

# DataONE - Enabling Long-Term Archive and Reuse of Data for the Earth Sciences

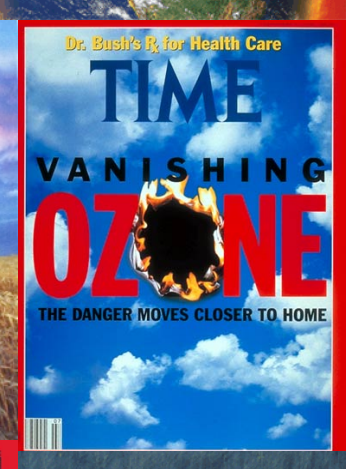
Botany 2014

Dave Vieglais

University of Kansas





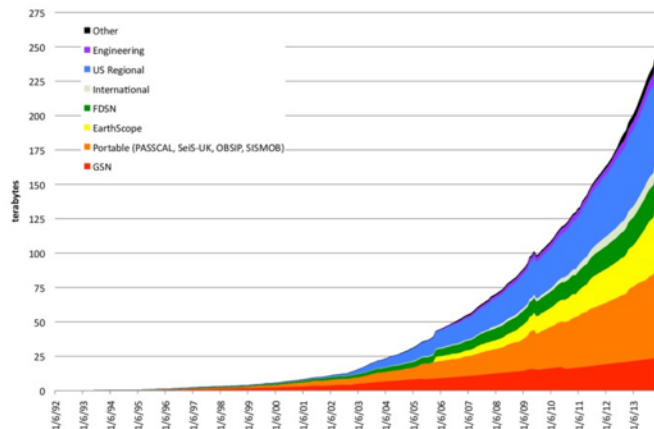




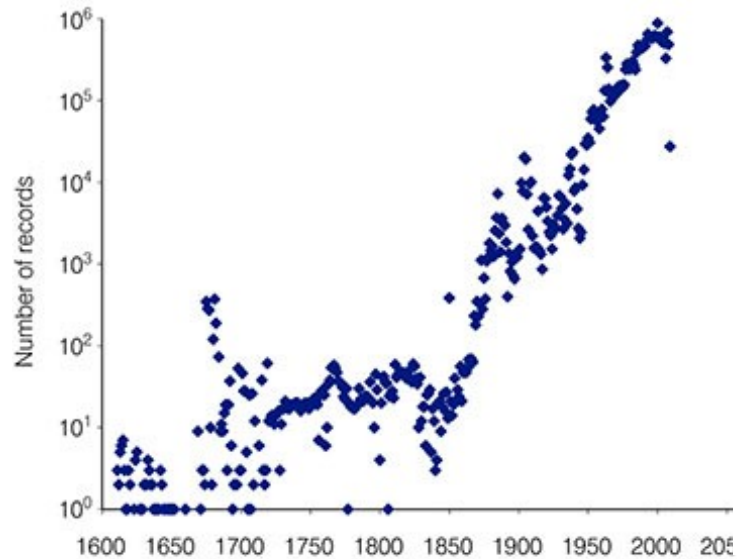


# Increasing Digital Data Resources

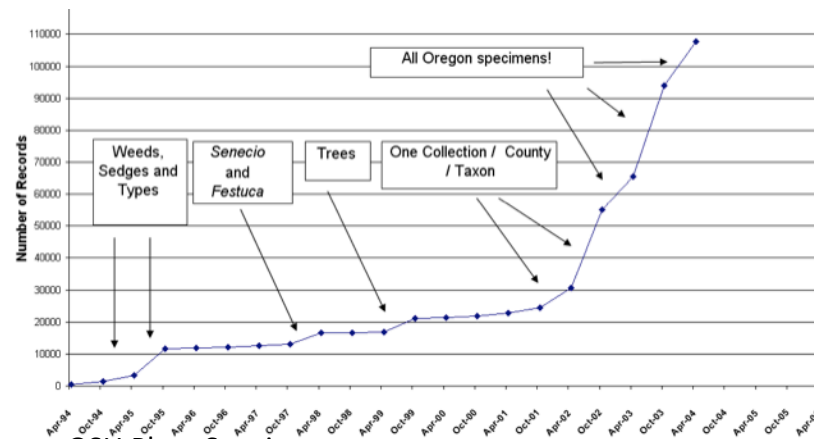
IRIS DMC Archive  
as of 1 Nov 2013



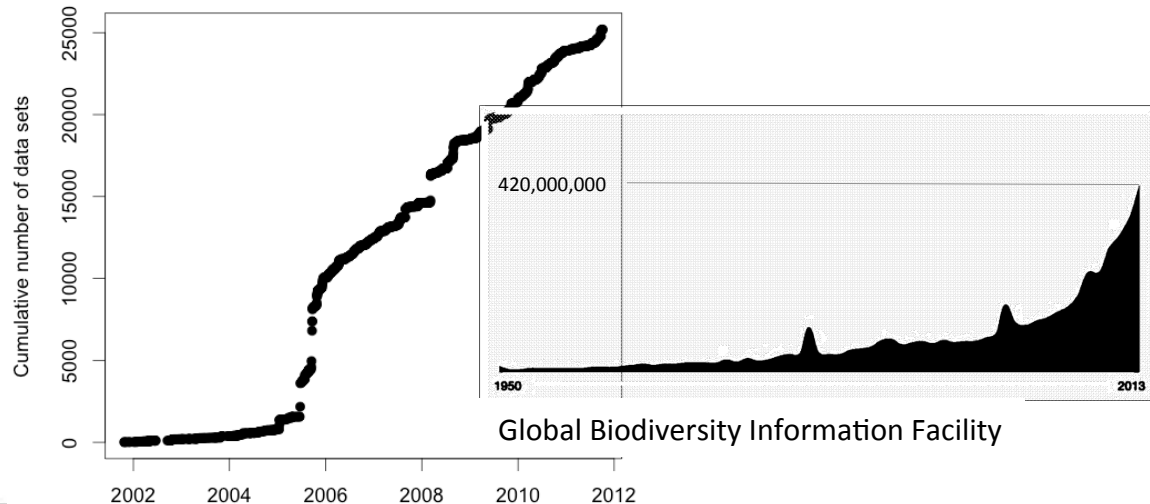
From: Incorporated Research Institutions for Seismology



Ocean-Biogeographic Information System



OSU Plant Specimens



Global Biodiversity Information Facility

Knowledge Network for Biocomplexity

EXECUTIVE OFFICE OF THE PRESIDENT  
OFFICE OF SCIENCE AND TECHNOLOGY POLICY  
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren  
Director

**SUBJECT:** Increasing Access to the Results of Federally Funded Scientific Research

## 1. Policy Principles

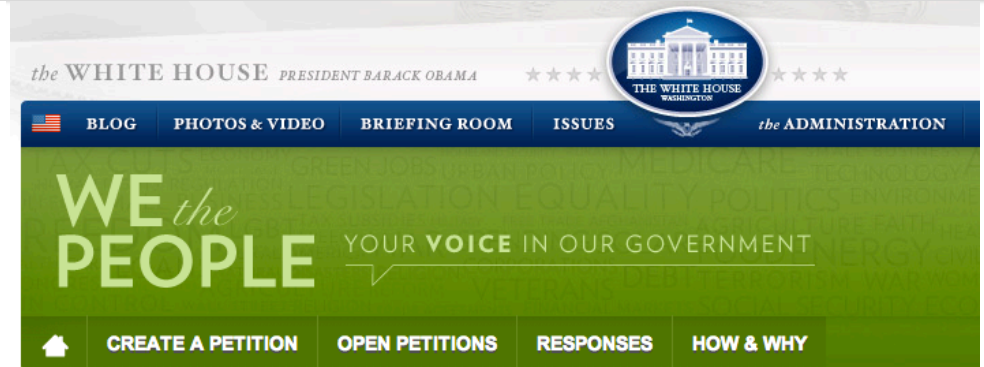
The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather

To that end, I have [issued a memorandum today \(.pdf\)](#) to Federal agencies that directs those with more than \$100 million in research and development expenditures to develop plans to make the results of federally-funded research publicly available free of charge within 12 months after original publication.

...the memorandum requires that agencies start to address the need to improve upon the management and sharing of scientific data produced with Federal funding.



OFFICIAL OFFICE OF SCIENCE AND TECHNOLOGY POLICY RESPONSE TO

Require free access over the Internet to scientific journal articles arising from taxpayer-funded research.

## Increasing Public Access to the Results of Scientific Research

By Dr. John Holdren

Thank you for [your participation](#) in the We the People platform. The Obama Administration agrees that citizens deserve easy access to the results of research their tax dollars have paid for. As you may



