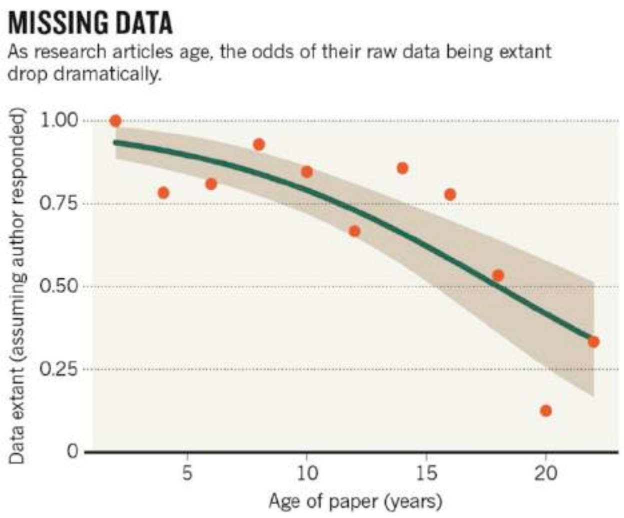
DataONE: Current Services, New Tools and Future Developments

Amber Budden
Director for Community Engagement and Outreach

Dave Vieglais
Director for Development and Operations

dataone.org

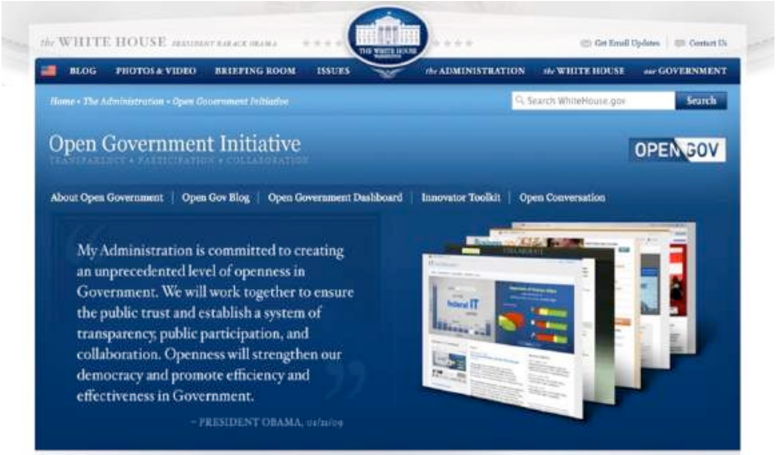
Science and Data Challenges



MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.

Vines, T. H. *et al.* Curr. Biol. http://dx.doi.org/10.1016/j.cub.2013.11.014 (2013).



the WHITE HOUSE PRESIDENT BARACK OBAMA

Get Email Updates
Contact Us
BLOG
PHOTOS & VIDEO
BRIEFING ROOM
ISSUES
the ADMINISTRATION
the WHITE HOUSE
our GOVERNMENT
Home • The Administration • Open Government Initiative
Search WhiteHouse.gov
Search

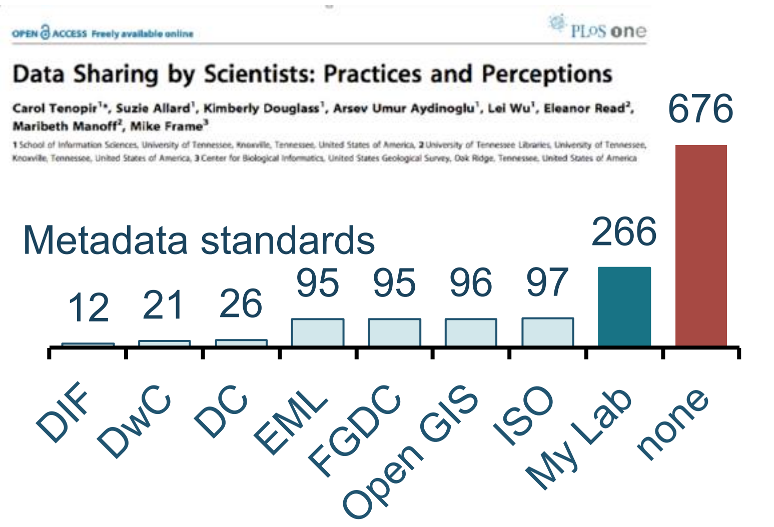
Open Government Initiative

TRANSPARENCY • PARTICIPATION • COLLABORATION

OPEN GOV
About Open Government | Open Gov Blog | Open Government Dashboard | Innovator Toolkit | Open Conversation

My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.

– PRESIDENT OBAMA, 01/21/09



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]

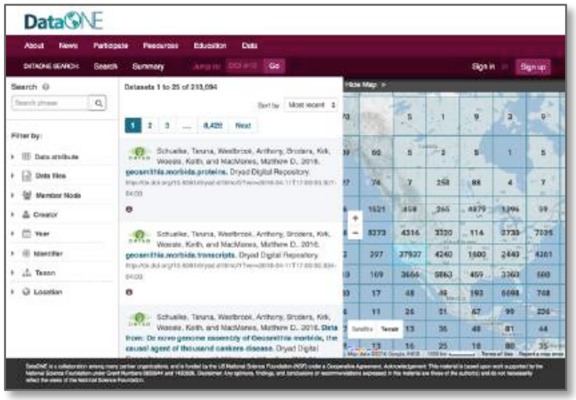
1 School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, 2 University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, 3 Center for Biological Informatics, United States Geological Survey, Oak Ridge, Tennessee,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

DataONE
Vision and Mission

Providing universal access to data about life on earth and the environment that sustains it

1. Building community

2. Developing sustainable data discovery and interoperability solutions



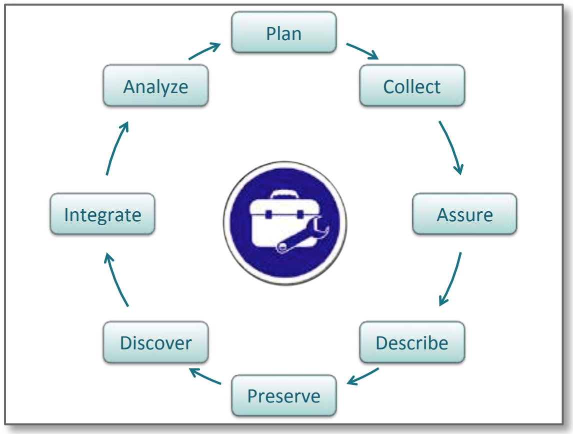
Datasets 1 to 25 of 210,694

Schuelke, Taruna, Westbrook, Anthony, Broders, Kirk, Woeste, Keith, and MacManes, Matthew D.. 2016. geosmithia.morbida.proteins. Dryad Digital Repository.

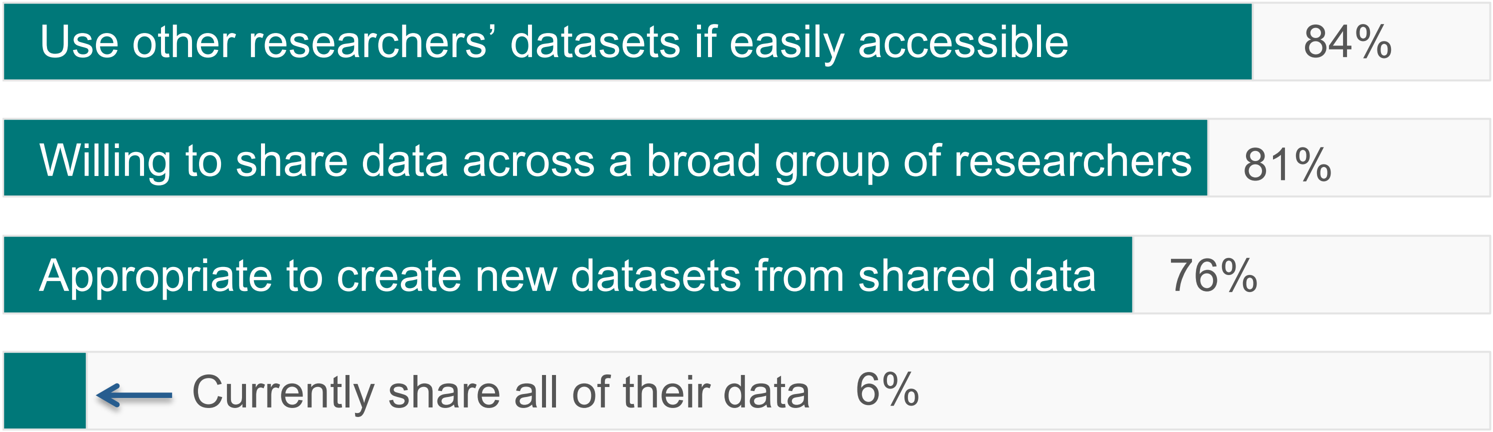
Schuelke, Taruna, Westbrook, Anthony, Broders, Kirk, Woeste, Keith, and MacManes, Matthew D.. 2016. geosmithia.morbida.transcripts. Dryad Digital Repository.

Schuelke, Taruna, Westbrook, Anthony, Broders, Kirk, Woeste, Keith, and MacManes, Matthew D.. 2016. Data from: De novo genome assembly of Geosmithia morbida, the causal agent of thousand cankers disease. Dryad Digital

3. Supporting researcher tools and services



Scientists want to share data



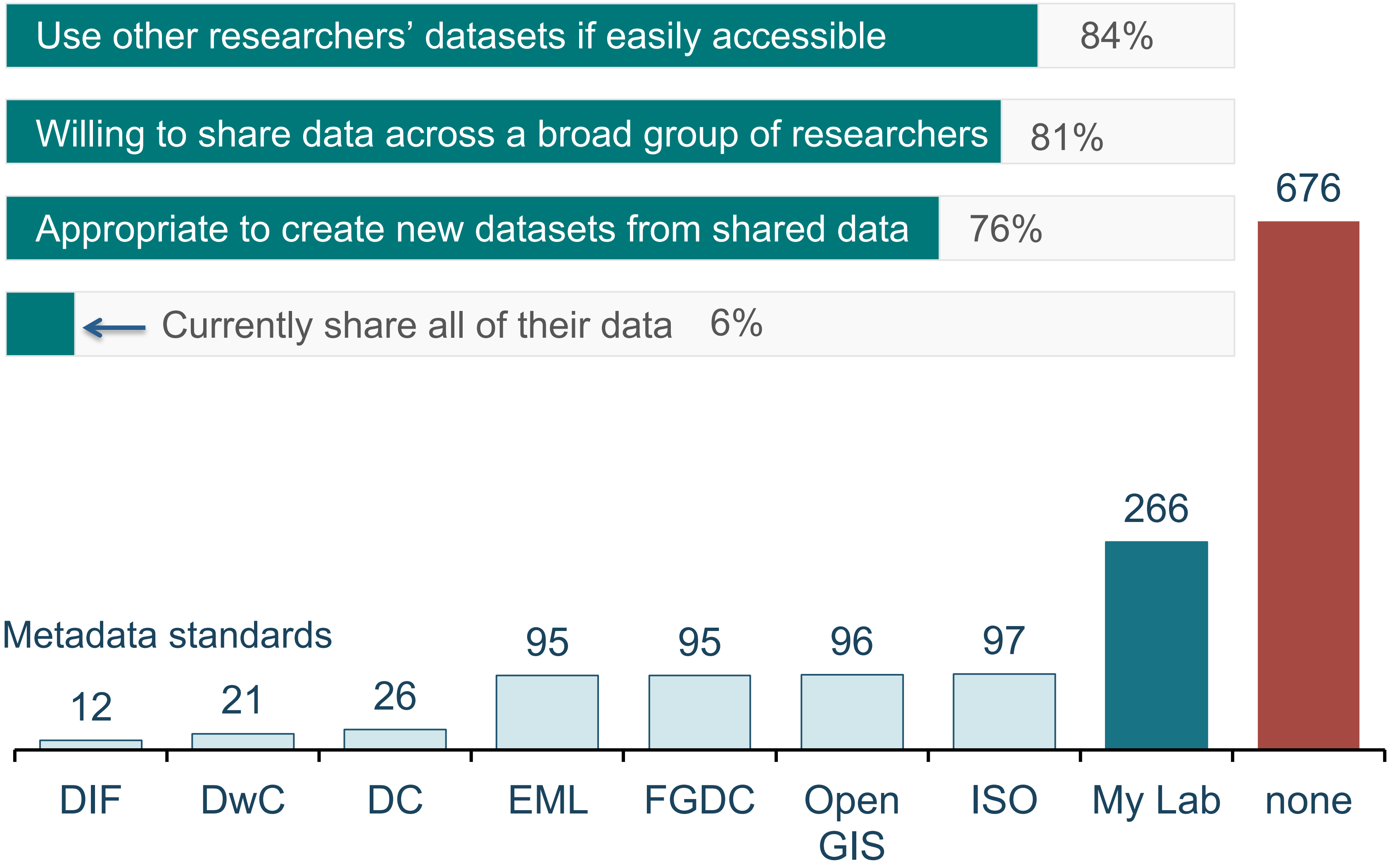
Use other researchers' datasets if easily accessible 84%

Willing to share data across a broad group of researchers 81%

Appropriate to create new datasets from shared data 76%

← Currently share all of their data 6%

Scientists want to share data



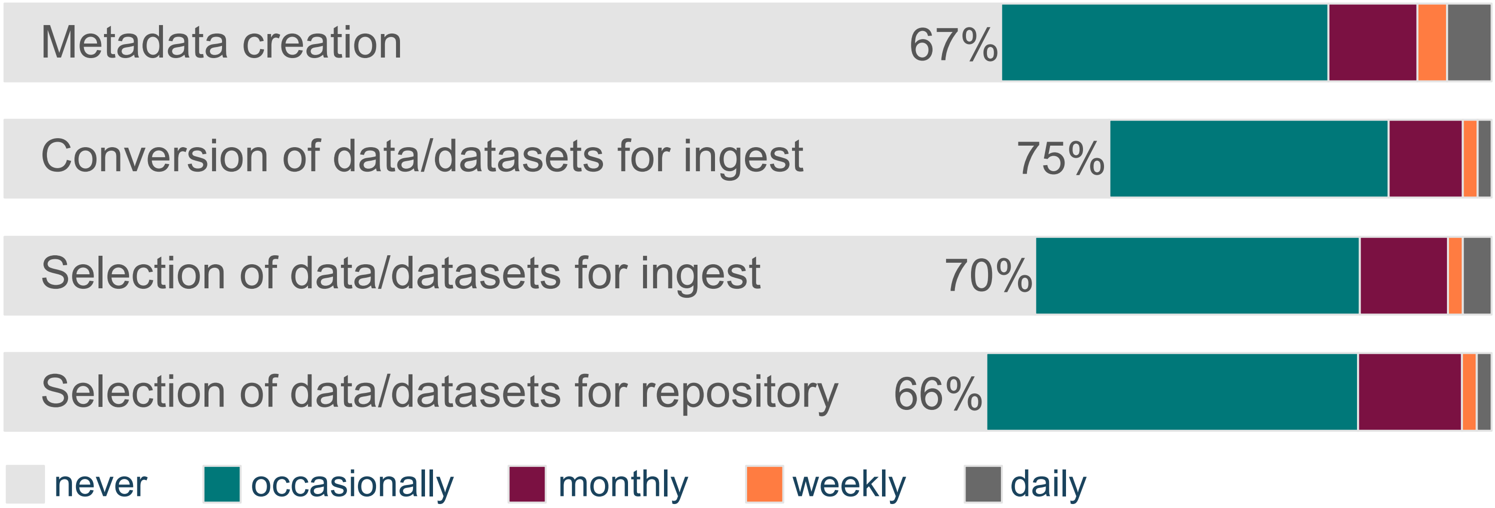
Use other researchers’ datasets if easily accessible 84%

