









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

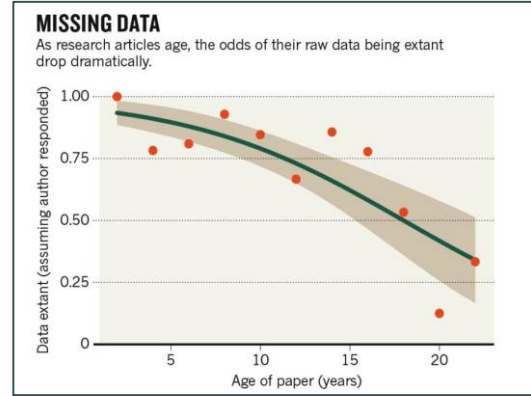
What environmental data are you looking for?



ADVANCED SEARCH

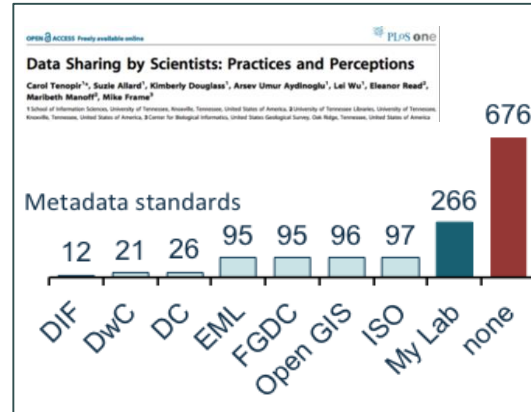


# Science and Informatics Challenges



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



## **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 21<sup>st</sup> 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):**

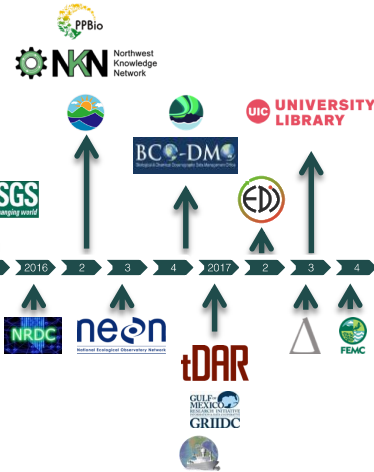
March 21, 2008

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

## Data Table, Image, and Other Data Details

4 sources

Source Data

Alkane.csv

Citation

[View »](#)

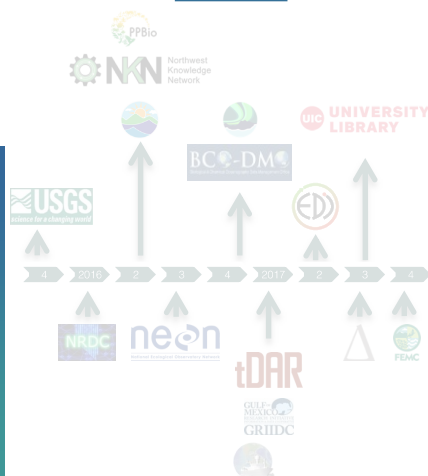
This data was generated by `</>`  
`DataDownload.R`.

This data prov\_hasDerivations the data  
you are currently viewing, `</>`  
`Total_Aromatic_Alkanes_PWS.csv` and  
the data you are currently viewing, `</>`  
`Total_Aromatic_Alkanes_PWS.csv`.

This data was used by `</>`  
`Total_PAH_and_Alkanes_GoA_Hydroca  
rbons_Clean.R`.

# Three Primary Goals

Building an  
Empowered  
and Engaged  
Community



Enabling  
Reproducible  
Science through  
Tools and Services

Developing  
sustainable data  
discovery and  
interoperability  
solutions

## Data Table, Image, and Other Data Details

4 sources

Source Data

Alkane.csv

Citation

View >

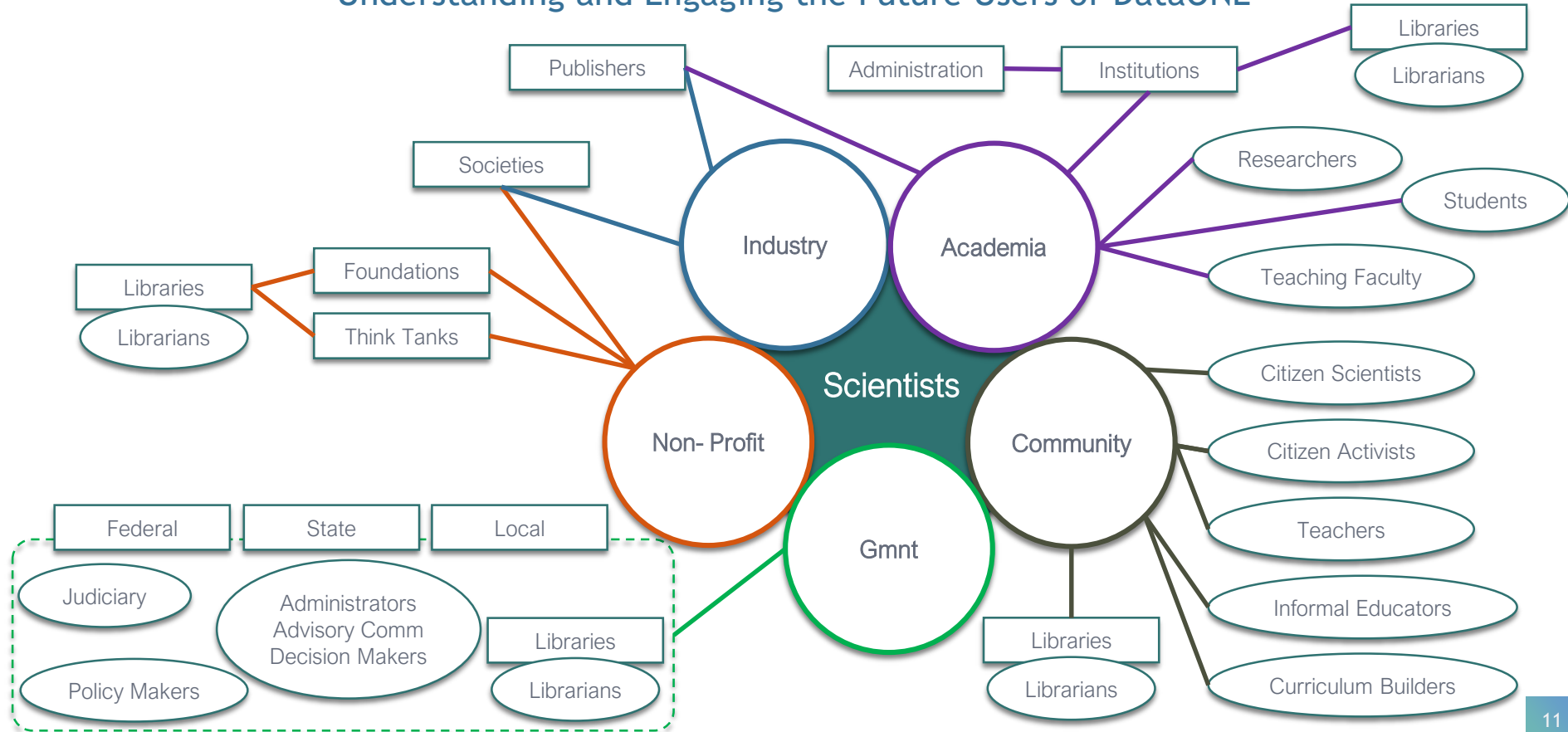
This data was generated by [DataDownload.R](#).

This data prov\_hasDerivations the data you are currently viewing, [Total\\_Aromatic\\_Alkanes\\_PWS.csv](#) and the data you are currently viewing, [Total\\_Aromatic\\_Alkanes\\_PWS.csv](#).

This data was used by [Total\\_PAH\\_and\\_Alkanes\\_GoA\\_Hydrocarbons\\_Clean.R](#).

