Apr 3rd, 12:00 AM

Panel Discussion presentation: "Data-Intensive Science with High Performance Computing Leveraging"

John W. Cobb
Oak Ridge National Laboratory

Data-Intensive Science with High Performance Computing leveraging

Presented to

Fifth Annual University of Massachusetts and New England Area Librarian e-Science Symposium Afternoon Panel

John W. Cobb, Ph.D.
Physicist
Computer Science & Mathematics Division

Shrewsbury, Massachusetts
April 3, 2013

U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Acknowledgements

- DataONE project (PI Michener, U. New Mexico)
- Oak Ridge National Laboratory and the Oak Ridge Leadership Computing Facility
- Cornell Lab of Ornithology eBrid project and S. Kelling, D. Fink, K. Webb, T. Damalou, (Cornell)
- Collaborators: M. Jones (UCSB) C. Tenopir (UTK), S. Allard (UTK), B. Wilson (ORNL/UTK), D. Vieglais (Kansas)

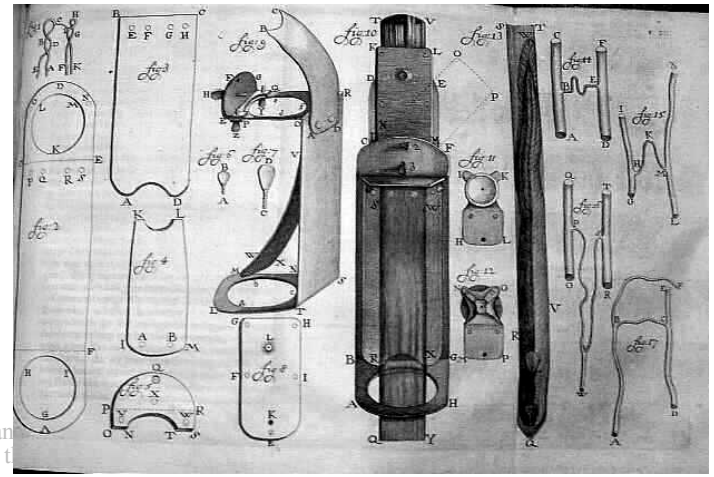
DataONE Community

Outline

- Data Begets Science
- The data lifecycle – the workflow of data driven science
- Data at Scale
- HPC at Scale
- Pathfinder exemplar: eBird occurrence maps
- Data management challenges
- DataONE project
- Dryad
- Role of libraries as data repositories
- DMPTool
- Open data movement

Data Gives Birth to Scientific Revolutions

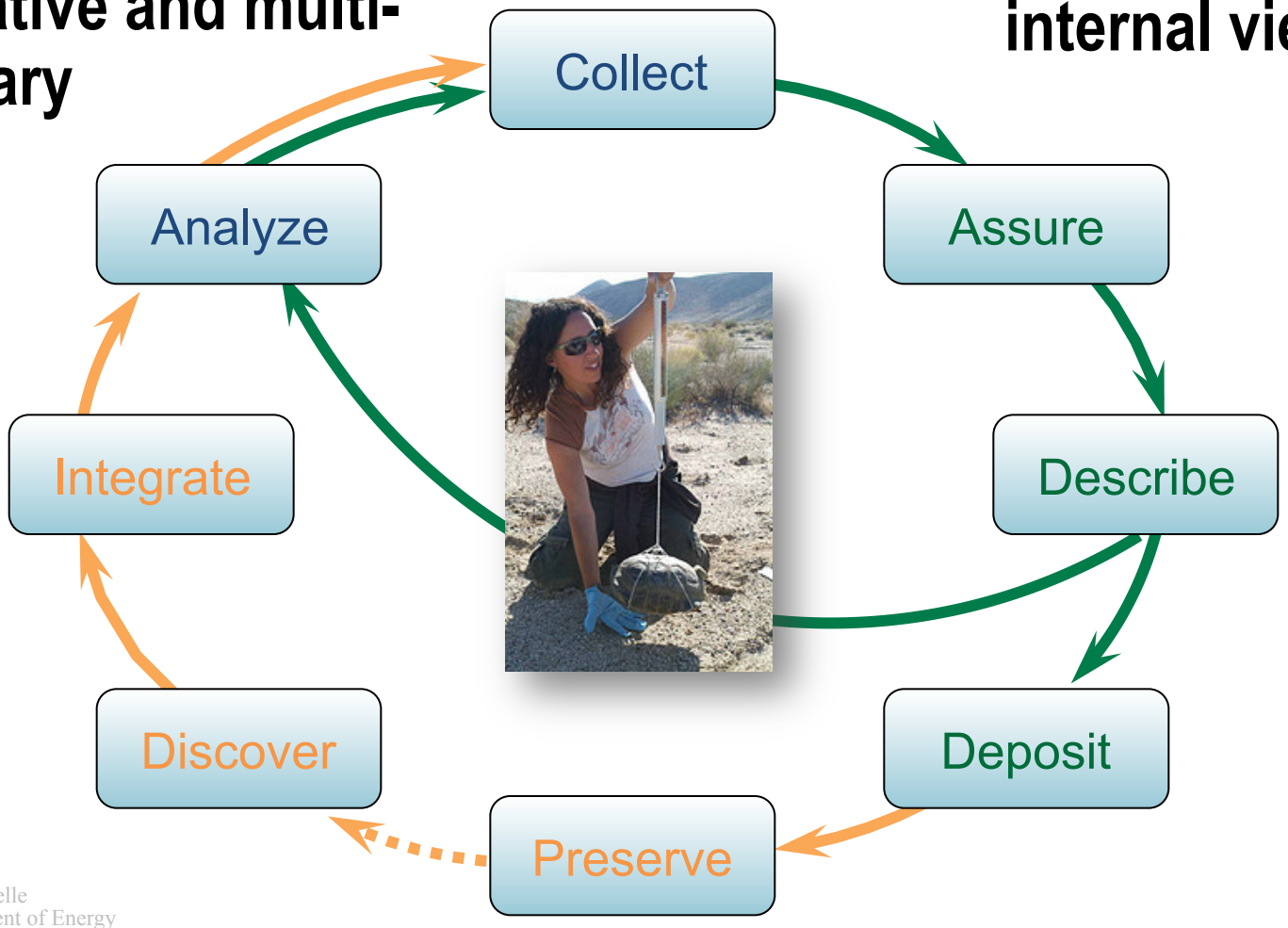
- Kepler's laws were divined by careful examination of Brahe's recorded observations
- Leeuwenhoek's founding of microbiology was triggered by observations with newly developed microscope.