Willing to share data across a broad group of researchers 81%

Appropriate to create new datasets from shared data 76%

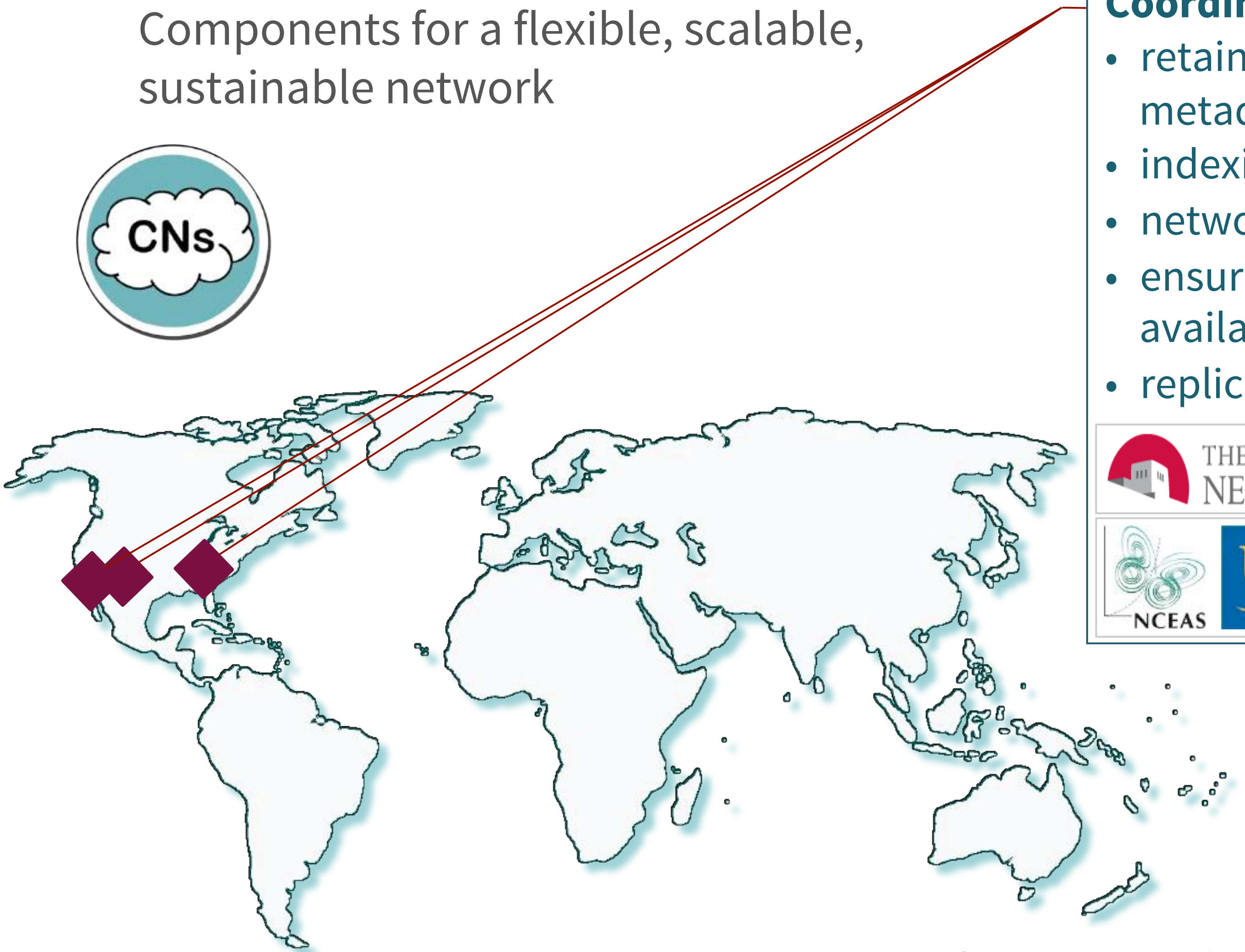
← Currently share all of their data 6%

Libraries not yet providing data services



DataONE Cyberinfrastructure
Coordinating Nodes

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

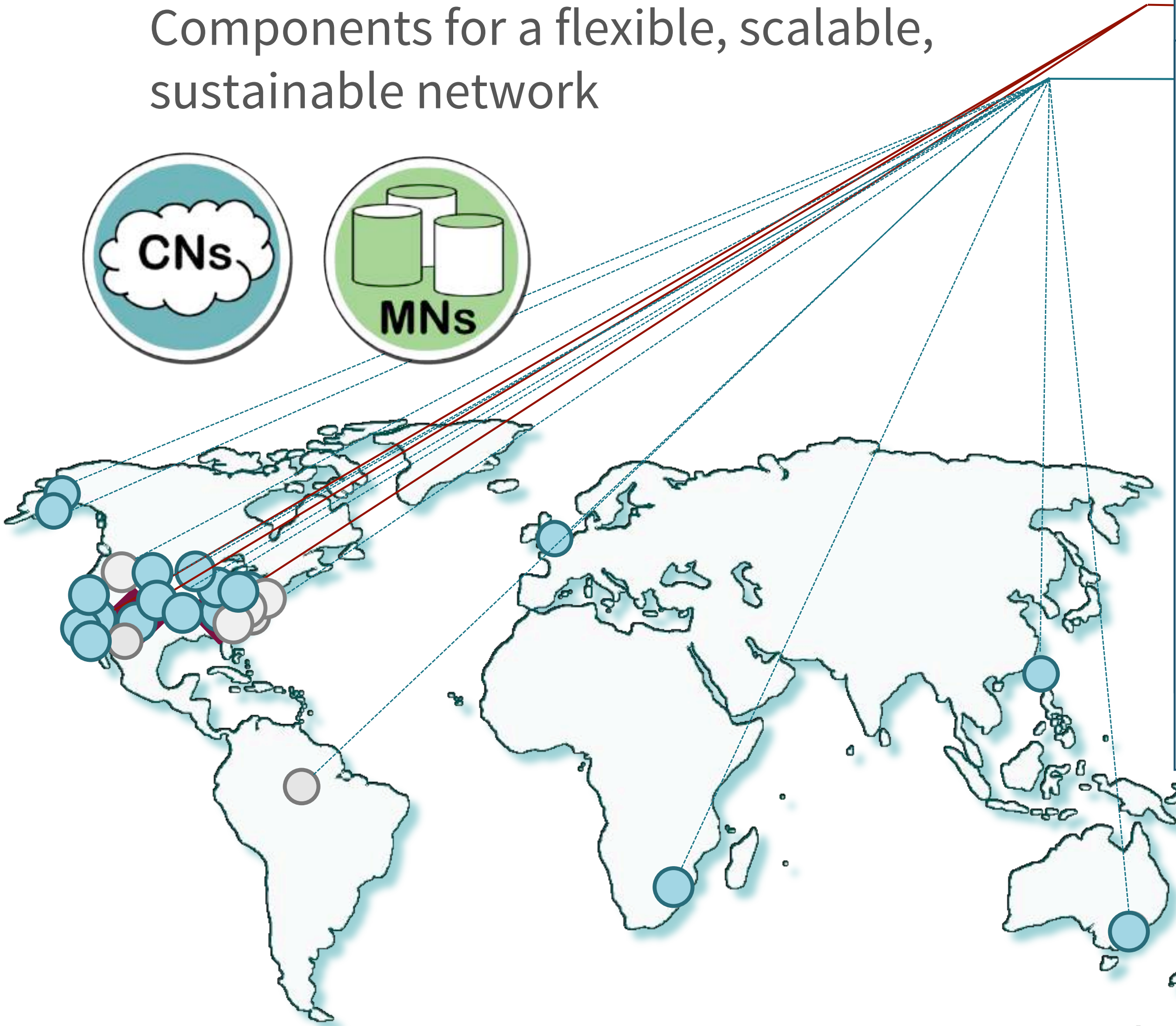
UCSB

UT

www.dataone.org/coordinating-nodes

DataONE Cyberinfrastructure
Member Nodes

Components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

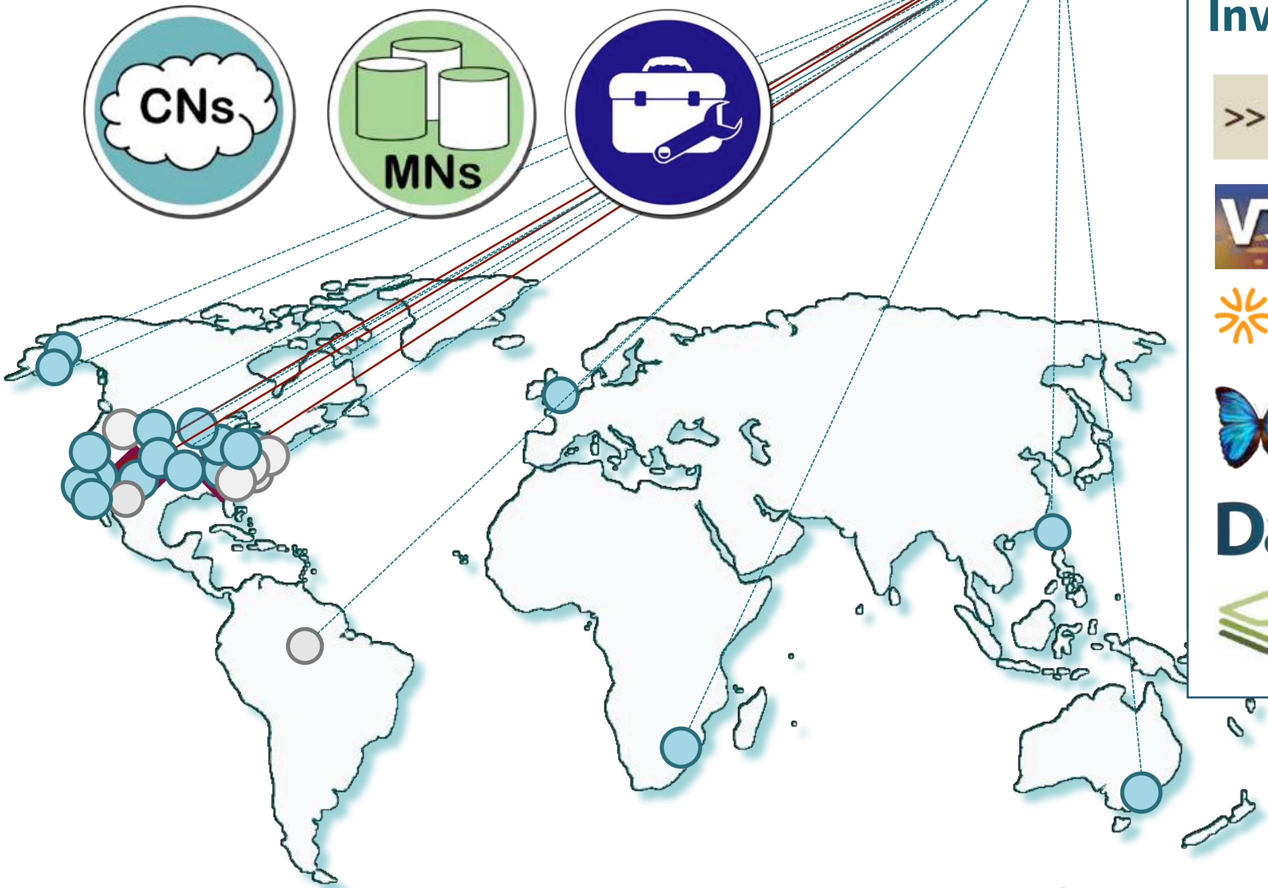
- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data

KU BIODIVERSITY INSTITUTE The University of Kansas
AOOS Alaska Ocean Observing System
UC3Merritt
knb
PISCO
DRYAD
USA npn National Phenology Network
SEAD
LTER

www.dataone.org/member-nodes

DataONE Cyberinfrastructure
Investigator Toolkit

Components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

Investigator Toolkit

```
>> command line interface
```

VT python MENDELEY

dash DATA SHARING MADE EASY zotero

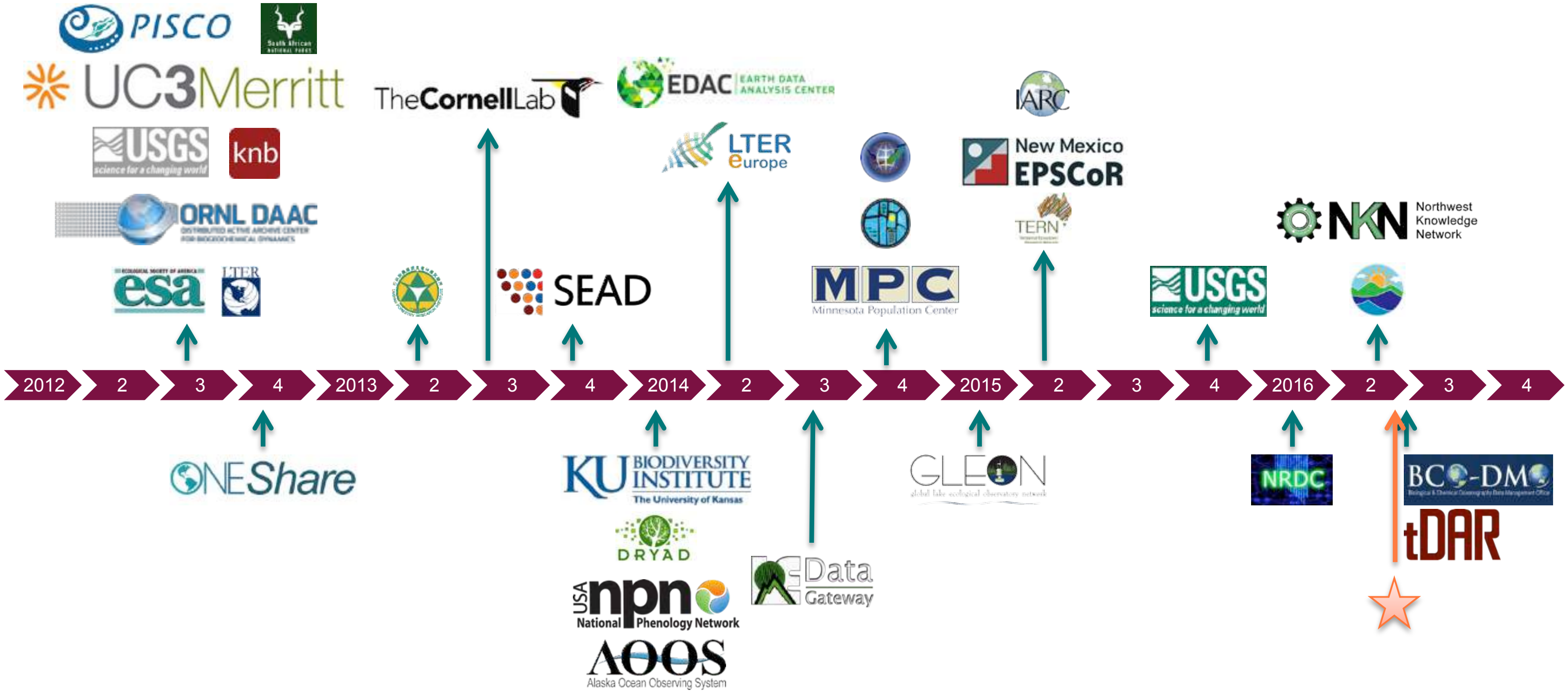
Kepler ONE R

DataONESearch

DMPTool

www.dataone.org/investigator-toolkit

DataONE Member Nodes
Current and Upcoming


PISCO
South African National Parks
UC3Merritt
USGS science for a changing world
knb
ORNL DAAC Distributed Active Archive Center for Biogeochemical Dynamics
esa
LTER
TheCornellLab
EDAC Earth Data Analysis Center
SEAD
LTER Europe
MPC Minnesota Population Center
IARC
New Mexico EPSCoR
TERN
USGS science for a changing world
NKN Northwest Knowledge Network
2012
2
3
4
2013
2
3
4
2014
2
3
4
2015
2
3
4
2016
2
3
4
ONEShare
KU Biodiversity Institute The University of Kansas
DRYAD
USA npn National Phenology Network
AOOS Alaska Ocean Observing System
Data Gateway
GLEON global lake ecological observatory network
NRDC
BCO-DMO
tDAR

Upcoming Member Nodes

DFC DataNet FEDERATION CONSORTIUM

neon

GBIF

GULF of MEXICO RESEARCH INITIATIVE GRIIDC

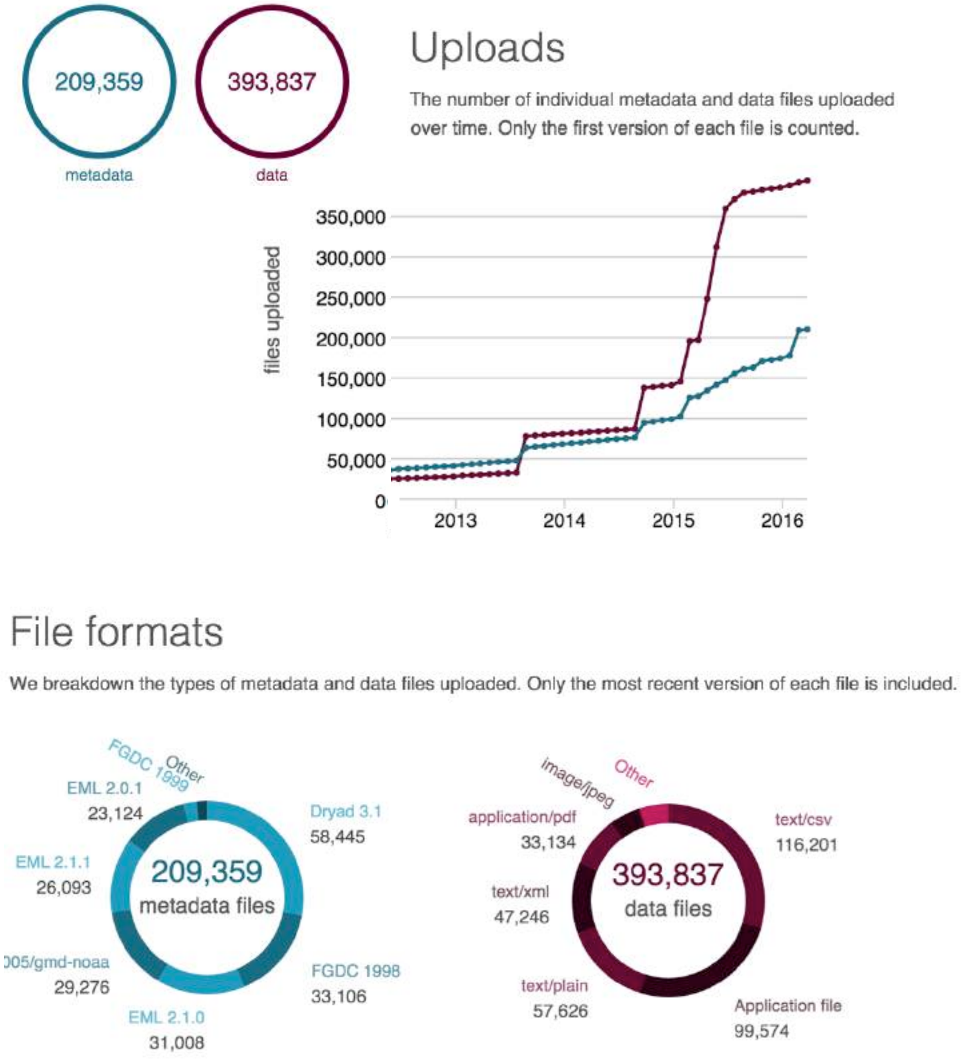
GCOOS GULF OF MEXICO COASTAL OCEAN OBSERVING SYSTEM

SAEON

PPBio

DataONE

Data Holdings



Uploads

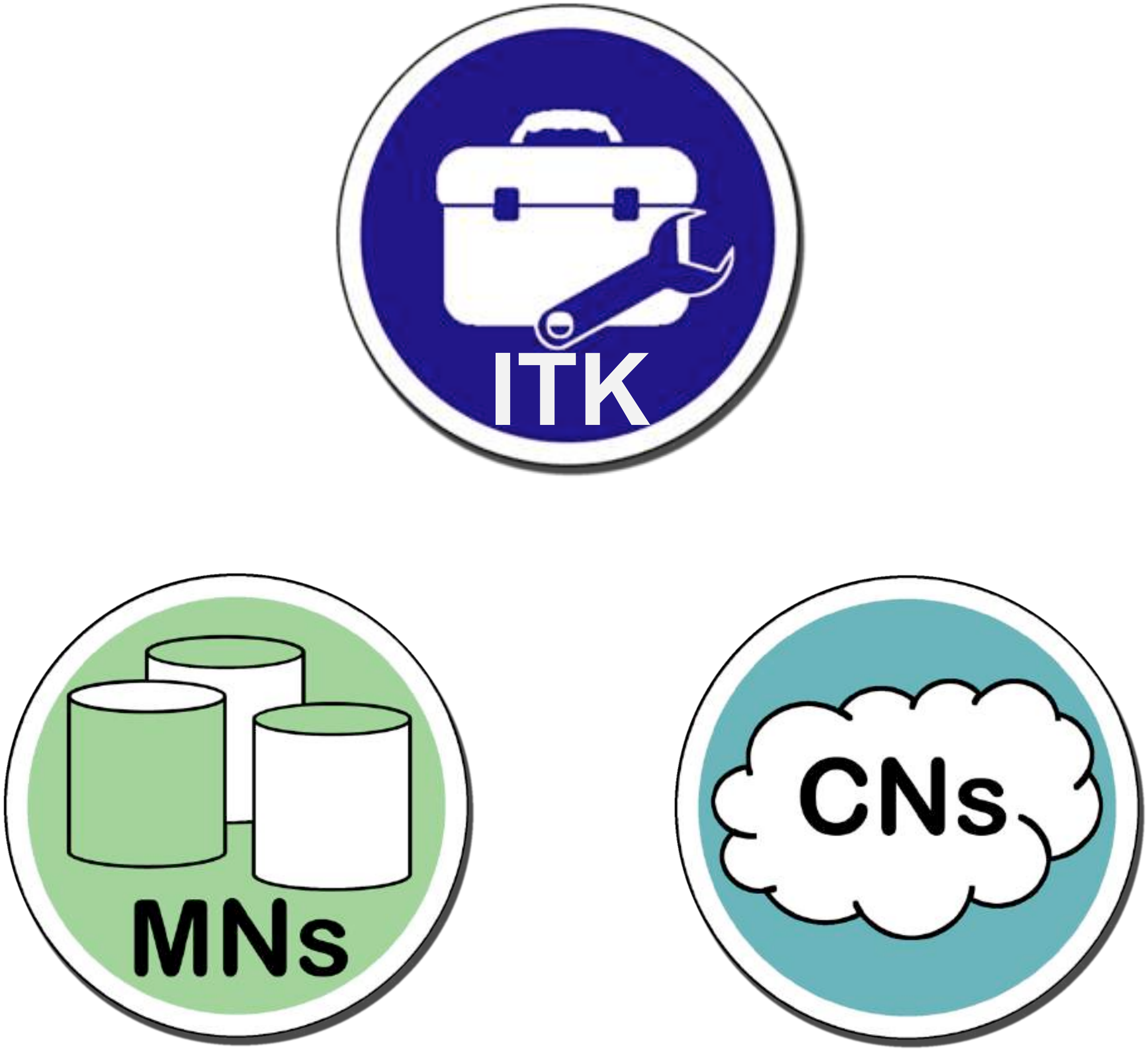
The number of individual metadata and data files uploaded over time. Only the first version of each file is counted.

File formats

We breakdown the types of metadata and data files uploaded. Only the most recent version of each file is included.

