

DataONE: An interoperable data repositories case study

John W. Cobb

R&D Staff and DataONE Leadership Team Member
Oak Ridge National Laboratory

HUBbub 2012 , the HUBzero conference

Indianapolis, IN
24 September 2012

DataONE



Acknowledgment:

- Authorship: This talk represents work of the entire DataONE extended team.
- It especially draws upon slide material from
 - Bill Michener, UNM (esp. recent DataONE AHM Sept. 18, 2012)
 - Amber Budden – DataONE Ass. Dir. For CE
- DataONE is an NSF supported project (OCI-0830944)



Hubs and data repositories

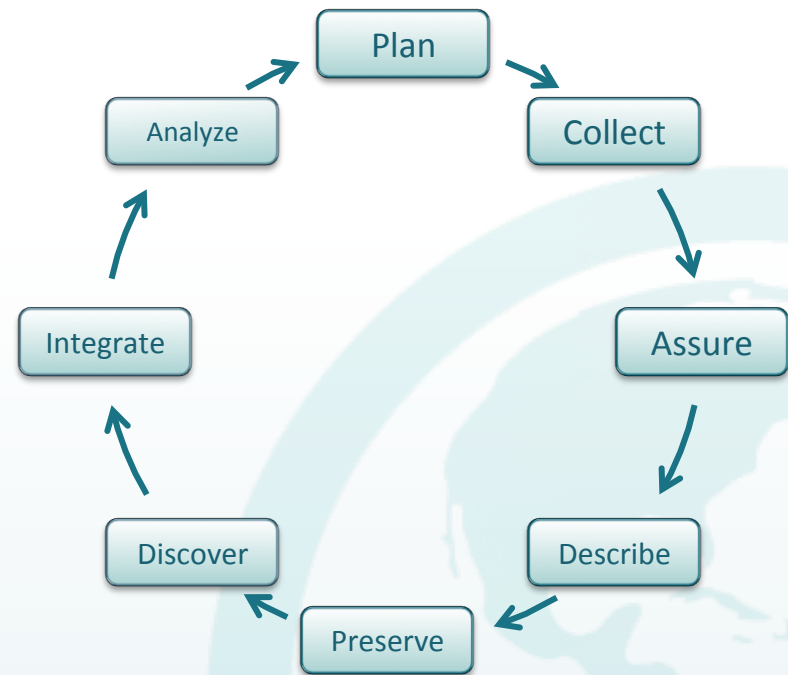
- A personal view
(apologies for a possibly mis-informed speaker)
- HUB-roots (history and pre-history)
 - PUNCH: web portal for running tools
(DOI: 10.1109/40.846308)
 - -> NanoHUB: Application orchestration environment
 - + RAPPTURE: Rapid Application porting and development
 - + Frameless VNC windows –
seamless hosted environment on clients!
 - + Rich collaborative environment and rich user experience !! (“wishlist”)
 - Repurpose: Hubzero -> hubs explode
(ex. NEESHUB a critical advantage for largest research award in Purdue history)
- Now (and recent past) turn to Hub+Data Integration. Some successes already
- Opportunity: Richer interactions between HUB’s and multiple data repositories
- Perhaps for example: Enable multi-project collaboration within PURR?
- Or: Integrate NEES DB’s with SCEC simulations and IRIS waveforms?

Multiple data repository access?

- HUB + Database exists
 - HUB + external data repository access use case.
 - But What if?
-
- Access multiple (possibly external) repositories from within a HUB environment?
 - Access multiple external repositories with similar data? Say aggregate all data from state hydrologists? C.f. driNET <http://drinet.hubzero.org>
 - Integrate disparate data sets for new and novel analysis.
Recall Noshir Contractor's comments this morning: teaming and interdisciplinary work has increased impact (Wuchty, Jones, Uzzi)
 - Enable reproducible analysis and synthesis via a automated workflow to create synthetic data products
 - Programmatic access
 - More integration (more than just raw search terms a la Google)
 - ...
 - What do you want to discover today? (to paraphrase Microsoft)

DataONE motivation

- DataONE is a project to address these issues
- Build (assemble/aggregate) data repository interoperability
- Advance state of the practice data lifecycle management
 - Planning
 - Deposition
 - Metadata generation
 - Semantic integration
 - Workflow and provenance
 - Analysis
 - Synthesis
- Focus on a broad science area
- Deploy a working CI and grow it
- DataONE – Data Observation Network Earth



The Economist

The euro crisis, continued
Attacking the Fed
What's up with North Korea
Germany's model Mittal-management
Saving Fiat from Italy

How to live with climate change

SCIENCE

Barack Obama v business
How to fix the euro
Great managers: born, not made?
Cleaning up sport
North Korea: thanks Dad

The world's lungs
Forests, and how to save them
A 14-PAGE SPECIAL REPORT

The Economist

Brazil as the next oil giant
God help Italy
London's funny but sad election
The return of Disney
Cancer's link to stem cells

TIME

Torching the Amazon
Can the rain forest be saved?

BusinessWeek

GLOBAL WARMING
Why Business Is Taking It So Seriously
BY JOEY CAREY (P. 60)

The Economist

Obama's Working Class Wives
Why More Women Are Choosing C-Sections
Can Richard Branson Save the Airline Industry?

TIME

Stopping climate change
A 14-PAGE SPECIAL REPORT

The Economist

Living with Cancer
Beyond Baghdad: Where The Enemy Has Its Own Surge
The Sopranos' Last Song: What Exit Will Tony Take?

TIME

The Global Warming Survival Guide
51 Things You Can Do to Make a Difference

The Economist

East Asia's economies, five years on
Why Arab countries have failed
THE GLOBAL ENVIRONMENT SURVEY, AFTER PAGE 50

CO₂AL
Environmental enemy No. 1

TIME

Dr. Bush's Rx for Health Care

TIME

VANISHING OZONE
THE DANGER MOVES CLOSER TO HOME

TIME

Why California Is Burning

TIME

Our Filthy Seas

TIME

How to Save the Earth

The hot and wild weather is a sign of things to come. But fresh ideas and new technology can cool us down and make this a **GREEN CENTURY**

