SUPPORTING DATA DISCOVERY AND ACCESS THROUGH SOCIAL AND TECHNICAL INFRASTRUCTURE

AMBER E BUDDEN

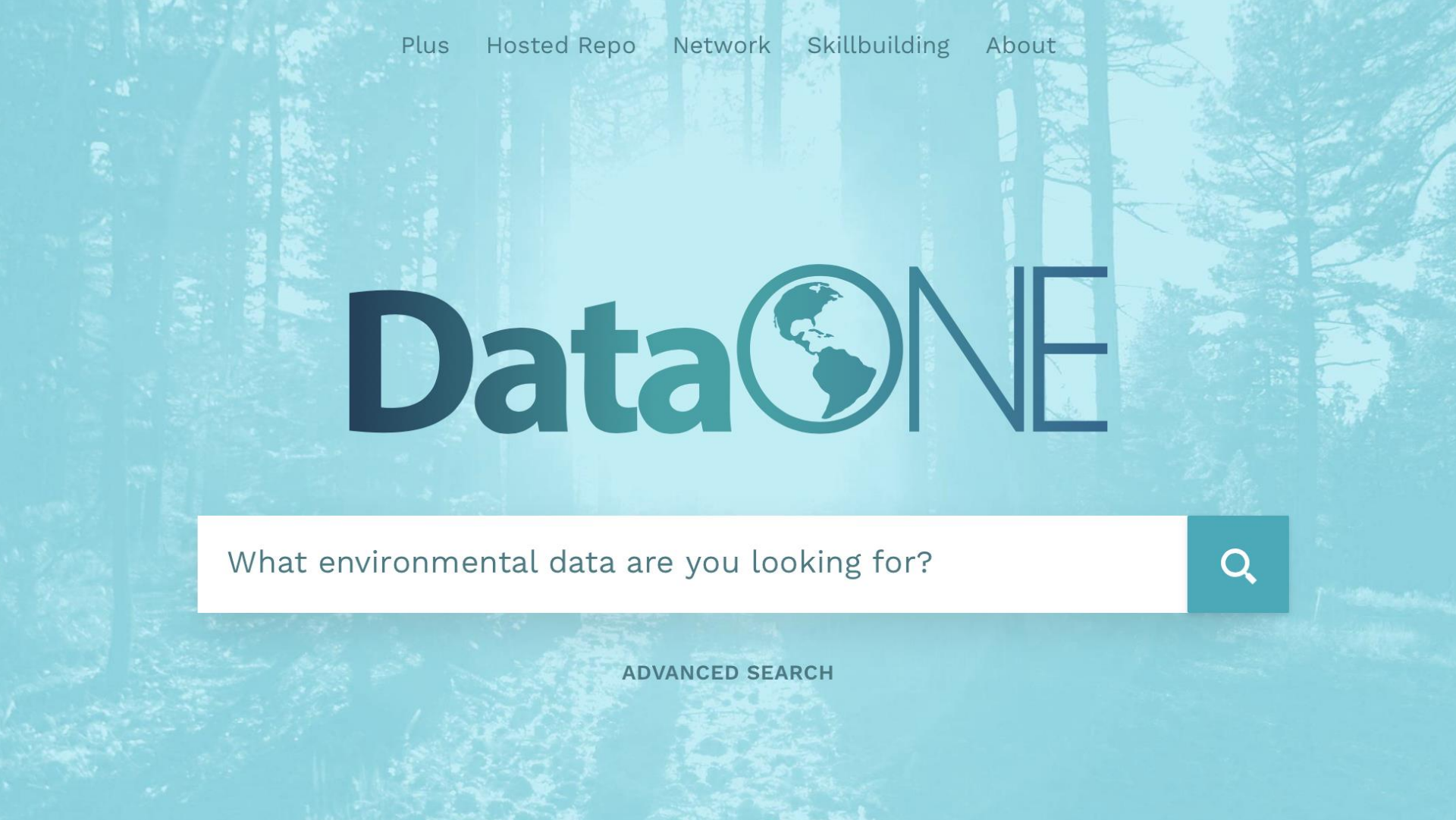
Director for Learning and Outreach, National Center for Ecological Analysis and Synthesis
Director of Community Engagement and Outreach, DataONE

@aebudden @DataONEorg @nceas

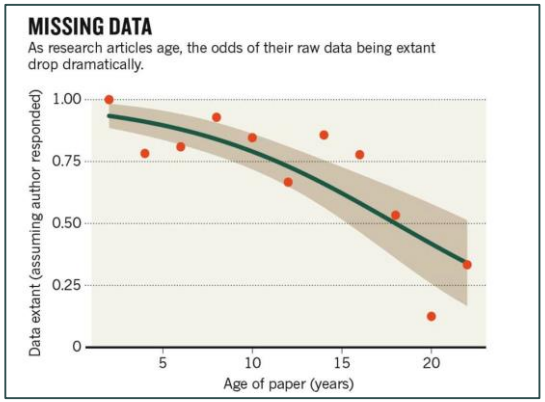
0000-0003-2885-3980

DataONE NCEAS National Center for Ecological Analysis and Synthesis

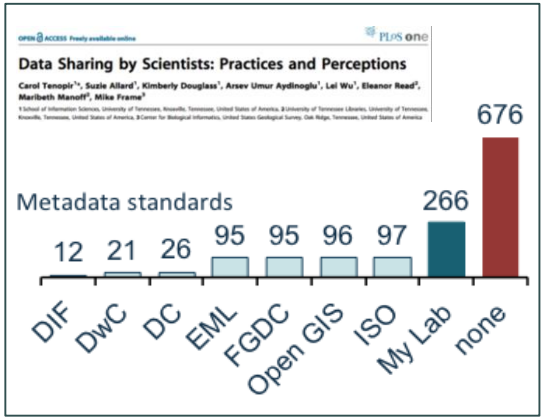
eResearch NZ 2020


Plus
Hosted Repo
Network
Skillbuilding
About
DataONE
What environmental data are you looking for?
ADVANCED SEARCH

Science and Informatics Challenges



Vines, T. H. *et al.* Curr. Biol. (2013)



- Require data integration across broad scales
- Traditional publication does not make data available by default
- Data within repositories can be siloed, limiting discovery
- Wide need for community training

NSF DataNet

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

PROGRAM SOLICITATION
NSF 07-601

National Science Foundation

Office of Cyberinfrastructure

Directorate for Computer & Information Science & Engineering

Preliminary Proposal Due Date(s) ***(required)*** (due by 5 p.m. proposer's local time)***:***

January 07, 2008

November 13, 2008

Full Proposal Target Date(s):

March 21, 2008

May 15, 2009

NSF DataNet

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

I. INTRODUCTION

Chapter 3 (Data, Data Analysis, and Visualization) of NSF's Cyberinfrastructure Vision for 21st Century Discovery (https://www.nsf.gov/pubs/2007/nsf0728/index.jsp) presents a vision in which "science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved." The goal of this solicitation is to catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable.

January 07, 2008

November 13, 2008

Full Proposal Target Date(s):

March 21, 2008

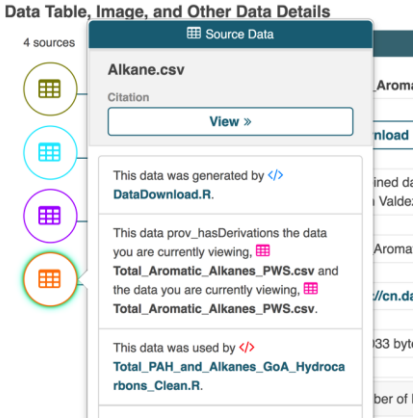
May 15, 2009

Three Primary Goals

Building an Empowered and Engaged Community

Enabling Reproducible Science through Tools and Services

Developing sustainable data discovery and interoperability solutions



Three Primary Goals

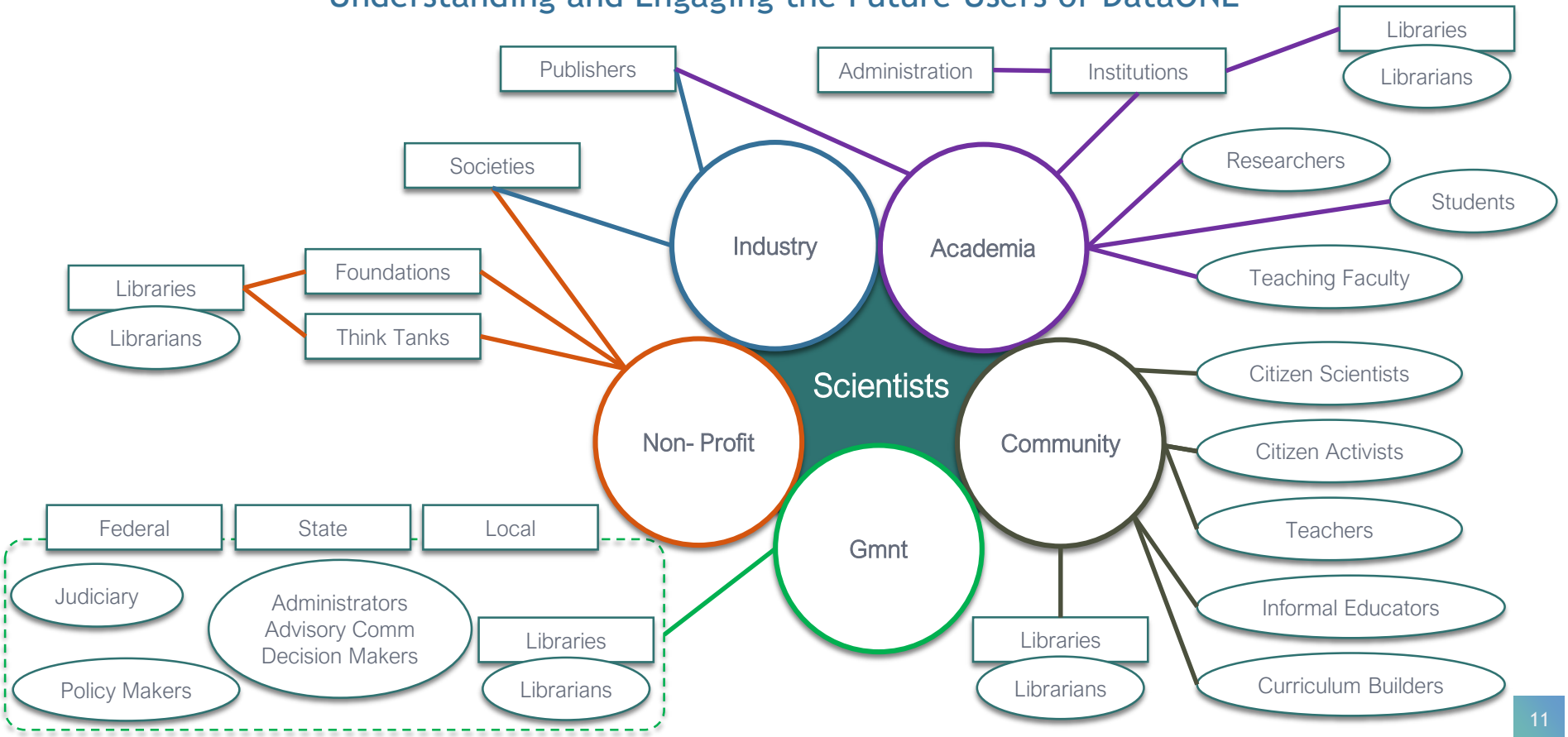
Building an Empowered and Engaged Community

Enabling Reproducible Science through Tools and Services

Developing sustainable data discovery and interoperability solutions

Science is Best Served by an Open and Inclusive Global Community

Understanding and Engaging the Future Users of DataONE



Federated Security

Preservation and Metadata

Provenance and Workflows

Semantics and Integration

Distributed Storage

Sociocultural Barriers to Data Sharing

Community Engagement and Education

Citizen Science / PPSR

Sustainability and Governance

Usability and Assessment

Exploration, Visualization and Analysis

Working Groups

Foundation for target research, development and education activities

Scientists Tell Us

1st Scientist Survey (2011)

PLOS ONE | TENTH ANNIVERSARY

Data Sharing by Scientists: Practices and Perceptions

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

Published: June 29, 2011 • https://doi.org/10.1371/journal.pone.0021101

750 Save	472 Citation
53,319 View	75 Share

2nd Scientist Survey (2015)

PLOS ONE | TENTH ANNIVERSARY

Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide

Carol Tenopir, Elizabeth D. Dalton, Suzie Allard, Mike Frame, Ivanka Pjesivac, Ben Birch, Danielle Pollock, Kristina Dorsett

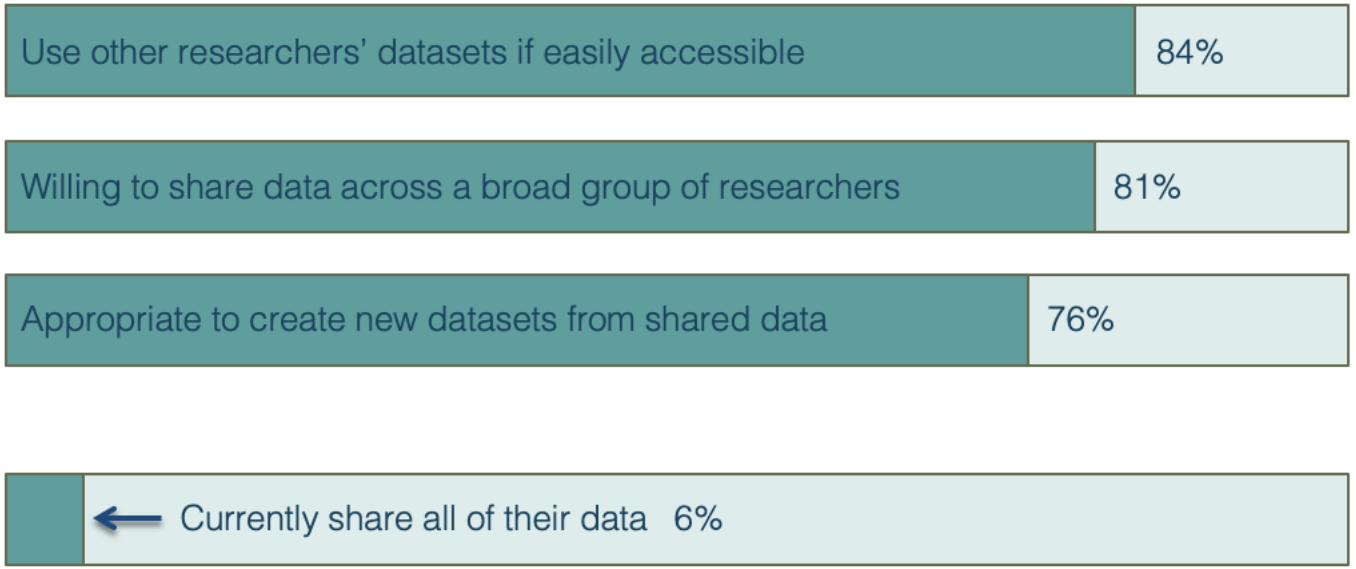
Published: August 26, 2015 • https://doi.org/10.1371/journal.pone.0134826