The data lifecycle: the workflow of science

The conduct of science is collaborative and multi-disciplinary

Refined DataONE internal view



User Matrix (DataONE)

Different team members care about different things

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

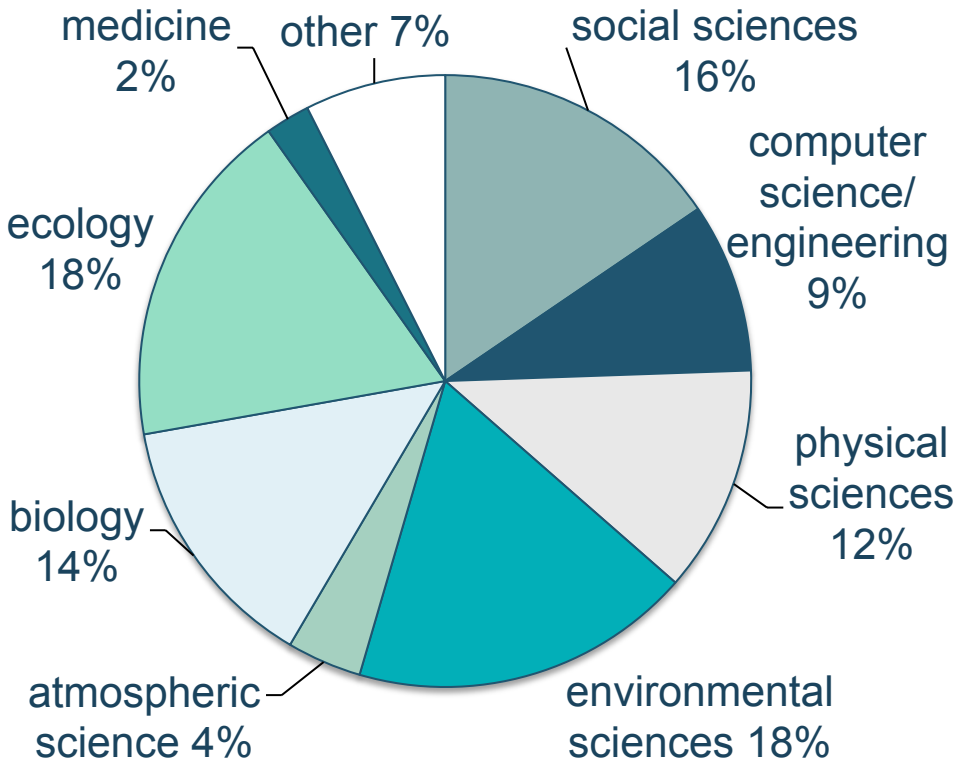
Can we share data along the data lifecyle?

