

Metadata Zoo

Dataset Metadata

Rebecca Koskela
Executive Director, DataONE

euroCRIS
September 9, 2013

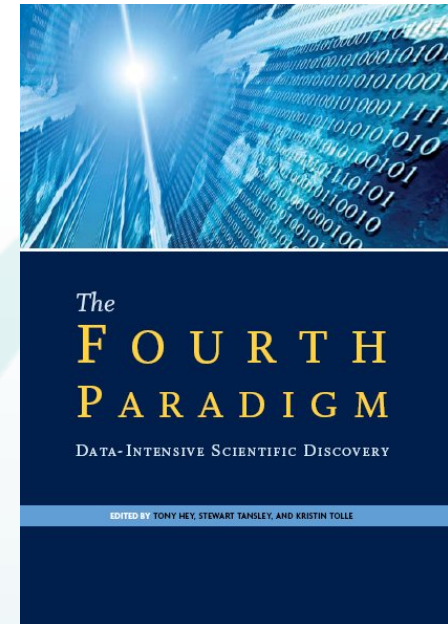


Outline

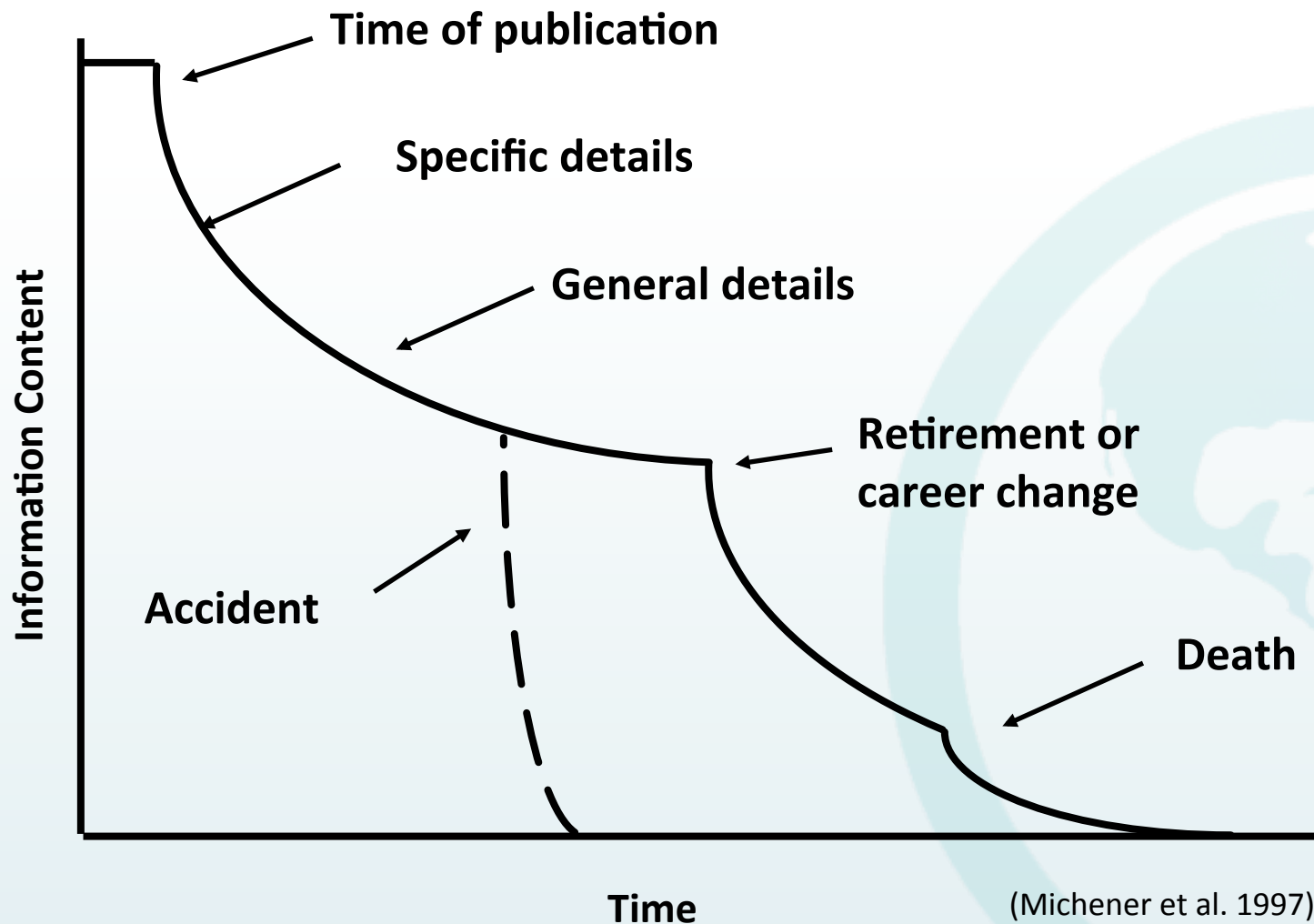
- Data Challenges
- Metadata Solution
- DataONE addressing the Data Challenge
- Enabling Scientific Discovery Example

Data Management Challenges

- Entropy
- Discovery
- Heterogeneity



Data Entropy



Data Discovery



100s
Repositories



1,000s
Institutions



10,000s
Scientists



100,000s
Citizen
Scientists

Data Heterogeneity

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

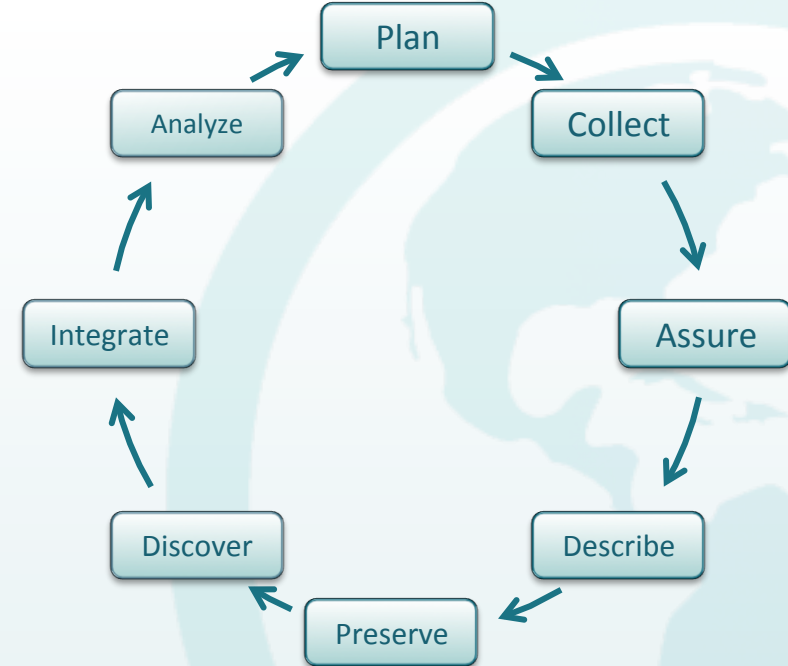
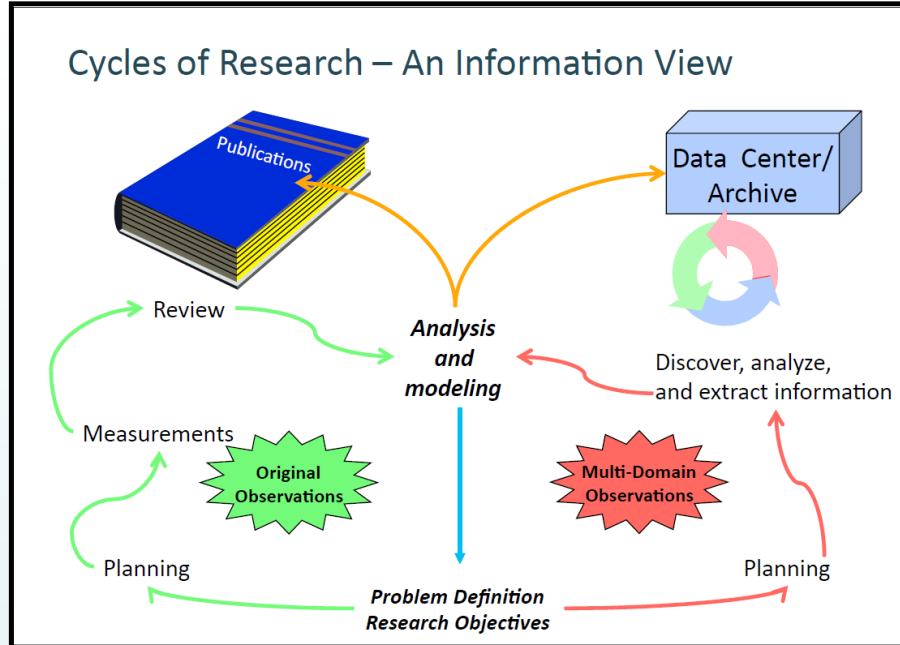
format
normalized
using
metadata

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

density
calculated
using
metadata

Jones et al. 2007

Research and data life cycle integration



Many iterations of data to final data product

Open Science Movement

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren 
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather

To that end, I have [issued a memorandum today \(pdf\)](#) to Federal agencies that directs those with more than \$100 million in research and development expenditures to develop plans to make the results of federally-funded research publicly available free of charge within 12 months after original publication.

