# Semantic Challenges and Opportunities in DataONE

Presented by

**Line C. Pouchard**

DataONE Integration and Semantics Working Group co-founder

Scientific Data Group

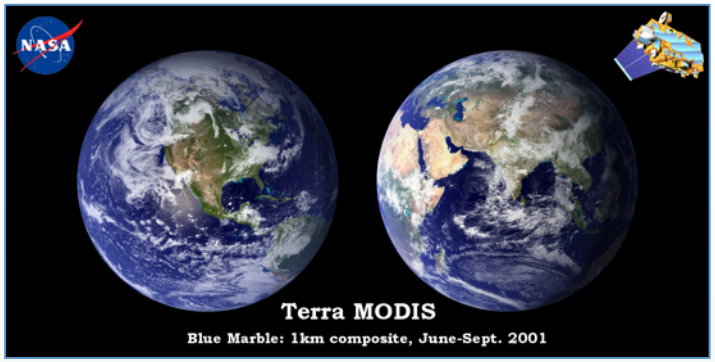
Oak Ridge National Laboratory

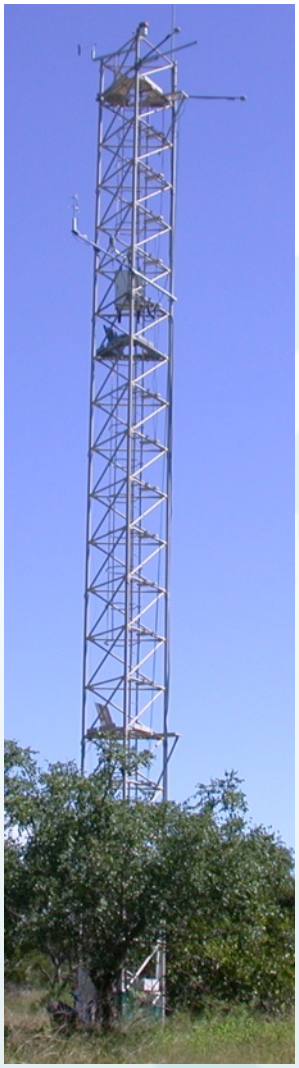
DataONE

Linked Science Workshop Keynote, Boston, November 12, 2012

# Data deluge

Sensors, sensor networks, and
remote sensing gather observations;
Data management and stewardship



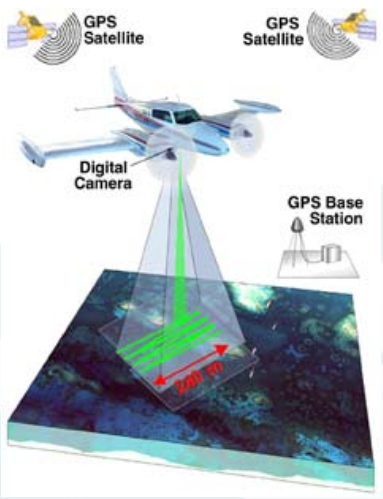
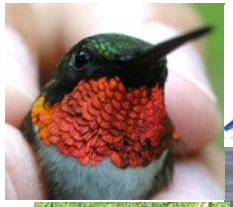
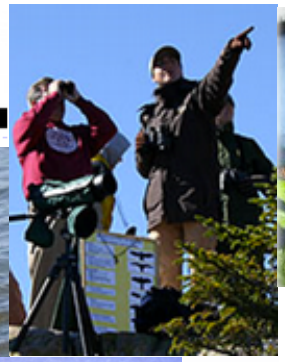
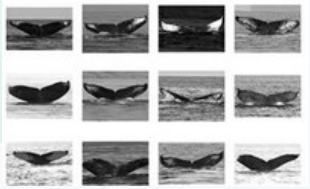


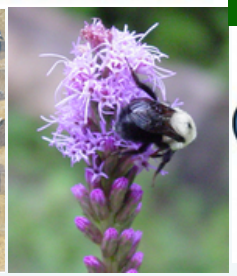
Photo courtesy of www.carboafrica.net

# Proliferation of citizen science









# DataONE is Foremost a Federation

## Member Nodes (MNs)

- Heart of the federation
- Harness the power of local curation

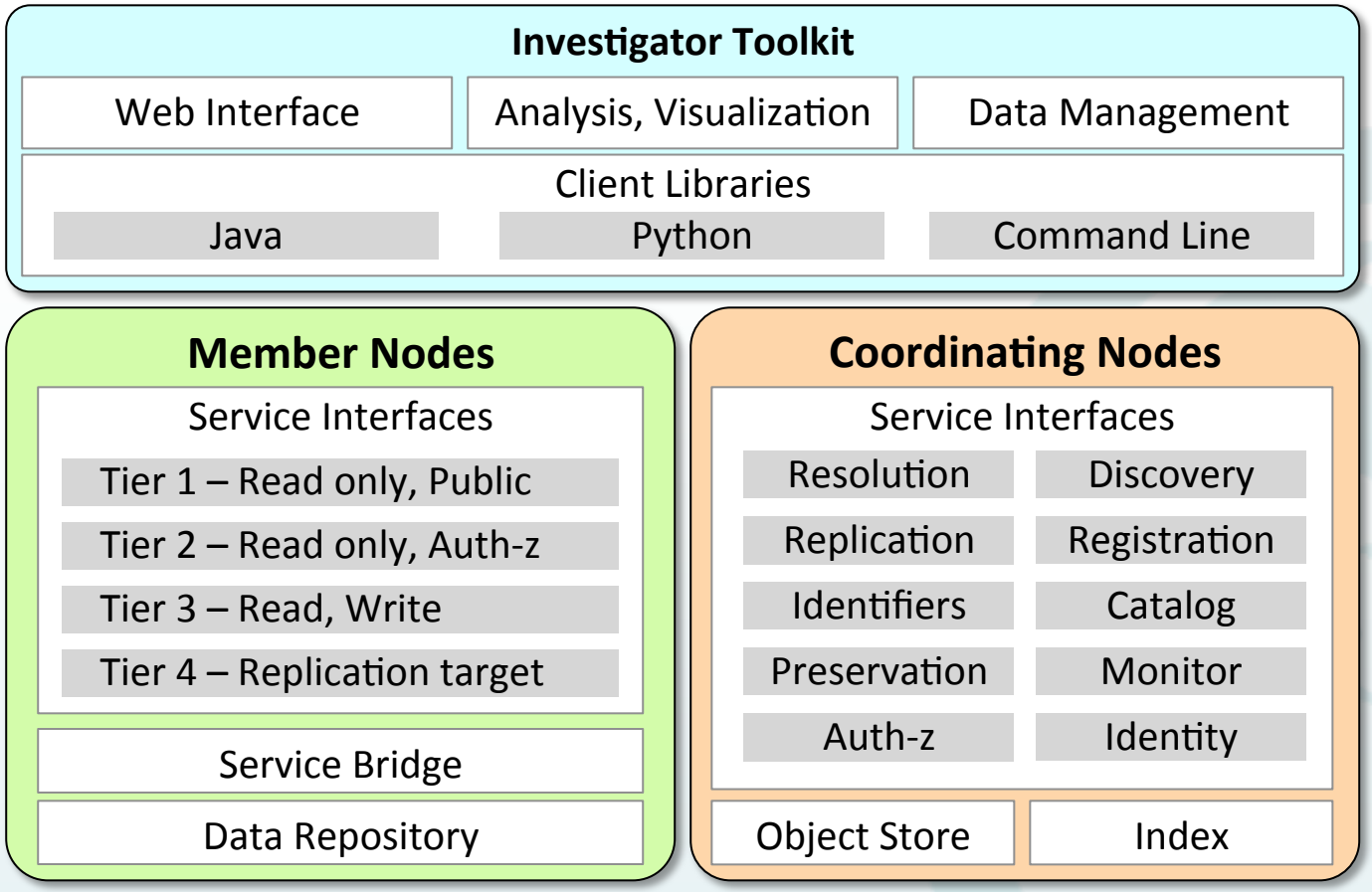
## Coordinating Nodes (CNs)

- Services to link Member Nodes

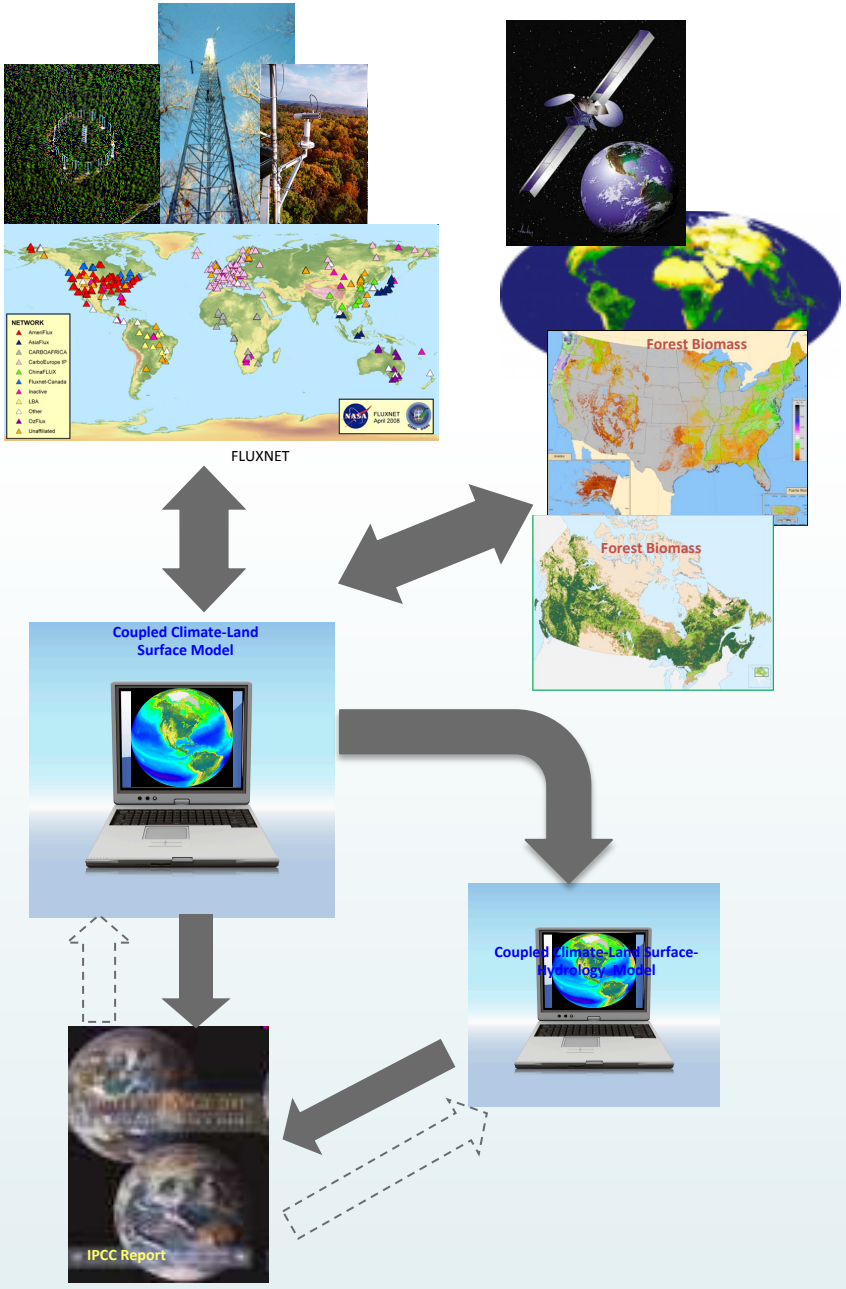
## Investigator Toolkit (ITK)