www.search.dataone.org/#profile

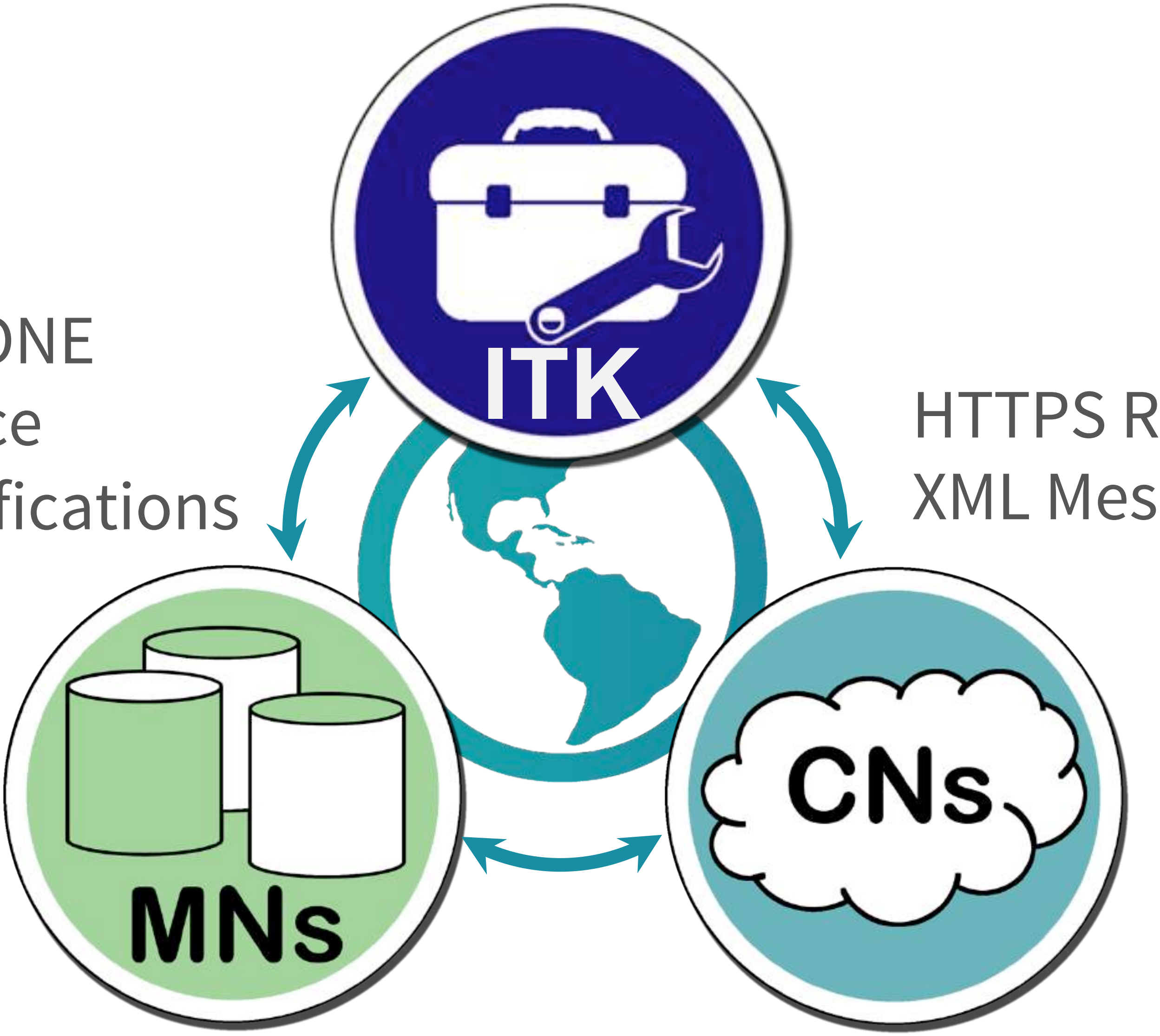
Three Components



Three Components

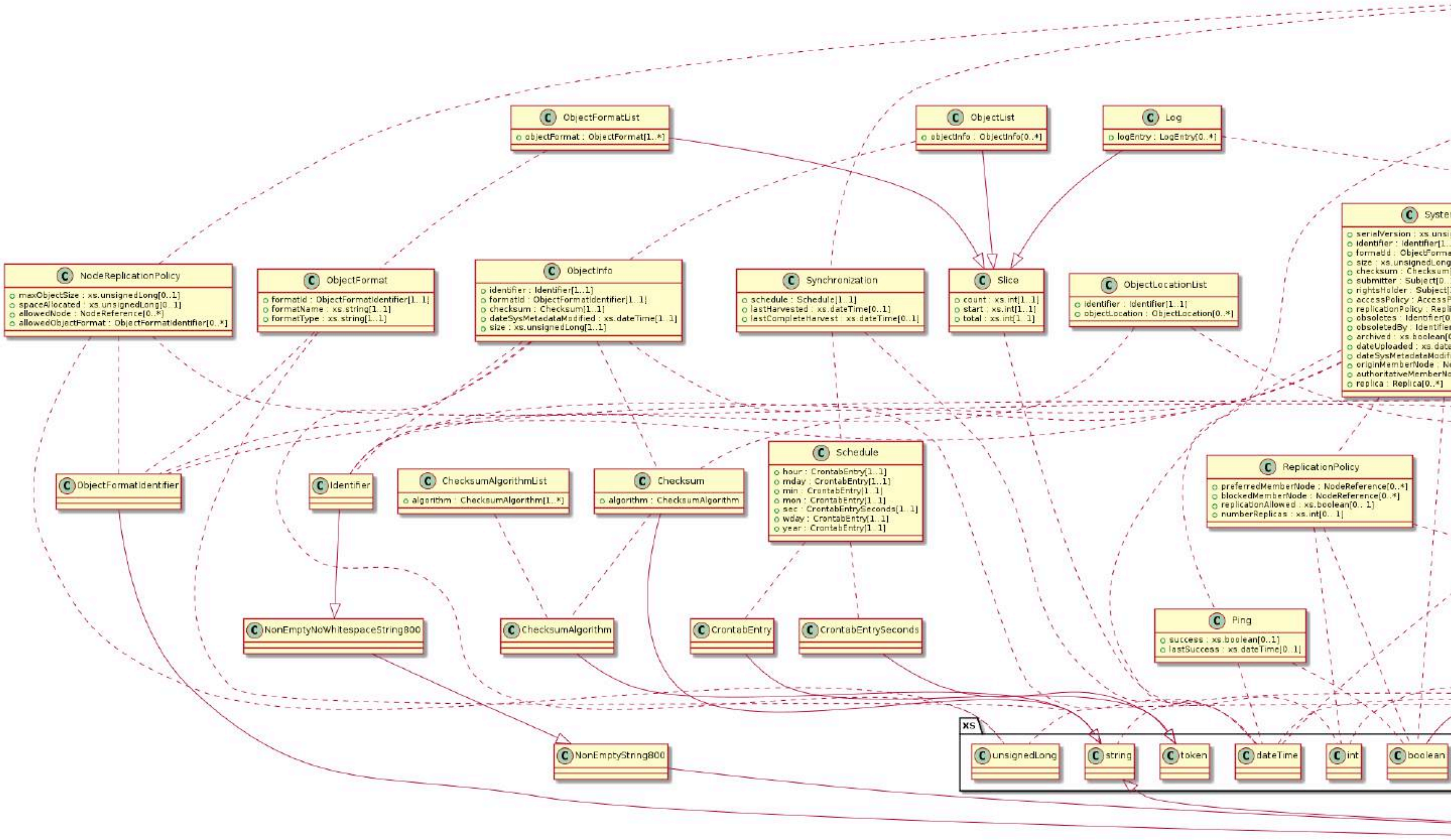
Integration

DataONE
Service
Specifications

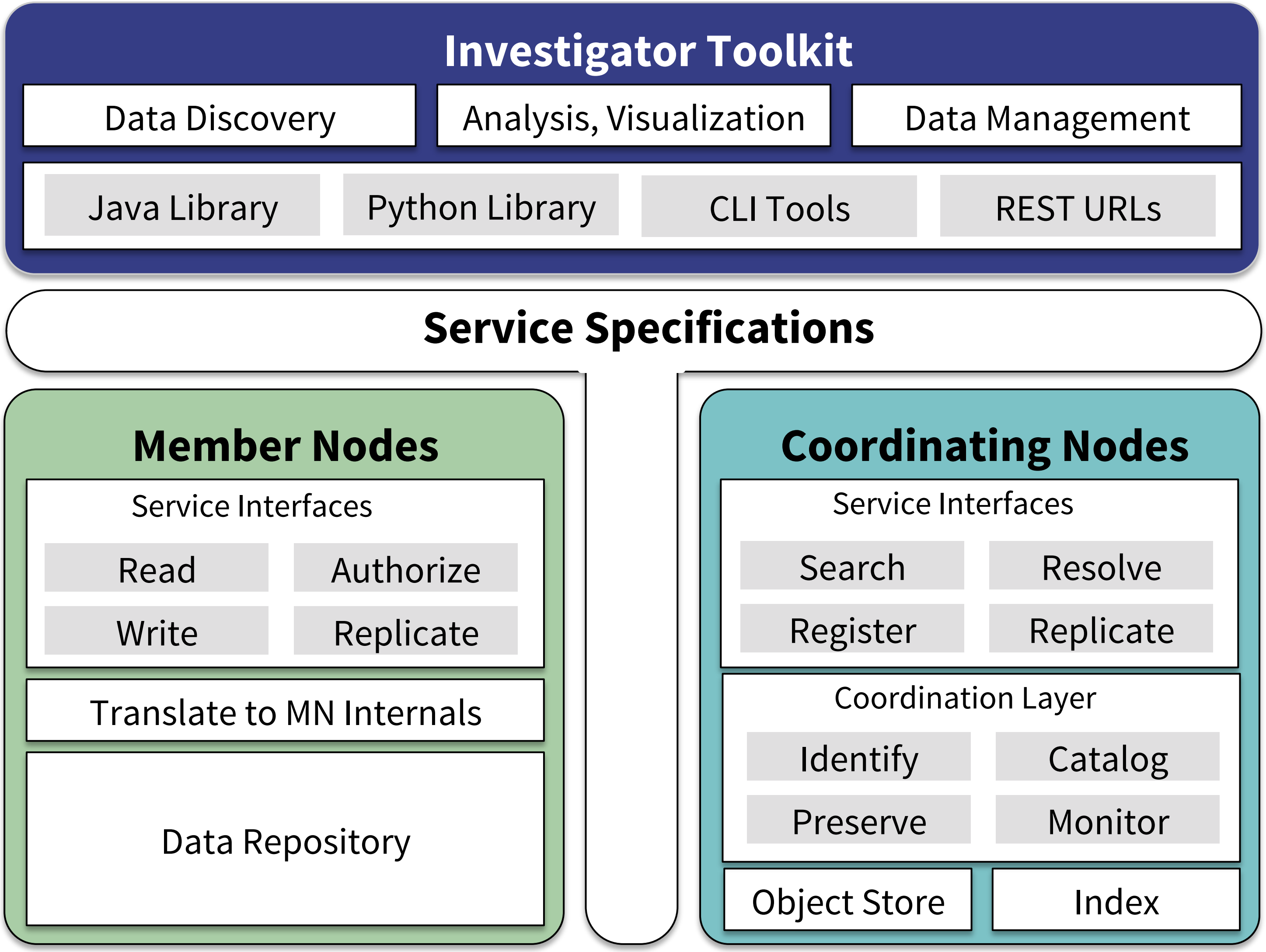


HTTPS REST
XML Messages

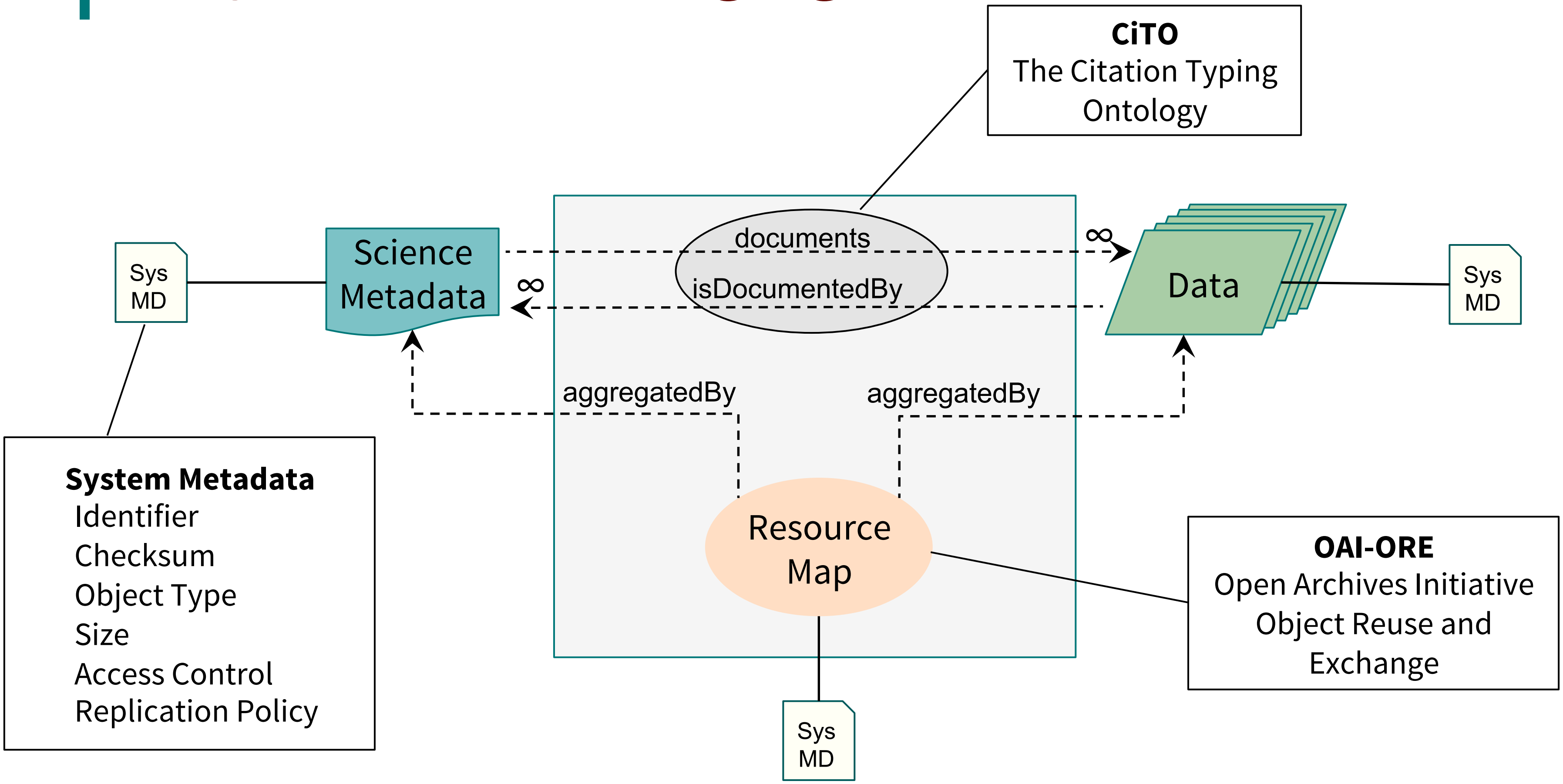
purl.dataone.org/architecture


ObjectFormatList
objectFormat : ObjectFormat[1..*]
ObjectList
objectInfo : ObjectInfo[0..*]
Log
logEntry : LogEntry[0..*]
Syster
serielVersion : xs.unsi
identifier : Identifier[1..
formatId : ObjectForma
size : xs.unsignedLong
checksum : Checksum
submitter : Subject[0..
rightsHolder : Subject[
accessPolicy : AccessP
replicationPolicy : Repli
obsoletes : Identifier[0
obsoletedBy : Identifier
archived : xs.boolean[0
dateUploaded : xs.date
dateSysMetadataModifi
originMemberNode : No
authoritativeMemberNo
replica : Replica[0..*]
NodeReplicationPolicy
maxObjectSize : xs.unsignedLong[0..1]
spaceAllocated : xs.unsignedLong[0..1]
allowedNode : NodeReference[0..*]
allowedObjectFormat : ObjectFormatIdentifier[0..*]
ObjectFormat
formatId : ObjectFormatIdentifier[1..1]
formatName : xs.string[1..1]
formatType : xs.string[1..1]
ObjectInfo
identifier : Identifier[1..1]
formatId : ObjectFormatIdentifier[1..1]
checksum : Checksum[1..1]
dateSysMetadataModified : xs.dateTime[1..1]
size : xs.unsignedLong[1..1]
Synchronization
schedule : Schedule[1..1]
lastHarvested : xs.dateTime[0..1]
lastCompleteHarvest : xs.dateTime[0..1]
Slice
count : xs.int[1..1]
start : xs.int[1..1]
total : xs.int[1..1]
ObjectLocationList
identifier : Identifier[1..1]
objectLocation : ObjectLocation[0..*]
ObjectFormatIdentifier
Identifier
ChecksumAlgorithmList
algorithm : ChecksumAlgorithm[1..*]
Checksum
algorithm : ChecksumAlgorithm
Schedule
hour : CrontabEntry[1..1]
mday : CrontabEntry[1..1]
min : CrontabEntry[1..1]
mon : CrontabEntry[1..1]
sec : CrontabEntrySeconds[1..1]
wday : CrontabEntry[1..1]
year : CrontabEntry[1..1]
ReplicationPolicy
preferredMemberNode : NodeReference[0..*]
blockedMemberNode : NodeReference[0..*]
replicationAllowed : xs.boolean[0..1]
numberReplicas : xs.int[0..1]
NonEmptyNoWhitespaceString800
ChecksumAlgorithm
CrontabEntry
CrontabEntrySeconds
Ping
success : xs.boolean[0..1]
lastSuccess : xs.dateTime[0..1]
NonEmptyString800
xs
unsignedLong
string
token
dateTime
int
boolean

Shown in a High Level Design



Objects in DataONE
Objects and Packaging

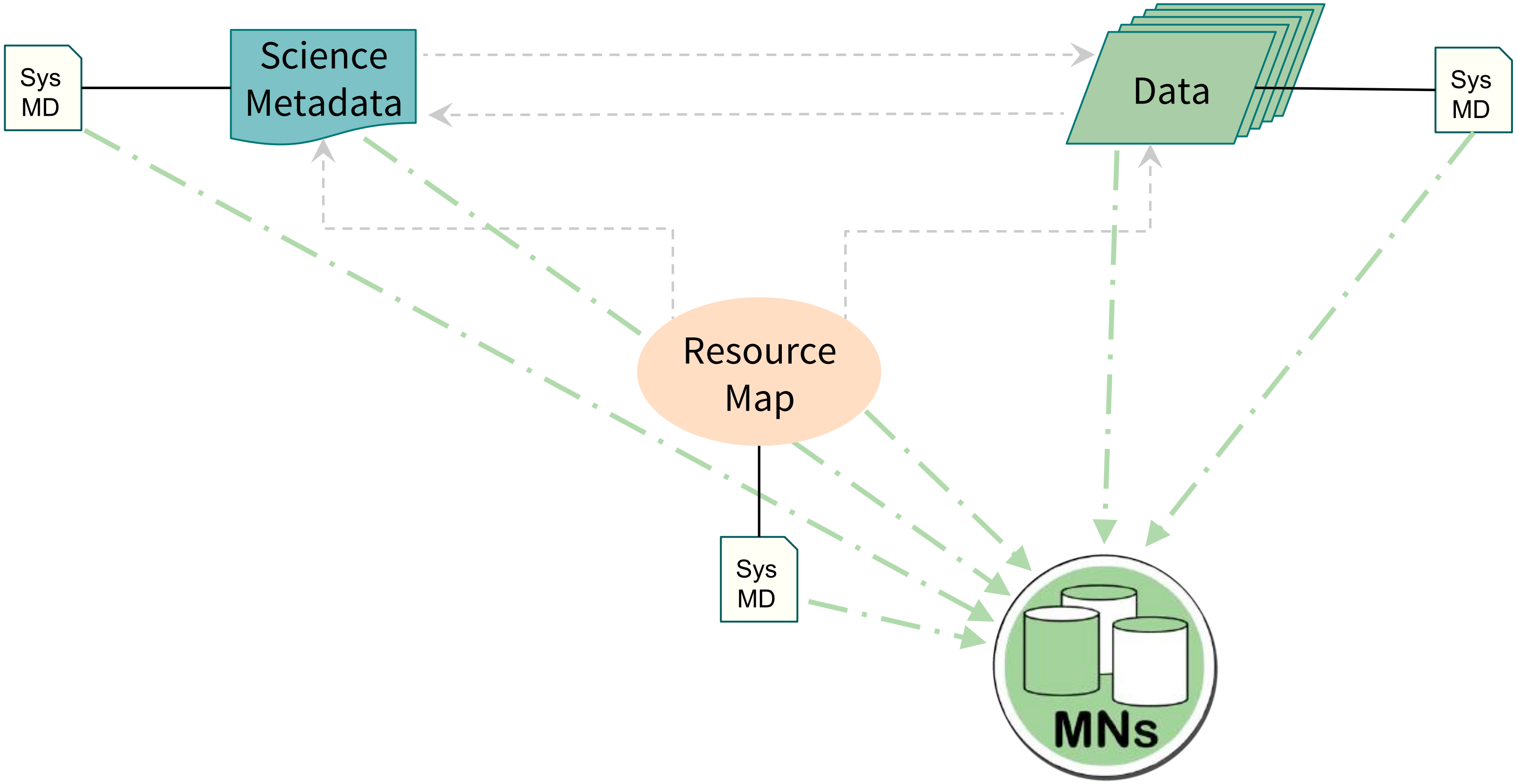


CiTO
The Citation Typing Ontology

System Metadata
- Identifier
- Checksum
- Object Type
- Size
- Access Control
- Replication Policy

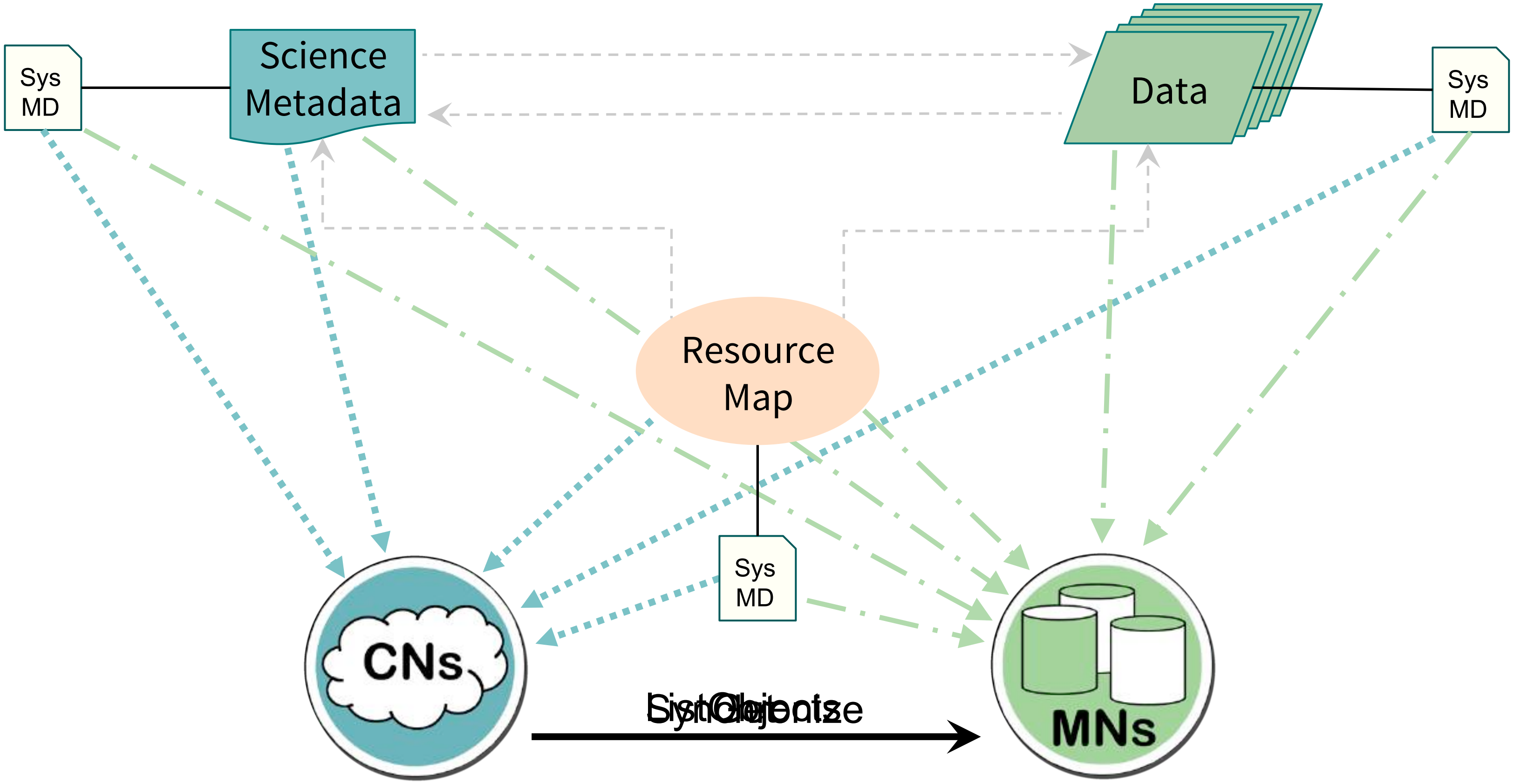
OAI-ORE
Open Archives Initiative Object Reuse and Exchange