# 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

## 1<sup>st</sup> Scientist Survey (2011)



### 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

## 2<sup>nd</sup> Scientist Survey (2015)



### 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



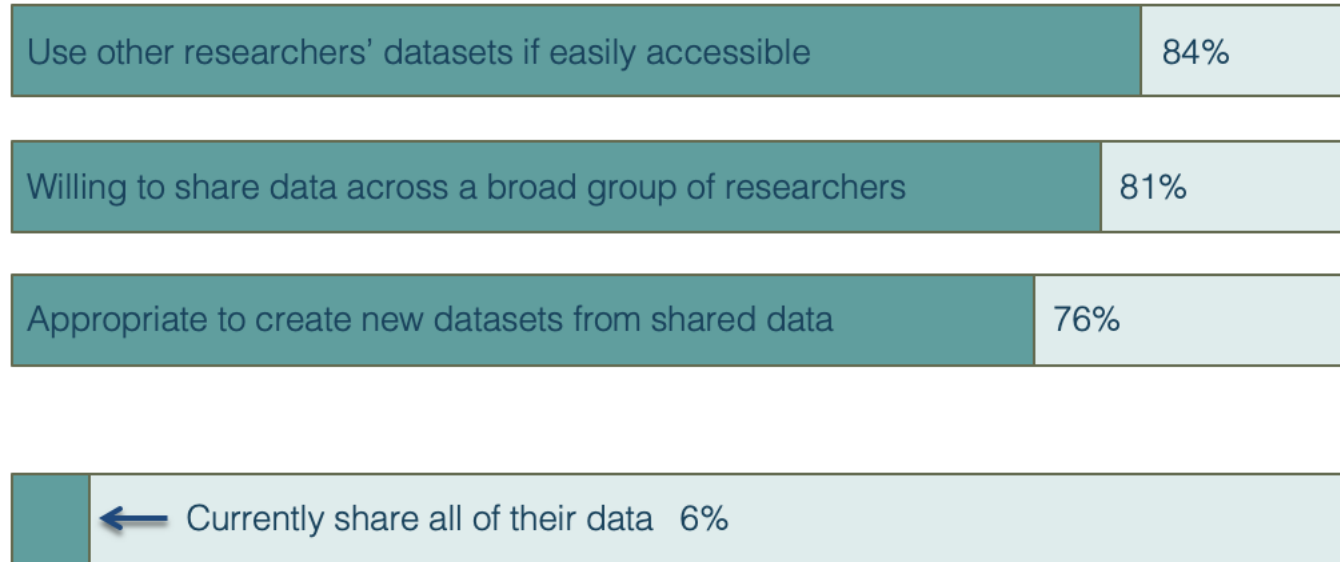
# Scientists Want to Share Data

---



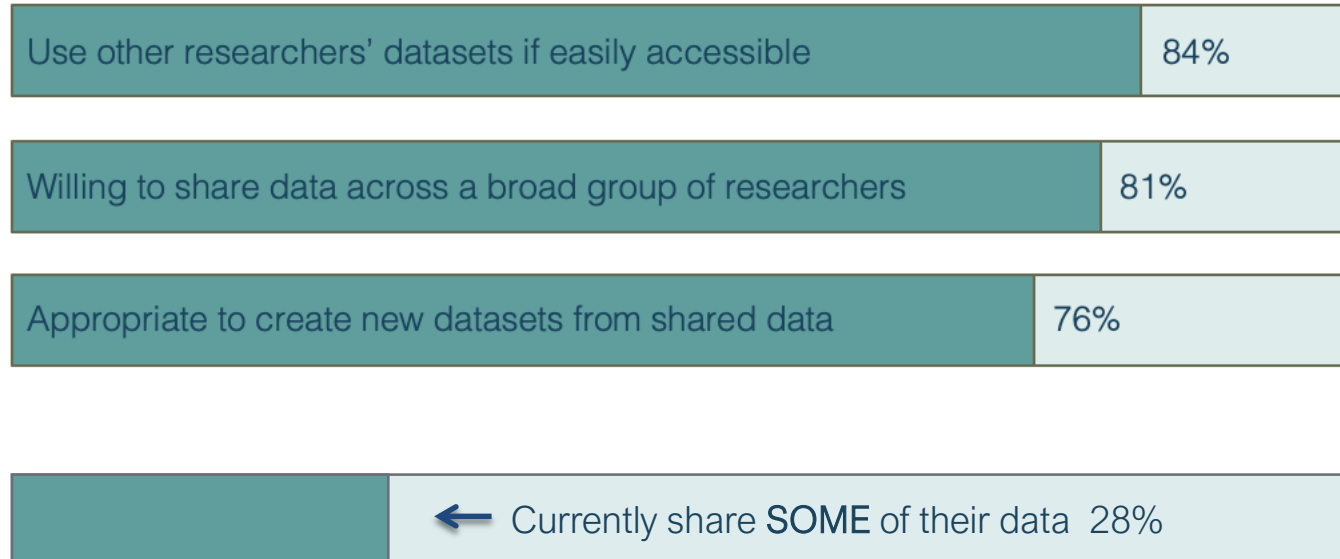
# Scientists Want to Share Data

---

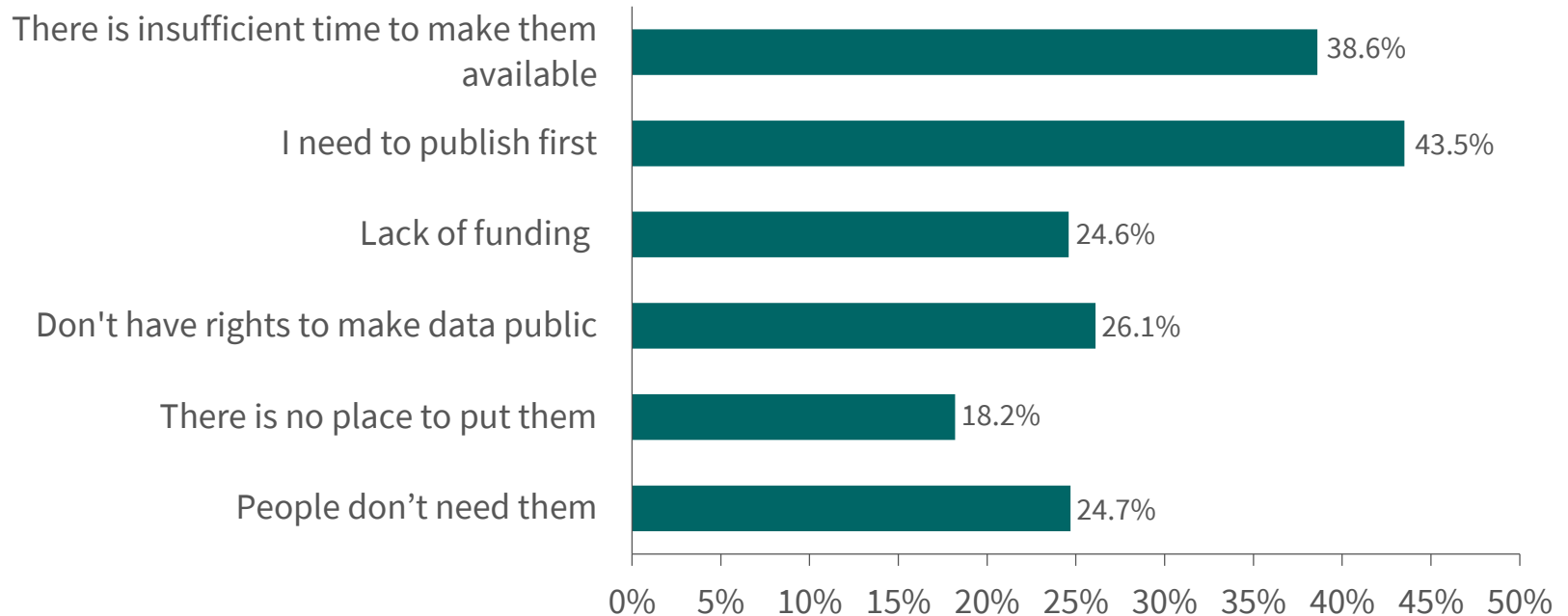


# Scientists Want to Share Data

---



# Perceived Barriers to Data Sharing

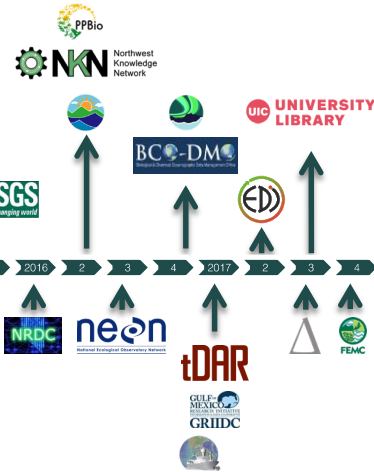
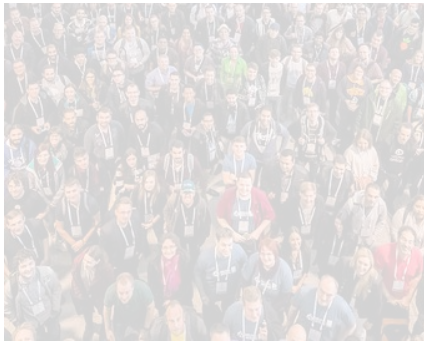


# Perceived Barriers to Data Sharing



# Three Primary Goals

Building an  
Empowered  
and Engaged  
Community



Enabling  
Reproducible  
Science through  
Tools and Services

Developing  
sustainable data  
discovery and  
interoperability  
solutions

## Data Table, Image, and Other Data Details

4 sources

Source Data

Alkane.csv

Citation

View >

This data was generated by `</>`  
DataDownload.R.

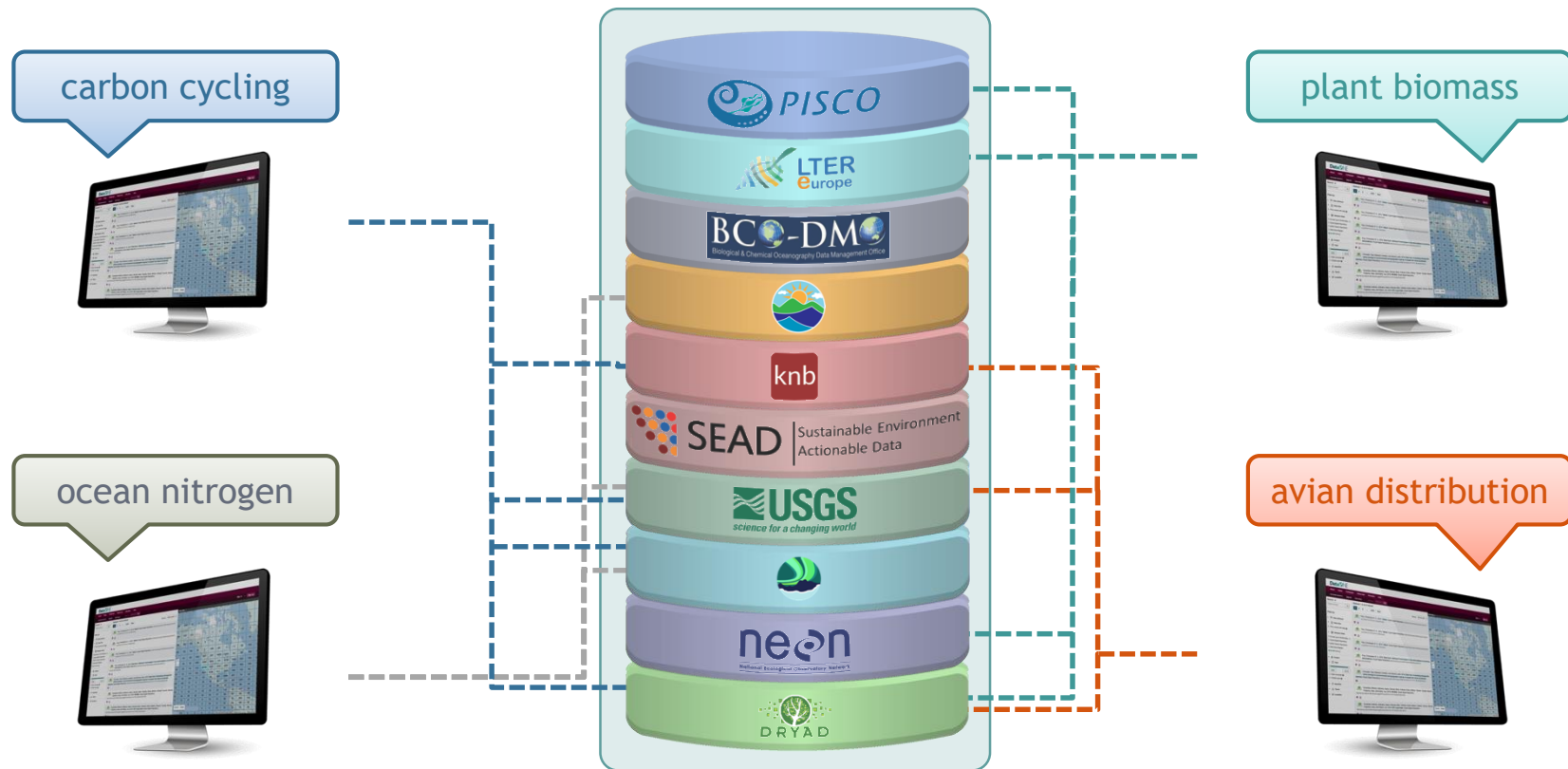
This data prov\_hasDerivations the data  
you are currently viewing, `</>`  
Total\_Aromatic\_Alkanes\_PWS.csv and  
the data you are currently viewing, `</>`  
Total\_Aromatic\_Alkanes\_PWS.csv.

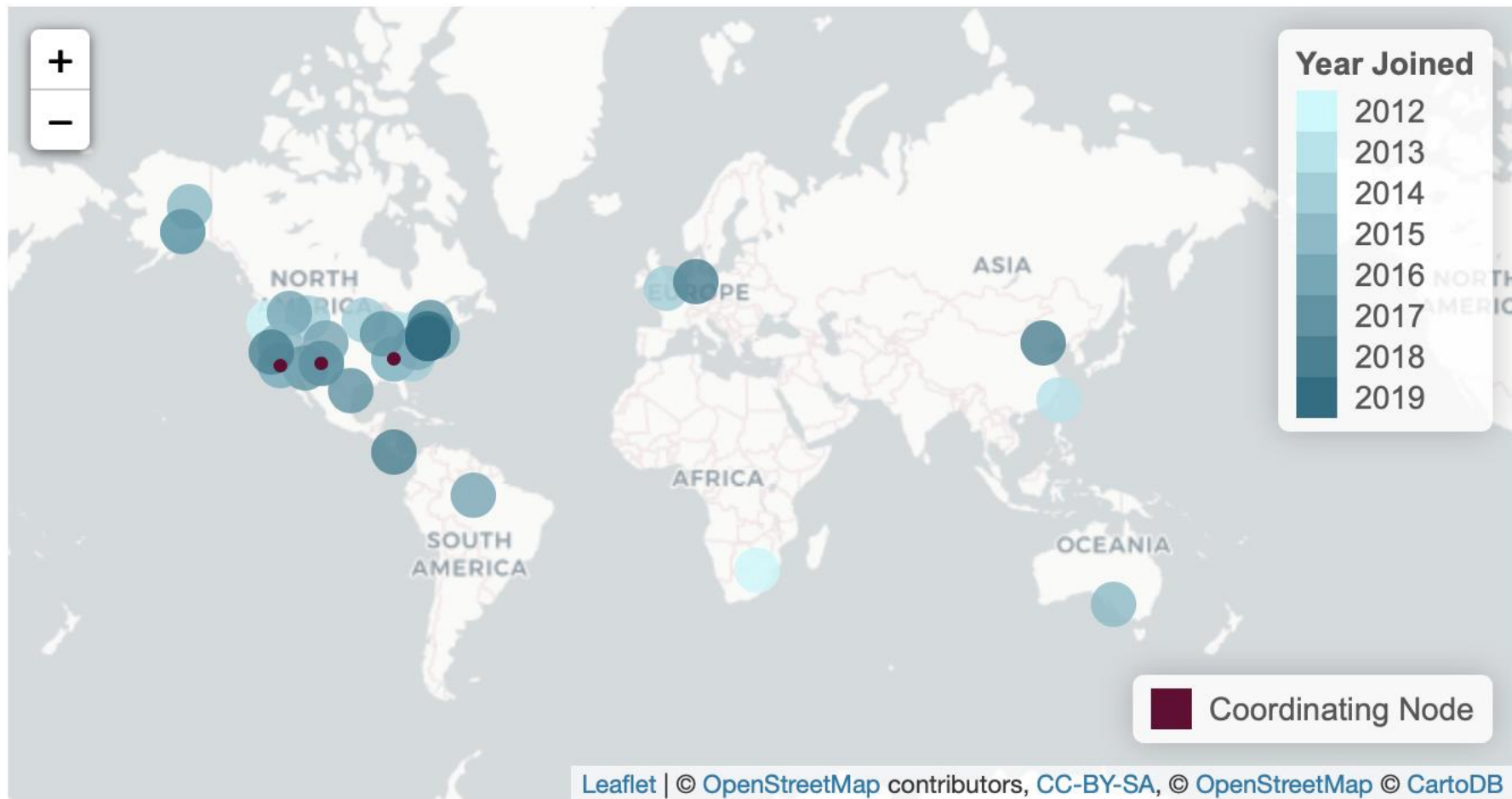
This data was used by `</>`  
Total\_PAH\_and\_Alkanes\_GoA\_Hydroc  
rbons\_Clean.R.