- Tools for the whole data lifecycle

# Three Major Components



# DataONE as Linked Science




- Discovery and access from heterogeneous sources
  - Simulations, models, experiments, remote sensing, GIS, molecular and omics databases, publications
- Metadata and semantics integration
- Workflows, scenario development, data and process re-use, provenance
- Engaging communities of scientists, educators, librarians, developers, volunteers
- Relies upon cyber-infrastructure promoting open source
- Complex systems of systems, networks of projects, repositories, archives, publishers

# Deployed Infrastructure

## Coordinating Nodes

- ORC: ORNL + UTK
- UCSB
- UNM

## Investigator Toolkit

- Java and Python libs
- Command line
- ONEMercury
- (R plugin)
- (Morpho)
- (ONEDrive)

## Member Nodes

### Production

- KNB
- ESA
- SANParks
- USGS CSAS
- ORNL DAAC
- LTER
- CDL
- PISCO

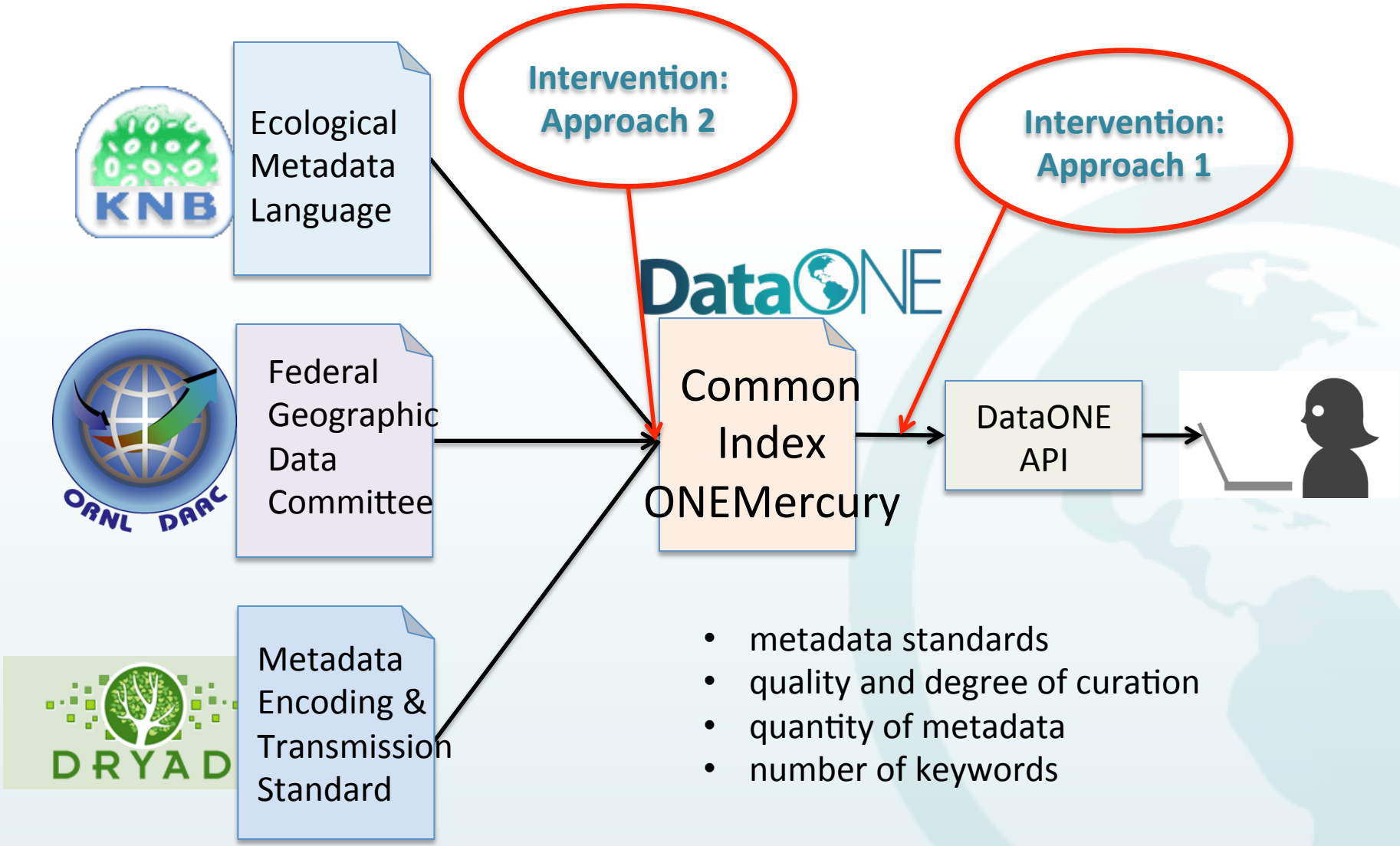
### Testing

- AKN
- TFRI
- Dryad
- EDAC
- KUBI

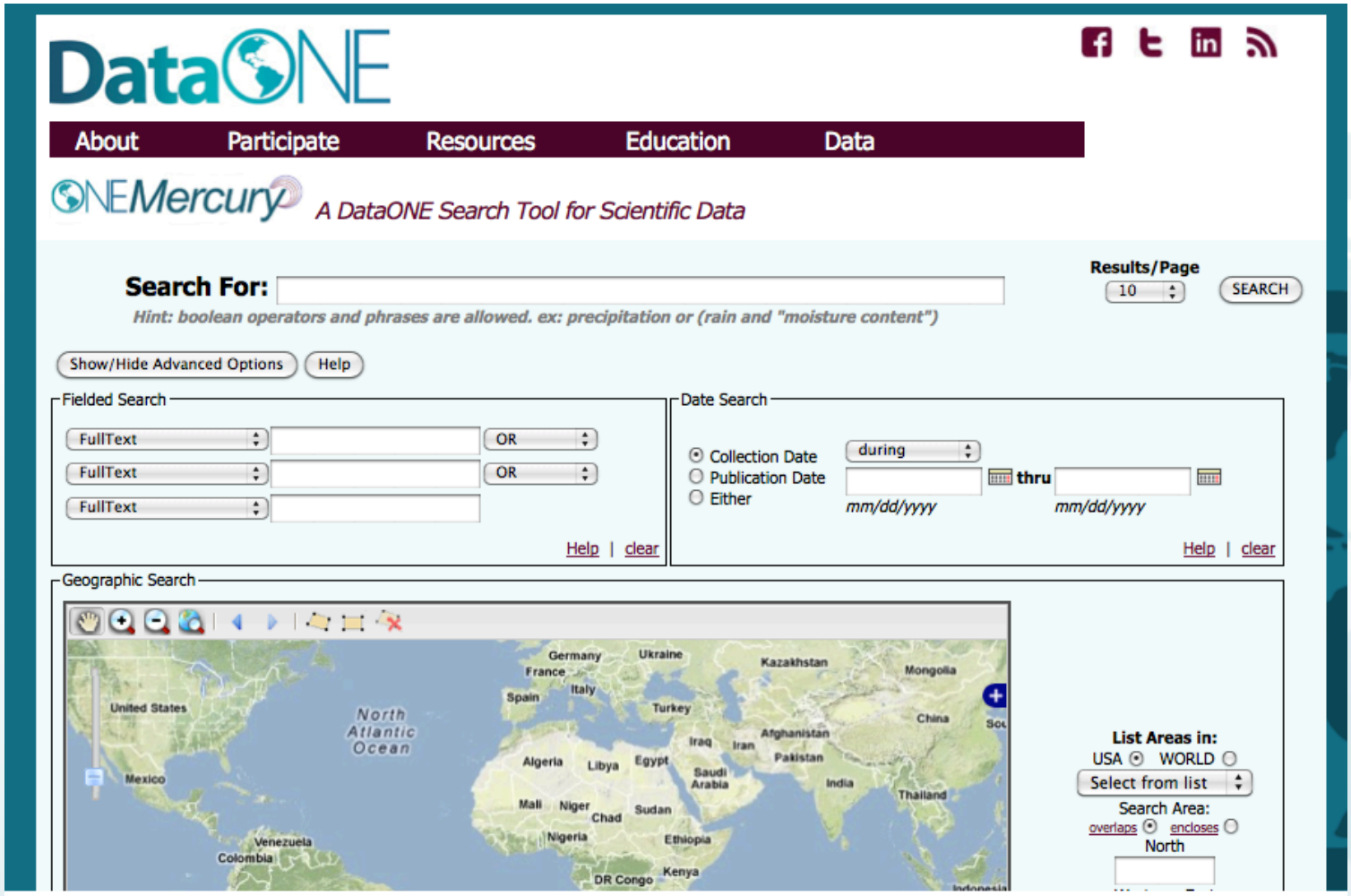
# Diversity of Member Nodes: a Key Factor

| | ORNL-DAAC | Dryad | KNB |
|---|---|---|---|
| Community | Agency repository | Journal consortium | Research network |
| Data | Ecology and biogeochemical dynamics | Biosciences | Biodiversity, ecology, environment |
| Size | ~ 1,000 data products, ~ 1 TB | ~ 1,000 data products, ~ 5 GB | 25,000 data products, 100s GBs |
| Services | Tools for data preservation, replication, discovery, access, subsetting, and visualization | Tools for data preservation, replication, discovery and access | Tools for data preservation, replication, discovery, access, management, and visualization |
| Metadata stds. | FGDC subset | Dublin Core application profile | EML, FGDC |
| Degree of curation | High | Medium | Low |
| Data submission | Staff-assisted submission and curation of final data product | Web-based data submission at time of journal article submission | Self-submission via desktop tool at any time |
| Sponsor | NASA | NSF/JISC, societies, publishers | NSF |

