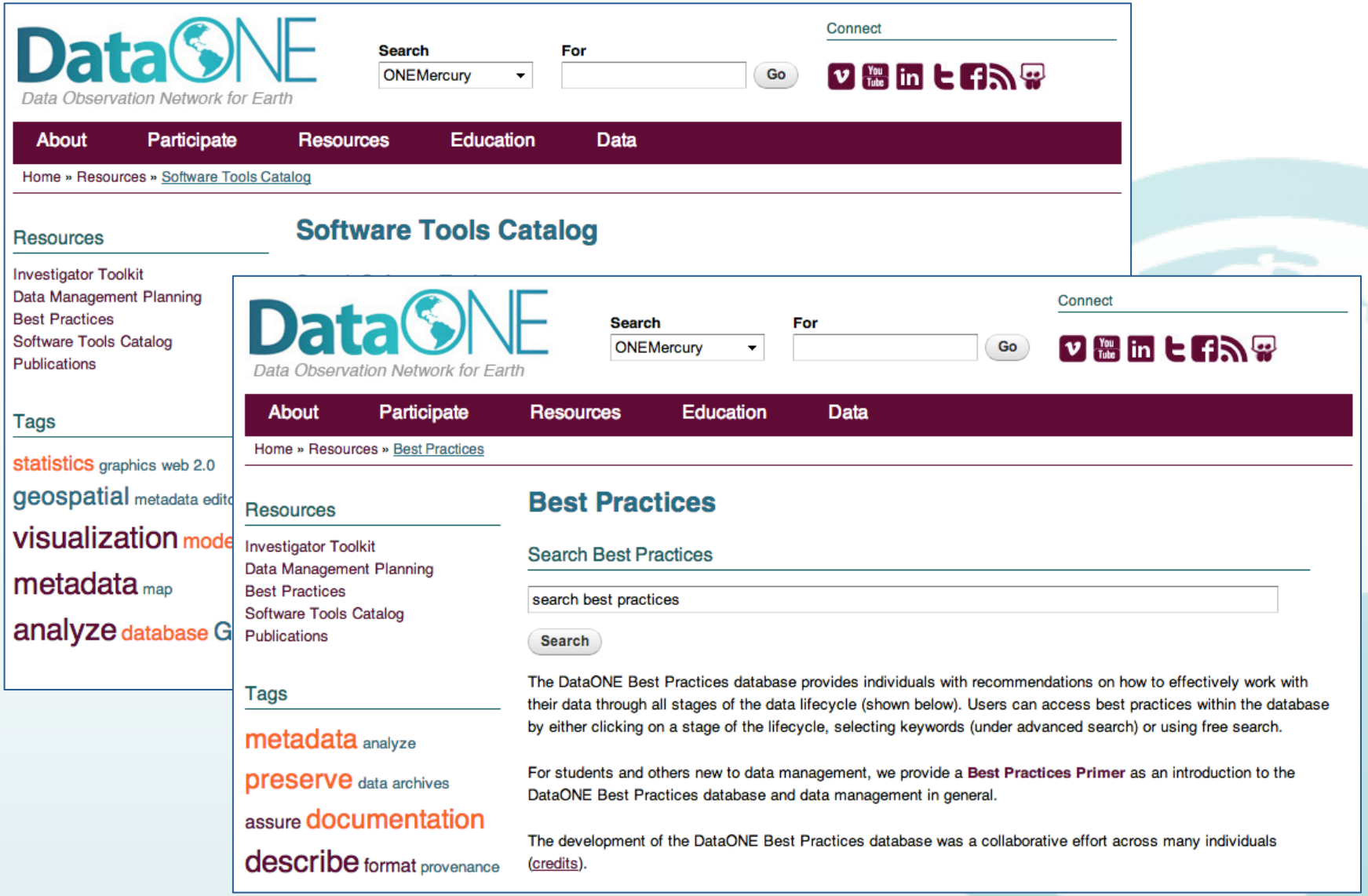

but don't know how to and, if they do, want to get proper credit for doing so.



# 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



# Making it easy to describe data

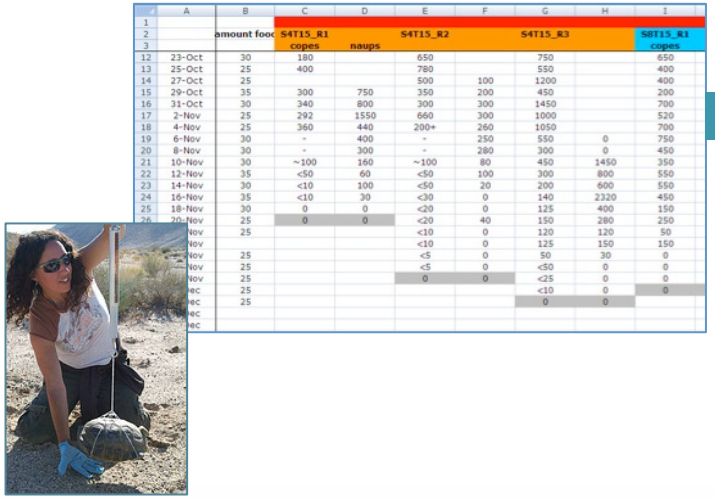
DATAUP

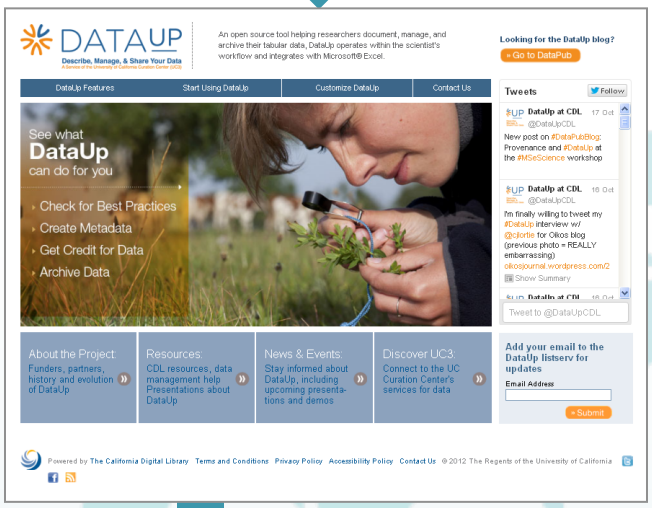
Intercept researchers where **they already work**