# Federated Search

## Data Discovery and Access from Multiple Repositories





Search phrase

Location: new zealand

Geographic region 

1 2 3 ... 74 Next

Sort by Most recent



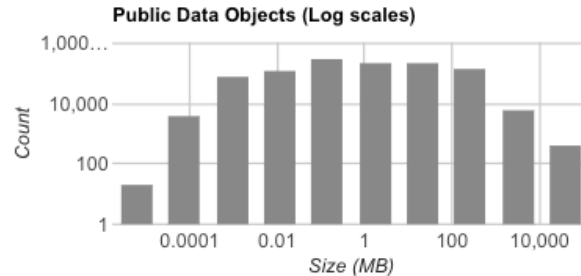
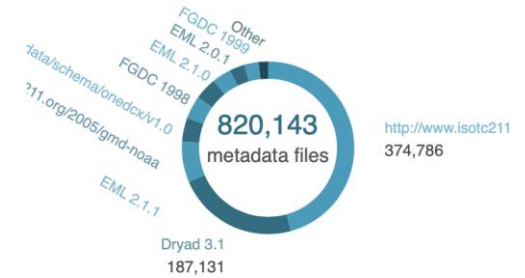
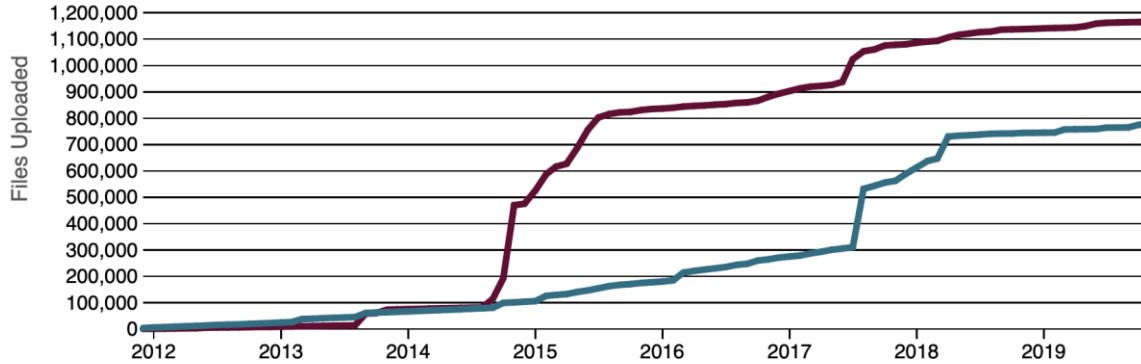
## Terrain

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

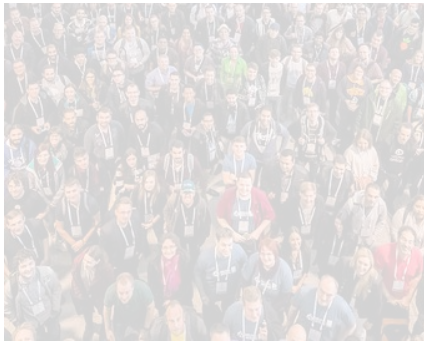


# Building the Federation

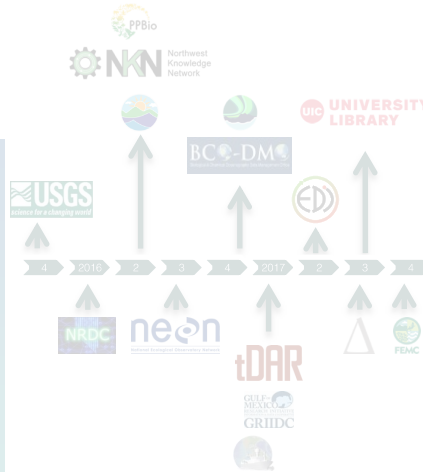


# Three Primary Goals

Building an  
Empowered  
and Engaged  
Community



Developing  
sustainable data  
discovery and  
interoperability  
solutions



Enabling  
Reproducible  
Science through  
Tools and Services

## Data Table, Image, and Other Data Details

4 sources

Source Data

**Alkane.csv**

Citation

[View »](#)


This data was generated by `</>`  
**DataDownload.R.**

This data prov\_hasDerivations the data  
you are currently viewing, **Total\_Aromatic\_Alkanes\_PWS.csv** and  
the data you are currently viewing, **Total\_Aromatic\_Alkanes\_PWS.csv.**

This data was used by `</>`  
**Total\_PAH\_and\_Alkanes\_GoA\_Hydroca  
rbons\_Clean.R.**




# Federated Search Across Repositories



- About
- News
- Participate
- Resources
- Education
- Data

DATAONE SEARCH: Search Summary Jump to: DOI or ID Go

Search 


Search phrase


Filter by:


- Data attribute
 


density, length, etc.
- Data files
- Member Node
- Creator
 


Name
- Year
- Identifier
- Taxon
- Location


 Fabes, Jez, Seligman, William, Barrett, Carolyn, McKee, Stuart, and Griffiths, John. 2017. **Data from: Does the implementation of a novel intensive care discharge risk score and nurse-led inpatient review tool improve outcome? A prospective cohort study in two intensive care units in the United Kingdom.** Dryad Digital Repository. <https://doi.org/10.5061/dryad.7f77q?ver=2017-11-22T09:11:59.034-05:00>.


 Park, Hyeree, Cho, Sooyoung, Woo, Hyeontaek, Park, Sue K., Shin, Hai-Rim, et al. 2017. **Data from: Fasting glucose and risk of colorectal cancer in the Korean Multi-center Cancer Cohort.** Dryad Digital Repository. <https://doi.org/10.5061/dryad.11858?ver=2017-11-22T09:03:35.154-05:00>.


 Jason Schatz, Carly Ziter, and Christopher Kucharik. 2016. **WSC - Temperature and relative humidity data from 150 locations in and around Madison, Wisconsin from 2012-2016.** LTER Network Member Node. <https://pasta.lternet.edu/package/metadata/eml/knb-lter-ni/324/12>.


 Heather Bateman, Dan Childers, Madhusudan Katti, Eyal Shochat, and Paige Warren. 2017. **Point-count bird censusing: long-term monitoring of bird abundance and diversity in central Arizona-Phoenix, ongoing since 2000.** LTER Network Member Node. <https://pasta.lternet.edu/package/metadata/eml/knb-lter-cap/46/15>.

 Peter Bergholz. 2017. **Escherichia coli isolated from surface soil in an agricultural landscape.** Knowledge Network for Biocomplexity. knb.92105.1.

 Márcia Cristina Mendes Marques, Anamaria Cequinell, Elivane Salete Capellessio, and Fernanda C. G. Cardoso. 2017. **valiação da regeneração natural do estrato arbóreo em um gradiente edáfico-sucessional em Floresta Atlântica, Antonina, Paraná, Brasil.** Programa de Pesquisa em Biodiversidade (PBBio). PBBioMA.31.3.

 Melinda Nicewonger, Kristal Verhulst, Murat Aydin, and Eric Steven Saltzman. 2017. **Ethane measurements in air extracted from Greenland and Antarctic ice cores.** Arctic Data Center. [um.uuid:748cd06a-1563-4981-a117-0257be1130eb](https://doi.org/10.5061/dryad.11858?ver=2017-11-22T09:03:35.154-05:00).

 Carlo, Michael. 2017. **Carloetal\_EcoLett\_labandfield\_rearing\_data2014and2015.** Dryad Digital Repository. <https://doi.org/10.5061/dryad.pr1h0?ver=2017-11-21T17:02:43.516-05:00>.

 Carlo, Michael A., Riddell, Eric A., Levy, Ofir, and Sears, Michael W. 2017. **Data from: Recurrent sublethal warming reduces embryonic survival, inhibits juvenile growth, and alters species distribution projections under climate change.** Dryad Digital Repository. <https://doi.org/10.5061/dryad.pr1h0?ver=2017-11-21T17:02:39.643-05:00>.

Hide Map >

3	12	34	253	143	6	6	7	2	8	66	7	13	16	3	43
1	24	342	4891	114	24	11	6	15	16	13	21	26	201	2	34
4	175	685	2738	267	6	19	2	10	5	12	63	139	63	32	2113
14	379	345	959	605	102	7	14	12	4	6	13	164	30	42	40
54	57	185	205	302	277	17	270	93	12	18	40	164	129	89	48
14	51	12	69	114	2081	1140	437	5249	1588	90	73	219	192	83	65
1	31	37	62	63	9773	5027	3862	1009	3894	11912	401	202	280	71	56
17	6	52	57	124	537	40838	12633	2951	3753	8644	508	166	194	228	201
4	35	35	53	149	247	4987	17360	11378	6787	1053	464	89	112	149	173
10	79	265	122	240	26	66	89	590	8039	1172	95	92	213	107	138
2	44	1080	719	33	13	40	117	313	352	394	1370	40	39	51	156
4	25	34	17	24	14	34	84	112	150	98	145	59	50	86	78
9	20	87	22	42	32	34	40	44	316	96	130	60	83	114	58
13	13	37	44	151	68	68	69	69	85	66	125	88	76	95	112
11	2	17	16	150	42	16	12	21	83	38	386	135	40	46	51
6	11	7	3	23	10	3	23	21	67	80	194	49	33	19	23
	95	12	10			9	8	26	24	85	20	83	44	5	22
2	7		184	13	7	13	1	8	8	32	19	15	24	22	12
4	12		11		2	2	4		3	6	10	98	14	24	8
					7			6		2	12	20	2	49	18

Google Satellite Terrain 1000 km

Map data ©2017 Google, INEGI 1000 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 0803644 and 1430268. 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 v1.14.11)

# Federated Search Across Repositories

**DataONE**

About News Participate Resources Education Data

DATONE SEARCH: Search Summary Jump to: DOI or ID Go

**Search** ?

Search phrase

**Filter by:**

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

**My Search**

**Creator: Peterson**

**Filter by:**

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

**Search results:**

- Fabes, Jez, Seligman, William, Barrett, Carolyn, McKee**  
Intensive care discharge risk score and nurse-led in units in the United Kingdom. Dryad Digital Repository. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Park, Hyeree, Cho, Sooyoung, Woo, Hyeontaek, Park, In**  
The Korean Multi-center Cancer Cohort. Dryad Digital Repository. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Jason Schatz, Carly Ziter, and Christopher Kucharik**  
Madison, Wisconsin from 2012-2016. LTER Network. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Heather Bateman, Dan Childers, Madhusudan Katti, Eyal**  
bird abundance and diversity in central Arizona-Phoenix. <https://pasta.hinet.net/package/metadata/em/knb-ter-cap4>
- Peter Bergholz**  
2017. *Escherichia coli* isolated from a... <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Márcia Cristina Mendes Marques, Anamaria Cequinel, E**  
natura do estrato arbóreo em um gradiente edáfico-a Biodiversidade (PPBio). PPBio/MA.31.3. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Melinda Nienwonger, Kristal Verhulst, Murat Aydin, and**  
Antarctic ice cores. Arctic Data Center. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Carlo, Michael**  
2017. *Carlo*. EcoLett. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>
- Carlo, Michael A., Riddell, Eric A., Levy, Ofir, and Sears**  
inhibits juvenile growth, and alters species distribution. <https://doi.org/10.5061/dryad.pr1h02?ver=2017-11-21T17:02:43.51005:00>