Science

13 June 2008 \$10

FORESTS IN FLUX

TIME

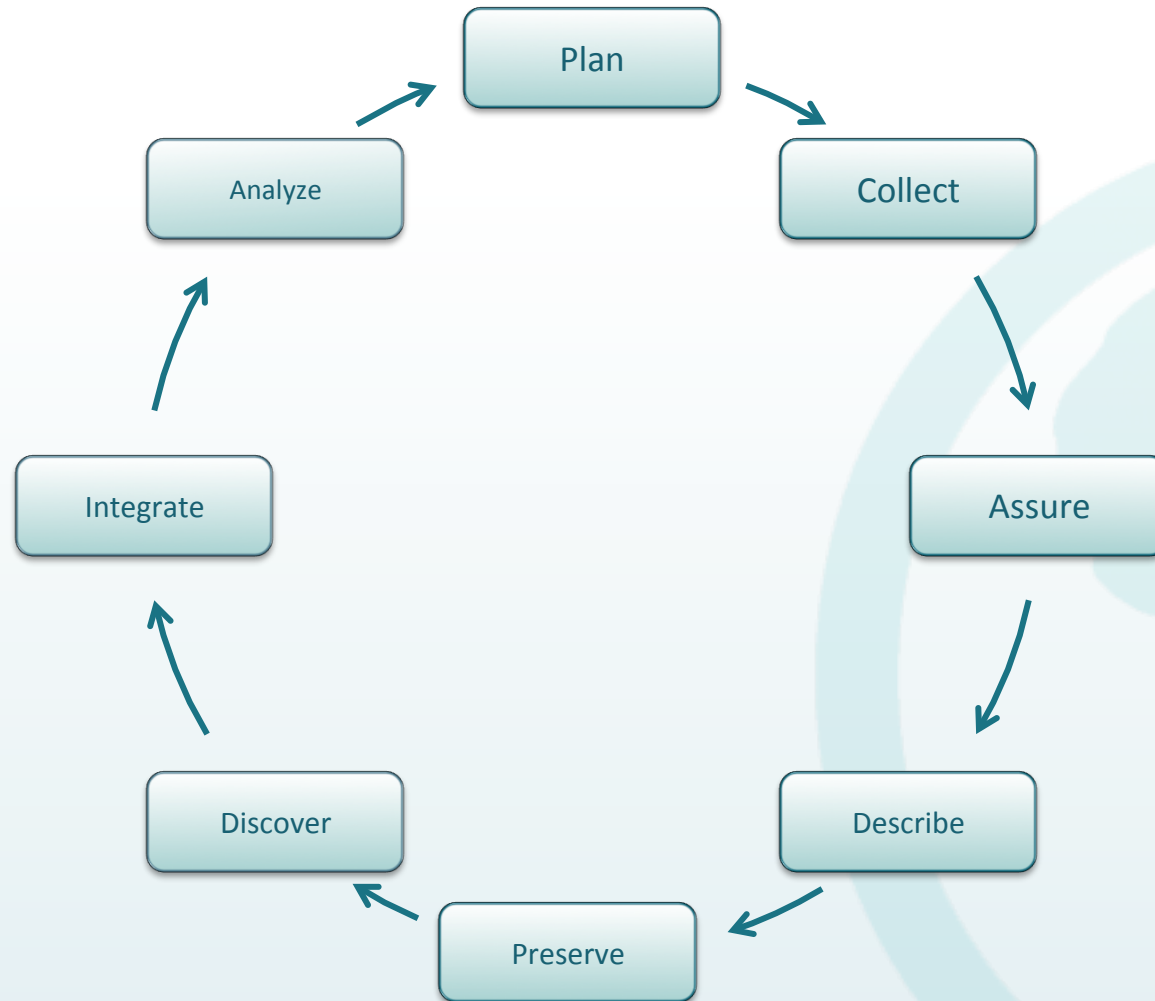
Why California Is Burning

TIME

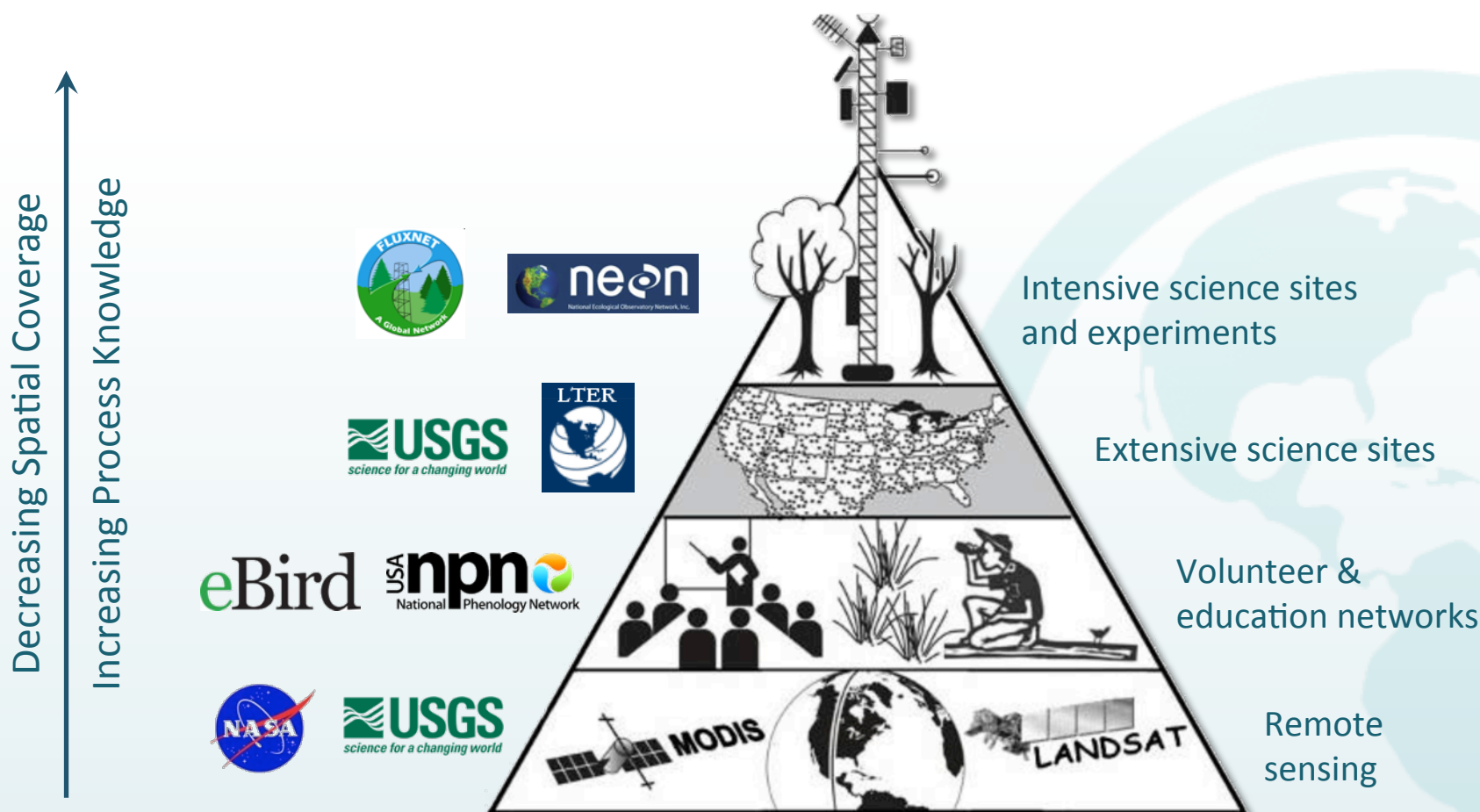
Our Filthy Seas

AAAS

Pressing issues for the digital data lifecycle



Multiple data sources – mutually reinforcing



Adapted from CENR-OSTP

Scattered data sources

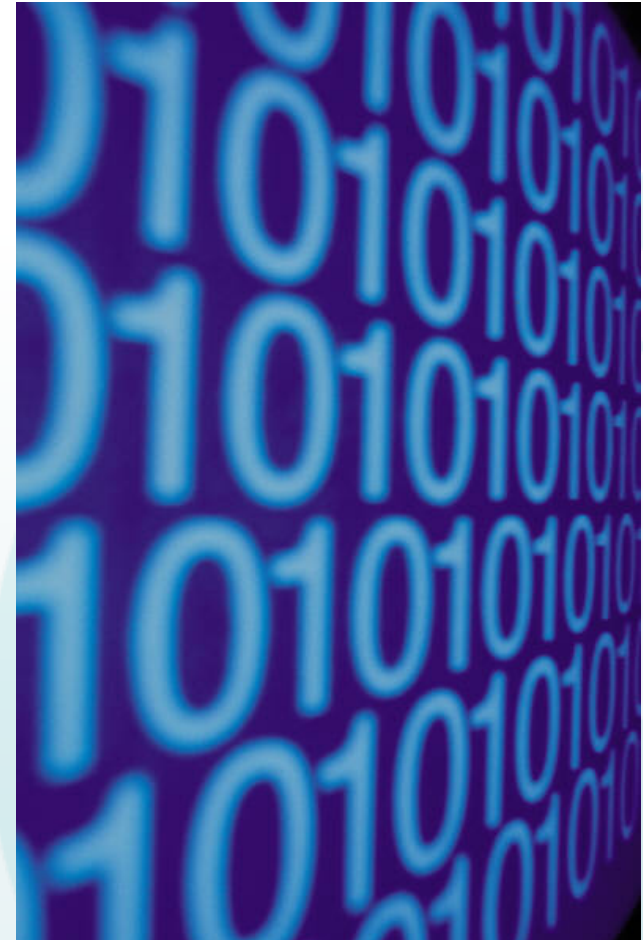
“finding the needle in the haystack”

Data are massively dispersed

- Ecological field stations and research centers (100s)
- Natural history museums and biocollection facilities (100s)
- Agency data collections (100s to 1000s)
- Individual scientists (1000s to 10,000s to 100,000s)

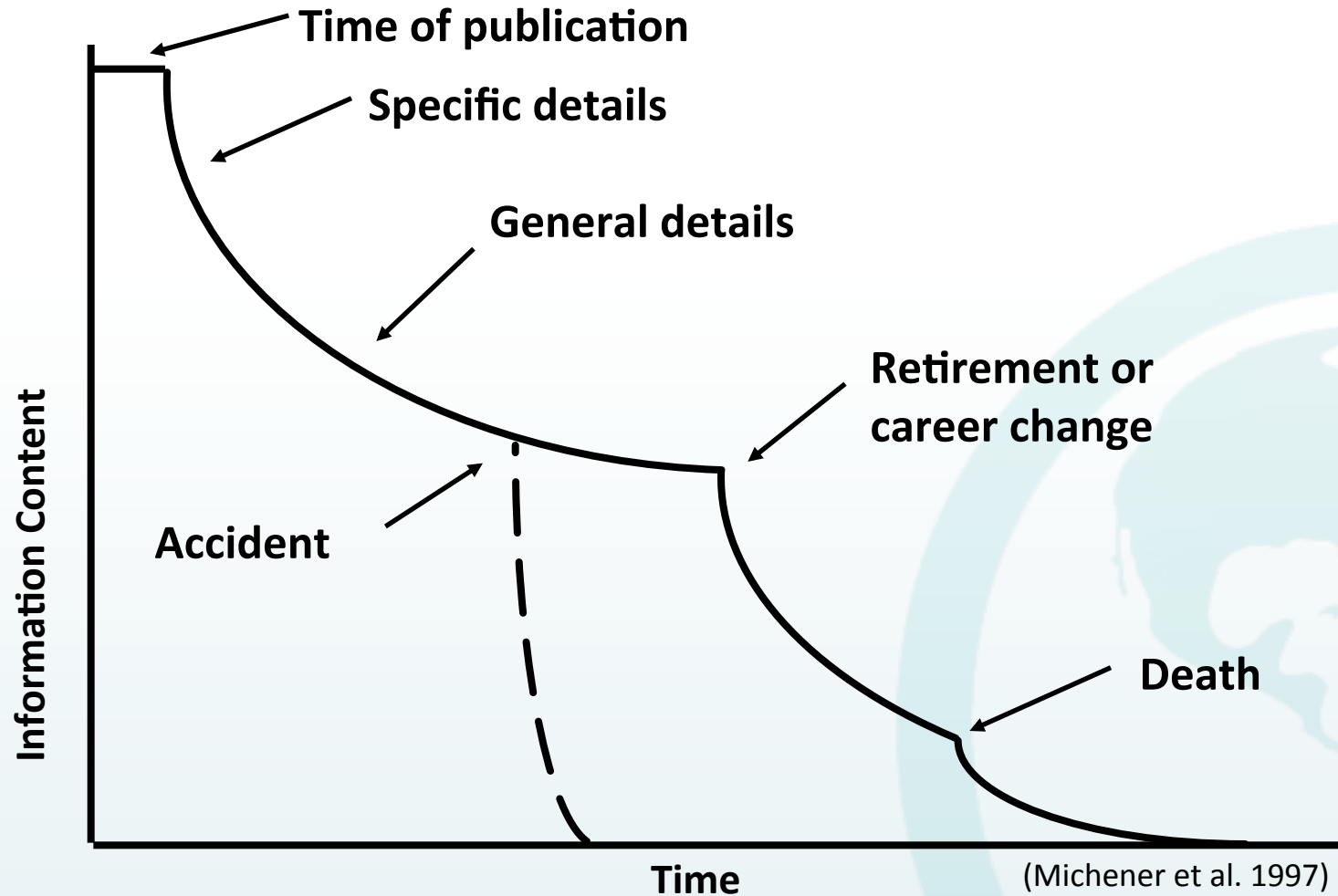


Data Preservation and Planning



Preservation: Poor data practice

“data entropy”