...the memorandum requires that agencies start to address the need to improve upon the management and sharing of scientific data produced with Federal funding.



the WHITE HOUSE

PRESIDENT BARACK OBAMA

★★★★

★★★★



BLOG

PHOTOS & VIDEO

BRIEFING ROOM

ISSUES

the ADMINISTRATION

WE *the* PEOPLE

GREEN JOBS

EDUCATION

TECHNOLOGY

LEGISLATION

EQUALITY

POLITICS

ENVIRONMENT

YOUR VOICE IN OUR GOVERNMENT



CREATE A PETITION

OPEN PETITIONS

RESPONSES

HOW & WHY



Metadata as a Solution

Descriptive

- Subject matter of content

Structural

- Content types and attributes

Administrative

- Who, when, how content created
- Defines who can access, how it can be used

IDENTIFY

- keywords
- geographic location
- time period
- attributes

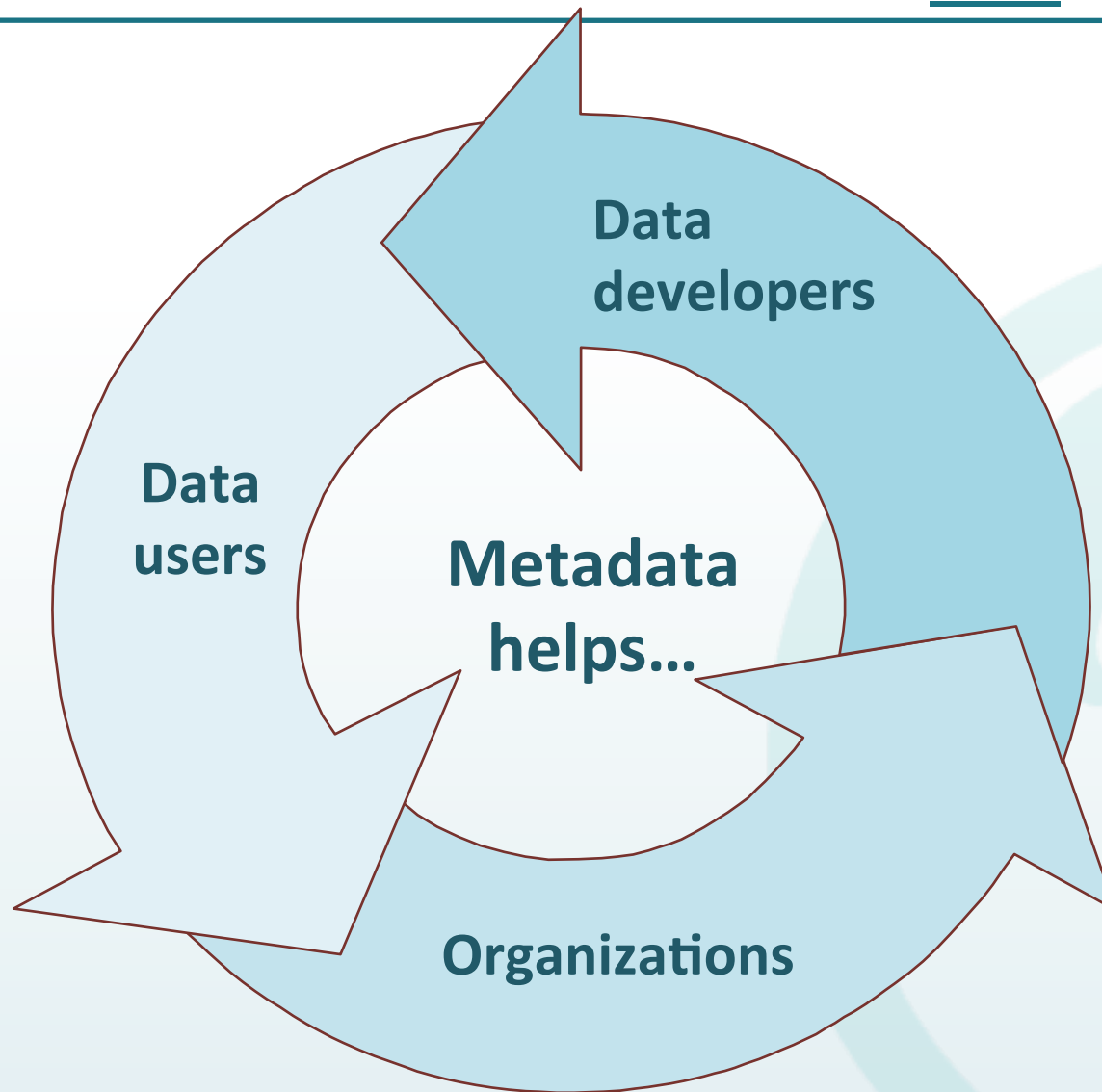
ASSESS

- use constraints
- access constraints
- data quality
- availability/pricing

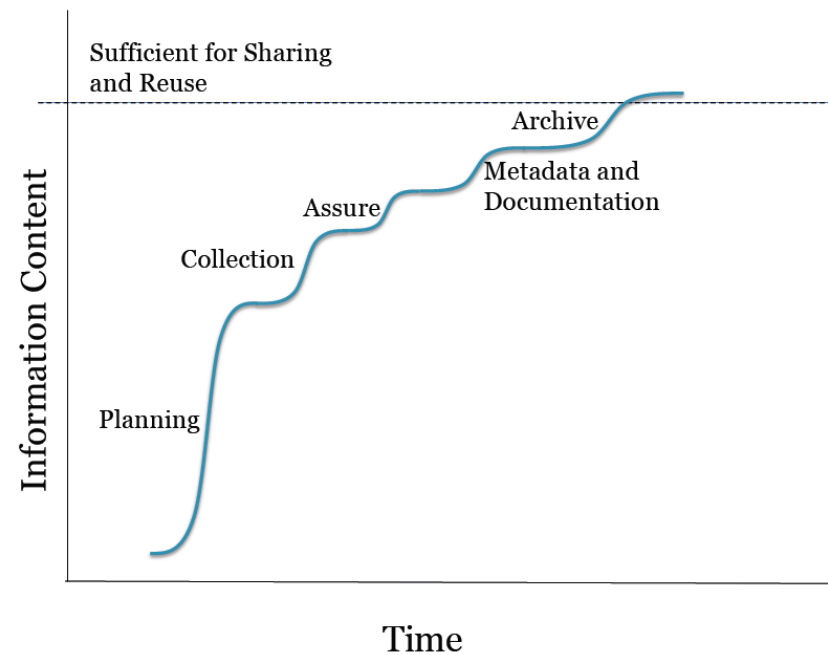
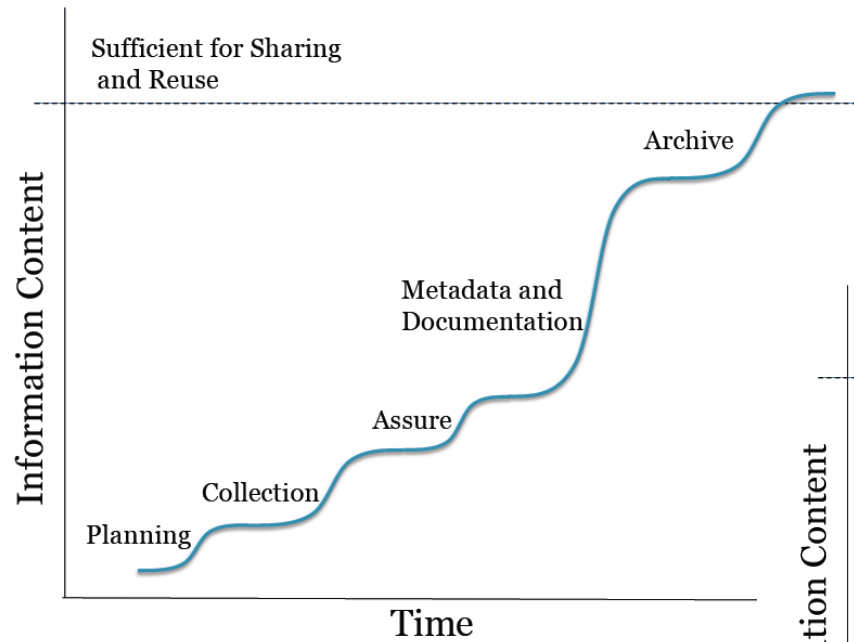
ACCESS