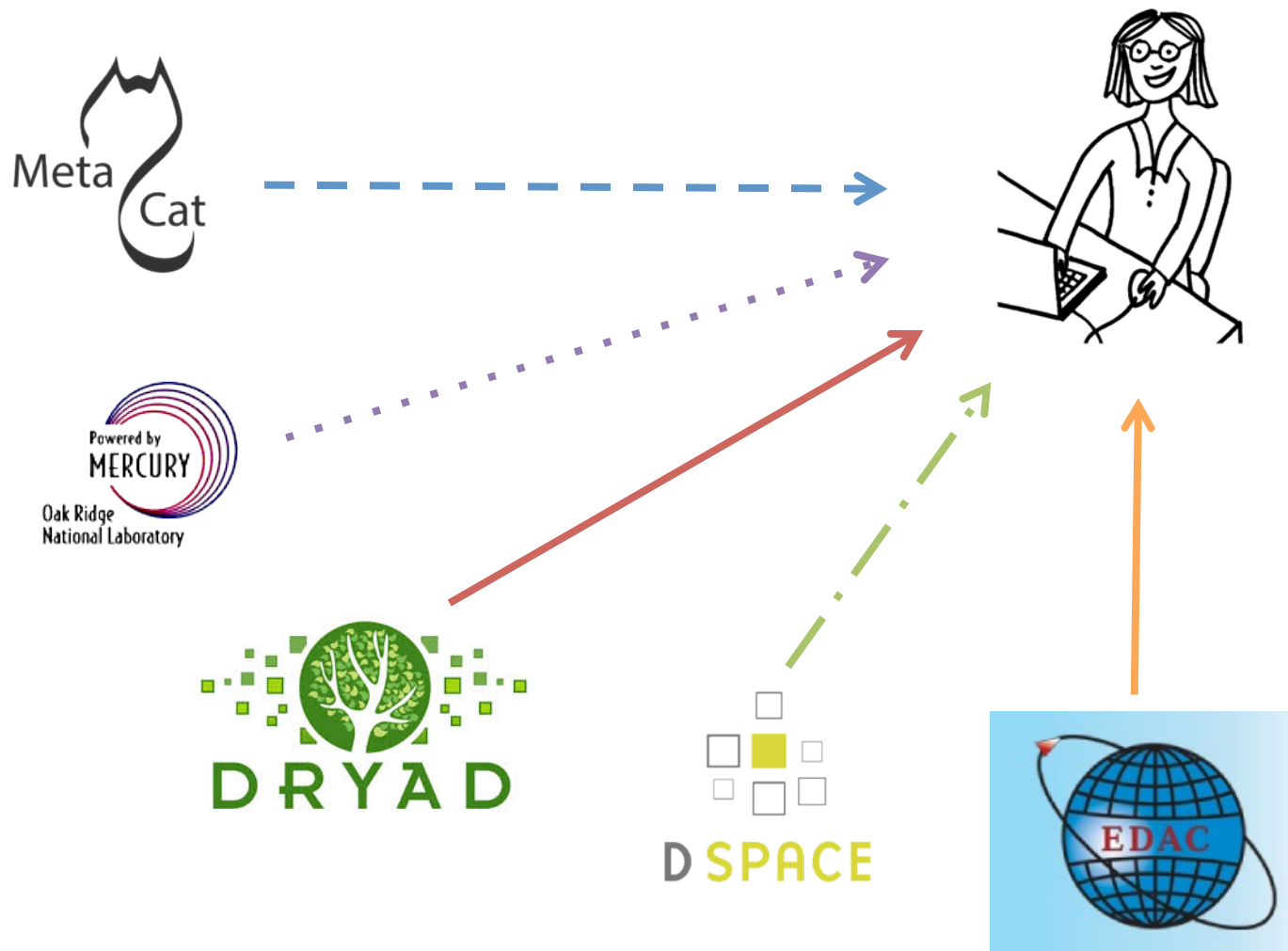


# Many Repository Solutions





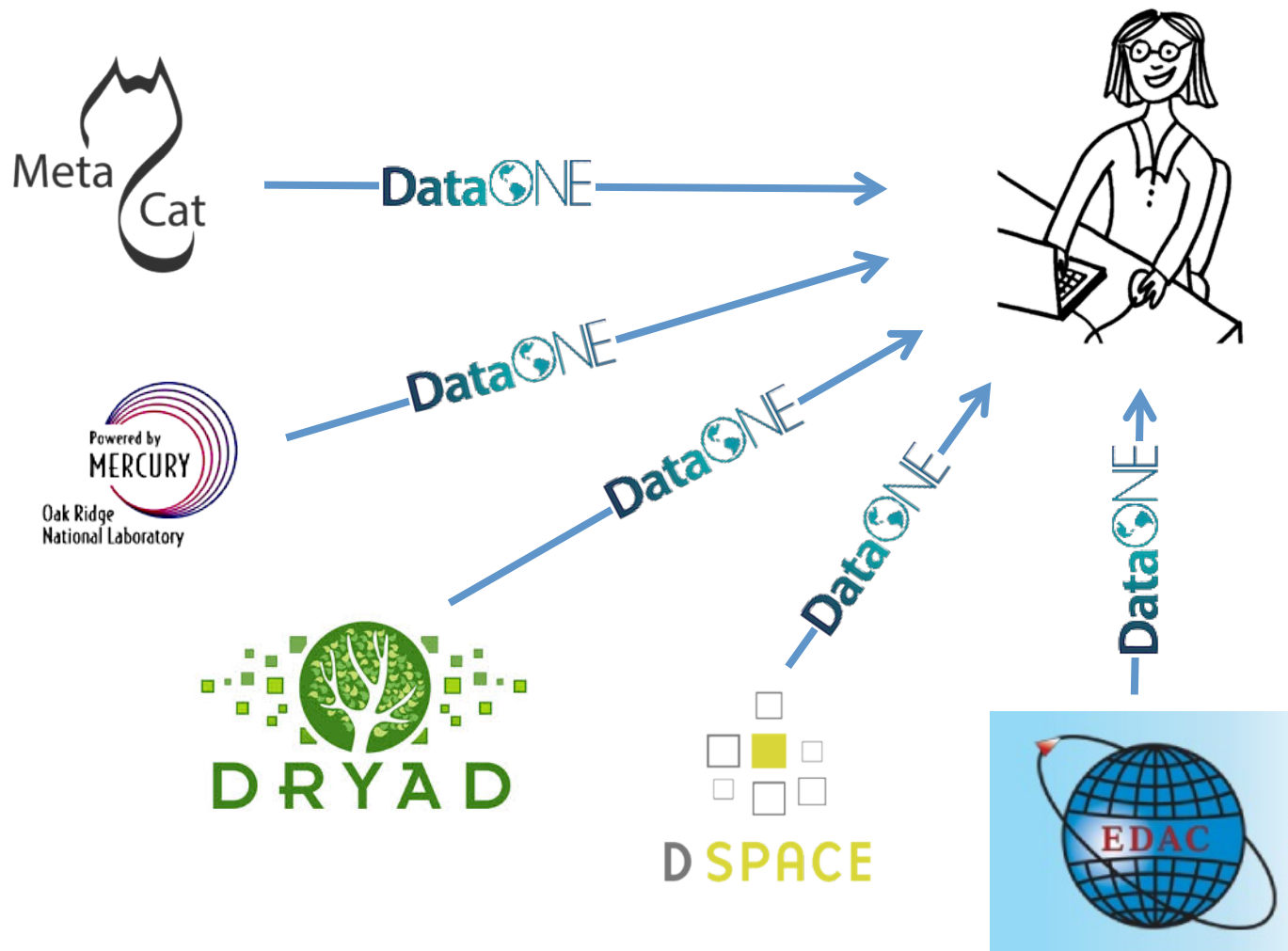
# Diversity Can Be Challenging



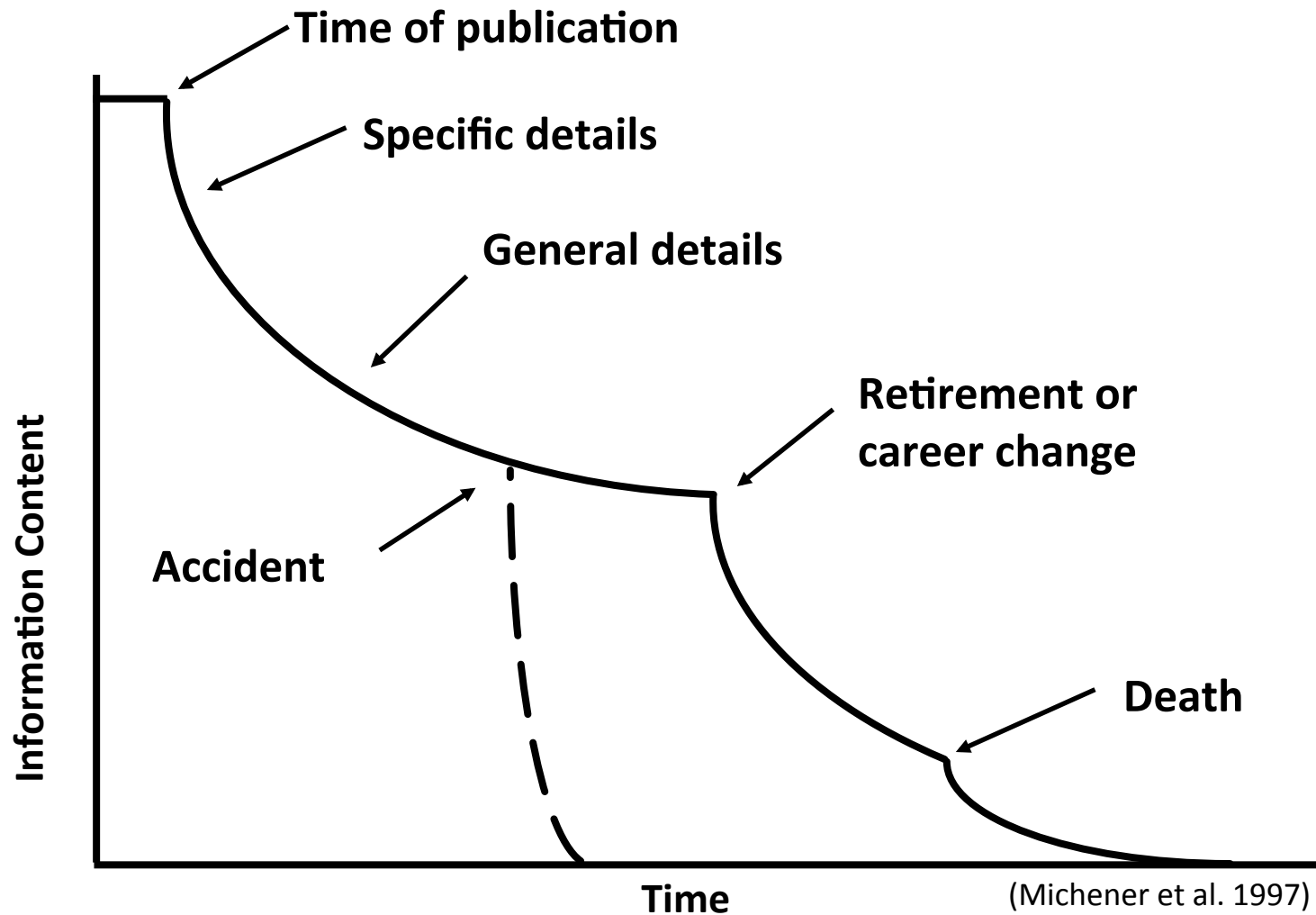




# DataONE Enables Consistency



# Data entropy







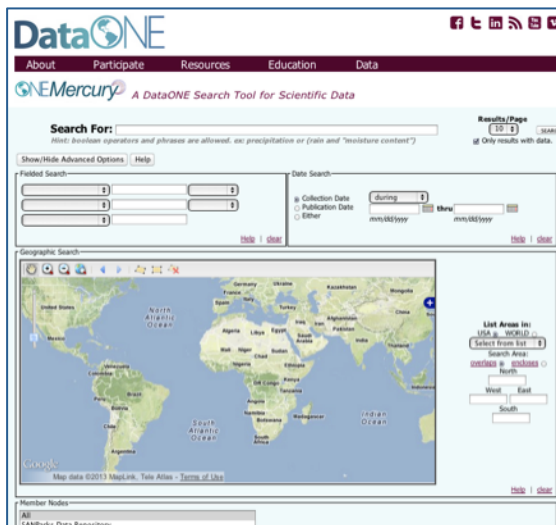
# The DataONE Vision and Approach

*Providing universal access to data about life on earth and the environment that sustains it, as well as the tools needed by researchers*

