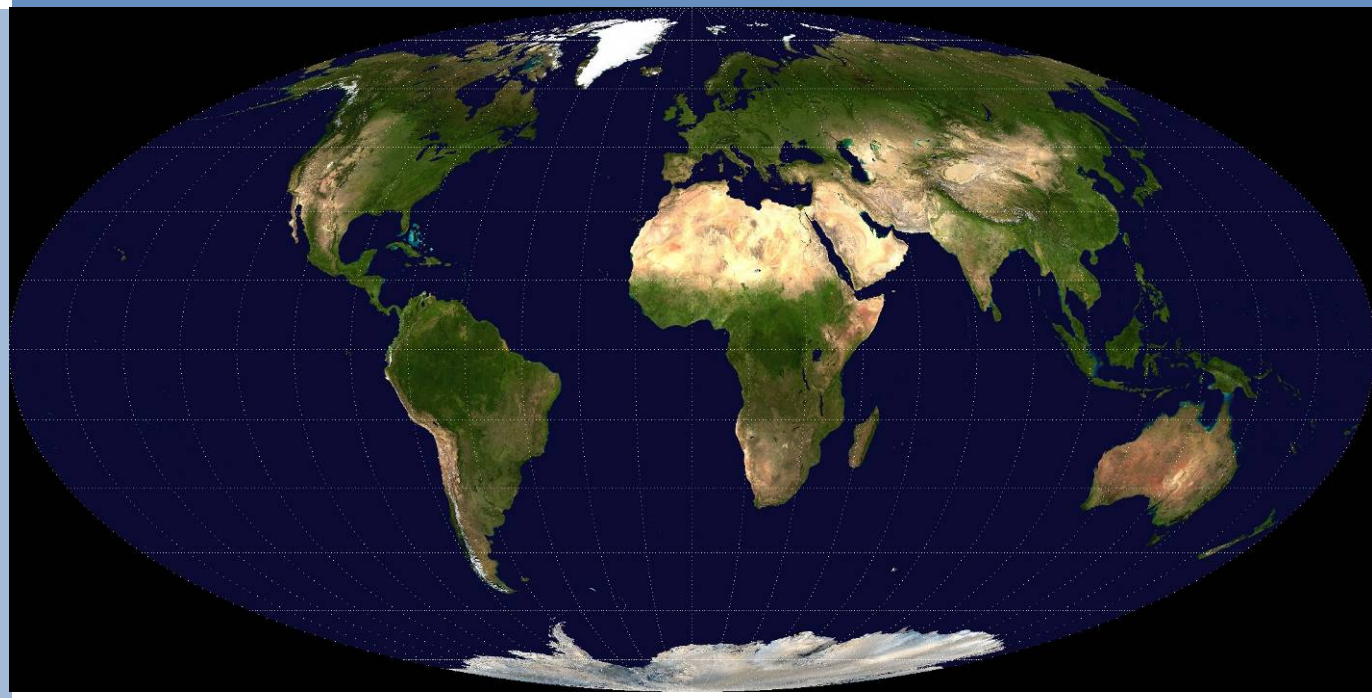
DataONE: Enabling Data-Intensive Biological and Environmental Research

-- Bill Michener (University Libraries, University of New Mexico)
-- Mike Frame (USGS National Biological Information Infrastructure



Road map to today's talk

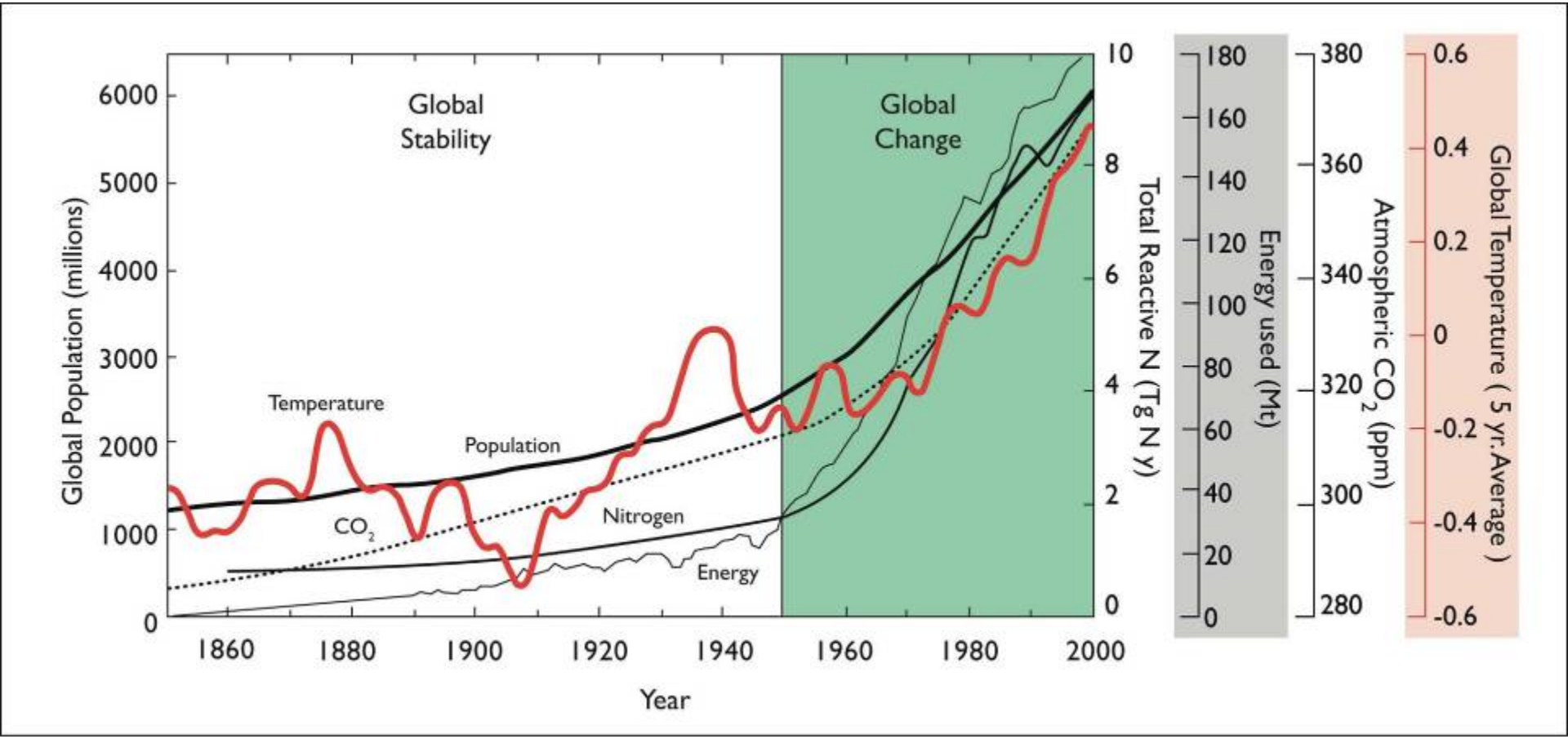
- Environmental science challenges
- Cyberinfrastructure challenges
- DataONE: A solution
- Benefits to science

Road map to today's talk

- Environmental science challenges
- Cyberinfrastructure challenges
- DataONE: A solution
- Benefits to science

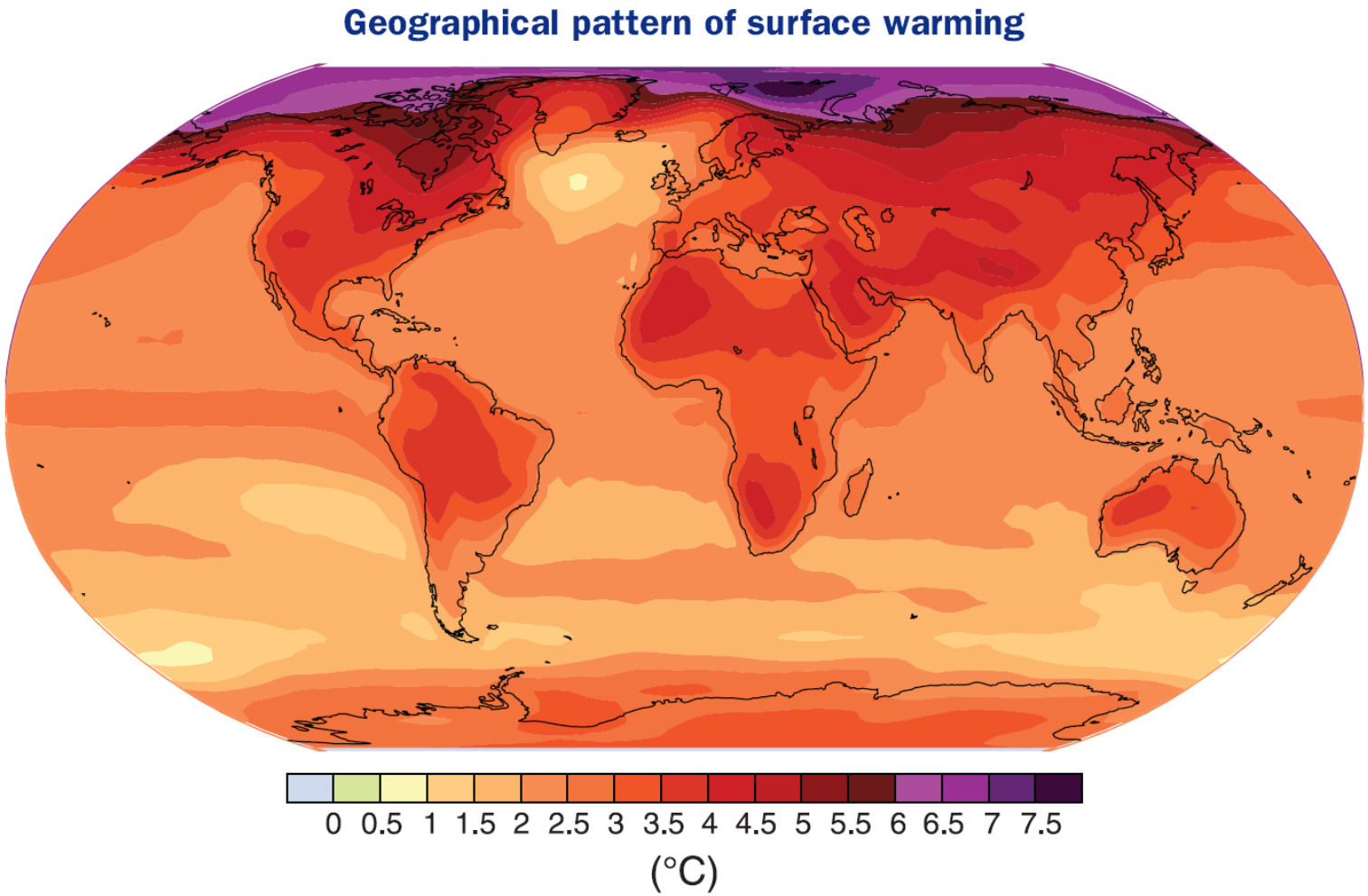
Living with Cancer
The changing science
Beyond Baghdad: Where The Enemy Has Its Own Surge
The Sopranos' Last Song: What Exit Will Tony Take?
TIME
SPECIAL DOUBLE ISSUE
The Global Warming Survival Guide
51 Things You Can Do to Make a Diffe
2ND-QTR SIZZLE PROFITS AT 900 COMPANIES (P.74)
PAYING FOR COLLEGE BEWARE OF THOSE HIGH 529 FEES (P.96)
TERRORISM WHAT COMPANIES STILL NEED TO DO (P.26)
BusinessWeek
GLOBAL WARMING
Why Business Is Taking It So Seriously
BY JOHN CAREY (P.60)
SPECIAL REPORT GLOBAL WARMING
TIME
BE WORRIED. BE VERY WORRIED.
Climate change isn't some vague future problem—it's already damaging the planet at an alarming pace. Here's how it affects you, your kids and their kids as well
EARTH AT THE TIPPING POINT
HOW IT THREATENS YOUR HEALTH
Adapted for A NEW GENERATION from the New York Times Bestseller
an inconvenient truth
the crisis of global warming
AL GORE
nature
Eocene global warming
Hydrothermal vents prompt methane release
SPECIAL REPORT
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
TIME
Where's the Beach?
America's Vanishing Coastline
TIME
The Heat Is On
How the Earth's Climate Is Changing
Why the Ozone Hole Is Growing
SPECIAL REPORT Fighting for Global Markets
Dr. Bush's Rx for Health Care
TIME
VANISHING OZONE
THE DANGER MOVES CLOSER TO HOME
THE BIG DRY
W YORKER
GLOBAL WARMING CARTOONS - 2007 - 210
CHINA
JAPAN
RUSSIA
PACIFIC OCEAN
TEXAS
UTAH
LAS VEGAS
ST LOUIS
CHICAGO
PENNSYLVANIA
ATLANTIC OCEAN
GLOBAL WARMING
Climbing temperatures. Melting glaciers. Rising seas. All over the earth we're feeling the heat. Why isn't Washington?
JOE KLEIN ON IRAQ
BROKEBACK MOUNTAIN: GIDDY-YEP, I'M GAY
TIME
New Orleans Blues
It's worse than you think. Three months after Katrina, the city still suffers
BY CATHY BOOTH THOMAS