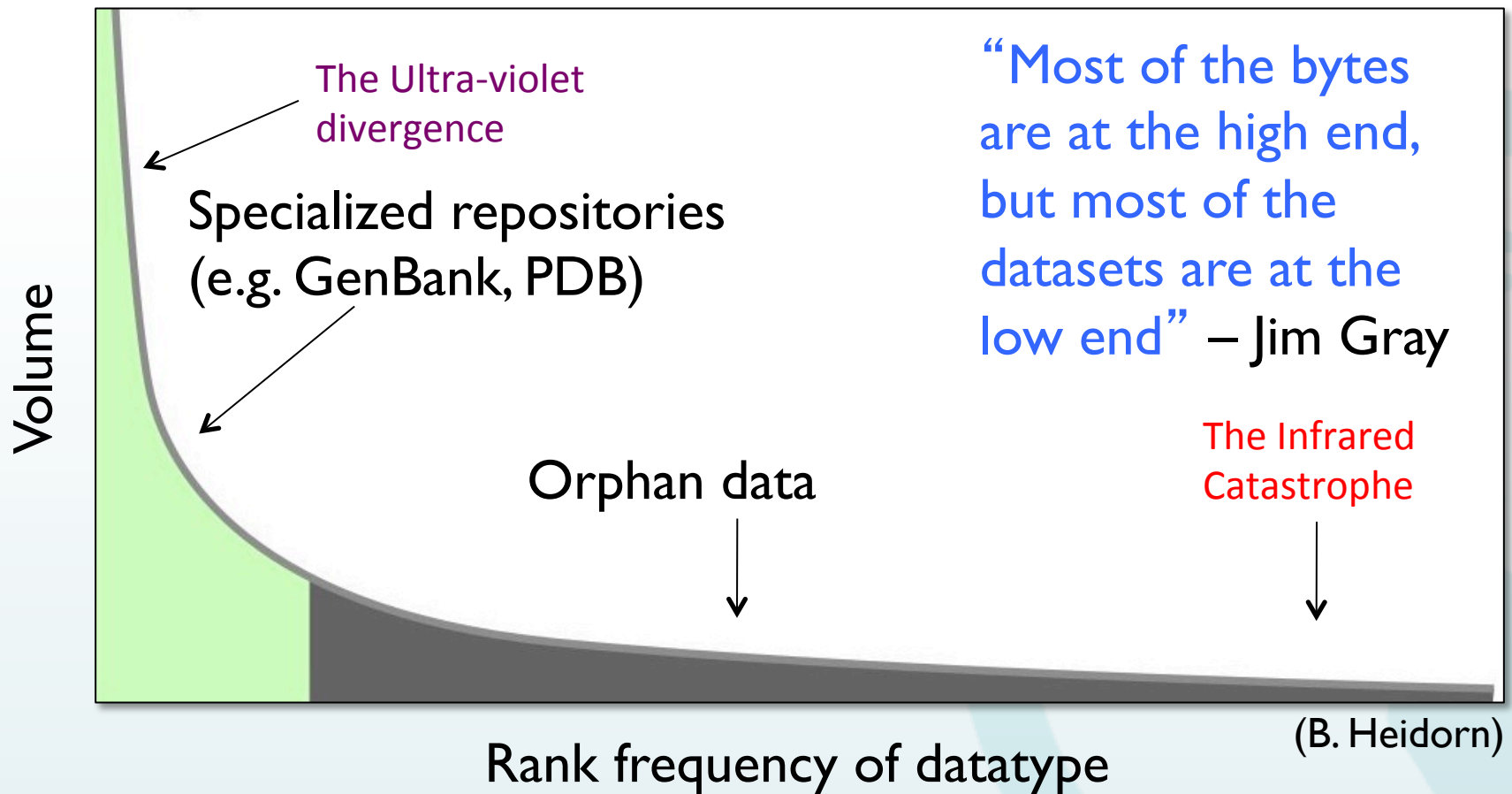


Preservation: Data longevity

Study	Resource Type	Resource Half-life
Rumsey (2002)	Legal Citations	1.4 years
Harter and Kim (1996)	Scholarly Article Citations	1.5 years
Koehler (1999 and 2002)	Random Web Pages	2.0 years
Spinellis (2003)	Computer Science Citations	4.0 years
Markwell and Brooks (2002)	Biological Science Education Resources	4.6 years
Nelson and Allen (2002)	Digital Library Object	24.5 years

Koehler, W. (2004) *Information Research* 9(2): 174.

The Long Tail of Orphan Data



Data deluge and interoperability

“the flood of increasingly heterogeneous data”

Data are heterogeneous

- Syntax
 - (format)
- Schema
 - (model)
- Semantics
 - (meaning)

Study A

METADATA (from EML)	
Study A:	White Mountains
Area col. units:	sq. meter
PIRU	= <i>Picea rubens</i>
BEPA	= <i>Betula papyifera</i>

date	site	species	area	count
10/1/1993	N654	PIRU	2	26
10/3/1994	N654	PIRU	2	29
10/1/1993	N654	BEPA	1	3

Study B

METADATA (from EML)	
Study B:	Green Mountains
Area sampled:	1 sq. meter
picrub	= <i>Picea rubens</i>
betpap	= <i>Betula papyifera</i>

date	site	picrub	betpap
31 Oct 1993	1	13.5	1.6
14 Nov 1994	1	8.4	1.8

Integrated Data

study	date	site	species	density
A	10/1/1993	N654	<i>Picea Rubens</i>	13.0
A	10/3/1994	N654	<i>Picea Rubens</i>	14.5
A	10/1/1993	N654	<i>Betula papyifera</i>	3.0
B	10/31/1993	1	<i>Picea Rubens</i>	13.5
B	10/31/1993	1	<i>Betula papyifera</i>	1.6
B	11/14/1994	1	<i>Picea Rubens</i>	8.4
B	11/14/1994	1	<i>Betula papyifera</i>	1.8

metadata
'promoted'
to become
data

format
normalized
using
metadata

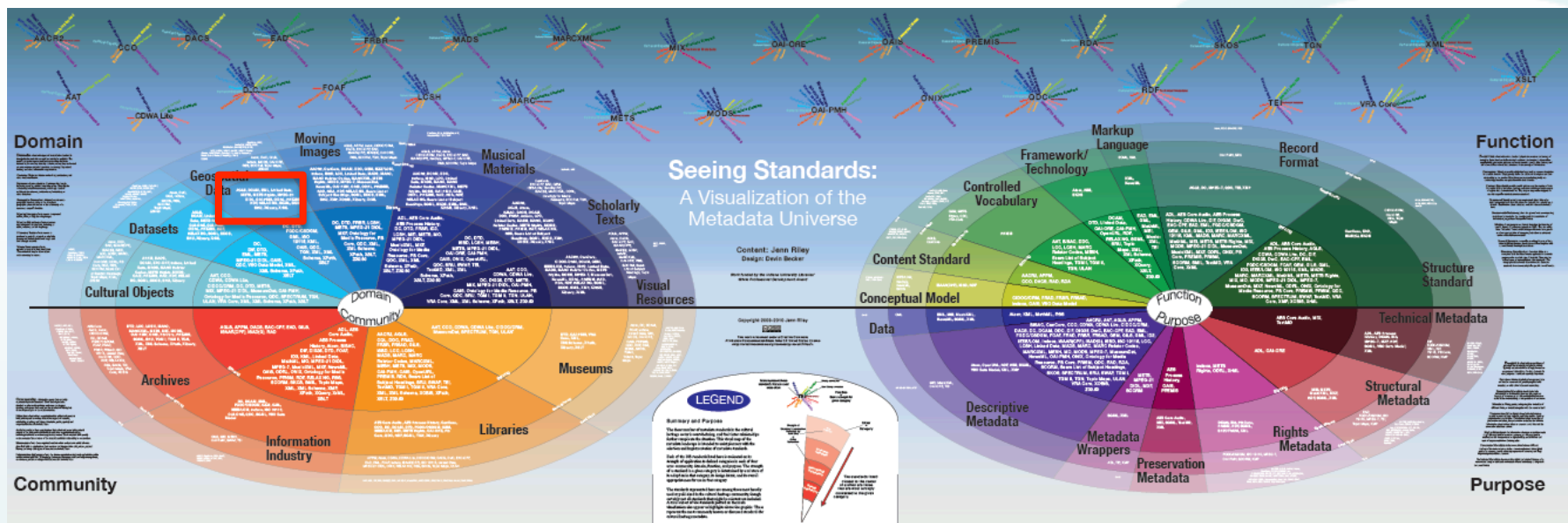
species metadata
from study B
is now data
(picrub/betpap
column headings)

density
calculated
using
metadata

Jones et al. 2007

Metadata universe (multi-verse)

- There are a multitude of metadata standards
- Discipline and sub-discipline specific
- Each with different terms and context



Source: Jenn Riley, Indiana U. Digital Librarian

<http://www.dlib.indiana.edu/~jenrile/metadatamap/> Via John Kunze, Cal. Dig. Lib

Each dot is its own standard !