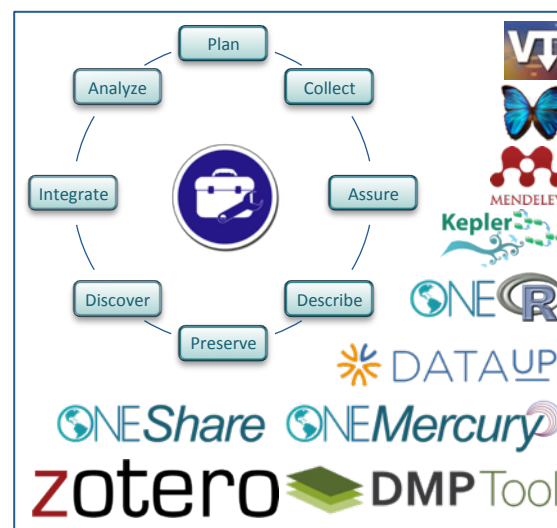
## 1. Building community



## 2. Developing sustainable data discovery and interoperability solutions



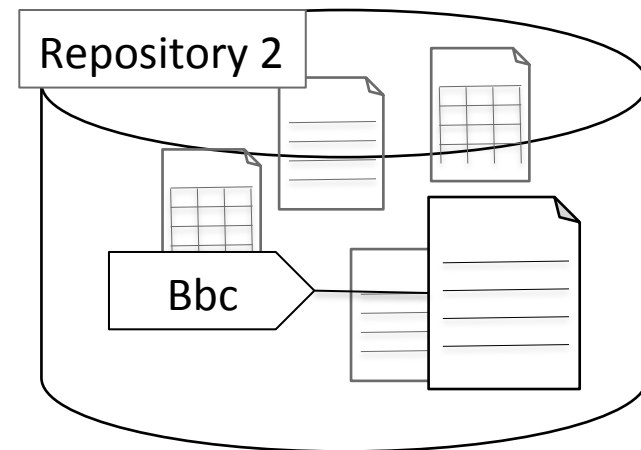
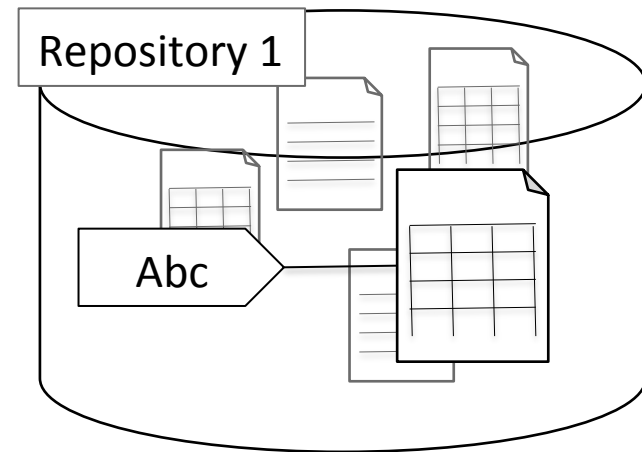
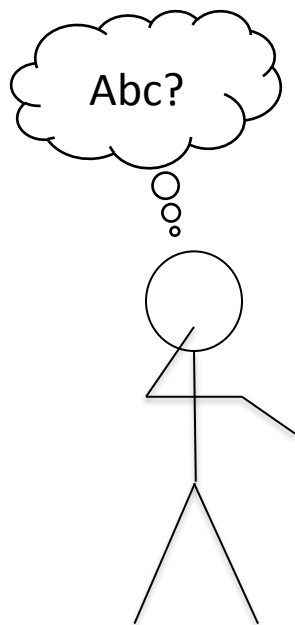
## 3. Enabling science through tools and services





# Data Access Infrastructure

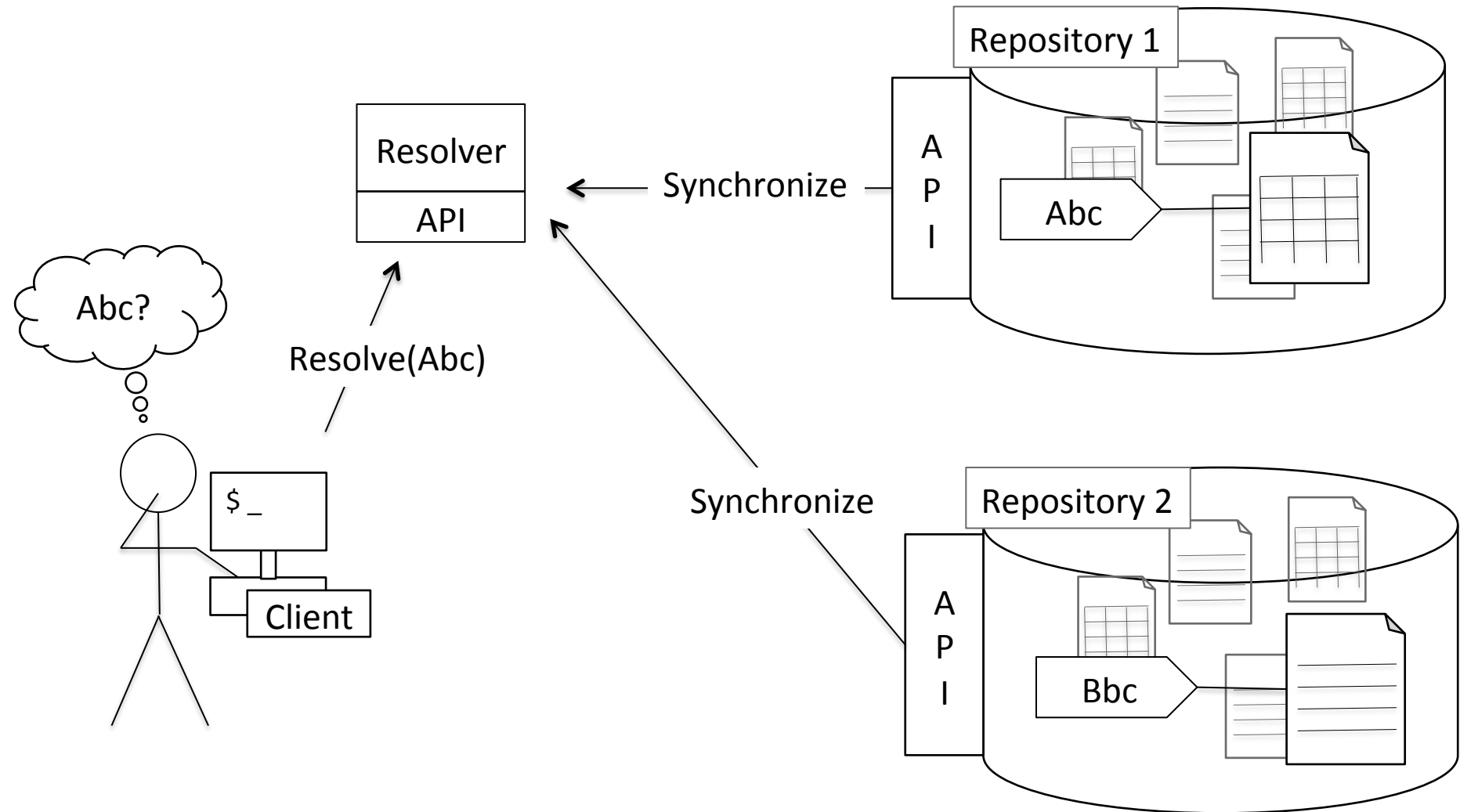
Where is the content?  
How do I retrieve it?  
Did it change since cited?  
Did someone delete it?