**Map**

Hide Map >

3 12 34 253 143 6 6 7 2 8 66 7 13 16 3 43

1 24 342 4891 114 24 11 6 15 16 13 21 26 201 2 34

4 175 685 2738 267 6 19 2 10 5 12 63 139 63 32 2113

14 379 345 959 605 102 7 14 12 4 6 13 164 30 42 40

54 57 185 205 302 277 17 270 93 12 18 40 164 129 89 48

14 51 12 69 114 2081 1140 437 5249 1588 90 73 219 192 83 65

1 31 37 62 63 9773 5027 3862 1009 3894 11912 401 202 280 71 56

17 6 52 57 124 537 40838 12633 2951 3753 8644 508 166 194 228 201

4 35 35 53 149 247 4987 17360 1378 6787 1053 464 89 112 149 173

10 79 265 122 240 26 66 89 590 8039 1172 95 92 213 107 138

2 44 1080 719 33 13 40 117 313 352 394 1370 40 39 51 156

4 25 34 17 24 14 34 84 112 150 98 145 59 50 86 78

9 20 87 22 42 32 34 40 44 316 96 130 130 60 83 114 58

13 13 37 44 151 68 68 69 69 85 66 125 88 76 95 112

11 2 17 16 150 42 16 12 21 83 38 386 135 40 46 51

6 11 7 3 23 10 3 23 21 67 80 194 49 33 19 23

95 12 10 9 8 26 24 85 20 83 44 5 22

2 7 184 13 7 13 1 8 8 32 19 15 24 22 12

4 12 11 2 2 4 3 6 10 98 14 24 8

7 6 1

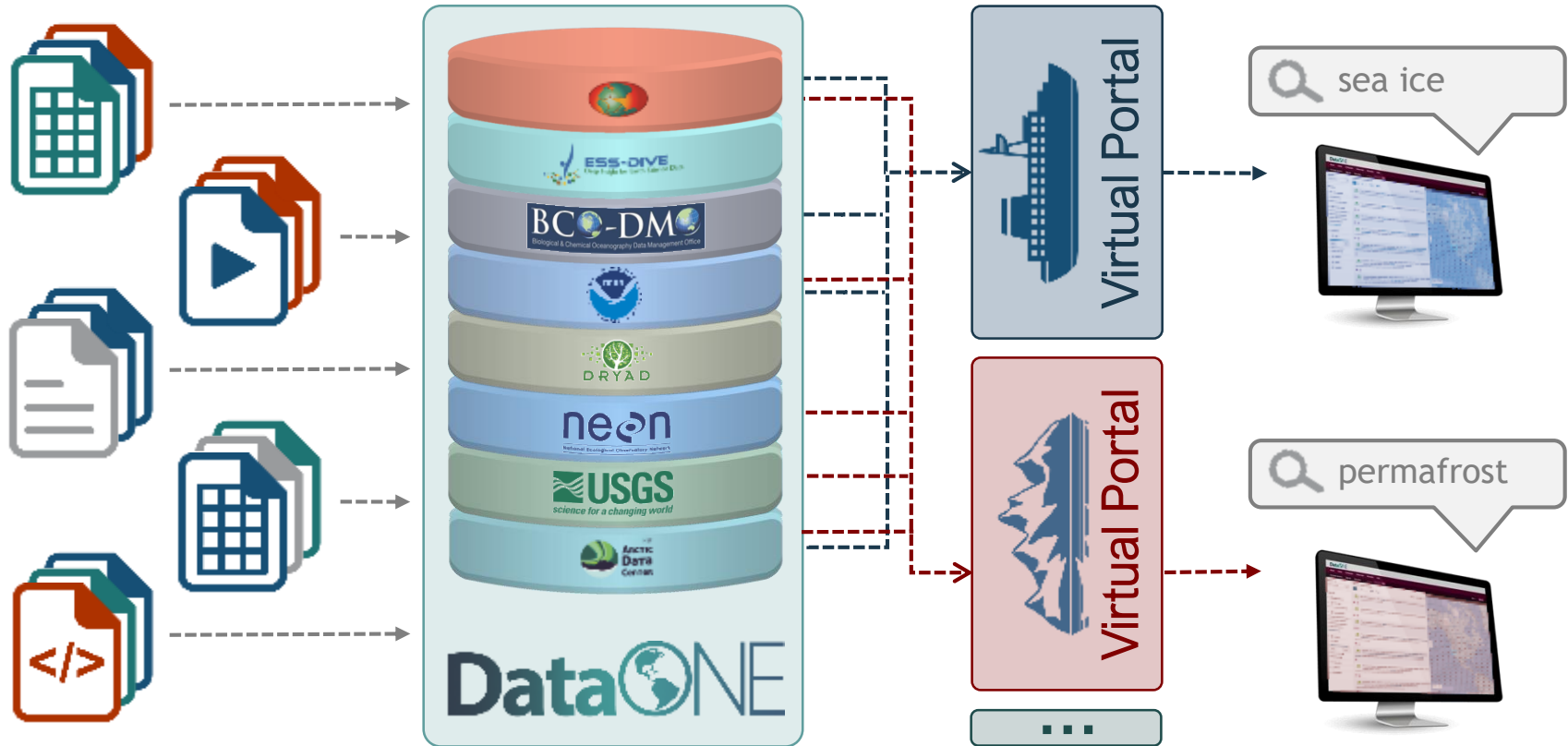
1 1

Map data ©2017 Google, INEGI 1000 km Terms of Use

Foundation (RFP) under a Cooperative Agreement. Acknowledgment: This material is based upon work supported by the  
ings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily

# 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



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

**knb** 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.

**knb** 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/F12W1J77.

**knb** 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.

**knb** 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.

**knb** 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 >

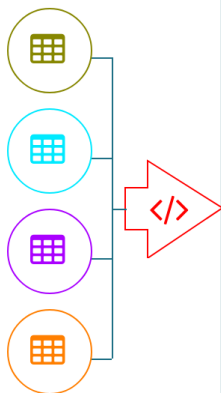
	11	39	7	
8	45	49	17	1
7	47	50	51	18
2	10	17	1	
				16

# Provenance Display

## Provenance Tracking and Management Services and Tools

### Data Table, Image, and Other Data Details

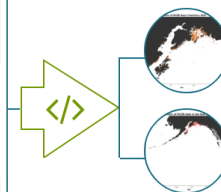
4 sources



#### Data Table

Entity Name	Total_Aromatic_Alkanes_PWS.csv						
	<a href="#">Download</a>						
Description	Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK						
Object Name	Total_Aromatic_Alkanes_PWS.csv						
Online Distribution Info	<a href="https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9">https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9</a>						
Size	2801033 byte						
Text Format	<table><tr><td>Number of Header Lines</td><td>1</td></tr><tr><td>Record Delimiter</td><td>#x0A</td></tr><tr><td>Attribute Orientation</td><td>column</td></tr></table>	Number of Header Lines	1	Record Delimiter	#x0A	Attribute Orientation	column
Number of Header Lines	1						
Record Delimiter	#x0A						
Attribute Orientation	column						

2 derivations



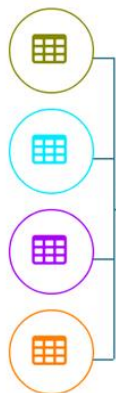


# Provenance Display

## Provenance Tracking and Management Services and Tools

### Data Table, Image, and Other Data Details

4 sources



**Source Program**

**Total\_PAH\_and\_Alkanes\_GoA\_Hydrocarbons\_Clean.R**

Citation

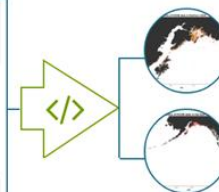
[View »](#)

This program generated the data you are currently viewing, **Total\_Aromatic\_Alkanes\_PWS.csv**.

This program used **PAH.csv**, **Sample.csv**, **Non-EVOS\_SINs.csv** and (and 1 more ).

Alkanes_PWS.csv
from PAH, Alkane and Sample tables documenting samples collected after the spill in Prince William Sound, AK
Alkanes_PWS.csv
<a href="https://doi.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9">https://doi.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9</a>

2 derivations

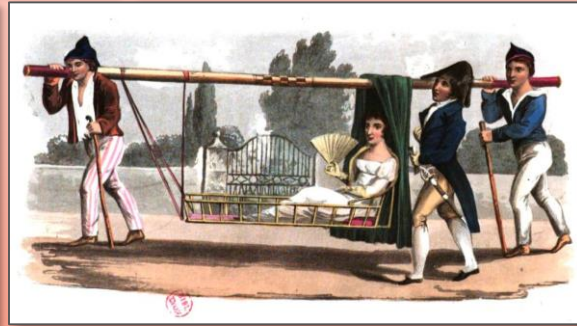
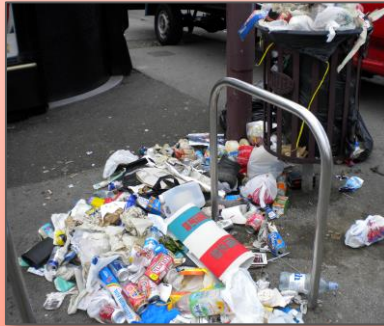


Text Format

Number of Header Lines	1
Record Delimiter	#x0A
Attribute Orientation	column

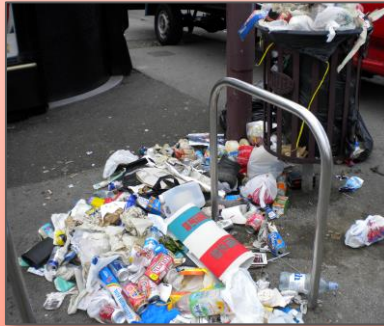
# Semantic Search

LITTER?



# Semantic Search

LITTER?



# Semantic Search

## Semantic Measurement Search Capabilities

### Attribute Information

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

Name litter

Leaf Litter Carbon Pool

+ Add tag

Label

Definition

Storage Type

Measurement Type

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



Tags

# 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.



Citations

1



Downloads

2.5K



Views

279



Copy Citation

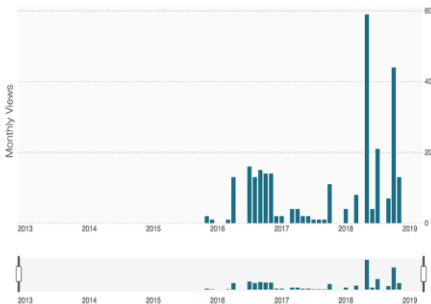
279 Views

For all versions of this data set, the number of times that all or part of this data set was viewed over time.

These view counts are **COUNTER** compliant, meaning that views from some Internet robots and repeat views within a certain time window are excluded.

Drag the slider to visualize a specific time window for the view events.

279 Views  
From Dec 2012 to Mar 2019



Downloads

Citations

1 Citations

Daisuke Hirano, Yasushi Fukamachi, Eiji Watanabe, Kay I. Ohshima, Katsushi Iwamoto, et al. 2016. **A wind-driven, hybrid latent and sensible heat coastal polynya off Barrow, Alaska.** Journal of Geophysical Research: Oceans. Vol. 121. pp. 980-997. <https://doi.org/10.1002/2015JC011318>.

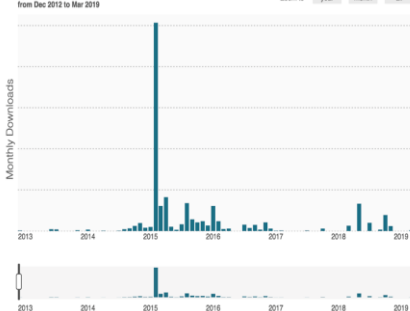
Downloads

For all versions of this data set, the number of times that all or part of this data set was downloaded over time.

These download counts are **COUNTER** compliant, meaning that downloads from some Internet robots and repeat downloads within a certain time window are excluded.

Drag the slider to visualize a specific time window for the download events.

2,489 Downloads  
From Dec 2012 to Mar 2019



Citations

Views

MAKE  
DATA  
COUNT  
  
COUNTER  
CONSISTENT CREDIBLE COMPARABLE

# 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 ▾

**Identification:** 88% complete



**Discovery:** 100% complete



**Interpretation:** 100% complete



▶ **Passed 14 checks out of 20 (informational checks not included).**

▶ **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

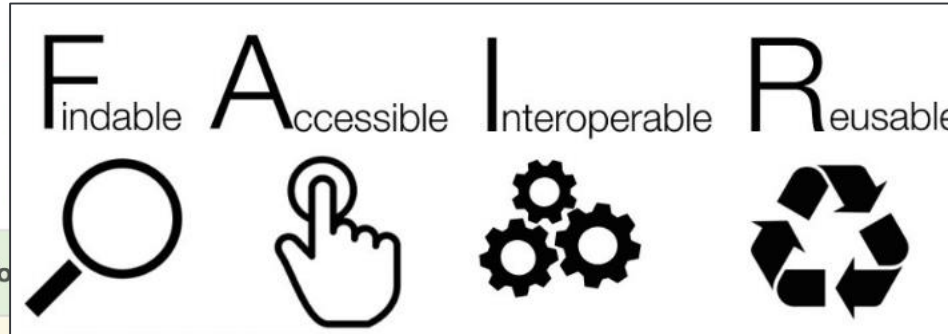
FAILURE

# 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

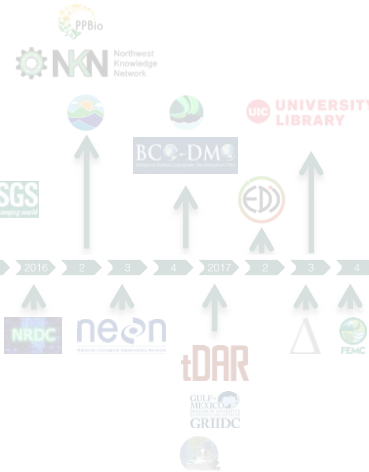
REQUIRED

FAILURE



# Three Primary Goals

Building an  
Empowered  
and Engaged  
Community



Enabling  
Reproducible  
Science through  
Tools and Services

Developing  
sustainable data  
discovery and  
interoperability  
solutions

## Data Table, Image, and Other Data Details

4 sources

Alkane.csv

Citation

View >

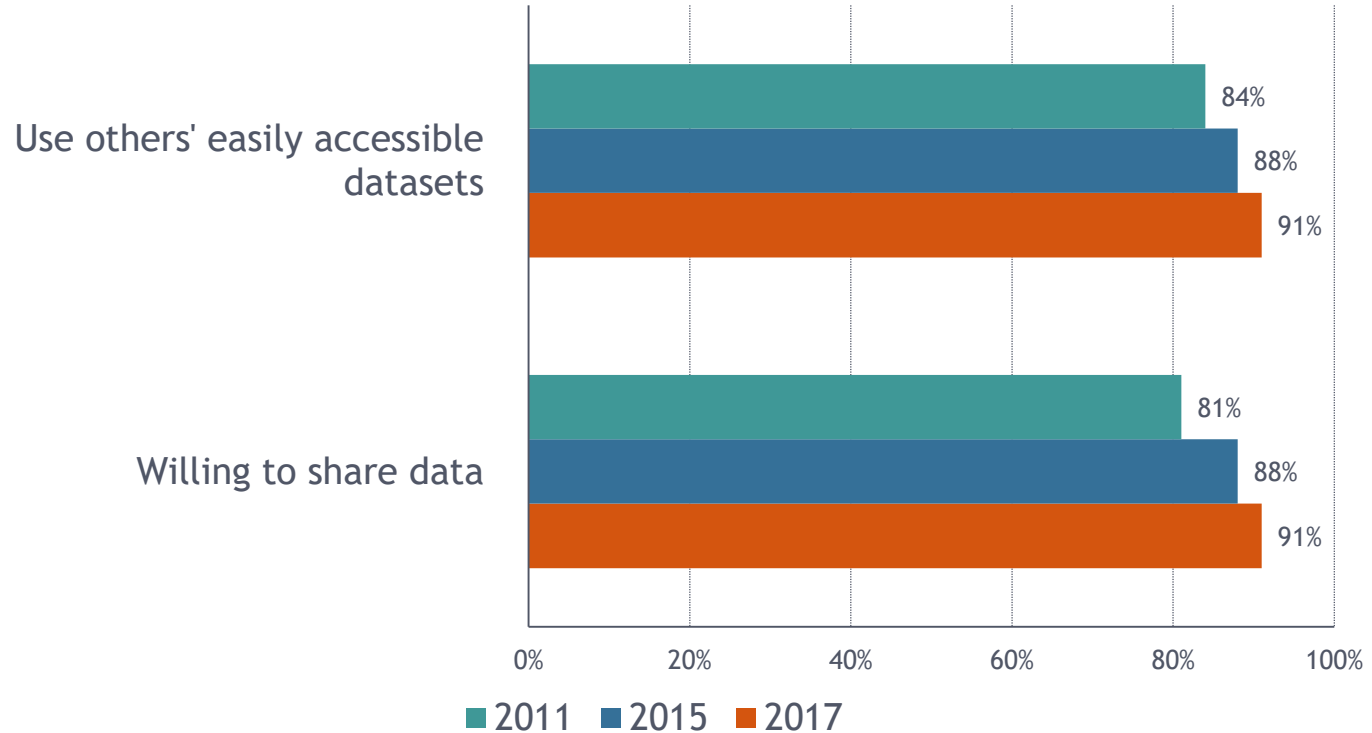
This data was generated by `</>`  
DataDownload.R.