Objects in DataONE
Everything on Member Nodes

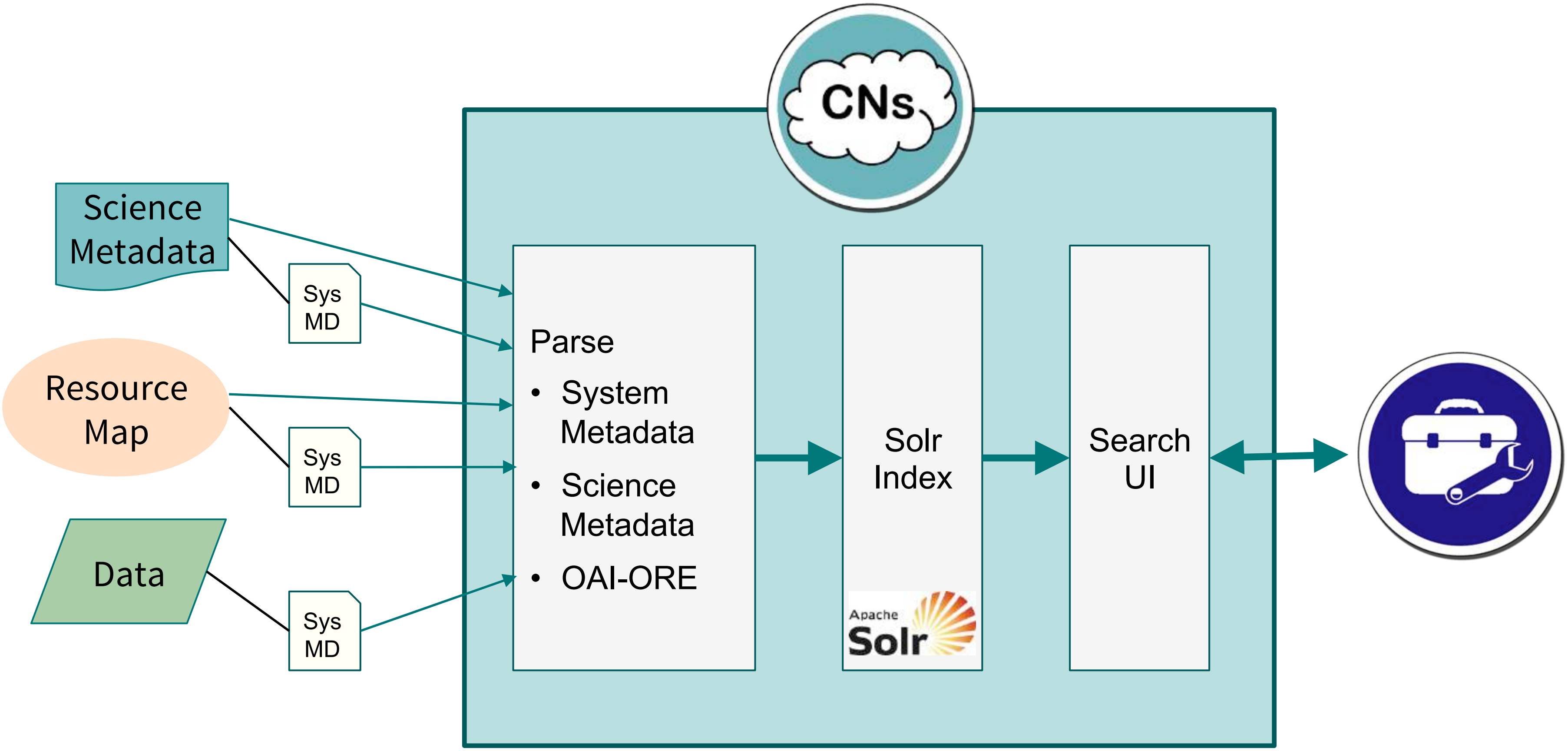


Objects in DataONE

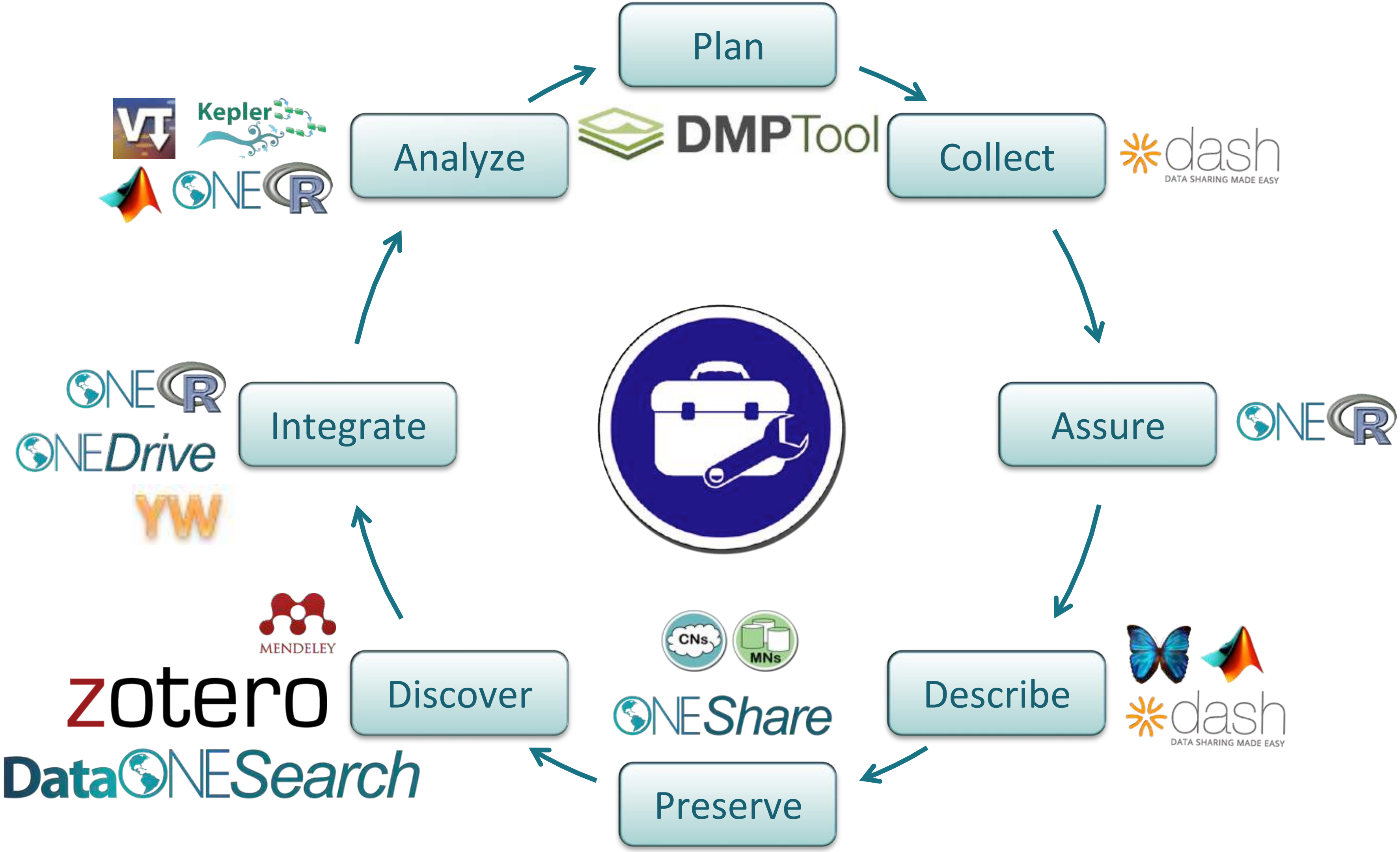
Synchronization to Coordinating Nodes



Coordinating Node Processing
Indexing for Discovery

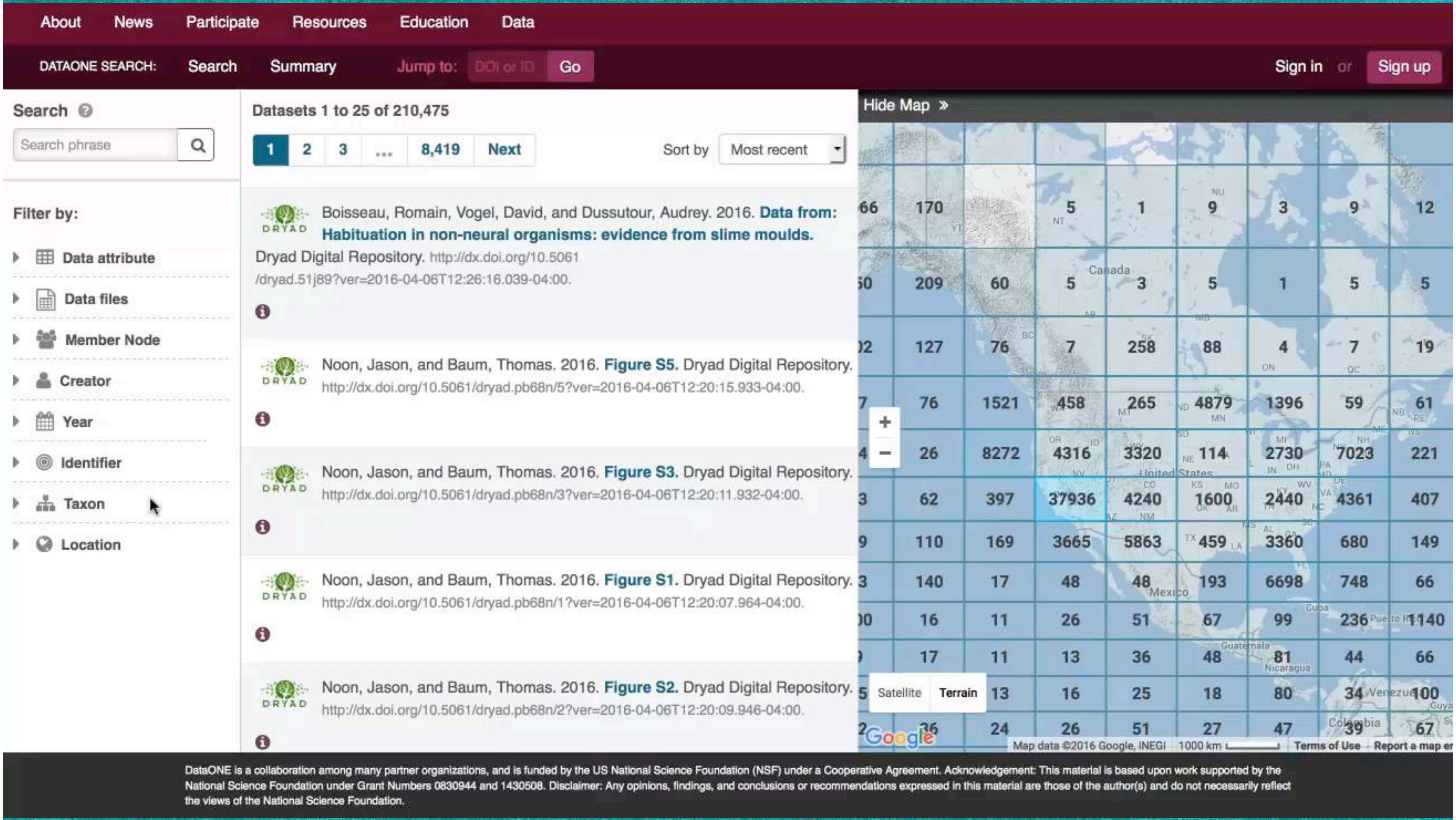


Data Life Cycle



DataONE Search

bit.ly/DWS_S2_04



Datasets 1 to 25 of 210,475

Boisseau, Romain, Vogel, David, and Dussutour, Audrey. 2016. **Data from: Habituation in non-neural organisms: evidence from slime moulds.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.51j89?ver=2016-04-06T12:26:16.039-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S5.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/5?ver=2016-04-06T12:20:15.933-04:00.

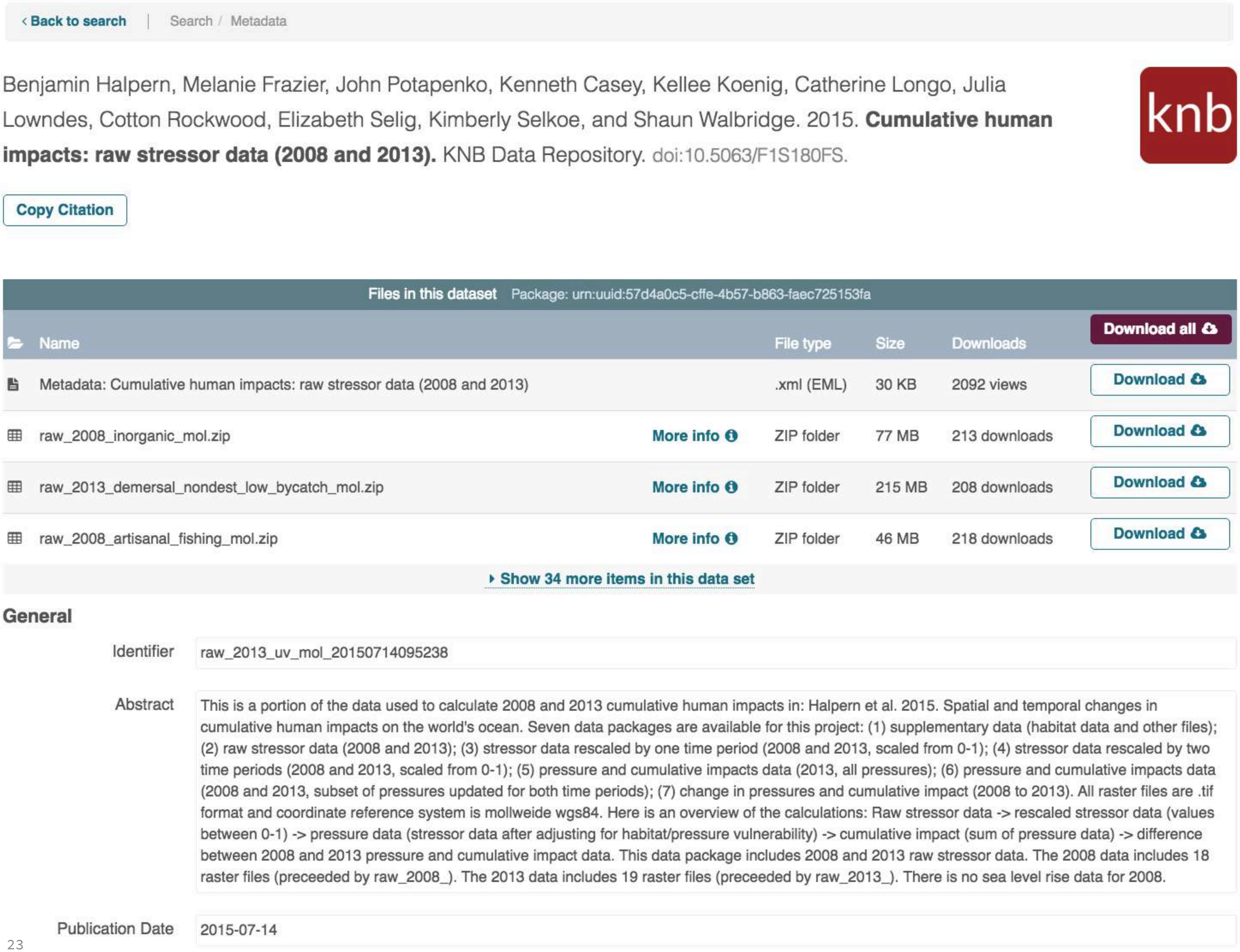
Noon, Jason, and Baum, Thomas. 2016. **Figure S3.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/3?ver=2016-04-06T12:20:11.932-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S1.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/1?ver=2016-04-06T12:20:07.964-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S2.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/2?ver=2016-04-06T12:20:09.946-04:00.

66	170		5	1	9	3	9	12
50	209	60	5	3	5	1	5	5
02	127	76	7	258	88	4	7	19
7	76	1521	458	265	4879	1396	59	61
4	26	8272	4316	3320	114	2730	7023	221
3	62	397	37936	4240	1600	2440	4361	407
9	110	169	3665	5863	459	3360	680	149
3	140	17	48	48	193	6698	748	66
00	16	11	26	51	67	99	236	1140
	17	11	13	36	48	81	44	66
5		13	16	25	18	80	34	100
2	36	24	26	51	27	47	39	67

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.



Benjamin Halpern, Melanie Frazier, John Potapenko, Kenneth Casey, Kellee Koenig, Catherine Longo, Julia Lowndes, Cotton Rockwood, Elizabeth Selig, Kimberly Selkoe, and Shaun Walbridge. 2015. **Cumulative human impacts: raw stressor data (2008 and 2013).** KNB Data Repository. doi:10.5063/F1S180FS.

Files in this dataset Package: urn:uuid:57d4a0c5-cffe-4b57-b863-faec725153fa

Name		File type	Size	Downloads	Download all
Metadata: Cumulative human impacts: raw stressor data (2008 and 2013)		.xml (EML)	30 KB	2092 views	Download
raw_2008_inorganic_mol.zip	More info	ZIP folder	77 MB	213 downloads	Download
raw_2013_demersal_nondest_low_bycatch_mol.zip	More info	ZIP folder	215 MB	208 downloads	Download
raw_2008_artisanal_fishing_mol.zip	More info	ZIP folder	46 MB	218 downloads	Download

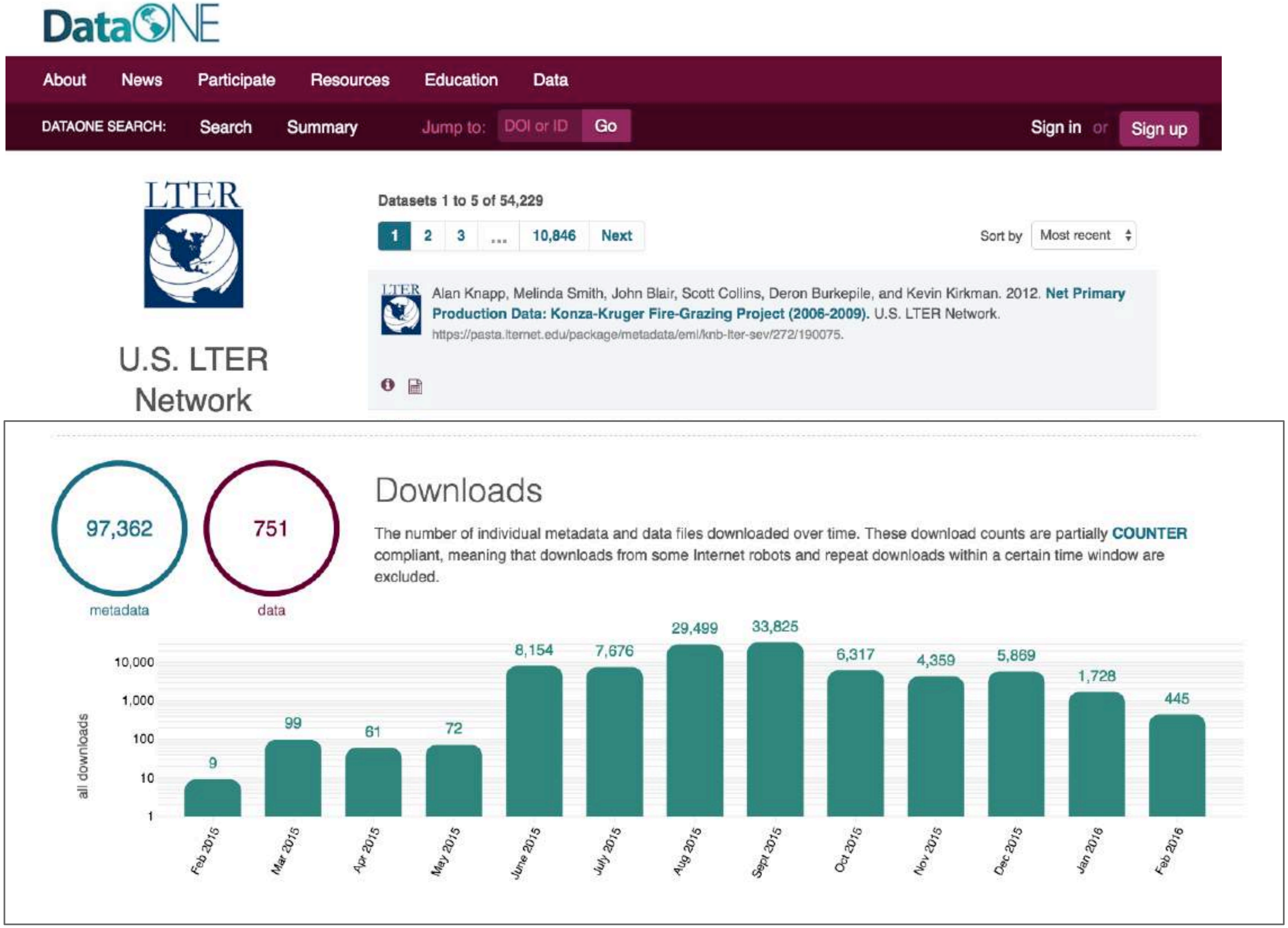
General

Identifier: raw_2013_uv_mol_20150714095238

Abstract: This is a portion of the data used to calculate 2008 and 2013 cumulative human impacts in: Halpern et al. 2015. Spatial and temporal changes in cumulative human impacts on the world's ocean. Seven data packages are available for this project: (1) supplementary data (habitat data and other files); (2) raw stressor data (2008 and 2013); (3) stressor data rescaled by one time period (2008 and 2013, scaled from 0-1); (4) stressor data rescaled by two time periods (2008 and 2013, scaled from 0-1); (5) pressure and cumulative impacts data (2013, all pressures); (6) pressure and cumulative impacts data (2008 and 2013, subset of pressures updated for both time periods); (7) change in pressures and cumulative impact (2008 to 2013). All raster files are .tif format and coordinate reference system is mollweide wgs84. Here is an overview of the calculations: Raw stressor data -> rescaled stressor data (values between 0-1) -> pressure data (stressor data after adjusting for habitat/pressure vulnerability) -> cumulative impact (sum of pressure data) -> difference between 2008 and 2013 pressure and cumulative impact data. This data package includes 2008 and 2013 raw stressor data. The 2008 data includes 18 raster files (preceeded by raw_2008_). The 2013 data includes 19 raster files (preceeded by raw_2013_). There is no sea level rise data for 2008.