# Goal: Improve DataONE Discovery Services



# Discovery Services based upon ONE Mercury



ONEMercury *A DataONE Search Tool for Scientific Data*

Return to Search | Email Query | Bookmark Query | RSS Feed for Query | Help

**Query: text : water**

Hide Filters

| Filter by author | Filter by project | Filter by keywords | Filter by Originator |
|---|---|---|---|
| Bruce Menge (6459) | Partnership for Interdisciplinary (14809) | temperature (15631) | Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777) |
| Margaret McManus (3465) | Florida Coastal Everglades (512) | Oceanographic Sensor Data (15359) | PISCO (1309) |
| Libe Washburn (3112) | Olympic Coast National (510) | continental shelf (15359) | Monterey Bay National Marine Sanctuary (952) |
| Jack Barth (1612) | Arctic Long-Term Ecological (339) | Temperature (15275) | Olympic Coast National Marine |
| Mary Sue Brancato (510) | | United States of America (15146) | |
| Anne Giblin (298) | | | |
| Pete Raimondi (237) | | | |

**Sort By:** Relevance | Date | Member Node

Viewing Documents 1 - 10 out of 22815
Prev 1 2 3 4 5 6 7 8 9 10 Next

**Filter by Member Node**

- PISCO MN (15460)
- LTER Network Member Node (4289)
- Merritt D1 Member Node (1845)
- Knowledge Network for Biocomplexity (764)
- ORNL DAAC (378)
- SANParks Data Repository (42)
- ESA Data Registry (21)
- USGS Core Sciences Clearinghouse (16)

**THE EFFECT OF OVERLAPPING PIOSPHERES ON LANDSCAPE HETEROGENEITY** 02/09/2004 - 05/09/2004

*Datasource:* SANPARKS DATA REPOSITORY

In 1933, Kruger National Park implemented artificial sources of surface water. Many studies have been conducted on the effect these waterholes have on herbivore distribution and the related impacts. One such finding is that piospheres, patches created by herbivores through their grazing, browsing and trampling activities focusing around a water source (Owen-Smith 1996 cited in Gaylard et al 2002) occur around waterholes and contribute to the patchiness in the landscape. The aim of this study was to determine if the proportion of increaser II grass species would drop below 50% and be replaced b...

View full metadata | Data Files (0)

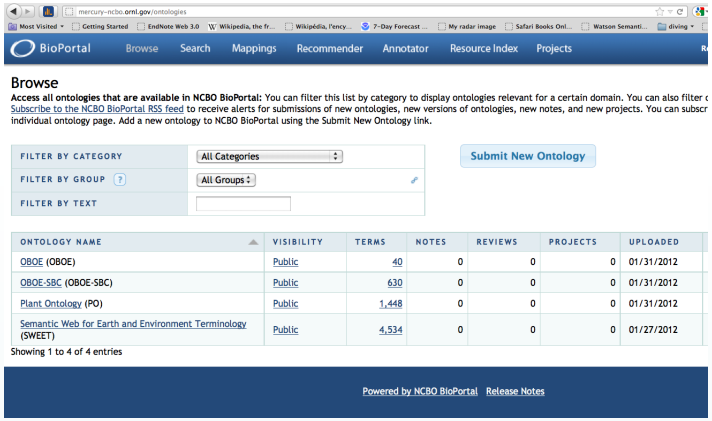
**POD! WATER-QUALITY_DAILY WATERTEMP (1984-06) AND EMP WATER QUALITY** 01/01/1975 - 01/01/2007

*Datasource:* KNOWLEDGE NETWORK FOR BIOCOMPLEXITY

Daily Water Temp 1984-2006 %26#226;%26#128;%26#147; average daily water temperature at three locations. EMP water quality parameters: EMP continuous %26#226;%26#128;%26#147; continuous sampling data available since 1998 EMP discrete %26#226;%26#128;%26#147; monthly sampling data available since 1975 The study area includes the Delta within its legal boundaries, Suisun Bay and Suisun Marsh, and northeastern San Pablo Bay bounded by a line between Pinole Point on the east and the Solano County line on the north shore. The EMP sampling sites range from San Pablo Bay east through the upper Estua...

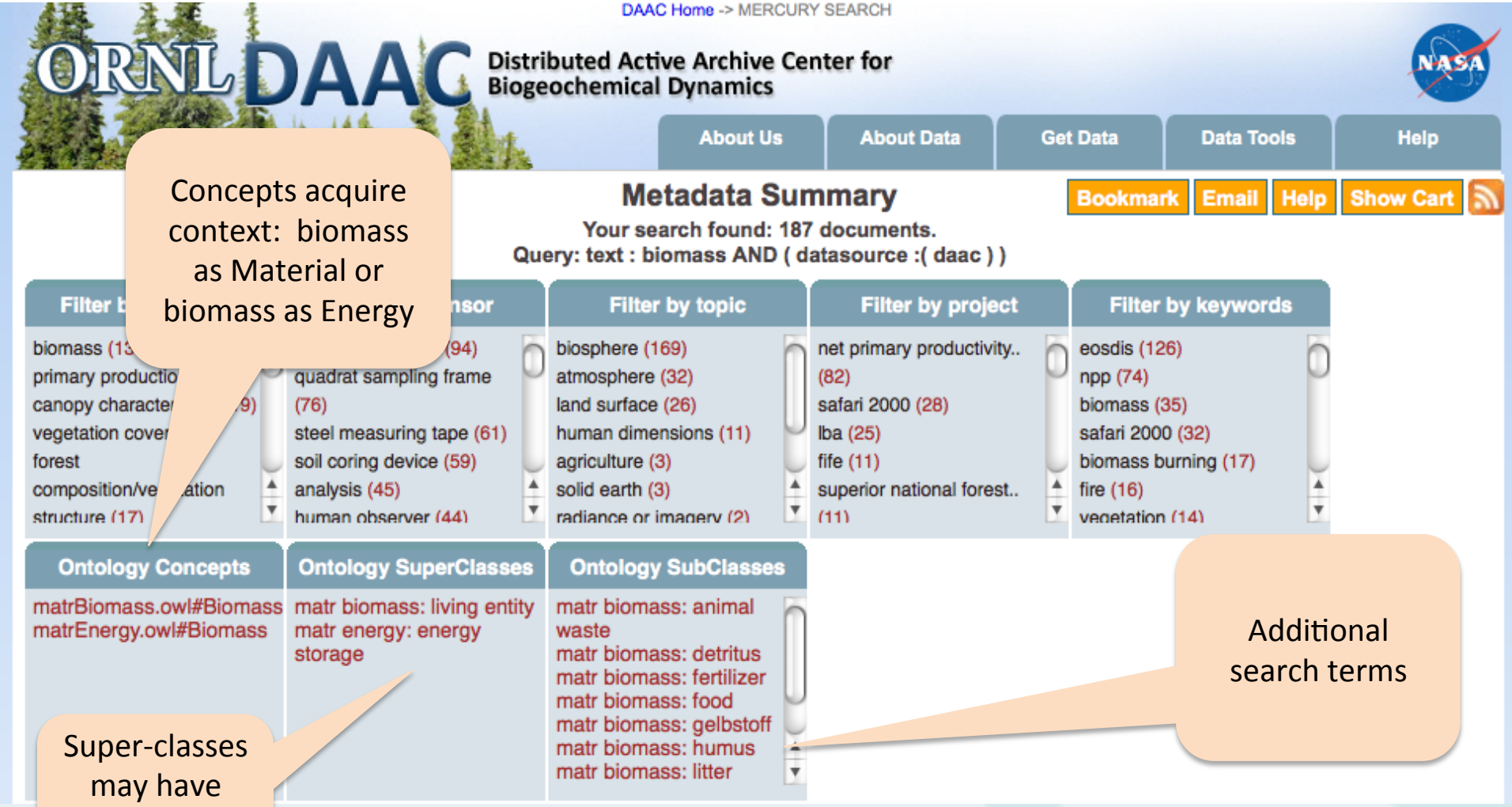
View full metadata | Data Files (2)

# Ontology-based Discovery: Ontology integration into query-results



1. we had a metadata collection and discovery engine (Mercury)
2. we installed the NCBO ontology repository instance on a DataONE node
3. we populated it with Earth Sciences ontologies
    - the NASA-JPL Semantic Web for Earth and Environmental Terminology
4. we queried the ontologies and returned results using REST services

# Ontology-based search results

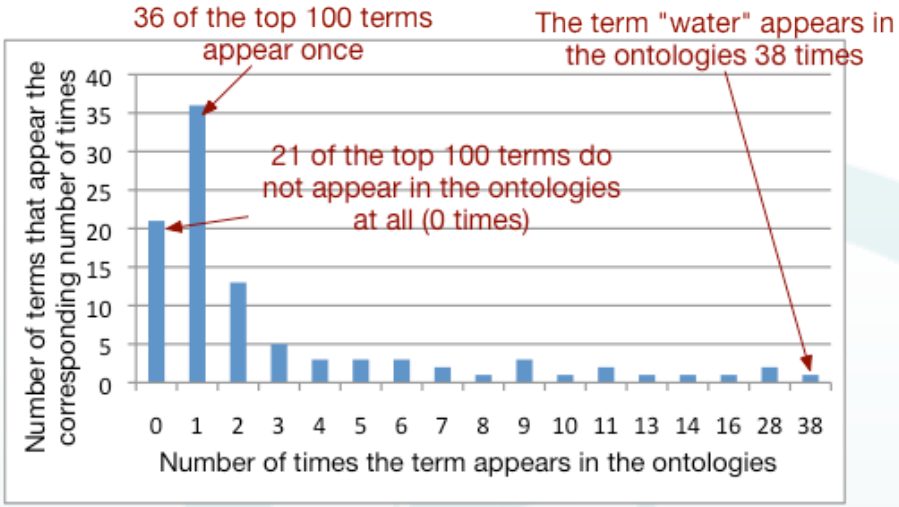


# Limitations and remaining challenges

Display is confusion, too many facets

SWEET provides a good basis but further extensions are needed

Many ontologies provide too few relationships

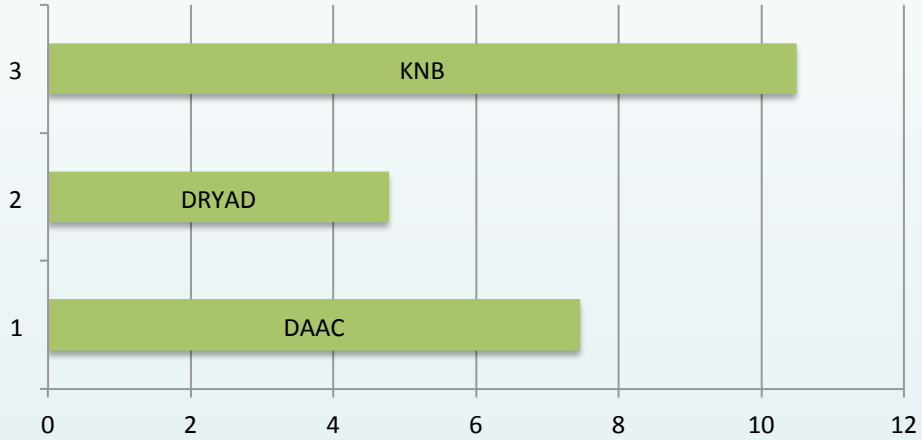



**Conclusion:**

**Adding ontology entities to a keyword index helps with recall but cannot substitute for semantic annotations of the metadata documents themselves**