190 Save	75 Citation
16,386 View	176 Share

Scientists Want to Share Data

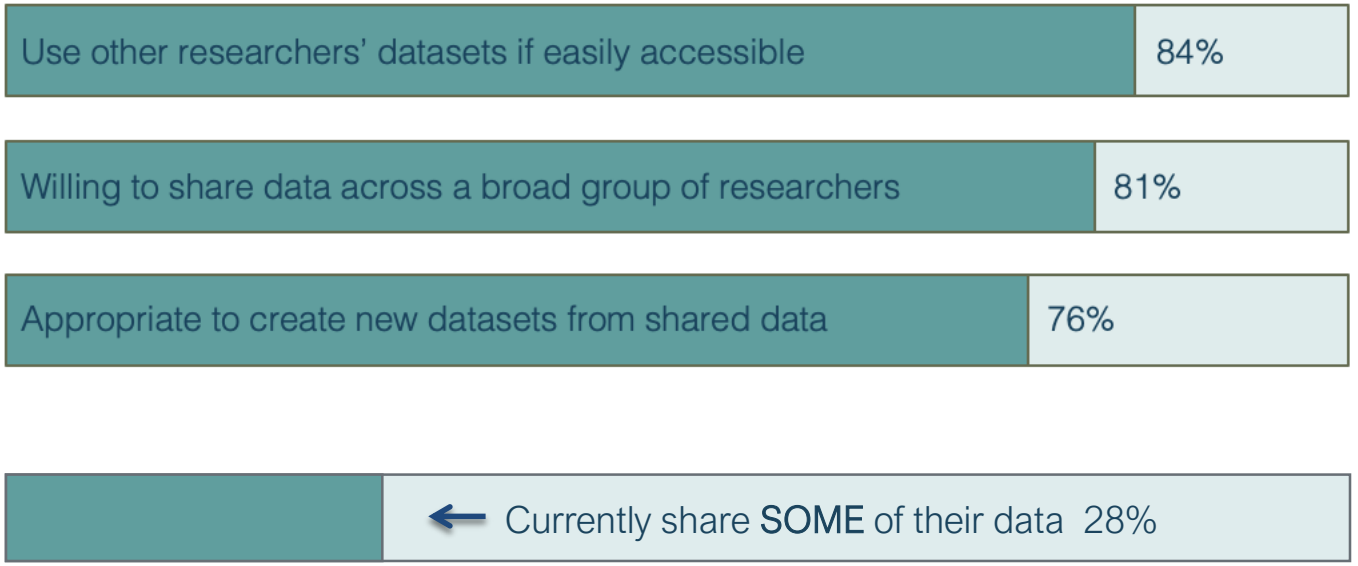
Tenopir et al 2011

Scientists Want to Share Data



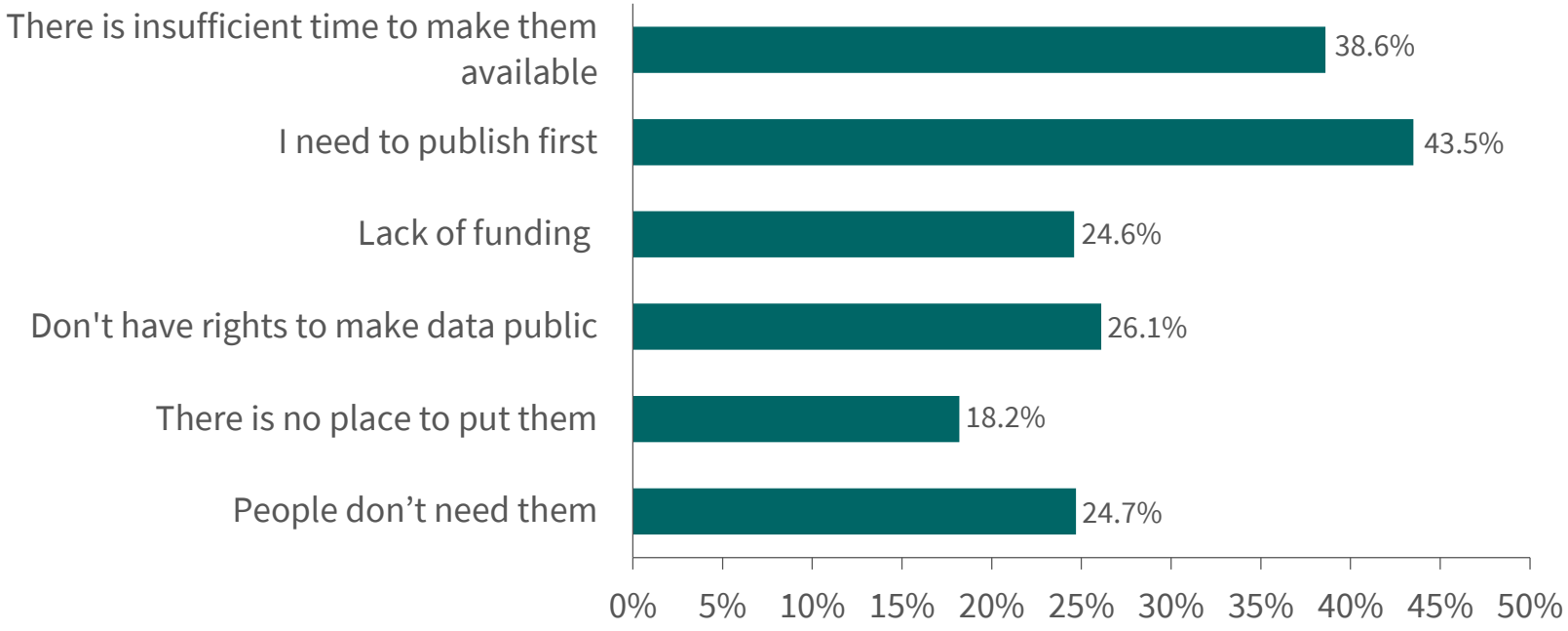
Tenopir et al 2011

Scientists Want to Share Data



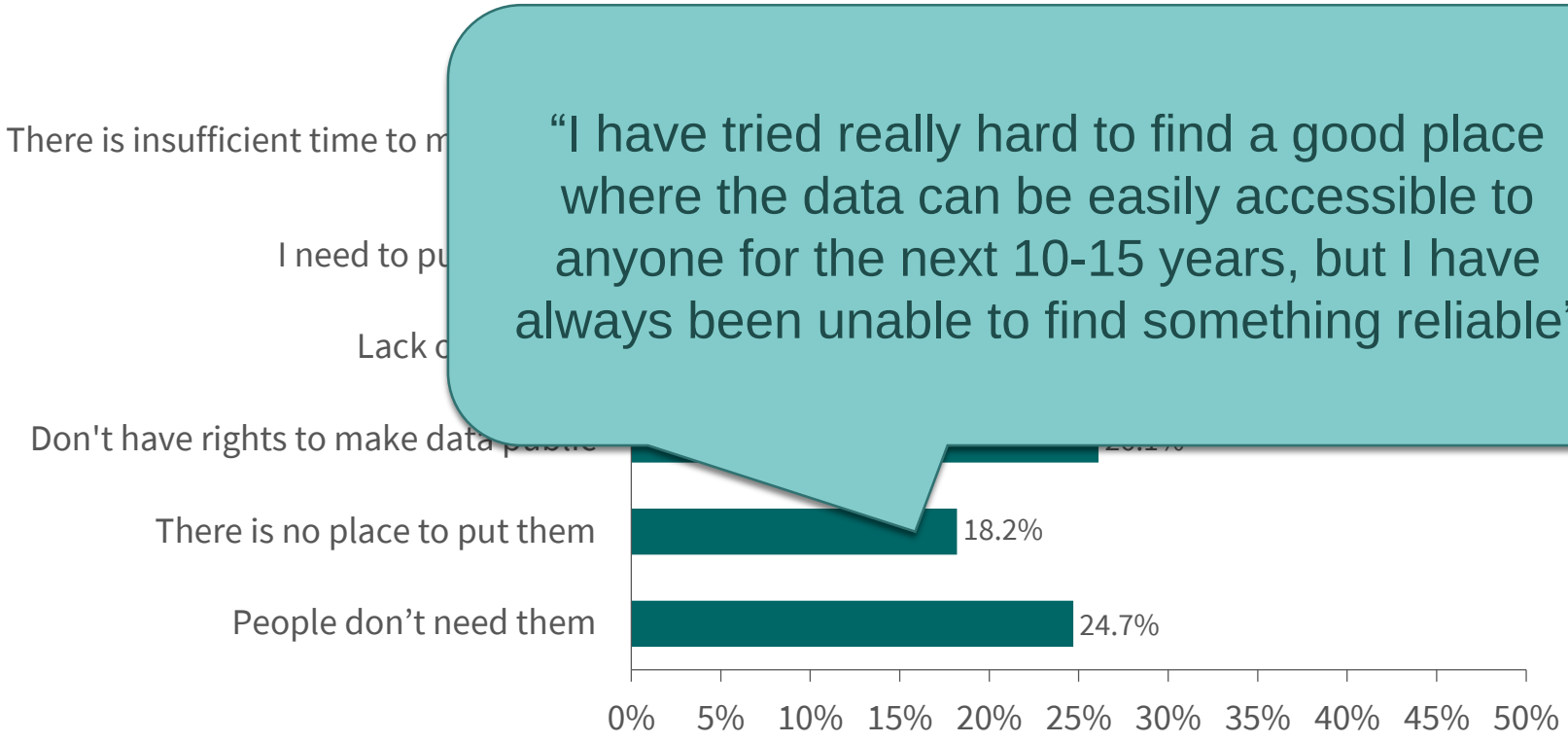
Tenopir et al 2011

Perceived Barriers to Data Sharing



Tenopir et al 2011

Perceived Barriers to Data Sharing



Tenopir et al 2011

Three Primary Goals

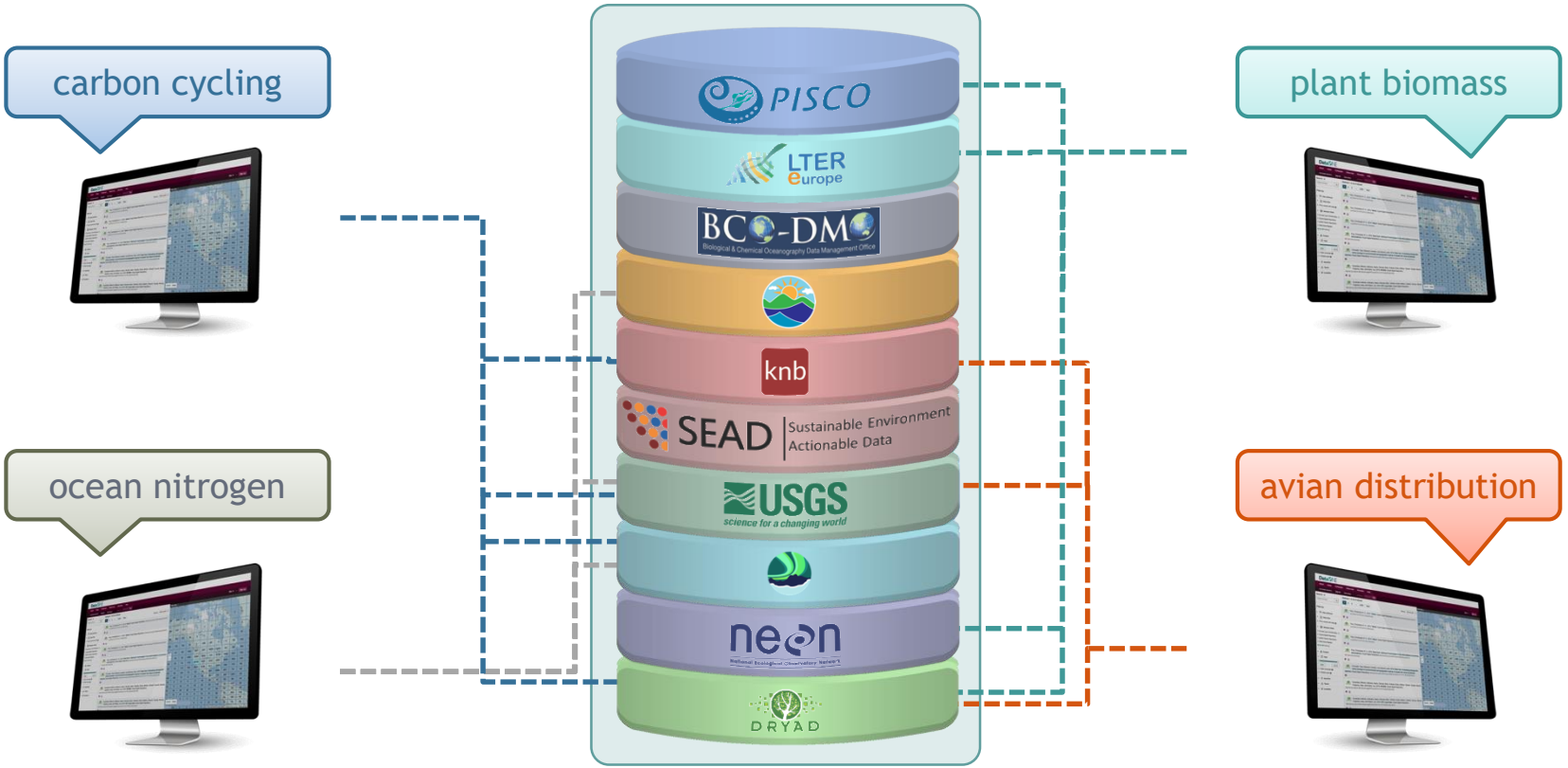
Building an Empowered and Engaged Community

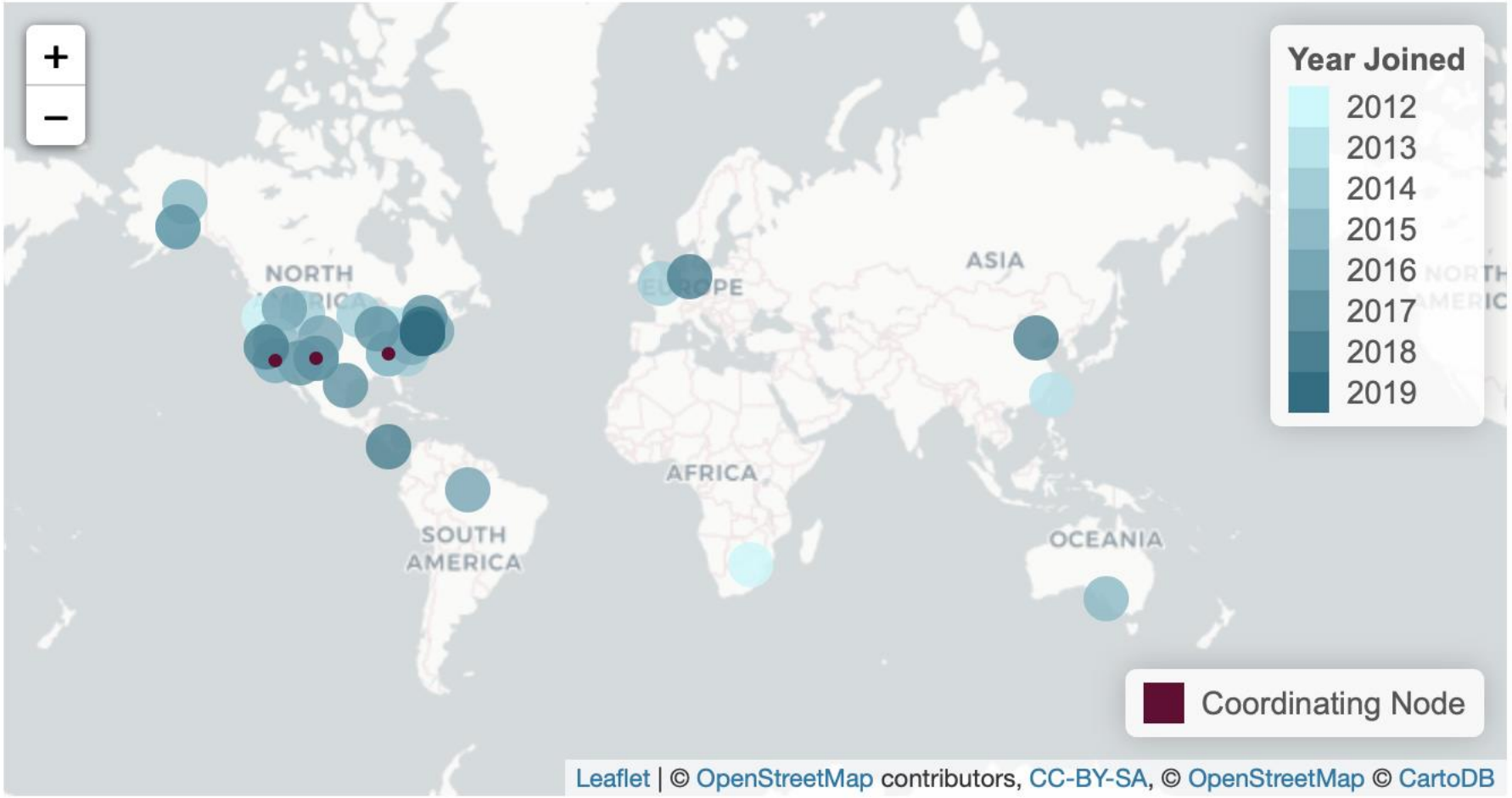
Enabling Reproducible Science through Tools and Services

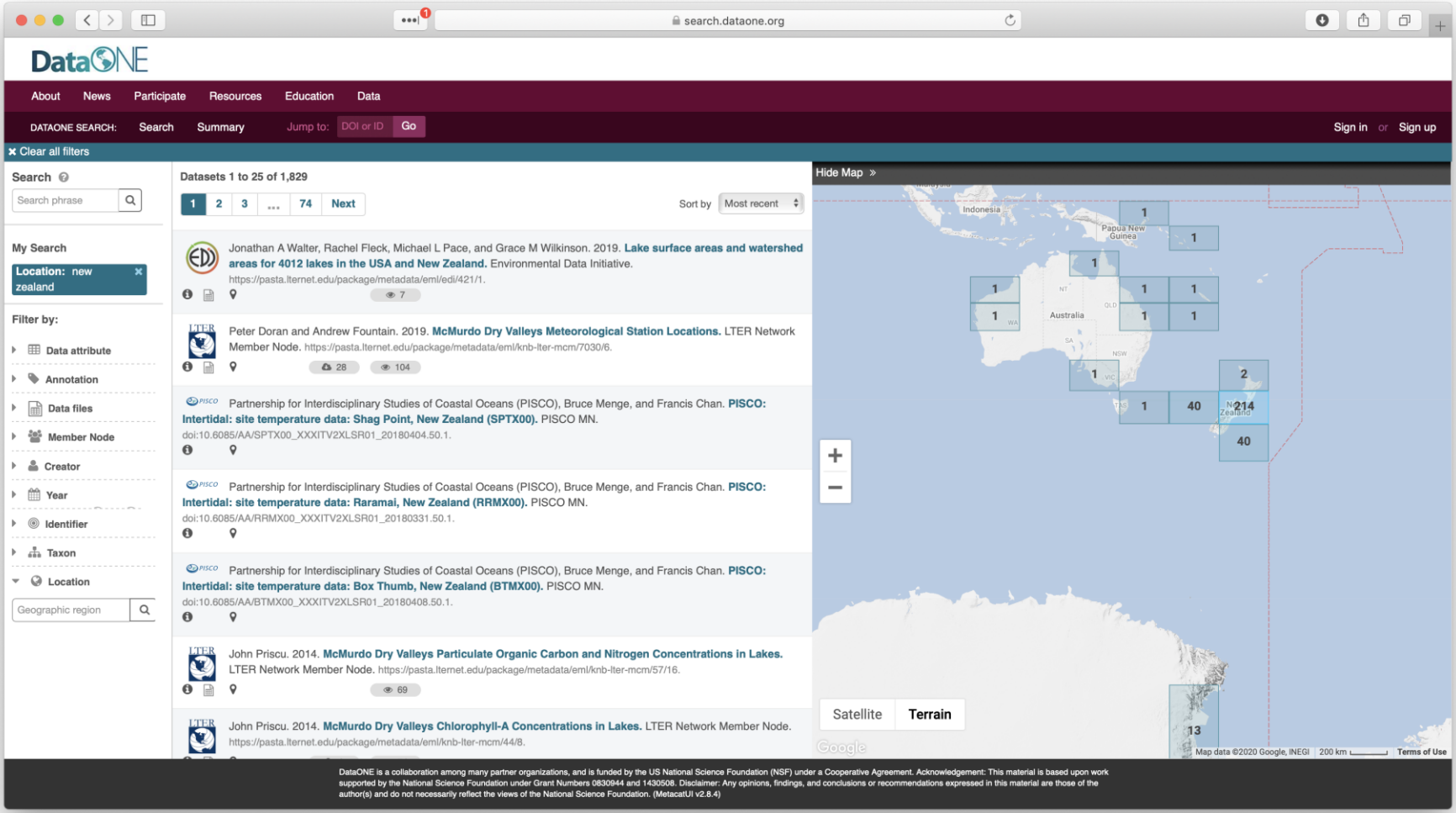
Developing sustainable data discovery and interoperability solutions