Global change

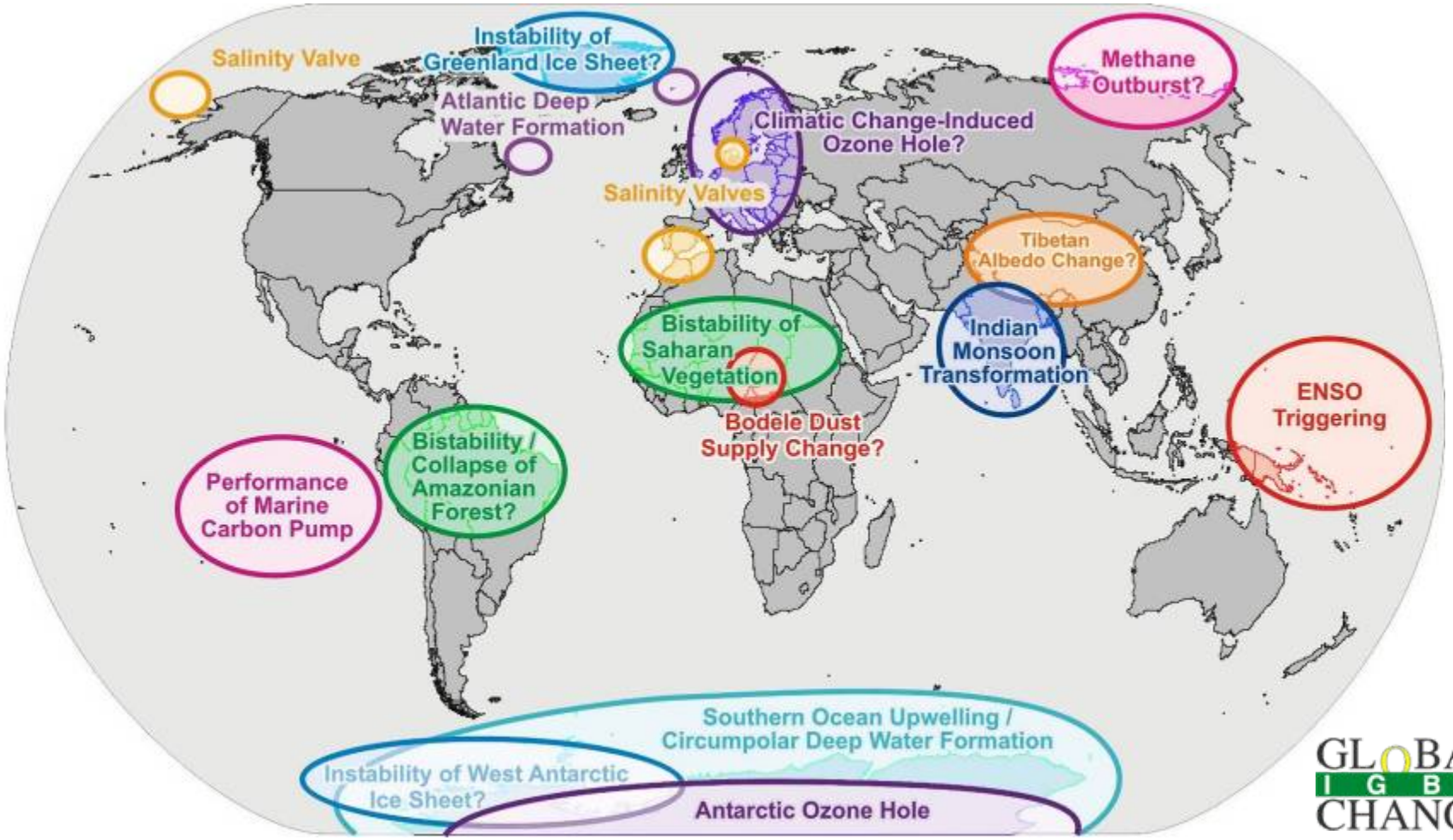


Smith, Knapp, Collins. In press.