Microsoft Research

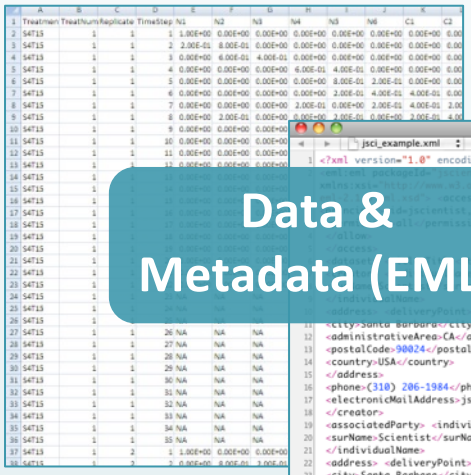
University of California
CDL
California Digital Library

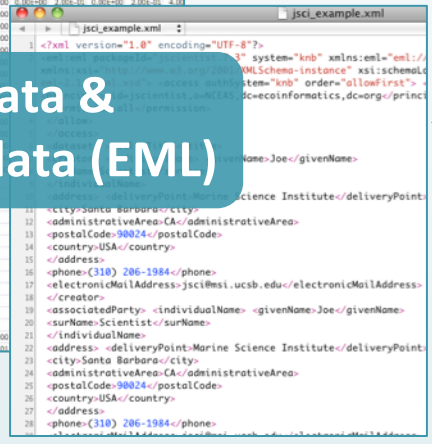
GORDON AND BETTY MOORE FOUNDATION



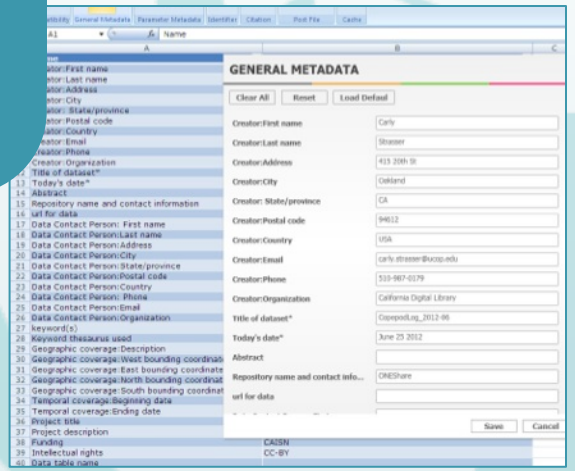



ONEShare

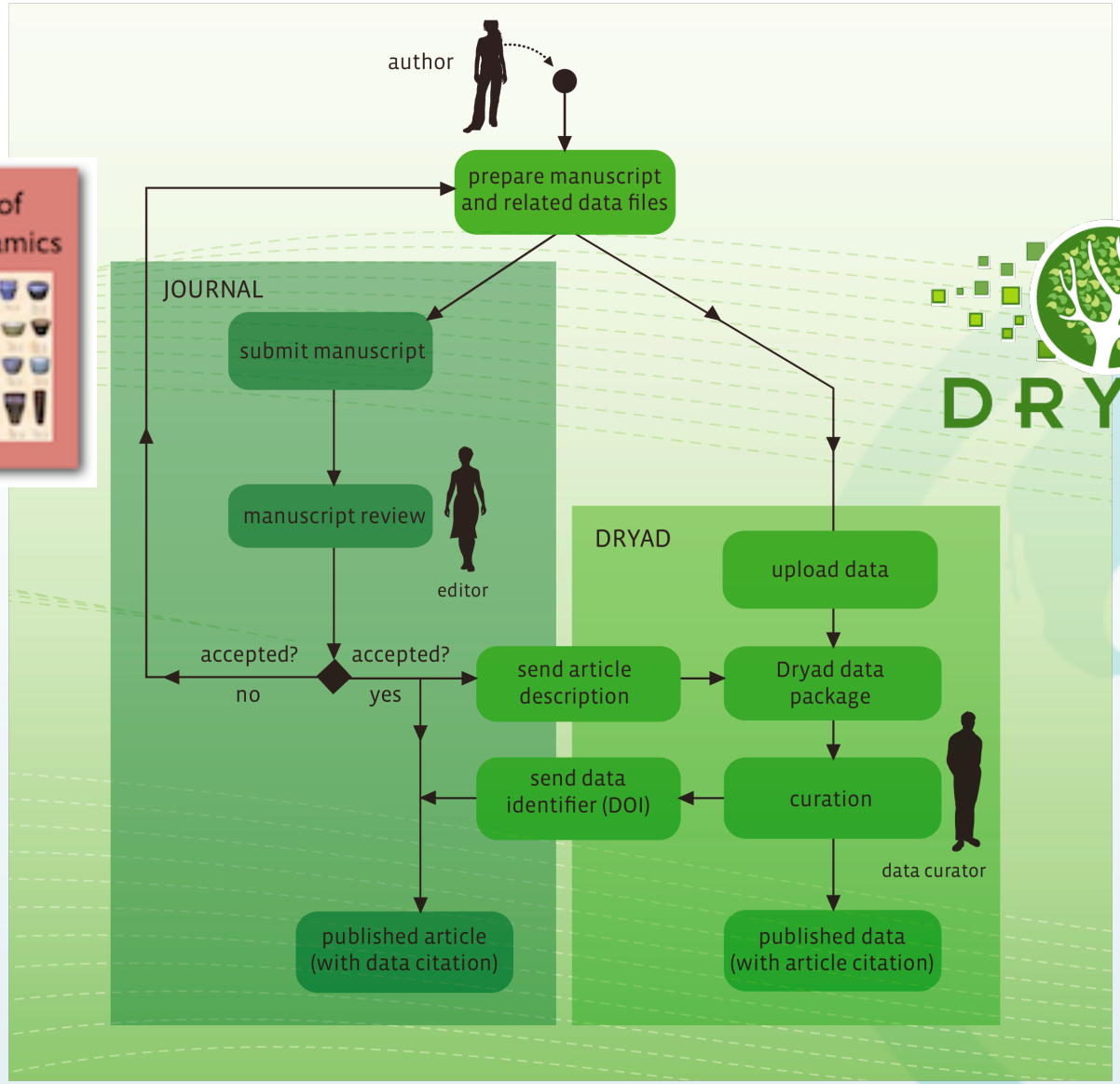





Data & Metadata (EML)