Federated Search

Data Discovery and Access from Multiple Repositories

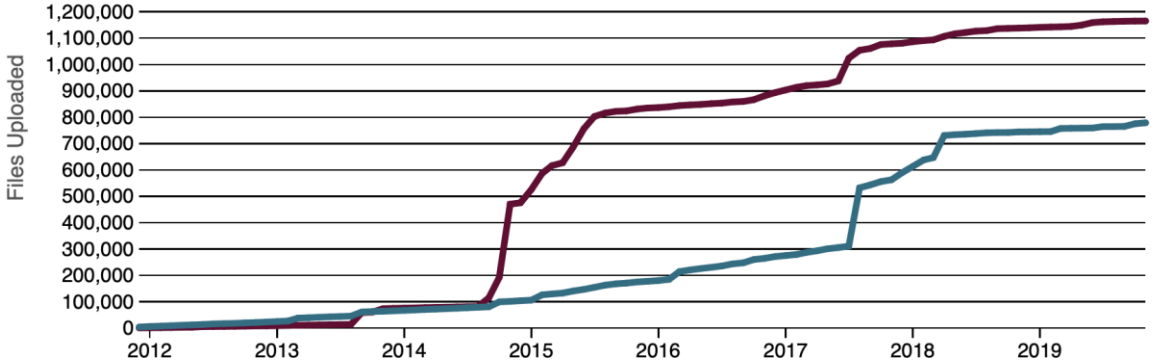


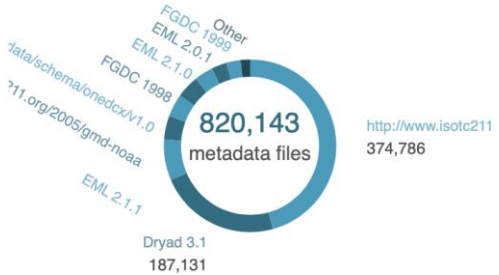

+
−
Year Joined
2012
2013
2014
2015
2016
2017
2018
2019
NORTH
AMERICA
EUROPE
ASIA
AFRICA
SOUTH
AMERICA
OCEANIA
Coordinating Node
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap © CartoDB

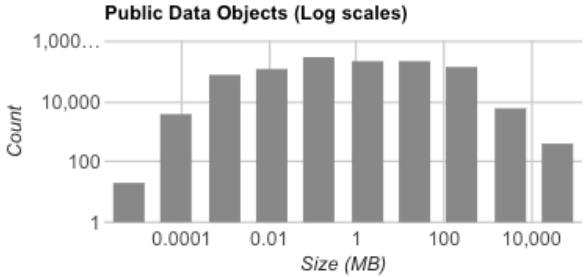

search.dataone.org
DataONE
About
News
Participate
Resources
Education
Data
DATAONE SEARCH:
Search
Summary
Jump to:
DOI or ID
Go
Sign in
or
Sign up
Clear all filters
Search
Search phrase
My Search
Location: new zealand
Filter by:
Data attribute
Annotation
Data files
Member Node
Creator
Year
Identifier
Taxon
Location
Geographic region
Datasets 1 to 25 of 1,829
1
2
3
...
74
Next
Sort by
Most recent
Jonathan A Walter, Rachel Fleck, Michael L Pace, and Grace M Wilkinson. 2019. Lake surface areas and watershed areas for 4012 lakes in the USA and New Zealand. Environmental Data Initiative. https://pasta.lternet.edu/package/metadata/eml/edi/421/1.
Peter Doran and Andrew Fountain. 2019. McMurdo Dry Valleys Meteorological Station Locations. LTER Network Member Node. https://pasta.lternet.edu/package/metadata/eml/knb-lter-mcm/7030/6.
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan. PISCO: Intertidal: site temperature data: Shag Point, New Zealand (SPTX00). PISCO MN. doi:10.6085/AA/SPTX00_XXXITV2XLSR01_20180404.50.1.
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan. PISCO: Intertidal: site temperature data: Raramai, New Zealand (RRMX00). PISCO MN. doi:10.6085/AA/RRMX00_XXXITV2XLSR01_20180331.50.1.
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan. PISCO: Intertidal: site temperature data: Box Thumb, New Zealand (BTMX00). PISCO MN. doi:10.6085/AA/BTMX00_XXXITV2XLSR01_20180408.50.1.
John Priscu. 2014. McMurdo Dry Valleys Particulate Organic Carbon and Nitrogen Concentrations in Lakes. LTER Network Member Node. https://pasta.lternet.edu/package/metadata/eml/knb-lter-mcm/57/16.
John Priscu. 2014. McMurdo Dry Valleys Chlorophyll-A Concentrations in Lakes. LTER Network Member Node. https://pasta.lternet.edu/package/metadata/eml/knb-lter-mcm/44/8.
Hide Map
Indonesia
Papua New Guinea
Australia
Satellite
Terrain
Map data ©2020 Google, INEGI
200 km
Terms of Use
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. (MetacatUI v2.8.4)

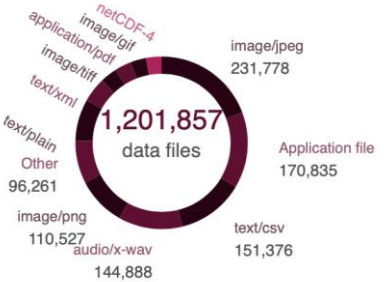
Network Growth

Steadily growing, stable infrastructure

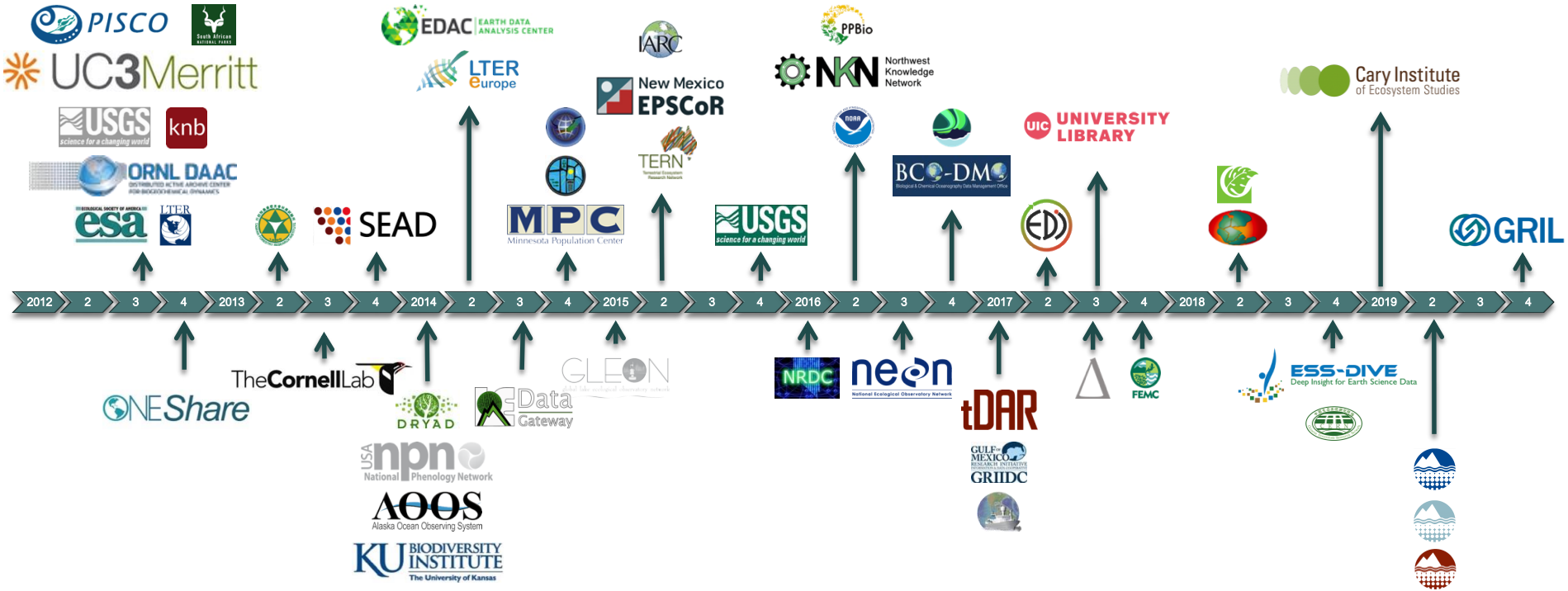








Building the Federation

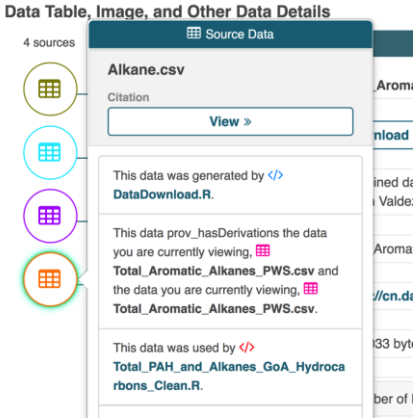


Three Primary Goals

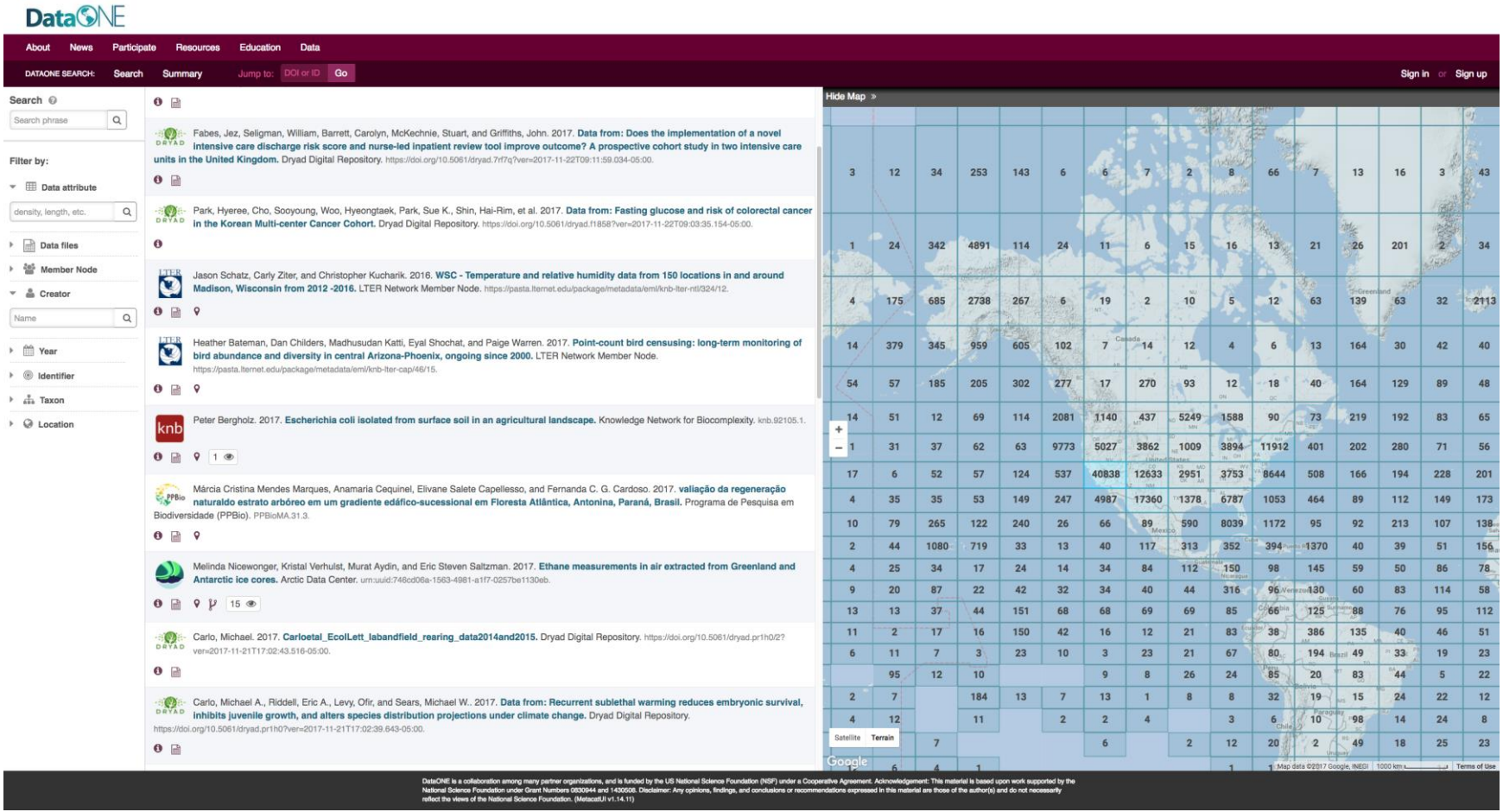
Building an Empowered and Engaged Community

Enabling Reproducible Science through Tools and Services

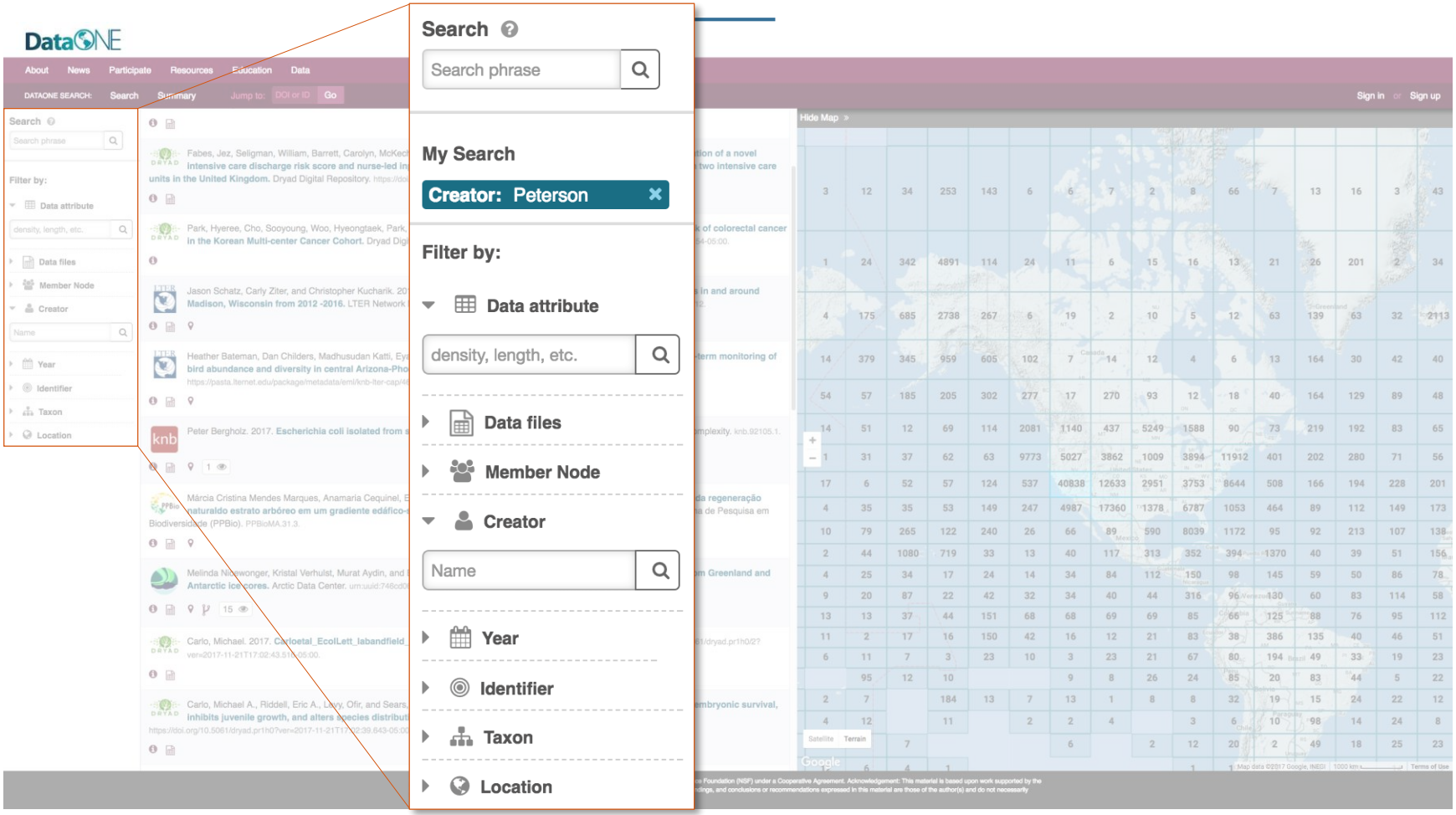
Developing sustainable data discovery and interoperability solutions