This data prov\_hasDerivations the data  
you are currently viewing, `</>`  
Total\_Aromatic\_Alkanes\_PWS.csv and  
the data you are currently viewing, `</>`  
Total\_Aromatic\_Alkanes\_PWS.csv.

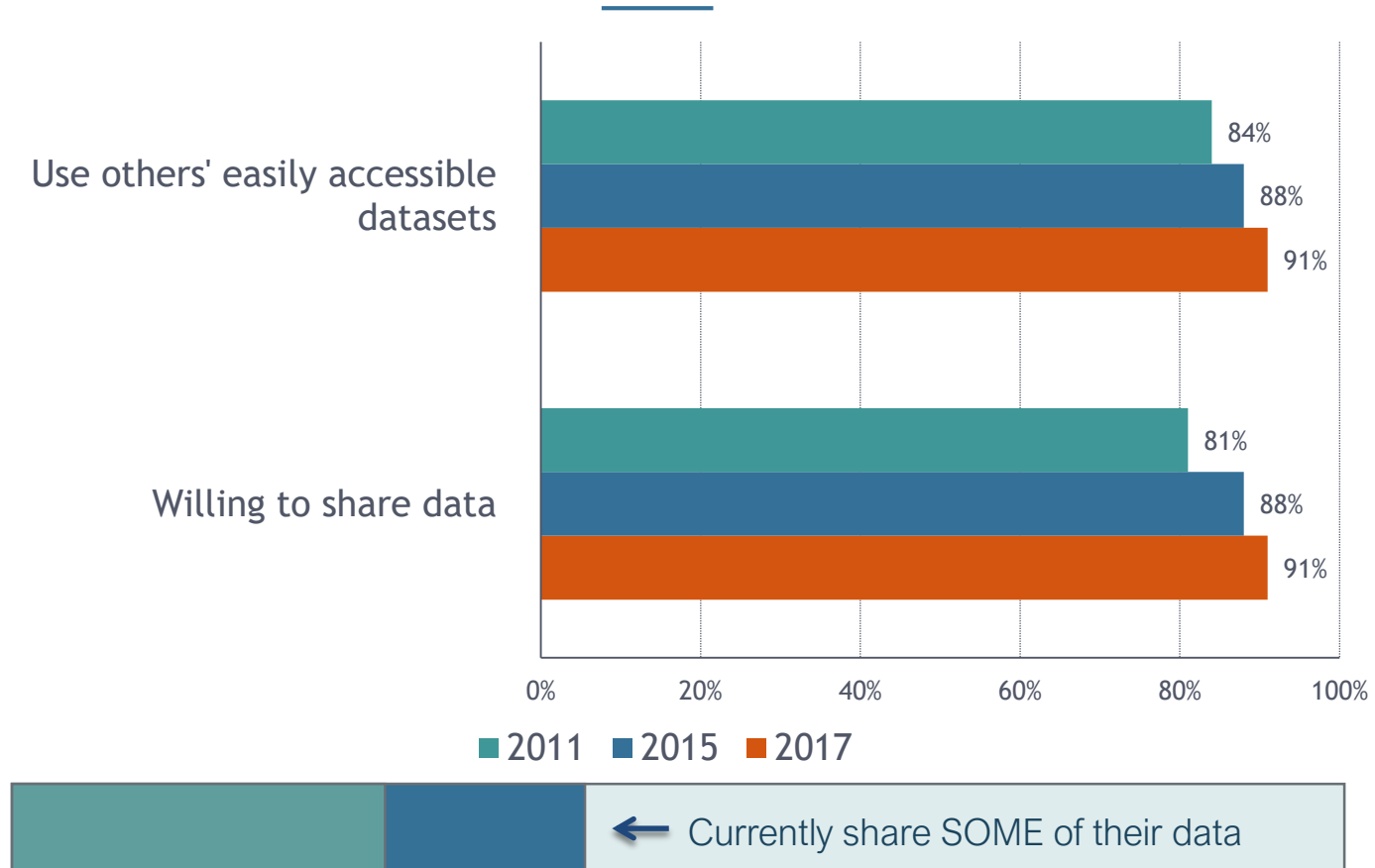
This data was used by `</>`  
Total\_PAH\_and\_Alkanes\_GoA\_Hydroca  
rbons\_Clean.R.



# Culture of Data Sharing is Improving

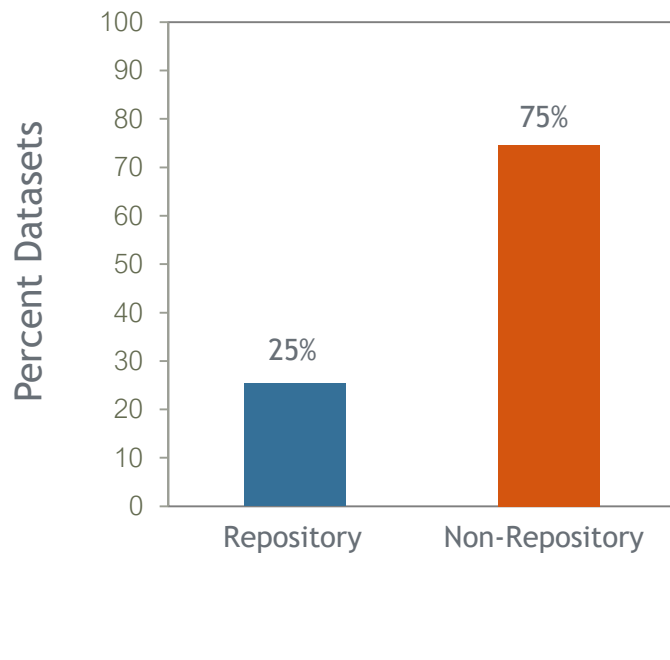


# Culture of Data Sharing is Improving



# Culture of Data Sharing is Improving

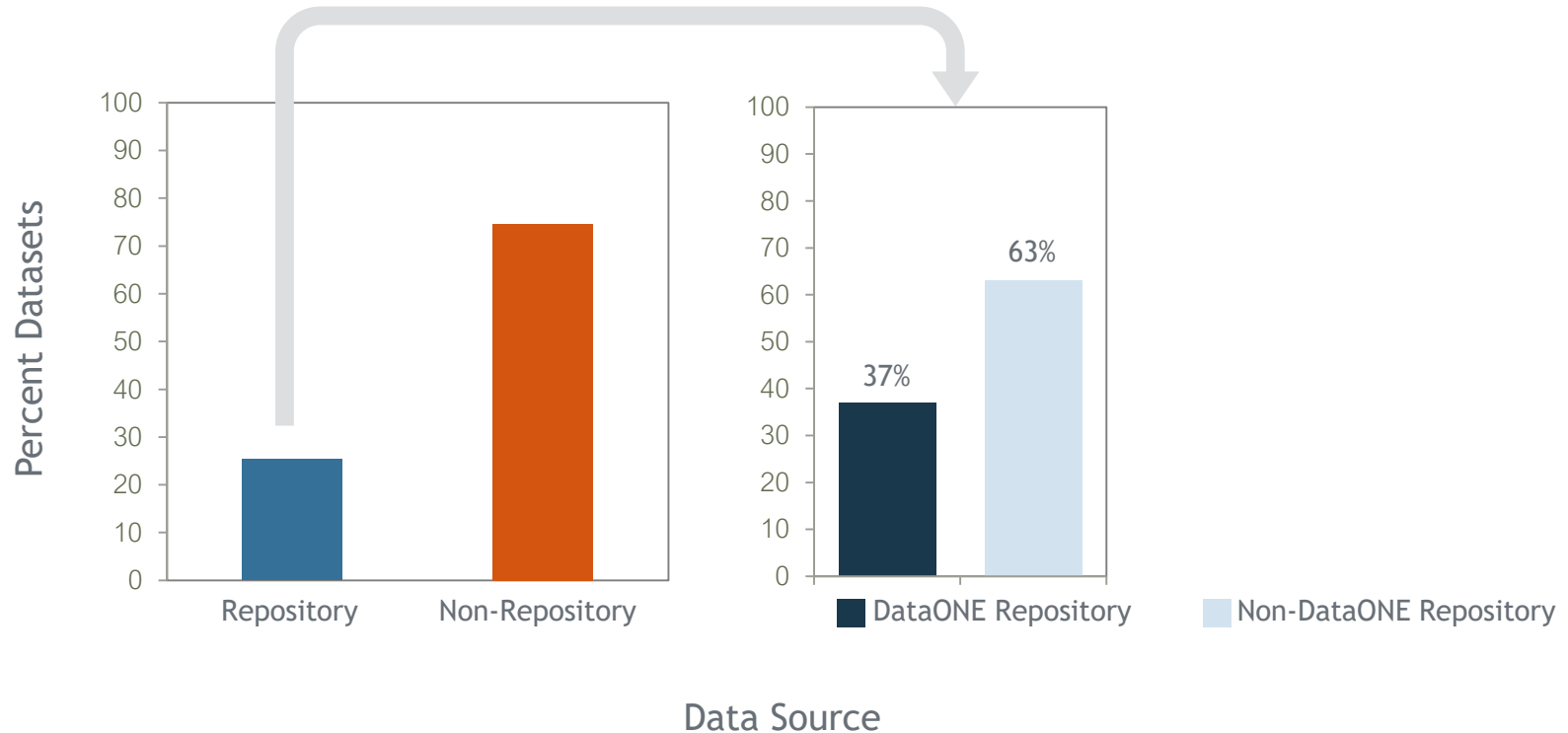
But there is still opportunity for growth



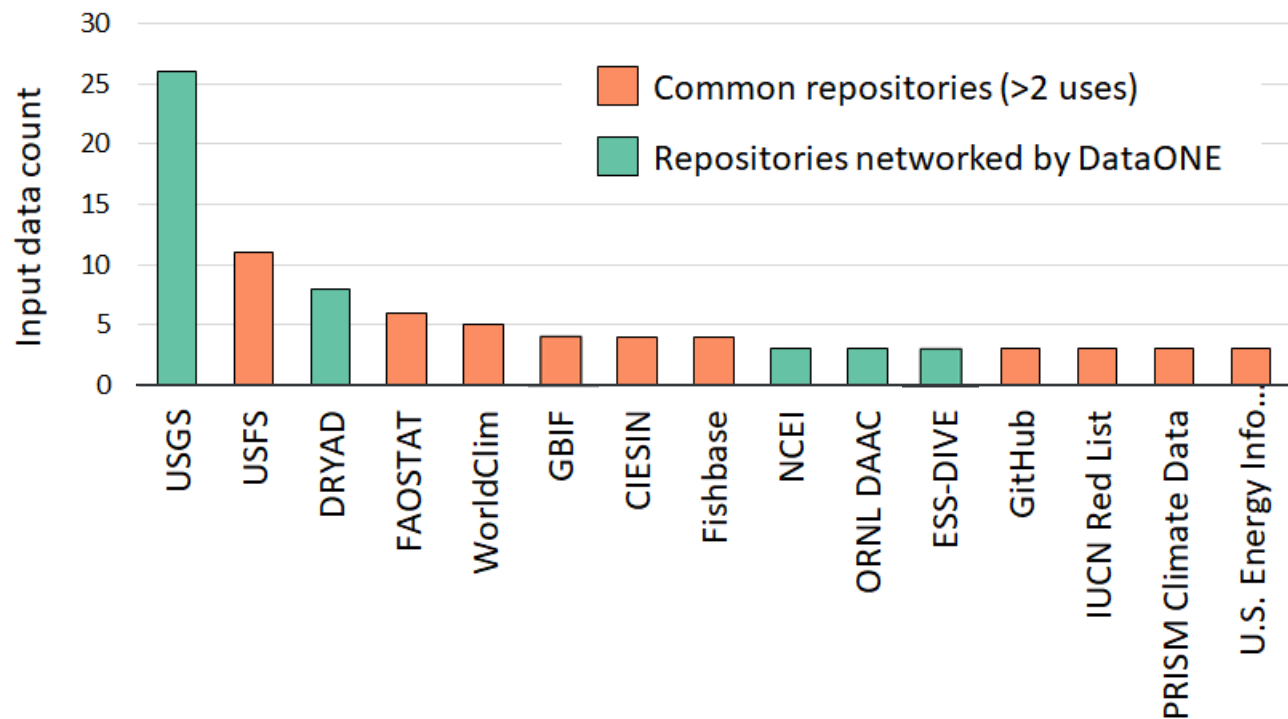
- Synthesis research
- 80 publications
- 505 datasets
- Average 6.4 reused datasets per study