Critical areas in the Earth's system



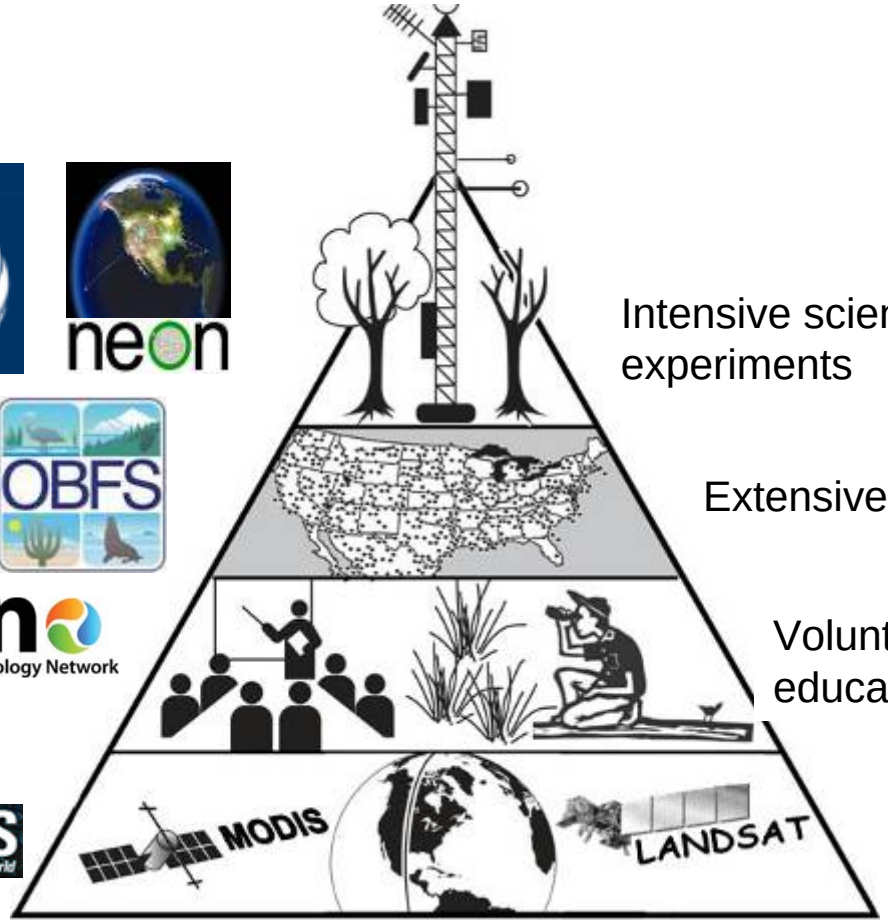
Critical areas in the Earth's system



Building the knowledge pyramid

Decreasing Spatial Coverage

Increasing Process Knowledge



Intensive science sites and experiments

Extensive science sites

Volunteer & education networks

Remote sensing

Adapted from CENR-OSTP

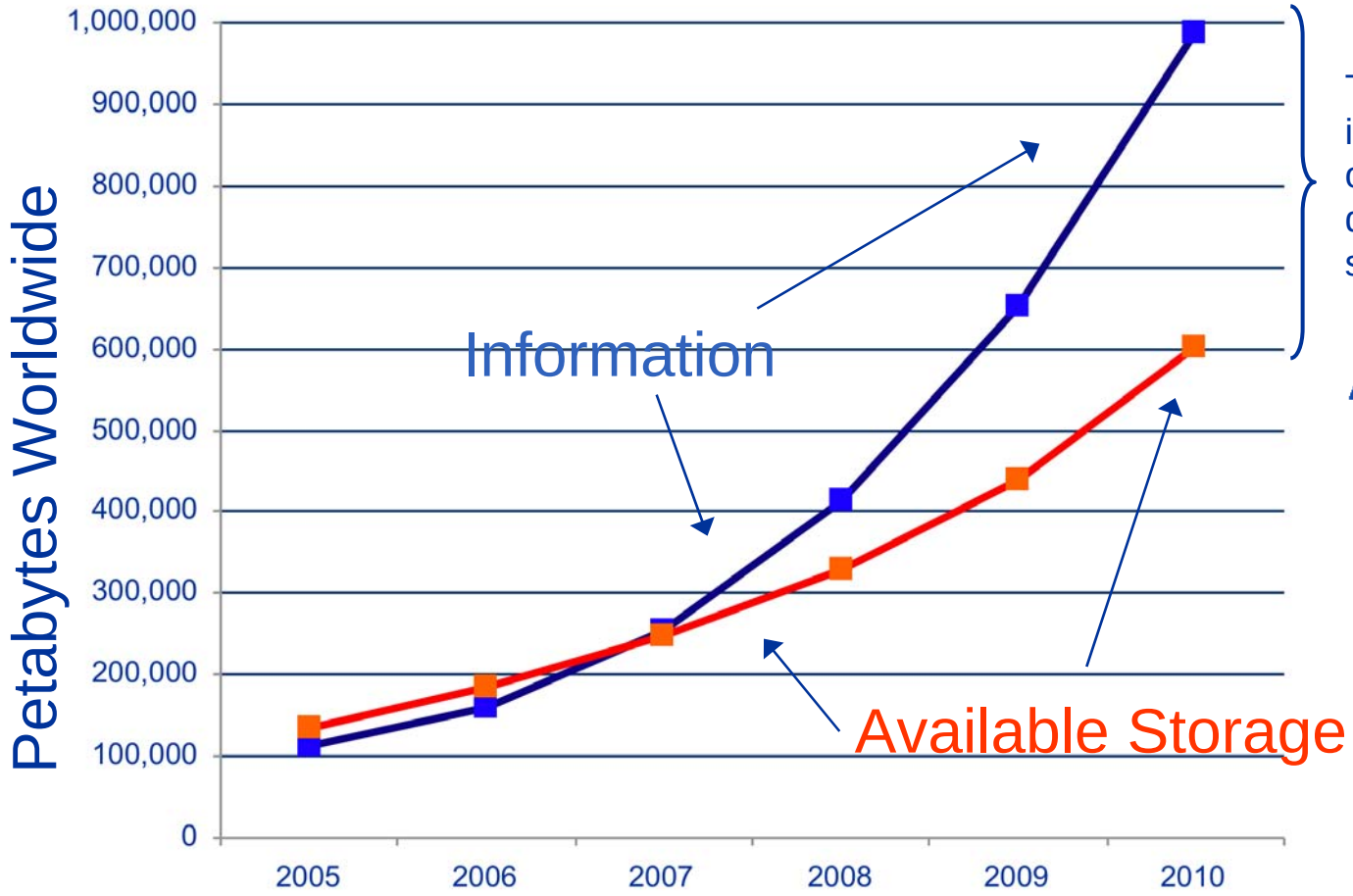
Road map to today's talk

- Environmental science challenges
- Cyberinfrastructure challenges
- DataONE: A solution
- Benefits to science

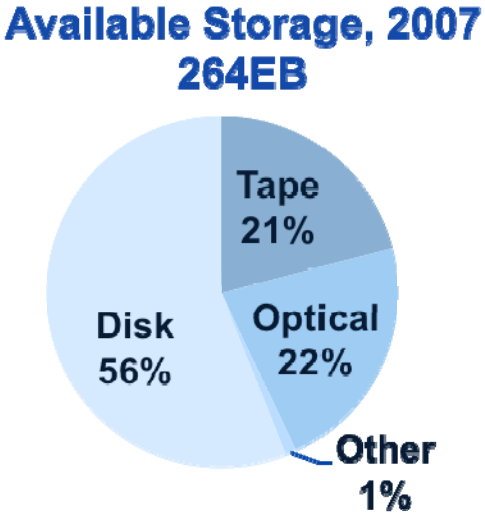
Data loss

- Natural disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment
- Loss of financial stability
- Changes in user expectations and requirements

Data Loss
(including orphaned data)



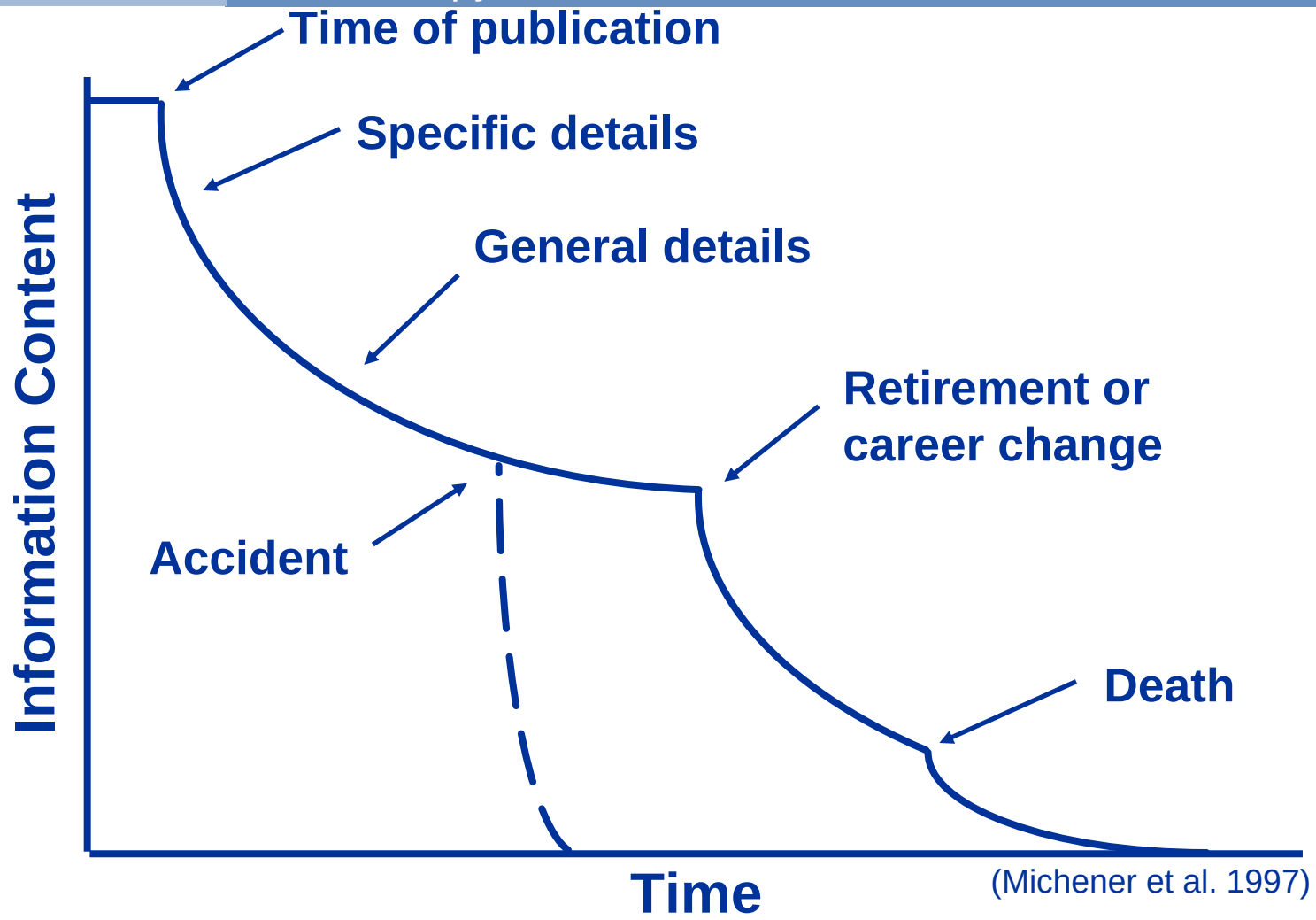
Transient information or unfilled demand for storage



Source: John Gantz, IDC Corporation: The Expanding Digital Universe

Poor data practice

"data entropy"



(Michener et al. 1997)

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.

Road map to today's talk

- Environmental science challenges
- Cyberinfrastructure challenges
- DataONE: A solution
  - Building on existing CI
  - Creating new CI
  - Changing science culture and institutions
- Benefits to science