Federated Search Across Repositories

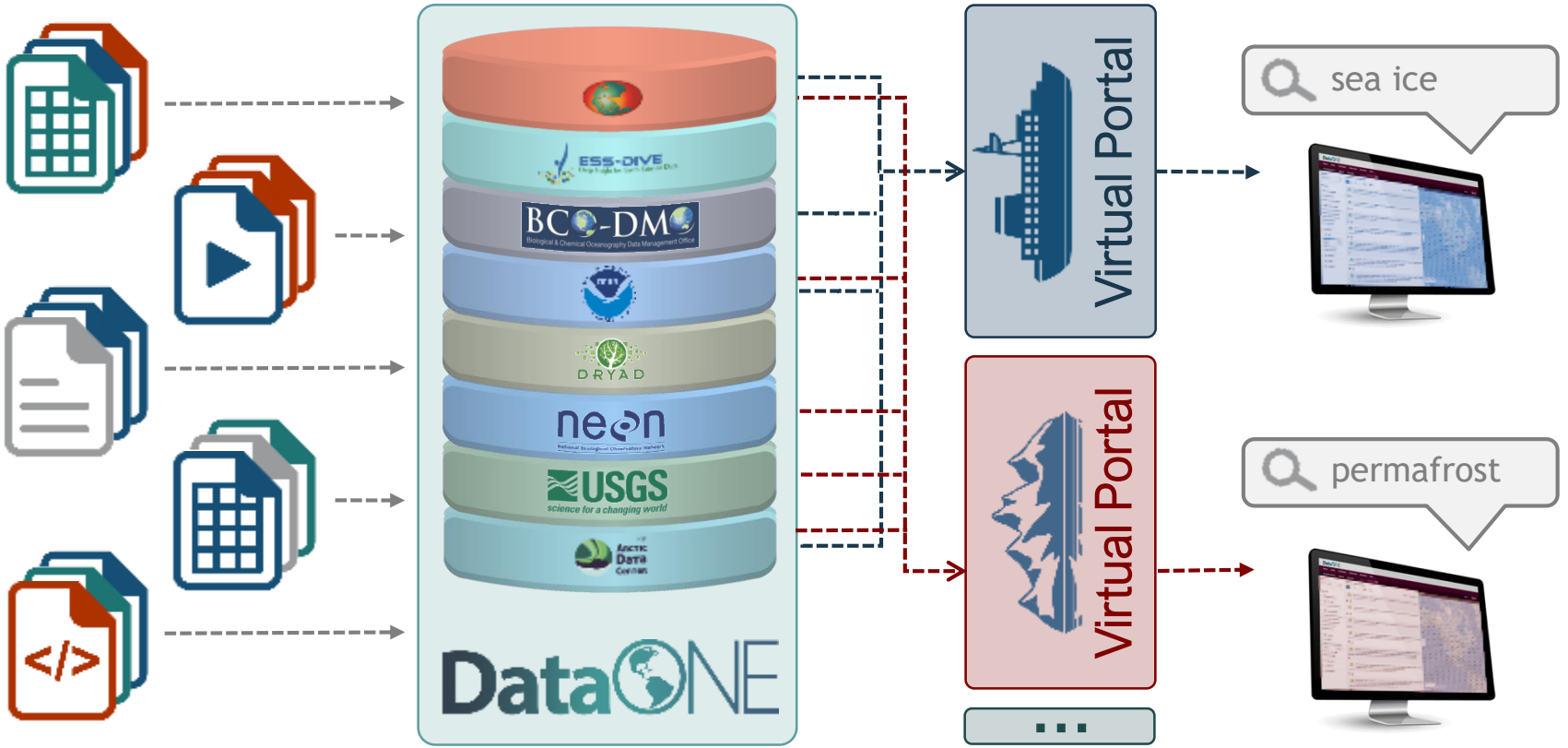


Federated Search Across Repositories



Data Portals

Data Aggregated from Repositories Across DataONE



Portal Services

Customizable Services by Organization, Theme and Region

Theme 

Project/ Organization 

Region 

Repository 

Custom branding and storytelling

Data Level Metrics

Aggregated views, downloads and citations

Metadata Quality

Customized checks of metadata quality

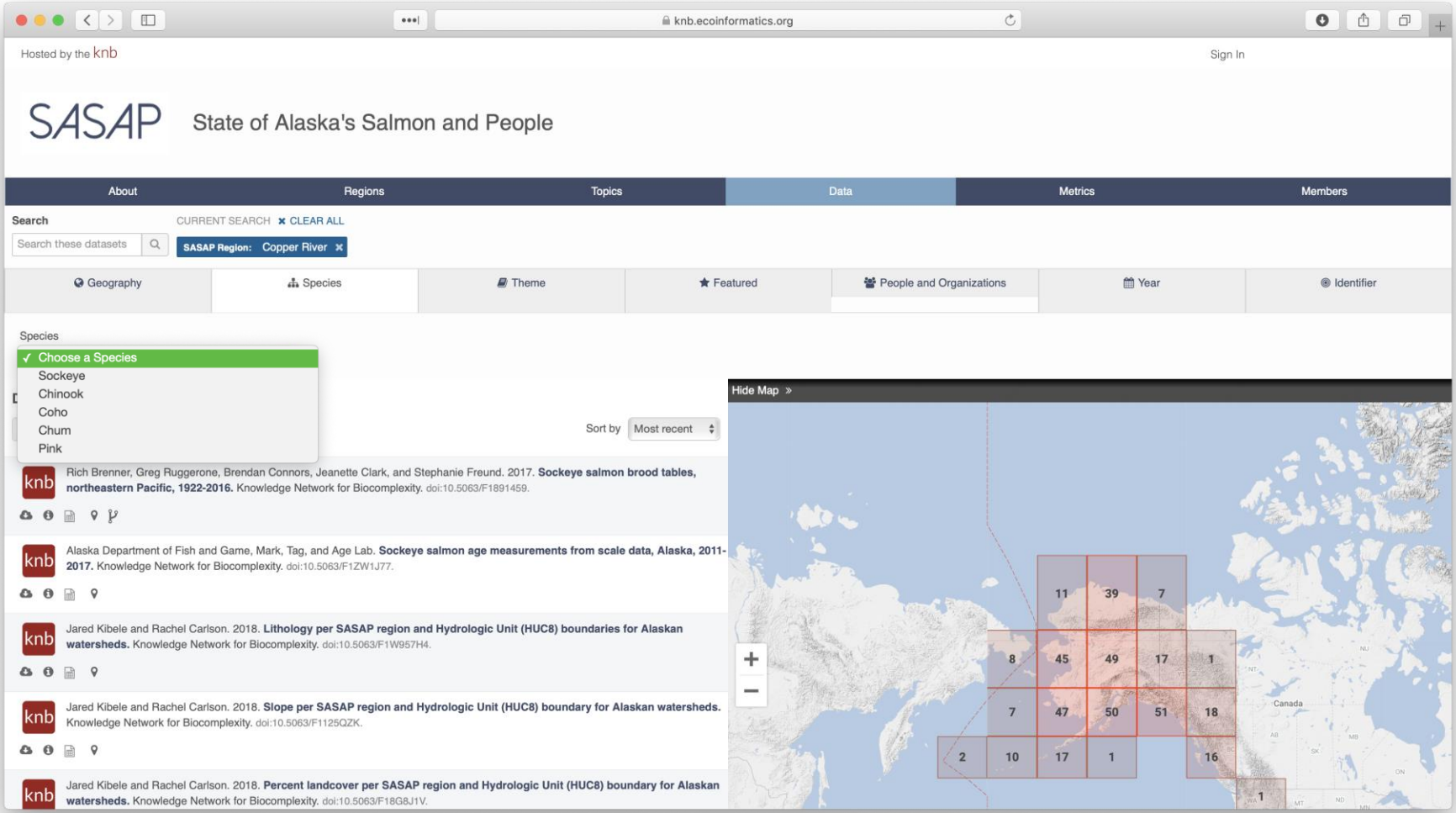
Metadata Creation

Metadata authoring and data upload tools

Customized Search

Community specific search features for enhanced discovery

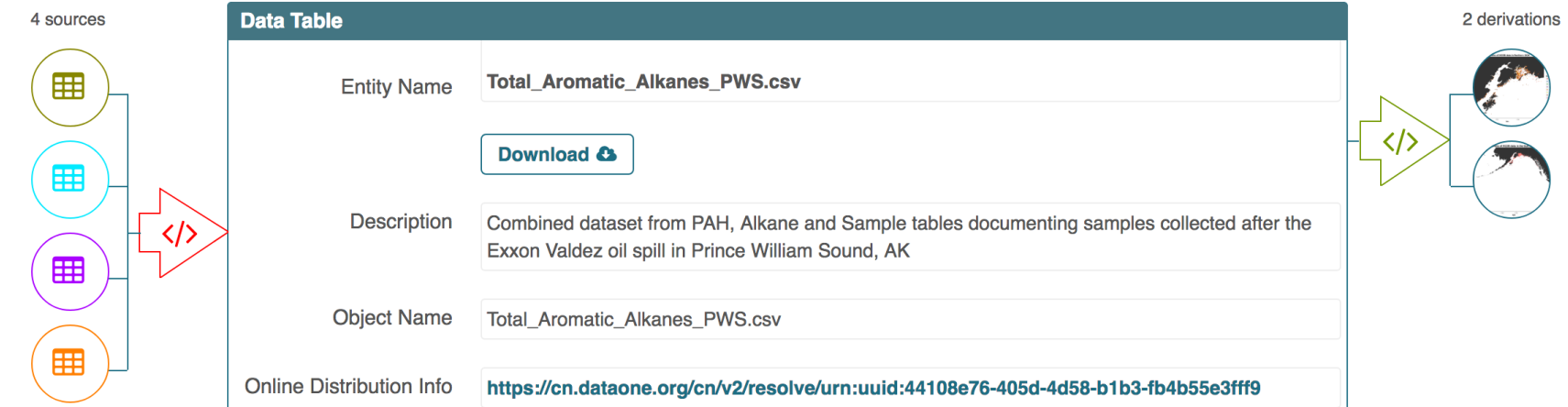
• • •


knb.ecoinformatics.org
Hosted by the knb
Sign In
SASAP
State of Alaska's Salmon and People
About
Regions
Topics
Data
Metrics
Members
Search
CURRENT SEARCH
CLEAR ALL
Search these datasets
SASAP Region: Copper River
Geography
Species
Theme
Featured
People and Organizations
Year
Identifier
Species
Choose a Species
Sockeye
Chinook
Coho
Chum
Pink
Sort by
Most recent
Hide Map
Rich Brenner, Greg Ruggerone, Brendan Connors, Jeanette Clark, and Stephanie Freund. 2017. Sockeye salmon brood tables, northeastern Pacific, 1922-2016. Knowledge Network for Biocomplexity. doi:10.5063/F1891459.
Alaska Department of Fish and Game, Mark, Tag, and Age Lab. Sockeye salmon age measurements from scale data, Alaska, 2011-2017. Knowledge Network for Biocomplexity. doi:10.5063/F1ZW1J77.
Jared Kibele and Rachel Carlson. 2018. Lithology per SASAP region and Hydrologic Unit (HUC8) boundaries for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F1W957H4.
Jared Kibele and Rachel Carlson. 2018. Slope per SASAP region and Hydrologic Unit (HUC8) boundary for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F1125QZK.
Jared Kibele and Rachel Carlson. 2018. Percent landcover per SASAP region and Hydrologic Unit (HUC8) boundary for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F18G8J1V.
11
39
7
8
45
49
17
1
7
47
50
51
18
2
10
17
1
16
1
Canada

Provenance Display

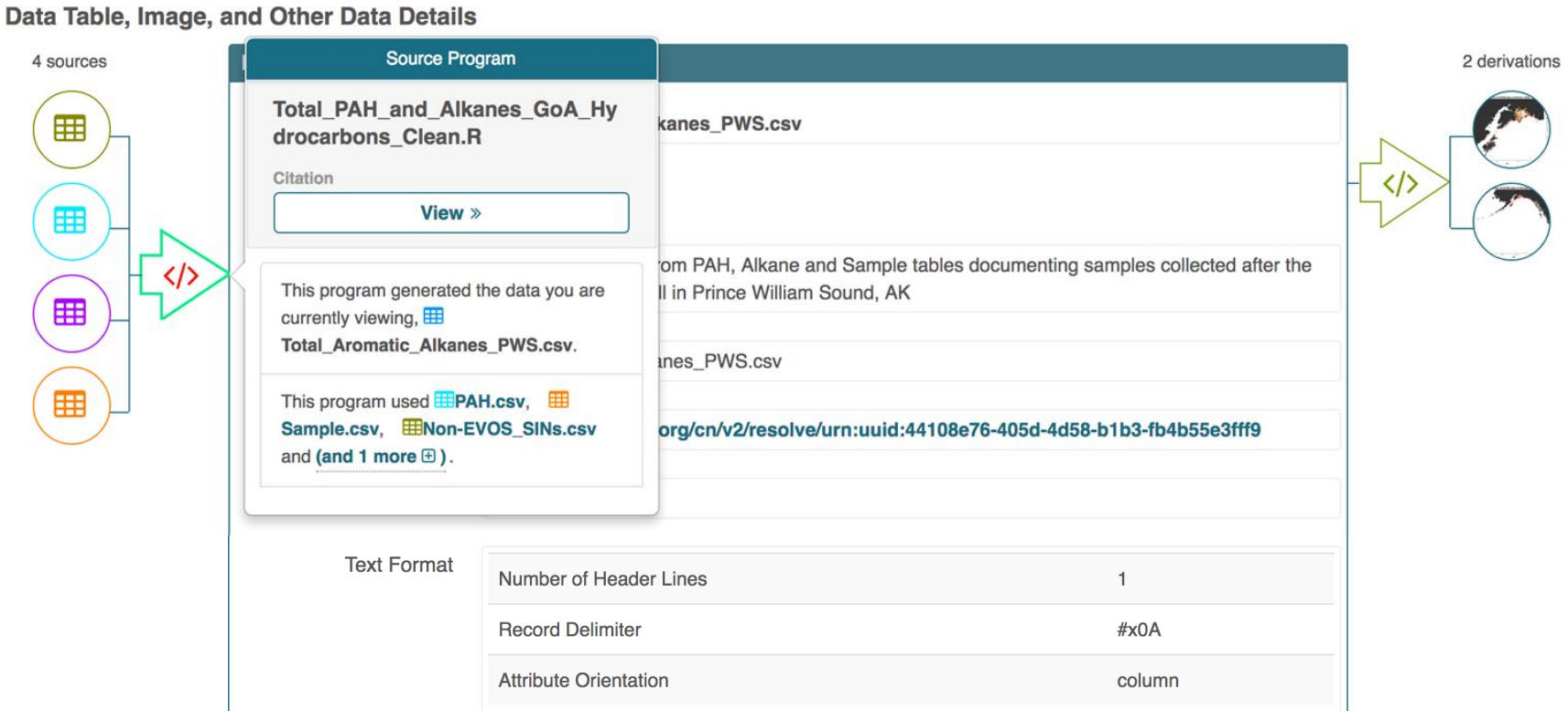
Provenance Tracking and Management Services and Tools

Data Table, Image, and Other Data Details



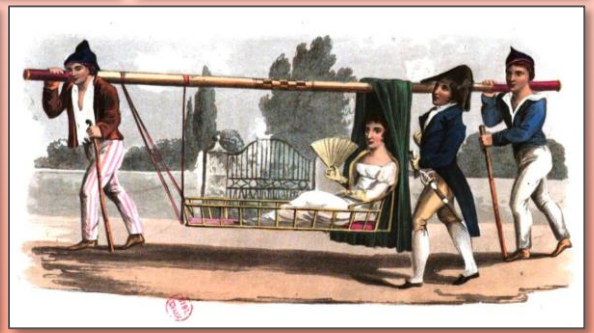
Provenance Display

Provenance Tracking and Management Services and Tools

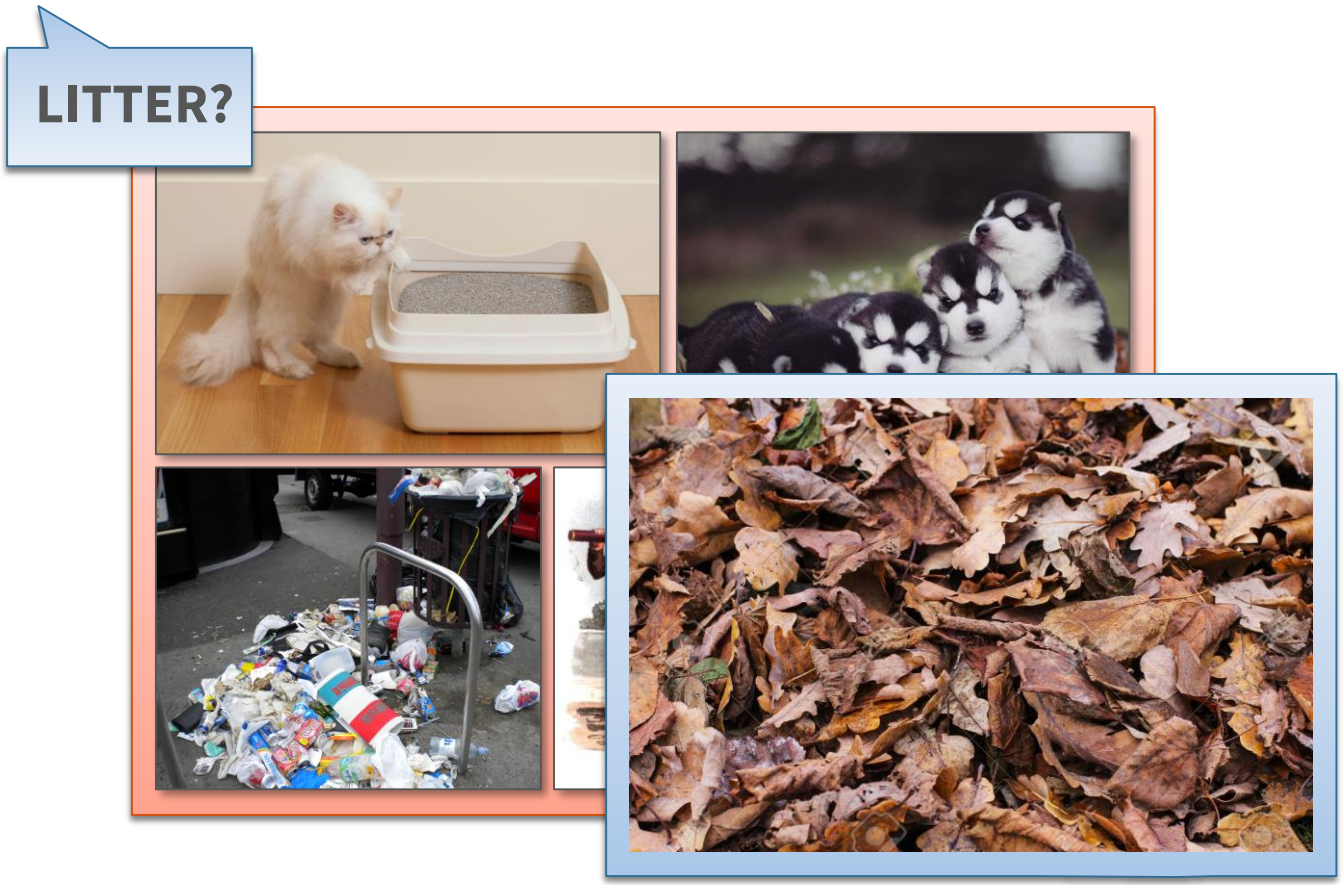


Semantic Search

LITTER?