# Credit: Dryad repository for journal data & metadata



DRYAD

# 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

**Dryad data package**

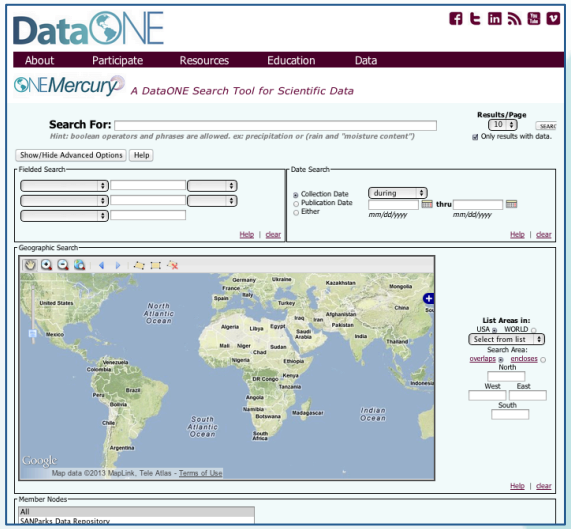
Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, Frazier M, Venter JC, Eisen JA (2011) Data from: Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in phylogenetic trees of phylogenetic marker genes. Dryad Digital Repository. doi:10.5061/dryad.8384

# Promoting data discovery

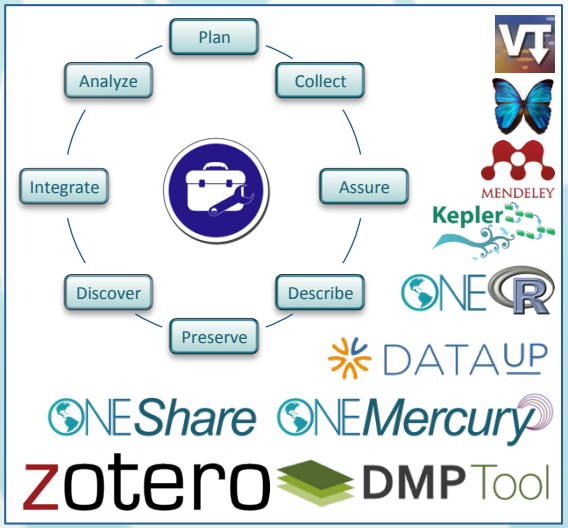
*Provide universal access to data about life on earth and the environment*