- online access
- order process
- contacts

Dataset Metadata Has Value to ALL



Proper Dataset Curation Enables Data Reuse



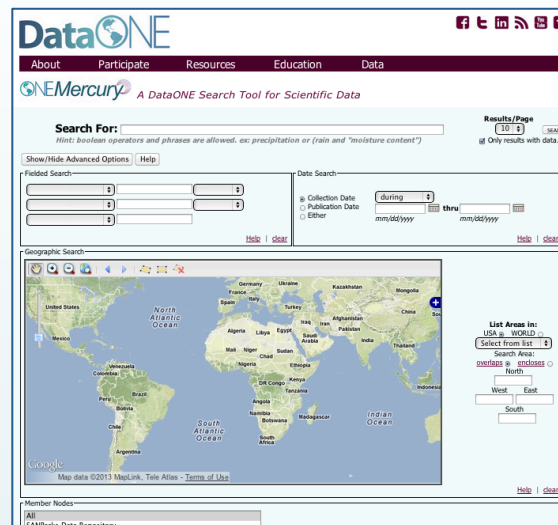
DataONE addressing the Data Challenge

Provide universal access to data about life on earth and the environment

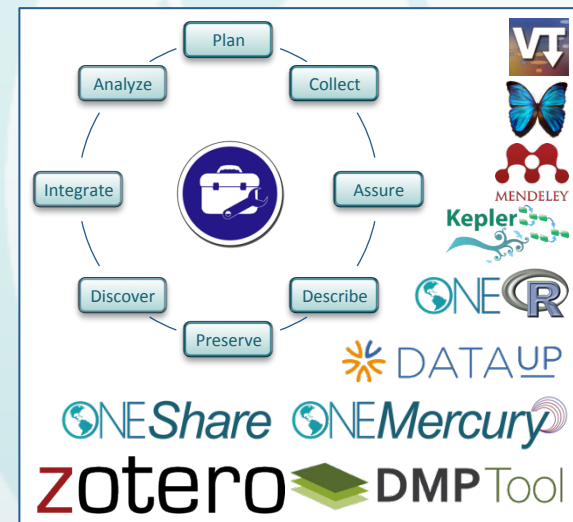
1. Building community



2. Developing sustainable data discovery and interoperability solutions



3. Enabling science through tools and services

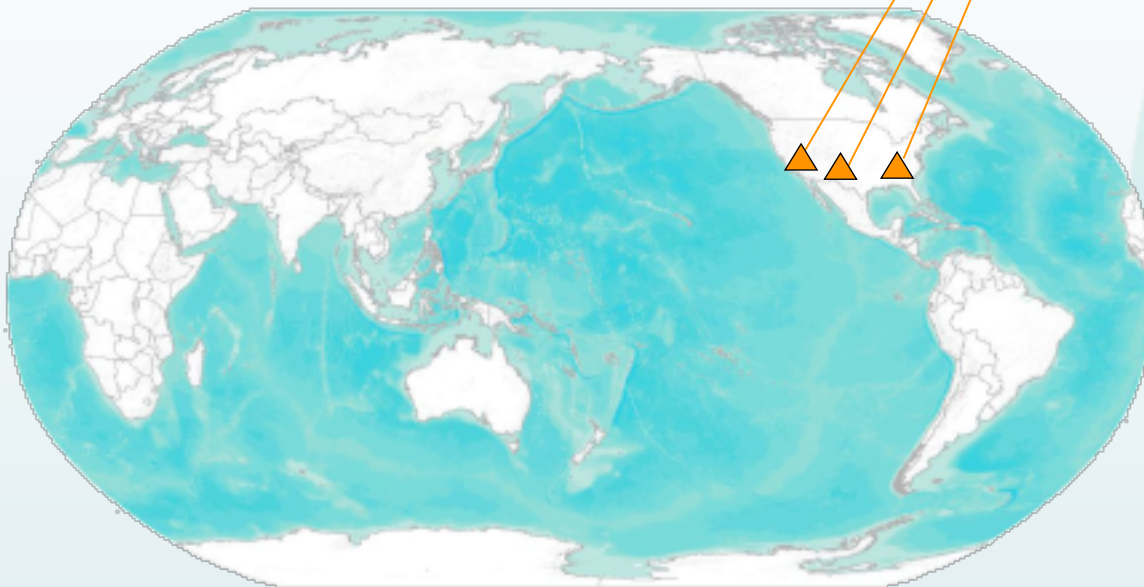


Discovery via DataONE

Three major components for a flexible, scalable, sustainable network

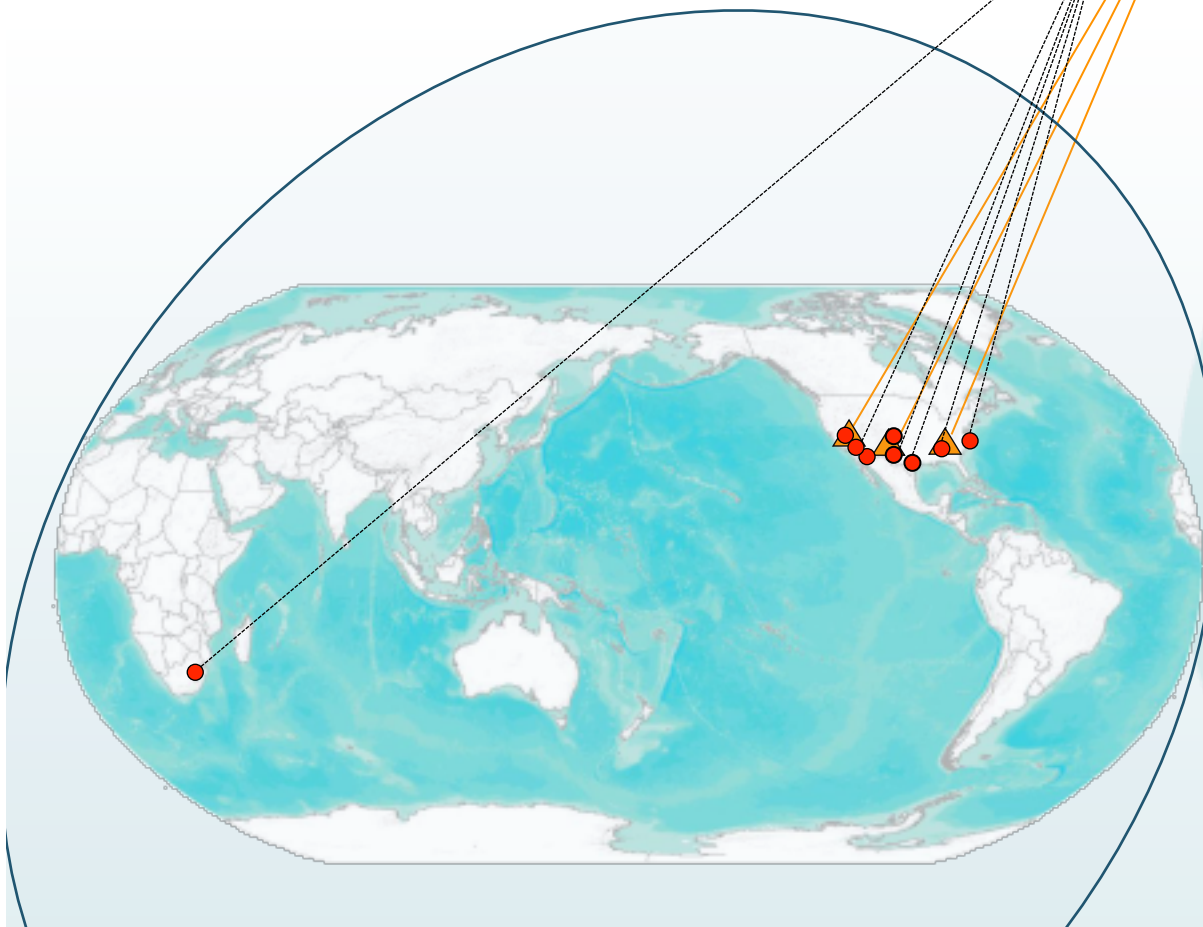
Coordinating Nodes

- retain complete metadata catalog
- indexing for search
- network-wide services
- ensure content availability (preservation)
- replication services