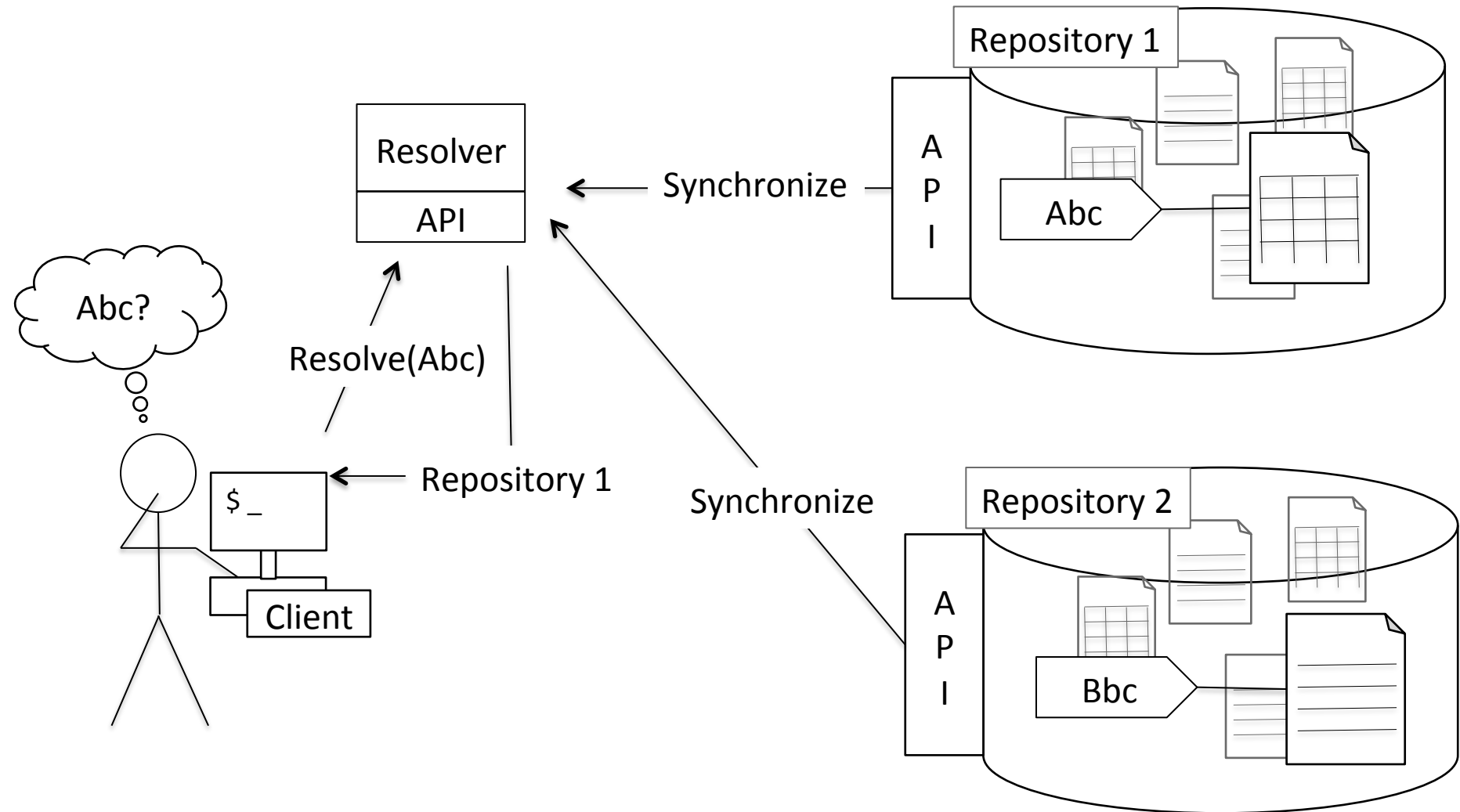


# Resolve process



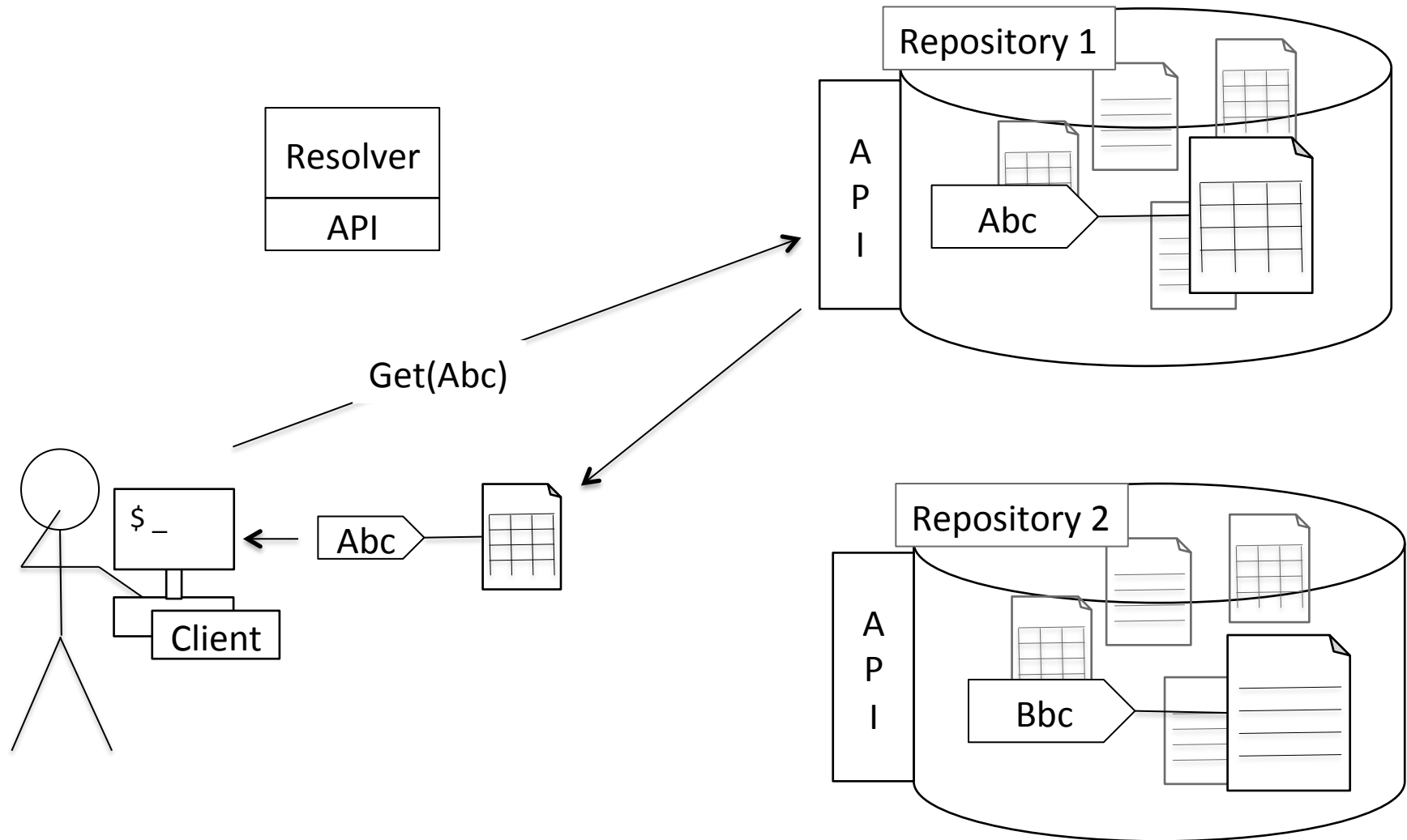


# Resolve cont'd



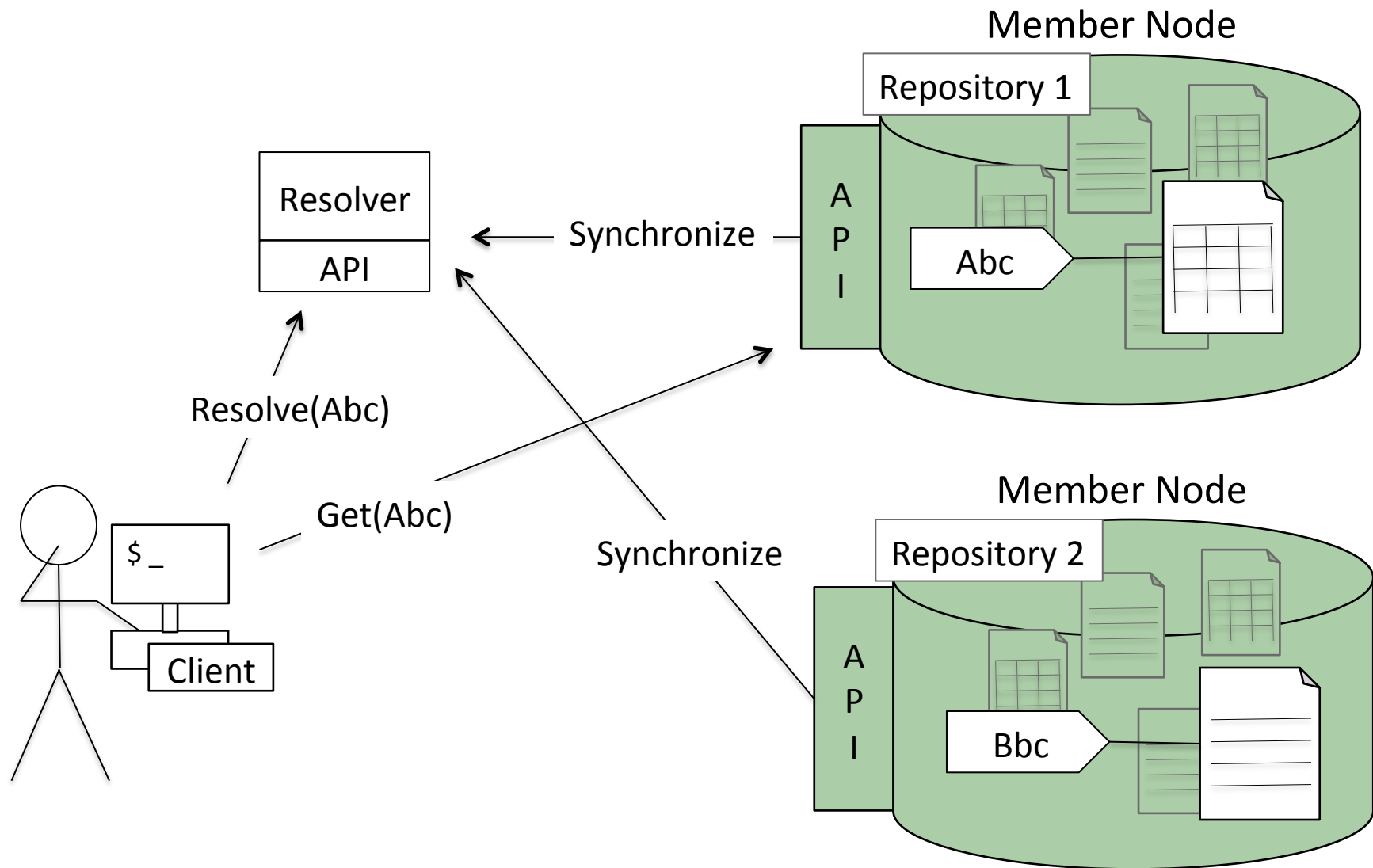


# Content Retrieval





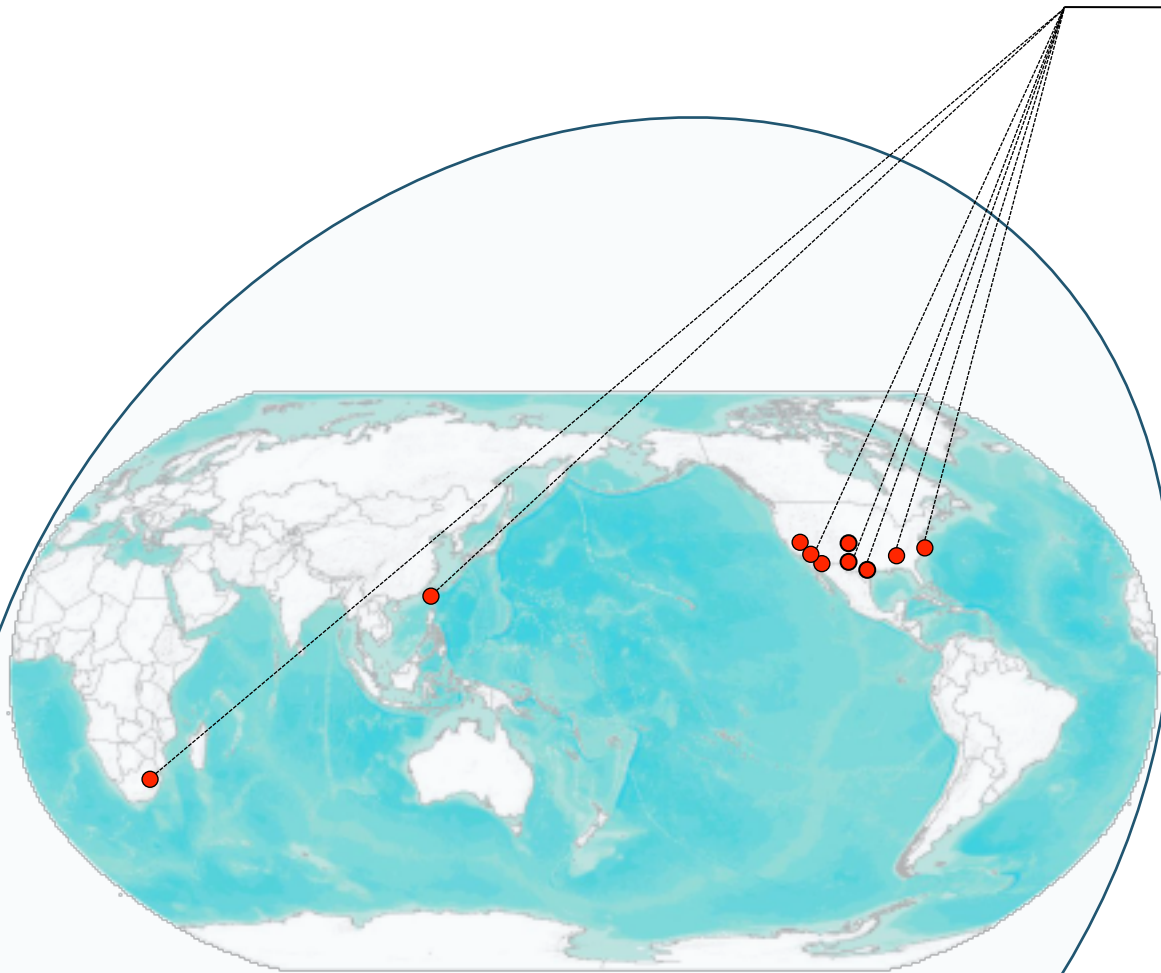
# DataONE – Member Nodes





# Member Nodes

The data curators and providers