1. Building community

2. Developing sustainable data discovery and interoperability solutions

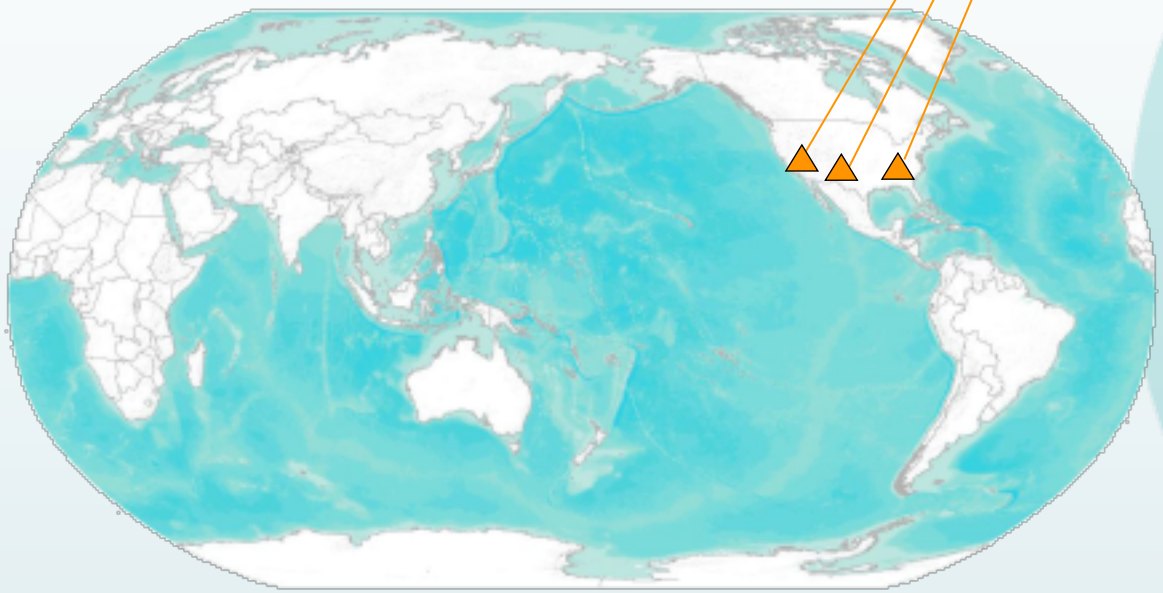



3. Enabling science through tools and services



# 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

THE UNIVERSITY of NEW MEXICO

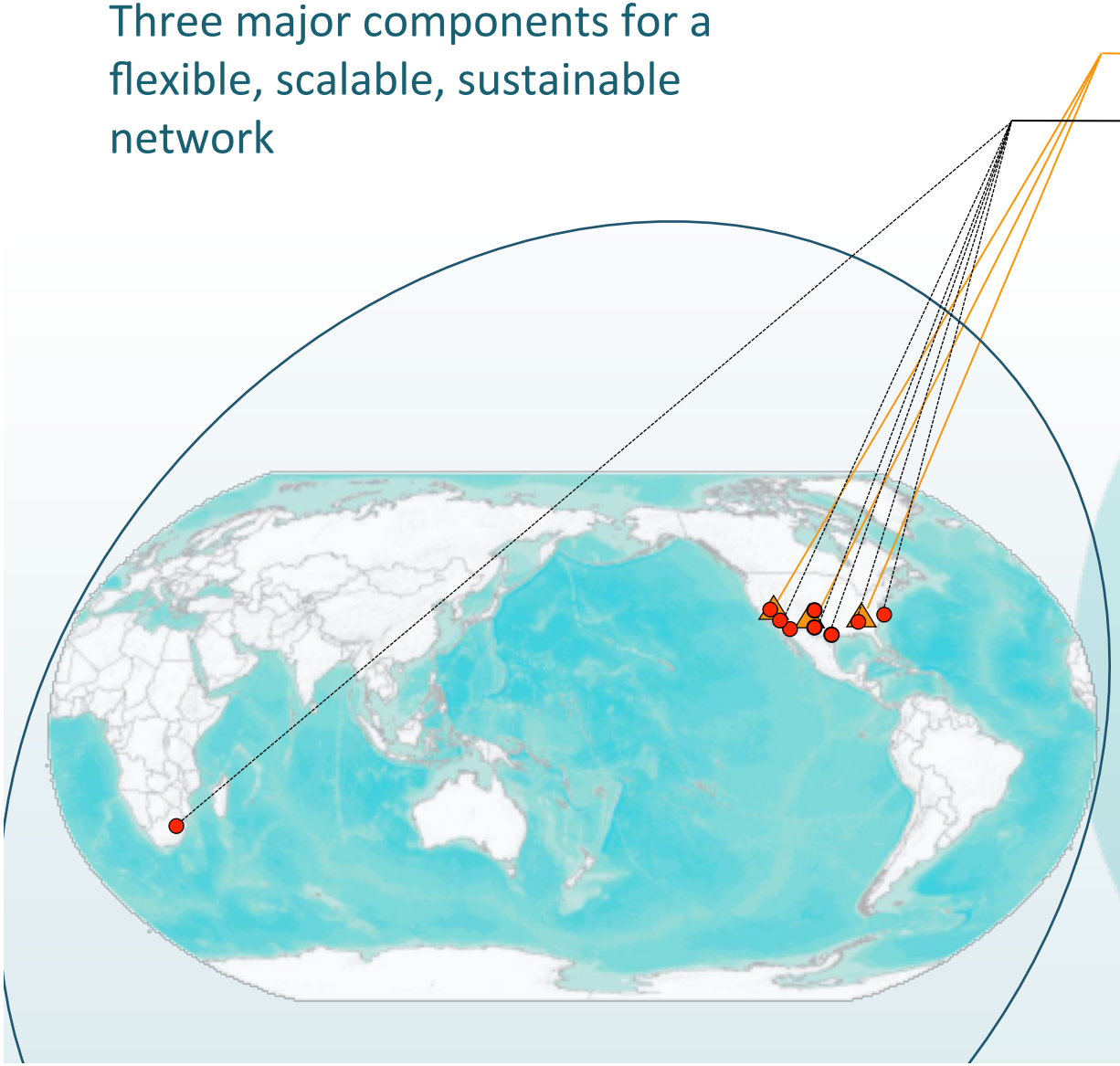
OAK RIDGE National Laboratory

NCEAS

UCSB

# DataONE

Three major components for a flexible, scalable, sustainable network