Existing biological data archives

ESA's Ecological Archive

Distributed Active Archive Center

National Biological Information Infrastructure

Fire Research & Management Exchange System

Long Term Ecological Research Network

Knowledge Network for Biocomplexity

Examples of data holdings

Metadata Interoperability Across Data Holdings

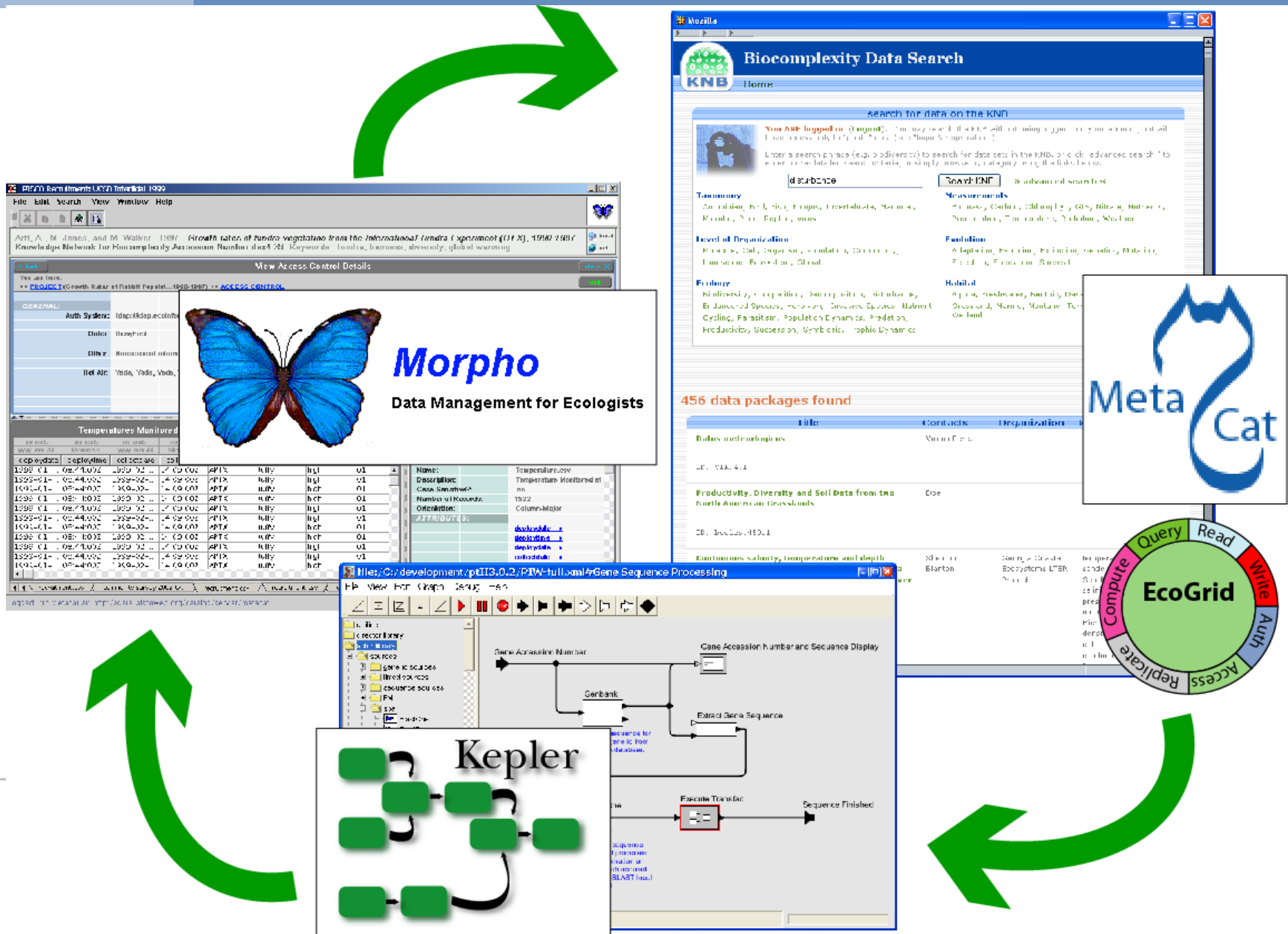
Data Archive	Types of Data Managed	Metadata Standard(s)
nbii National Biological Information Infrastructure	Biodiversity, taxonomic, ecological	BDP, DwC, DC, OGIS
DAAC for biogeochemical dynamics Oak Ridge National Laboratory	Biogeochemical dynamics, terrestrial ecological Earth observation imagery	DIF, BDP, ECHO
The US Long Term Ecological Research Network	Ecological, biodiversity, biophysical, social, genomics, and taxonomic	EML
Avian Knowledge Network	Avian populations and molecular biology	DwC
ALA ATLAS OF LIVING AUSTRALIA	Biological and taxonomic	DC subset
South African Environmental Observation Network	Biophysical, biodiversity, disturbance, and Earth observation imagery	EML
TERN TAIWAN ECOLOGICAL RESEARCH NETWORK	Biodiversity, biotic structure, function/process, biogeochemical, climate, and hydrologic	EML

BDP=Biological Data Profile DC subset=Dublin Core subset DwC=Darwin Core

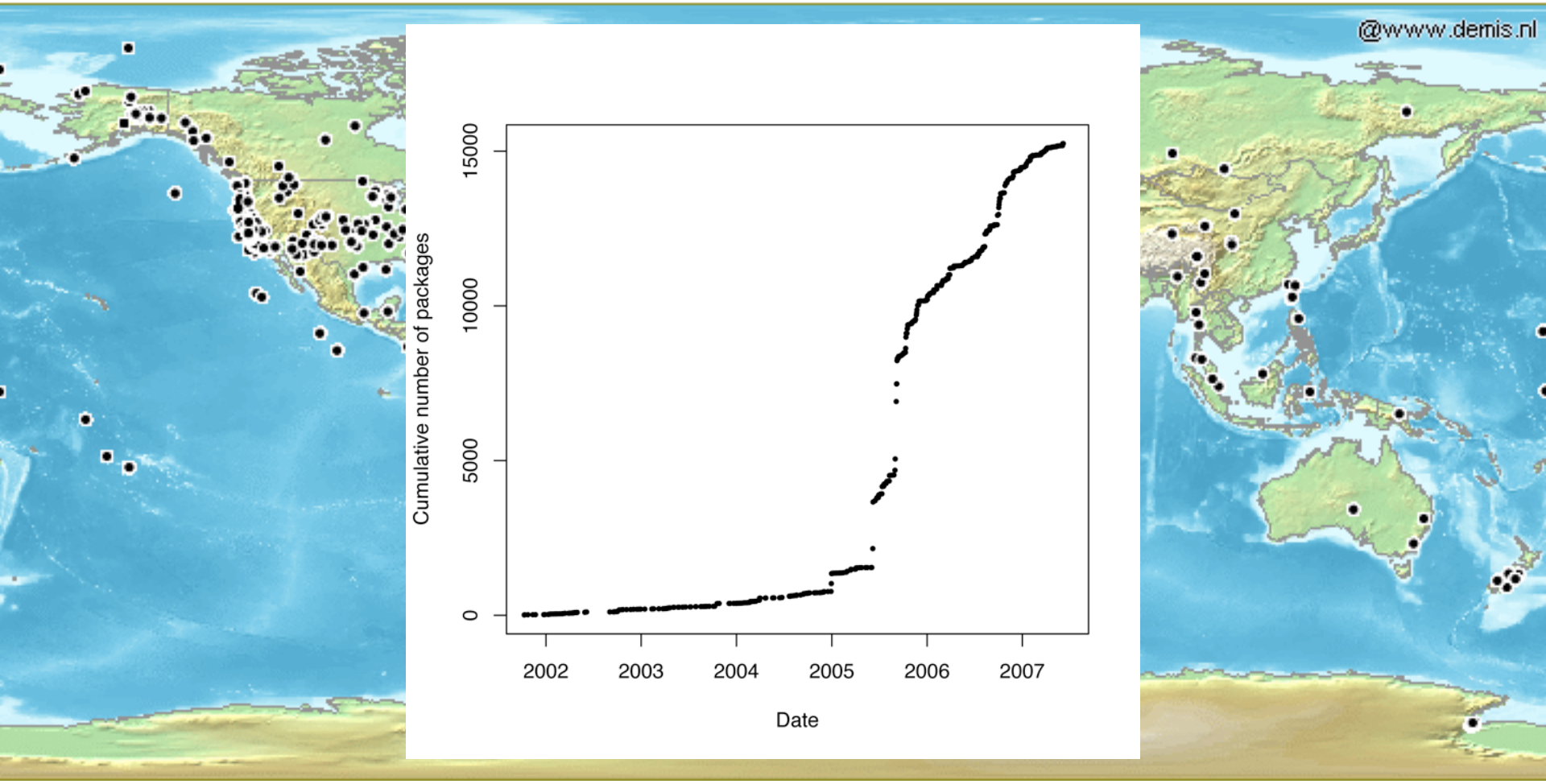
DC=Dublin Core DIF=Directory Interchange Format ECHO=EOS ClearingHOuse

EML=Ecological Metadata Language OGIS=OpenGIS

Existing CI: Enabling tools

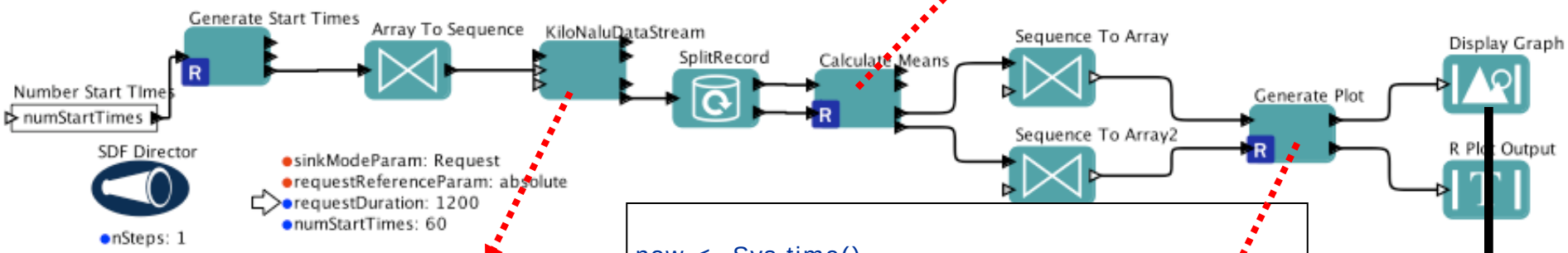


Existing CI: Metacat data distribution



Existing CI: Kepler - Kilo Nalu workflow

Support application scripts in R, Matlab, etc.

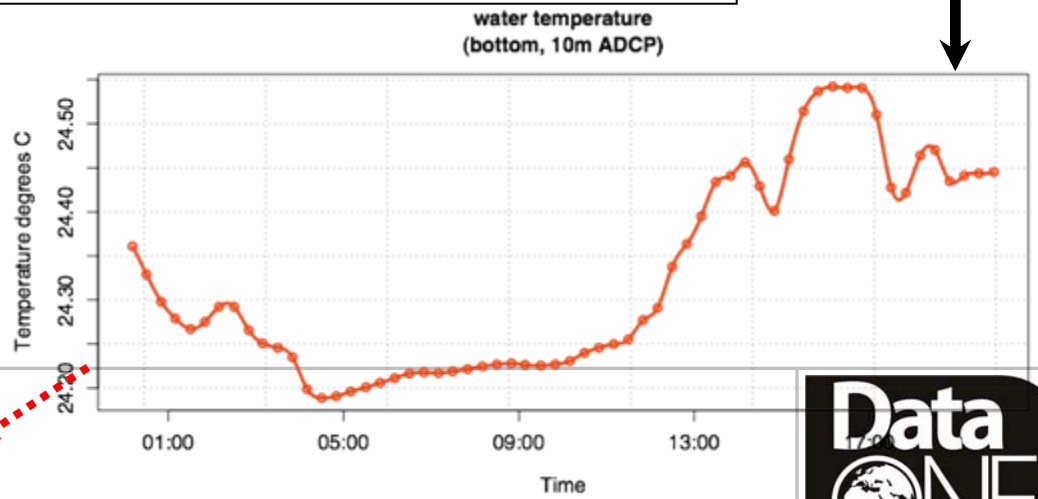


Streaming Data from observatory DataTurbine Server

```
now <- Sys.time()
Epoch <- now - as.numeric(now)
timeval <-Epoch + timestamps
posixtmedian = media
mediantime = as.num
meantemp = mean(da
```