## Member Nodes

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



The**Cornell**Lab 



UC3Merritt



PISCO



行政院農業委員會 林業試驗所  
TAIWAN FORESTRY RESEARCH INSTITUTE



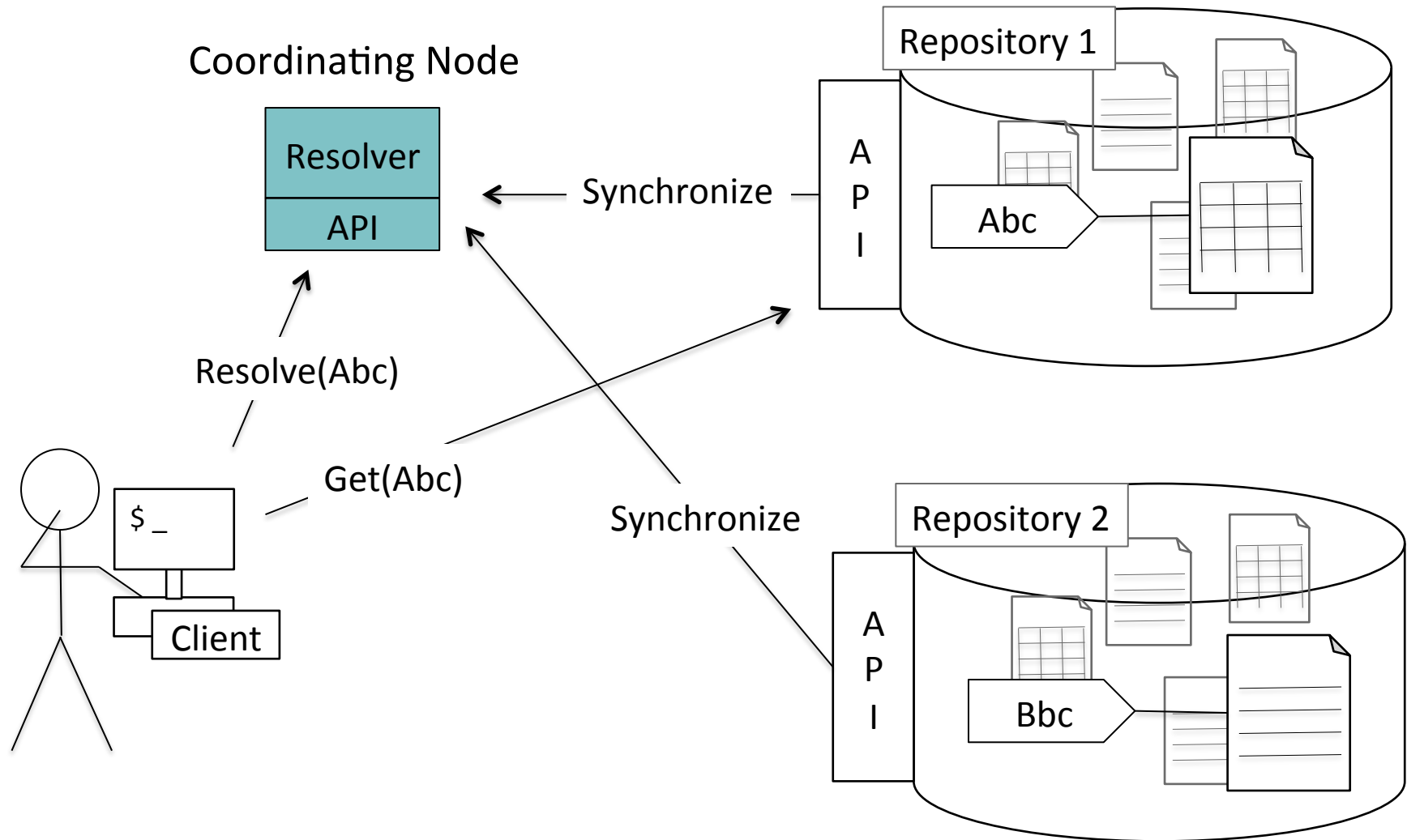
SEAD | Sustainable Environment  
Actionable Data