Discovery via DataONE

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

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



The **Cornell** Lab 



UC3 Merritt



PISCO



USGS
science for a changing world



ECOLOGICAL SOCIETY OF AMERICA
esa



KNB



ONE Share

Discovery via DataONE

Three major components for a flexible, scalable, sustainable network

Coordinating Nodes

Member Nodes

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



The **Cornell** Lab 



UC3Merritt



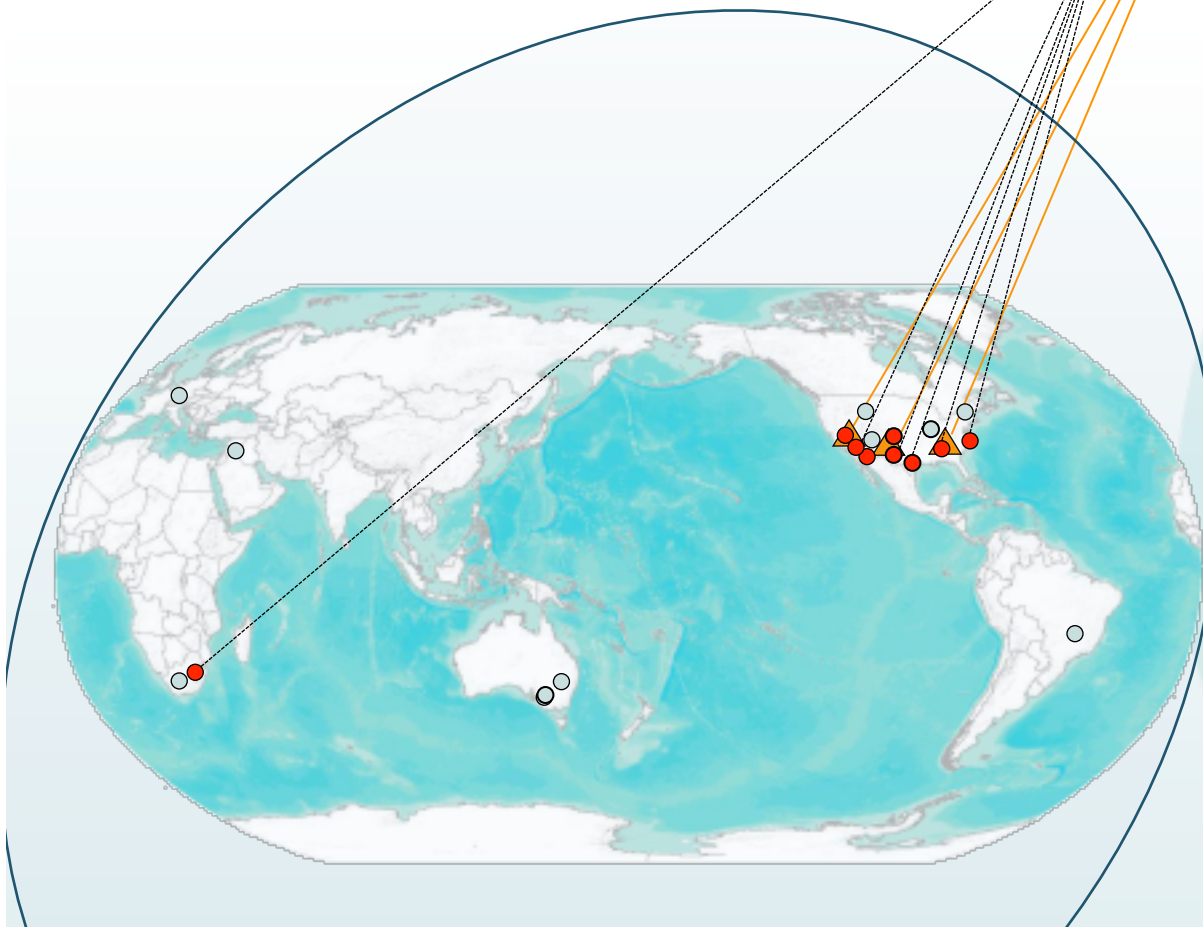
PISCO



ECOLOGICAL SOCIETY OF AMERICA
esa

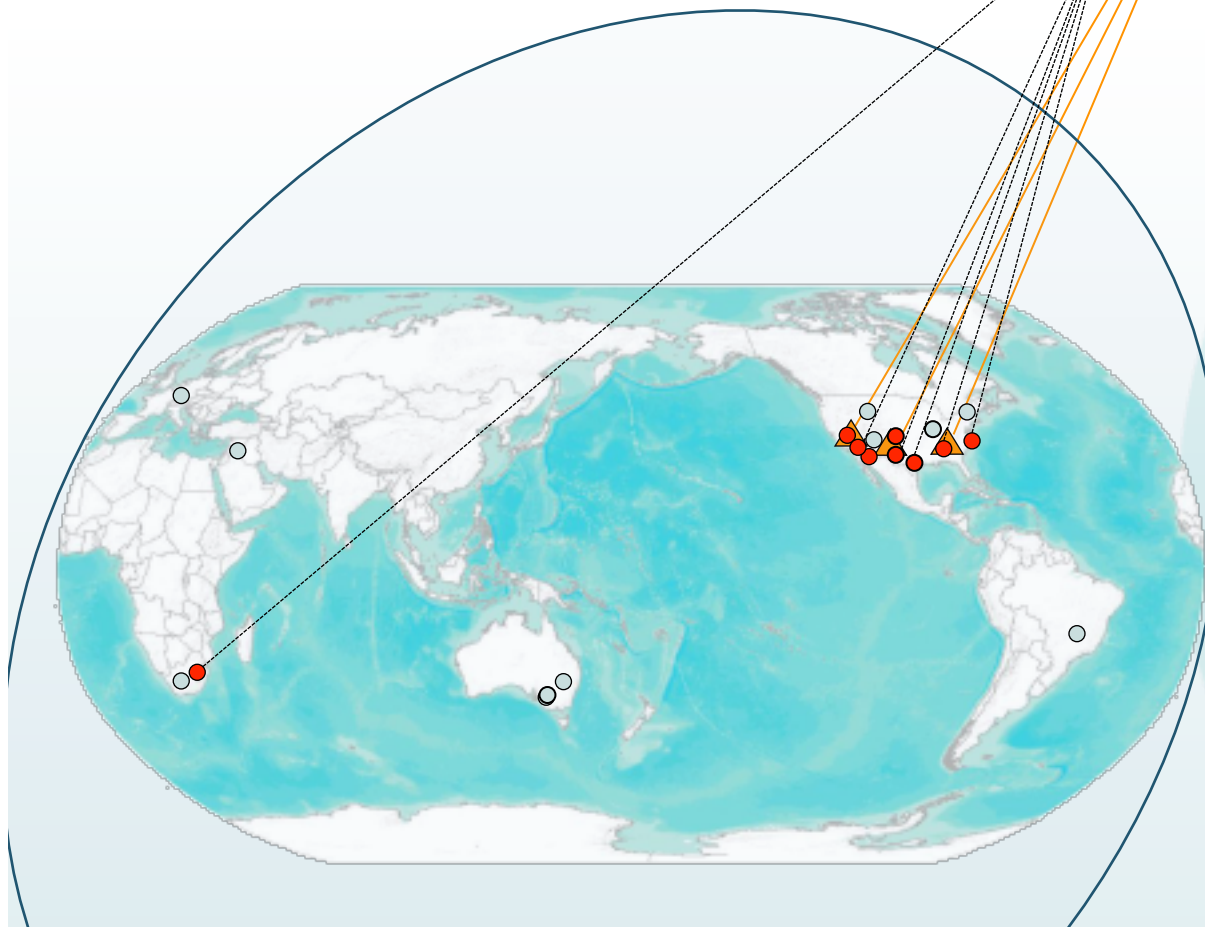


ONE Share



Discovery via DataONE

Three major components for a flexible, scalable, sustainable network



Coordinating Nodes

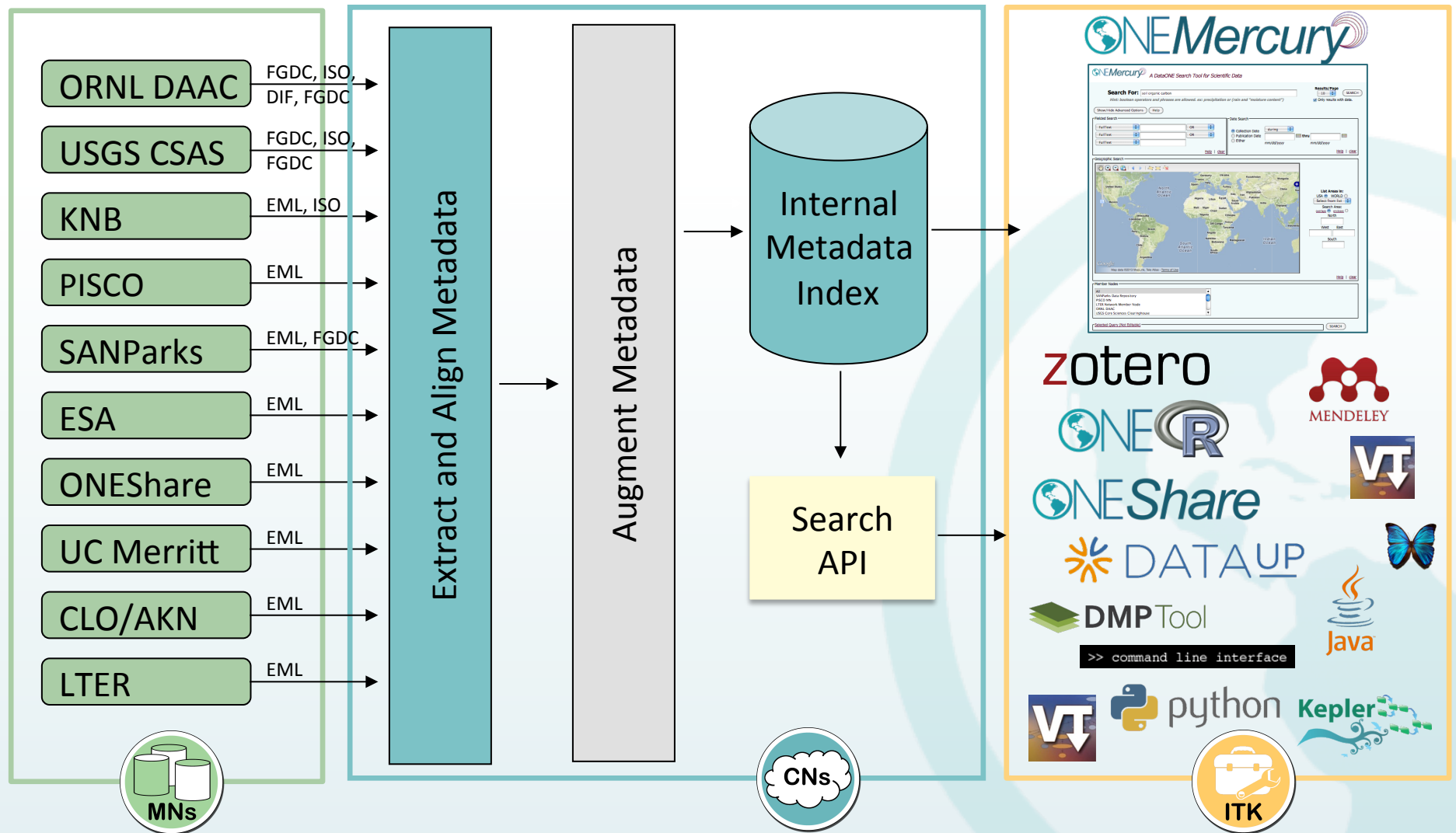
Member Nodes

Investigator Toolkit

>> command line interface



Enable Data Discovery



Community

- Listening
- Understanding
- Engaging



Listening

stakeholder surveys



scientists



library's & librarians