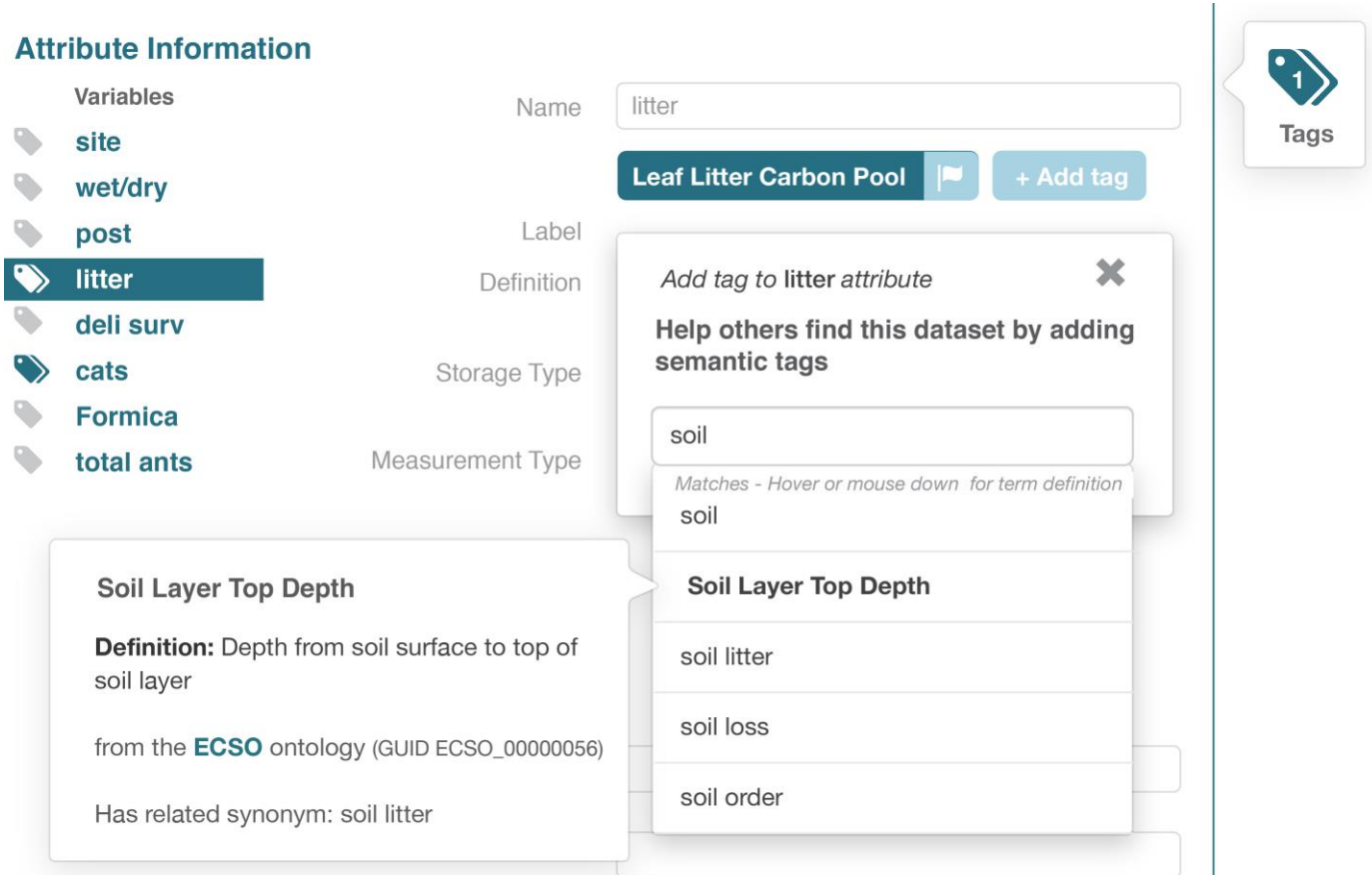


Semantic Search



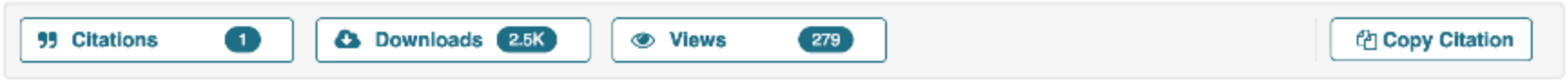
Semantic Search

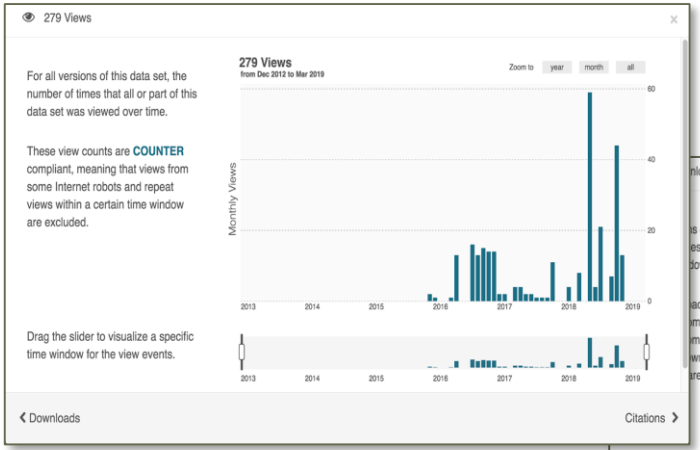
Semantic Measurement Search Capabilities

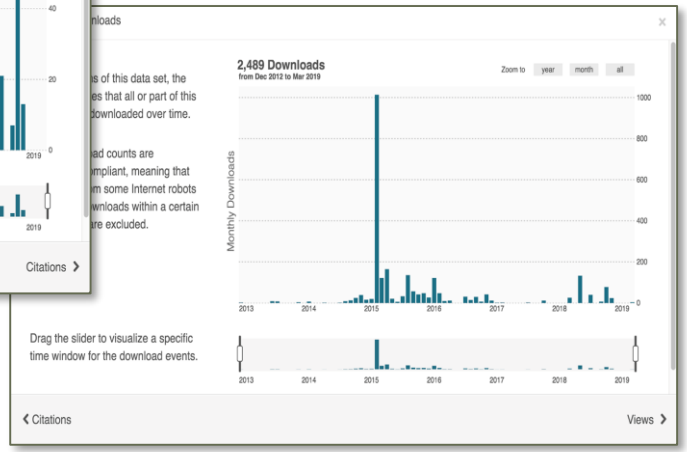


Data Usage and Citation

Hajo Eicken, Rolf Gradinger, Thomas Heinrichs, Mark Johnson, Amy Lovecraft, et al. 2011. Mooring Temperature/Conductivity & Temperature/Pressure data. Arctic Data Center. **doi:10.18739/A2CZ3244X**.



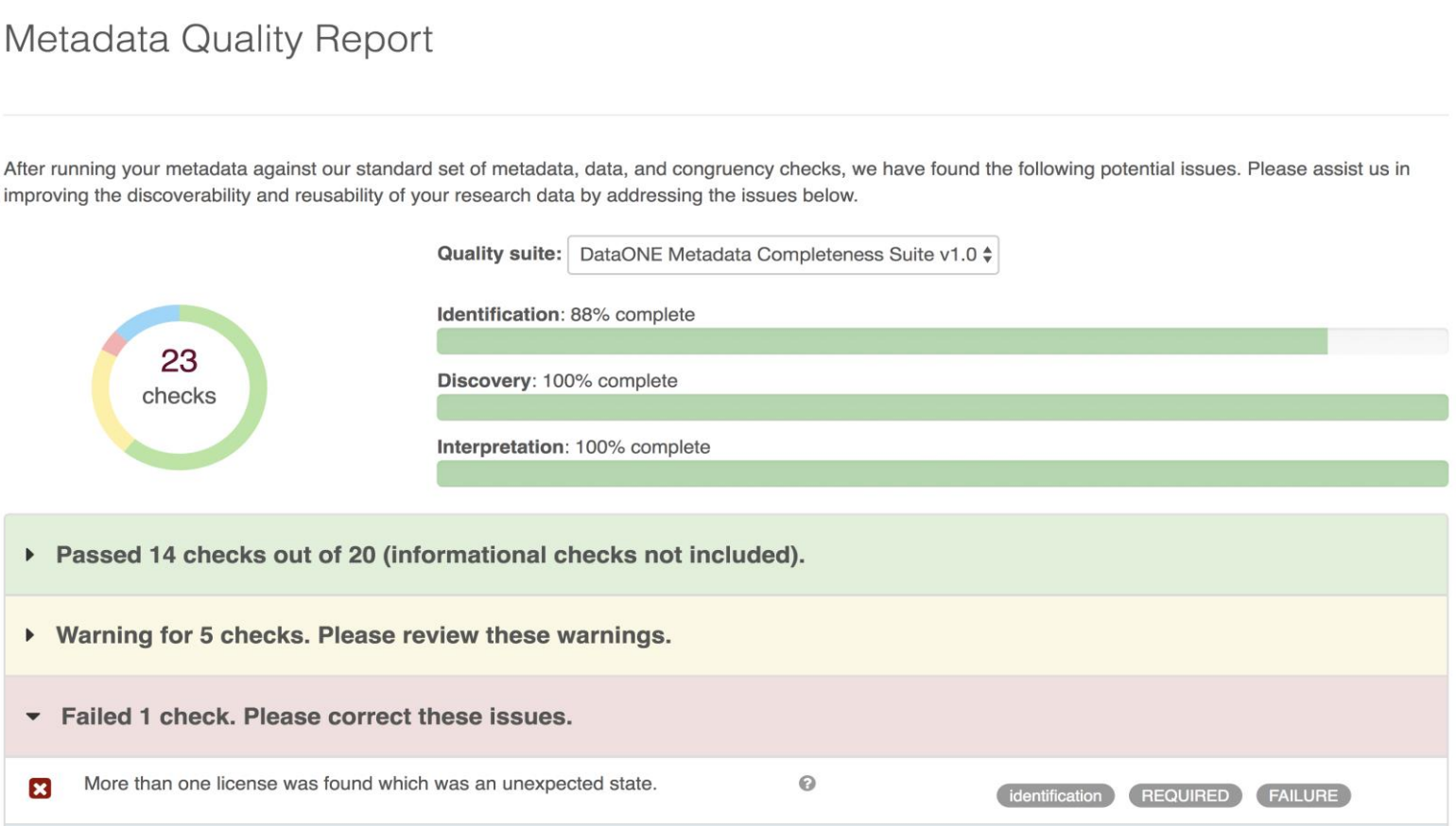




MAKE DATA COUNT

COUNTER
CONSISTENT CREDIBLE COMPARABLE

Metadata Quality Improvement

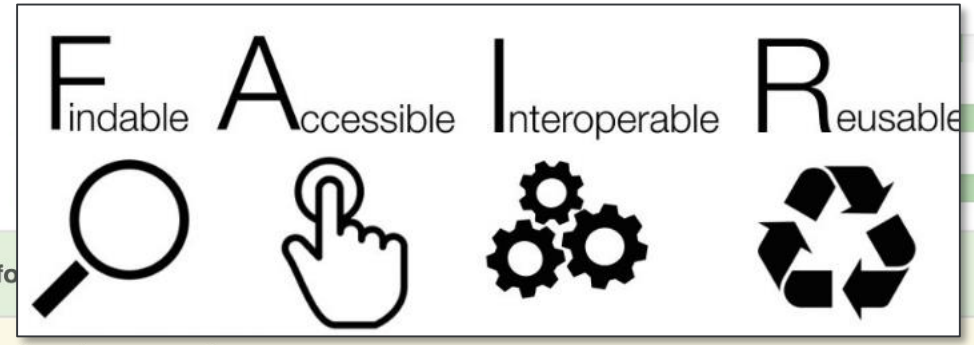


Metadata Quality Improvement

Metadata Quality Report

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.

Quality suite: DataONE Metadata Completeness Suite v1.0



▸ **Passed 14 checks out of 20 (info**

▸ **Warning for 5 checks. Please review these warnings.**

▾ **Failed 1 check. Please correct these issues.**

More than one license was found which was an unexpected state. identification REQUIRED FAILURE

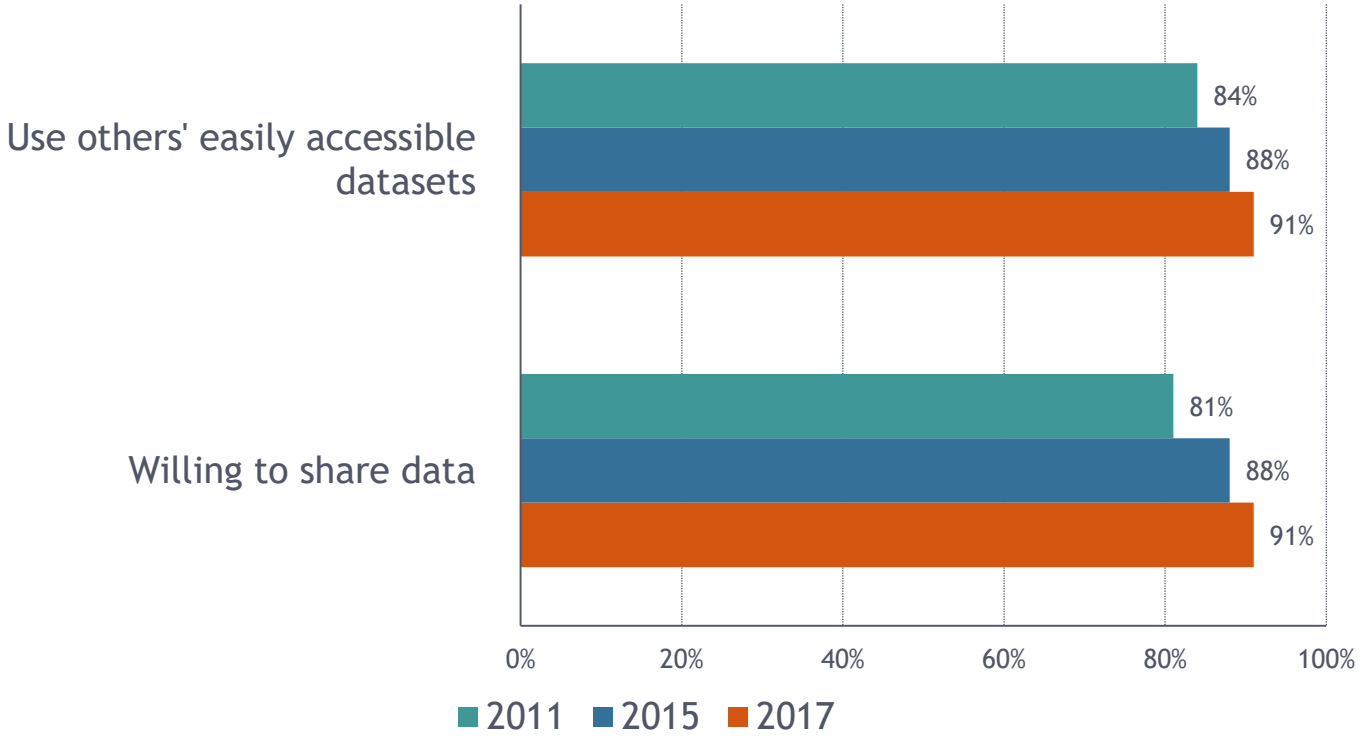
Three Primary Goals

Building an Empowered and Engaged Community

Enabling Reproducible Science through Tools and Services

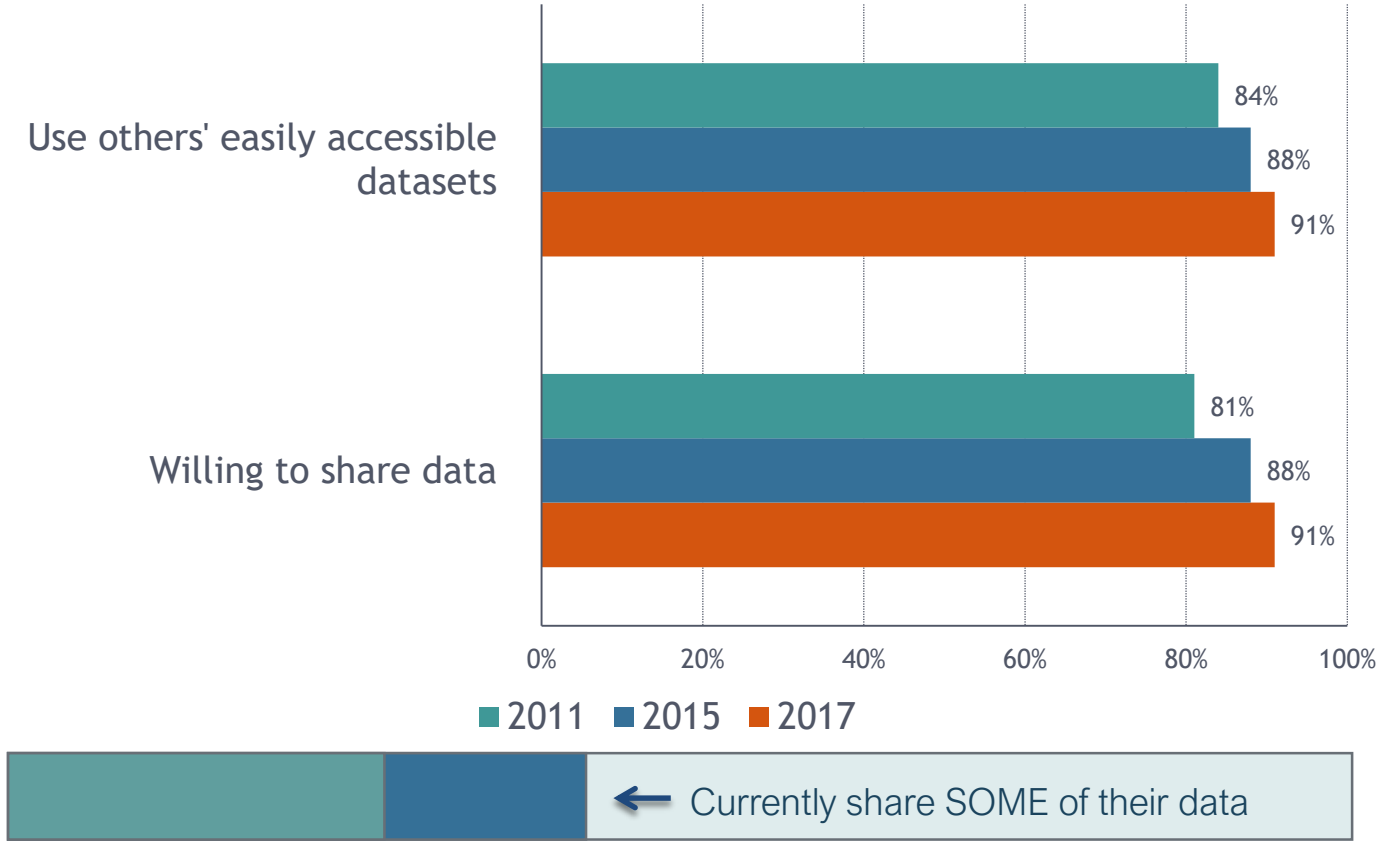
Developing sustainable data discovery and interoperability solutions