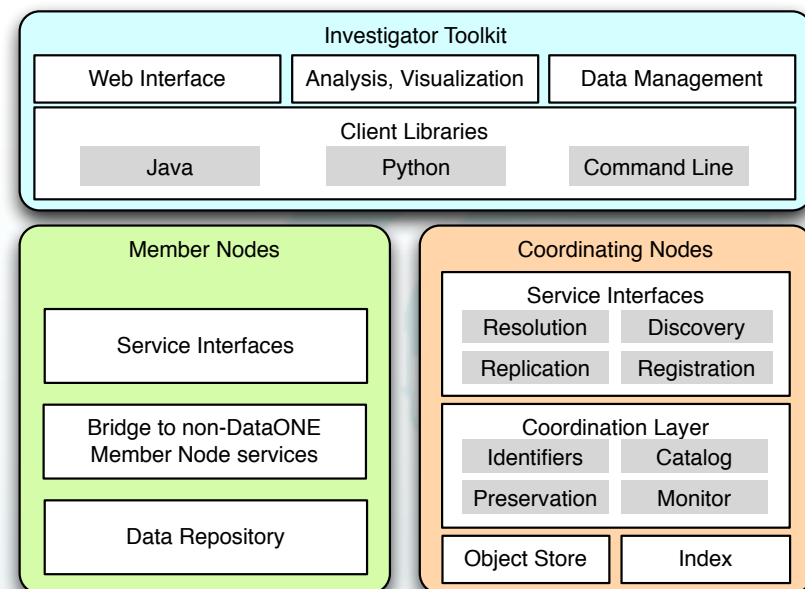
Geospatial Data

AGLS, DCAM, EML, Linked Data,
METS, METS Rights, MPEG-21
DIDL, OAI-PMH, ODRL, PREMIS
RDF, RELAX NG, SGML, SKOS
SRU, XQuery, XrML

“...billions and billions of worlds ...” – Carl Sagan

DataONE CI architectural Elements

- Hard-core cyberinfrastructure (CI)
 - CI Member Node (MN) data repositories
 - Coordinating Node (CN) global metadata repo's
 - Simple, but powerful REST API/SPI for universal access
 - Investigator toolkit (ITK) software tools to allow access to the data repository collective via familiar access idioms
- Cultural and wetware issues
 - Educational Materials
 - Best practices
 - Workshops and tutorials
 - Surveys and assessments
 - Scientist, policymaker, citizen engagement
 - Collaboration, governance, and sustainability

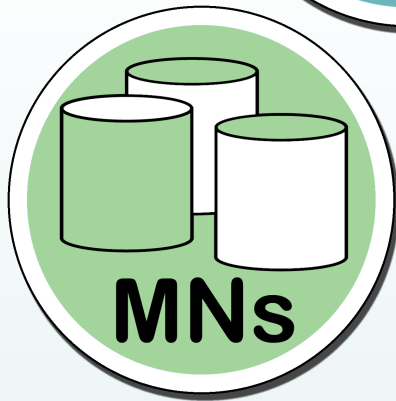


<http://mule1.dataone.org/ArchitectureDocs-current/>

A User's View

Depositing Data with DataONE

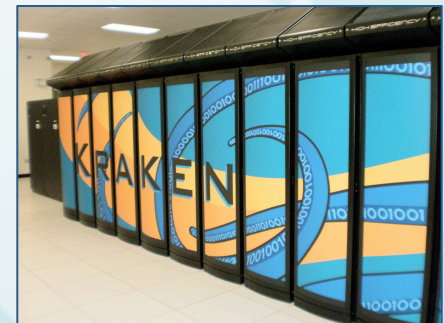
Key Cyberinfrastructure Elements



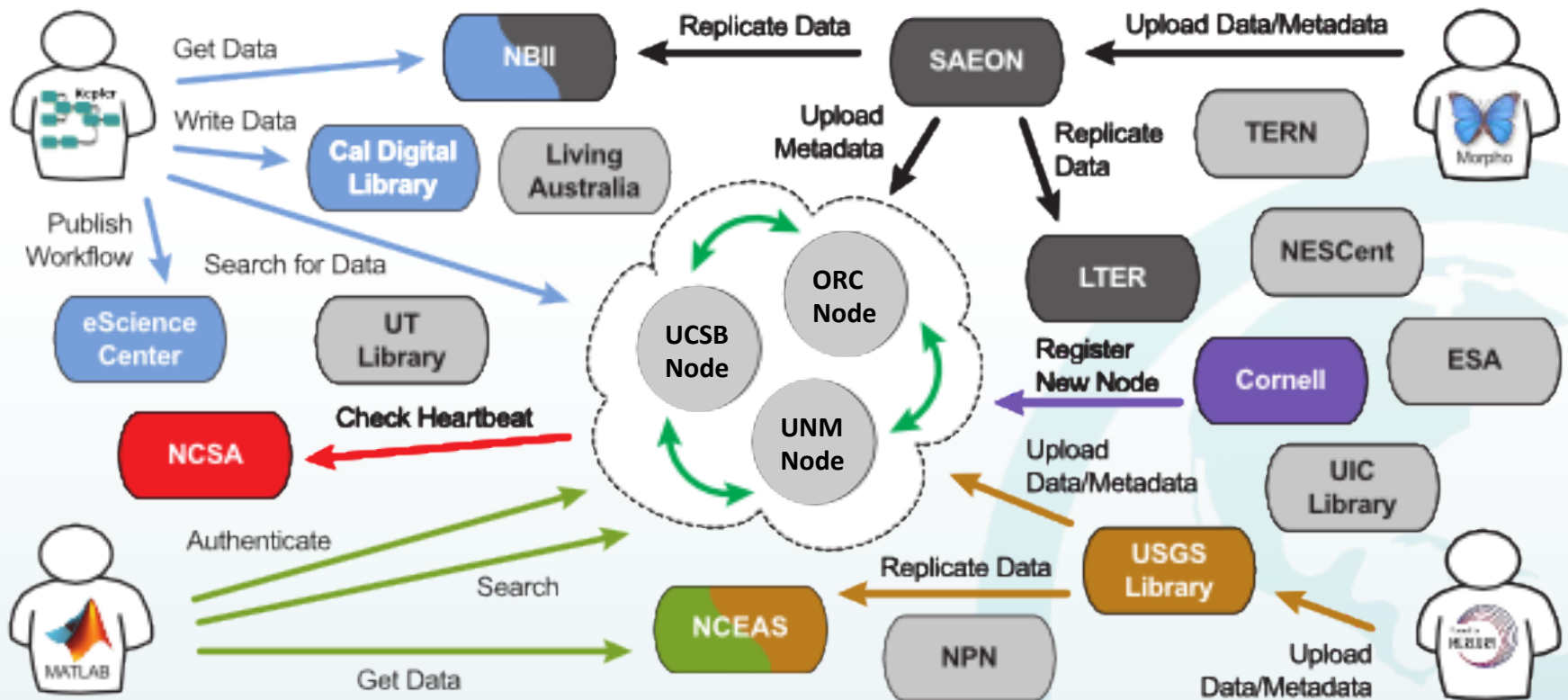
- Unique identifiers
- Search and deliver
- Replication
- Federated identity

Usable by
People and their Agents

Interoperability



Supporting the data lifecycle



The data lifecycle



1. Deposition/acquisition/ingest
2. Curation and metadata management
3. Protection, including privacy
4. Discovery, access, use, and dissemination
5. Interoperability, standards, and integration
6. Evaluation, analysis, and visualization



DataONE Supports Data Preservation

Three major components for a flexible, scalable, sustainable network

Member Nodes

Coordinating Nodes

Investigator Toolkit



DataONE satisfies arch requirements

- Enables integration of multiple geographically diverse and metadata diverse repositories
- Presents collective search results across multiple repository
- Provides a unified API/SPI for search and programmatic interface
<http://mule1.dataone.org/ArchitectureDocs-current/>
- DataONE content has unique identifiers (DOI's) for referencable/citable data objects
- Supports both **large datasets** and the **long-tails**

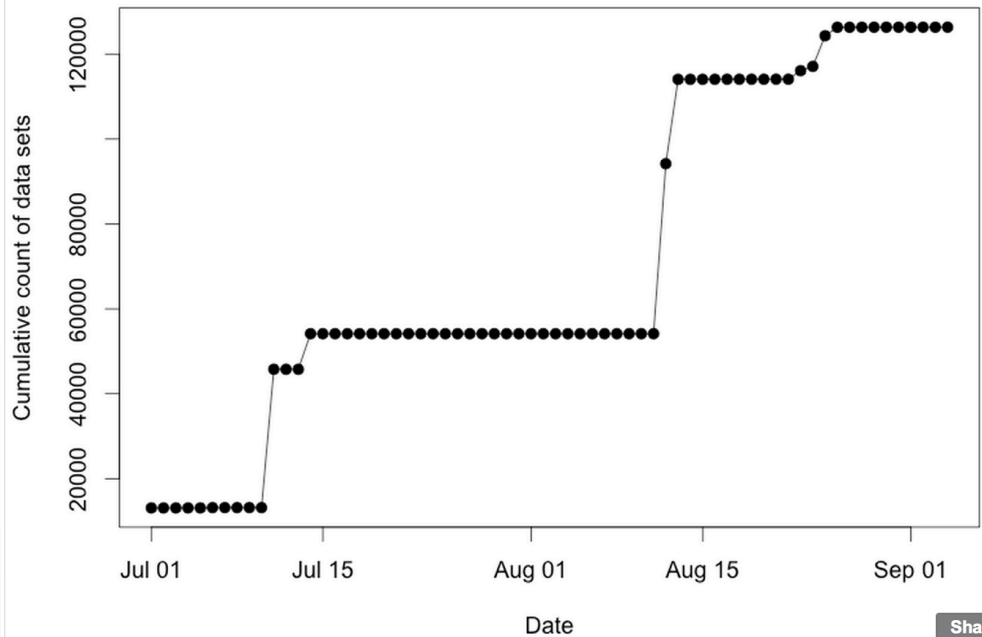
DataONE spurs innovation

- Enables new analysis and synthesis efforts by integrating tasks across repositories
- Provides means for data replication and basis for repositories to build “data wills” or “data trust” plans
- Provides a platform to develop advanced interoperable workflow tools and semantic integration tools

DataONE: current state/recent progress



DataONE: Supporting Scientific Data Preservation, Discovery, and Innovation



Current Member Nodes:



Coming Soon:



Current Tools: DMPTool



Tools Coming Soon:



Data Management Planning Tool



DMP TOOL

Build your data management plan

<https://dmp.cdlib.org/>


[Contact Us](#) | [Sign up](#) | [Login](#)

[Home](#) | [About DMP Tool](#) | [DMP News](#) | [My Plans](#) | [Help](#)



Create ready to use data management plans for specific funding agencies

Sign up and start building your data management plan now!