# Culture of Data Sharing is Improving

But there is still opportunity for growth



# Informing DataONE Design



# Building and Engaging Community

## Community Informed Design

Surveys

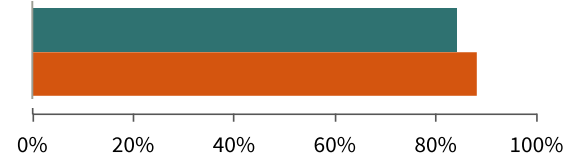
User scenarios

Personas

Usability assessments

### Perception

Use others' datasets if their data were easily accessible

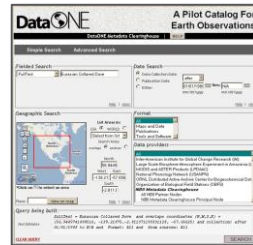


### Satisfaction

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



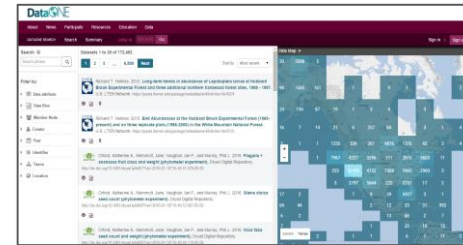
Satisfaction → 2010 2014



2009



2012



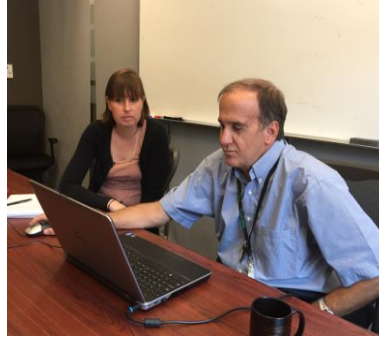
2016



# DataONE Community

## Annual Community Meeting

Inform  
development  
prioritization and  
direction of  
DataONE



User testing  
and feedback



Community  
challenges and  
shared solutions



# Training

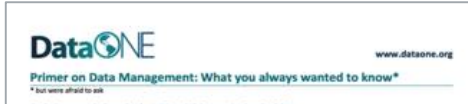
## Leadership in Data Management Education

45

Webinars

~4100

Webinar  
Attendees



Carly Strasser, Robert Cook, William Michener

Contents

1. Objective of This Primer

2. Why Manage Data?

2.1. It will benefit you and your colleagues

2.2. It will benefit the scientific community

2.3. Journals and sponsors want you to

3. How To Use This Primer

4. The Data Life Cycle: An Overview

5. Data Management Throughout the Data Life

5.1 Plan

5.2 Collect

5.3 Assess

5.4 Describe: Data Documentation

5.5 Preserve

5.6 Discover, Integrate, and Analyze

6. Conclusion

7. Acknowledgements

8. References

9. Glossary

1. Objective of This Primer

The goal of data management is to produce self-colleagues who has not been involved with your able to use it effectively and properly? This primer practices that will enable you to develop a data organizer, manage, describe, preserve and share

2. Why Manage Data?

2.1. It will benefit you and your colleagues

Establishing how you will collect, document, organize your research project has many benefits. You will research by investing the time and energy before easier for you to find, use, and analyze, and it will your data. In the long term, following good data with the project can find, understand, and use it recommending appropriate ways to cite your data and their use [1].

DataONE Best Practices Primer



### 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 final product can be complex.

#### Reproducibility

...is at the core of the scientific process. If results are not reproducible, they lose credibility. Good documentation of the data and the analysis are essential

#### Workflows

**Definition:** Precise description of the procedures used in a project. Can be formal or informal.

#### Informal workflow

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

- Inputs and outputs
- Transformation rules or analytical processes
- Decision points
- Arrows indicating direction of process flow



#### Informal Workflow Example

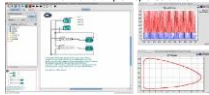


#### Formal Workflow

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
- Allows sharing and reuse of individual steps or overall workflow

#### Formal workflow example: Kepler software



Greg Wilson



Liz Ferguson



Stephanie Hampton



Fernando Pérez



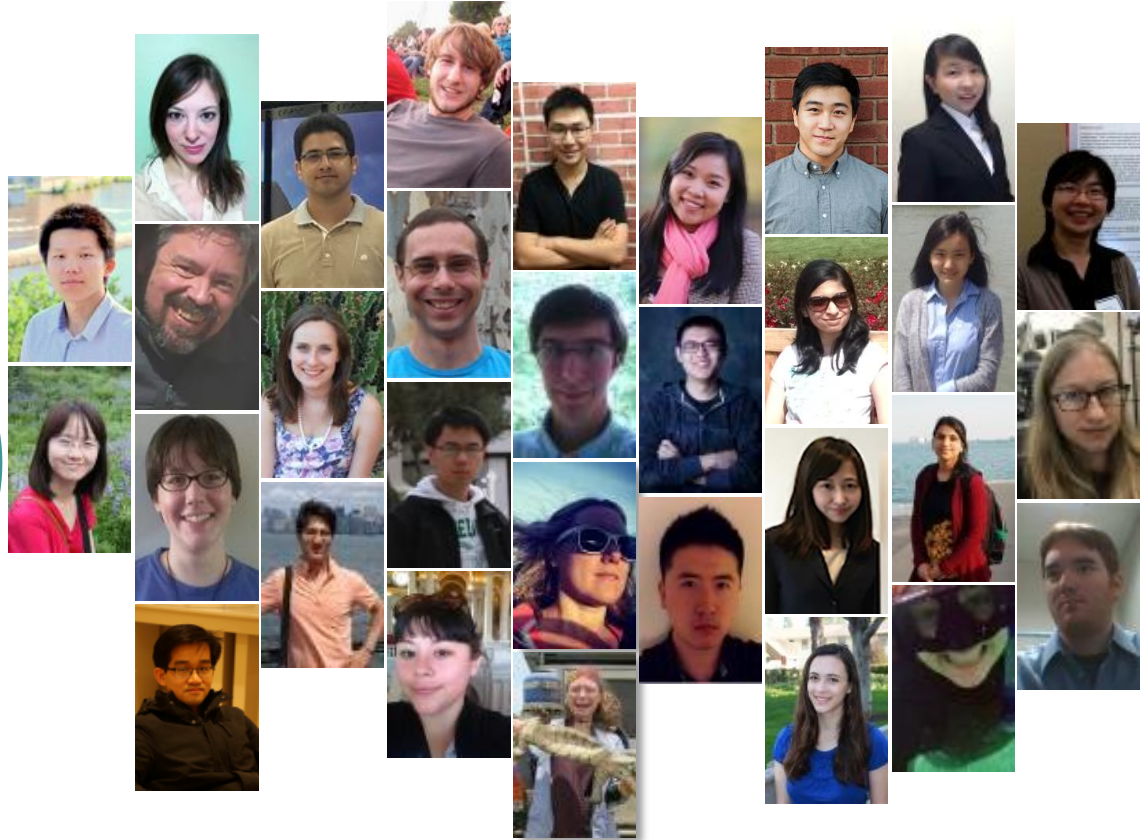
# Broadening Participation

## DataONE Summer Internship Program

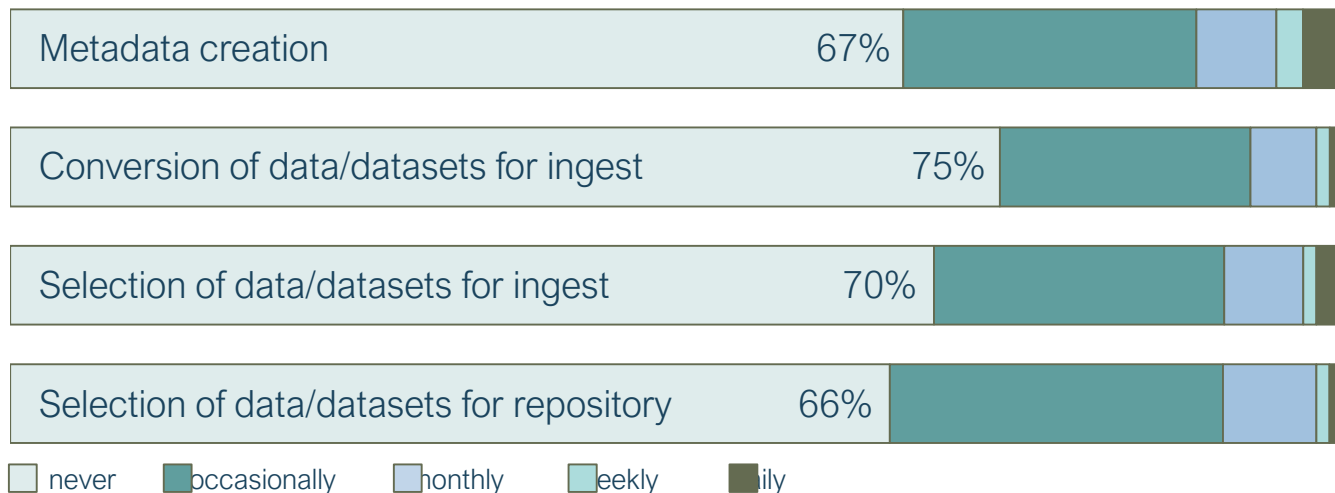
- 2009 - 4 interns
- 2010 - 4 interns
- 2011 - 8 interns
- 2012 - 6 interns
- 2013 - 8 interns
- 2014 - 10 interns
- 2015 - 4 interns
- 2016 - 5 interns
- 2017 - 6 interns
- 2018 - 4 interns
- 2019 - 5 interns

64  
Interns  
Mentored

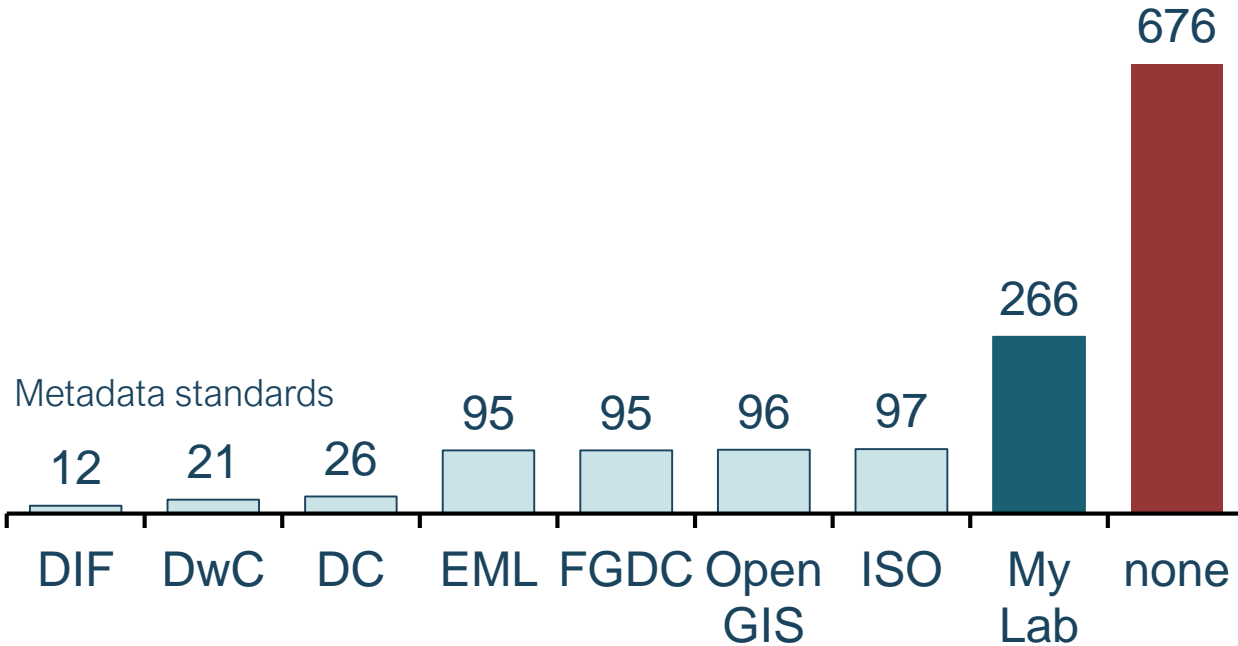
\*2 interns sponsored by other organizations