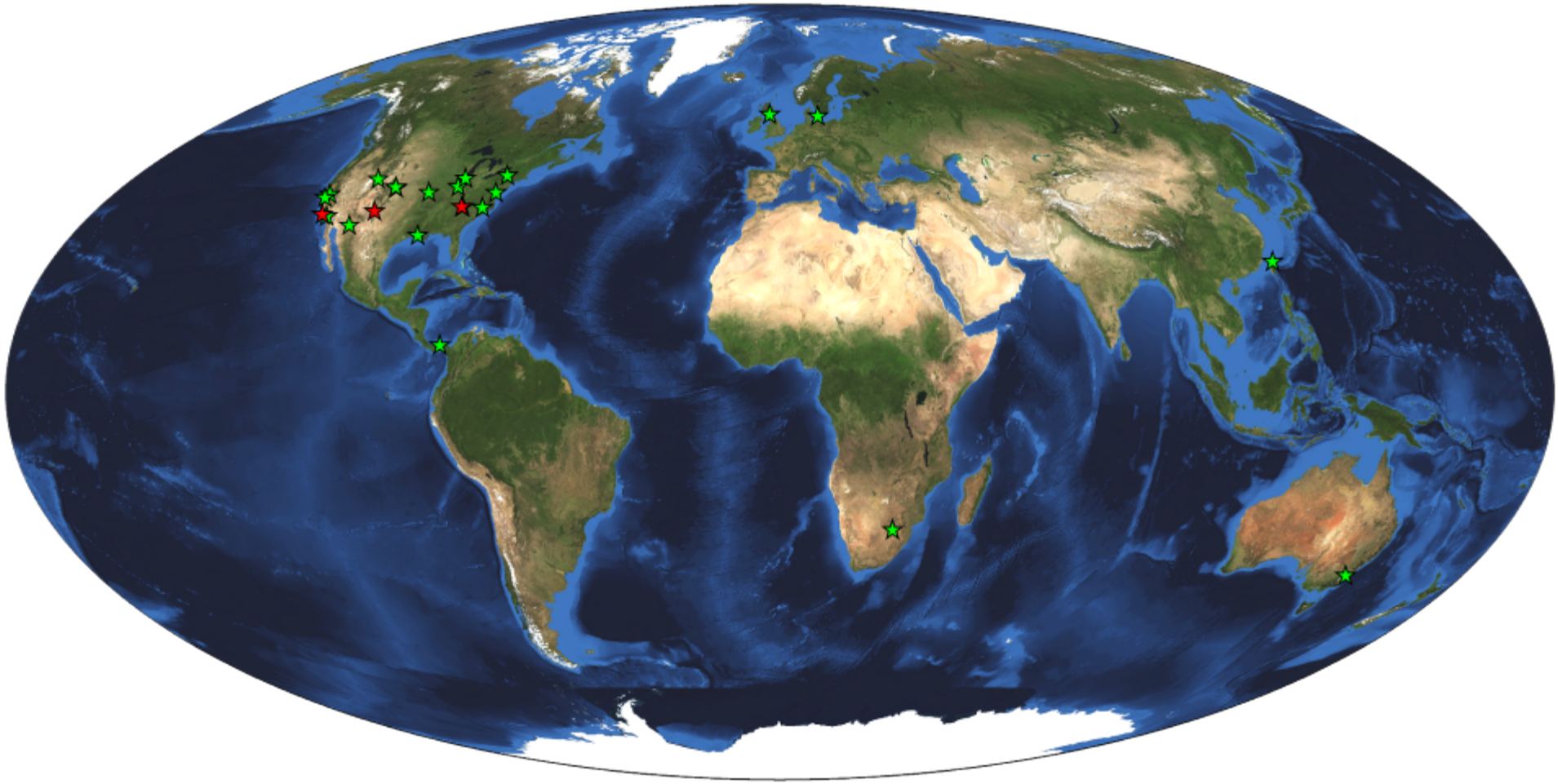
Modular components, easily saved and shared

Publish to workflow repository with accession number
Documents the linkage between publication, analysis, and data



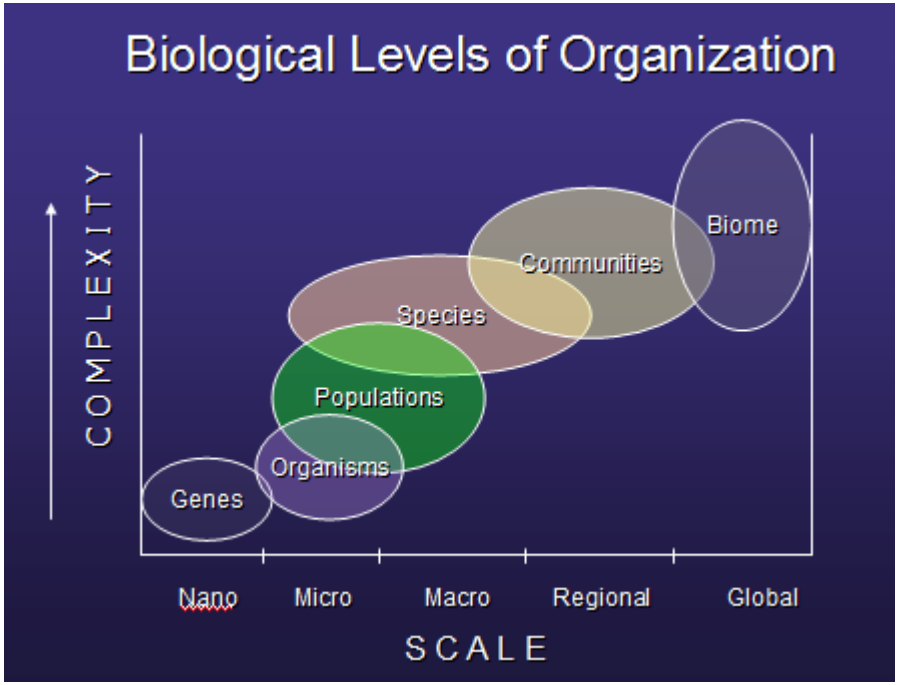
Graphs and derived data can be archived and displayed

DataONE – Building new global CI



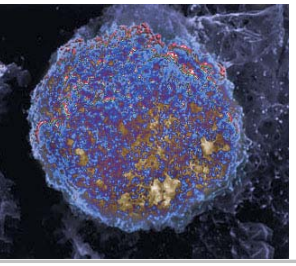
Data types

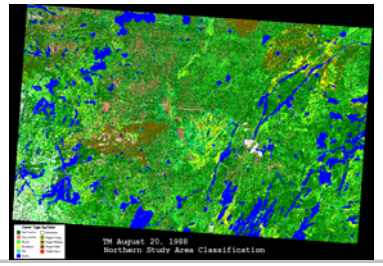
- Biological (genes to biomes)
- Environmental
  - Atmospheric
  - Ecological
  - Hydrological
  - Oceanographic



Data sources

- Research networks and environmental observatories
- Biological specimens
- Individual scientists
- Citizen scientists' data
- Natural resources and conservation data
- Observational data
- Global and continental land cover/land change and biogeochemical data