USA npn  
National Phenology Network



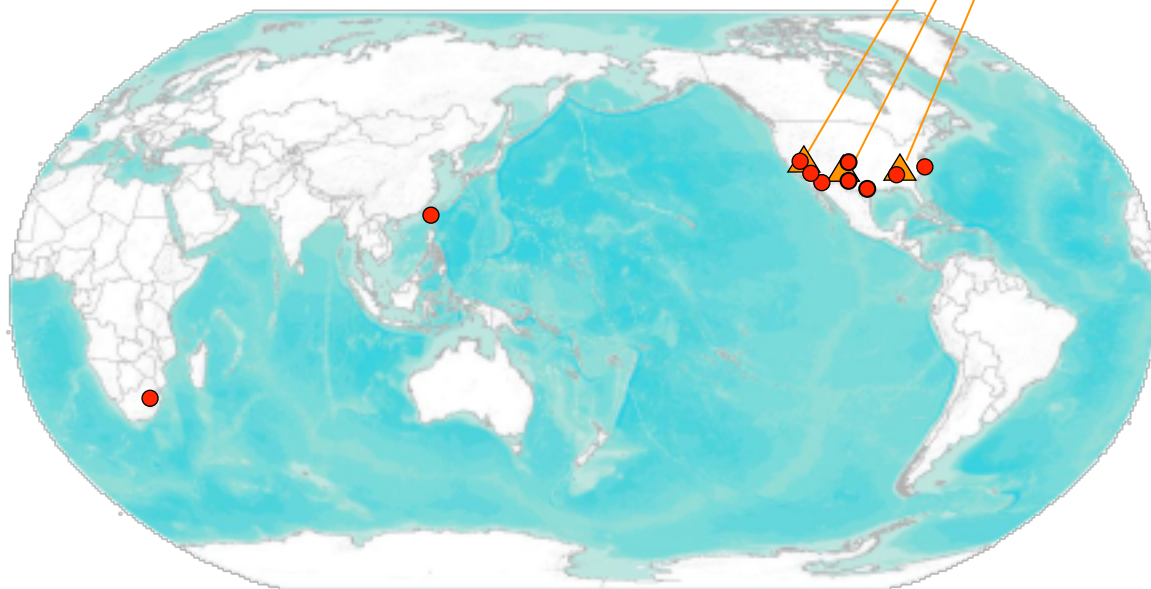
# DataONE – Coordinating Node





# Coordinating Nodes

Core services facilitating access to and management of data held by Member Nodes

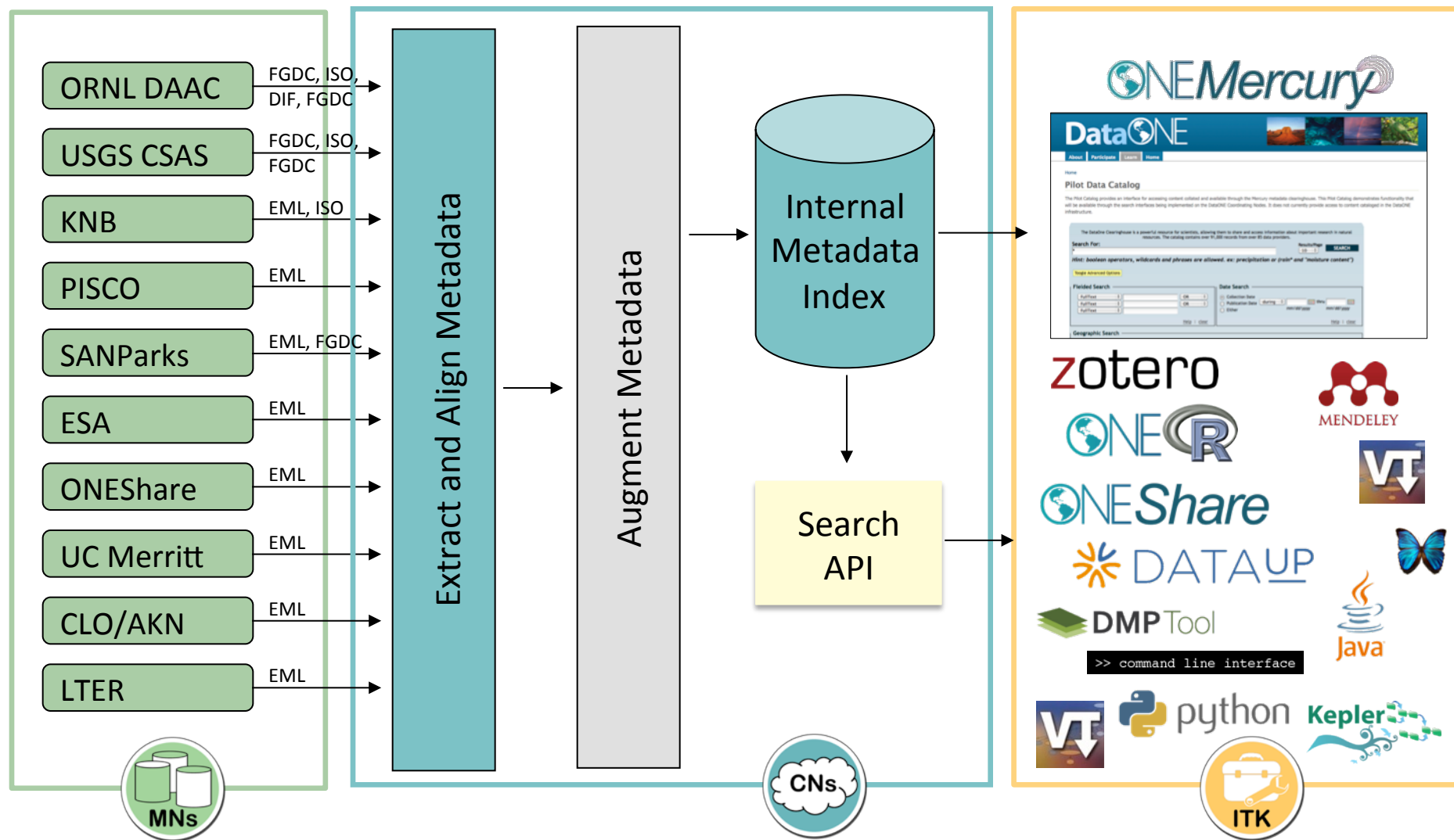


## Coordinating Nodes

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



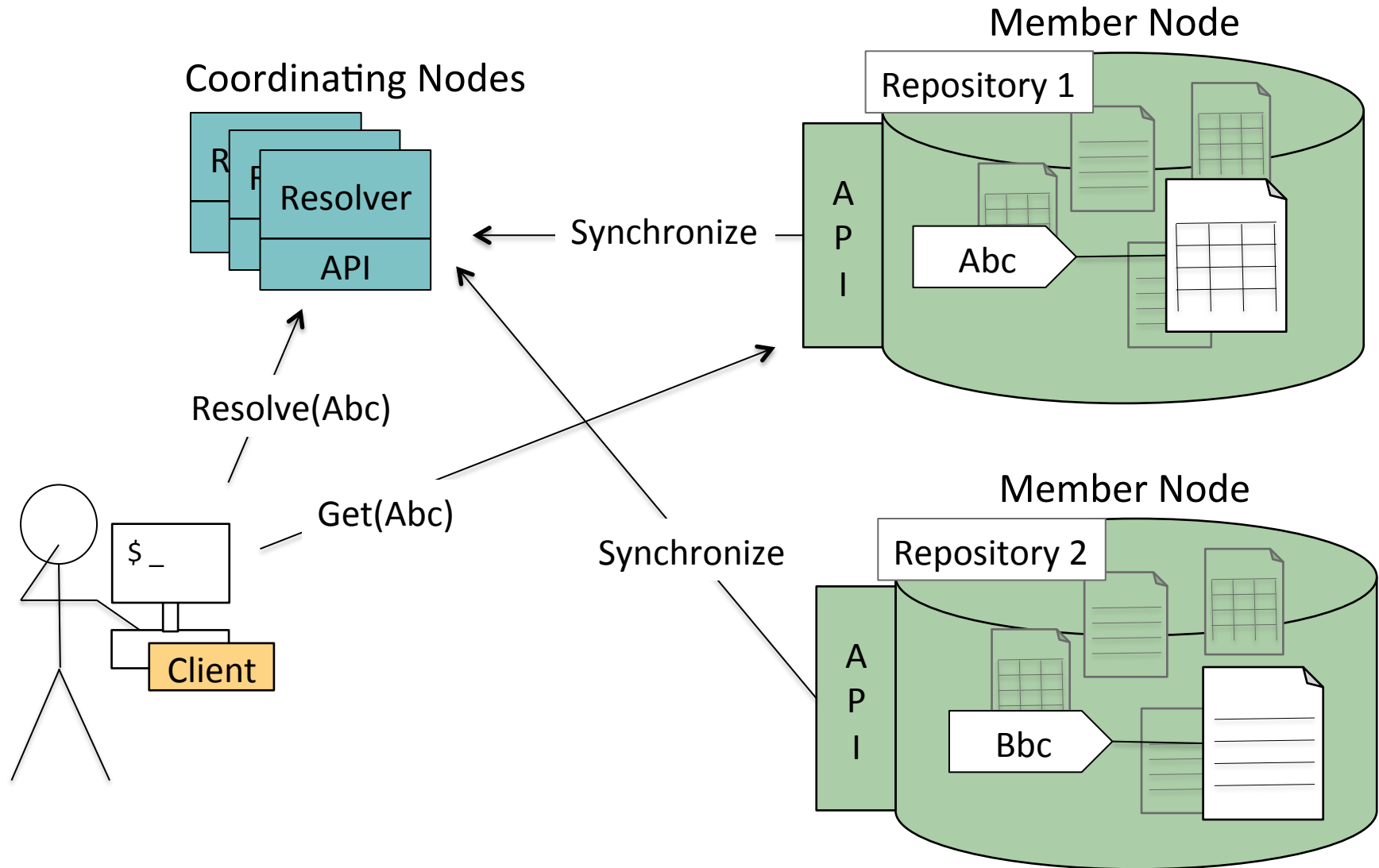
# Enable Data Discovery







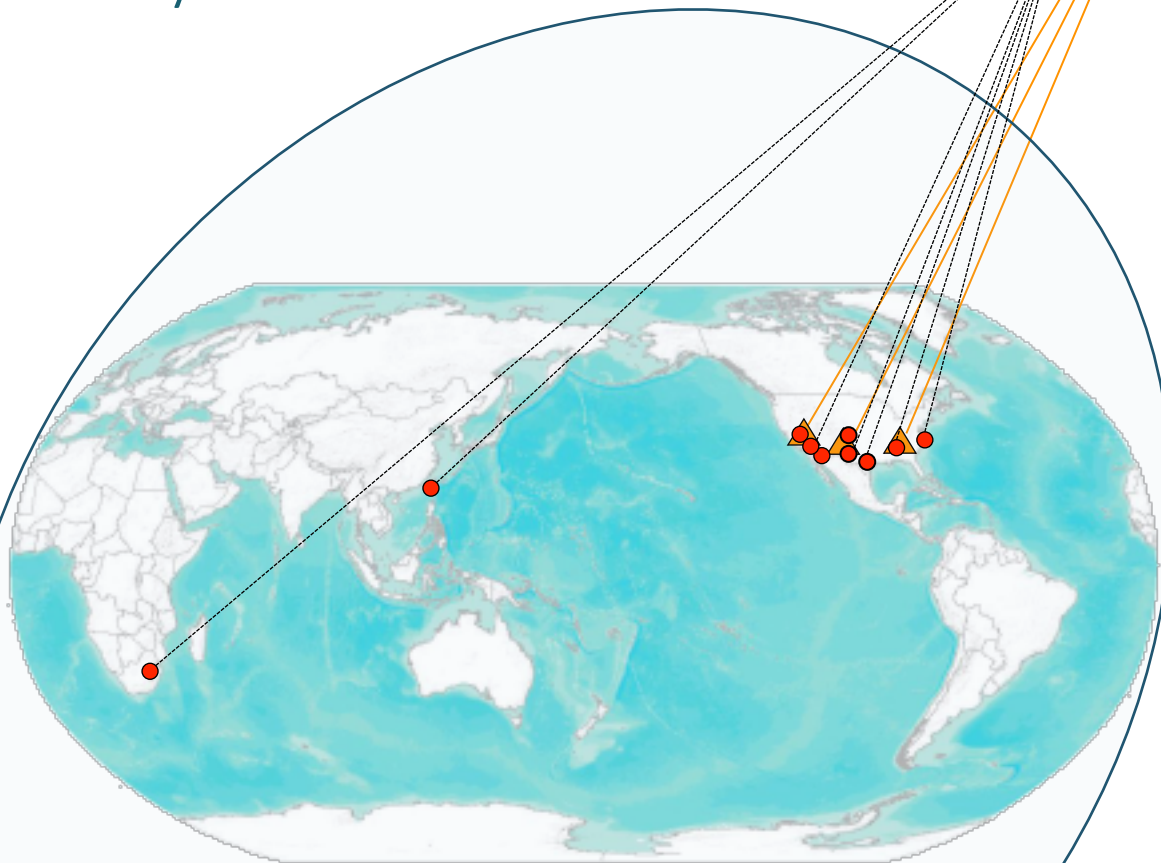
# DataONE – Investigator Tools





# Investigator Tools

Libraries, web interfaces,  
desktop tools for data  
management, discovery, and  
analysis