Publication Date: 2015-07-14

Member Node Profiles



U.S. LTER Network

Datasets 1 to 5 of 54,229

Alan Knapp, Melinda Smith, John Blair, Scott Collins, Deron Burkepile, and Kevin Kirkman. 2012. **Net Primary Production Data: Konza-Kruger Fire-Grazing Project (2006-2009).** U.S. LTER Network. https://pasta.lternet.edu/package/metadata/eml/knb-lter-sev/272/190075.

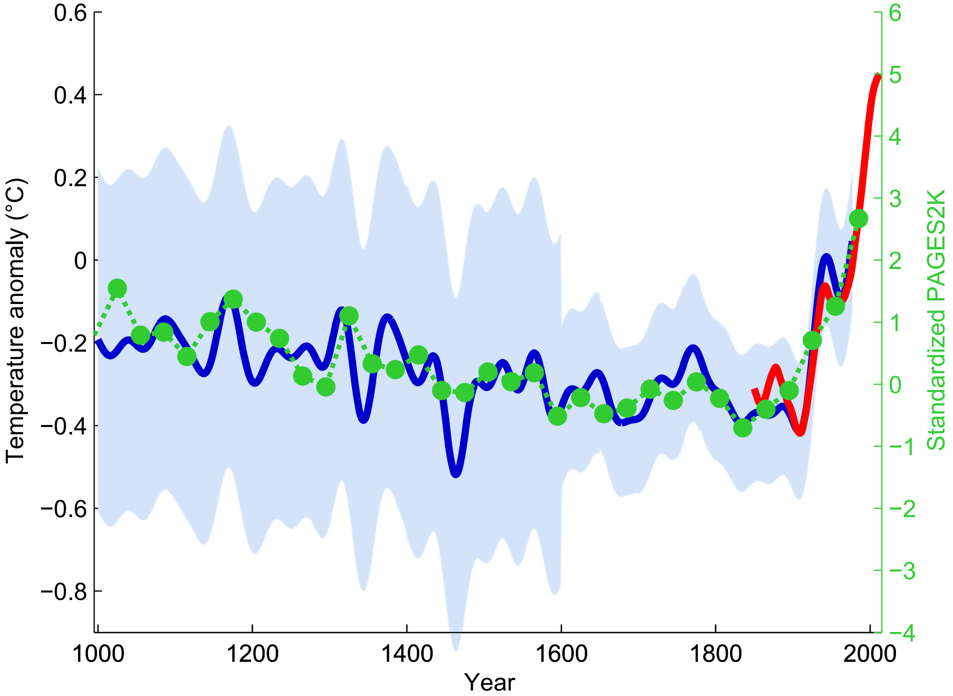
97,362 metadata

751 data

Downloads

The number of individual metadata and data files downloaded over time. These download counts are partially **COUNTER** compliant, meaning that downloads from some Internet robots and repeat downloads within a certain time window are excluded.

Use Provenance for
Transparency, Reproducibility



What ***input data*** went into this study?

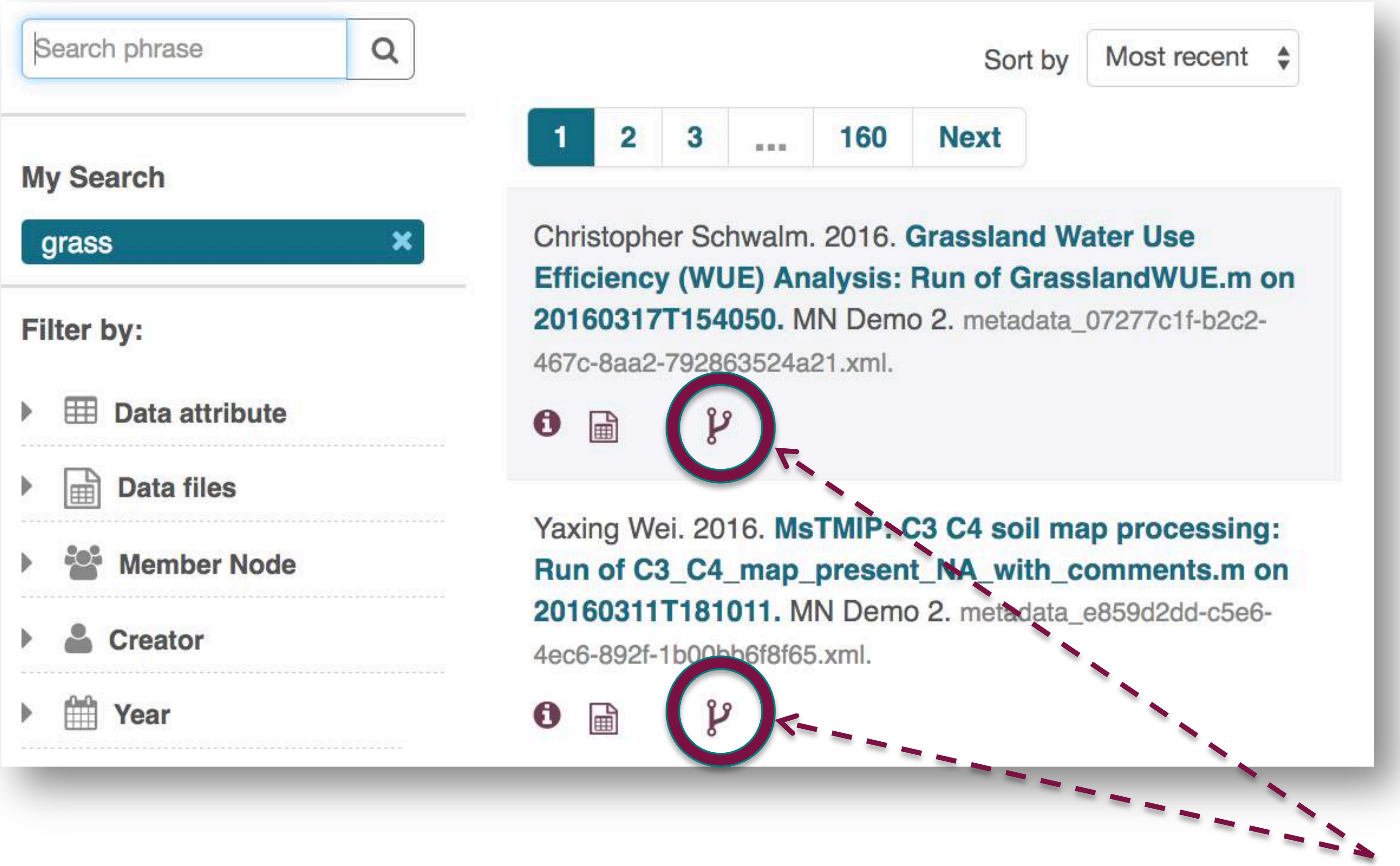
What ***methods*** were used?

… with what ***parameter*** settings, ***calibrations***, …?

*Can we **trust** the data and methods?*

- **Provenance** (*lineage*): track **origin** and **processing history** of data ➔ trust, data quality ~ audit trail for attribution, credit
- **Discovery** of data, methodologies, experiments

Dataset Provenance



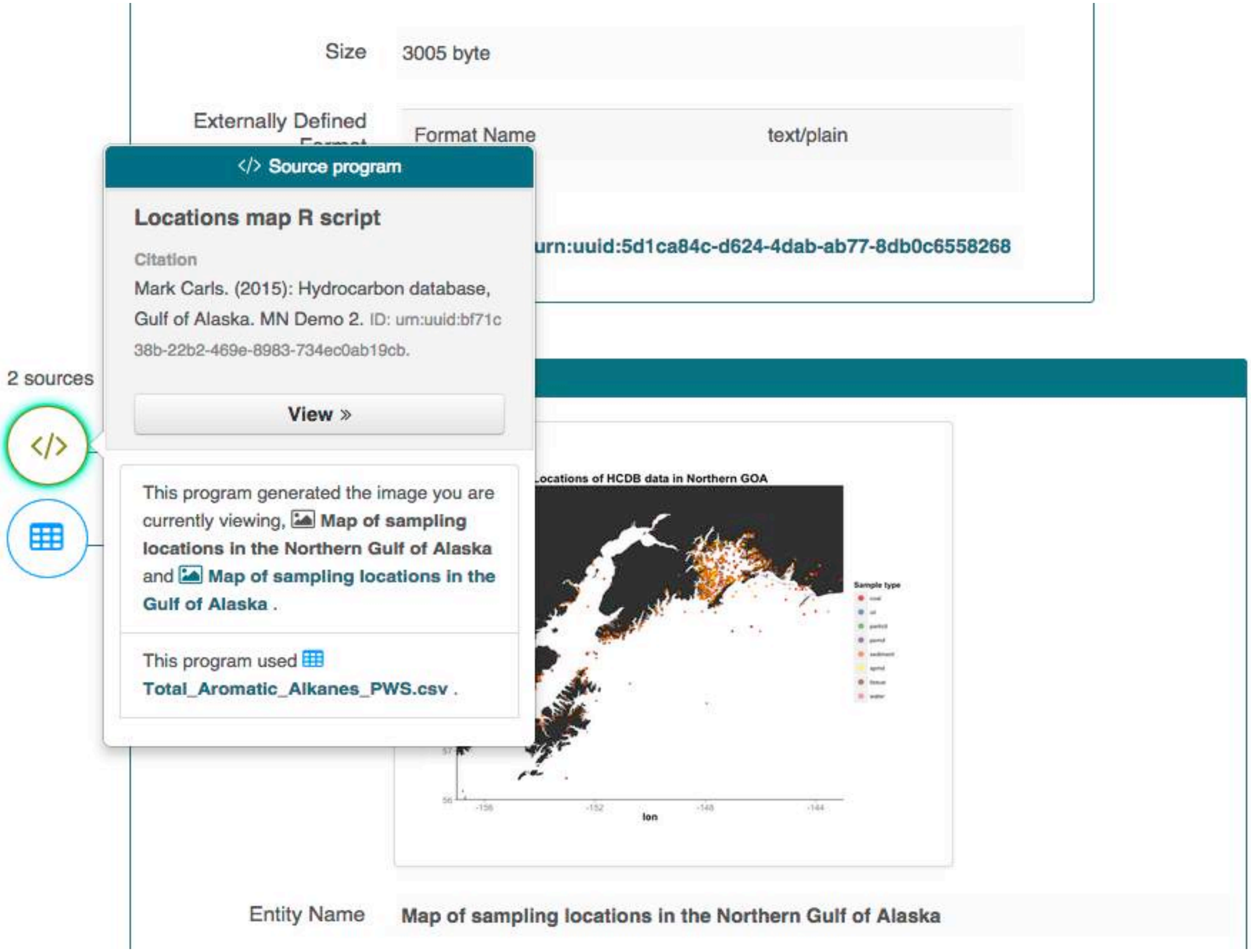
Christopher Schwalm. 2016. **Grassland Water Use Efficiency (WUE) Analysis: Run of GrasslandWUE.m on 20160317T154050.** MN Demo 2. metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml.

Yaxing Wei. 2016. **MsTMIP: C3 C4 soil map processing: Run of C3_C4_map_present_NA_with_comments.m on 20160311T181011.** MN Demo 2. metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml.

Record contains provenance information

Provenance

… of Figures



Size 3005 byte

Externally Defined Format

Format Name text/plain

urn:uuid:5d1ca84c-d624-4dab-ab77-8db0c6558268

2 sources

</> Source program

Locations map R script

Citation

Mark Carls. (2015): Hydrocarbon database, Gulf of Alaska. MN Demo 2. ID: urn:uuid:bf71c38b-22b2-469e-8983-734ec0ab19cb.

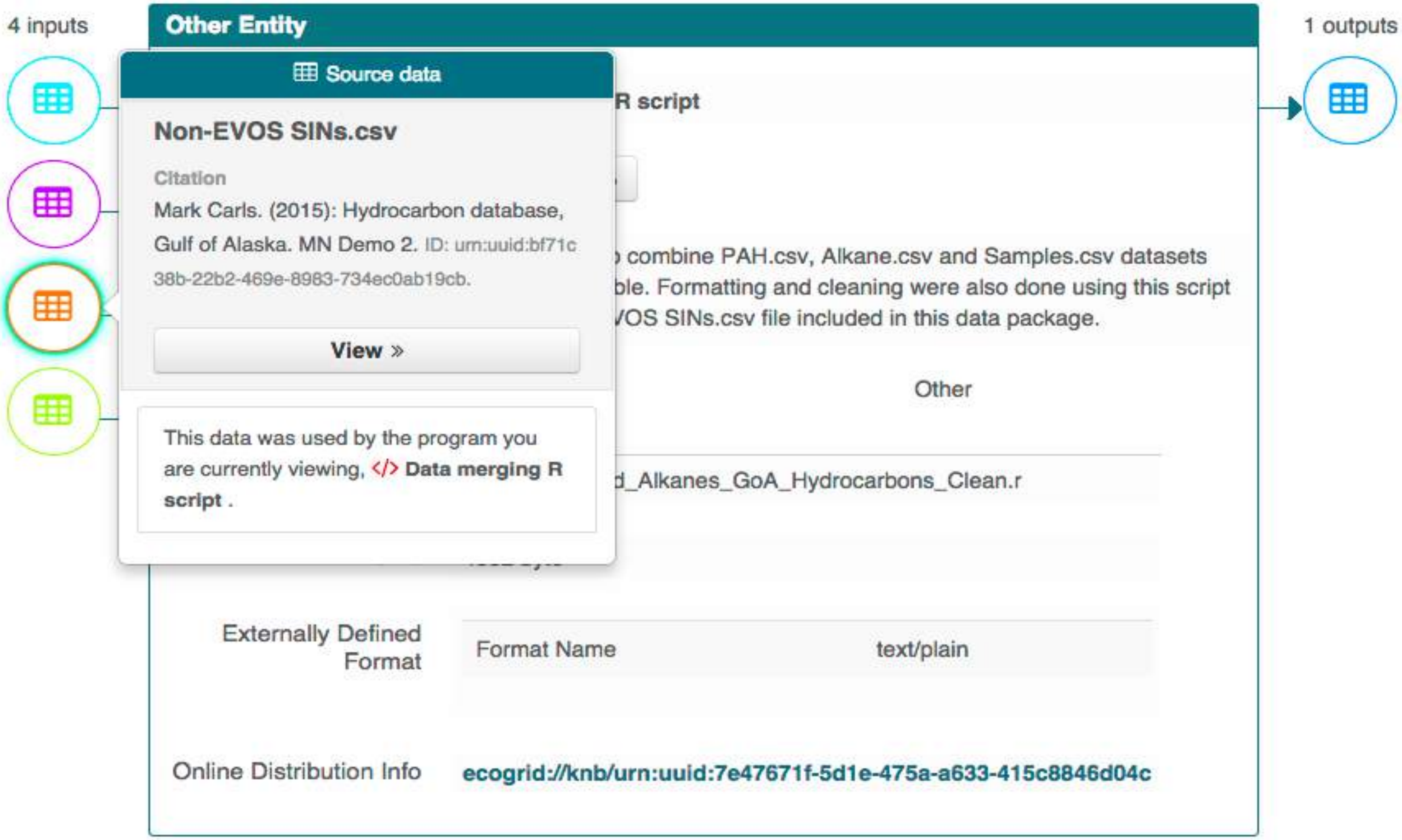
View »

This program generated the image you are currently viewing, **Map of sampling locations in the Northern Gulf of Alaska** and **Map of sampling locations in the Gulf of Alaska** .

This program used **Total_Aromatic_Alkanes_PWS.csv** .

Entity Name **Map of sampling locations in the Northern Gulf of Alaska**

Provenance
… of Data



4 inputs

Other Entity

Source data

Non-EVOS SINs.csv

Citation

Mark Carls. (2015): Hydrocarbon database, Gulf of Alaska. MN Demo 2. ID: urn:uuid:bf71c38b-22b2-469e-8983-734ec0ab19cb.

This data was used by the program you are currently viewing, </> **Data merging R script** .

R script

combine PAH.csv, Alkane.csv and Samples.csv datasets
ble. Formatting and cleaning were also done using this script
/OS SINs.csv file included in this data package.