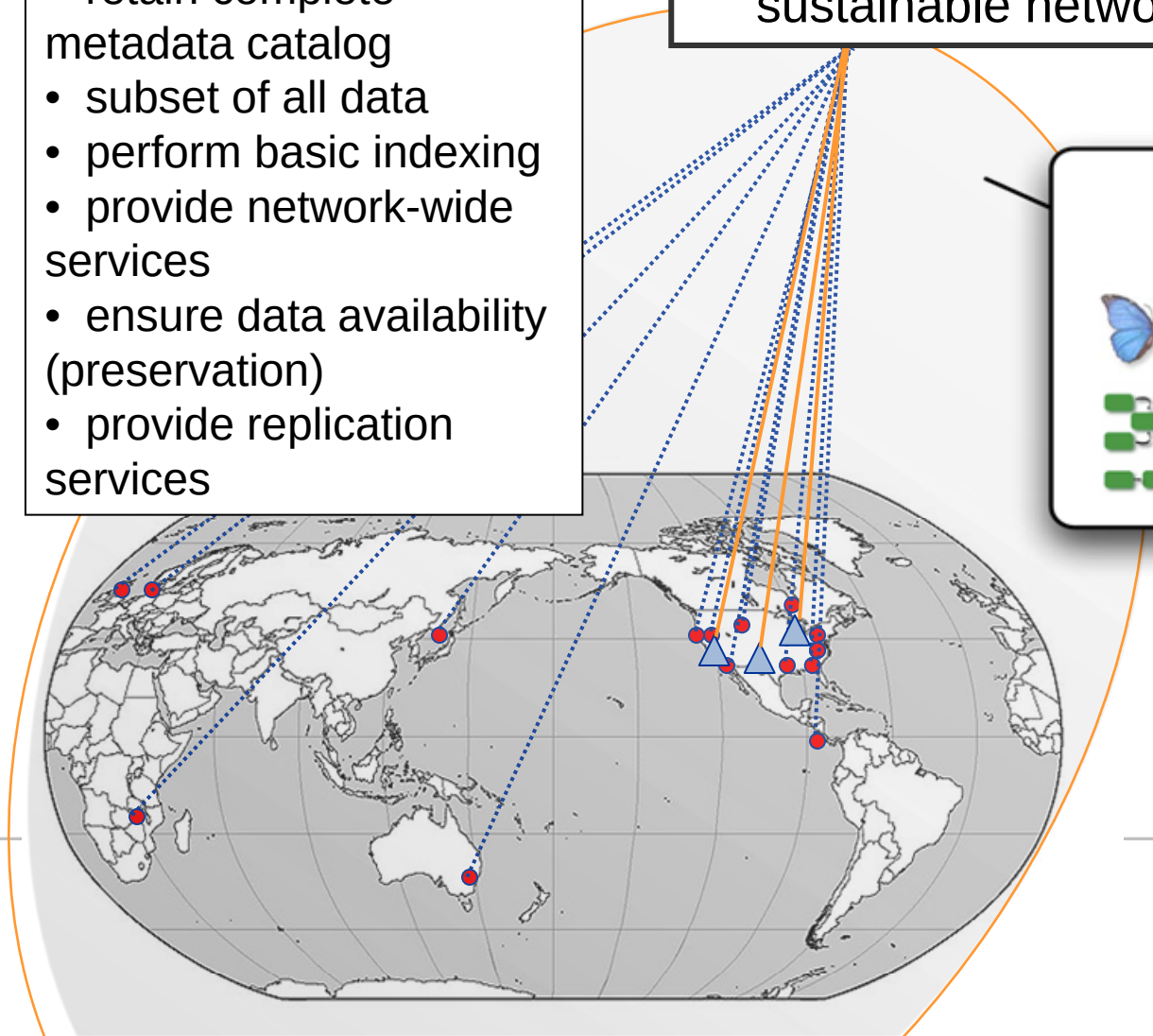


Distributed framework

Coordinating Nodes

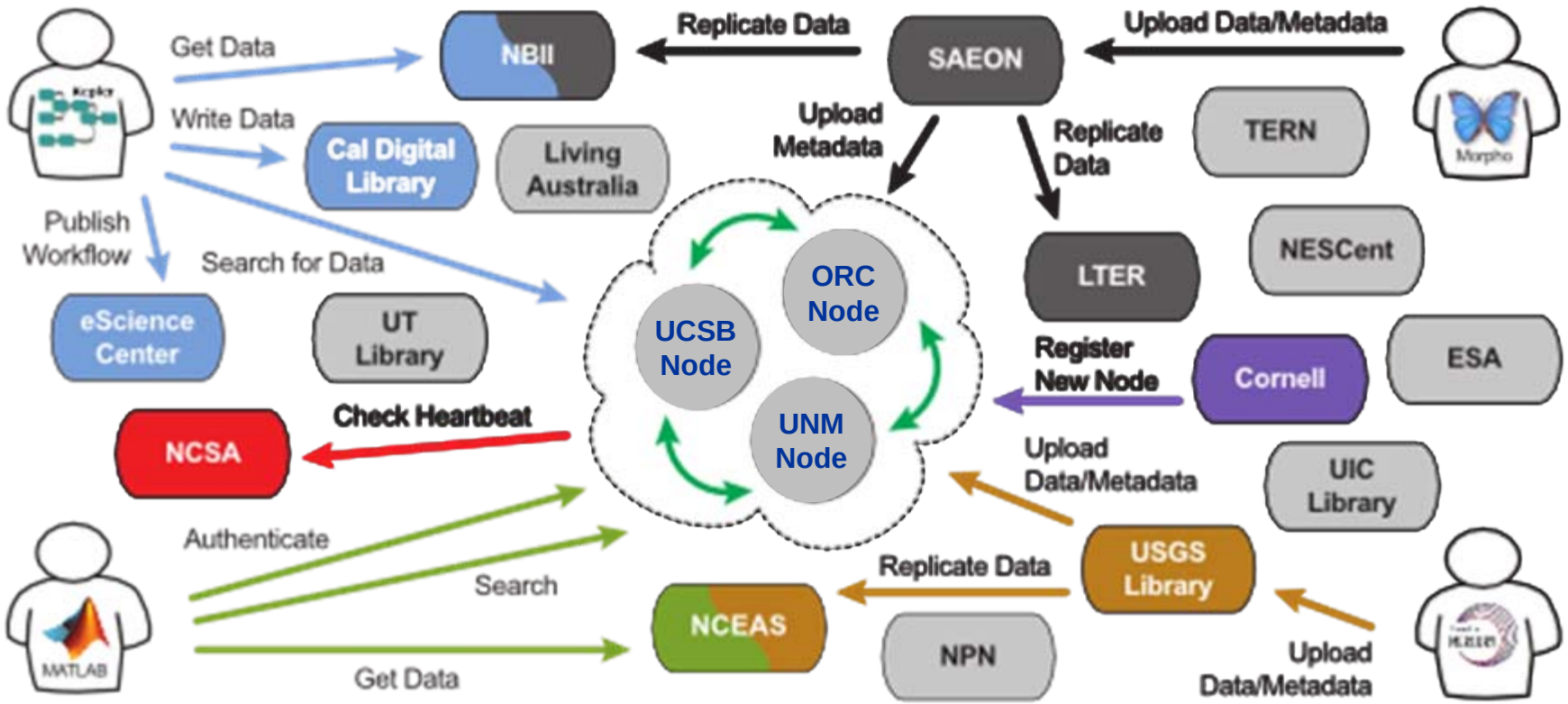
- retain complete metadata catalog
- subset of all data
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

Flexible, scalable, sustainable network



Investigator 1..N Toolkit

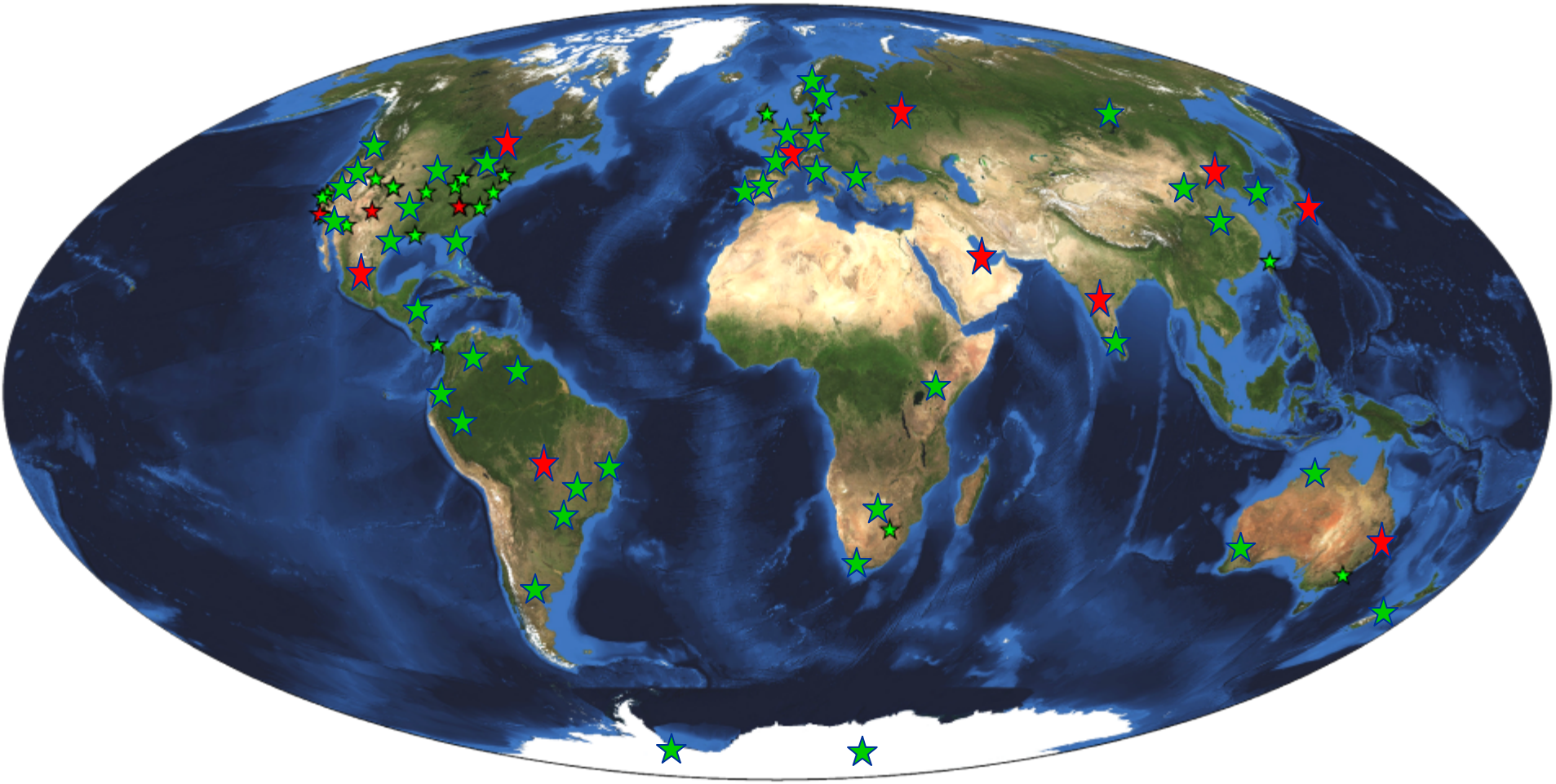
Supporting the data lifecycle



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 – Building new global CI



Providing one-stop shopping for data

Pilot Catalog: Simple Search Interface

(searches entire metadata record)

>48,000 Data Set Records

NBII Metadata Clearinghouse (39,550)

Long Term Ecological Research (LTER) Network (6,897)

ORNL Distributed Active Archive Center for Biogeochemical Data (810)

Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA) (783)