# Approach 2: Enrich MN Metadata

| | DAAC | DRYAD | KNB |
|---|---|---|---|
| Number of Documents | 978 | 1,729 | 24,249 |
| Total Number of Keywords | 7,294 | 8,266 | 254,525 |
| Average Number of Keyword/ Document | 7.46 | 4.78 | 10.49 |




- Each MN provides keywords
- Suggest additional keywords for each document
- Treated like a tag recommendation problem


15

# Example Scenario- Hydro-Eco Use Case

- *"I want data for streamflow and water quality constituent concentrations co-located with fish and aquatic macroinvertebrate species occurrence observations to inform studies of hydrology and water quality impacts on aquatic organisms."*
- Requires location and access of data from many sources (KNB/LTER, CUAHSI, ORNL DAAC, Dryad)


16

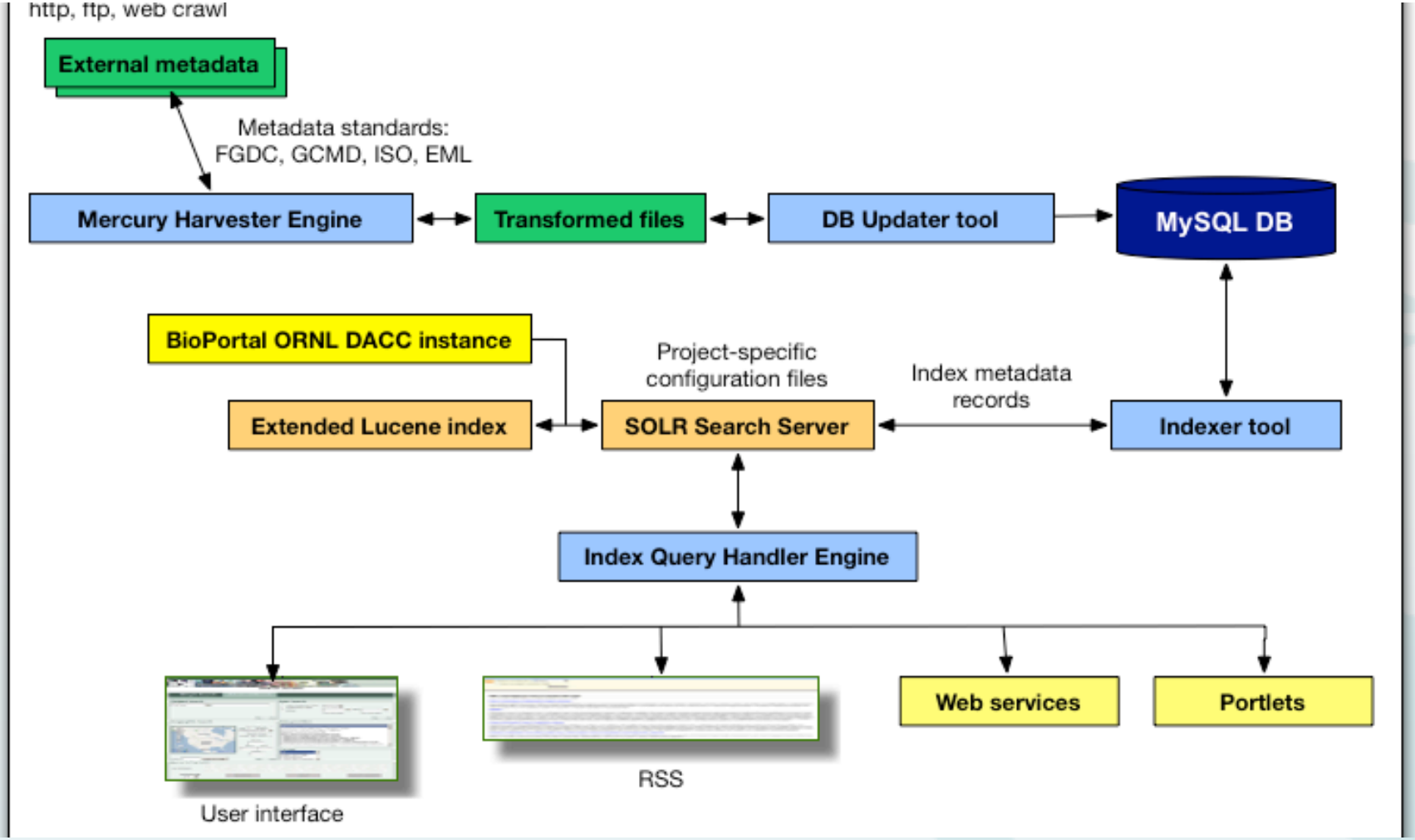
# Integration and Semantics Working Group

Jeff Horsburgh, USU
Deborah McGuinness, RPI
Luis Bermudez, OGC
Damian Gessler, iPlant
Carl Lagoze, U. Michigan
Hilmar Lapp, NESCENT
Carlos Rueda, MMI
Natasha Noy, Stanford
Paulo Pinheiro Da Silva, PNNL
Line Pouchard, ORNL
Mark Schildhauer, NCEAS
Patrice Seyed, UNM / RPI

DataONE

# Acknowledgements

# Architecture: Ontology-based discovery



# Precision and recall

| | Actual Positive | Actual Negative |
|---|---|---|
| Predicted Positive | Tp | Fp |
| Predicted Negative | Fn | Tn |

Precision: $Tp/Tp+Fp$

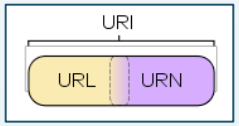
Recall: $Tp/■Tp+Fn$

- Precision: the proportion of Actual positive in the recommended set
- Recall: the proportion of Actual that are recommended over the entire dataset

# CN Services: Identify objects

Goal: Uniquely identify data or metadata objects

- Support the several identifier types widely used
- Identifiers assigned by Member Nodes
- Uniqueness ensured by Coordinating Nodes
- Resolution through Coordinating Nodes




GUID

{3F2504E0-4…

LSID

PURL

# Value of DataONE

- Discovery and access: Enabling discovery and universal access to data about life on earth from around the world
- Data integration and synthesis: Providing transformational tools that enable cross-cutting research
- Education and training: Providing essential skills (e.g., data management training, best practices) for scientific enquiry
- Building community: Combining expertise and resources across diverse communities to collectively educate, advocate, and support stewardship of scientific data
- Data Sharing: Providing incentives and infrastructure for sharing data from federally funded researchers