Culture of Data Sharing is Improving



Tenopir et al 2011, Tenopir et al 2015, Tenopir et al in press

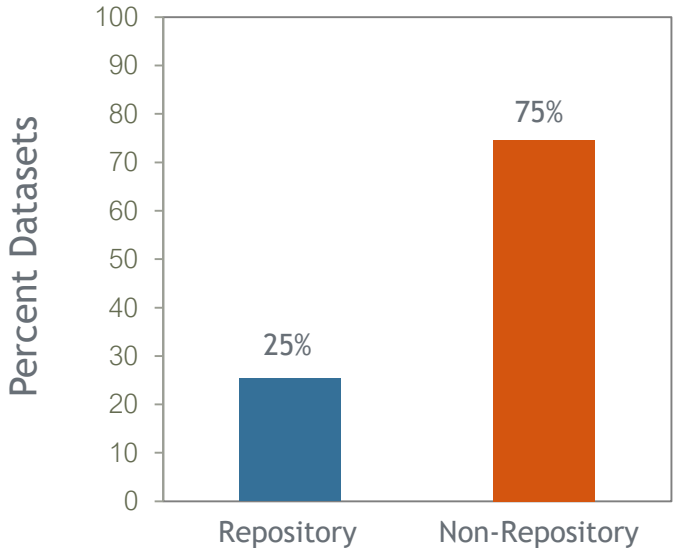
Culture of Data Sharing is Improving



Tenopir et al 2011, Tenopir et al 2015, Tenopir et al in press

Culture of Data Sharing is Improving

But there is still opportunity for growth

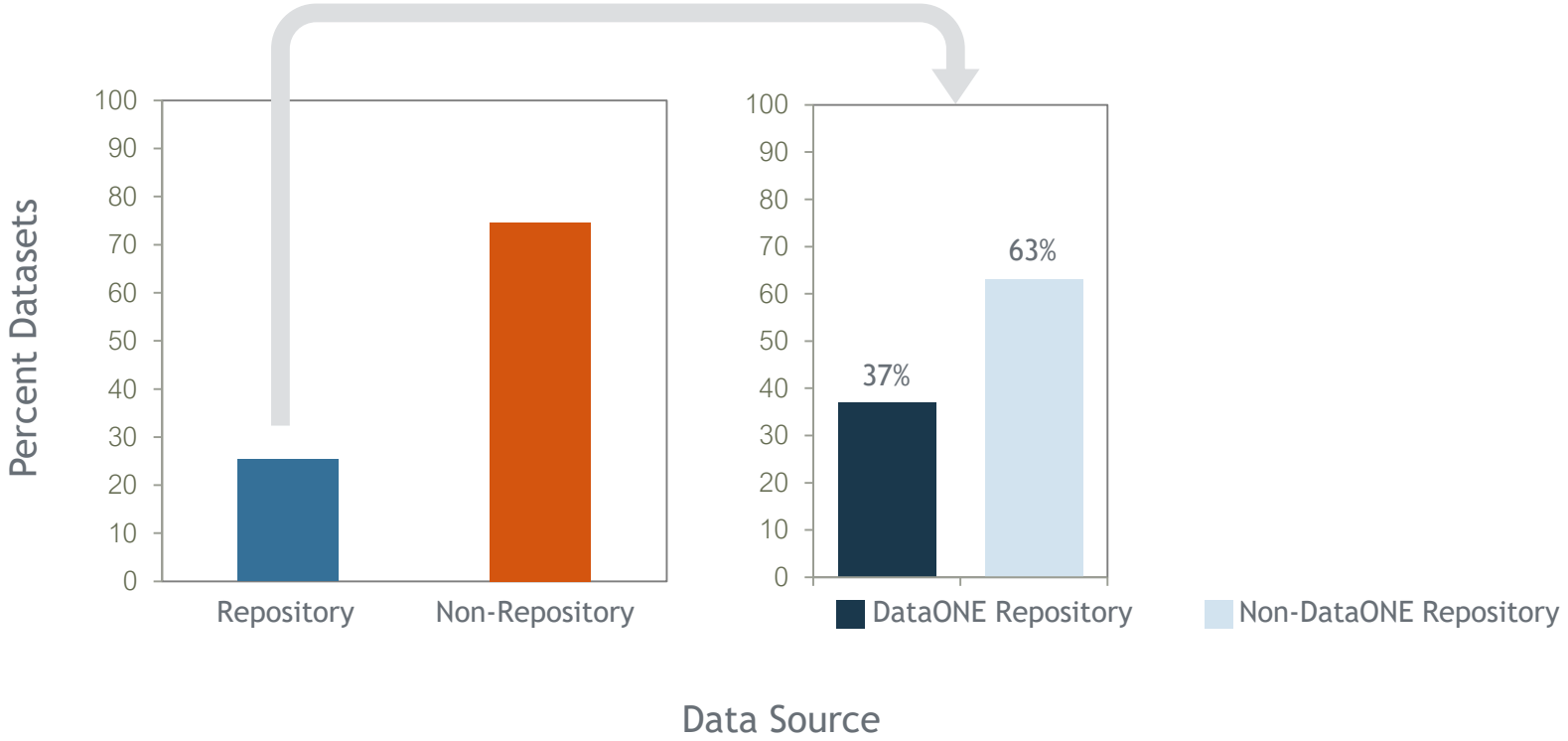


- Synthesis research
- 80 publications
- 505 datasets
- Average 6.4 reused datasets per study

Budden et al, in prep

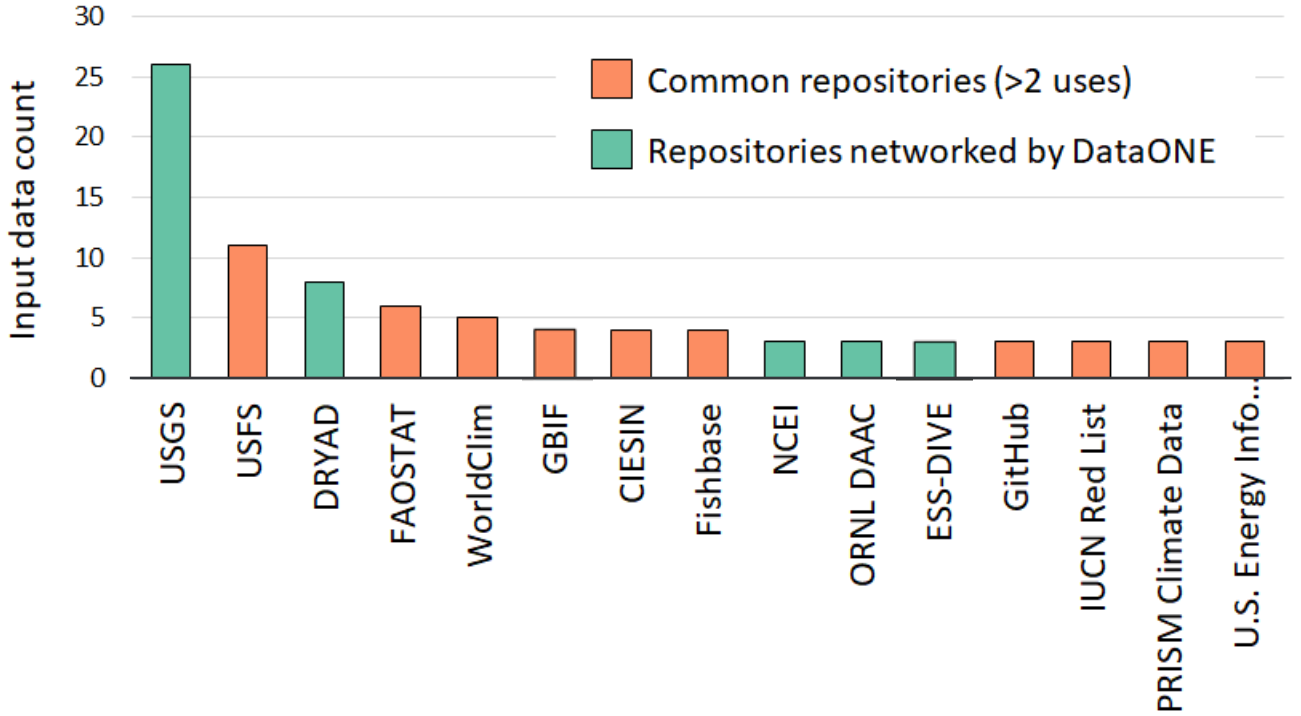
Culture of Data Sharing is Improving

But there is still opportunity for growth



Budden et al, in prep

Informing DataONE Design



Budden et al, in prep

Building and Engaging Community

Community Informed Design

Surveys

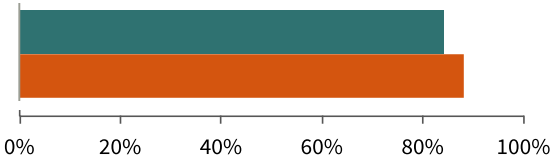
User scenarios

Personas

Usability assessments

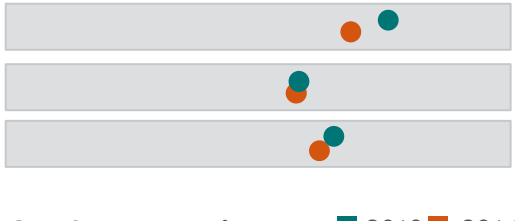
Perception

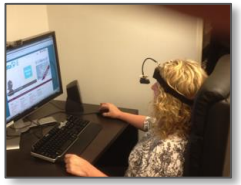
Use others' datasets if their data were easily accessible

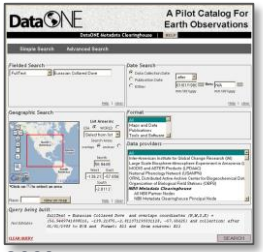


Satisfaction

Process for searching for own data
Tools for preparing metadata
Tools for preparing documentation

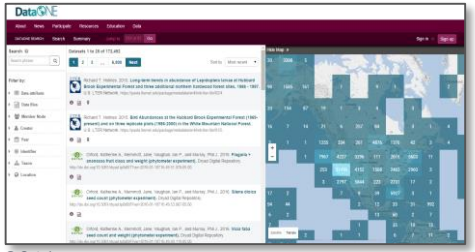






2009

2012



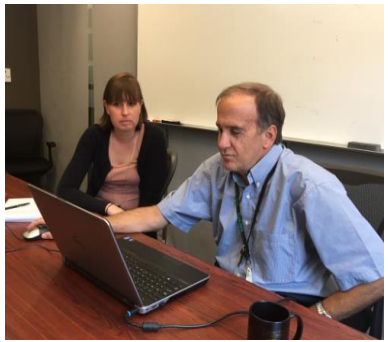
2016

Tenopir et al 2011, Tenopir et al 2015

DataONE Community

Annual Community Meeting

Inform development prioritization and direction of DataONE

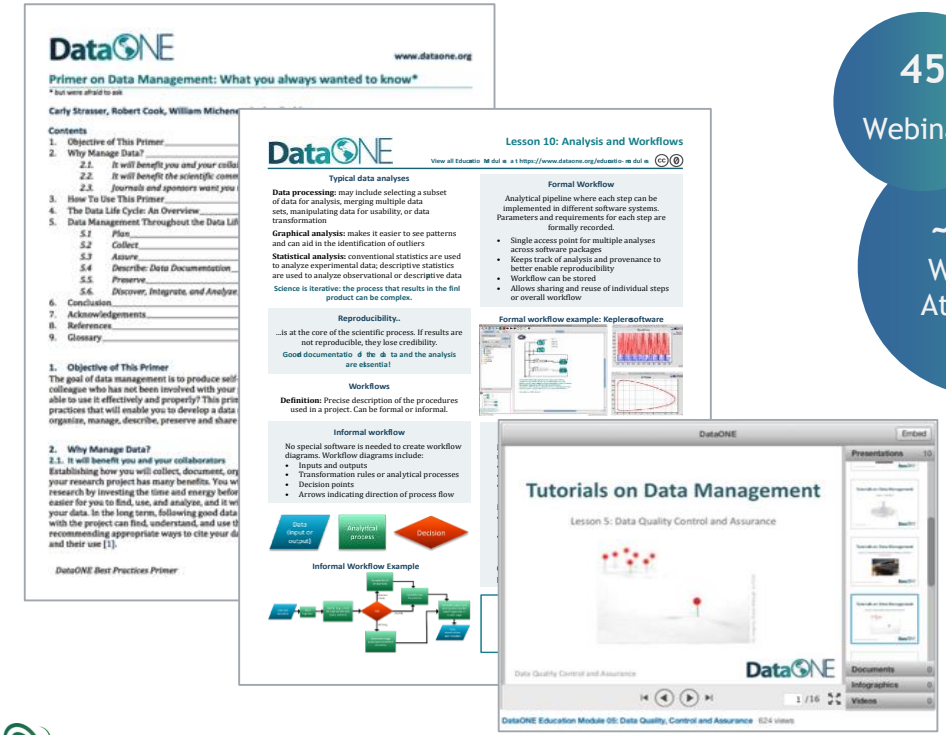


User testing and feedback

Community challenges and shared solutions

Training

Leadership in Data Management Education



45 Webinars

~4100 Webinar Attendees

Greg Wilson

Liz Ferguson

Stephanie Hampton

Fernando Pérez

NCEAS National Center for Ecological Analysis and Synthesis

DataONE

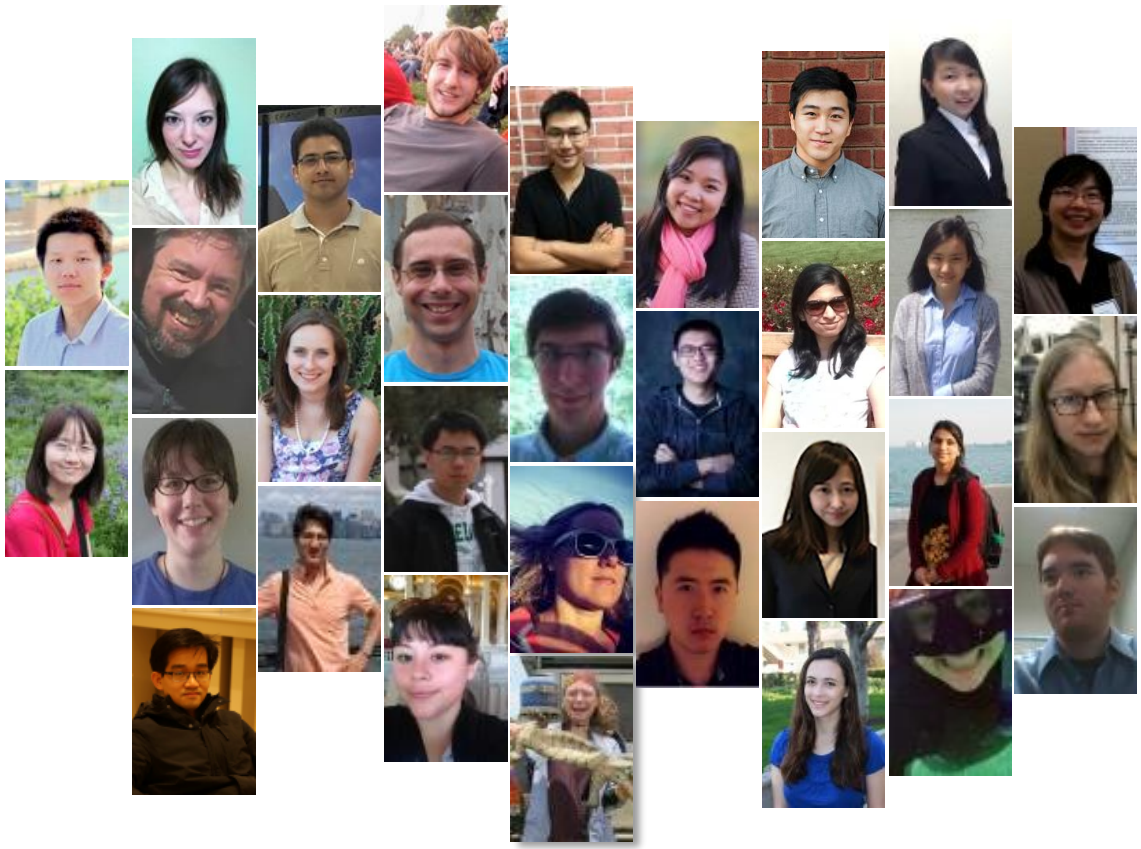
Broadening Participation

DataONE Summer Internship Program

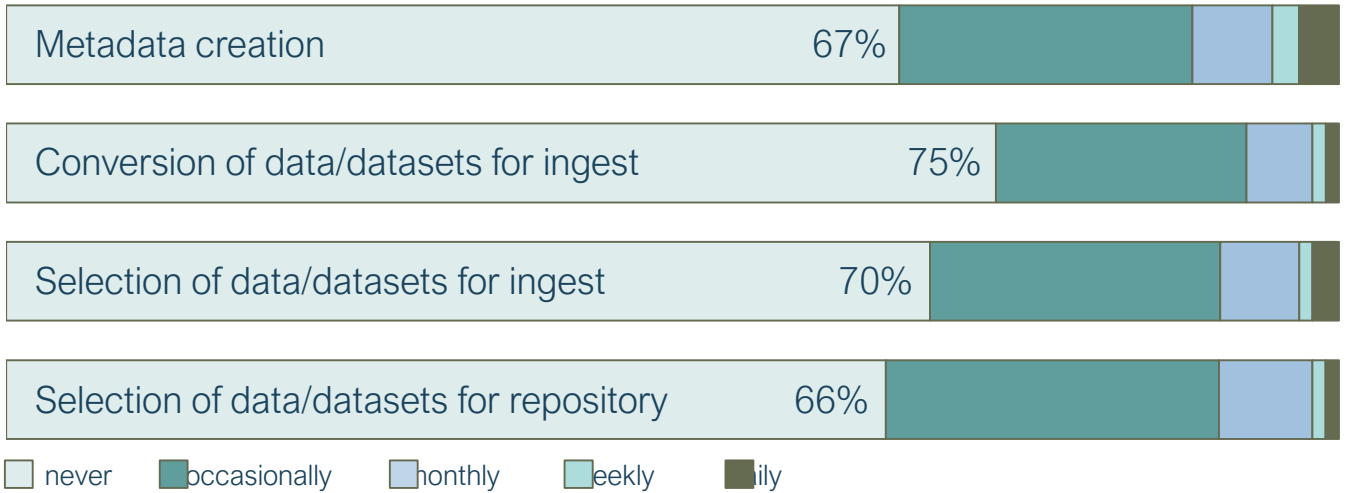
- 2009 – 4 interns
- 2010 – 4 interns
- 2011 – 8 interns
- 2012 – 6 interns
- 2013 – 8 interns
- 2014 – 10 interns
- 2015 – 4 int
- 2016 – 5
- 2017 – 6
- 2018 –
- 2019 – 5