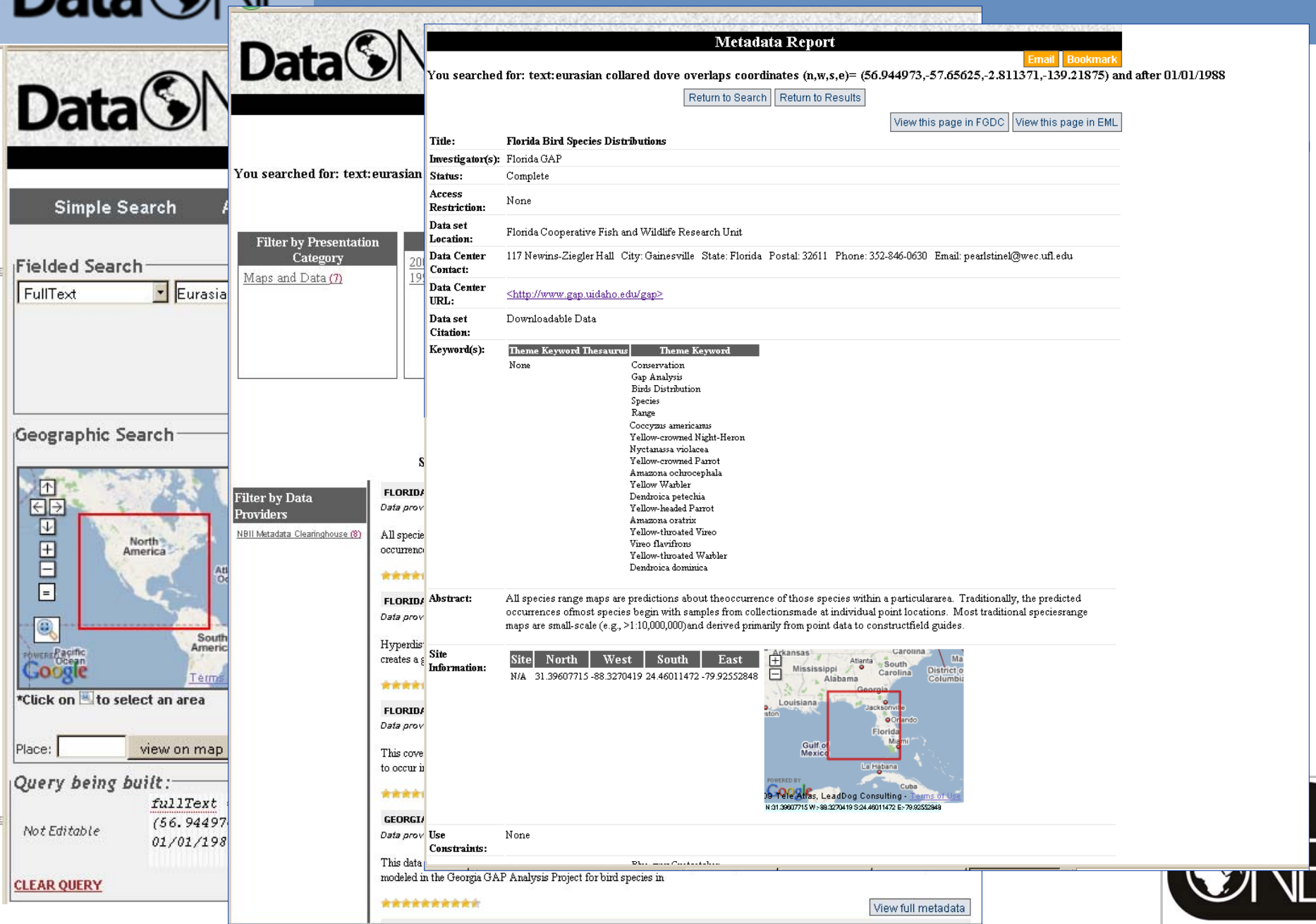
Organization of Biological Field Stations (124)

Inter-American Institute for Global Change Research (IAI) (79)

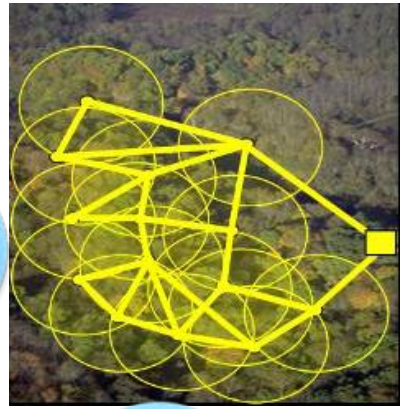
MODIS and ASTER Products (LPDAAC) (38)

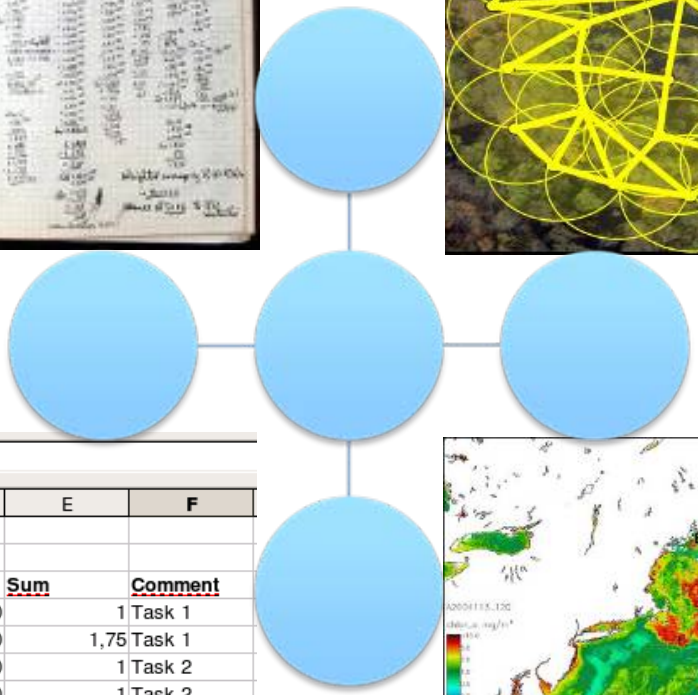
National Phenology Network (USANPN) (29)

Eurasian Collared Dove in North America



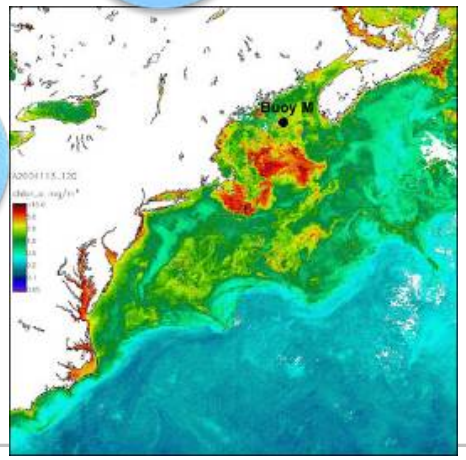
Enhancing data interoperability: ontologies and semantic mediation





F16

	A	B	C	D	E	F
1						
2						
3	Date	Start time	End time	Pause	Sum	Comment
4	2007-05-07	9,25	10,25	0	1	Task 1
5	2007-05-07	10,75	12,50	0	1,75	Task 1
6	2007-05-07	18,00	19,00	0	1	Task 2
7	2007-05-08	9,25	10,25	0	1	Task 2
8	2007-05-08	14,50	15,50	0	1	Task 3
9	2007-05-08	8,75	9,25	0	0,5	Task 3
10	2007-05-14	21,75	22,25	0	0,5	Task 3
11	2007-05-14	22,50	23,00	0	0,5	Task 3
12	2007-05-15	11,75	12,75	0	1	Task 3

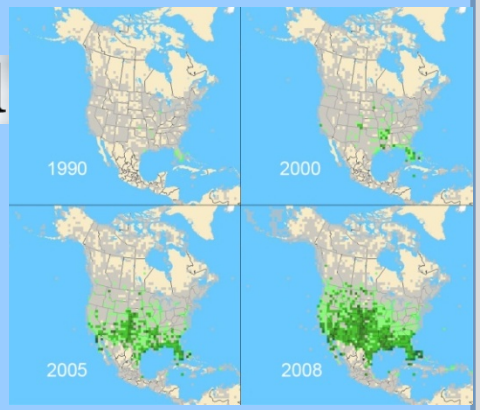


Use Case Scenario: Forecasting the spread of an invasive species across North America

The Eurasian Collared Dove (*Streptopelia decaocto*) was introduced in the Bahamas in 1988. Since then it has spread across North America. There is concern that the Eurasian Collared Dove could compete with native dove species (i.e. Mourning Dove, *Zenaida macrura,* or White-winged Dove, *Zenaida asiatica*) both of which are economically beneficial. A analyst would like to predict how the invasive dove will spread over the next 20 years, and how it might impact the ranges of the other dove species.

First, the **analyst** searches DataONE **clearinghouse** to **discover** and **access** data on the distribution of the dove species. They find that a continent-wide network of citizen scientists has gathered information on the occurrence of Eurasian Collared Dove since it was introduced.

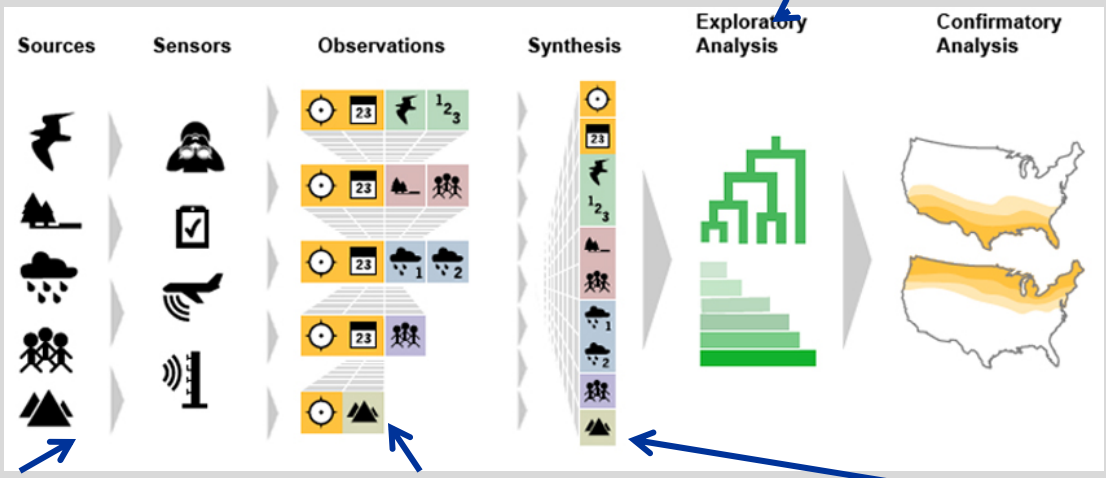
eBird



Range expansion of Eurasian Collared Dove

Second, analysis **workflows** are used to first explore and then predict patterns of species occurrence.

Exploratory analysis techniques that identify the factors that best predict species occurrence and drive hypotheses generation and predictive analysis.

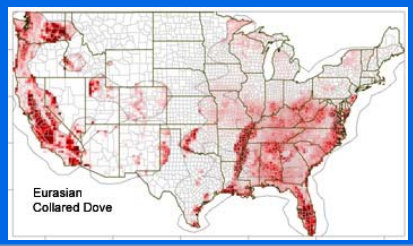


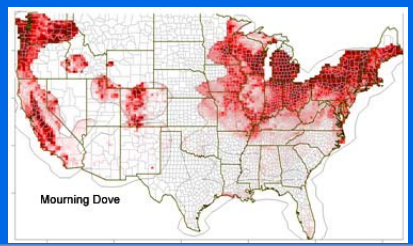
Sources of species observations are linked to landscape, climate, geographical and human factors.

Observations are available through the distributed network of DataONE data providers.

Data are organized via a core semantic model for observational data making **data synthesis** straightforward.