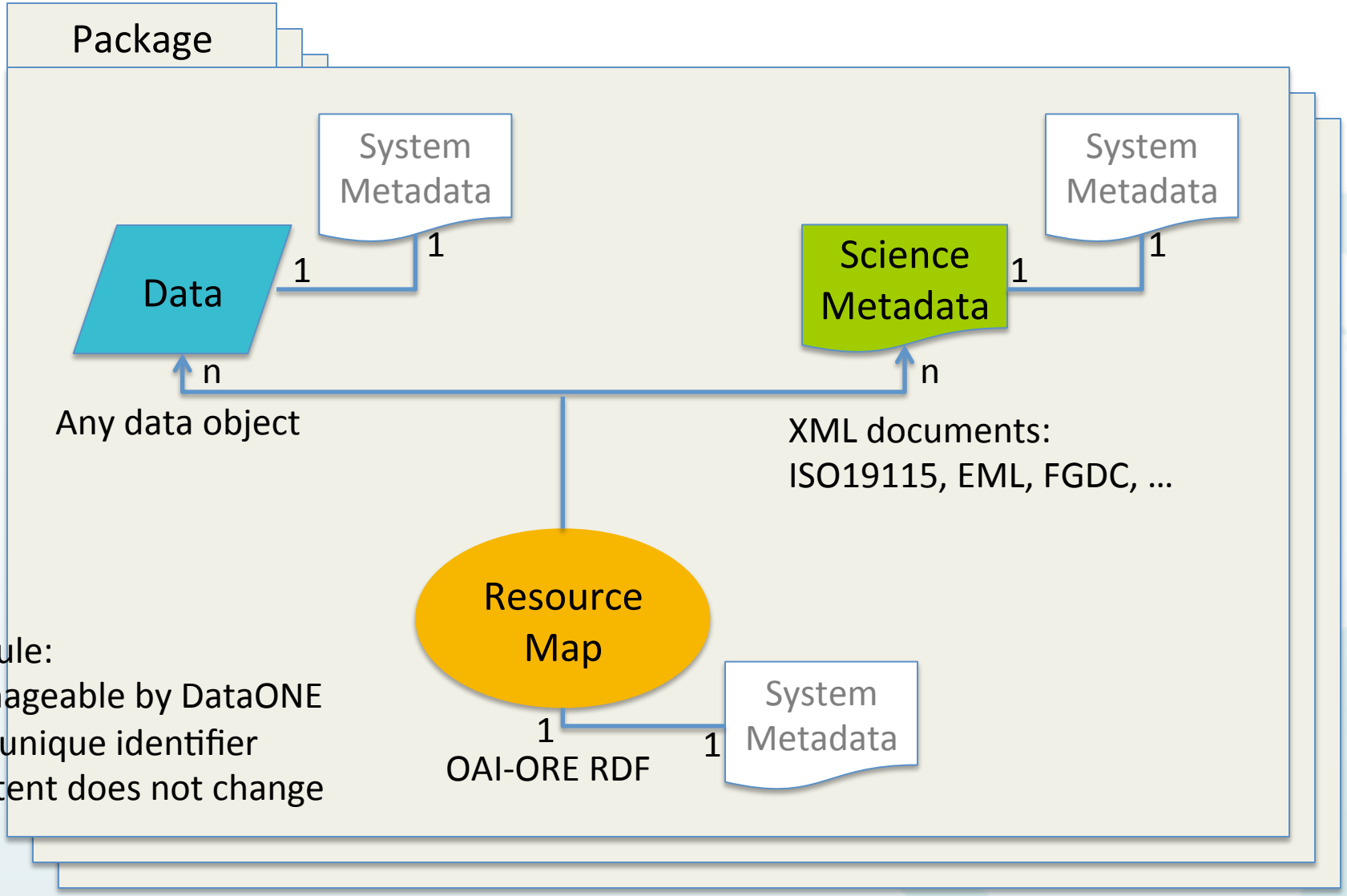


22

# DataONE Team and Sponsors

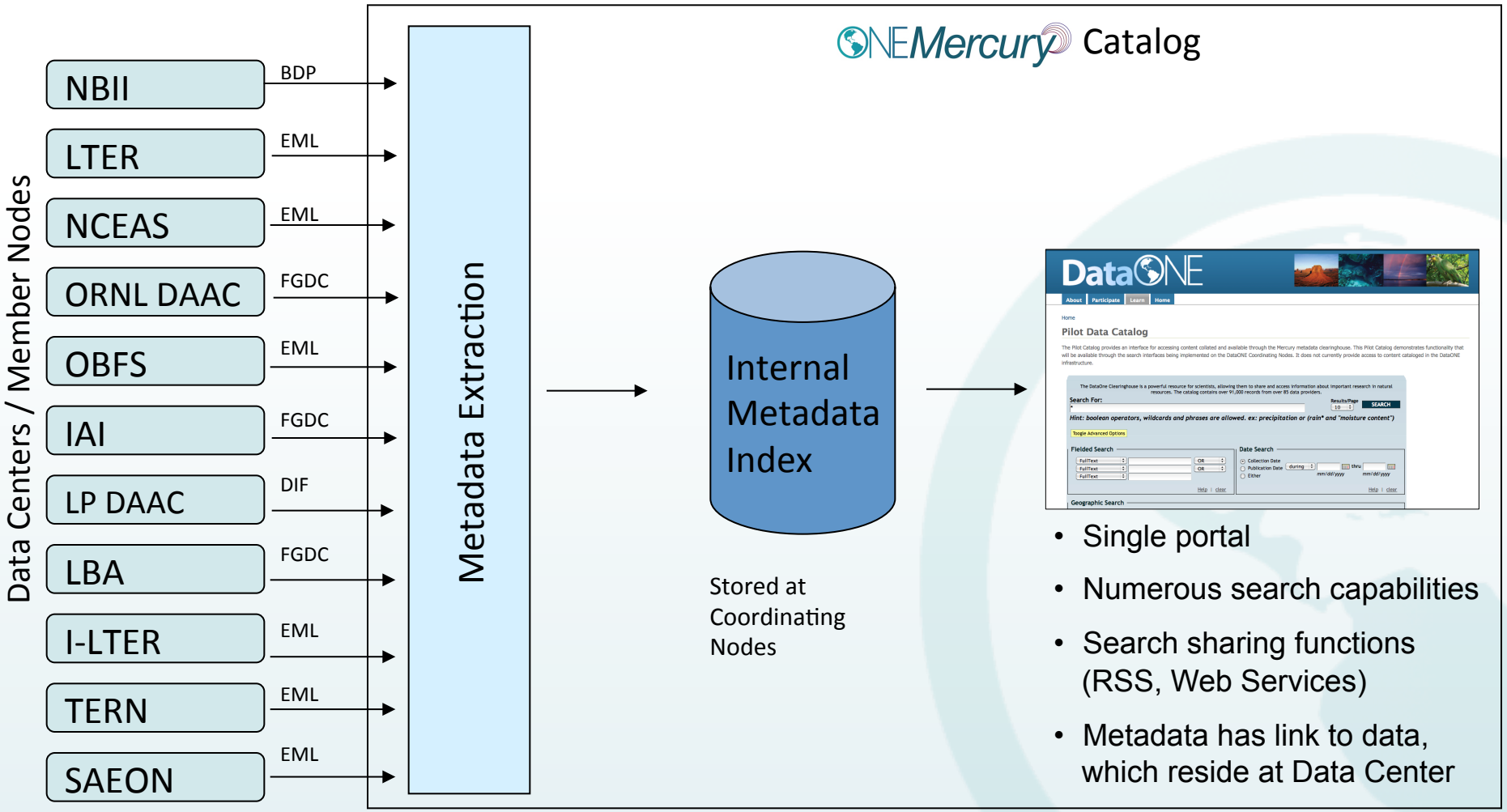
- Amber Budden, Roger Dahl, Rebecca Koskela, Bill Michener, Robert Nahf, Mark Servilla
- Dave Vieglais
- Suzie Allard, Carol Tenopir, Maribeth Manoff, Robert Waltz, Bruce Wilson
- John Cobb, Bob Cook, Giri Palanismy, Line Pouchard
- Patricia Cruse, John Kunze
- Sky Bristol, Mike Frame, Richard Huffine, Viv Hutchison, Jeff Morisette, Jake Weltzin, Lisa Zolly
- Chad Berkley, Stephanie Hampton, Matt Jones
- Paul Allen, Rick Bonney, Steve Kelling
- Ryan Scherle, Todd Vision
- Randy Butler
- Ewa Deelman
- Peter Honeyman
- Jeff Horsburgh
- Robert Sandusky
- Bertram Ludaescher
- Peter Buneman
- Cliff Duke
- Carole Goble
- Donald Hobern
- David DeRoure

# What is an object? DataONE Data Model

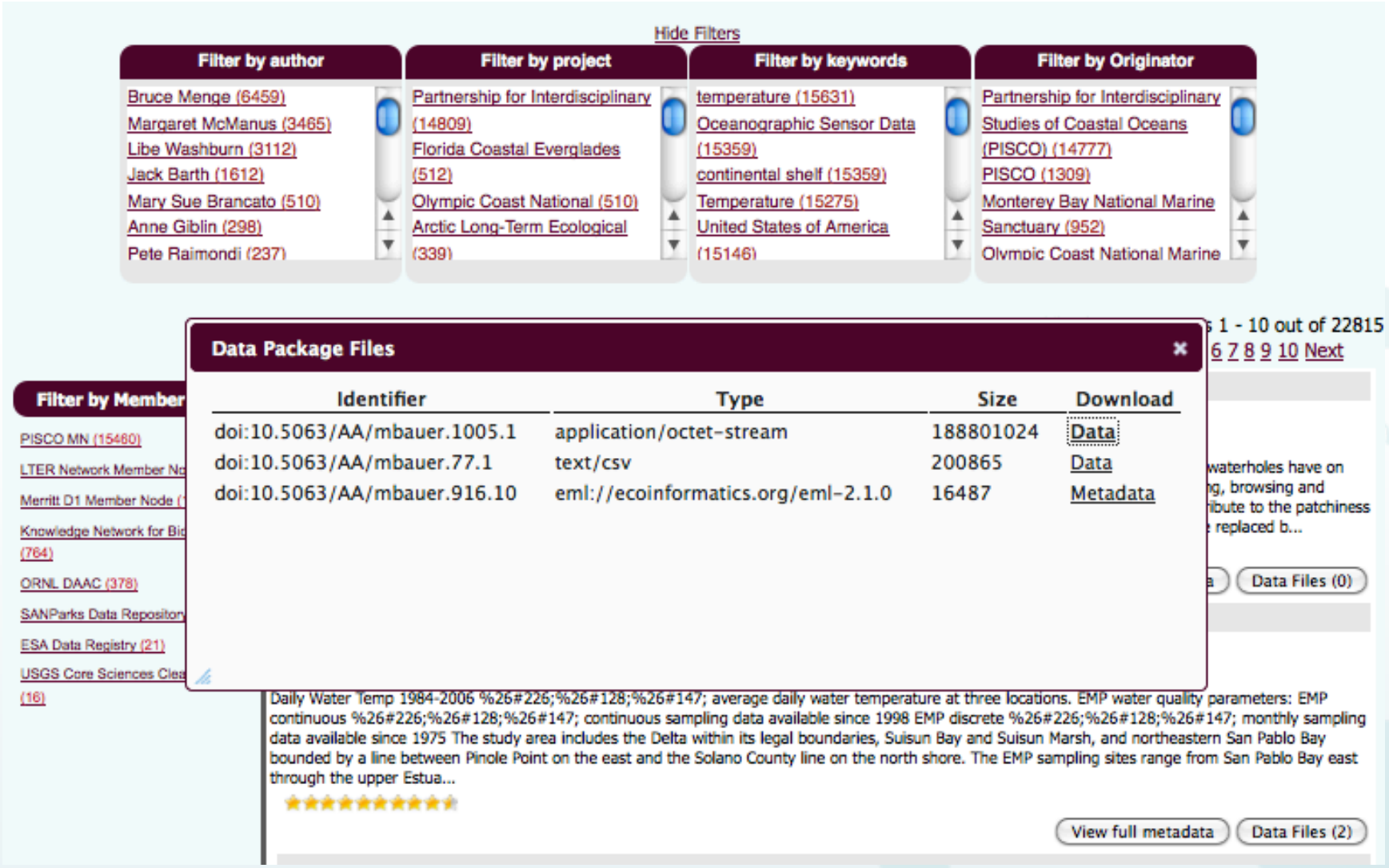


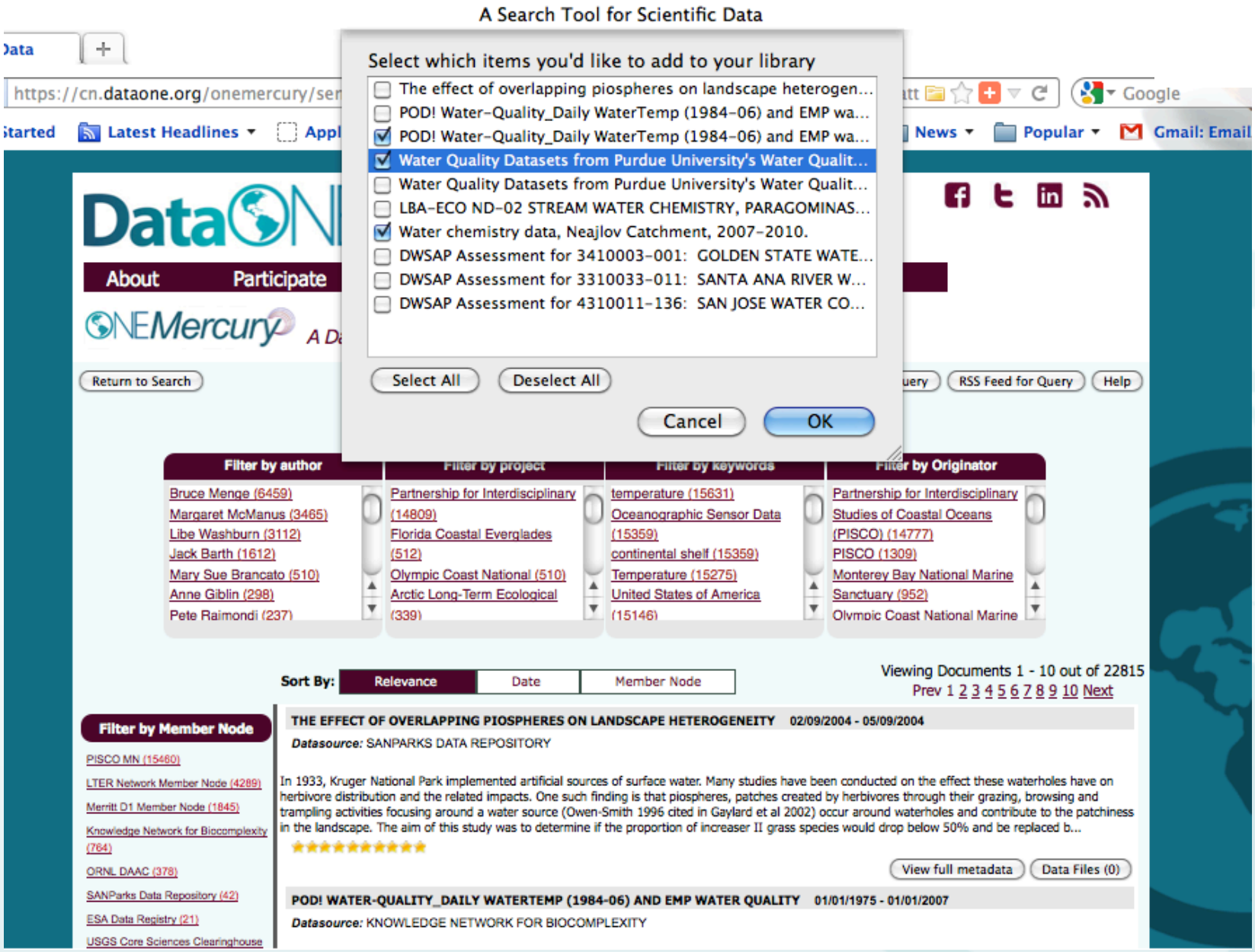

Granule:
- Manageable by DataONE
- Has unique identifier
- Content does not change