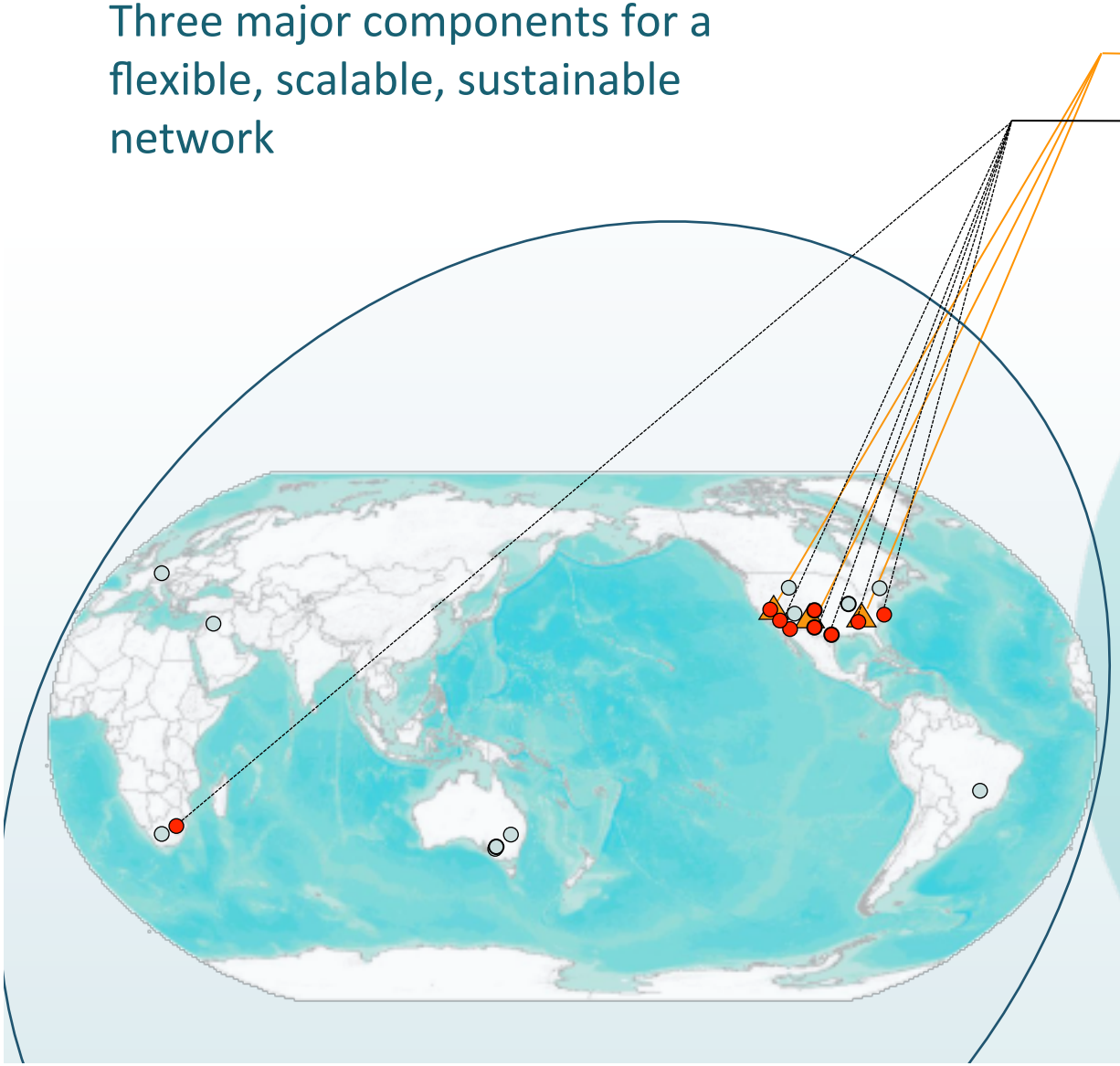
**Coordinating Nodes**

**Member Nodes**

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

# DataONE

Three major components for a flexible, scalable, sustainable network



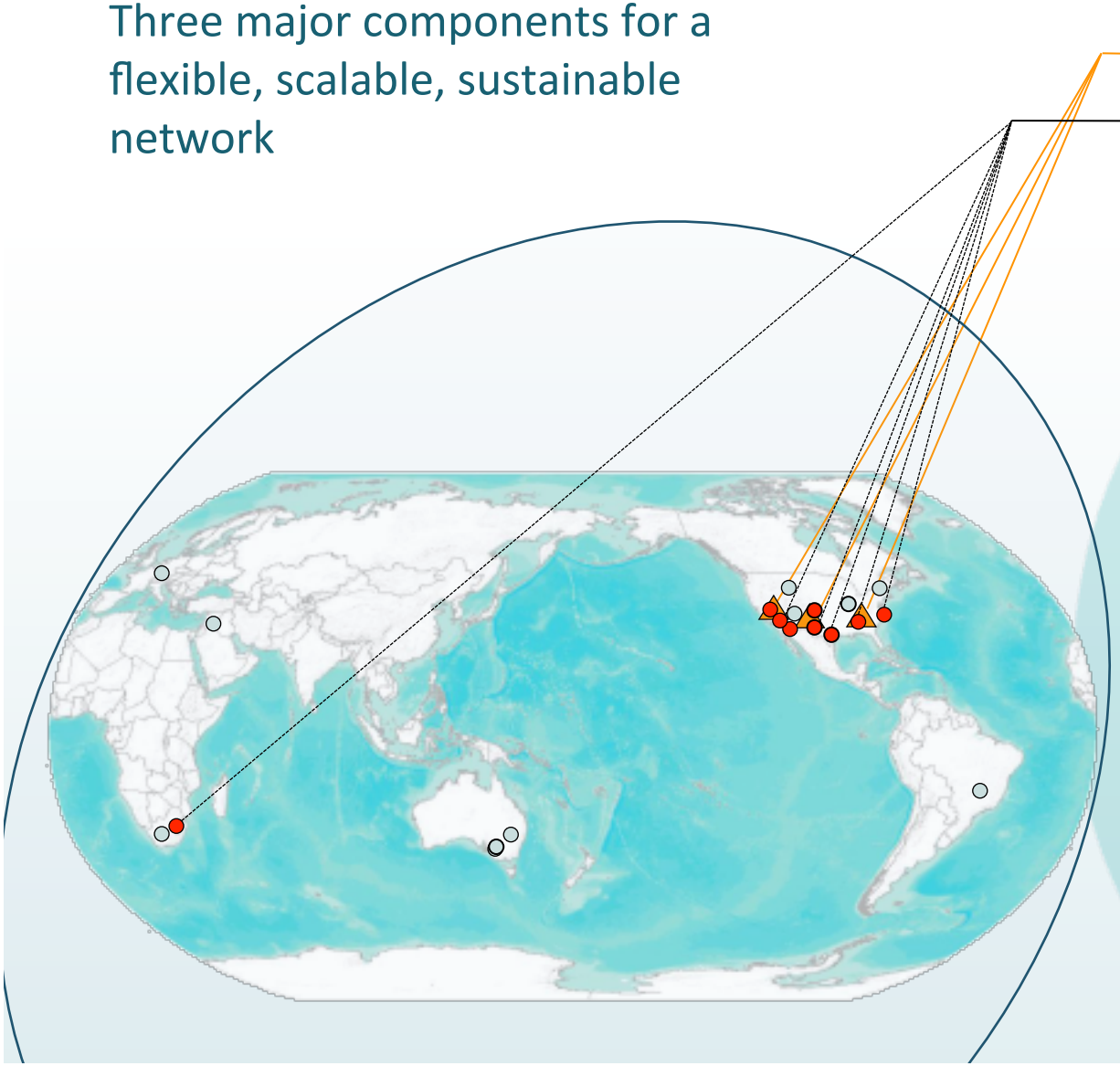
**Coordinating Nodes**

**Member Nodes**

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

# DataONE

Three major components for a flexible, scalable, sustainable network