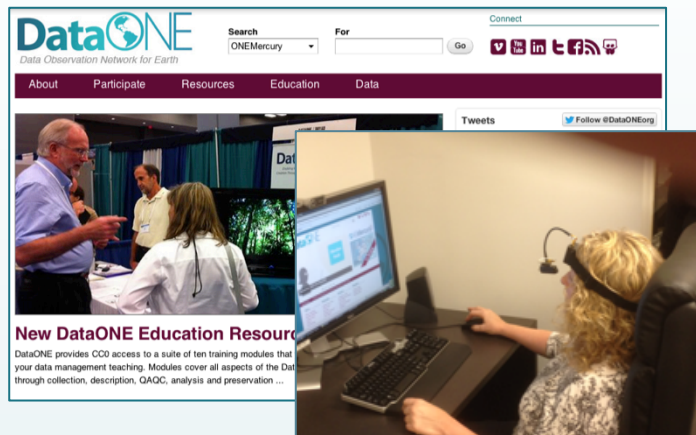


data managers



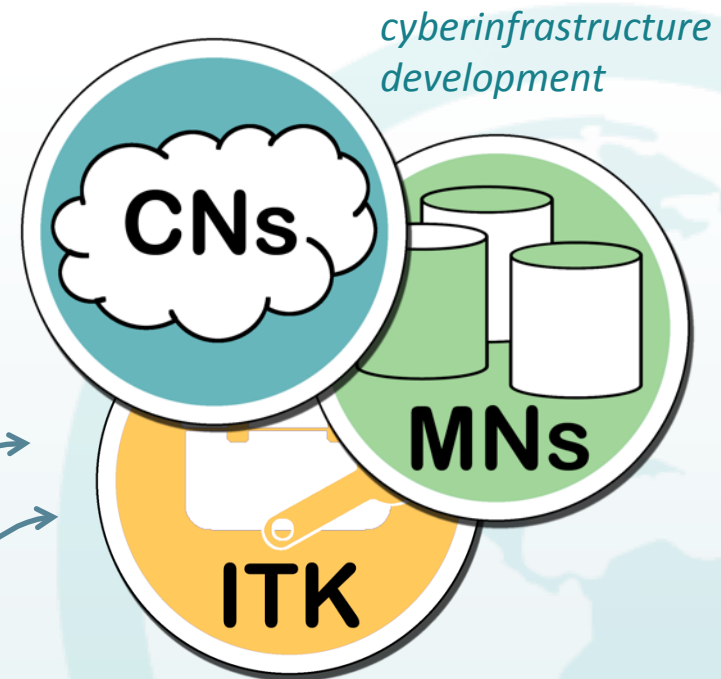
persona and scenario development

usability testing

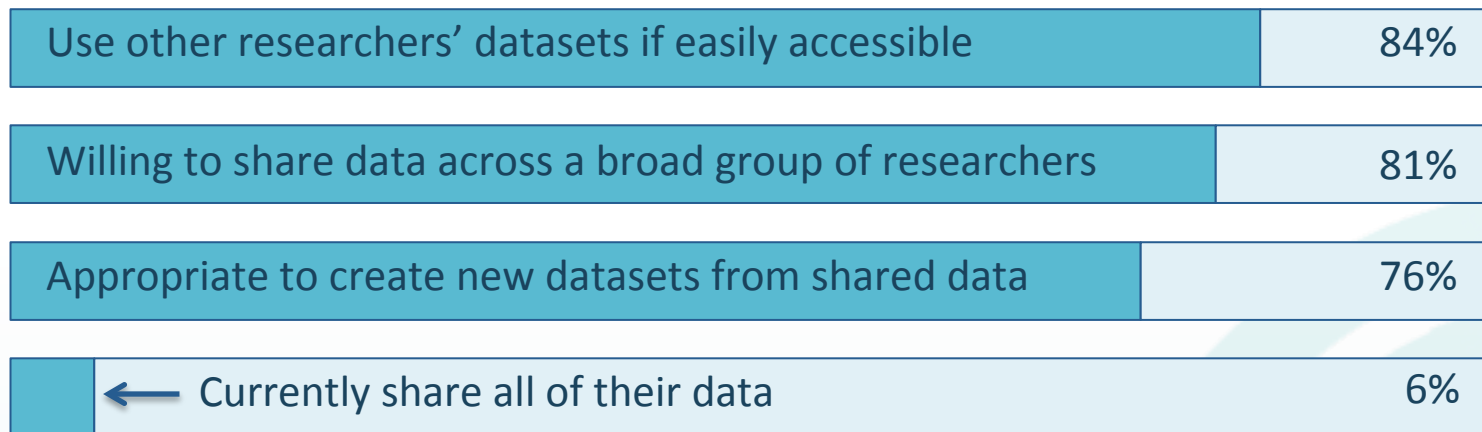


external assessments / surveys

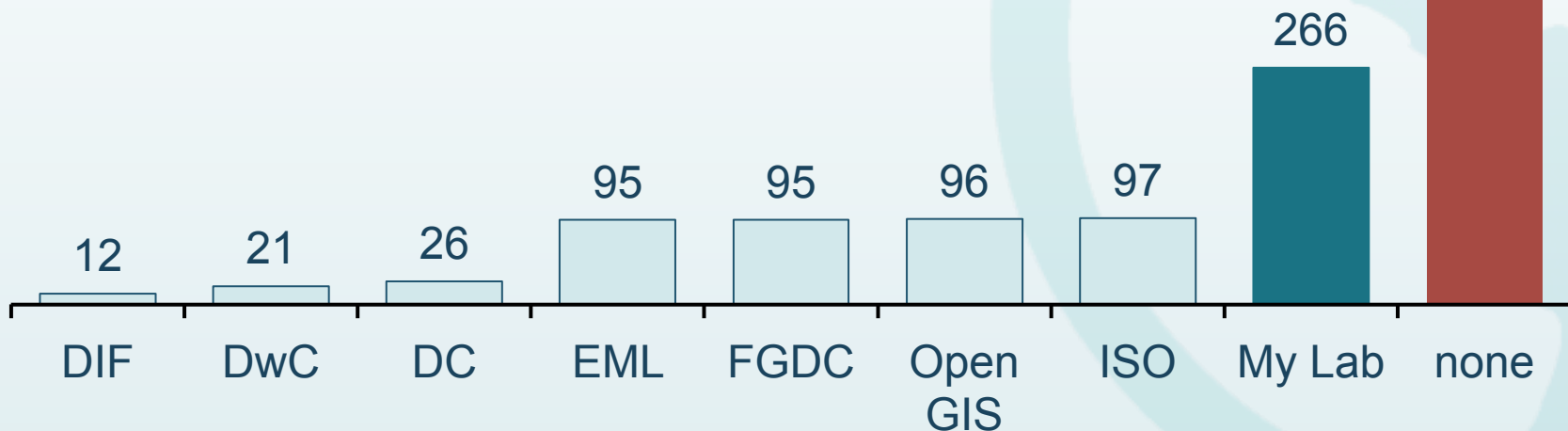
		Articles	For Authors	About Us	Search		
		advanced search					
 OPEN ACCESS	 PEER-REVIEWED	10,244	21	189	40	VIEWS	CITATIONS
RESEARCH ARTICLE				ACADEMIC BOOKMARKS	SOCIAL SHARES		
Data Sharing by Scientists: Practices and Perceptions							
Carol Tenopir  , Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, Mike Frame							



Scientists want to share data



Metadata standards



Understanding

OPEN ACCESS Freely available online

PLOS one

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

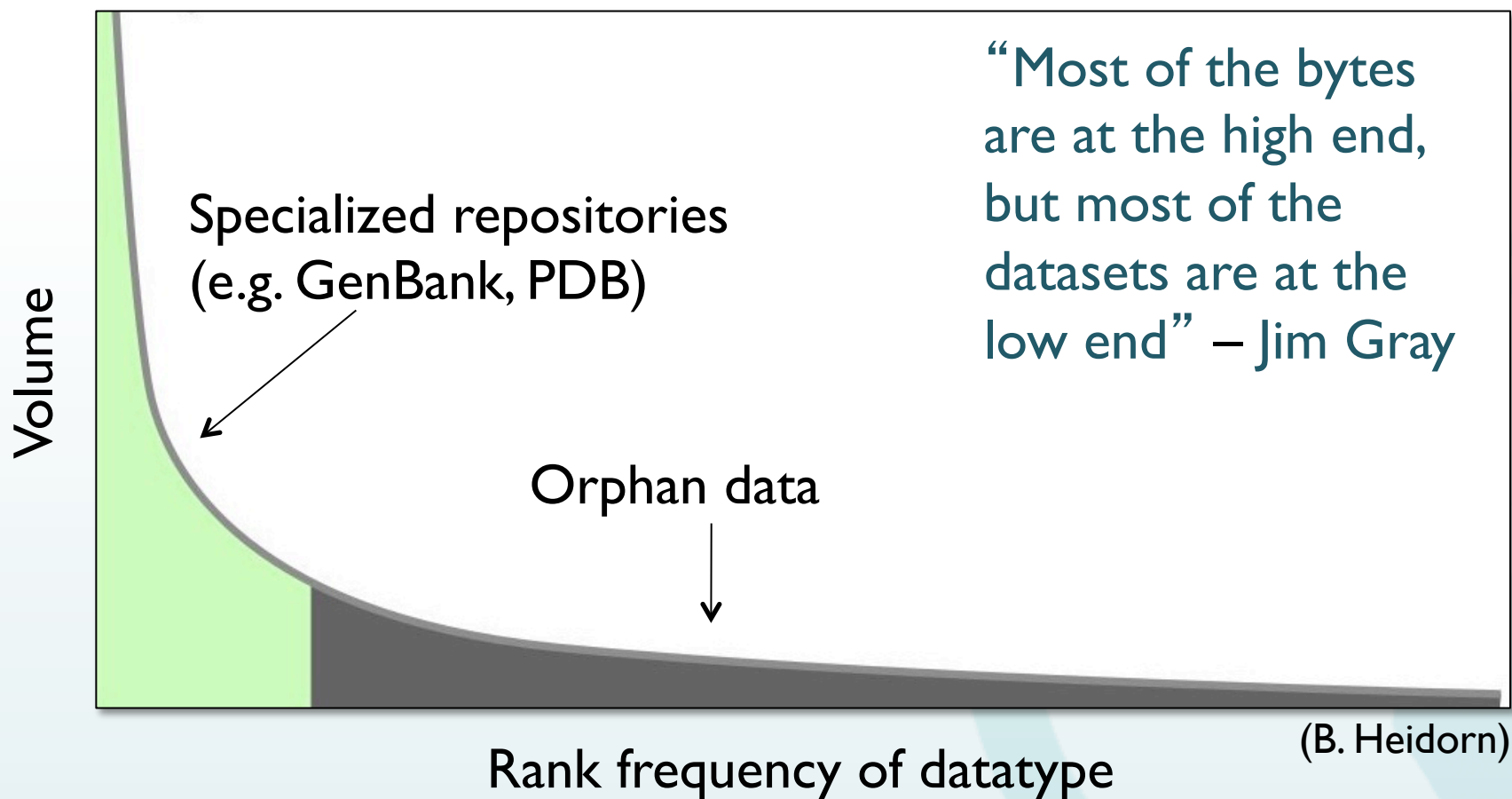
Abstract

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-

- 2/3rd report that organizational help and support is lacking

The Long Tail Science





Intercept researchers where **they already work**