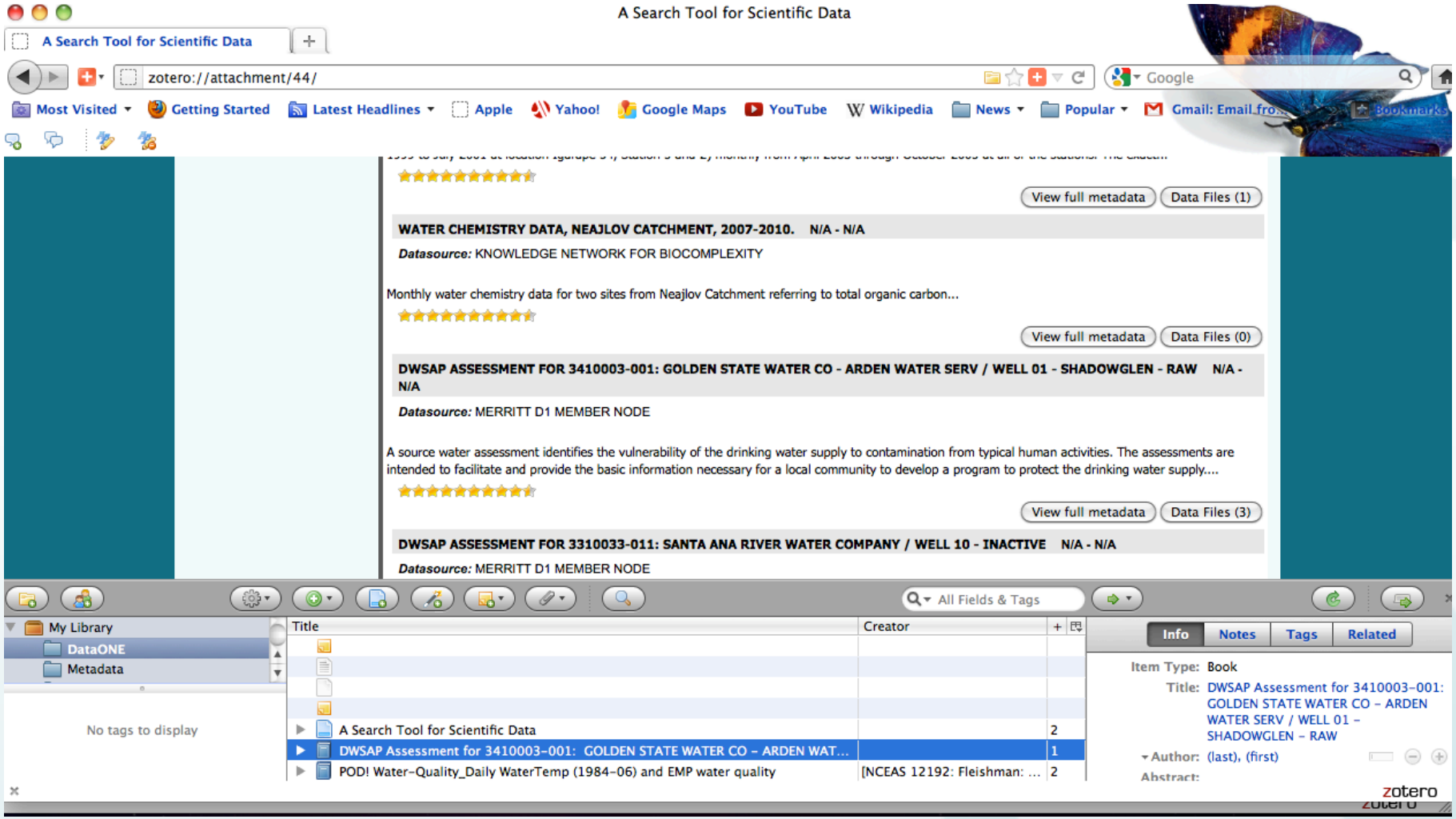
# ONEMercury Architecture



ONEMercury
Results/Page
10
SEARCH
Search For:
Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain* and "moisture content")
Show/Hide Advanced Options
HELP
Fielded Search
FullText
OR
FullText
OR
FullText
Help | clear
Date Search
Collection Date
Publication Date
Either
during
thru
mm/dd/yyyy
mm/dd/yyyy
Help | clear
Geographic Search
North America
Atlantic Ocean
Europe
Asia
Africa
South America
Pacific Ocean
Indian Ocean
Australia
Pacific Ocean
Antarctica
POWERED BY Google
Terms of Use
List Areas in:
USA
WORLD
Select from list
Search Area:
overlaps
encloses
North
West
East
South
Place Name:
view on map
Help | clear
Content Type
All
Maps and Data
Publications
Tools and Software
Member Nodes
All
ORNL Distributed Active Archive Center for Biogeochemical Dynamics (ORN
Dryad
NBII Metadata Clearinghouse
All NBII Partner Nodes
NBII Metadata Clearinghouse Principal Node
Selected Query (Not Editable)


Hide Filters
Filter by author
Bruce Menge (6459)
Margaret McManus (3465)
Libe Washburn (3112)
Jack Barth (1612)
Mary Sue Brancato (510)
Anne Giblin (298)
Pete Raimondi (237)
Filter by project
Partnership for Interdisciplinary (14809)
Florida Coastal Everglades (512)
Olympic Coast National (510)
Arctic Long-Term Ecological (339)
Filter by keywords
temperature (15631)
Oceanographic Sensor Data (15359)
continental shelf (15359)
Temperature (15275)
United States of America (15146)
Filter by Originator
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777)
PISCO (1309)
Monterey Bay National Marine Sanctuary (952)
Olympic Coast National Marine
Data Package Files
Identifier
Type
Size
Download
doi:10.5063/AA/mbauer.1005.1
application/octet-stream
188801024
Data
doi:10.5063/AA/mbauer.77.1
text/csv
200865
Data
doi:10.5063/AA/mbauer.916.10
eml://ecoinformatics.org/eml-2.1.0
16487
Metadata
1 - 10 out of 22815
6 7 8 9 10 Next
Filter by Member
PISCO MN (15460)
LTER Network Member No
Merritt D1 Member Node (
Knowledge Network for Bio
(764)
ORNL DAAC (378)
SANParks Data Repository
ESA Data Registry (21)
USGS Core Sciences Clea
(16)
waterholes have on
g, browsing and
ibute to the patchiness
replaced b...
Data Files (0)
Daily Water Temp 1984-2006 %26#226;%26#128;%26#147; average daily water temperature at three locations. EMP water quality parameters: EMP continuous %26#226;%26#128;%26#147; continuous sampling data available since 1998 EMP discrete %26#226;%26#128;%26#147; monthly sampling data available since 1975 The study area includes the Delta within its legal boundaries, Suisun Bay and Suisun Marsh, and northeastern San Pablo Bay bounded by a line between Pinole Point on the east and the Solano County line on the north shore. The EMP sampling sites range from San Pablo Bay east through the upper Estua...
View full metadata
Data Files (2)


A Search Tool for Scientific Data
Select which items you'd like to add to your library
The effect of overlapping piospheres on landscape heterogen...
POD! Water-Quality_Daily WaterTemp (1984-06) and EMP wa...
POD! Water-Quality_Daily WaterTemp (1984-06) and EMP wa...
Water Quality Datasets from Purdue University's Water Qualit...
Water Quality Datasets from Purdue University's Water Qualit...
LBA-ECO ND-02 STREAM WATER CHEMISTRY, PARAGOMINAS...
Water chemistry data, Neajlov Catchment, 2007-2010.
DWSAP Assessment for 3410003-001: GOLDEN STATE WATE...
DWSAP Assessment for 3310033-011: SANTA ANA RIVER W...
DWSAP Assessment for 4310011-136: SAN JOSE WATER CO...
Select All
Deselect All
Cancel
OK
About
Participate
Return to Search
RSS Feed for Query
Help
Filter by author
Bruce Menge (6459)
Margaret McManus (3465)
Libe Washburn (3112)
Jack Barth (1612)
Mary Sue Brancato (510)
Anne Giblin (298)
Pete Raimondi (237)
Filter by project
Partnership for Interdisciplinary (14809)
Florida Coastal Everglades (512)
Olympic Coast National (510)
Arctic Long-Term Ecological (339)
Filter by keywords
temperature (15631)
Oceanographic Sensor Data (15359)
continental shelf (15359)
Temperature (15275)
United States of America (15146)
Filter by Originator
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO) (14777)
PISCO (1309)
Monterey Bay National Marine Sanctuary (952)
Olympic Coast National Marine
Sort By:
Relevance
Date
Member Node
Viewing Documents 1 - 10 out of 22815
Prev 1 2 3 4 5 6 7 8 9 10 Next
Filter by Member Node
PISCO MN (15460)
LTER Network Member Node (4289)
Merritt D1 Member Node (1845)
Knowledge Network for Biocomplexity (764)
ORNL DAAC (378)
SANParks Data Repository (42)
ESA Data Registry (21)
USGS Core Sciences Clearinghouse
THE EFFECT OF OVERLAPPING PIOSPHERES ON LANDSCAPE HETEROGENEITY 02/09/2004 - 05/09/2004
Datasource: SANPARKS DATA REPOSITORY
In 1933, Kruger National Park implemented artificial sources of surface water. Many studies have been conducted on the effect these waterholes have on herbivore distribution and the related impacts. One such finding is that piospheres, patches created by herbivores through their grazing, browsing and trampling activities focusing around a water source (Owen-Smith 1996 cited in Gaylard et al 2002) occur around waterholes and contribute to the patchiness in the landscape. The aim of this study was to determine if the proportion of increaser II grass species would drop below 50% and be replaced b...
View full metadata
Data Files (0)
POD! WATER-QUALITY_DAILY WATERTEMP (1984-06) AND EMP WATER QUALITY 01/01/1975 - 01/01/2007
Datasource: KNOWLEDGE NETWORK FOR BIOCOMPLEXITY