Third, accurate long-range forecasting **models** for each species are presented. **Predicted Ranges of 3 Dove species in 2025.**





Note: maps are examples of possible outcomes, and not actual representations of range.

Workflow sharing portal

myexperiment beta

Wiki | Mailing List | Publications

Log in | Register | Give us Feedback

Home | Users | Groups | Workflows | Files | Blogs | Forums

All | Search

Home » Tags » View: kegg

BOOKMARK

Tag Results

Items tagged with "kegg" (10)

Workflow

Uploader:

Paul Fisher

Mouse Pathways and Gene annotations for QTL Phenotype (v3)

View

Download (v3)

Created: 03/10/07 @ 18:36:02 | **Updated:** 20/02/08 @ 16:05:44

License: Creative Commons Attribution-Share Alike 3.0 License

This workflow searches for genes which reside in a QTL (Quantitative Trait Loci) region in the mouse, Mus musculus. The workflow requires an input of: a chromosome name or number; a QTL start base pair position; QTL end base pair position. Data is then extracted from BioMart to annotate each of the genes found in this region. The Entrez and UniProt identifiers are then sent to KEGG to obtain KEGG gene identifiers. The KEGG gene identifiers are then used to searcg for pathways in the KEGG path...

Rating: 4.50/5 (2 ratings) | **Versions:** 3 | **Reviews:** 0 | **Comments:** 0 | **Citations:** 1

Viewed: 673 times | **Downloaded:** 778 times

Tags (11):

data-driven | disease | genotype | kegg | nbiconworkflows | pathway | pathway-driven | pathways | phenotype | qtl | shim

New/Upload

Workflow | GO

Log in / Register

Username or Email:

Password:

Remember me:

OR

Use OpenID:

(eg: name.openid.com)

Log in

Need an account?
Click here to register

Forgot Password?

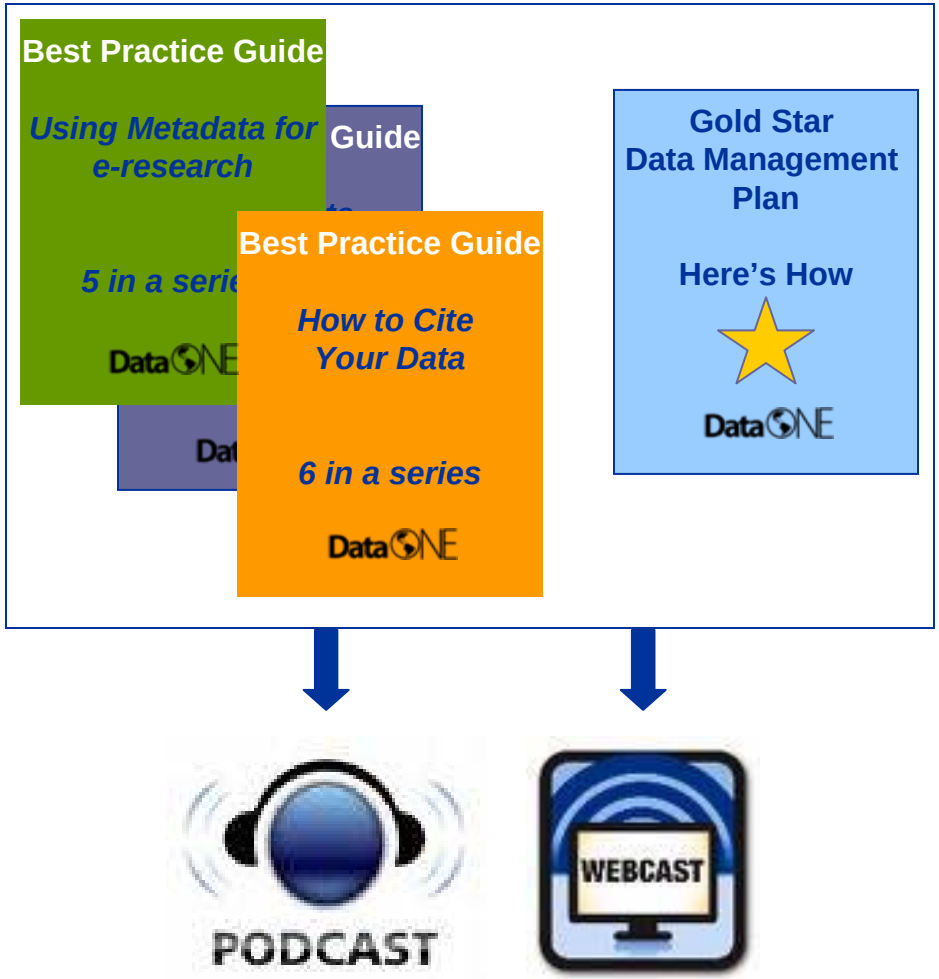
Changing science culture and institutions

1. Education and training
2. Engaging citizens in science
3. Building global communities of practice

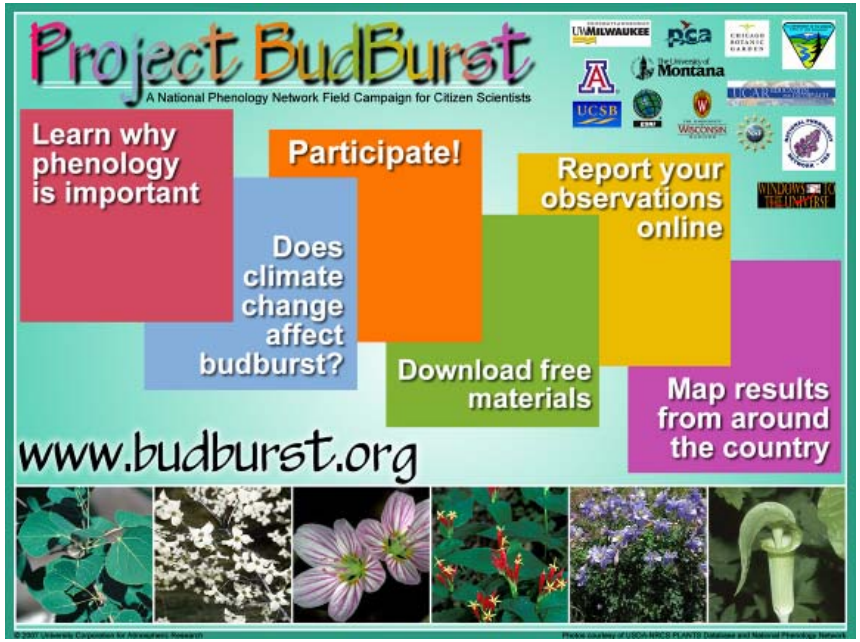
Education and training

Career Long Learning:

- best practice guides
- exemplary data management plans
- podcasts, web-casts
- workshops and seminars
- downloadable curricula



Engaging citizens in science



eBird

www.CitizenScience.org

USA npn National Phenology Network

Building global communities of practice: … creating long-lived CI enterprises,

- Broad, active community engagement
  - Involvement of library and science educators engaging new generations of students in best practices
  - Existing outreach and education programs
- Transparent, participatory governance
- Adoption/creation of innovative and sustainable business and organizational models

… engaging diverse partners.

- Libraries & digital libraries
- Academic institutions
- Research networks
- NSF- and government-funded synthesis & supercomputer centers/networks
- Governmental organizations
- International organizations
- Data and metadata archives
- Professional societies
- NGOs
- Commercial sector

Road map to today's talk

- Environmental science challenges
- Cyberinfrastructure challenges
- DataONE: A solution
- Benefits to science

What shapes DataONE?

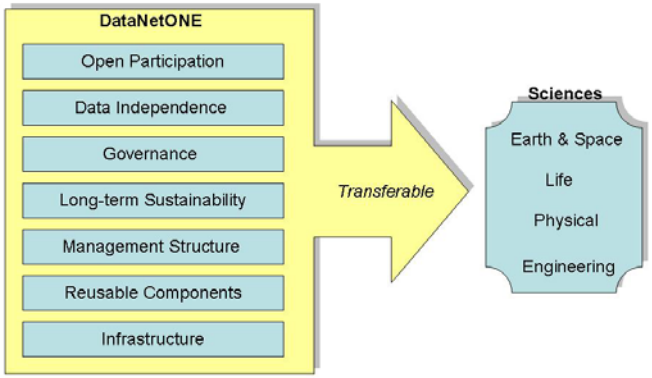
- Challenges associated with climate variability
- Community needs good data
- Good data
  - builds good science
  - makes possible wise management
  - enables sound decisions
- Good data needs
  - good technical infrastructure
  - sound organization
  - community engagement (you)

Data Matter !

Why is this relevant to USGS

- USGS Science Strategy – needs substantial data and information
- USGS & DataONE share similar Cyber infrastructure needs
  - Architecture
  - Portals
  - Distributed approaches
  - Replication
  - Secure, controlled access
  - Authentication methods
  - Tools deployed, and supported
  - Data discovery & interoperability methods
  - Standards developed, deployed
- Life Cycle Data Management tools (i.e Investigator toolkit)
- Closer collaboration with science & information partners
  - NSF, LTER, ORNL, DOE, NCEAS, etc.
- R&D activities in the areas of CS, IS, SS, GIS, Env., etc.
- Opportunity for broad USGS participation
- Complementary to USGS NBII, Library, Climate Center, Data Archiving, etc. initiatives

DataNetONE Domain Independence
Transferable Models



Current Status

- Draft Strategic Plan
- Draft Implementation Plan
- Several Workshop Reports
- Draft Charters for Working Groups
- Draft Fact Sheet
- Web site in Development
- Upcoming Events:
  - PM Training (April)
  - USGS and NASA briefings (May)
  - CI Meetings (June)
  - Formal Announcement by NSF
  - Community Engagement & Outreach – (July)
  - External Advisory Board meeting – (August/September)
  - All hands kick-off meeting (October)

Thanks!

Management Team:

Suzie Allard – *UT*
John Cobb – *ORNL*
Bob Cook – *ORNL*
Patricia Cruse – *CDL*
Mike Frame – *USGS*
Stephanie Hampton – *UCSB*
Viv Hutchison – *USGS*
Matt Jones – *UCSB*
Steve Kelling – *Cornell*
Kathleen Smith - *Duke*
Carol Tenopir – *UT*
Bruce Wilson – *Joint ORNL – UT*

Projects and Funding Sources:

DataONE Partners
Kepler-CORE Team
SEEK & KNB Teams