Microsoft®

Research



University of California

CDL

California Digital Library

GORDON AND BETTY
MOORE
FOUNDATION

Data  **ONE**

	A	B	C	D	E	F	G	H	I
1			\$4T15_R1		\$4T15_R2		\$4T15_R3		\$4T15_R1
2		amount for	copies	naups					
3									
4	23-Oct	30	180		650		750		650
5	25-Oct	25	400		780		550		400
6	27-Oct	25			500		1200		400
7	29-Oct	35	300	750	350	100			200
8	31-Oct	30	340	800	300	300	1450		700
9	2-Nov	25	292	1150	660		1000		520
10	4-Nov	25	360	440	200+	260	1050		700
11	6-Nov	30	-	400	-	250	550	0	750
12	10-Nov	30	-	300	-	280	300	0	450
13	12-Nov	30	~100	160	<100	80	450	1450	550
14	12-Nov	35	<50	60	<50	100	300	800	550
15	14-Nov	30	<10	100	<50	20	200	600	450
16	16-Nov	35	<10	30	<10	0	140	2100	550
17	18-Nov	30	0	0	<20	0	125	400	150
18	20-Nov	25	0	0	<20	40	150	280	250
19	22-Nov	25			<10	0	120	50	
20	24-Nov				<10	0	125	150	150
21	26-Nov				<5	0	50	30	0
22	30-Nov				<5	0		0	0
23	31-Dec	25			0	0	<25	0	0
24	2-Dec	25					<10	0	0
25	4-Dec	25						0	0
26	6-Dec								
27	8-Dec								