A Search Tool for Scientific Data
A Search Tool for Scientific Data
zotero://attachment/44/
Google
Most Visited
Getting Started
Latest Headlines
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News
Popular
Gmail: Email.fro...
Bookmarks
View full metadata
Data Files (1)
WATER CHEMISTRY DATA, NEAJLOV CATCHMENT, 2007-2010. N/A - N/A
Datasource: KNOWLEDGE NETWORK FOR BIOCOMPLEXITY
Monthly water chemistry data for two sites from Neajlov Catchment referring to total organic carbon...
View full metadata
Data Files (0)
DWSAP ASSESSMENT FOR 3410003-001: GOLDEN STATE WATER CO - ARDEN WATER SERV / WELL 01 - SHADOWGLEN - RAW N/A - N/A
Datasource: MERRITT D1 MEMBER NODE
A source water assessment identifies the vulnerability of the drinking water supply to contamination from typical human activities. The assessments are intended to facilitate and provide the basic information necessary for a local community to develop a program to protect the drinking water supply....
View full metadata
Data Files (3)
DWSAP ASSESSMENT FOR 3310033-011: SANTA ANA RIVER WATER COMPANY / WELL 10 - INACTIVE N/A - N/A
Datasource: MERRITT D1 MEMBER NODE
All Fields & Tags
My Library
DataONE
Metadata
No tags to display
Title
Creator
A Search Tool for Scientific Data
2
DWSAP Assessment for 3410003-001: GOLDEN STATE WATER CO - ARDEN WAT...
1
POD! Water-Quality_Daily WaterTemp (1984-06) and EMP water quality
[NCEAS 12192: Fleishman: ...
2
Info
Notes
Tags
Related
Item Type: Book
Title: DWSAP Assessment for 3410003-001: GOLDEN STATE WATER CO - ARDEN WATER SERV / WELL 01 - SHADOWGLEN - RAW
Author: (last), (first)
Abstract:
zotero

# ORNL DAAC as a DataONE Member Node

DataONE

CDL California Digital Library

neon National Ecological Observatory Network, Inc.

USGS science for a changing world

NESCent

NSF

**Investigator Toolkit**

MENDELEY RESEARCH NETWORKS

VEGBANK

Specify 6

Kepler

Java

my experiment

MATLAB

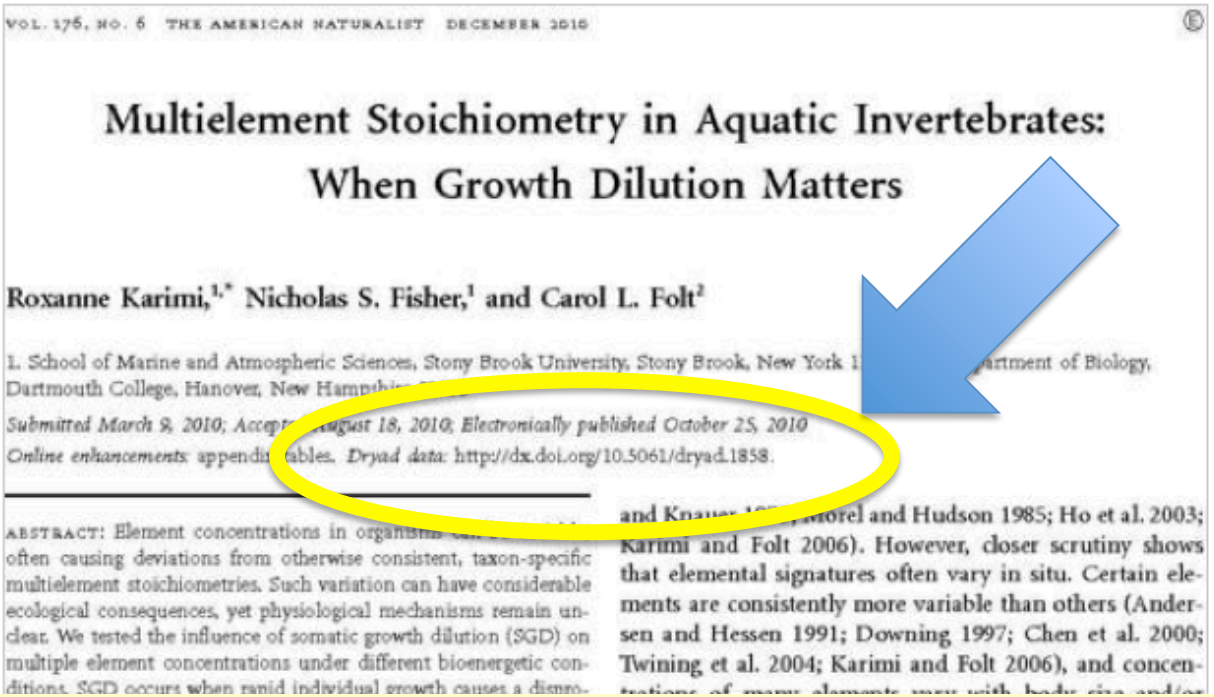
Excel:mac

zotero

ORNL DAAC Member Node

DataONE Users

# Provide credit for data publication



VOL. 176, NO. 6 THE AMERICAN NATURALIST DECEMBER 2010

## Multielement Stoichiometry in Aquatic Invertebrates: When Growth Dilution Matters

Roxanne Karimi,[1,*] Nicholas S. Fisher,[1] and Carol L. Folt[2]

1. School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 1... Department of Biology, Dartmouth College, Hanover, New Hampshi...

*Submitted March 9, 2010; Accept... August 18, 2010; Electronically published October 25, 2010*

*Online enhancements:* appendi... tables. *Dryad data:* http://dx.doi.org/10.5061/dryad.1858.

ABSTRACT: Element concentrations in organisms... often causing deviations from otherwise consistent, taxon-specific multielement stoichiometries. Such variation can have considerable ecological consequences, yet physiological mechanisms remain unclear. We tested the influence of somatic growth dilution (SGD) on multiple element concentrations under different bioenergetic conditions. SGD occurs when rapid individual growth causes a dispro...

...and Knauer 19... Morel and Hudson 1985; Ho et al. 2003; Karimi and Folt 2006). However, closer scrutiny shows that elemental signatures often vary in situ. Certain elements are consistently more variable than others (Andersen and Hessen 1991; Downing 1997; Chen et al. 2000; Twining et al. 2004; Karimi and Folt 2006), and concentrations of many elements may with body size and/or

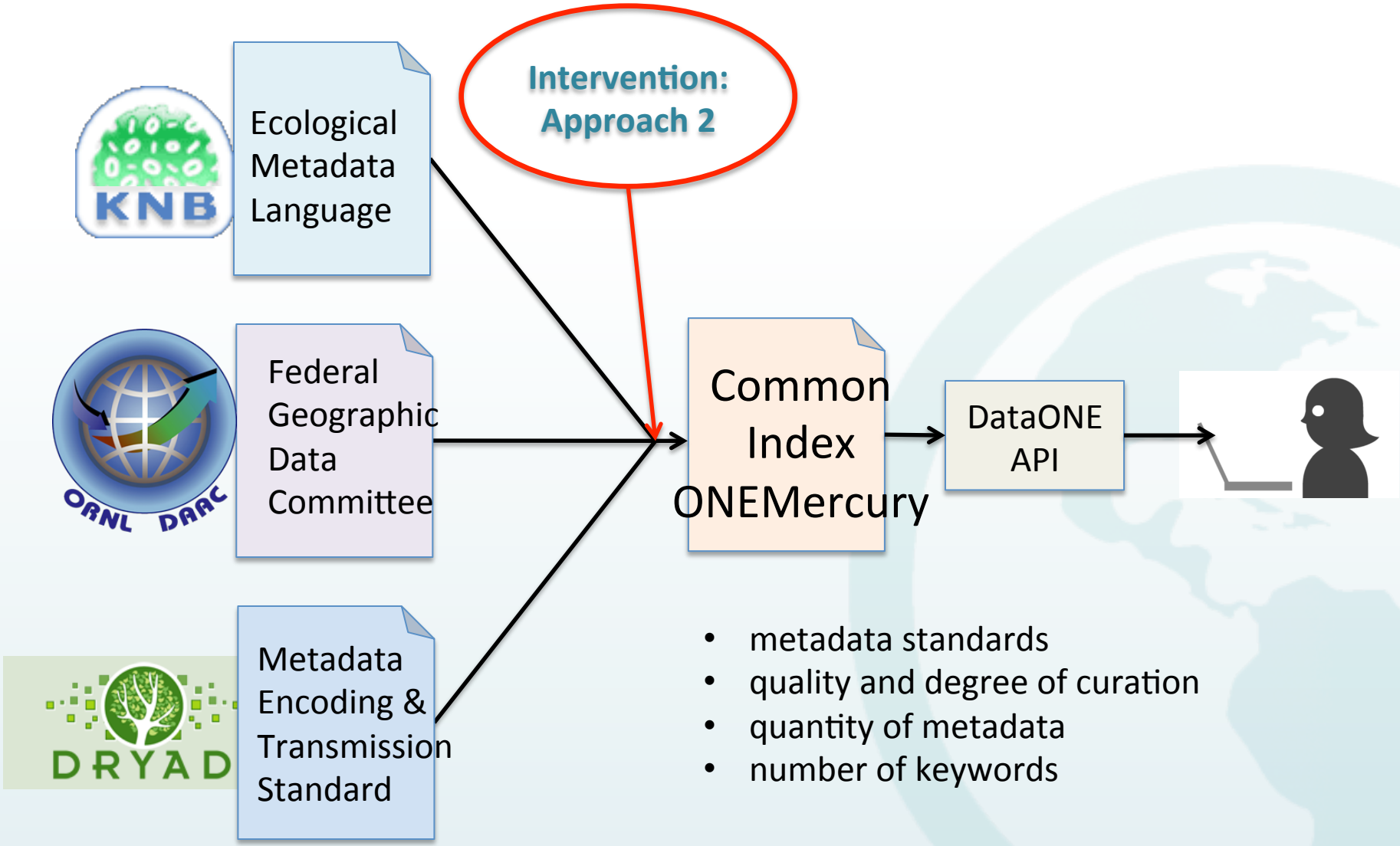
When using this data, please cite the original article:

> Ally D, Ritland K, Otto SP (2008) Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Molecular Ecology 17(22): 4897-4911. doi:10.1111/j.1365-294X.2008.03962.x