See a plan created with the DMP Tool

Recent DMP News

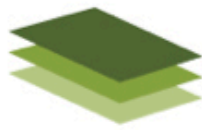
- [Open Access and Climate Research Data](#)
- [Data, Data Everywhere...A Deluge of Data Management Articles](#)
- [University of Illinois at Urbana-Champaign joins DMP Tool partners](#)
- [Funder X now available in DMP Tool](#)

[more news >](#)

The DMP Tool allows you to:

- Meet funder requirements for data management plans.
- Get step by step instructions and guidance for your data management plan as you build it.
- In many cases, get institution specific advice and assistance.

copyright 2011
[Privacy statement](#) | [Terms of use](#)



NSF-GEN: Generic: Plan description

The NSF-GEN: Generic plan will cover the subject areas listed to the left.

You can save a plan in progress and return later to finish or edit.

Progress

Click on a section below to edit it at any time.

● = complete

Plan description

- 1. Types of data produced
- 2. Data and metadata standards
- 3. Policies for access and sharing
- 4. Policies for re-use, redistribution
- 5. Plans for archiving & preservation

Plan Name: (required)

Please give your plan a name to help you identify it in the future

Solicitation Number:

Comment:

Provide any notes you want to appear on your My Plans page. This will not appear in the document

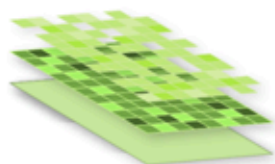
B I U abc x₂ x² ABC ↶ ↷ ↶ ↷ ☰ ☰ ☰ ☰ ☰

Resources

General

[NSF Data Sharing Policy](#)

[NSF Data Management Plan Requirements](#)



DMP Tool

Build your data management plan

Progress

Sections marked with a check are complete. You can navigate to a section and edit at any time.

NSF-GEN: Generic

Cover page

- ✓ 1. [Types of data produced](#)
- ✓ 2. [Data and metadata standards](#)
- ✓ 3. [Policies for access and sharing](#)
- ✓ 4. [Policies for re-use, redistribution](#)
- ✓ 5. **[Plans for archiving & preservation](#)**

Plans for archiving data, samples, and other research products, and for preservation of access to them.

Suggested answer text; copy and paste as needed:

As advised by University of Virginia Library staff members, I plan on depositing my research data in the UVA institutional repository – Libra. I will submit the necessary metadata and other resources to make my data accessible for future users. In accordance with the University of Virginia policy RES-002, "Policy: Laboratory Notebook and Recordkeeping," the data will be preserved for a minimum of five years upon completion of the project. However the current preservation plan for Libra will be to preserve the data indefinitely. The Libra backup plan provides for data redundancy including off-site storage.

As advised by University of Virginia Library staff members, I plan on depositing my research data in the UVA institutional repository – Libra. I will submit the necessary metadata and other resources to make my data accessible for future users. In accordance with the University of Virginia policy RES-002, "Policy: Laboratory Notebook and Recordkeeping," the data will be preserved for a minimum of five years upon completion of the project. However the current preservation plan for Libra will be to preserve the data indefinitely. The Libra backup plan provides for data redundancy including off-site storage.

Resources

University of Virginia

[UVA Scientific Data Consulting Group](#)

[Archiving & Sharing Data Guidance](#)

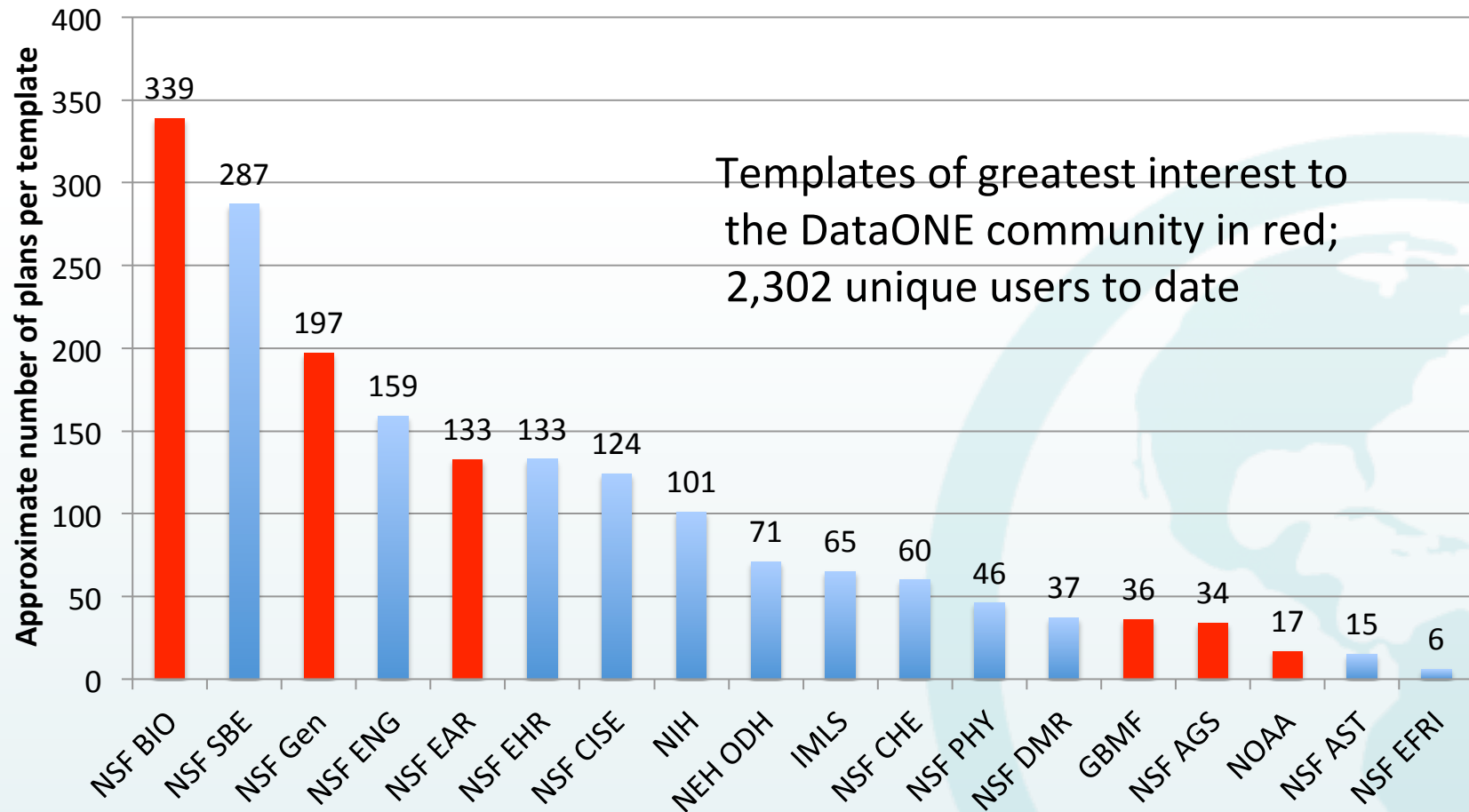
[UVA Policy RES-002: "Laboratory Notebook and Recordkeeping"](#)

General

[NSF Data Sharing Policy](#)

[NSF Data Management Plan Requirements](#)

Plans per template (as of June 2012)



	A	B	C	D	E	F	G	H	I
1			amount foot	S4T15_R1	S4T15_R2	S4T15_R3			S8T15_R1
2				copas	naups				copas
3									
12	23-Oct	30	180		650		750		650
13	25-Oct	25	400		780		550		400
14	27-Oct	25			500	100	1200		400
15	29-Oct	35	300	750	350	200	450		200
16	31-Oct	30	340	800	300	300	1450		700
17	2-Nov	25	292	1550	660	300	1000		520
18	4-Nov	25	360	440	200+	260	1050		700
19	6-Nov	30	-	400	-	250	550	0	750
20	8-Nov	30		300		280	300	0	450
21	10-Nov	30	~100	160	~100	80	450	1450	350
22	12-Nov	35	<50	60	<50	100	300	800	550
23	14-Nov	30	<10	100	<50	20	200	600	550
24	16-Nov	35	<10	30	<30	0	140	2320	450
25	18-Nov	30	0	0	<20	0	125	400	150
26	20-Nov	25	0	0	<20	40	150	280	250
27	22-Nov	25			<10	0	120	120	50
28	24-Nov	25			<10	0	125	150	150
29	26-Nov	25			<5	0	50	30	0
30	28-Nov	25			<5	0	<50	0	0
31	30-Nov	25			0	0	<25	0	0
2-Dec	25						<10	0	0
4-Dec	25						0	0	
6-Dec									
8-Dec									