**Coordinating Nodes**

**Member Nodes**

**Investigator Toolkit**

```
>> command line interface
```

python

Java

ONEDrive

DMPTool

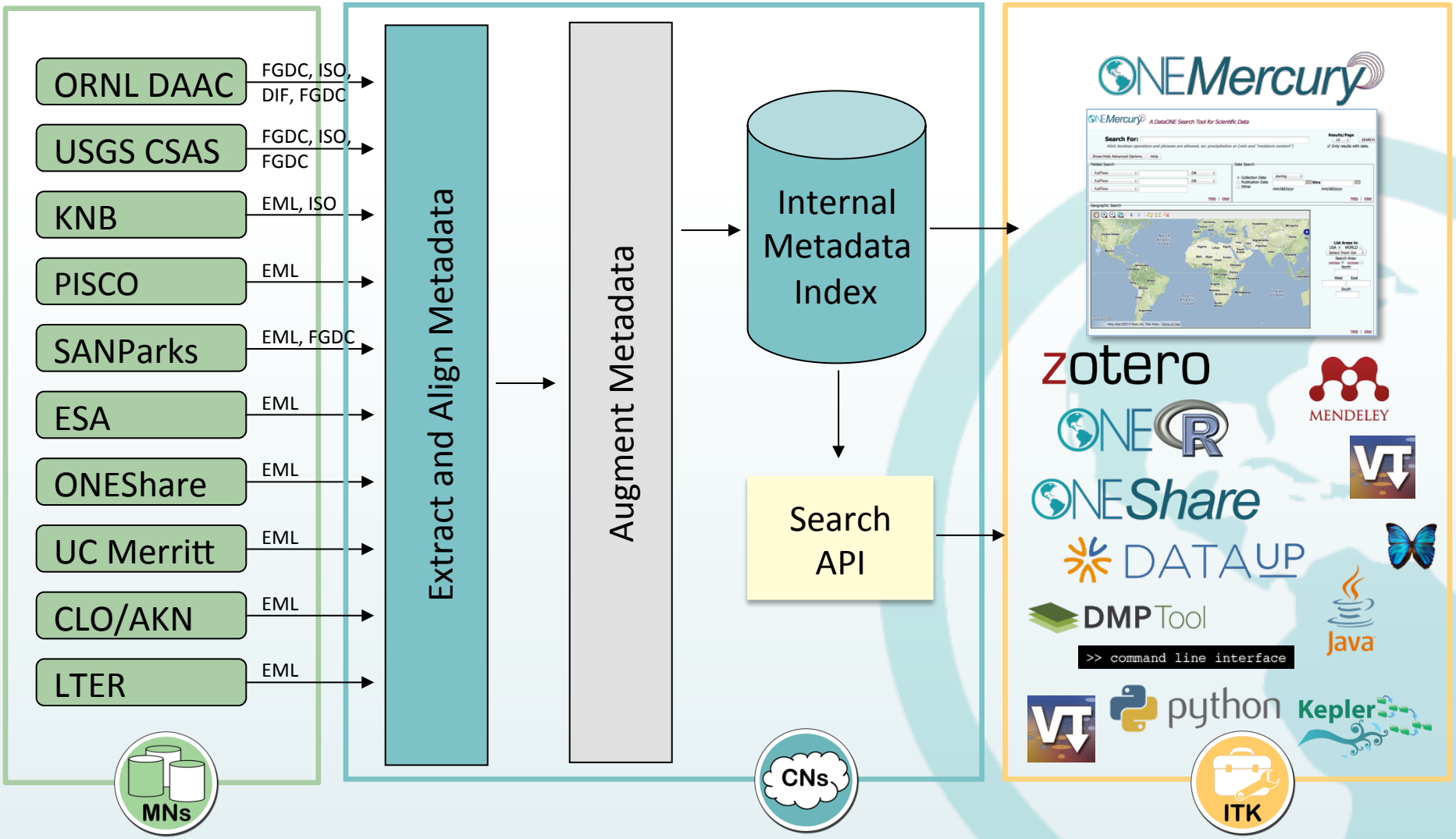
ONEMercury

DATAUP

Kepler

ONE R

# DataONE: Enabling data discovery



Information Center for the Environment (**ICE**) UC Davis

agencies

citizens

faculty

ICE Users

DataONE

NESCent

CDL California Digital Library

neon National Ecological Observatory Network, Inc.