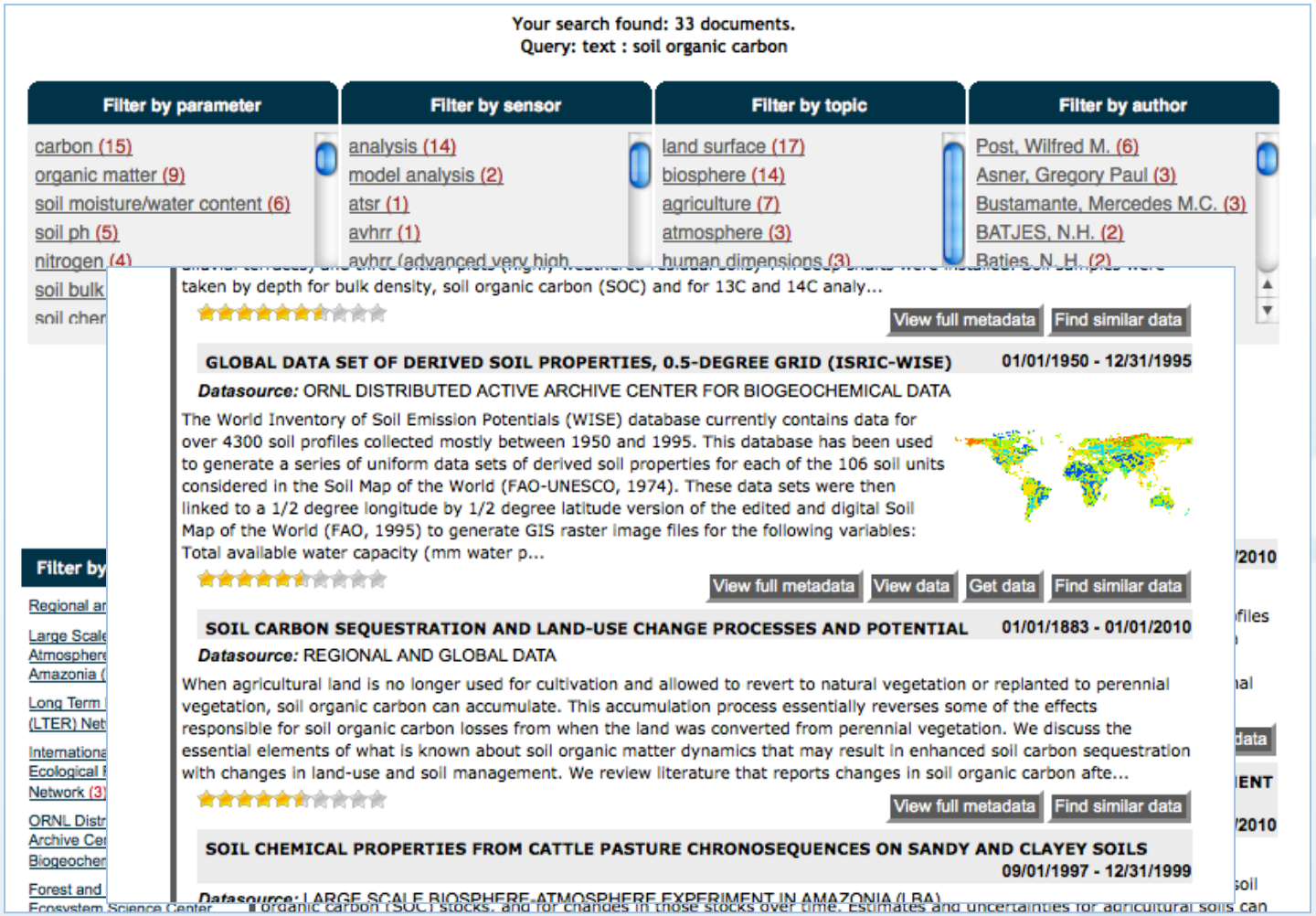
Additionally, please cite the Dryad data package:

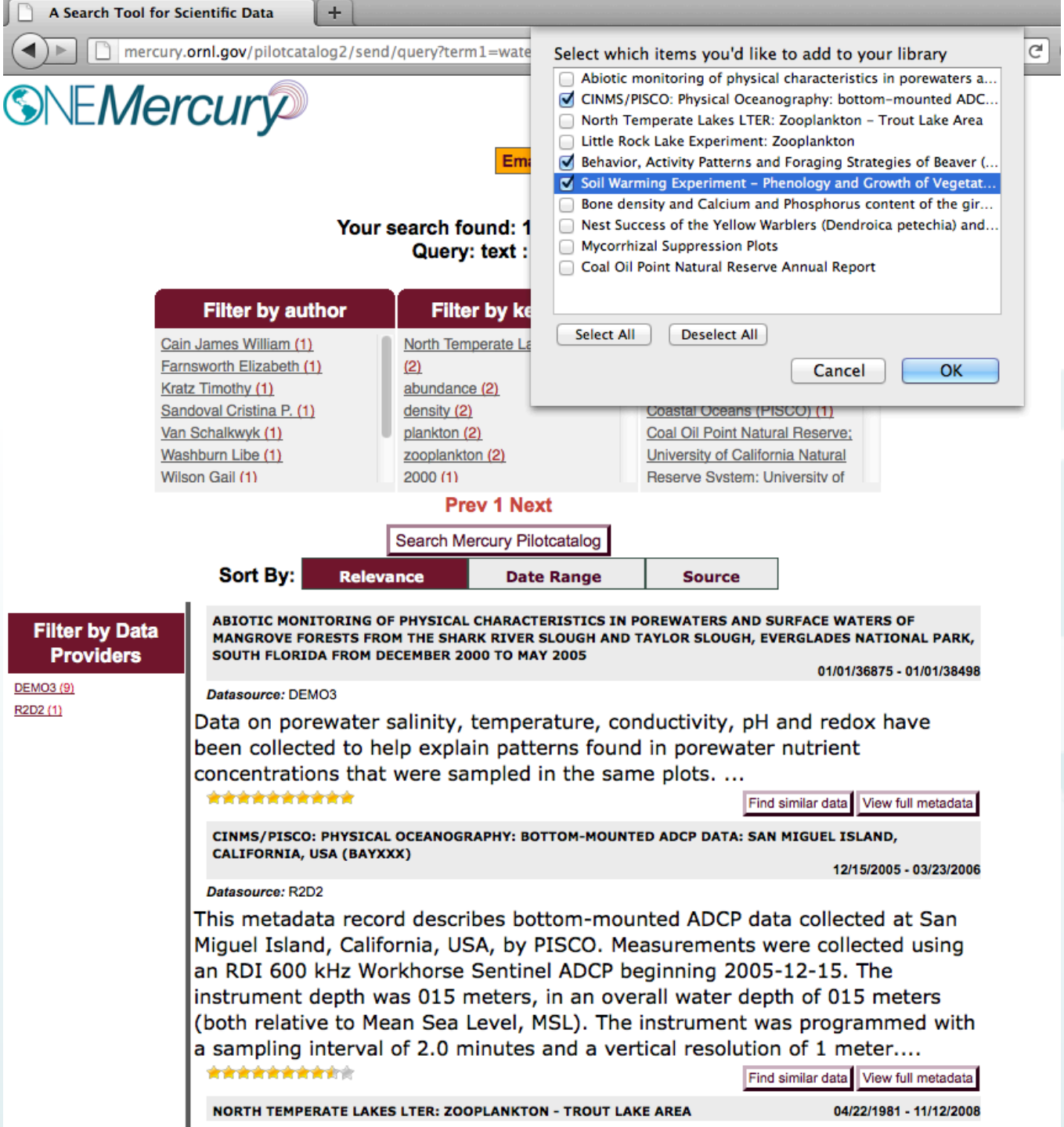
> Ally D, Ritland K, Otto SP (2008) Data from: Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Dryad Digital Repository. doi:10.5061/dryad.7898

# Goal: Improve DataONE Discovery Services



# Quickly discover content




A Search Tool for Scientific Data
mercury.ornl.gov/pilotcatalog2/send/query?term1=wate
ONEMercury
Select which items you'd like to add to your library
Abiotic monitoring of physical characteristics in porewaters a...
CINMS/PISCO: Physical Oceanography: bottom-mounted ADC...
North Temperate Lakes LTER: Zooplankton - Trout Lake Area
Little Rock Lake Experiment: Zooplankton
Behavior, Activity Patterns and Foraging Strategies of Beaver (...
Soil Warming Experiment - Phenology and Growth of Vegetat...
Bone density and Calcium and Phosphorus content of the gir...
Nest Success of the Yellow Warblers (Dendroica petechia) and...
Mycorrhizal Suppression Plots
Coal Oil Point Natural Reserve Annual Report
Select All
Deselect All
Cancel
OK
Your search found: 1
Query: text :
Filter by author
Cain James William (1)
Farnsworth Elizabeth (1)
Kratz Timothy (1)
Sandoval Cristina P. (1)
Van Schalkwyk (1)
Washburn Libe (1)
Wilson Gail (1)
North Temperate La
(2)
abundance (2)
density (2)
plankton (2)
zooplankton (2)
2000 (1)
Coastal Oceans (PISCO) (1)
Coal Oil Point Natural Reserve;
University of California Natural
Reserve System: University of
Prev 1 Next
Search Mercury Pilotcatalog
Sort By:
Relevance
Date Range
Source
Filter by Data Providers
DEMO3 (9)
R2D2 (1)
ABIOTIC MONITORING OF PHYSICAL CHARACTERISTICS IN POREWATERS AND SURFACE WATERS OF MANGROVE FORESTS FROM THE SHARK RIVER SLOUGH AND TAYLOR SLOUGH, EVERGLADES NATIONAL PARK, SOUTH FLORIDA FROM DECEMBER 2000 TO MAY 2005
01/01/36875 - 01/01/38498
Datasource: DEMO3
Data on porewater salinity, temperature, conductivity, pH and redox have been collected to help explain patterns found in porewater nutrient concentrations that were sampled in the same plots. ...
Find similar data
View full metadata
CINMS/PISCO: PHYSICAL OCEANOGRAPHY: BOTTOM-MOUNTED ADCP DATA: SAN MIGUEL ISLAND, CALIFORNIA, USA (BAYXXX)
12/15/2005 - 03/23/2006
Datasource: R2D2
This metadata record describes bottom-mounted ADCP data collected at San Miguel Island, California, USA, by PISCO. Measurements were collected using an RDI 600 kHz Workhorse Sentinel ADCP beginning 2005-12-15. The instrument depth was 015 meters, in an overall water depth of 015 meters (both relative to Mean Sea Level, MSL). The instrument was programmed with a sampling interval of 2.0 minutes and a vertical resolution of 1 meter....
Find similar data
View full metadata
NORTH TEMPERATE LAKES LTER: ZOOPLANKTON - TROUT LAKE AREA
04/22/1981 - 11/12/2008