64
Interns
Mentored

*2 interns spo ther organizations

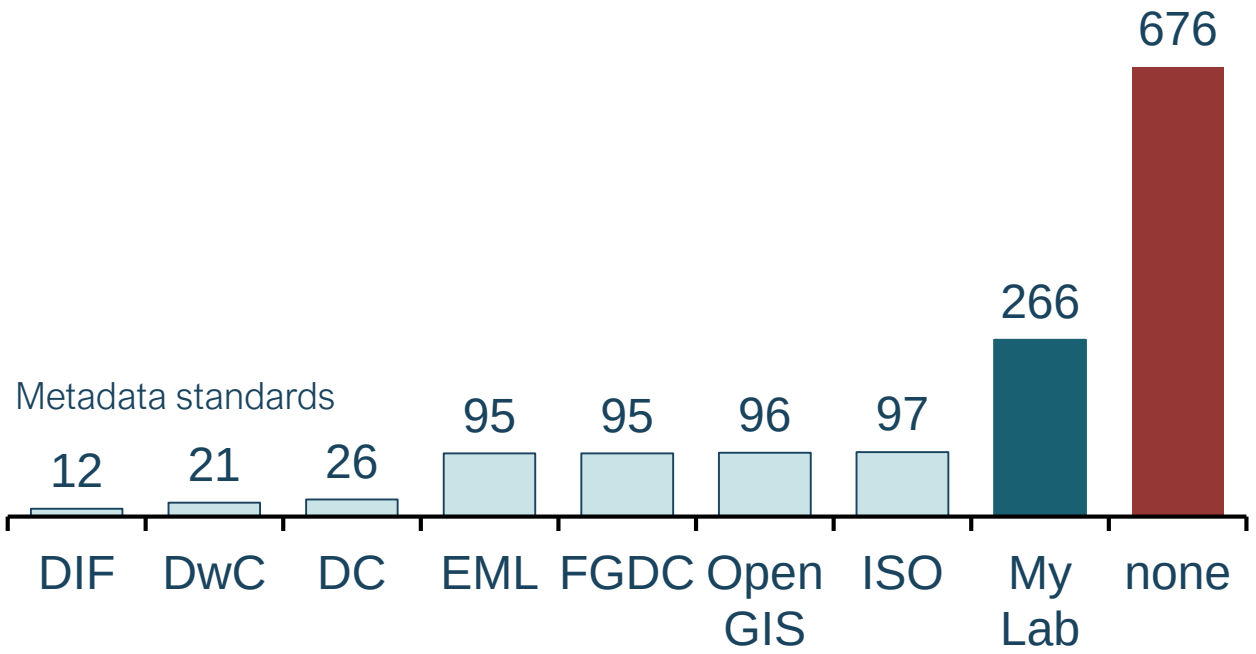


Demonstrated Need for Training



Tenopir et al 2014

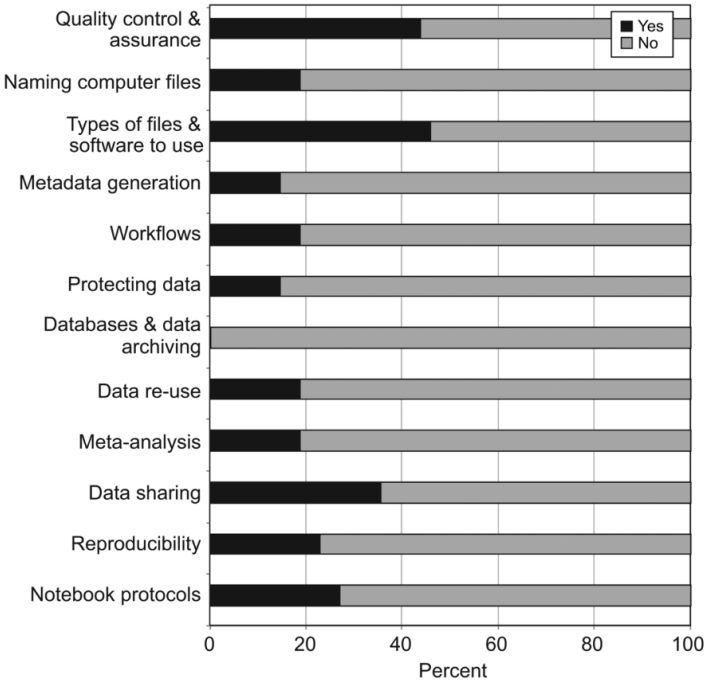
Demonstrated Need for Training



Tenopir et al 2011, Tenopir et al 2014

Status of Data Management Education

Survey of Ecology Courses

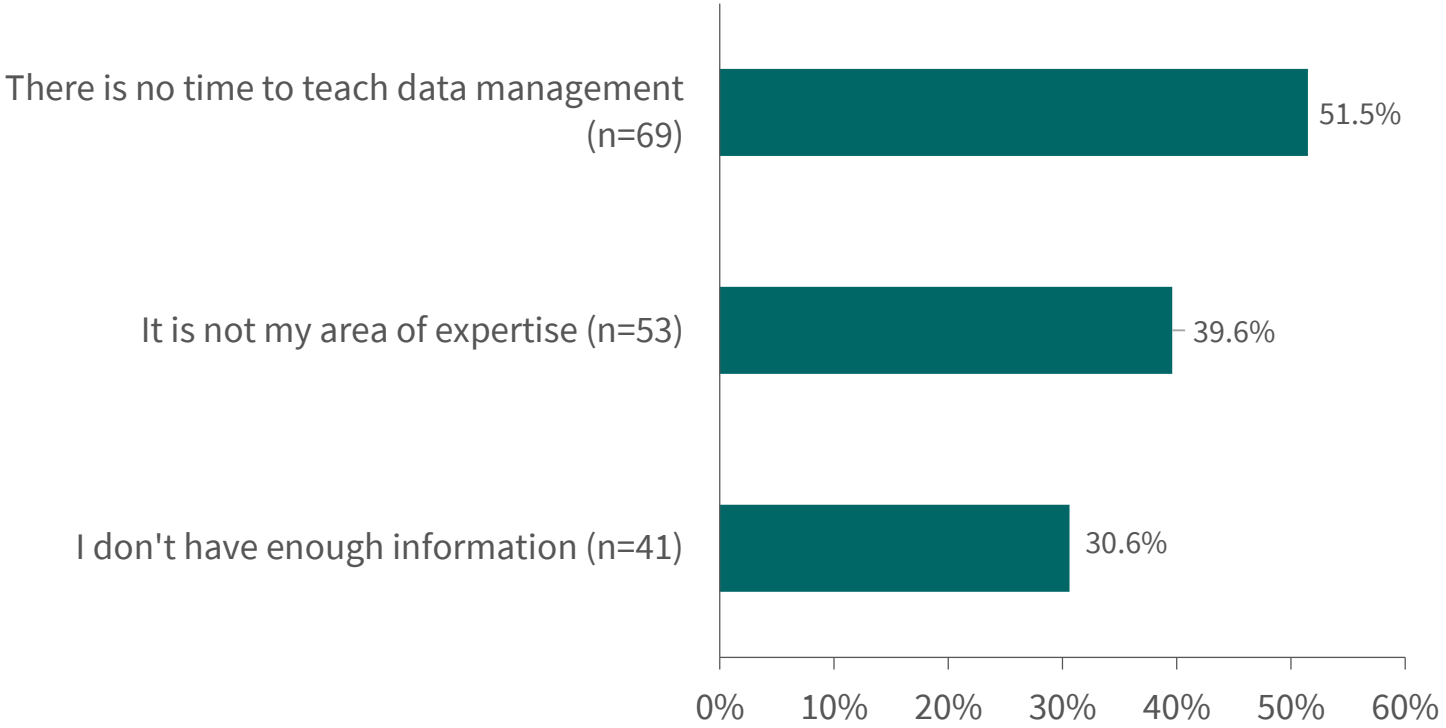


Percent of ecology courses that address and/or teach the data management topics listed

Strasser and Hampton 2012

Challenges in Data Management Training

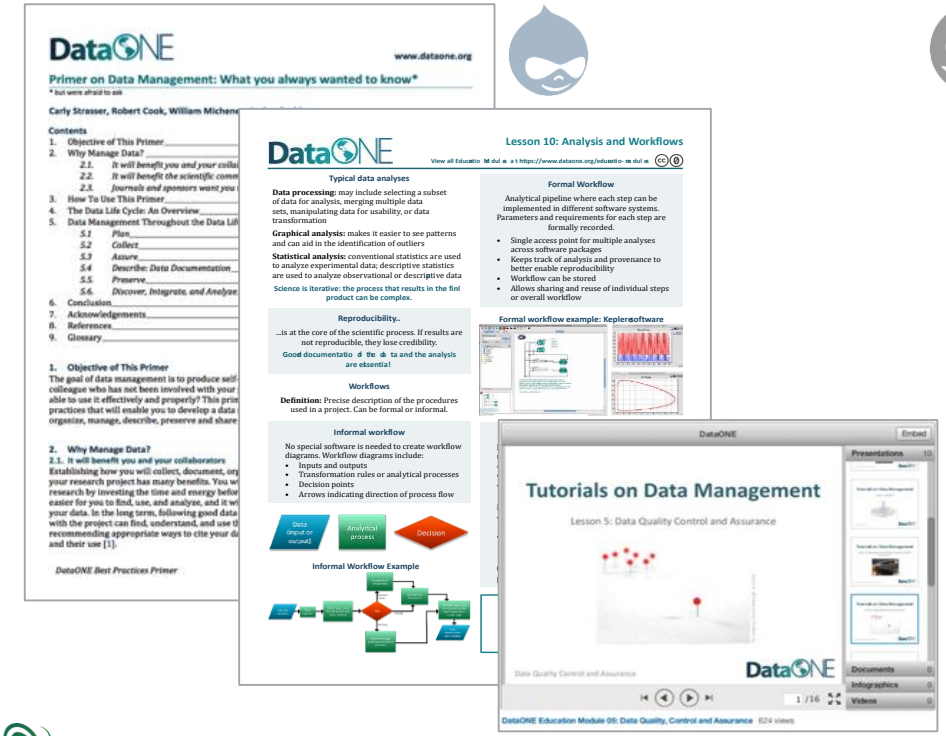
Survey of Educators

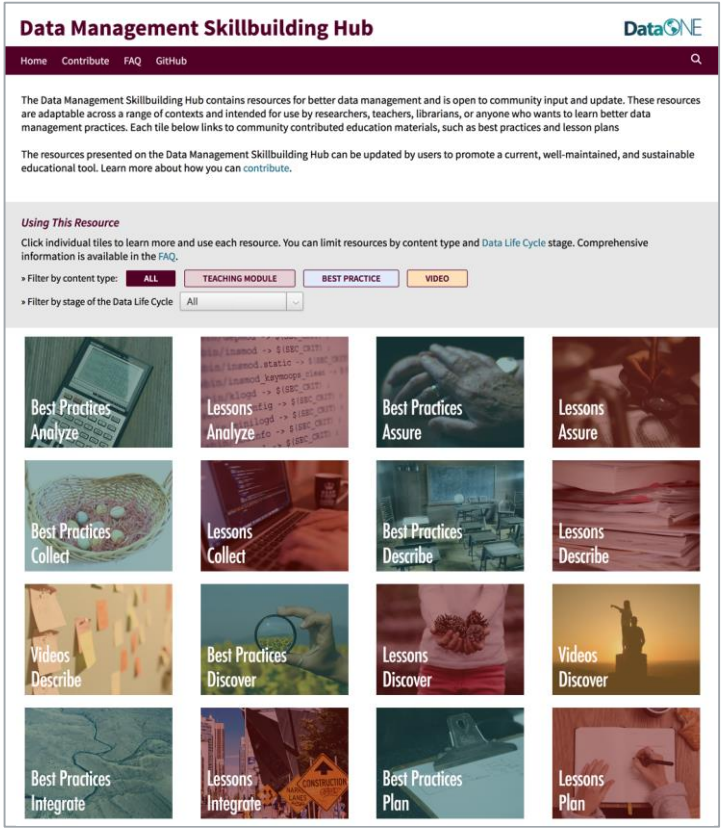


Strasser and Hampton 2012

Education Materials

The Evolution of Resources





NCEAS
National Center for Ecological Analysis and Synthesis

DataONE

NCEAS Learning Hub

National Center for Ecological Analysis and Synthesis

A knowledge-sharing community where researchers can learn the latest data skills and technologies to increase efficiency, productivity, transparency, and collaborative capacity.

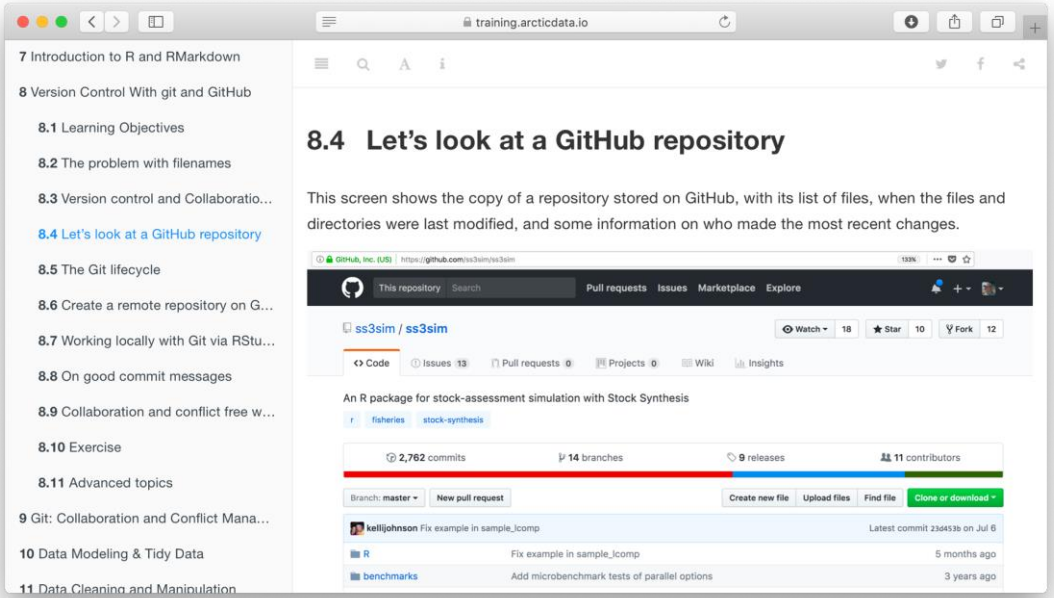
Courses: Fee-based and grant-supported intensive data science workshops

Mentored Programs: Experiential residential and remote learning programs to build skills in data and open science

Resources: Extensive online curricula, webinars, training materials and best practices

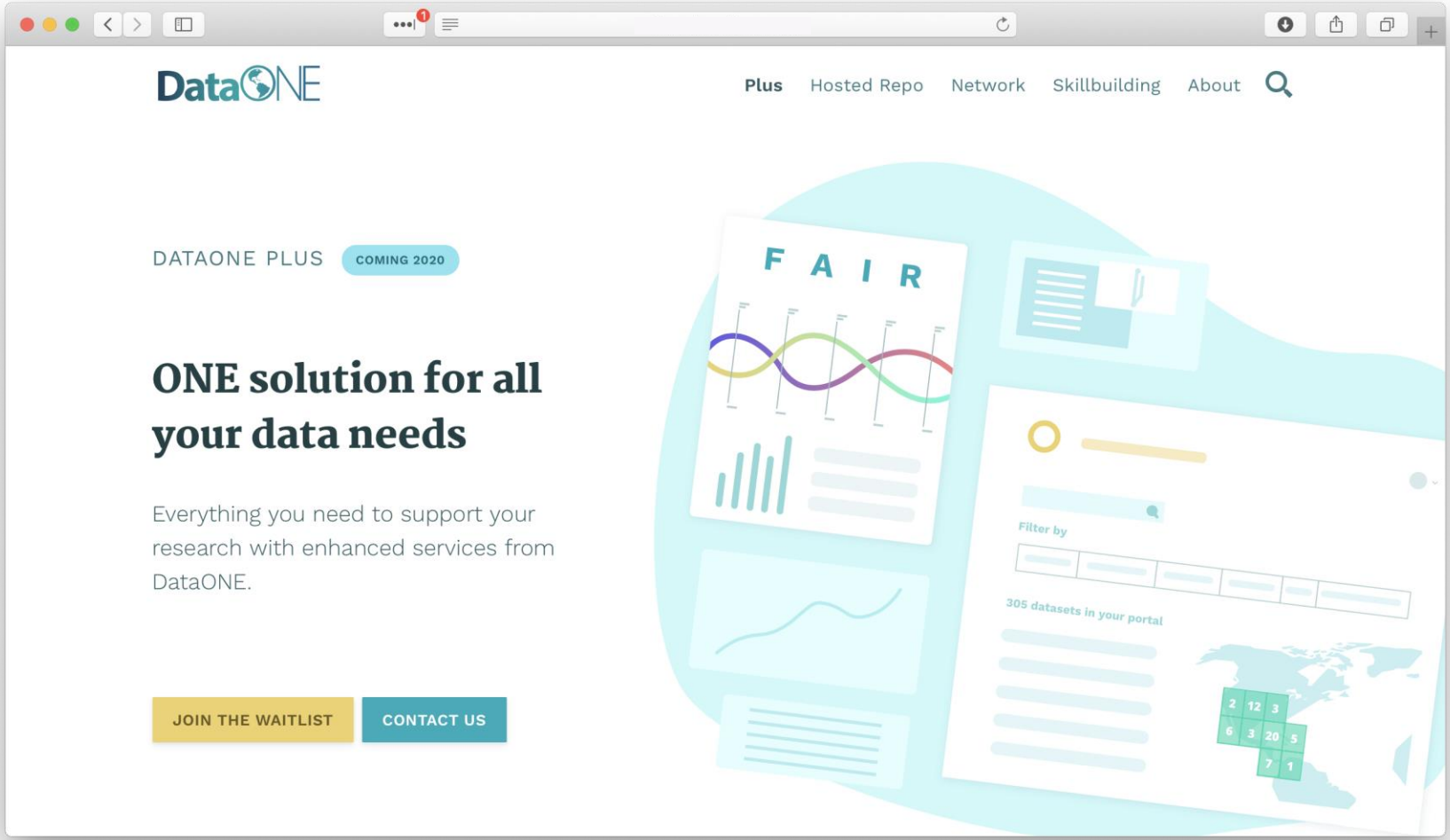
Partnerships: Customized workshops and collaborative initiatives in data science training

www.nceas.ucsb.edu/learning-hub


training.arcticdata.io
7 Introduction to R and RMarkdown
8 Version Control With git and GitHub
8.1 Learning Objectives
8.2 The problem with filenames
8.3 Version control and Collaboratio...
8.4 Let's look at a GitHub repository
8.5 The Git lifecycle
8.6 Create a remote repository on G...
8.7 Working locally with Git via RStu...
8.8 On good commit messages
8.9 Collaboration and conflict free w...
8.10 Exercise
8.11 Advanced topics
9 Git: Collaboration and Conflict Mana...
10 Data Modeling & Tidy Data
11 Data Cleaning and Manipulation
8.4 Let's look at a GitHub repository
This screen shows the copy of a repository stored on GitHub, with its list of files, when the files and directories were last modified, and some information on who made the most recent changes.
This repository
Search
Pull requests
Issues
Marketplace
Explore
ss3sim / ss3sim
Watch 18
Star 10
Fork 12
Code
Issues 13
Pull requests 0
Projects 0
Wiki
Insights
An R package for stock-assessment simulation with Stock Synthesis
r
fisheries
stock-synthesis
2,762 commits
14 branches
9 releases
11 contributors
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
kelliljohnson Fix example in sample_lcomp
Latest commit 23d453b on Jul 6
R
Fix example in sample_lcomp
5 months ago
benchmarks
Add microbenchmark tests of parallel options
3 years ago