## Coordinating Nodes

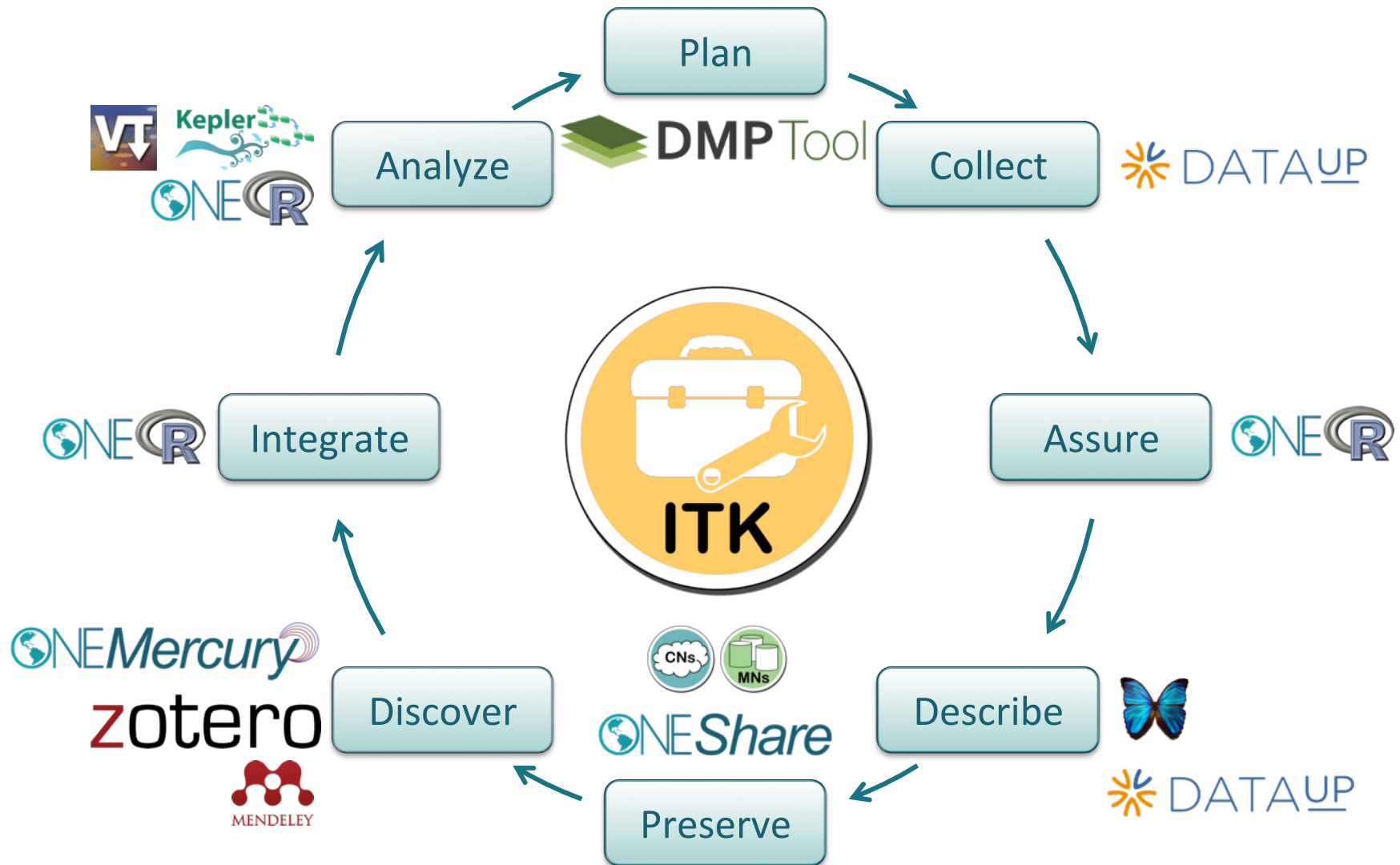
## Member Nodes

## Investigator Toolkit

>> command line interface



# Supporting the Full Data Life Cycle





# DataONE Components

## Investigator Toolkit

Data Discovery

Analysis, Visualization

Data Management

Java Library

Python Library

CLI Tools

REST URLs

## Service Specifications

### Member Nodes

#### Service Interfaces

Read

Authorize

Write

Replicate

Translate to MN Internals

Data Repository

### Coordinating Nodes

#### Service Interfaces

Search

Resolve

Register

Replicate

#### Coordination Layer

Identify

Catalog

Preserve

Monitor

Object Store

Index



# Community Engagement

*stakeholder surveys*



scientists



library's & librarians



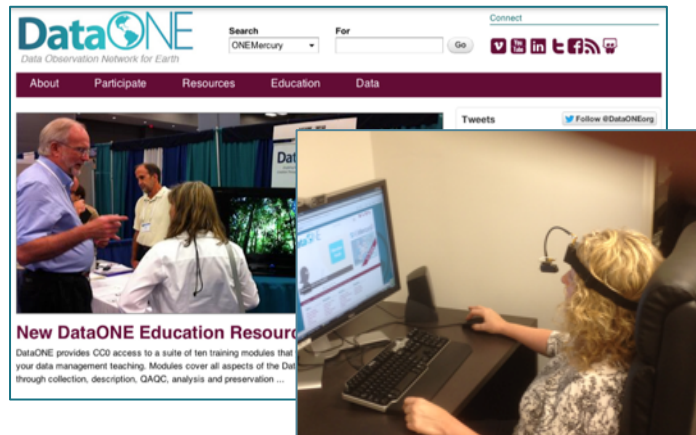
data managers



*persona and scenario development*

*cyberinfrastructure development*

*usability testing*



*external assessments / surveys*

Articles

For Authors

About Us

Search

advanced search

OPEN ACCESS

PEER-REVIEWED

10,244

VIEWS

21

CITATIONS

189

ACADEMIC BOOKMARKS

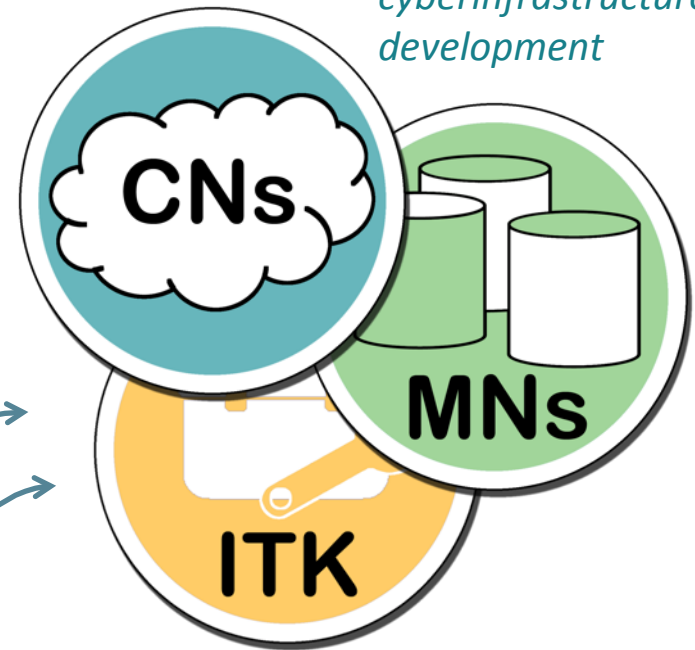
40

SOCIAL SHARES

RESEARCH ARTICLE

Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir , Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, Mike Frame



# Current Member Nodes

Today: 18 (+3) production Member Nodes



Next up:

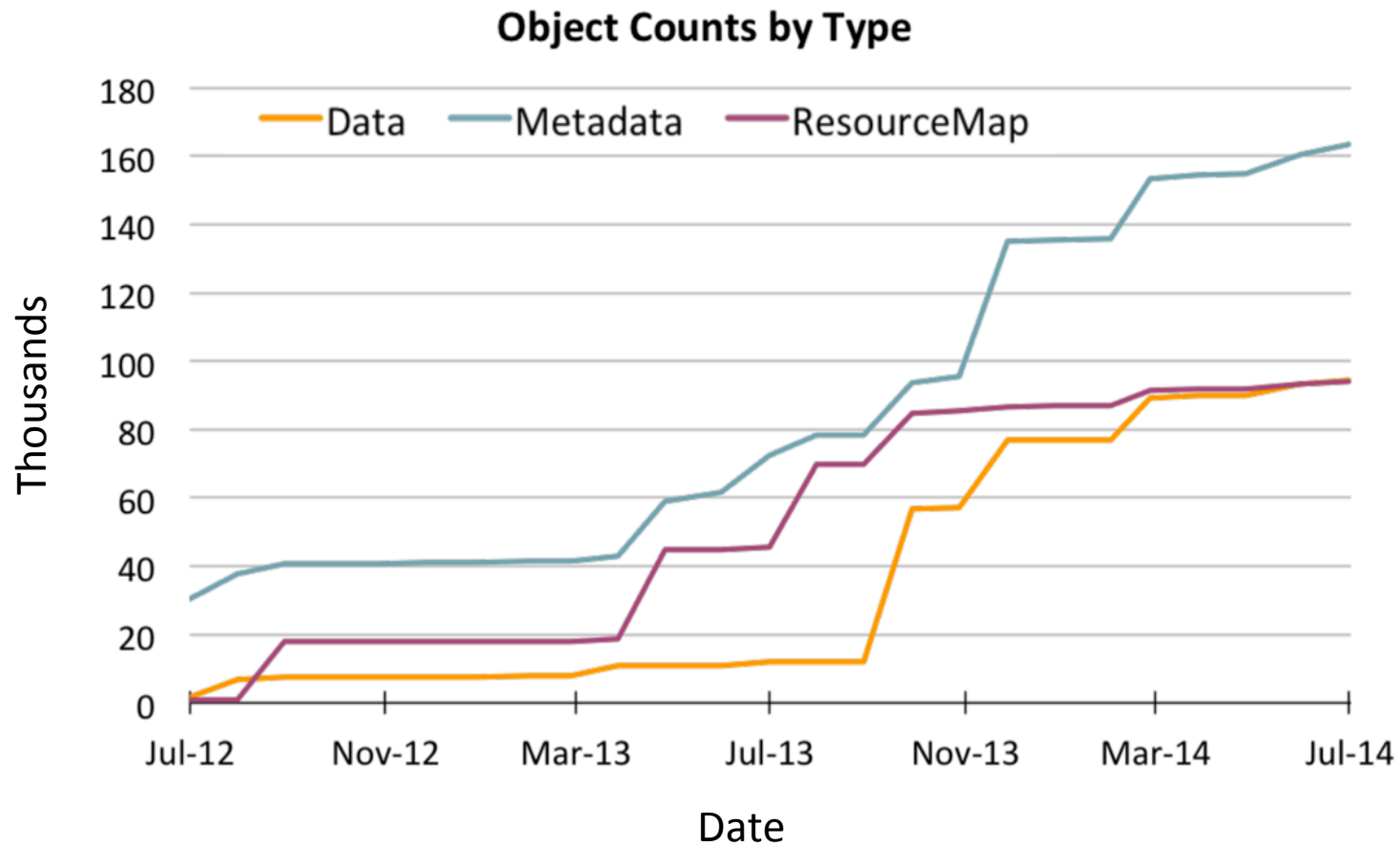


EDORA





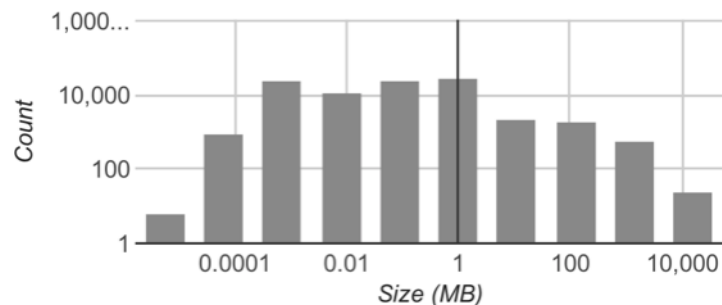
# Content Available



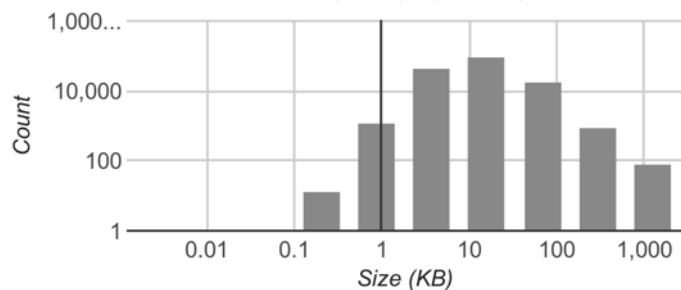


# Content Characteristics

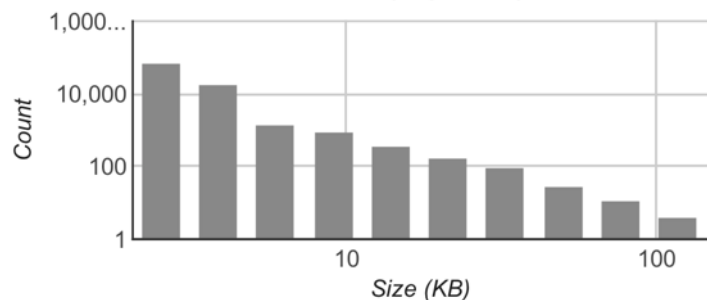
Public Data Objects (Log scales)



Public Metadata Objects (Log scales)



Public Resource Maps (Log scales)



## Content Totals:

- Objects = 478,584
- Data = 152,671
- Data Packages = 114,995

## Typical sizes:

- Data = 1MB
- Metadata = 10KB
- Resource Map = 1KB

Metadata EML variants: 80%

Data simple text / spreadsheet: 80%





# Data Packages and Observations

- DataONE
  - Relatively few, diverse data
  - Operates at collection level
- Collection repositories (e.g. eBird, GBIF)
  - Many, structurally homogeneous data
  - Operate at record level
- New feature to support data subsetting
- Emerging standard by W3C for CSV on the Web

eBird



W3C<sup>®</sup>



# Participation

- Work with an existing Member Node
- Setup a new Member Node with existing software
  - Metacat
  - Mercury
  - DSpace
  - OPeNDAP
  - Dryad
  - GMN (reference implementation)
- Benefit from common services, unique identifiers, content replication, discovery services

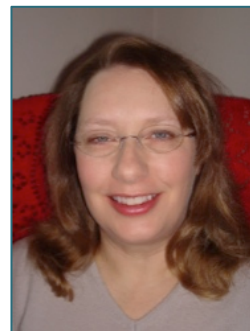




# Member Node Forum



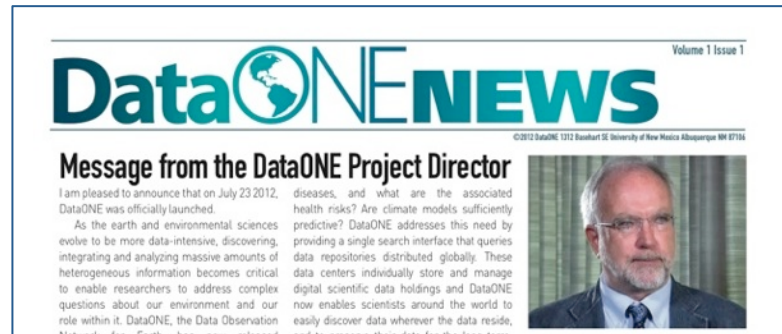
Member Node Forum Coordinators



- Member Node Forum enables current and prospective MNs to:
  - Interact with and learn from one another
  - Receive guidance and technical support
  - Advise on public facing materials and technical documentation

# Engagement & Education

dataone.org  
ask.dataone.org  
notebooks.dataone.org  
coffeehouse.dataone.org



Search: ONEMercury For: Go

Connect: YouTube Twitter LinkedIn Facebook RSS

About Participate Resources Education Data

**Summer Internship Program Now Open**

The DataONE 2014 Summer Internship Program is now open for applications. Project opportunities cover both cyberinfrastructure and community engagement activities. Read more read more

**Latest News**

- The USA National Phenology Network joins DataONE as a Member Node
- SEAD Virtual Archive joins DataONE as a Member Node
- View All News

**Find it Fast**

- ONEMercury
- Data Management Planning
- Best Practices
- Software Tools

**Subscribe**

**Search for Data**

Research Notes Intern Notebooks Coffeehouse DataONE.org

**Continue Scraping, Introduce Quality Control with Hashes**

Posted on February 18, 2014 by Tanner Jessel - No Comments

Continuation and completion of harvesting with quality control / assurance

Tags: archiving Citations community

**Coffeehouse**

A collection of data blog posts from around the web

Add your blog Most

**2014 NDSA Philly Regional Meeting: January 23-24**

by [Link] February 18, 2014

The following is a guest post by Nicole Scates, IT manager at The Library Company of Philadelphia, an NDSA member. Digital stewardship is a prime topic for small institutions trying to keep pace with the increasing demands for digital content. The Library Company of Philadelphia, a special collections library founded by Benjamin Franklin in 1731, [...]

Read more

tags people & groups badges

Hi there! Please sign in help

ALL UNANSWERED search or ask your question

ASK YOUR QUESTION

31 questions

Sort by: by date by activity by answers by votes RSS

Contributors

How do I prepare my data for addition to a DataONE repository?

1 vote 1 answer 50 views


data, curation data, management data, preparation metadata contribute





# Developer Resources

- [#dataone](http://irc.ecoinformatics.org)
- [developers@dataone.org](mailto:developers@dataone.org)
- [support@dataone.org](mailto:support@dataone.org)
- <http://mule1.dataone.org/ArchitectureDocs-current/>



Data Observation Network for Earth








Search

ONEMercury

For

Go

Connect

About

Participate

Resources

Education

Data

Home » Participate » Developer Resources

Participate

DataONE Users Group

Member Nodes

Internships

Developer Resources

Open Positions

## Developer Resources

DataONE welcomes contributions and collaborations with all who are interested in helping the goals of meeting the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. This page describes resources that are available to those interested in learning more about the cyberinfrastructure for DataONE, particularly for contributing to DataONE or using the products that DataONE has developed in other efforts. DataONE code is generally available under an [Apache 2.0 license](#) and our [code repository](#).

# DataONE Team and Sponsors



• Amber Budden, Roger Dahl, Rebecca Koskela, Bill Michener, Robert Nahf, Skye Roseboom, Mark Servilla



• Ewa Deelman



• Dave Viegais



• Deborah McGuinness



• Suzie Allard, Kimberly Douglass, Laura Moyers, Carol Tenopir, Robert Waltz, Bruce Wilson



• Jeff Horsburgh



• John Cobb, Bob Cook, Ranjeet Devarakonda, Giri Palanismany, Line Pouchard



• Robert Sandusky



• Patricia Cruse, John Kunze



• Bertram Ludascher



• Sky Bristol, Mike Frame, Richard Huffine, Viv Hutchison, Jeff Morisette, Jake Weltzin, Lisa Zolly



• Peter Honeyman



• Stephanie Hampton, Chris Jones, Matt Jones, Ben Leinfelder, Andrew Pippin, Mark Schildhauer, Jing Tao



• Cliff Duke



• Paul Allen, Rick Bonney, Steve Kelling



• Carole Goble



• Jane Greenberg, Ryan Scherle, Todd Vision



• Donald Hobern



• Randy Butler



• David DeRoure



• Paolo Missier