# Demonstrated Need for Training

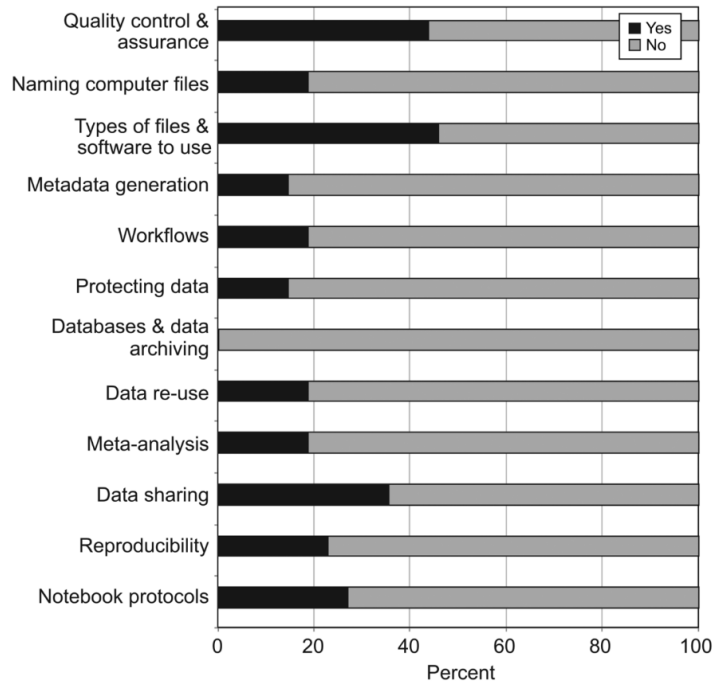


# Demonstrated Need for Training



# Status of Data Management Education

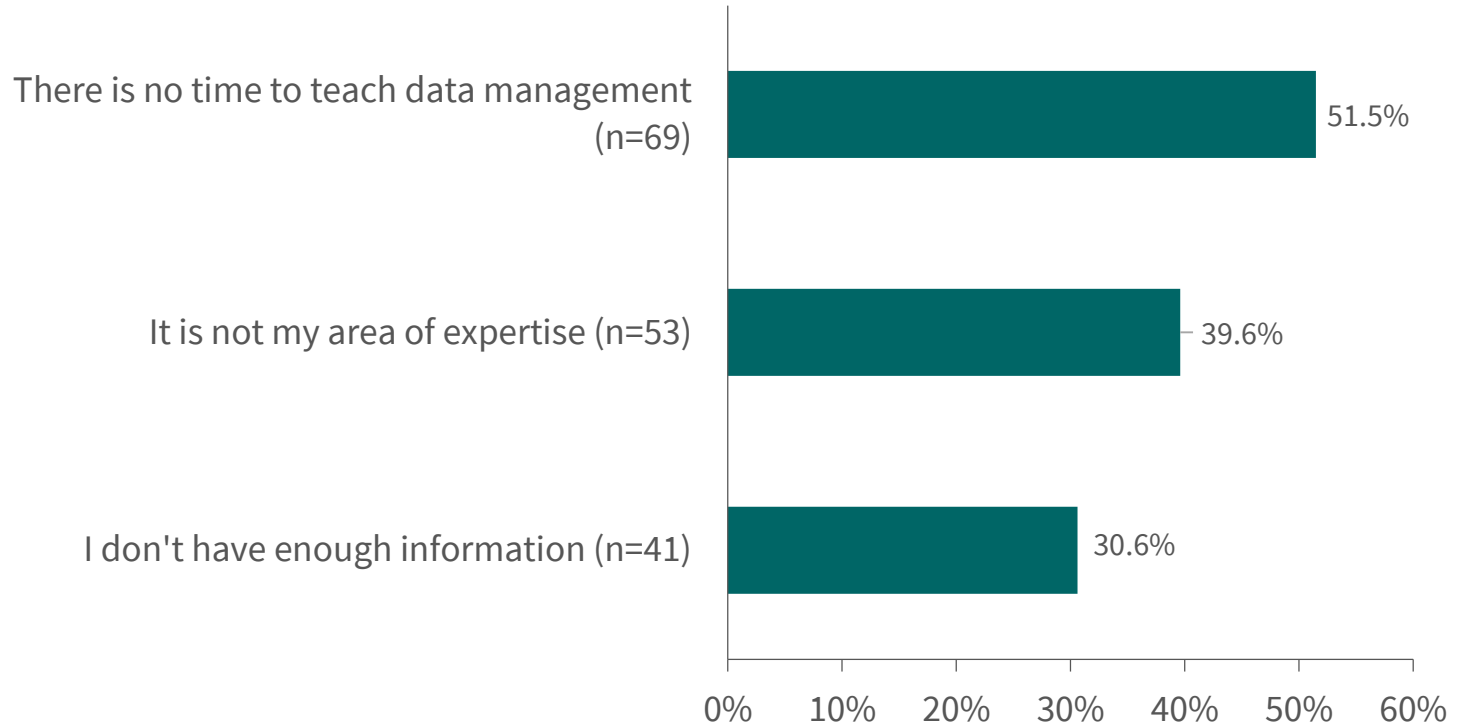
## Survey of Ecology Courses



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


# Challenges in Data Management Training

## Survey of Educators



# Education Materials

## The Evolution of Resources



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

\* but were afraid to ask

Carly Strasser, Robert Cook, William Michener

Contents

1. Objective of This Primer
2. Why Manage Data?
- 2.1. It will benefit you and your colleagues
- 2.2. It will benefit the scientific community
- 2.3. Journals and sponsors want you to
3. How To Use This Primer
4. The Data Life Cycle: An Overview
5. Data Management Throughout the Data Life Cycle
- 5.1. Plan
- 5.2. Collect
- 5.3. Assess
- 5.4. Describe: Data Documentation
- 5.5. Preserve
- 5.6. Discover, Integrate, and Analyze
6. Conclusion
7. Acknowledgements
8. References
9. Glossary

1. Objective of This Primer


The goal of data management is to produce self-colleagues who has not been involved with your able to use it effectively and properly? This primer practices that will enable you to develop a data organization, manage, describe, preserve and share

2. Why Manage Data?

2.1. It will benefit you and your colleagues

Establishing how you will collect, document, organize your research project has many benefits. You will research by investing the time and energy before easier for you to find, use, and analyze, and it will your data. In the long term, following good data with the project can find, understand, and use it and their use [1].

DataONE Best Practices Primer



Lesson 10: Analysis and Workflows

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 final product can be complex.

Reproducibility...

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

Good documentation of the data and the analysis are essential

Workflows

Definition: Precise description of the procedures used in a project. Can be formal or informal.

Informal workflow

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

- Inputs and outputs
- Transformation rules or analytical processes
- Decision points
- Arrows indicating direction of process flow

Informal Workflow Example

Formal workflow example: Kepler software

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
- Allows sharing and reuse of individual steps or overall workflow


Formal workflow example: Kepler software

Tutorials on Data Management

Lesson 5: Data Quality Control and Assurance

Data Quality Control and Assurance

DataONE Education Module 06: Data Quality, Control and Assurance 624 views



Data Management Skillbuilding Hub

Home Contribute FAQ GitHub

The Data Management Skillbuilding Hub contains resources for better data management and is open to community input and update. These resources are adaptable across a range of contexts and intended for use by researchers, teachers, librarians, or anyone who wants to learn better data management practices. Each tile below links to community contributed education materials, such as best practices and lesson plans

The resources presented on the Data Management Skillbuilding Hub can be updated by users to promote a current, well-maintained, and sustainable educational tool. Learn more about how you can [contribute](#).

Using This Resource

Click individual tiles to learn more and use each resource. You can limit resources by content type and Data Life Cycle stage. Comprehensive information is available in the [FAQ](#).

Filter by content type:

Filter by stage of the Data Life Cycle:

Best Practices Analyze	Lessons Analyze	Best Practices Assure	Lessons Assure
Best Practices Collect	Lessons Collect	Best Practices Describe	Lessons Describe
Videos Describe	Best Practices Discover	Lessons Discover	Videos Discover
Best Practices Integrate	Lessons Integrate	Best Practices Plan	Lessons Plan



# 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





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.