DataUp

- ✓ Check for best practices
- ✓ Create metadata
- ✓ Connect to ONEShare

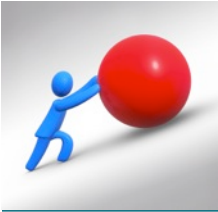
ONEShare

	A	B	C	D	E	F	G	H	I	J	K
1	Treatment	TreatNum	Replicate	TimeStep	N1	N2	N3	NA	N5	N6	C1
2	54715	1	1	1	1.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
3	54715	1	1	2	2.00E-01	8.00E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00	0.00E+00
4	54715	1	1	3	0.00E+00	6.00E-01	4.00E-01	0.00E+00	0.00E+00	0.00E+00	0.00E+00
5	54715	1	1	4	0.00E+00	0.00E+00	0.00E+00	6.00E-01	4.00E-01	0.00E+00	0.00E+00
6	54715	1	1	5	0.00E+00	0.00E+00	0.00E+00	0.00E+00	8.00E-01	2.00E-01	0.00E+00
7	54715	1	1	6	0.00E+00	0.00E+00	0.00E+00	0.00E+00	2.00E-01	4.00E-01	4.00E-01
8	54715	1	1	7	0.00E+00	0.00E+00	0.00E+00	2.00E-01	0.00E+00	2.00E-01	4.00E-01
9	54715	1	1	8	0.00E+00	2.00E-01	0.00E+00	2.00E+00	2.00E-01	0.00E+00	2.00E-01
10	54715	1	1	9	0.00E+00	0.00E+00	0.00E+00				
11	54715	1	1	10	0.00E+00	0.00E+00	0.00E+00				
12	54715	1	1	11	0.00E+00	0.00E+00	0.00E+00				
13	54715	1	1	12	0.00E+00	0.00E+00	0.00E+00	1			

```

<?xml version="1.0" encoding="UTF-8"?>
<xml:stylesheet href="http://www.jscimed.org/jscimed/xsl/ncml-schema.xsl" type="text/xsl"/>
<ncml:ncml xmlns="http://www.jscimed.org/ncml/" xmlns:xs="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.jscimed.org/ncml/ http://www.jscimed.org/ncml/ncml-schema.xsl">
  <ncml:header>
    <ncml:headerItem name="Title" value="JSCI Example" />
  </ncml:header>
  <ncml:body>
    <ncml:bodyItem name="Person" value="Joe" />
    <ncml:bodyItem name="Address" value="Marine Science Institute" />
    <ncml:bodyItem name="City" value="Santa Barbara" />
    <ncml:bodyItem name="State" value="CA" />
    <ncml:bodyItem name="PostalCode" value="93024" />
    <ncml:bodyItem name="Country" value="USA" />
    <ncml:bodyItem name="Phone" value="(310) 206-1984" />
    <ncml:bodyItem name="Email" value="jsci@msi.ucsb.edu" />
    <ncml:bodyItem name="Creator" value="Joe" />
    <ncml:bodyItem name="AssociatedParty" value="Joe" />
    <ncml:bodyItem name="SurName" value="Joe" />
    <ncml:bodyItem name="DeliveryPoint" value="Marine Science Institute" />
    <ncml:bodyItem name="City" value="Santa Barbara" />
    <ncml:bodyItem name="State" value="CA" />
    <ncml:bodyItem name="PostalCode" value="93024" />
    <ncml:bodyItem name="Country" value="USA" />
    <ncml:bodyItem name="Address" value="(310) 206-1984" />
  </ncml:body>
</ncml:ncml>
  
```

GENERAL METADATA	
First name	Carly
Last name	Strasser
Address	415 20th St
City	Oakland
State/province	CA
Postal code	94612
Country	USA
Phone	510-967-0179
Organization	California Digital Library
Title of dataset*	Copepodlog_2012-06
Keyword(s)	June 25 2012
Abstract	
Repository name and contact information	ONEShare
Data Contact Person: First name	
Data Contact Person: Last name	
Data Contact Person: Address	
Data Contact Person: City	
Data Contact Person: State/province	
Data Contact Person: Postal code	
Data Contact Person: Country	
Data Contact Person: Phone	
Data Contact Person: Email	
Data Contact Person: Organization	
Today's date*	
Keyword/thesaurus used	
Geographic coverage: Description	
Geographic coverage: West bounding coordinate	
Geographic coverage: East bounding coordinate	
Geographic coverage: North bounding coordinate	
Geographic coverage: South bounding coordinate	
Temporal coverage: Beginning date	
Temporal coverage: Ending date	
Project title	
Project description	
Funding	
Intellectual rights	
Data table name	



2. Data Discovery

Google

soil organic carbon



Search

About 1,700,000 results (0.29 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Santa Fe, NM

Change location

Show search tools

Ad related to soil organic carbon

Why this ad?

[Organic Soil | Lowes.com](#)

www.lowes.com/ - ★★★★★ 168 seller reviews

Find Gardening Supplies At Lowe's® Official Site Today. Shop Now!

1,471 people +1'd Lowe's Home Improvement

+ [Show map of 3458 Zafarano Dr, Santa Fe, NM](#)

[Scholarly articles for soil organic carbon](#)

[Total carbon, organic carbon, and organic matter](#) - Nelson - Cited by 6671

[Soil carbon pools and world life zones](#) - Post - Cited by 1428

[The vertical distribution of soil organic carbon and its ...](#) - Jobbágy - Cited by 935

[Soil carbon - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Soil_carbon

Soil organic matter, of which carbon is a major part, holds a great proportion of ...

Tillage and drainage both expose **soil organic matter** to oxygen and oxidation.

↳ [Overview - Soil carbon and soil health](#) - [Losses of soil carbon](#)

[Soil organic carbon](#)

www.eoearth.org/article/Soil_organic_carbon

by E Milne - [Related articles](#)

Dec 21, 2009 – **Soils** contain **carbon** (C) in both **organic** and inorganic forms. In most

soils (with the exception of calcareous **soils**) the majority of C is held as ...

[PDF] [The importance of soil organic matter](#)

<ftp://ftp.fao.org/agl/agll/docs/sb80e.pdf>

File Format: PDF/Adobe Acrobat - [Quick View](#)

by VT di Caracalla - [Related articles](#)

Human interventions that influence **soil organic matter**. 15. Practices that ... Evaluation of the organic matter content of a soil in Paraná. 20. 9. Reduction of dry ...

9) [Soil Organic Matter \(SOM\)](#) Department of Soil Science



The DataONE Federation



ORNL DAAC as a DataONE Member Node



Search For:

Results/Page

10

SEARCH

Hint: boolean operators 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 during thru

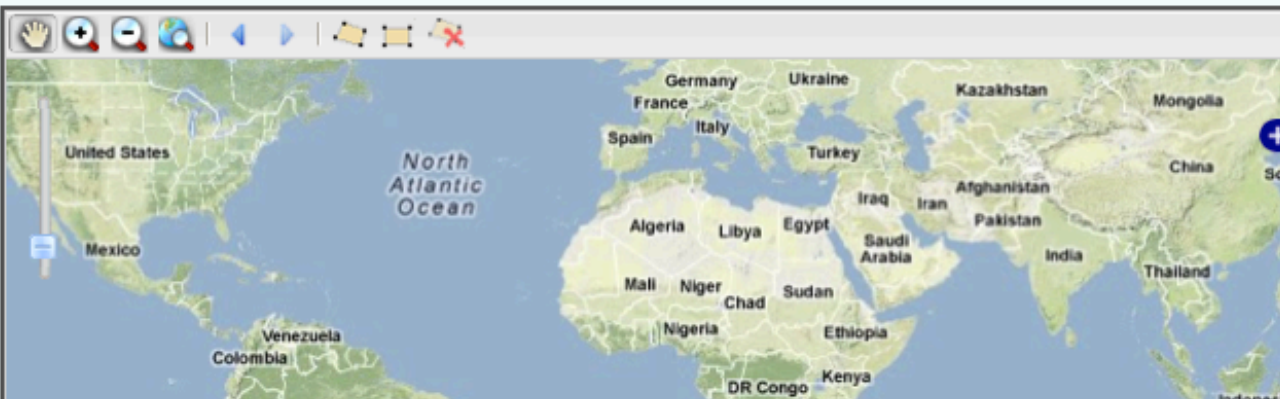
Publication Date

Either

mm/dd/yyyy mm/dd/yyyy

[Help](#) | [clear](#)

Geographic Search



List Areas in:

USA WORLD

Select from list

Search Area:

overlaps encloses

North

[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
Margaret McManus (3465)	Florida Coastal Everglades (512)	Oceanographic Sensor Data (15359)	Studies of Coastal Oceans (PISCO) (14777)
Libe Washburn (3112)	Olympic Coast National (510)	continental shelf (15359)	PISCO (1309)
Jack Barth (1612)	Arctic Long-Term Ecological (339)	Temperature (15275)	Monterey Bay National Marine Sanctuary (952)
Mary Sue Brancato (510)		United States of America (15146)	Olympic Coast National Marine
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 \(2\)](#)

[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
Margaret McManus (3465)	Florida Coastal Everglades (512)	Oceanographic Sensor Data (15359)	Studies of Coastal Oceans (PISCO) (14777)
Libe Washburn (3112)	Olympic Coast National (510)	continental shelf (15359)	PISCO (1309)
Jack Barth (1612)	Arctic Long-Term Ecological (339)	Temperature (15275)	Monterey Bay National Marine Sanctuary (952)
Mary Sue Brancato (510)		United States of America (15146)	Olympic Coast National Marine
Anne Giblin (298)			
Pete Raimondi (237)			

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

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 \(2\)](#)