DATAUP

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



Data & Metadata (EML)

[illegible]

File General Metadata Identifier Citation Post File Cache

▼ A Name B

1 First name
2 Last name
3 Address
4 City
5 State/province
6 Postal code
7 Country
8 Email
9 Phone
10 Organization
11 Title of dataset*
12 Today's date*
13 Abstract
14 Repository name and contact information
15 Data Contact Person: First name
16 Data Contact Person: Last name
17 Data Contact Person: Address
18 Data Contact Person: City
19 Data Contact Person: State/province
20 Data Contact Person: Postal code
21 Data Contact Person: Country
22 Data Contact Person: Phone
23 Data Contact Person: Email
24 Data Contact Person: Organization
25 Keywords(s)
26 Keyword Backlist used
27 Geographic coverage: Description
28 Geographic coverage: West bounding coordinate
29 Geographic coverage: East bounding coordinate
30 Geographic coverage: North bounding coordinate
31 Geographic coverage: South bounding coordinate
32 Temporal coverage: Beginning date
33 Temporal coverage: Ending date
34 Project title
35 Project description
36 Funding
37 Intellectual rights

GENERAL METADATA

Clear All Reset Load Default

Creator: First name Carly
Creator: Last name Smither
Creator: Address 413 20th St
Creator: City Oakland
Creator: State/province CA
Creator: Postal code 94612
Creator: Country USA
Creator: Email carly.smith@ucsf.edu
Creator: Phone 510-867-6179
Creator: Organization California Digital Library
Title of dataset* OpenPoplog_2012-08
Today's date* June 23 2012
Abstract
Repository name and contact info... ORCID iD
url for data
Save Cancel

CACSN
CC-BY

Enable Data Discovery

Ease of good search key for users

Metadata quality key enabler

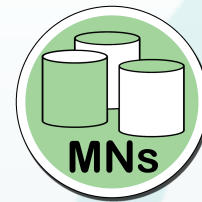
Scaling search is hard



Access search API
Promote quality metadata
Enable search refinement



Index & replicate metadata
Fast, robust search API
Validate metadata



Develop & curate metadata
Promote quality metadata



Semantics research collaboration
Drive usability improvements

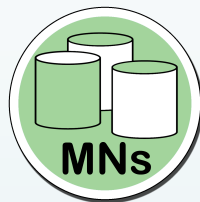
Preserve Data and Metadata

Support common metadata standards

Use checksums for invariance

Replicate metadata across all CNs

Replicate data across participating MNs



Assemble data packages
Curate data & metadata
Establish replication policies
Promote quality metadata



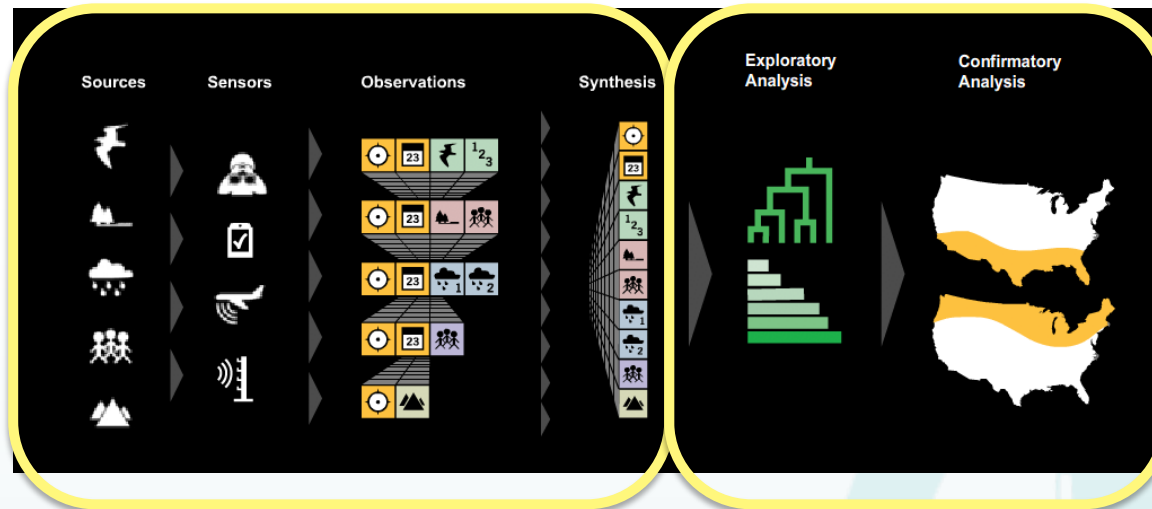
Manage system metadata
Replicate and preserve metadata
Enforce checksum consistency
Broker & maintain data replication

Enabling Scientific Discovery Example



Data Intensive Science in Biodiversity Research

*Data
Collection
Organization
Validation
Preservation,
& Access*

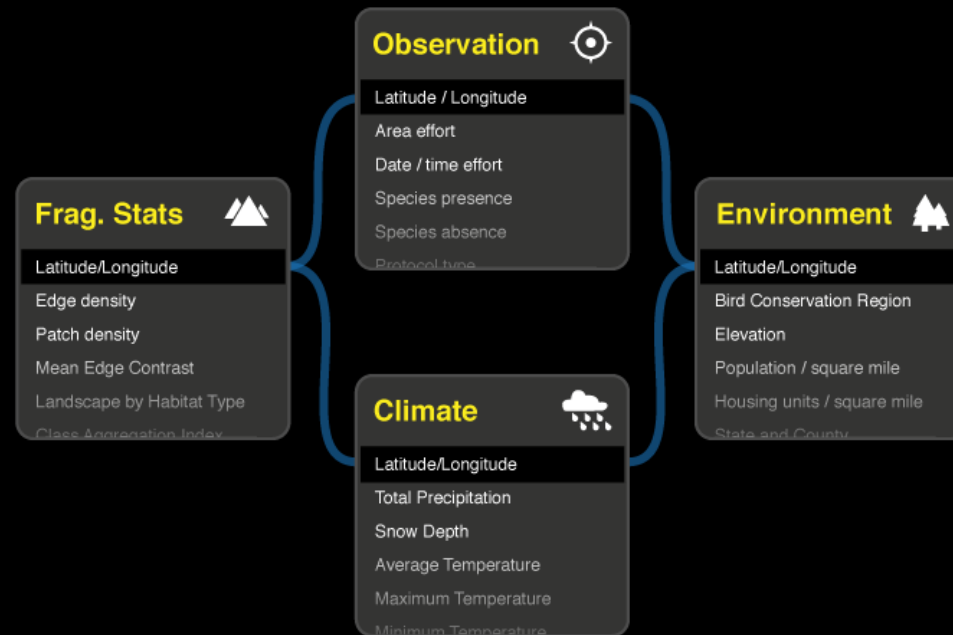
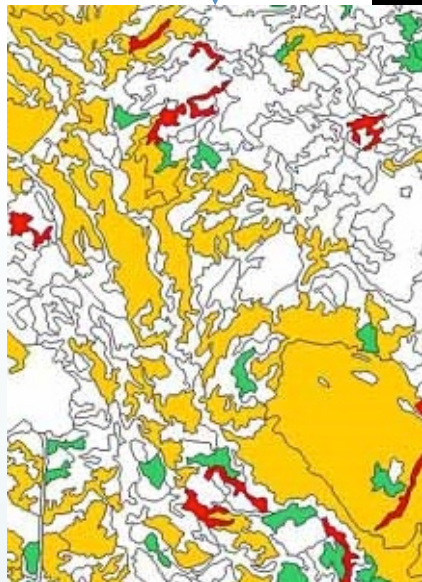