USGS science for a changing world

NSF

**Investigator Toolkit**

ONEDrive VT python

Java DATAUP ONE R

Kepler ONEMercury

DMPTool >> command line interface

UC3 Merritt Member Node

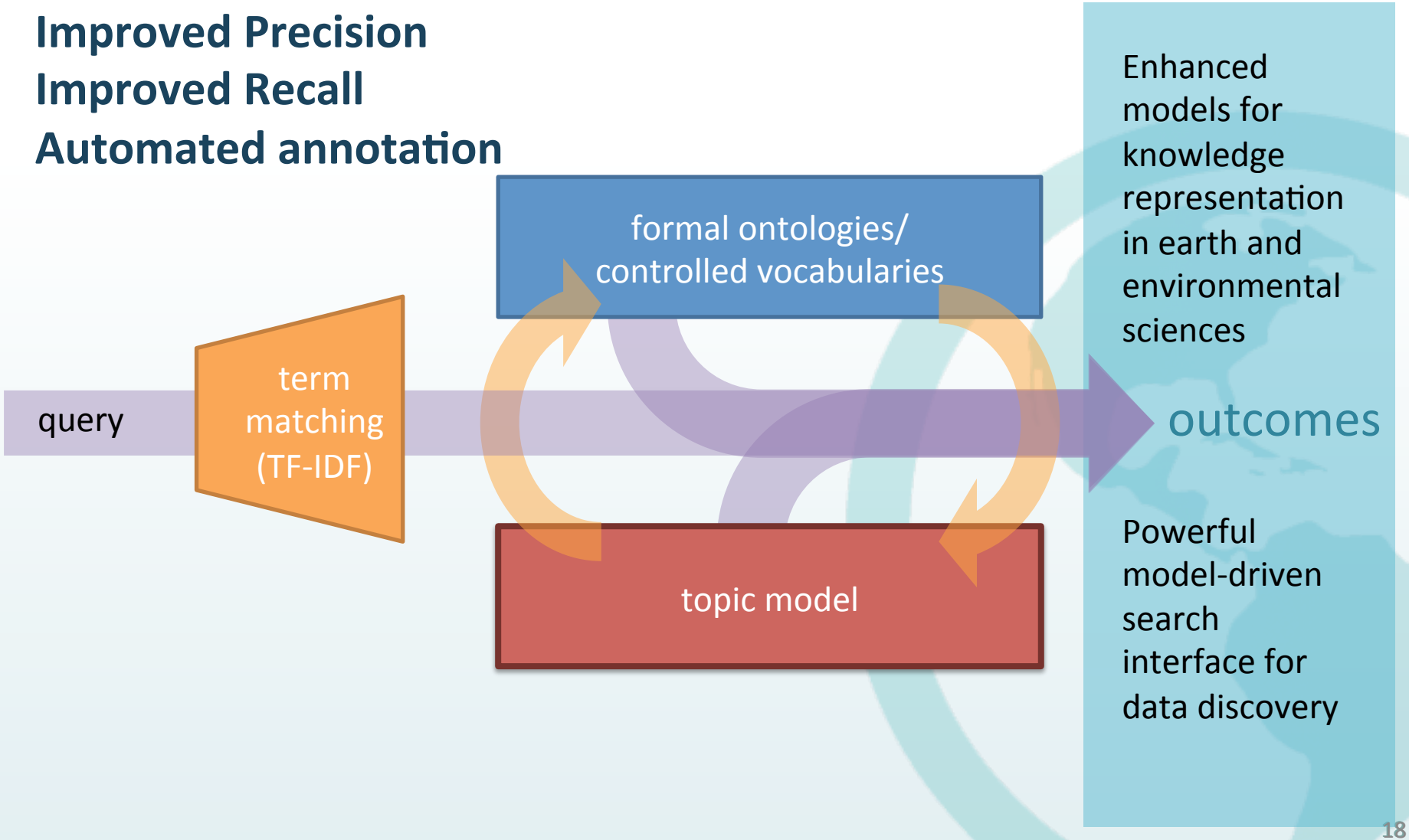
DataONE Users

# Some remaining challenges

- Semantic mediation
- Provenance
- Improving metadata quality over time

# Powerful Data Discovery via Semantics

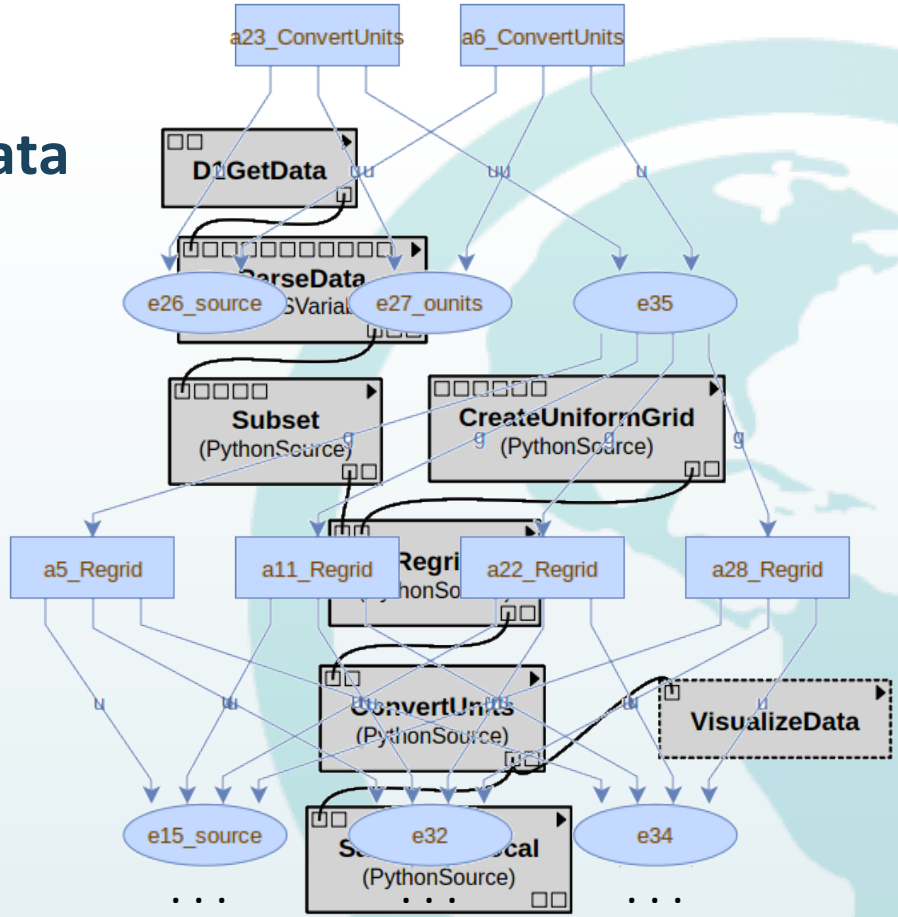
**Improved Precision**
**Improved Recall**
**Automated annotation**