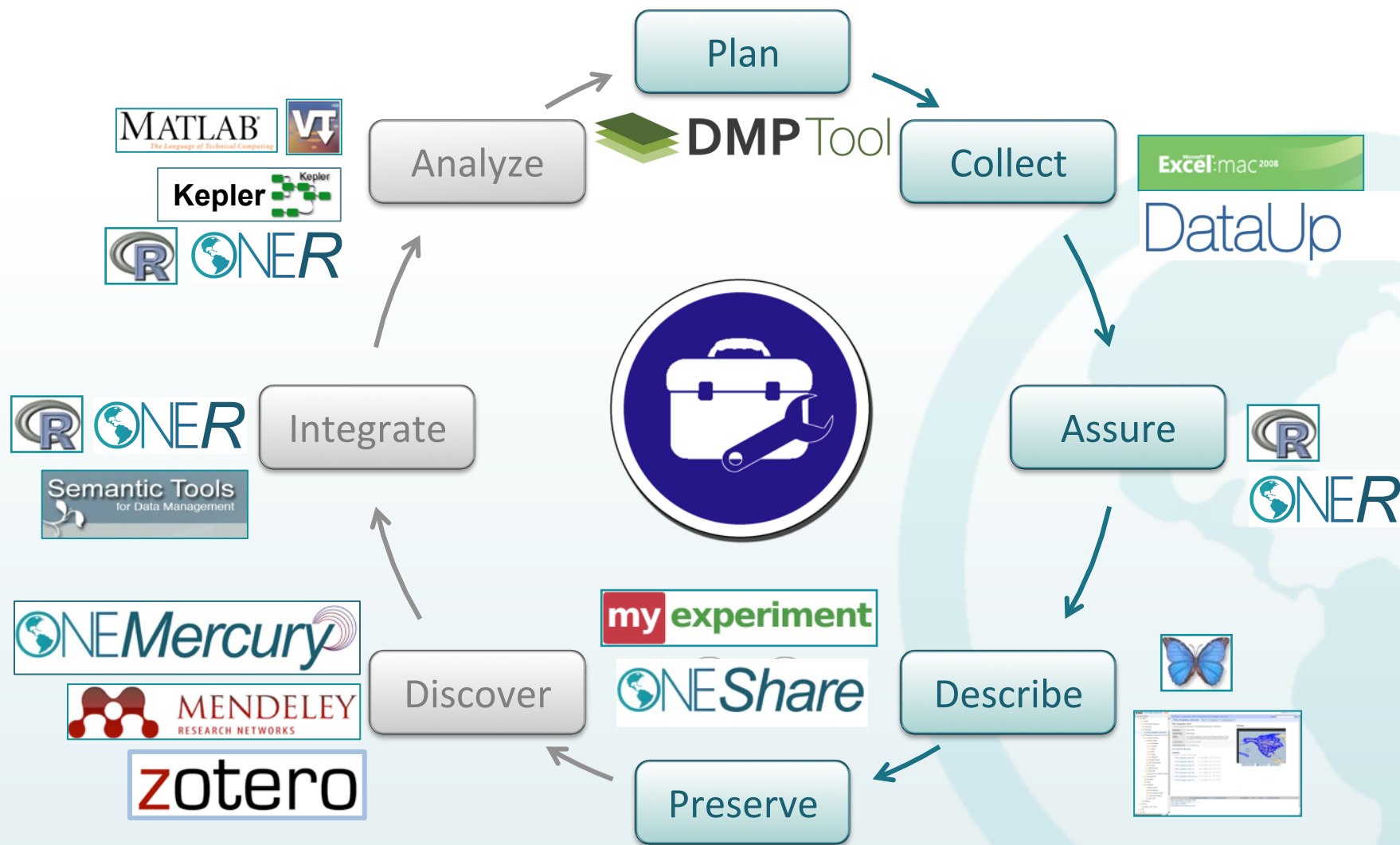
1 - 10 out of 22815

[6](#) [7](#) [8](#) [9](#) [10](#) [Next](#)

[Data Files \(0\)](#)



Investigator Toolkit Support



Exploration, Visualization, and Analysis



eBird



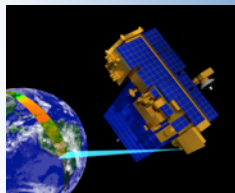
Land Cover



Meteorology

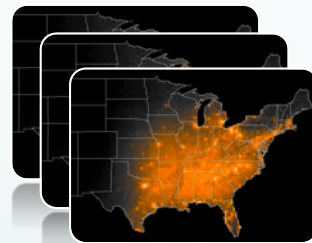


MODIS – Remote sensing data



The Cornell Lab of Ornithology

Diverse bird observations and environmental data from 300,00 locations in the US integrated and analyzed using High Performance Computing Resources



$$F(X, s, t) = \frac{1}{n(s, t)} \sum_{i=1}^n f_i(X, s, t) I(s, t \in \theta_i)$$

Spatio-Temporal Exploratory Model identifies factors affecting patterns of migration

DataONE

Model results

Occurrence of Indigo Bunting (2008)



- Examine patterns of migration
- Infer how climate change may affect bird migration



Public Participation in Scientific Research Conference: 4-5 August 2012 in Portland, Oregon USA prior to Ecological Society of America meeting (6-10 Aug.):

<http://www.birds.cornell.edu/citscitoolkit/conference/2012>



User Assessments

OPEN ACCESS Freely available online



Data Sharing by Scientists: Practices and Perceptions

Carol Tenopir^{1*}, Suzie Allard¹, Kimberly Douglass¹, Arsev Umur Aydinoglu¹, Lei Wu¹, Eleanor Read², Maribeth Manoff², Mike Frame³

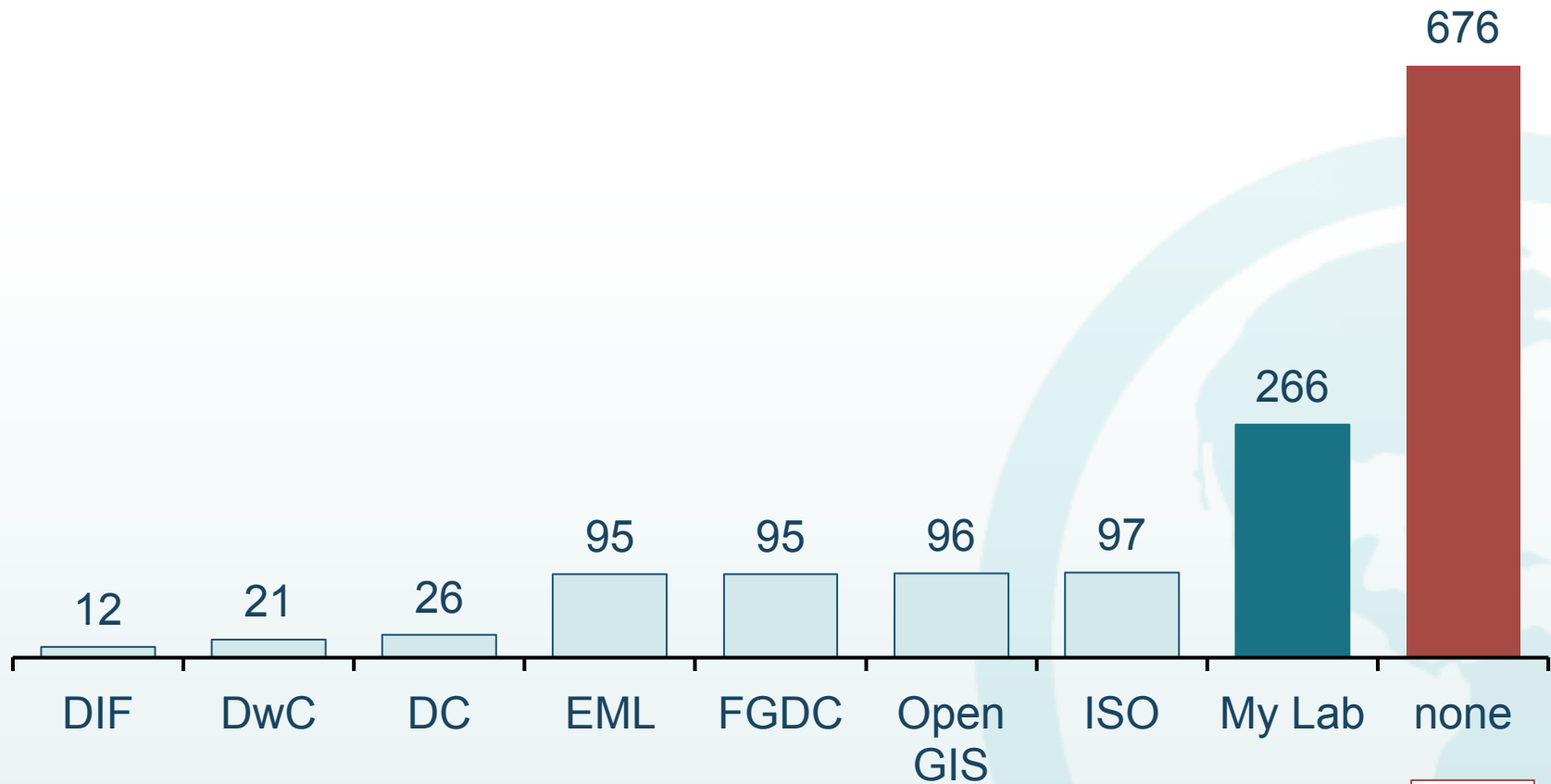
¹ School of Information Sciences, University of Tennessee, Knoxville, Tennessee, United States of America, ² University of Tennessee Libraries, University of Tennessee, Knoxville, Tennessee, United States of America, ³ Center for Biological Informatics, United States Geological Survey, Oak Ridge, Tennessee,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

Abstract

Background: Scientific research in the 21st century is more data intensive and collaborative than in the past. It is important to study the data practices of researchers – data accessibility, discovery, re-use, preservation and, particularly, data sharing. Data sharing is a valuable part of the scientific method allowing for verification of results and extending research from prior results.