Other

d_Alkanes_GoA_Hydrocarbons_Clean.r

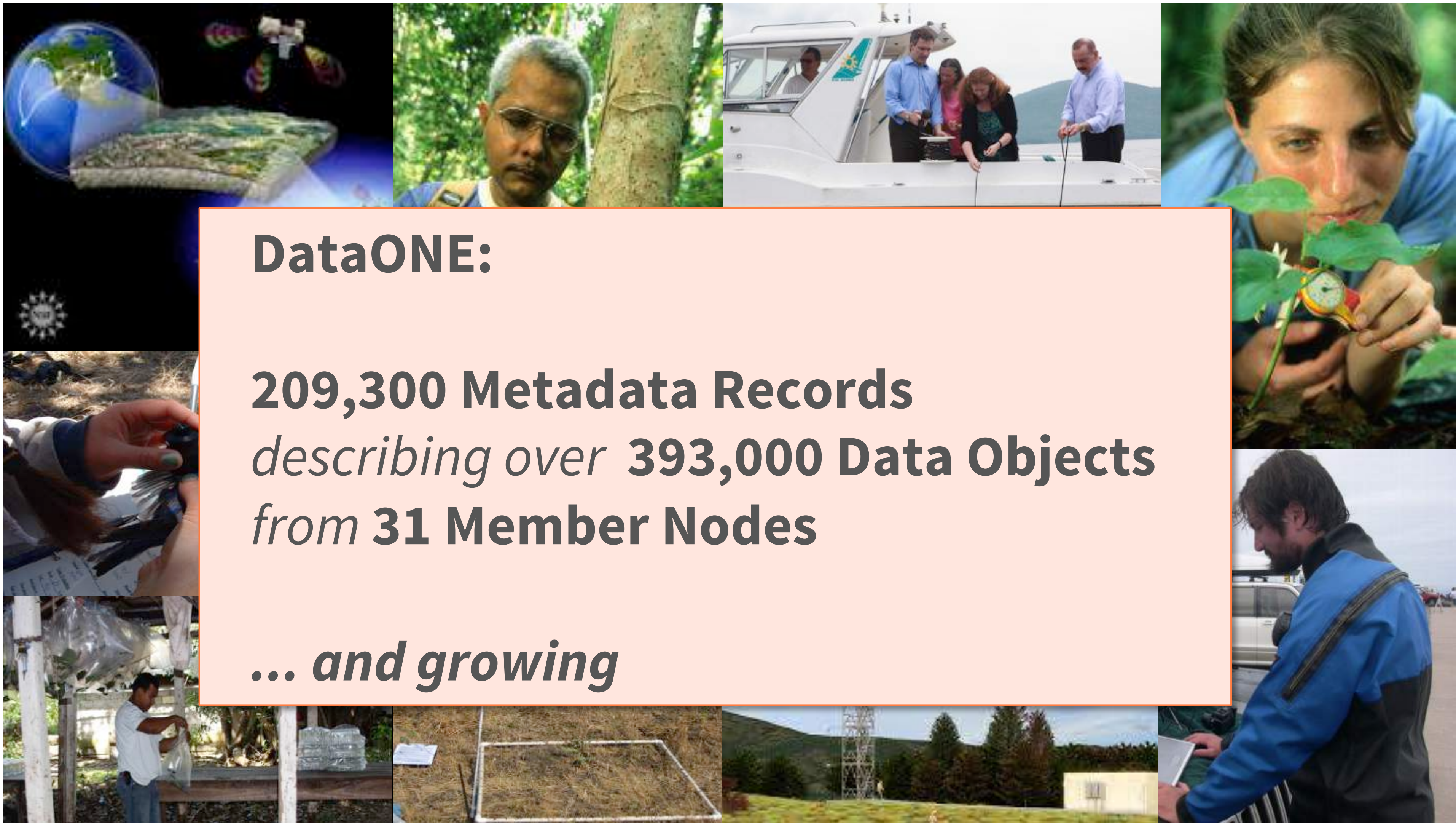
Externally Defined Format	Format Name	text/plain

Online Distribution Info ecogrid://knb/urn:uuid:7e47671f-5d1e-475a-a633-415c8846d04c

1 outputs

bit.ly/DWS_01_04

The Problem: Enabling researchers to effectively find data in DataONE



DataONE:

209,300 Metadata Records
describing over **393,000 Data Objects**
from **31 Member Nodes**

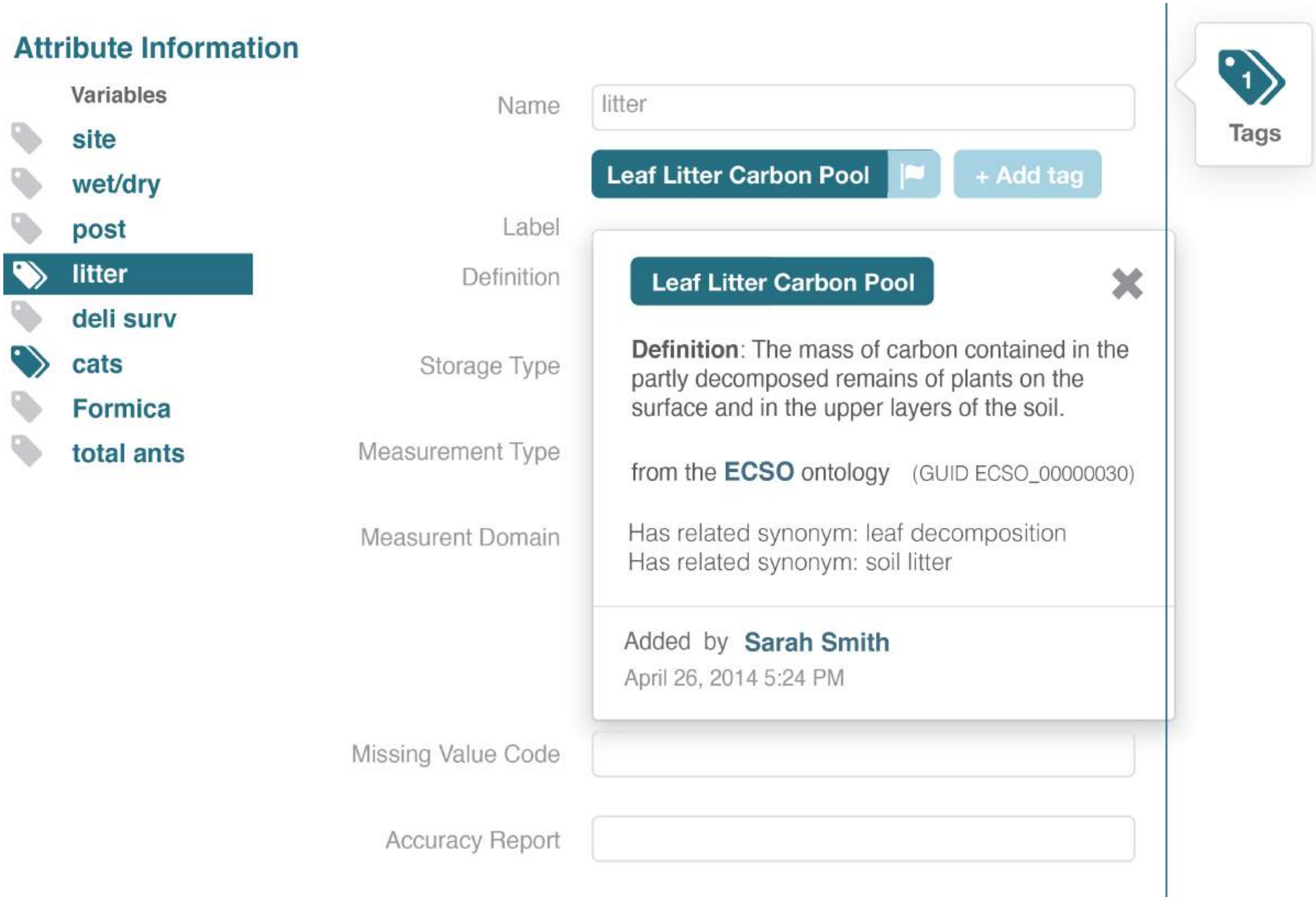
… and growing

Semantics

For greater clarity and consistency

Litter?

Displaying semantics of attribute labels



Attribute Information

Variables

- site
- wet/dry
- post
- **litter**
- deli surv
- cats
- Formica
- total ants

Name: litter

Leaf Litter Carbon Pool

Label

Definition

Storage Type

Measurement Type

Measurent Domain

Leaf Litter Carbon Pool

Definition: The mass of carbon contained in the partly decomposed remains of plants on the surface and in the upper layers of the soil.

from the **ECSO** ontology (GUID ECSO_00000030)

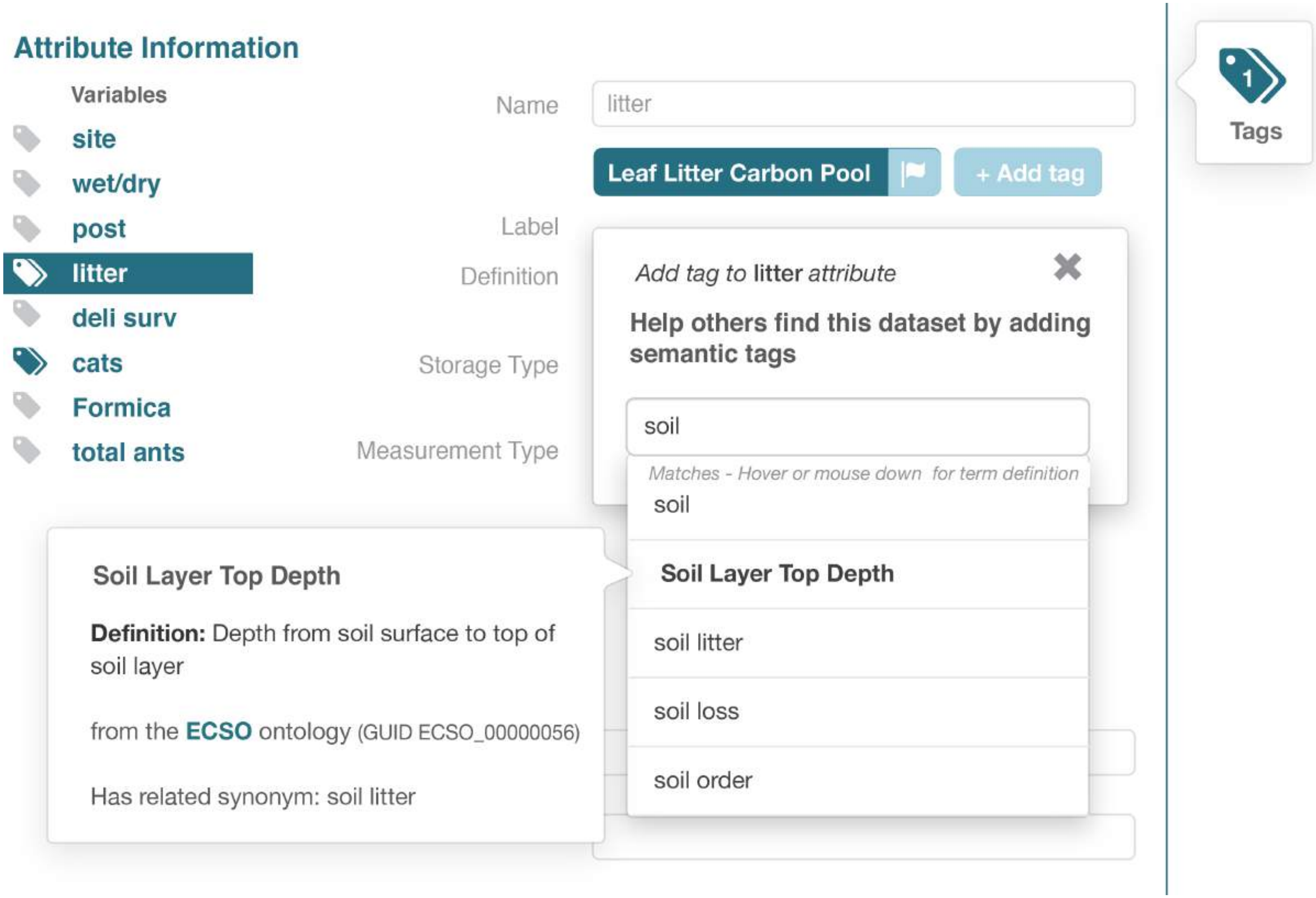
Has related synonym: leaf decomposition
Has related synonym: soil litter

Added by **Sarah Smith**
April 26, 2014 5:24 PM

Missing Value Code

Accuracy Report

Manual annotation UI



Name: litter

Add tag to litter *attribute*

Help others find this dataset by adding semantic tags

soil

Matches - Hover or mouse down for term definition

- soil
- **Soil Layer Top Depth**
- soil litter
- soil loss
- soil order

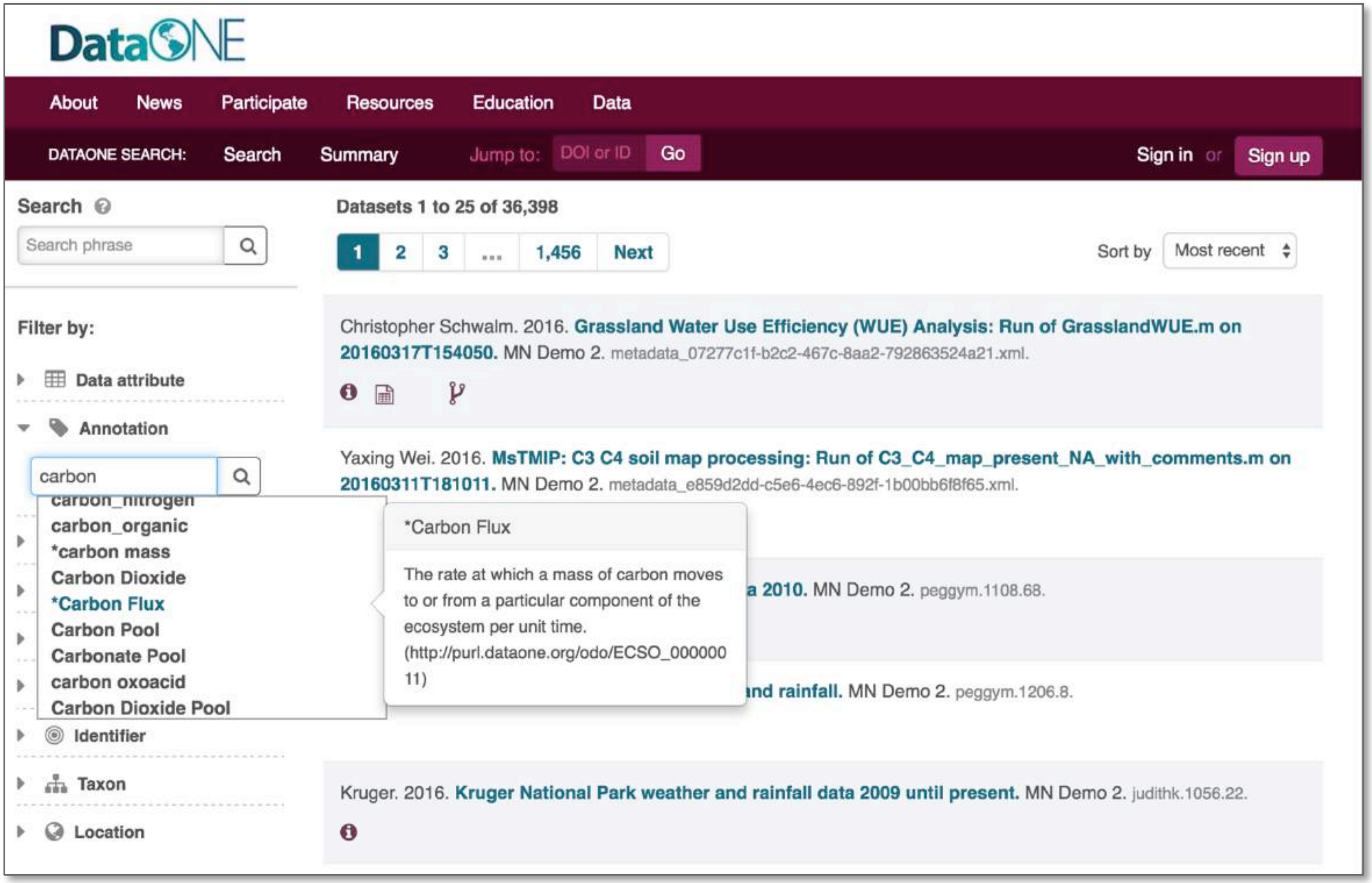
Soil Layer Top Depth

Definition: Depth from soil surface to top of soil layer

from the **ECSO** ontology (GUID ECSO_00000056)

Has related synonym: soil litter

Semantic search



Datasets 1 to 25 of 36,398

Annotation: carbon

carbon_nitrogen
carbon_organic
*carbon mass
Carbon Dioxide
*Carbon Flux
Carbon Pool
Carbonate Pool
carbon oxoacid
Carbon Dioxide Pool

*Carbon Flux

The rate at which a mass of carbon moves to or from a particular component of the ecosystem per unit time. (http://purl.dataone.org/odo/ECSO_00000011)

Christopher Schwalm. 2016. **Grassland Water Use Efficiency (WUE) Analysis: Run of GrasslandWUE.m on 20160317T154050.** MN Demo 2. metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml.

Yaxing Wei. 2016. **MsTMIP: C3 C4 soil map processing: Run of C3_C4_map_present_NA_with_comments.m on 20160311T181011.** MN Demo 2. metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml.

a 2010. MN Demo 2. peggym.1108.68.

nd rainfall. MN Demo 2. peggym.1206.8.

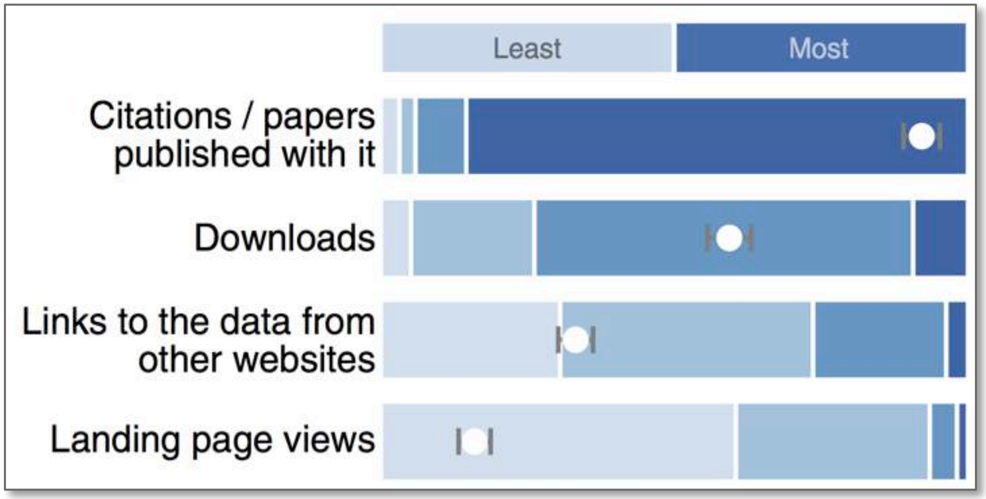
Kruger. 2016. **Kruger National Park weather and rainfall data 2009 until present.** MN Demo 2. judithk.1056.22.

Community Need
Data Use Metrics

Challenge: Data citation and usage reporting are rare, difficult to find, but highly valuable

Goal: Index the science literature to provide citation and usage metrics for data and software in DataONE

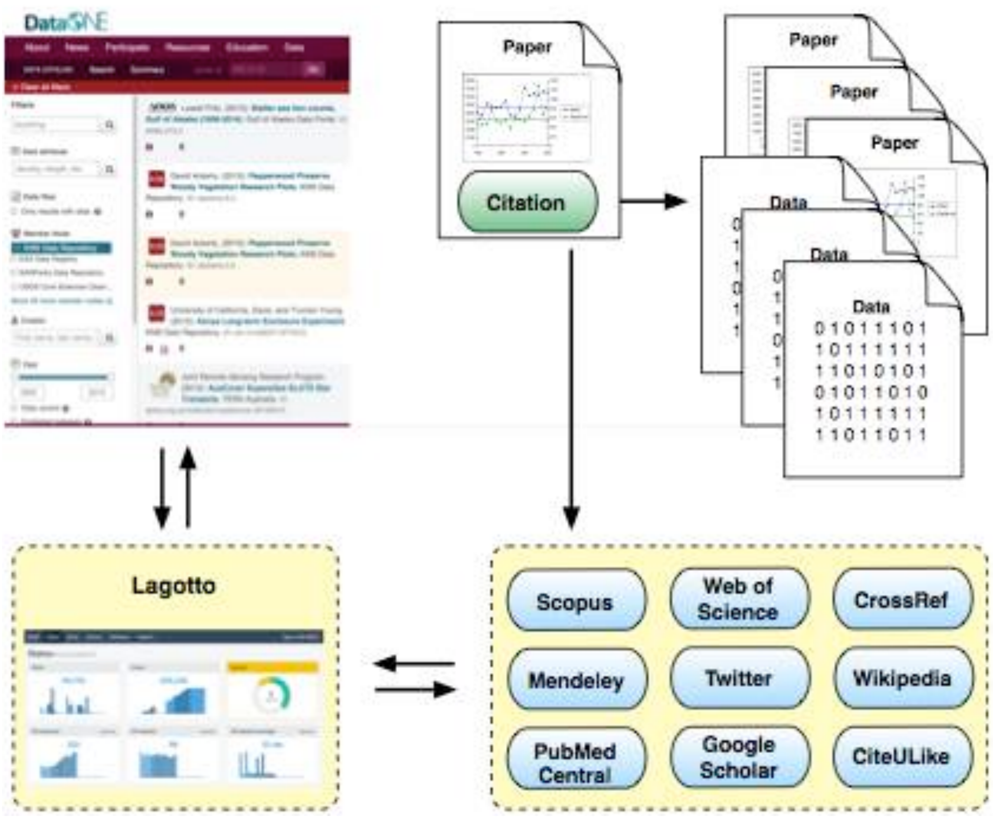
How interested would you be to know each of the following about the impact of your data?