*Data
Exploration,
Visualization,
And Analysis*

Kelling et al. 2009

The full data “life cycle” from data curation to scientific learning –
engage scientists, land-managers, policy makers, students,
educators, and the public

Joining Disparate Data

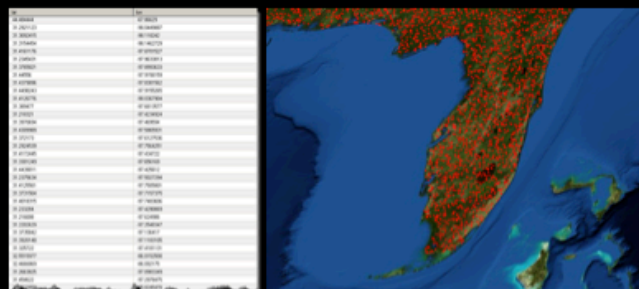


High Performance Computing and Climate Change:

Environmental Cues of Migration

eBird

Latitude, Longitude of bird Observations from eBird database providing presence/absence



- Observations from ~350,000 locations used in the extraction
- More than 15 million NDVI/EVI values extracted



Latitude, longitude, & time of observation

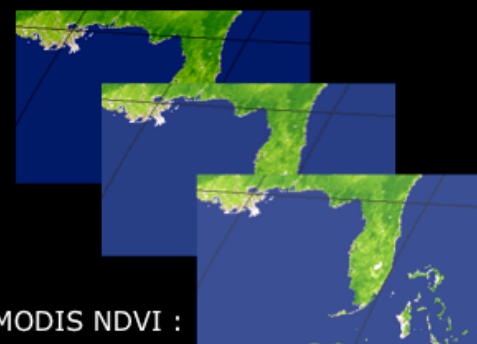


MODIS Web Service



MODIS NDVI value

ORNL DAAC

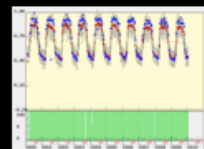


MODIS NDVI :

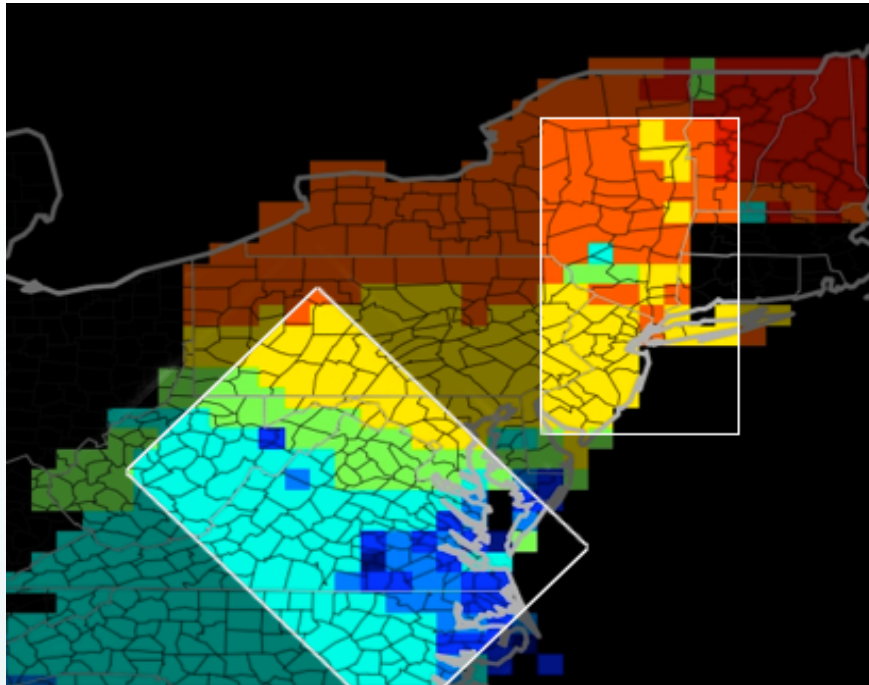
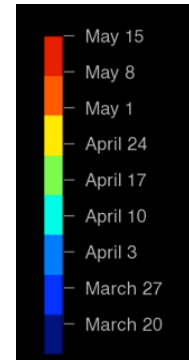
- 16 day composite value
- 250-m spatial resolution
- Data from 2000 onwards

MODIS NDVI bands extracted:

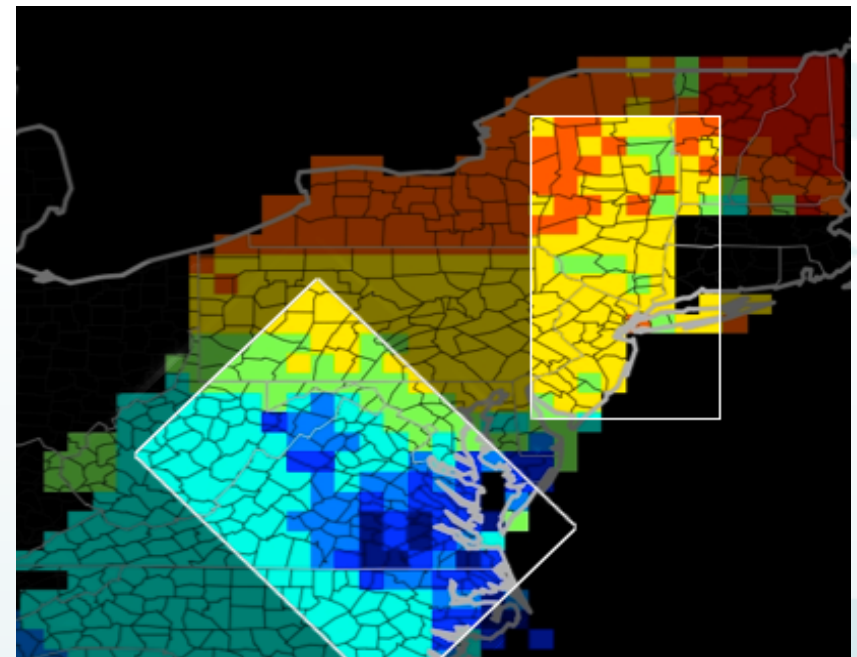
- 250m_16_days_EVI
- 250m_16_days_NDVI
- 250m_16_days_VI_Quality
- 250m_16_days_pixel_reliability



Climate Change: Environmental Cues of Migration



Spring Arrival Dates

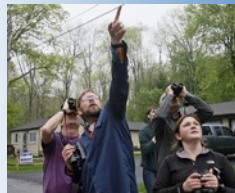


Expected Spring Arrival Dates
with "leaf out" two weeks early

Enabling Scientific Discovery



eBird



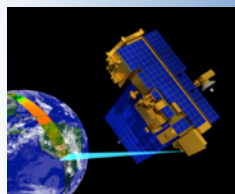
Land Cover



Meteorology

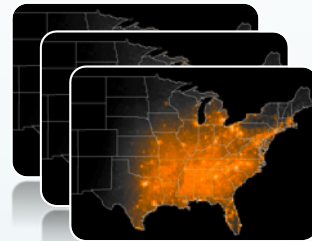


MODIS –
Remote
sensing data



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

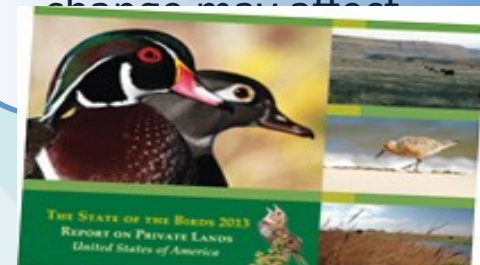
DataONE

Model results

Occurrence of Indigo Bunting (2008)



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



Thank you

Rebecca Koskela, Executive Director DataONE
rkoskela@unm.edu

www.dataone.org

@dataONEorg