Methodology/Principal Findings: A total of 1329 scientists participated in this survey exploring current data sharing practices and perceptions of the barriers and enablers of data sharing. Scientists do not make their data electronically available to others for various reasons, including insufficient time and lack of funding. Most respondents are satisfied with their current processes for the initial and short-term parts of the data or research lifecycle (collecting their research data; searching for, describing or cataloging, analyzing, and short-term storage of their data) but are not satisfied with long-term data preservation. Many organizations do not provide support to their researchers for data management both in the short-

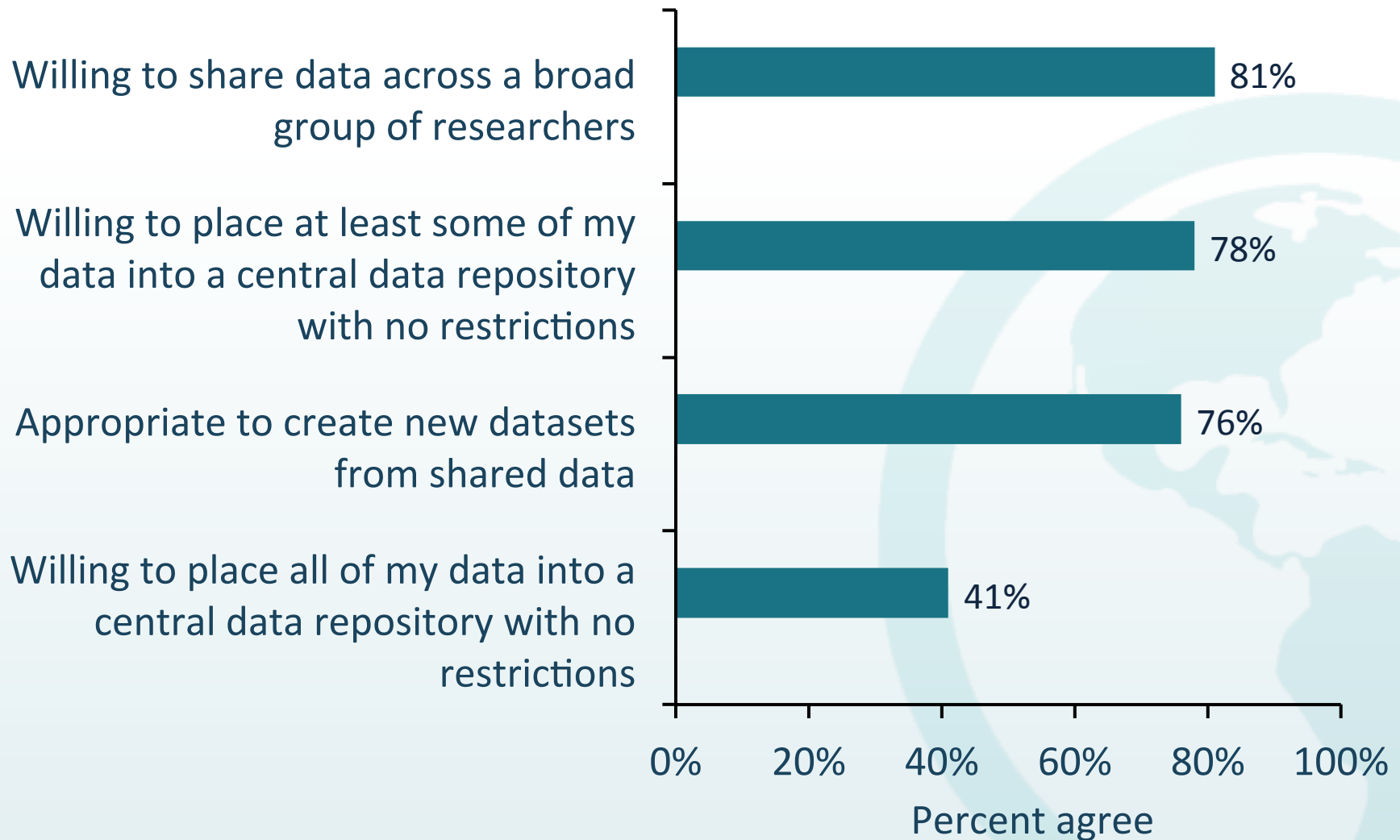
What standard do you currently use?











Metadata language



Many are interested in sharing data



User Matrix

	Data Service	Investigator ToolKit	Data Management Planning	Best Practices	Tools Database	Training	Curricula
Scientist							
Data Librarians							
Ecological Modeler							
Resource Manager							

Community Engagement



Search For



- About
- Participate
- Resources
- Education
- Data

Education

- Training Activities
- Education Modules
- Graduate Courses

Find it Fast

- Data Management Planning
- Best Practices
- Software Tools

Home » Education » Education Modules

Education Modules

Below are links to education modules in powerpoint format that you can download and incorporate into your teaching materials.

The topics covered include:

- Lesson 01: Why Data Management
- Lesson 02: Data Sharing
- Lesson 03: Data Management Planning
- Lesson 04: Data Entry and Manipulation
- Lesson 05: Data Quality Control and Assurance
- Lesson 06: Data Protection and Backups
- Lesson 07: Metadata
- Lesson 08: How to Write Good Quality Metadata
- Lesson 09: Data Citation
- Lesson 10: Analysis and Workflows

If you use or consider using these materials, we would be grateful if you would take the opportunity to provide *feedback*.

Credits: Heather Henkel, Viv Hutchison, Carly Strasser, Stacy Rebich Hespanha, Kristin Vanderbilt, Lynda Wayne

Best Practices and Software Tools

The image displays two overlapping screenshots of the DataONE website. The top screenshot shows the 'Best Practices' page, and the bottom screenshot shows the 'Software Tools Catalog' page. Both pages feature a search bar, navigation menu, and various resource links.

Top Screenshot: Best Practices Page

- Search:** DataONE Website
- Navigation:** About, Participate, Resources, Education, Data
- Page Title:** Best Practices
- Content:** The DataONE through all stage of the For student Practices d The develop
- Resources:** ONEMercury, Investigator Toolkit, Data Management Planning, Best Practices, Software Tools Catalog, Publications
- Tags:** provenance, documentation, metadata, access, assure plan, describe, format, preserve, analyze data, archives, quality
- Featured Practice:** Plan for effective multimedia management
- View All:** Multimedia data present

Bottom Screenshot: Software Tools Catalog Page

- Search:** DataONE Website
- Navigation:** About, Participate, Resources, Education, Data
- Page Title:** Software Tools Catalog
- Content:** Home » Resources » Software Tools Catalog. The Software Tools database is the product of two NSF-funded Informatics Education Planning Workshops hosted by DataONE. The database provides a brief description of a wide range of tools that are recommended for use by scientists and students, as well as additional information and links to further resources. Users can access tools within the database by selecting keywords (under advanced search) or using free search. The development of the DataONE Software Tools database was a collaborative effort across many individuals (credits).
- Resources:** ONEMercury, Investigator Toolkit, Data Management Planning, Best Practices, Software Tools Catalog, Publications
- Tags:** analyze, map, graphics, metadata, editor, statistics, web 2.0, metadata, database, GIS, visualization, models, geospatial
- View All Software Tools:** Search Software Tools, Search software tools, Search
- Advanced Search:** Advanced Search

Resources

ONEMercury
Investigator Toolkit
Data Management Planning
Best Practices
Software Tools Catalog
Publications

Tags

data archives

describe access

format preserve

documentation

quality metadata

assure analyze provenance

plan

[More](#)

Featured Practice

Create a data dictionary

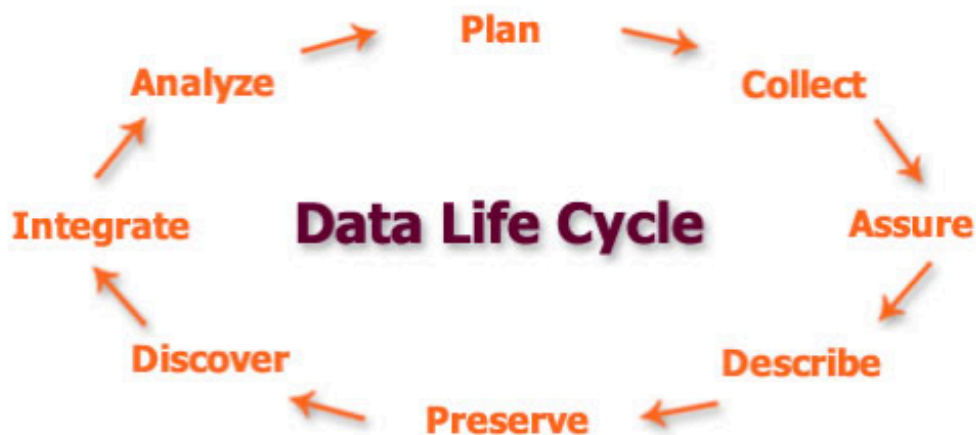
A data dictionary provides a detailed description for each element or variable in your dataset and data model. Data dictionaries are used to document important and useful information such as a descriptive name, the data type, allowed values, units, and text description.

Best Practices

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

For students and others new to data management, we provide a **Best Practices Primer** as an introduction to the DataONE Best Practices database and data management in general.

The development of the DataONE Best Practices database was a collaborative effort across many individuals ([credits](#)).



View All Best Practices

Search Best Practices

Search

DataONE: Next steps

- Member node growth
 - Number of member nodes
 - Increase the number and size of data sets
 - Sustainably
 - In terms of resource needs form MN's
 - In terms of resource demands on DataONE
- New Investigator toolkit tools (strategically)
- An increasing number of science use cases with more breakthrough science
- Also, re-purposing DataONE CI outside of Bio/Eco/Env areas in strategic collaborative partnerships

Ack: DataONE Team and Sponsors



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



• Ewa Deelman



• Dave Vieglais



• Deborah McGuinness



• Suzie Allard, Nick Dexter, Kimberly Douglass, Carol Tenopir, Robert Waltz, Bruce Wilson



• Jeff Horsburgh



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



• Robert Sandusky



• Patricia Cruse, John Kunze



• Bertram Ludaescher



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



• Peter Honeyman



• Stephanie Hampton, Chris Jones, Matt Jones, Ben Leinfelder, Andrew Pippin



• Cliff Duke



• Paul Allen, Rick Bonney, Steve Kelling



• Carole Goble



• Ryan Scherle, Todd Vision



• Donald Hobern



• Randy Butler



• David DeRoure



Questions?



Contact Points

John W. Cobb, Ph.D.
Oak Ridge

John W. Cobb, Ph.D.
Oak Ridge National Lab
cobbjw@ornl.gov
865.576.5439

<http://www.dataone.org/>
<http://docs.dataone.org>