Kratz and Strasser (2015) doi:10.1038/sdata.2015.39

Data Use Metrics
Approach

- Leverage ‘Making Data Count’ prototype
- Index usage and citation in papers and open access sources
- Powerful reports for users, repositories, and funders



* Kratz and Strasser (2015) “Making Data Count”, Scientific Data (Nature). doi: 10.1038/sdata.2015.39

Data Use Metrics

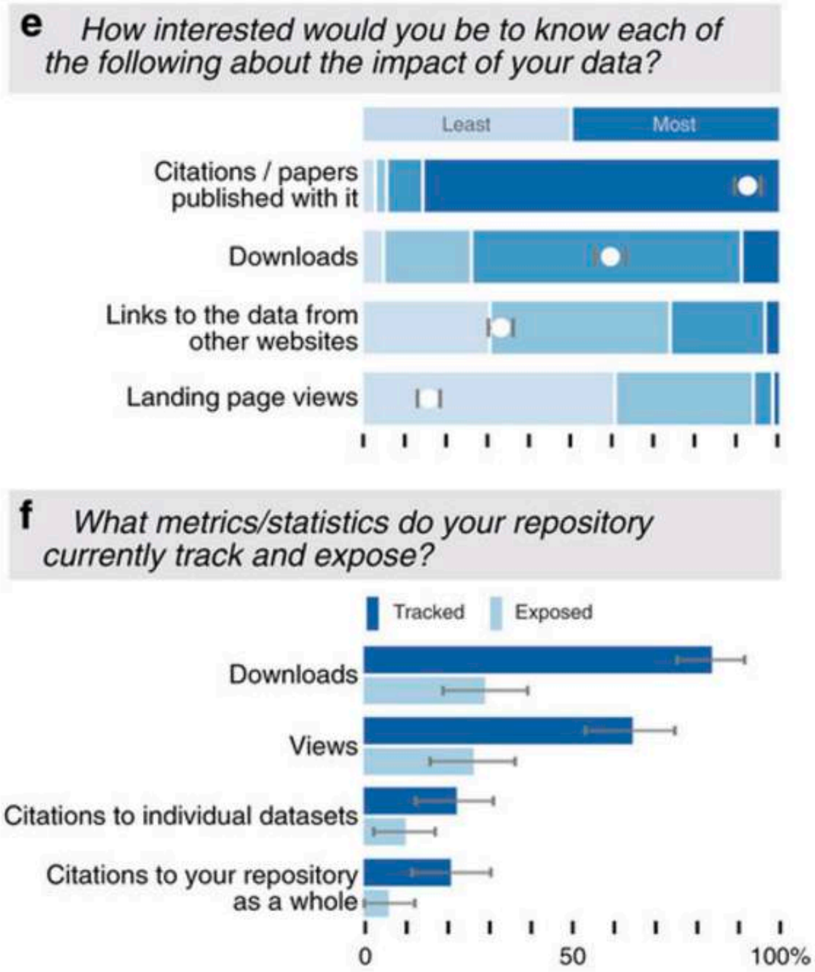
Outputs

For users and repositories:

- Citation and usage services
  - with DataCite
  - interactive displays, reports
- Notification services
  - when cited, by whom…

For funders:

- Per-award reports
- Program-wide reports
- Impact assessments



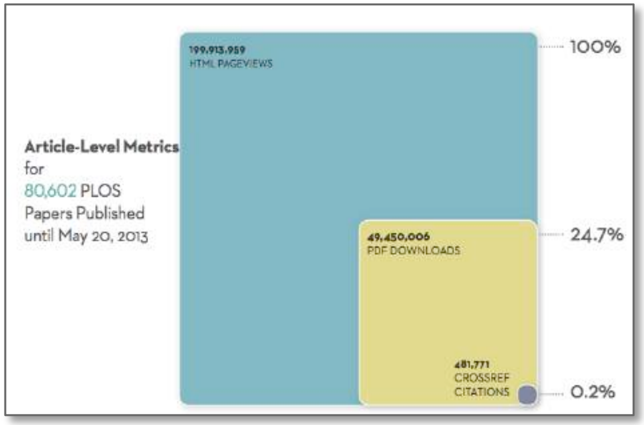
e *How interested would you be to know each of the following about the impact of your data?*

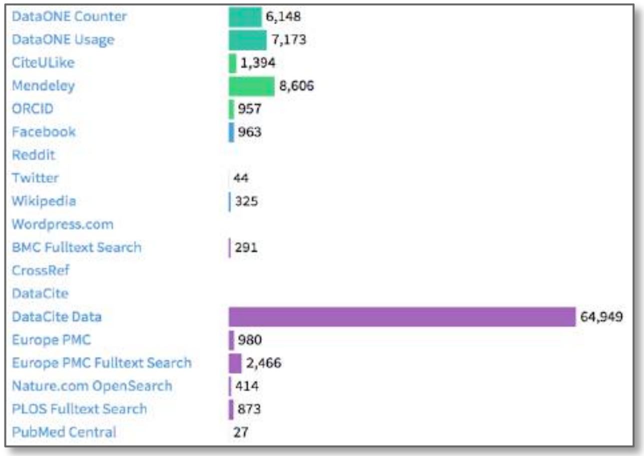
f *What metrics/statistics do your repository currently track and expose?*

graphic from Kratz and Strasser (2015). doi: 10.1038/sdata.2015.39

Data Use Metrics
Outcomes

- **Enable Greater Attribution**
  - Article level
  - micro-citation
- **Enhance Resource Discovery**
  - Greater motivation to share
  - More resources to explore
- **Build Community Engagement**
  - Awareness of others' work
- ***Promote Reproducible Science***





graphics from: Lin and Fenner, 2013 Information Standards Quarterly; doi: 10.3789/isqv25no2.2013.04

bit.ly/DWS_01_03

Technical Resources

Architecture and API Documentation

- purl.dataone.org/architecture

Mailing List

- developers@dataone.org

IRC

- irc.ecoinformatics.org #dataone

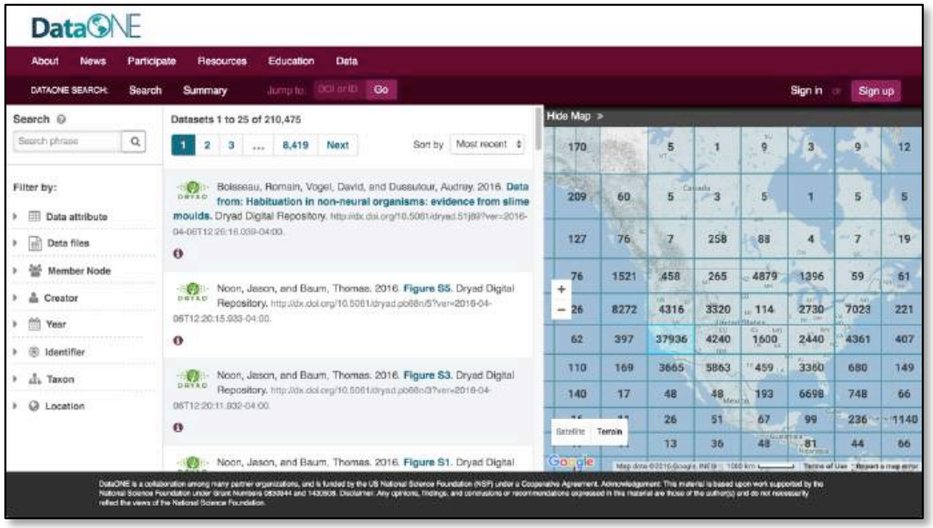
Subversion, GitHub

- repository.dataone.org/software/cicore
- github.com/DataONEorg

Previous Webinars:

- dataone.org/previous-webinars

Community Engagement
Education and Outreach



Search

Search phrase

Filter by:

- Data attribute
- Data files
- Member Node
- Creator
- Year
- Identifier
- Taxon
- Location

Datasets 1 to 25 of 210,475

1 2 3 ... 8,419 Next

Sort by Most recent

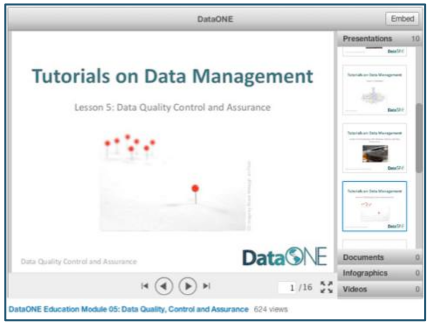
Boisseau, Romain, Vogel, David, and Dussutour, Audrey. 2016. **Data from: Habituation in non-neural organisms: evidence from slime moulds.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.51j89?ver=2016-04-06T12:26:16.039-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S5.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/5?ver=2016-04-06T12:20:15.933-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S3.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/3?ver=2016-04-06T12:20:11.832-04:00.

Noon, Jason, and Baum, Thomas. 2016. **Figure S1.** Dryad Digital

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

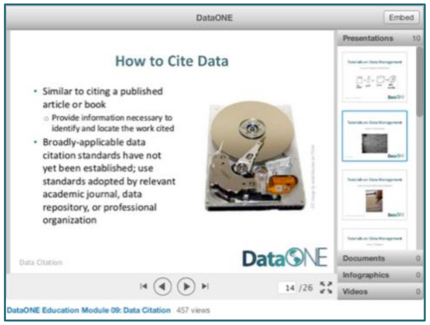


Tutorials on Data Management

Lesson 5: Data Quality Control and Assurance

Data Quality Control and Assurance

DataONE Education Module 05: Data Quality, Control and Assurance 624 views

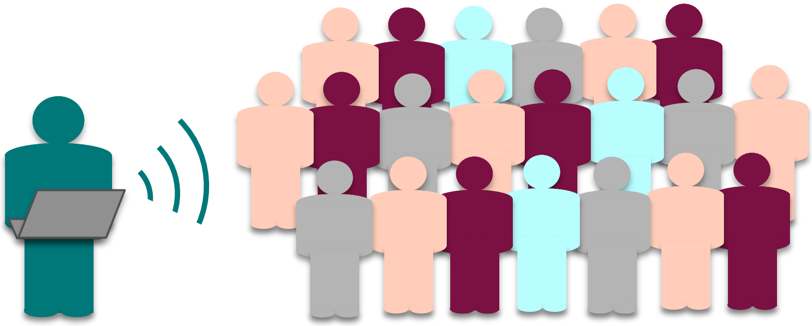


How to Cite Data

- Similar to citing a published article or book
  - Provide information necessary to identify and locate the work cited
- Broadly-applicable data citation standards have not yet been established; use standards adopted by relevant academic journal, data repository, or professional organization

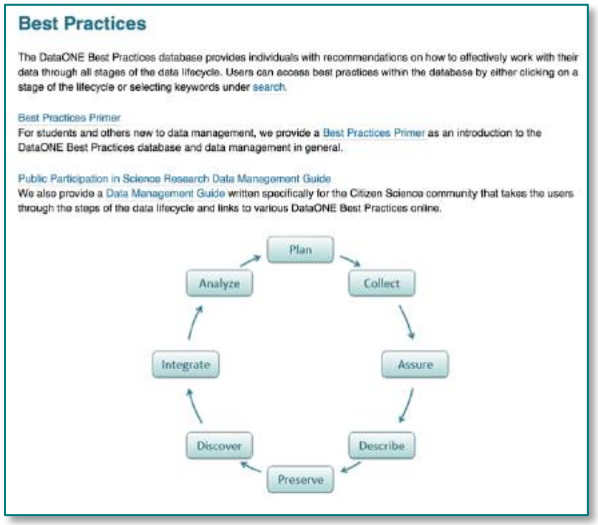
Data Citation

DataONE Education Module 09: Data Citation 457 views



COS ESIP iDigBio Integrated Digitized Biocollections SEAD

Best Practices
Database and Primer

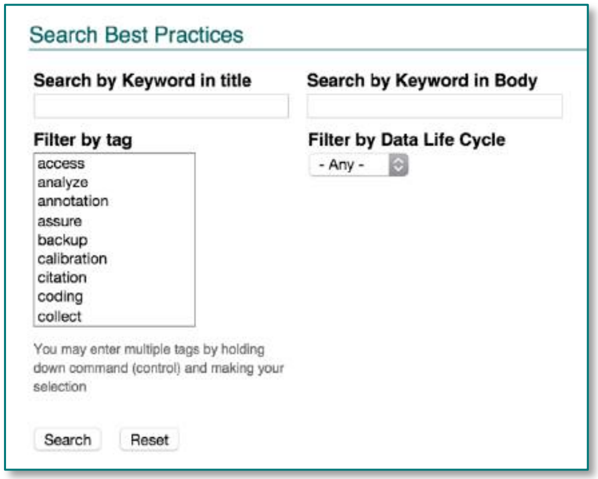


Best Practices

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

Best Practices Primer
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.

Public Participation in Science Research Data Management Guide
We also provide a Data Management Guide written specifically for the Citizen Science community that takes the users through the steps of the data lifecycle and links to various DataONE Best Practices online.



Search Best Practices

Search by Keyword in title

Search by Keyword in Body

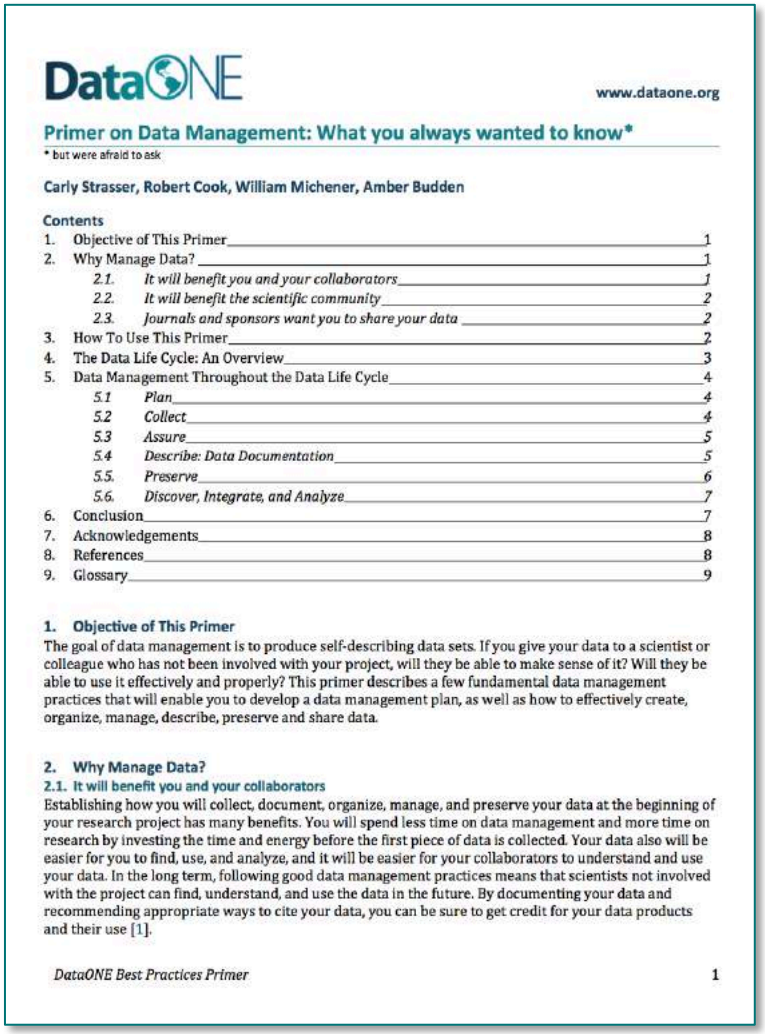
Filter by tag

access
analyze
annotation
assure
backup
calibration
citation
coding
collect

Filter by Data Life Cycle

- Any -

You may enter multiple tags by holding down command (control) and making your selection



DataONE

www.dataone.org

Primer on Data Management: What you always wanted to know*

* but were afraid to ask

Carly Strasser, Robert Cook, William Michener, Amber Budden

Contents

1. Objective of This Primer

The goal of data management is to produce self-describing data sets. If you give your data to a scientist or colleague who has not been involved with your project, will they be able to make sense of it? Will they be able to use it effectively and properly? This primer describes a few fundamental data management practices that will enable you to develop a data management plan, as well as how to effectively create, organize, manage, describe, preserve and share data.

2. Why Manage Data?

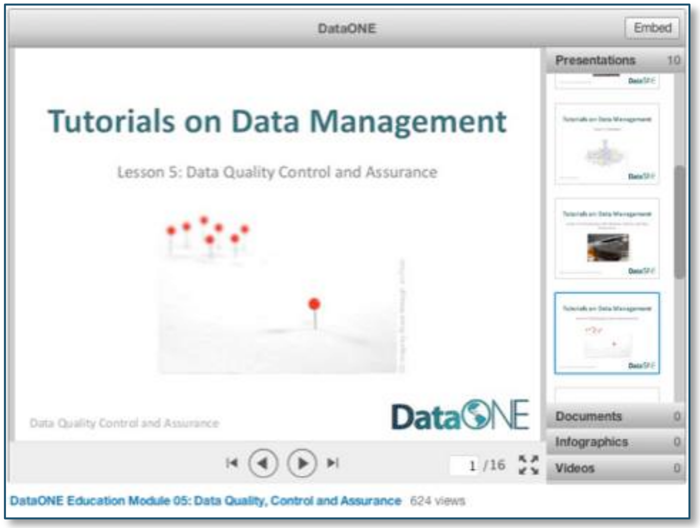
2.1. It will benefit you and your collaborators

Establishing how you will collect, document, organize, manage, and preserve your data at the beginning of your research project has many benefits. You will spend less time on data management and more time on research by investing the time and energy before the first piece of data is collected. Your data also will be easier for you to find, use, and analyze, and it will be easier for your collaborators to understand and use your data. In the long term, following good data management practices means that scientists not involved with the project can find, understand, and use the data in the future. By documenting your data and recommending appropriate ways to cite your data, you can be sure to get credit for your data products and their use [1].

DataONE Best Practices Primer 1

www.dataone.org/best-practices

Data Management Modules

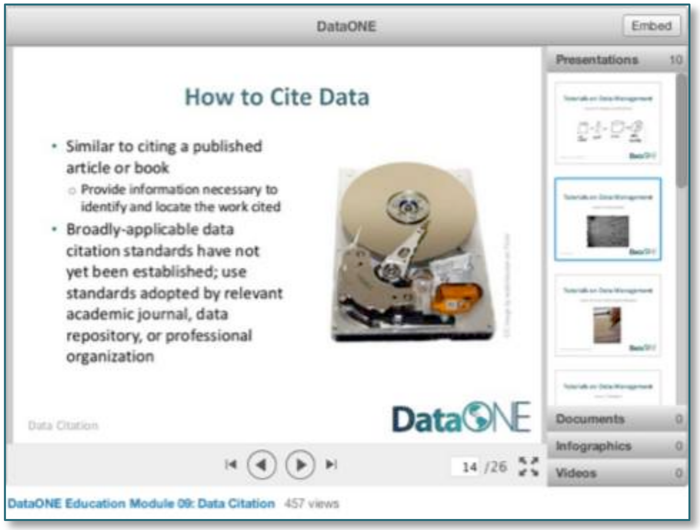


Tutorials on Data Management

Lesson 5: Data Quality Control and Assurance

Data Quality Control and Assurance

DataONE Education Module 05: Data Quality, Control and Assurance 624 views



How to Cite Data

- Similar to citing a published article or book
  - Provide information necessary to identify and locate the work cited
- Broadly-applicable data citation standards have not yet been established; use standards adopted by relevant academic journal, data repository, or professional organization

Data Citation

DataONE Education Module 09: Data Citation 457 views

Lesson 10: Analysis and Workflows

View all Education Modules at https://www.dataone.org/education-modules

Typical data analyses

Data processing: may include selecting a subset of data for analysis, merging multiple data sets, manipulating data for usability, or data transformation

Graphical analysis: makes it easier to see patterns and can aid in the identification of outliers

Statistical analysis: conventional statistics are used to analyze experimental data; descriptive statistics are used to analyze observational or descriptive data

Science is iterative: the process that results in the product can be complex.

Reproducibility..

...is at the core of the scientific process. If result not reproducible, they lose credibility.

Good documentation of the data and the anal are essential!

Workflows

Definition: Precise description of the procedu used in a project. Can be formal or informa

Informal workflow

No special software is needed to create work diagrams. Workflow diagrams include:

- Inputs and outputs
- Transformation rules or analytical proce
- Decision points
- Arrows indicating direction of process fl

Informal Workflow Example

Formal Workflow

Analytical pipeline where each step can be implemented in different software systems. Parameters and requirements for each step are formally recorded.

- Single access point for multiple analyses across software packages
- Keeps track of analysis and provenance to better enable reproducibility
- Workflow can be stored

Hands-on Activity 1: Accessing Data in the Literature

Associated DataONE Lecture: Lesson 1: *Why Data Management*

Objectives: Students recognize the value of accessibly archived data, by experiencing the challenges of accessing data from published papers.

Outcomes: (1) Students can explain why accessible data archiving is valuable. (2) Students can provide strategies for getting data from published papers, and anticipate challenges to accessing the data.

Time Needed: One hour out-of-class, 15 – 30 minutes in-class discussion.

URLs: Any resource for searching scientific literature (e.g. *Web of Science*, *Google Scholar*, *JSTOR*, *BioOne*).

Additional Files Needed: None

Key Reading: Carly A Strasser and Stephanie E Hampton. 2012. The fractured lab notebook: undergraduates and ecological data management training in the United States. *Ecosphere* 3:art116. doi: 10.1890/ES12-00139.1

Notes and Instructions for Instructors:

An intended take-home lesson of this activity is that access to valuable original data can become difficult or impossible in a short period of time after a paper is published, but this loss of accessibility is avoidable. How easy it is to access original data depends on the field; some fields have developed a culture of data sharing and data accessibility, including genetics, climate studies, and geography. Others do not have this tradition. Because of these field-specific cultures, students' success at accessing data will depend on the topic and question they chose.