ss3sim / ss3sim

Code Issues 13 Pull requests 0 Projects 0 Wiki Insights

An R package for stock-assessment simulation with Stock Synthesis

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

kelljohnson Fix example in sample\_icomp Latest commit 2304530 on Jul 6

R Fix example in sample\_icomp 5 months ago

benchmarks Add microbenchmark tests of parallel options 3 years ago



# DataONE: An Interoperable Federation

---



**Global**  
Data Coverage



**820K**  
Data Packages



**45**  
Webinars



**10**  
Education Modules



**44**  
Member Nodes



**100K**  
Contributors

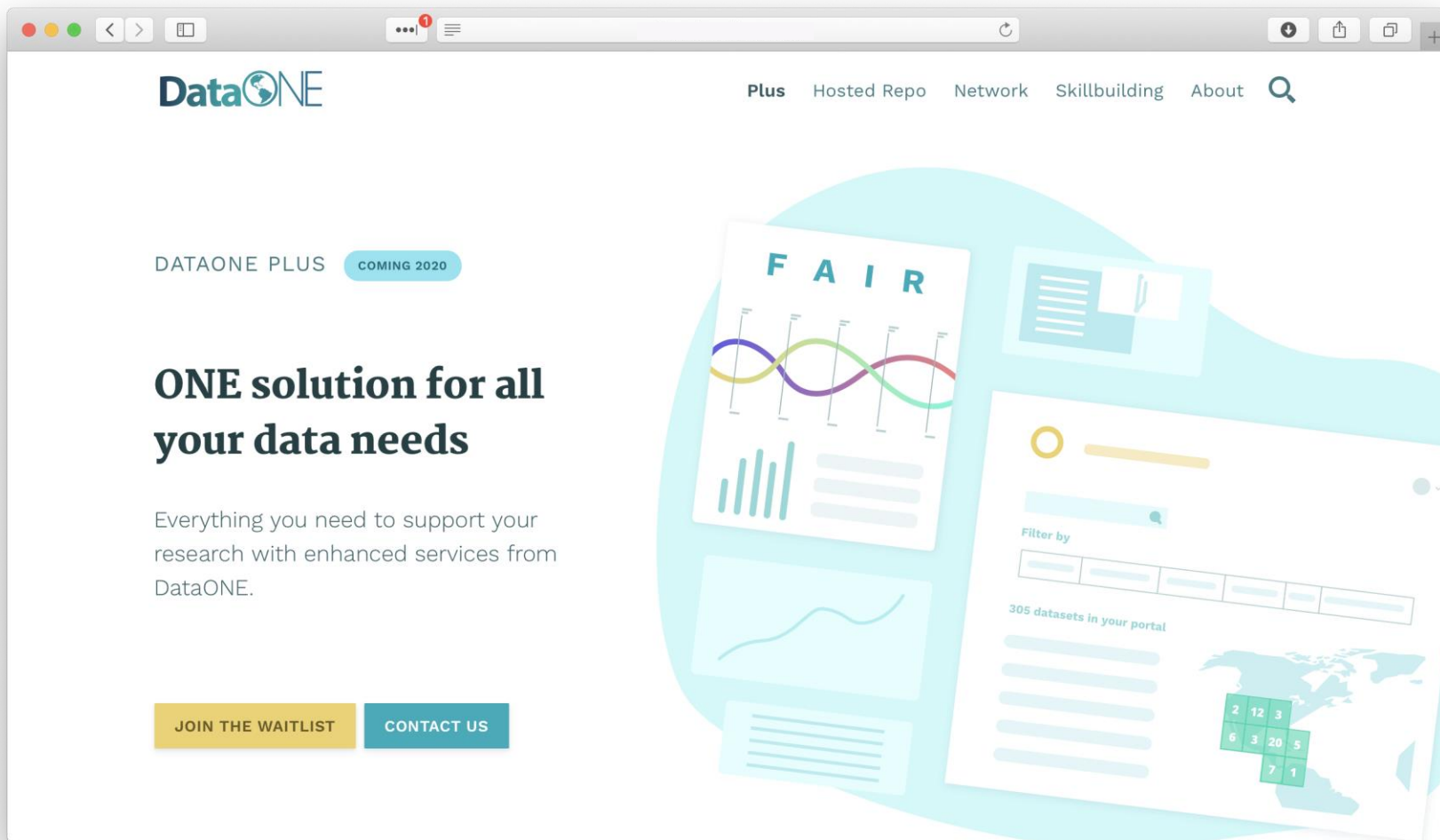


**20K**  
Users/Month



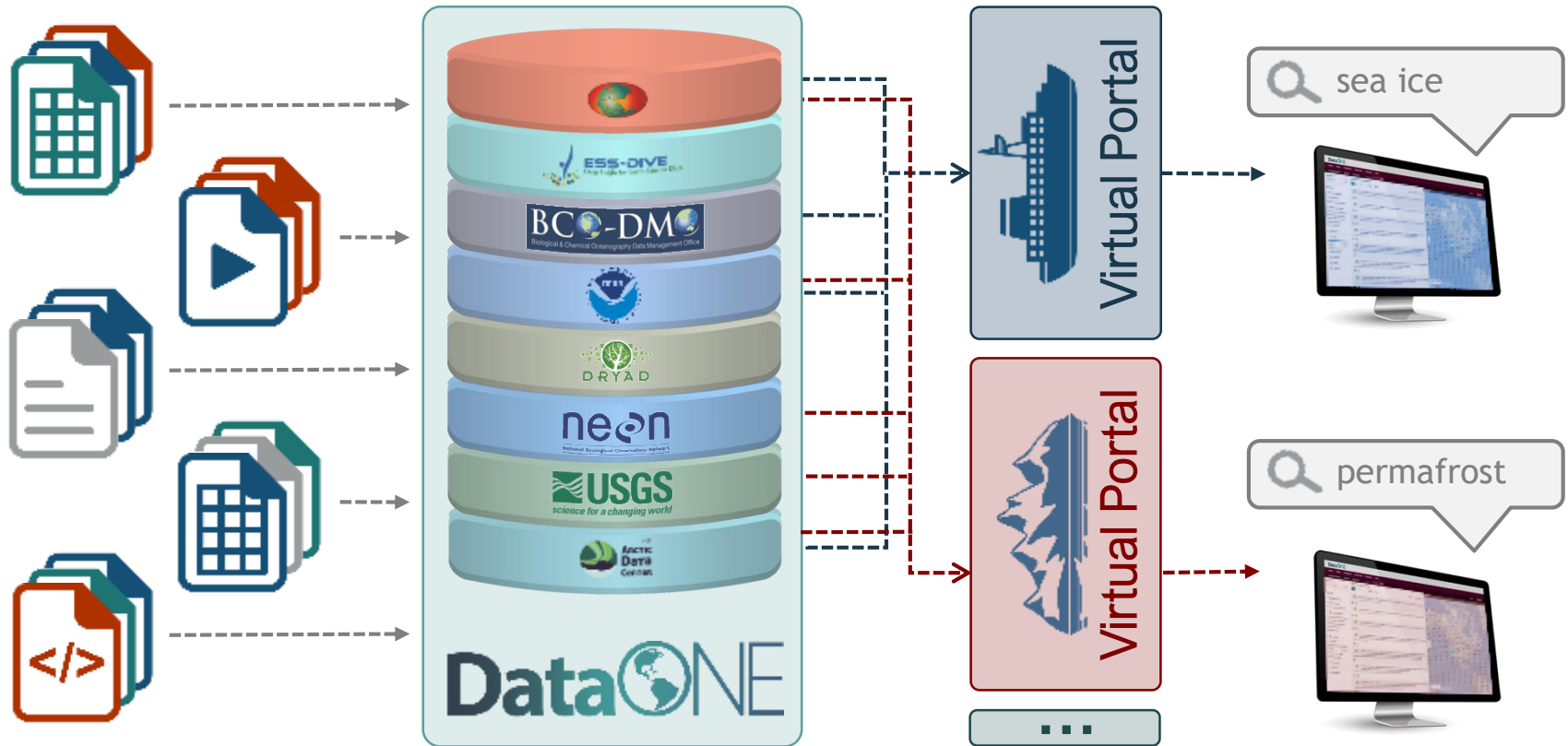
**5700+**  
Trained

# DataONE Plus



# Data Portals

Data Aggregated from Repositories Across DataONE



Hosted by the knb

Sign In

SASAPState of Alaska's Salmon and People

AboutRegionsTopicsDataMetricsMembers

Search

CURRENT SEARCH x CLEAR ALL

Search these datasets

SASAP Region: Copper River x

Geography

Species

Theme

Featured

People and Organizations

Year

Identifier

Species

Choose a Species

DATASETS 1 TO 25 OF 57

Sort byMost 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/F12W1J77.

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.

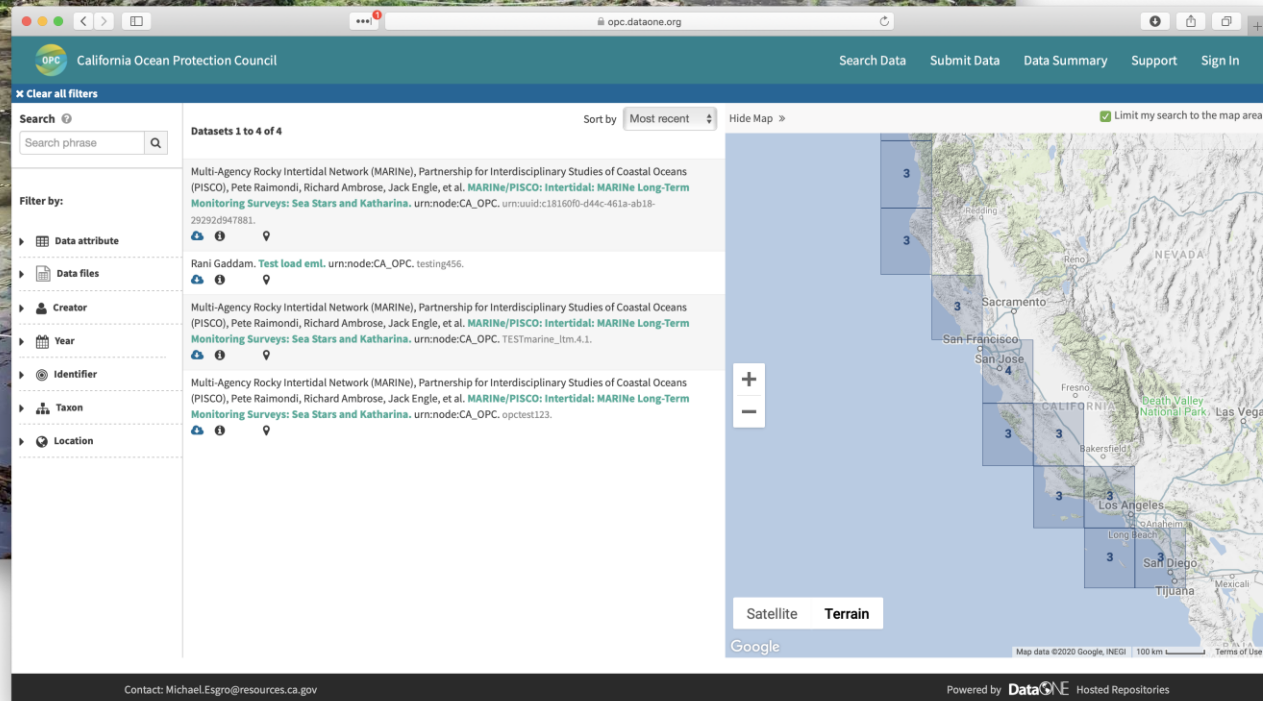
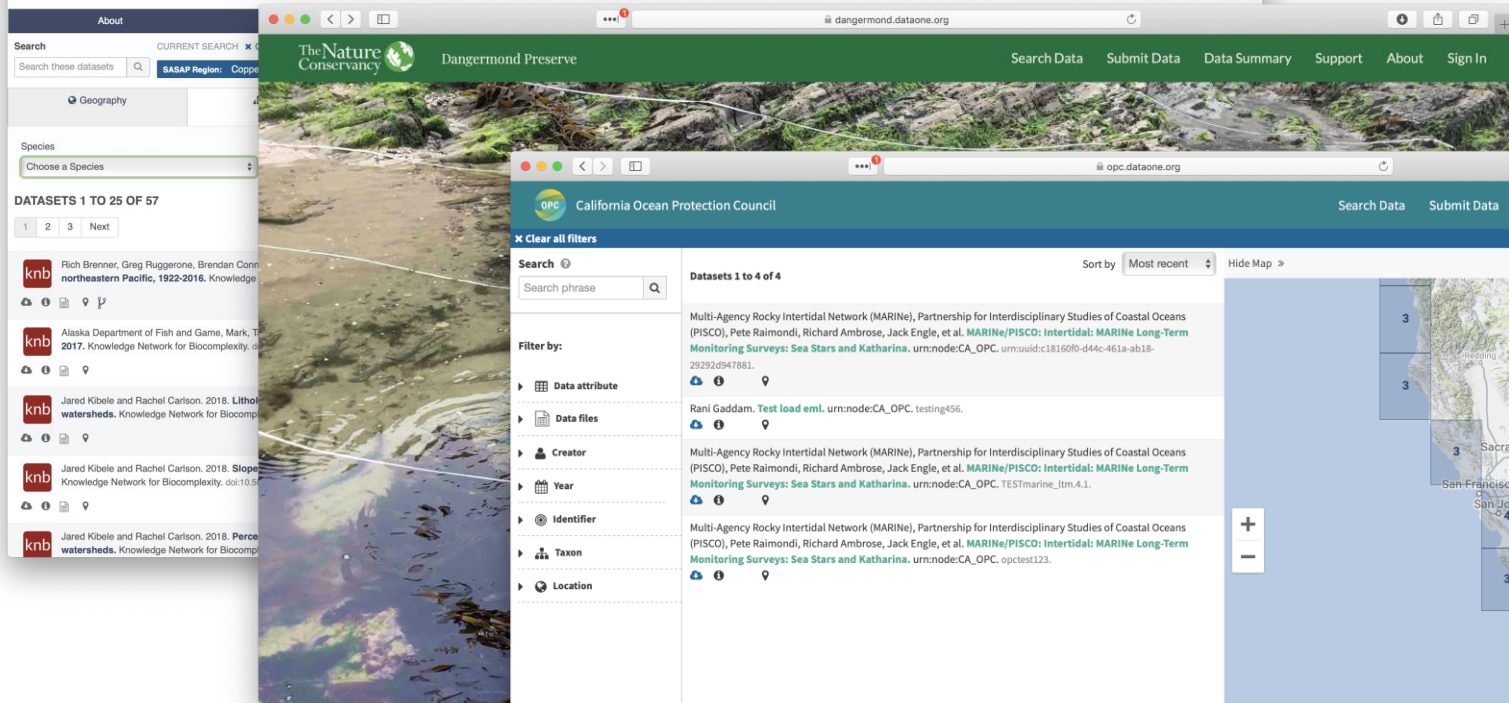
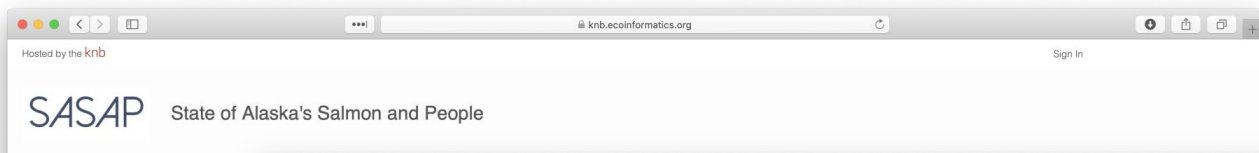
Jared Kibele and Rachel Carlson. 2018. Slope per SASAP region and Hydrologic Unit (HUC8) boundary for Alaskan watersheds. Knowledge Network for Biocomplexity. doi:10.5063/F1129QZK.

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.

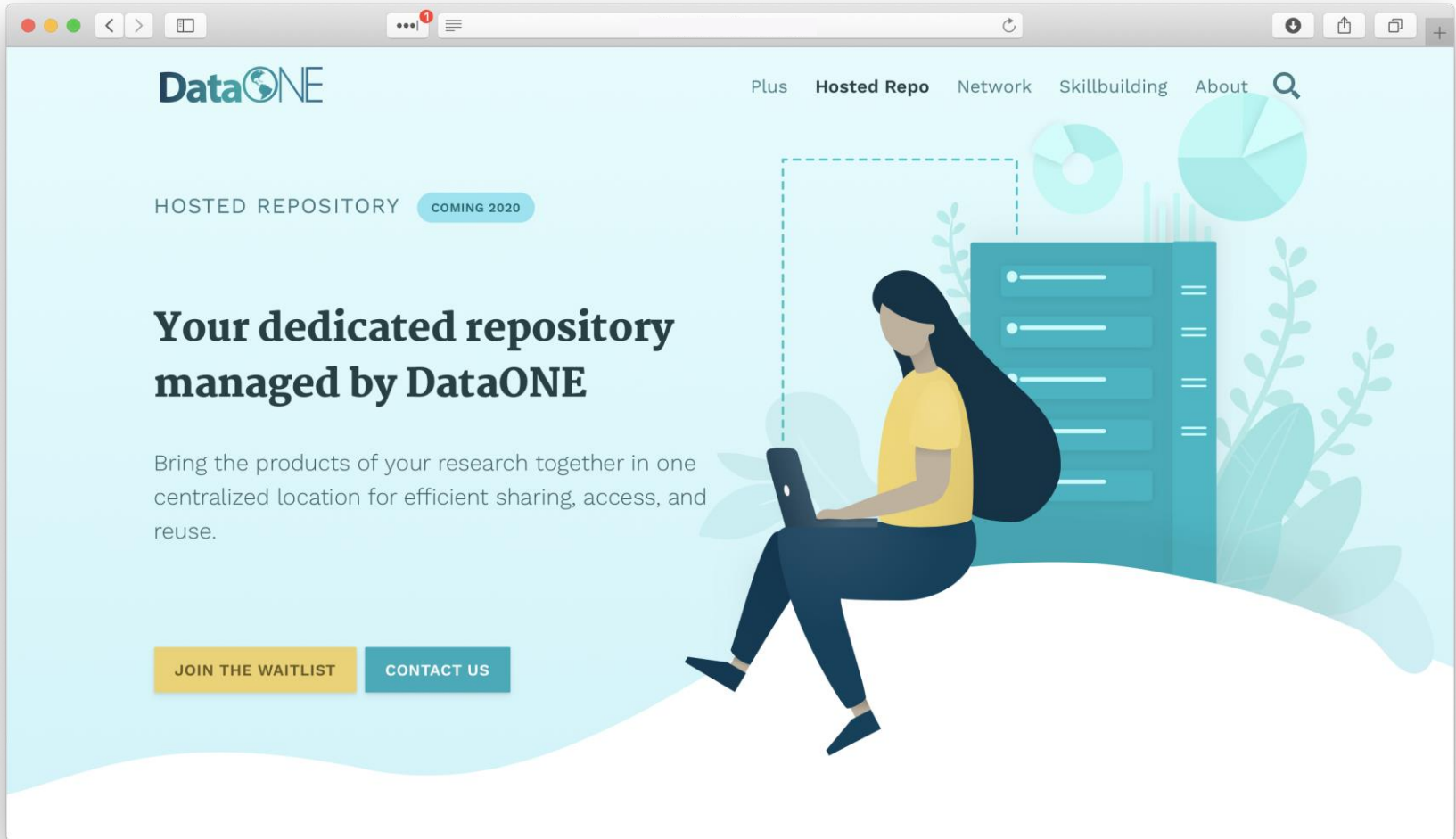
Hide Map

60





# 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 21<sup>st</sup> 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**