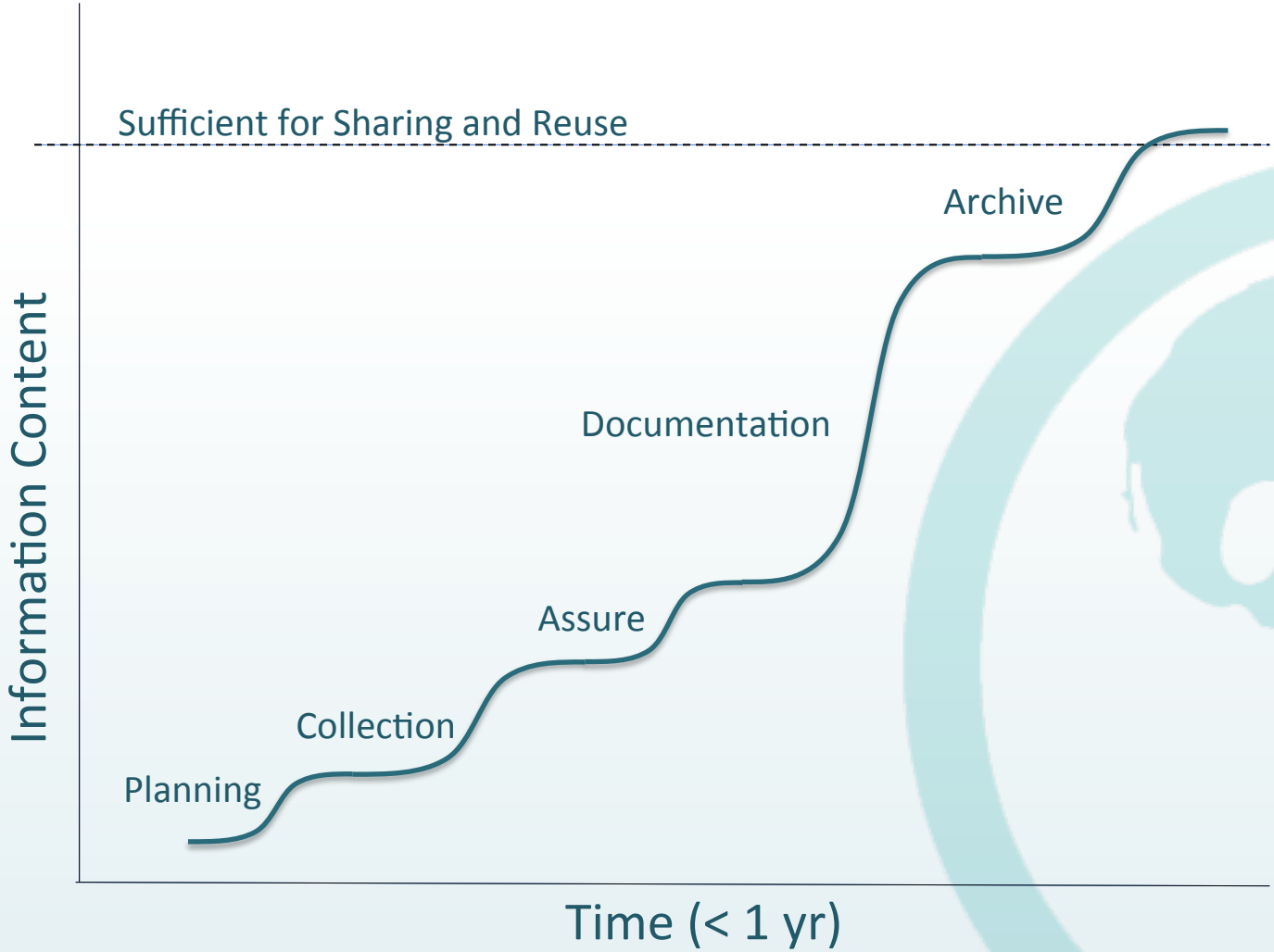
# Provenance

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

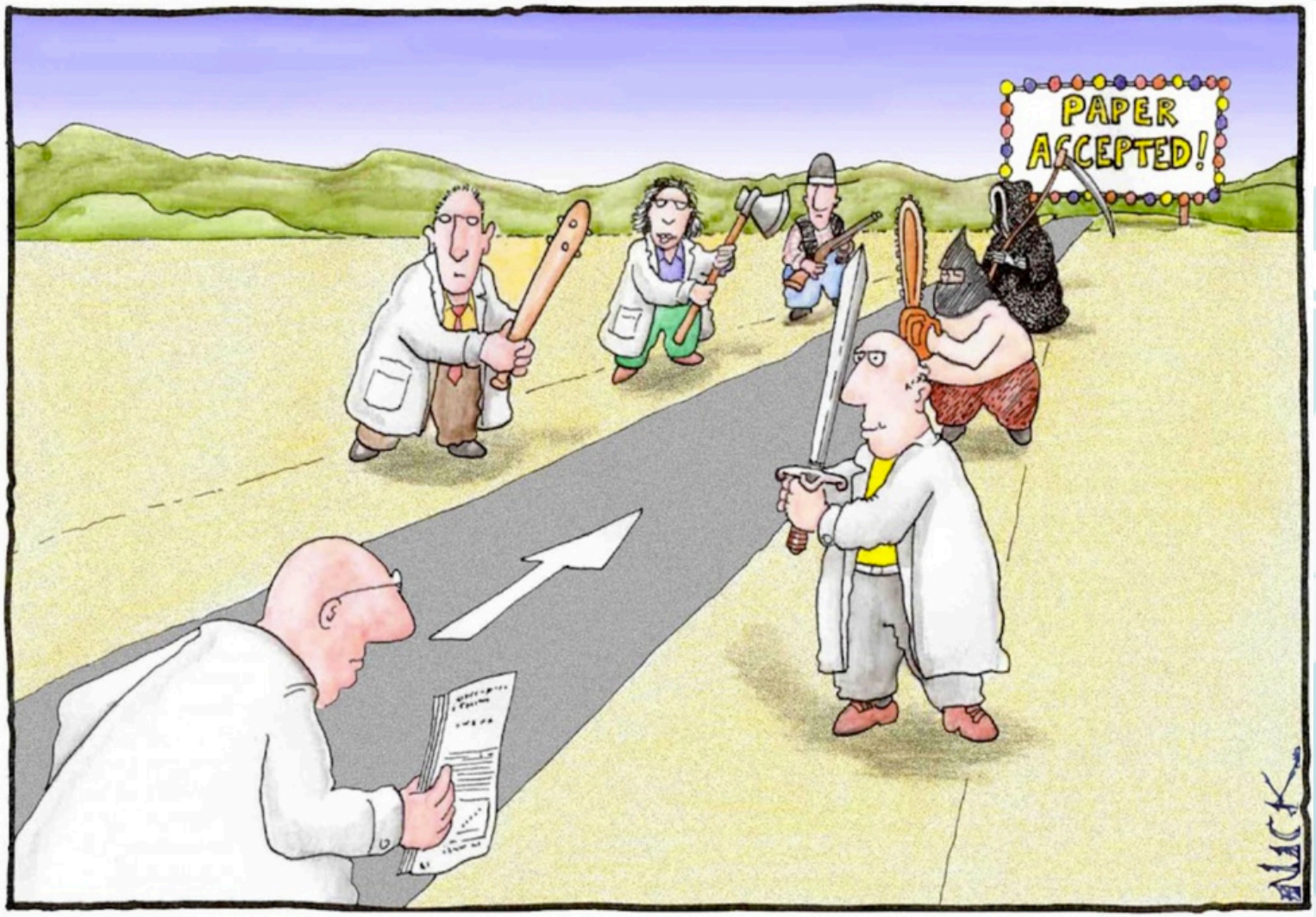
- Record **processing history**, **data lineage**
  - *dependency graph*
- W3C standard: PROV
- DataONE Extension: **D-PROV**
  - Workflow provenance
  - System agnostic!



# 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**