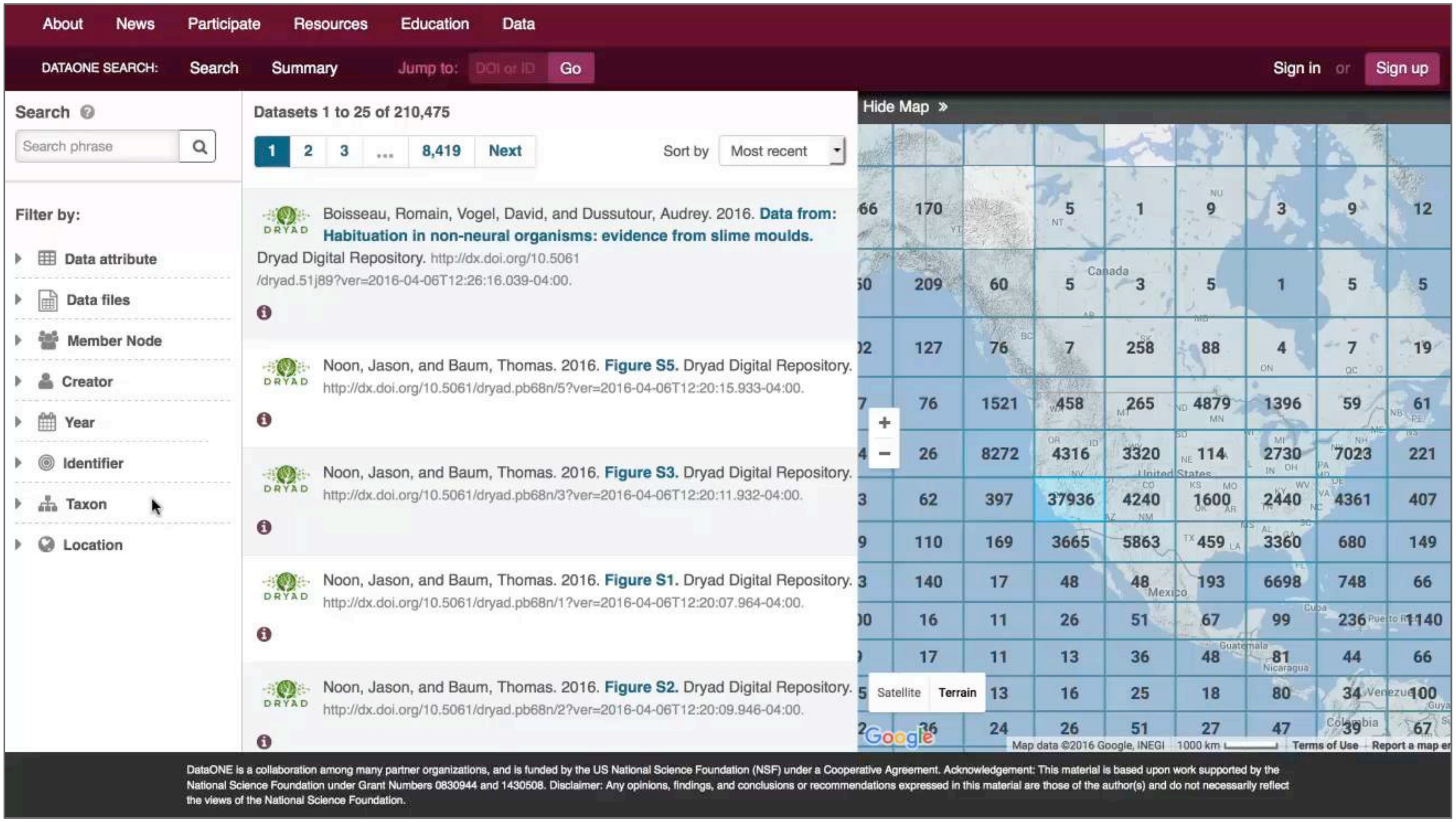
It may be worth reviewing with the students the different ways by which scientists access others' data: data tables or published data appendices within a paper, extracting (estimating) data from published graphs, online data archives or data streams (either restricted to journal subscribers or public), writing the author and requesting the data etc.

After students have completed the exercise (see *Student Instructions*, below), have students discuss the challenges that they faced in figuring out how to access data from the published literature that are relevant to their question, and ways the students came up with to deal with the challenges. This can be done as a 15 to 30 minute whole-class discussion or in small groups with a report-out. Things to note include whether accessibility to data varied depending on the question addressed, and whether accessibility depended on how long ago the paper was published. Perhaps culminate the discussion with questions about why data underlying

Hands–on Exercises for Data Management
http://www.dataone.org/education–modules 1

www.dataone.org/education-modules

Screencast Tutorials



Datasets 1 to 25 of 210,475

DRYAD Boisseau, Romain, Vogel, David, and Dussutour, Audrey. 2016. **Data from: Habituation in non-neural organisms: evidence from slime moulds.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.51j89?ver=2016-04-06T12:26:16.039-04:00.

DRYAD Noon, Jason, and Baum, Thomas. 2016. **Figure S5.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/5?ver=2016-04-06T12:20:15.933-04:00.

DRYAD Noon, Jason, and Baum, Thomas. 2016. **Figure S3.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/3?ver=2016-04-06T12:20:11.932-04:00.

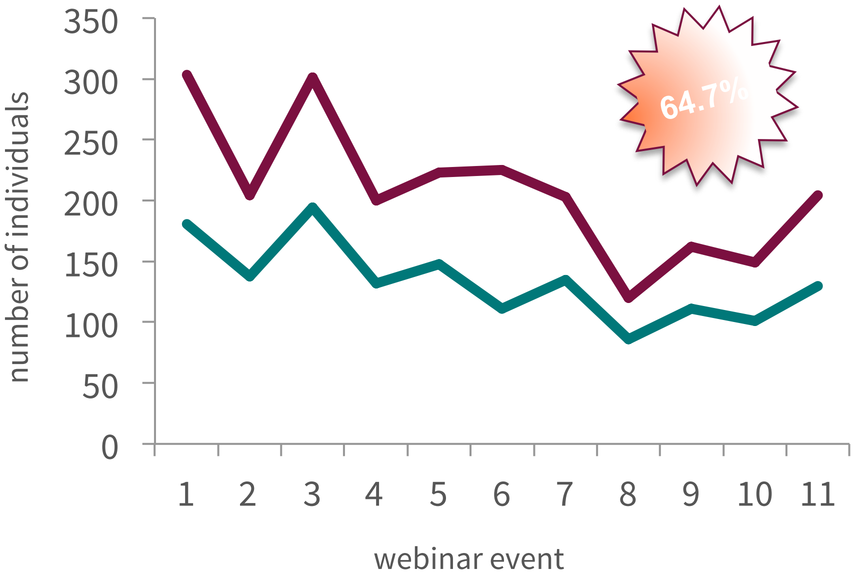
DRYAD Noon, Jason, and Baum, Thomas. 2016. **Figure S1.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/1?ver=2016-04-06T12:20:07.964-04:00.

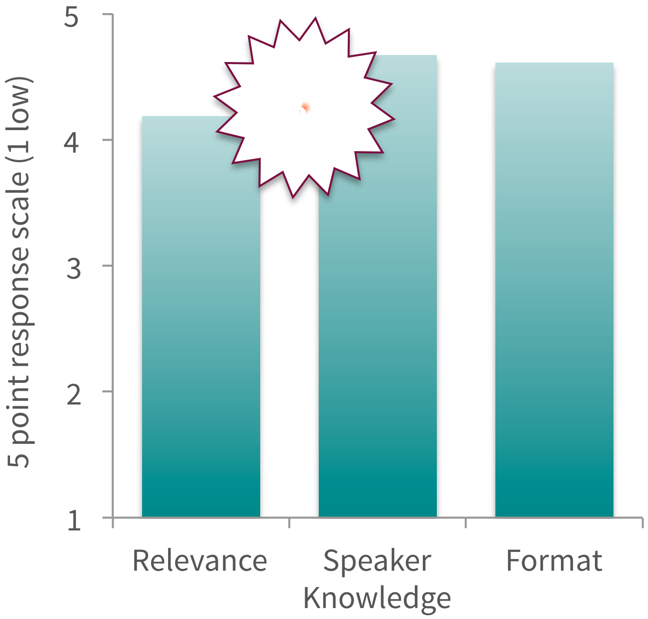
DRYAD Noon, Jason, and Baum, Thomas. 2016. **Figure S2.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.pb68n/2?ver=2016-04-06T12:20:09.946-04:00.

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

www.dataone.org/screencast-tutorials

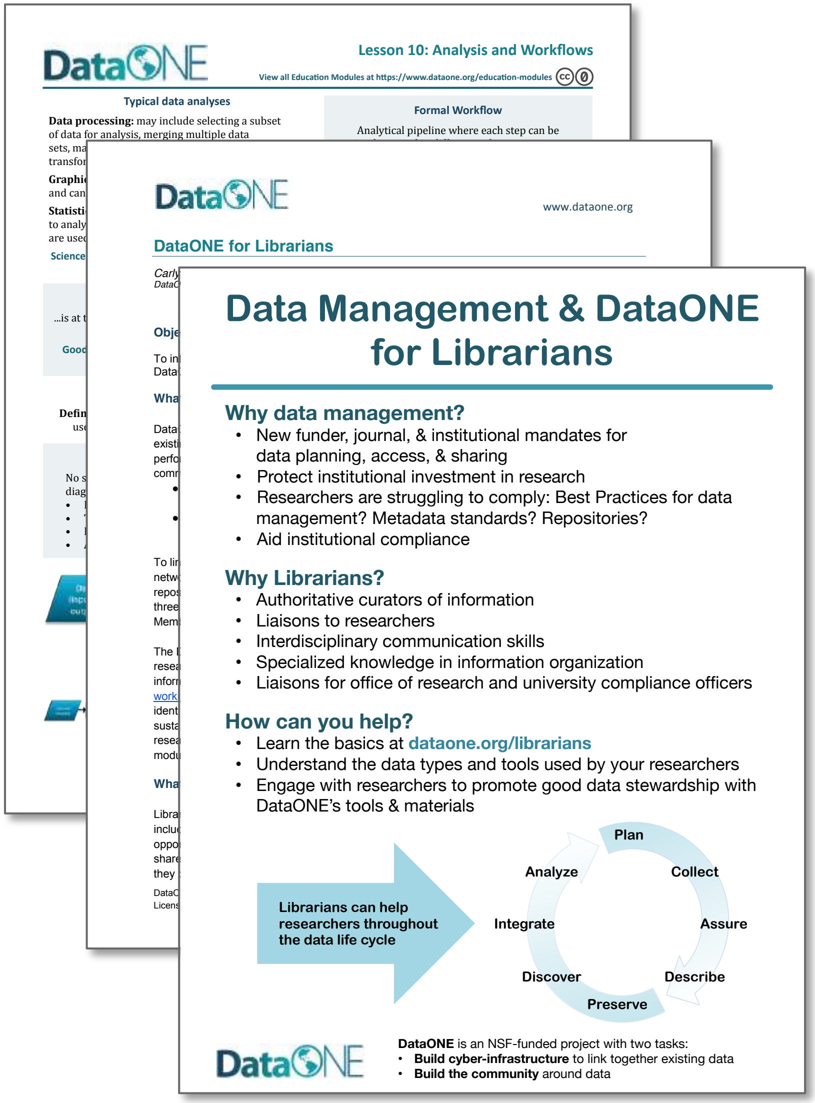
DataONE Webinar Series





www.dataone.org/webinars

Librarian Outreach Kit



DataONE

Lesson 10: Analysis and Workflows

View all Education Modules at https://www.dataone.org/education-modules

Typical data analyses

Data processing: may include selecting a subset of data for analysis, merging multiple data

Formal Workflow

Analytical pipeline where each step can be

DataONE

www.dataone.org

DataONE for Librarians

Data Management & DataONE for Librarians

Why data management?

- New funder, journal, & institutional mandates for data planning, access, & sharing
- Protect institutional investment in research
- Researchers are struggling to comply: Best Practices for data management? Metadata standards? Repositories?
- Aid institutional compliance

Why Librarians?

- Authoritative curators of information
- Liaisons to researchers
- Interdisciplinary communication skills
- Specialized knowledge in information organization
- Liaisons for office of research and university compliance officers

How can you help?

- Learn the basics at dataone.org/librarians
- Understand the data types and tools used by your researchers
- Engage with researchers to promote good data stewardship with DataONE's tools & materials

Librarians can help researchers throughout the data life cycle

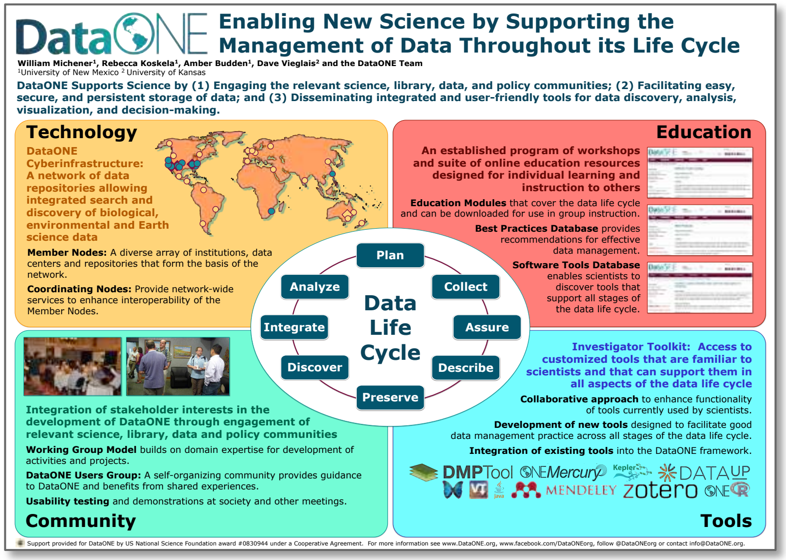
DataONE

DataONE is an NSF-funded project with two tasks:

- **Build cyber-infrastructure** to link together existing data
- **Build the community** around data

citeulike

DMPTool



Enabling New Science by Supporting the Management of Data Throughout its Life Cycle

William Michener[1], Rebecca Koskela[1], Amber Budden[1], Dave Vieglais[2] and the DataONE Team
[1]University of New Mexico [2]University of Kansas

DataONE Supports Science by (1) Engaging the relevant science, library, data, and policy communities; (2) Facilitating easy, secure, and persistent storage of data; and (3) Disseminating integrated and user-friendly tools for data discovery, analysis, visualization, and decision-making.

Technology

DataONE Cyberinfrastructure: A network of data repositories allowing integrated search and discovery of biological, environmental and Earth science data

Member Nodes: A diverse array of institutions, data centers and repositories that form the basis of the network.

Coordinating Nodes: Provide network-wide services to enhance interoperability of the Member Nodes.

Education

An established program of workshops and suite of online education resources designed for individual learning and instruction to others

Education Modules that cover the data life cycle and can be downloaded for use in group instruction.

Best Practices Database provides recommendations for effective data management.

Software Tools Database enables scientists to discover tools that support all stages of the data life cycle.

Integration of stakeholder interests in the development of DataONE through engagement of relevant science, library, data and policy communities

Working Group Model builds on domain expertise for development of activities and projects.

DataONE Users Group: A self-organizing community provides guidance to DataONE and benefits from shared experiences.

Usability testing and demonstrations at society and other meetings.

Community

Investigator Toolkit: Access to customized tools that are familiar to scientists and that can support them in all aspects of the data life cycle

Collaborative approach to enhance functionality of tools currently used by scientists.

Development of new tools designed to facilitate good data management practice across all stages of the data life cycle.

Integration of existing tools into the DataONE framework.

Tools

Support provided for DataONE by US National Science Foundation award #0830944 under a Cooperative Agreement. For more information see www.DataONE.org, www.facebook.com/DataONEorg, follow @DataONEorg or contact info@DataONE.org.

www.dataone.org/for-librarians

Other communication mechanisms

DataONE NEWS

Volume 4 Issue 1

©2015 DataONE 1312 Basehart SE University of New Mexico Albuquerque NM 87106

Making Open Science a Reality

The past two weeks have been marked by two events that directly or indirectly focus on making open science a reality. First, the Organization for Economic Co-Operation and Development (OECD) released a seminal report entitled "Making Open Science a Reality"[1]. The OECD consists of 34 countries including the United States, United Kingdom, Australia, Canada, France, Germany, Japan and New Zealand. The report is notable for at least three reasons. First, it eloquently summarizes the evidence for the benefits of open science (something I have touched upon in numerous prior newsletters). Second, it identifies the key actors in the open science space (e.g., researchers, government ministries, research funding agencies and private non-profit organizations and foundations, universities and public research institutes, libraries and repositories and data centers, and publishers and the business community) and explains their roles in promoting and enabling open science. Third, and most importantly, it details a number of key findings and policy messages to further open science that are highlighted below:

- *"Open science is a means and not an ends.* Open science strategies and policies are a means to support better quality science, increased collaboration, and engagement between research and society that can lead to higher social and economic impacts of public research.
- *Open science is more than open access to publications or data; it includes many aspects and stages of research processes.*
- *Policies to promote open data are less mature than those to promote open access to scientific publications.*
- *Open science policies should be principle-based but adapted to local realities.*
- *Better incentive mechanisms to promote data-sharing practices among researchers are needed.*
- *Data-related skill development is essential.*
- *Training of and awareness-raising among researchers is important for the development of an open science culture.*
- *Repositories and online platforms will not have impact if the information they contain is not of good quality.*
- *The long-term preservation costs of openly available research output need to be considered.*
- *Clear legal frameworks for the sharing of publications and reuse of data sets are needed at the national and international levels.*
- *Consultative approaches that involve all relevant actors for open science are a key component of successful open science strategies.*
- *International collaboration in the area of open science is necessary to address global challenges.*
- *Policy makers need to promote openness in science while at the same time preserving competition."*

It is reassuring to note that DataONE and its affiliated data repositories are making great strides in promoting open science by building infrastructure for discovery, use and long term preservation of data as well as solutions for improving data quality and metadata, and training the current and next generation of researchers.

The second notable recent event was the Research Data Alliance (RDA) Sixth Plenary Meeting and e-Infrastructures for Data Intensive Science Workshop that was held September 22-25, 2015 in Paris, France. DataONE was well represented at both meetings and played a key role in leading Working Groups and Interest Groups focused on metadata and brokering, and participating in other Groups focused on semantics, provenance and libraries. Over 500 individuals attended the meetings and participated in more than 50 Working and Interest Group activities. RDA is open to the broad community and if you have an interest in participating or tracking progress, I encourage you to visit https://rd-alliance.org.

In addition to the two events noted above, I also want to bring two other publications to your attention. First, the Journal of Librarianship and Scholarly Communication issued Volume 3, Issue 2[2] which includes numerous papers relevant to open science including institutional data management and data sharing practices, institutional data policies, and training programs and data workshops. Data scientists and librarians should find several articles worthy of their attention. Second, Carol Tenopir, Suzie Allard, Mike Frame and colleagues published a valuable article looking at changes in perceptions and practices related to data sharing that occurred from 2009/10 to 2013/14[3]. In particular, they noted many modest improvements in data sharing attitudes and practices that varied geographically and across age groups. They further emphasized the need for "organizations such as DataONE [that] will continue to assess, monitor, educate, and provide the infrastructure necessary to support such complex grand science challenges."

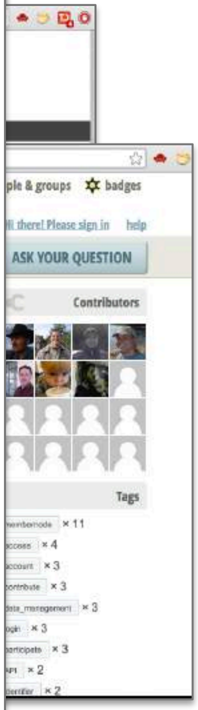
In the next newsletter issue, I will switch gears and look at how one can create an effective data management plan. In the meantime, please visit the RDA website and peruse the articles cited below for more information on data sharing and open science. ■

— *Bill Michener*
Principal Investigator

[1] OECD, 2015. Making Open Science a Reality. https://www.innovationpolicyplatform.org/content/open-science

[2] http://jlsc-pub.org/10/volume/3/issue/2/

[3] Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLoS ONE 10(8): e0134826. doi:10.1371/journal.pone.0134826



DataONE Users Group

- A self-organizing, independent group providing feedback to DataONE
- 310 members, 13 member Steering Committee, 2 Co-chairs
- Open participation and membership
- Annual summer meeting co-located with ESIP

www.dataone.org/dataone-users-group

Save the Date:
DataONE Users Group Meeting

Please save July 17-18, 2016 for the open DataONE Users Group meeting to be co-located with the Summer ESIP Federation Meeting at the Friday Center, Chapel Hill, North Carolina. The DataONE Users Group (DUG) meeting will be a 2-day event featuring plenary presentations, topical breakout sessions, and community-led discussions.

There is no registration fee to attend and participate in the DUG meeting.

Registration and hotel block will open in the spring, a few months before the meeting. Please visit https://www.dataone.org/dataone-users-group for updates and to join the DUG.

Meeting Theme and Objectives

The 2016 Meeting theme, "**Expanding Data Networks**," will focus on the new challenges and efforts in making data accessible, discoverable, and deliverable while promoting open data policies, standards, and compliance with funders' emerging data management requirements. A strong emphasis is on data synthesis and technological progress made in data network infrastructure.

The scientific program of the 2016 meeting will invite talks and posters on the following topics:

- Leveraging research data level metrics for large data repositories and data networks
- Integrating the needs and inputs of data users to advance and improve data discoverability
- Assessing the progress, impact, and success in promoting open data policies

DataONE encourages DataONE Member Nodes, data scientists, researchers, scientists, students and others to submit abstracts for posters and talks.

Abstract Submission for Posters and Talks

Please submit an abstract (250 words maximum) to dugchairs@dataone.org and indicate whether you prefer to present a talk or a poster. Talks will be approximately 10-20 minutes in duration, to be confirmed with development of the agenda. The poster session will be held the evening of Sunday July 17th during the reception event.

Submissions will be reviewed by the DataONE Users Group Steering Committee. Accepted abstracts will be published on the DataONE website.

Important dates

Abstract Submission Deadline**: April 15th 2016**
Author Notification: **May 15th 2016**

DUG Steering Committee: Felimon Gayanilo (co-chair), Plato Smith (co-chair), Steven Aulenbach, Amber Budden, Debora Drucker, Rebecca Koskela, Myrica McCune, Laura Moyers, Shannon Rauch, Robert Sandusky, Stephanie Simms, Heather Soyka

DataONE Users Group Meeting

July 17-18th 2016
Research Triangle, NC

Theme:

Expanding Data Networks

www.DataONE.org

@DataONEorg

facebook.com/DataONEorg

vimeo.com/DataONEorg

slideshare.net/DataONEorg

aebudden@dataone.unm.edu

vieglais@ku.edu