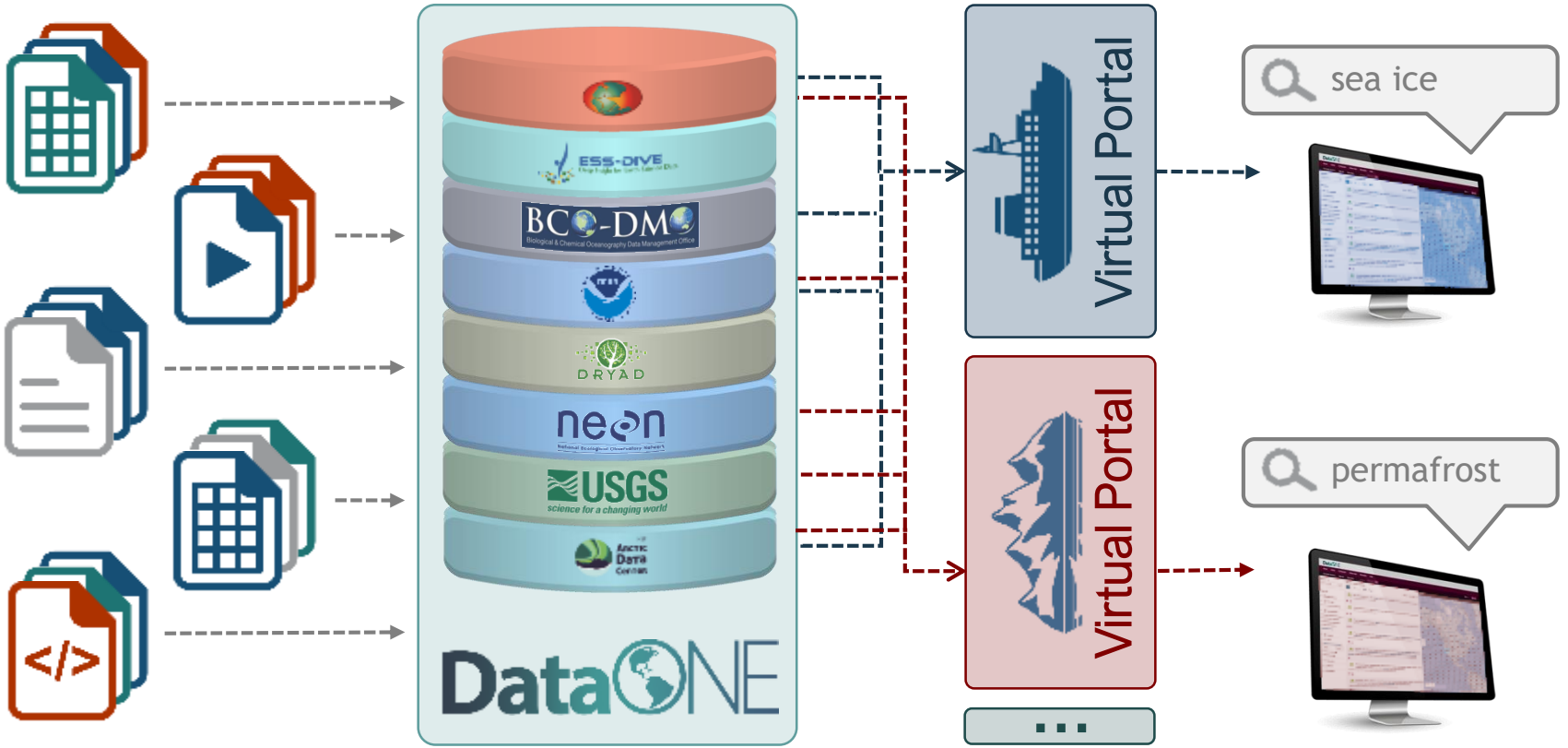
DataONE: An Interoperable Federation

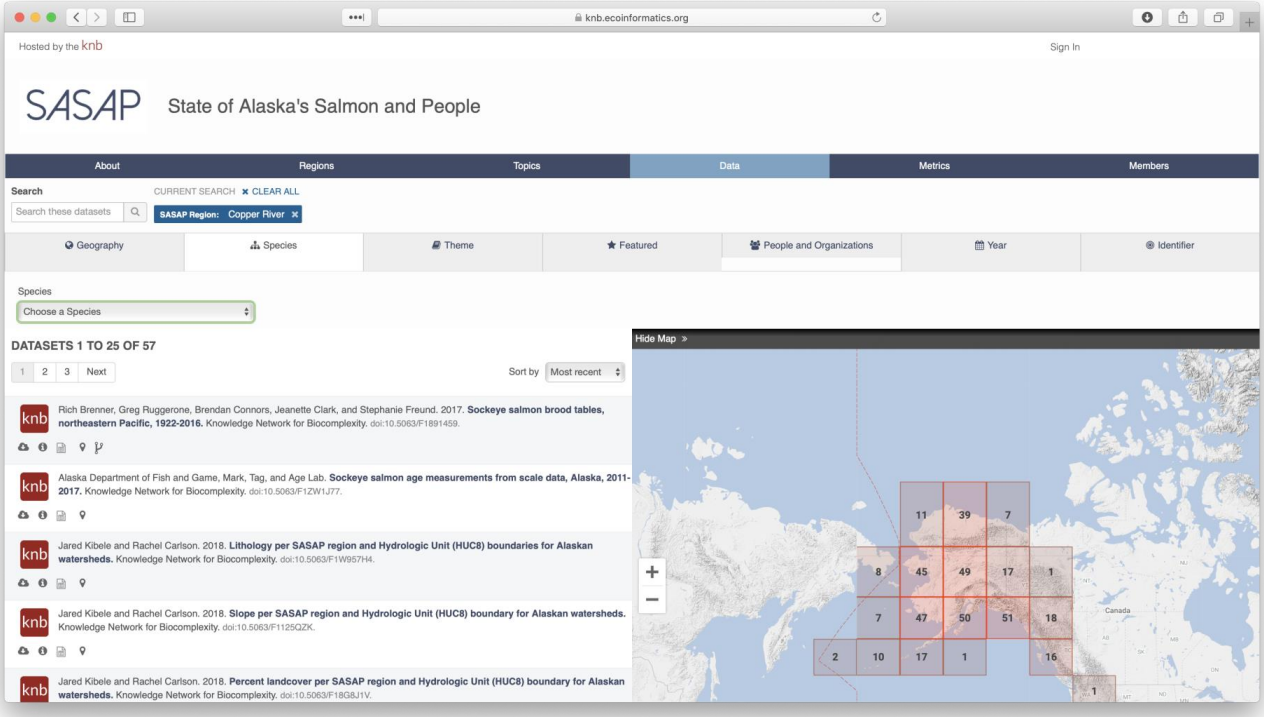
DataONE Plus

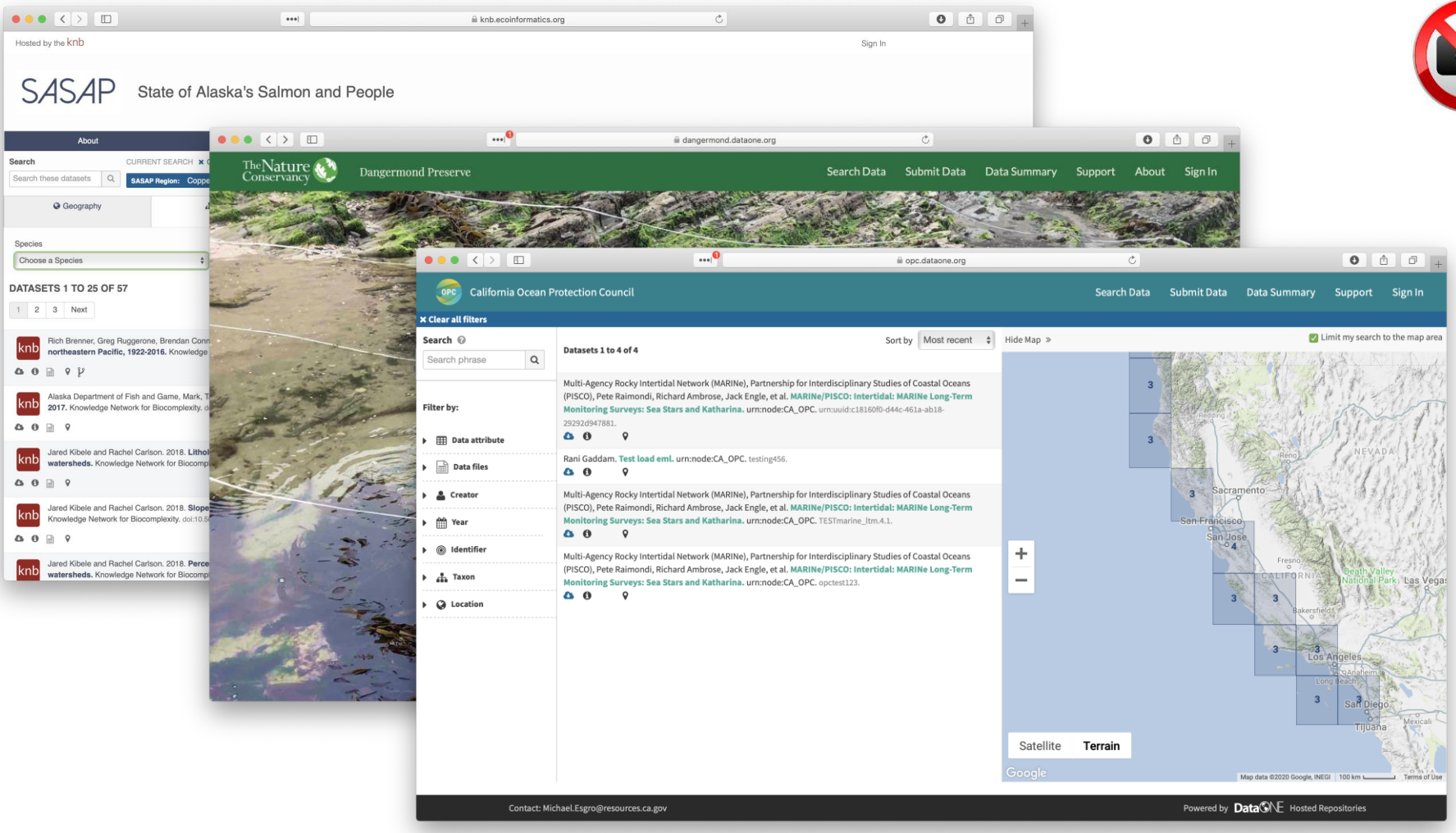


Data Portals

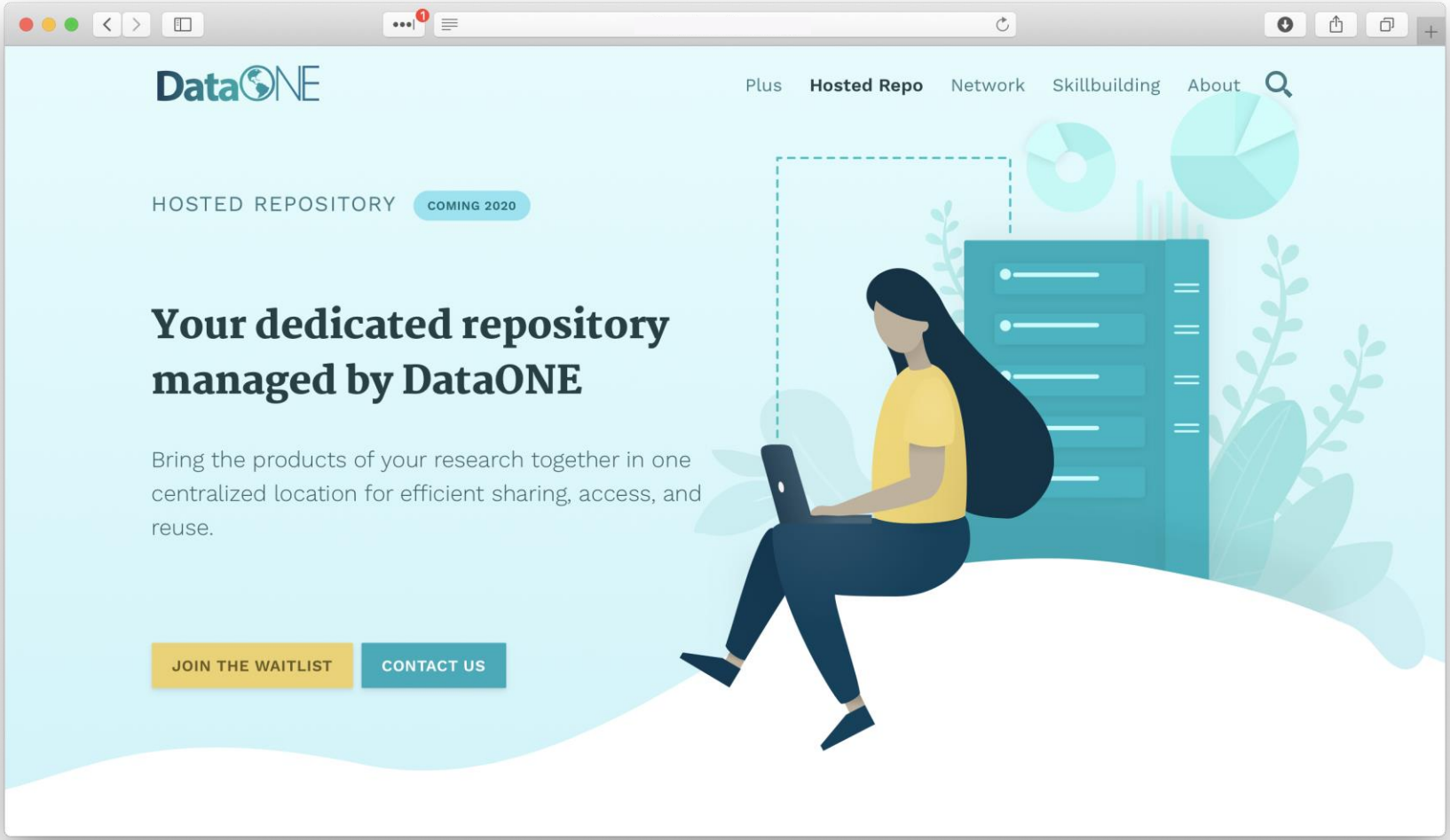
Data Aggregated from Repositories Across DataONE




knb.ecoinformatics.org
Hosted by the knb
Sign In
SASAP
State of Alaska's Salmon and People
About
Regions
Topics
Data
Metrics
Members
Search
CURRENT SEARCH
CLEAR ALL
Search these datasets
SASAP Region: Copper River
Geography
Species
Theme
Featured
People and Organizations
Year
Identifier
Species
Choose a Species
DATASETS 1 TO 25 OF 57
1
2
3
Next
Sort by
Most recent
Rich Brenner, Greg Ruggerone, Brendan Connors, Jeanette Clark, and Stephanie Freund. 2017. Sockeye salmon brood tables, northeastern Pacific, 1922-2016. Knowledge Network for Biocomplexity. doi:10.5063/F1891459.
Alaska Department of Fish and Game, Mark, Tag, and Age Lab. Sockeye salmon age measurements from scale data, Alaska, 2011-2017. Knowledge Network for Biocomplexity. doi:10.5063/F1ZW1J77.
Jared Kibele and Rachel Carlson. 2018. Lithology per SASAP region and Hydrologic Unit (HUC8) boundaries for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F1W957H4.
Jared Kibele and Rachel Carlson. 2018. Slope per SASAP region and Hydrologic Unit (HUC8) boundary for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F1125QZK.
Jared Kibele and Rachel Carlson. 2018. Percent landcover per SASAP region and Hydrologic Unit (HUC8) boundary for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F18G8J1V.
Hide Map
11
39
7
8
45
49
17
1
7
47
50
51
18
2
10
17
1
16
1
Canada


knb.ecoinformatics.org
Hosted by the knb
Sign In
SASAP
State of Alaska's Salmon and People
About
Search
CURRENT SEARCH
Search these datasets
SASAP Region:
Geography
Species
Choose a Species
DATASETS 1 TO 25 OF 57
1 2 3 Next
dangermond.dataone.org
The Nature Conservancy
Dangermond Preserve
Search Data
Submit Data
Data Summary
Support
About
Sign In
opc.dataone.org
California Ocean Protection Council
Search Data
Submit Data
Data Summary
Support
Sign In
Clear all filters
Search
Search phrase
Filter by:
Data attribute
Data files
Creator
Year
Identifier
Taxon
Location
Datasets 1 to 4 of 4
Sort by
Most recent
Hide Map
Limit my search to the map area
Multi-Agency Rocky Intertidal Network (MARINe), Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Pete Raimondi, Richard Ambrose, Jack Engle, et al. MARINe/PISCO: Intertidal: MARINe Long-Term Monitoring Surveys: Sea Stars and Katharina. urn:node:CA_OPC.
Rani Gaddam. Test load eml. urn:node:CA_OPC. testing456.
Multi-Agency Rocky Intertidal Network (MARINe), Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Pete Raimondi, Richard Ambrose, Jack Engle, et al. MARINe/PISCO: Intertidal: MARINe Long-Term Monitoring Surveys: Sea Stars and Katharina. urn:node:CA_OPC. TESTmarine_ltm.4.1.
Multi-Agency Rocky Intertidal Network (MARINe), Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Pete Raimondi, Richard Ambrose, Jack Engle, et al. MARINe/PISCO: Intertidal: MARINe Long-Term Monitoring Surveys: Sea Stars and Katharina. urn:node:CA_OPC. opctest123.
Satellite
Terrain
Google
Contact: Michael.Esgro@resources.ca.gov
Powered by DataONE Hosted Repositories

Hosted Repositories



NSF DataNet

Sustainable Digital Data Preservation and Access Network Partners (DataNet)

I. INTRODUCTION

Chapter 3 (Data, Data Analysis, and Visualization) of NSF's Cyberinfrastructure Vision for 21st Century Discovery (https://www.nsf.gov/pubs/2007/nsf0728/index.jsp) presents a vision in which "science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved." The goal of this solicitation is to catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable.

January 07, 2008

November 13, 2008

Full Proposal Target Date(s):

March 21, 2008

May 15, 2009

DataONE: Supporting Data Discovery and Access

Federating across networks increases data discovery

- Community informed design allows for positive user experience and technical interoperability

Continued training and outreach required for increased data literacy and data preservation

aebudden@nceas.ucsb.edu

dataone.org
search.dataone.org
www.nceas.ucsb.edu/learning-hub

NCEAS
National Center for Ecological Analysis and Synthesis

DataONE