- **Demographics**

Baseline assessment: scientists (2010)

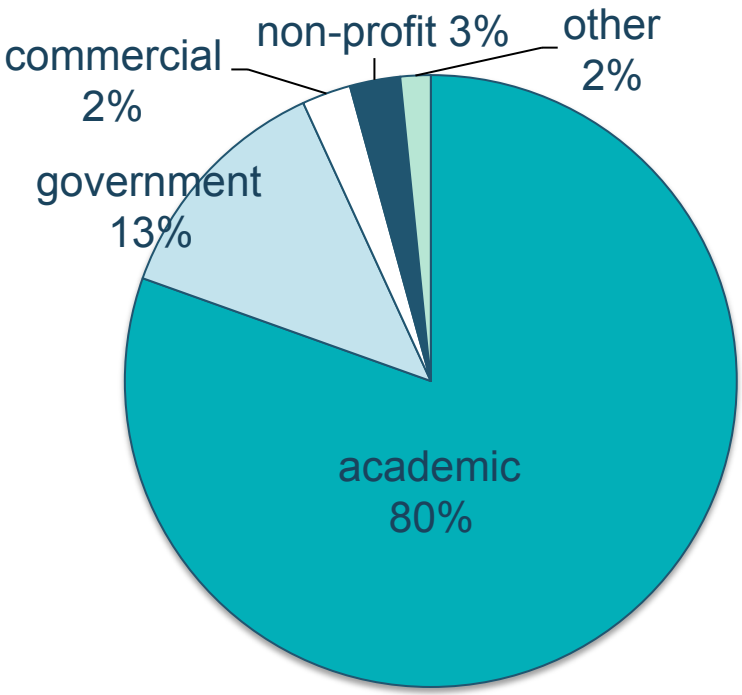
Tenopir, C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Frame M. 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE. 6(6)

Discipline



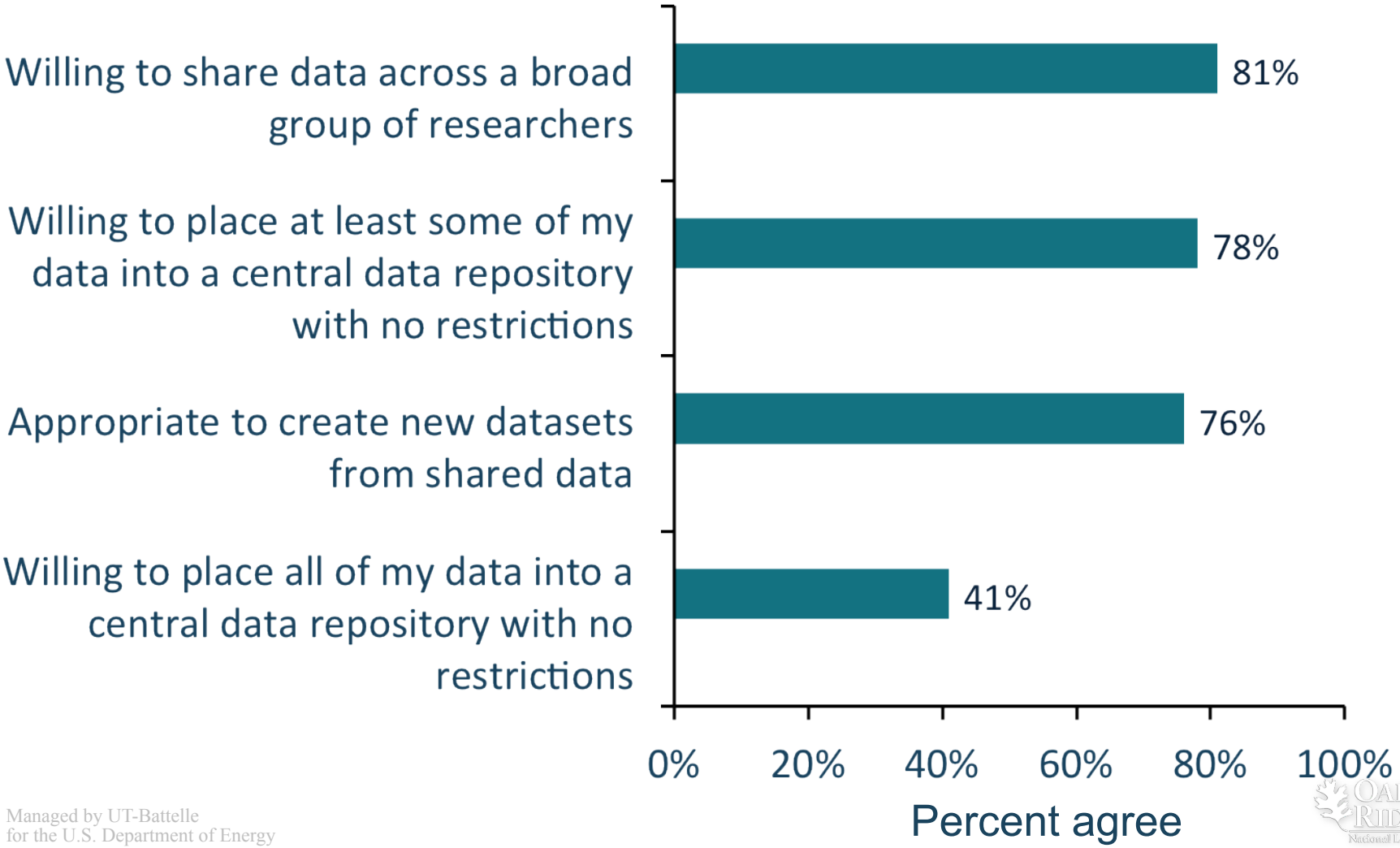
n=1317

Work Sector

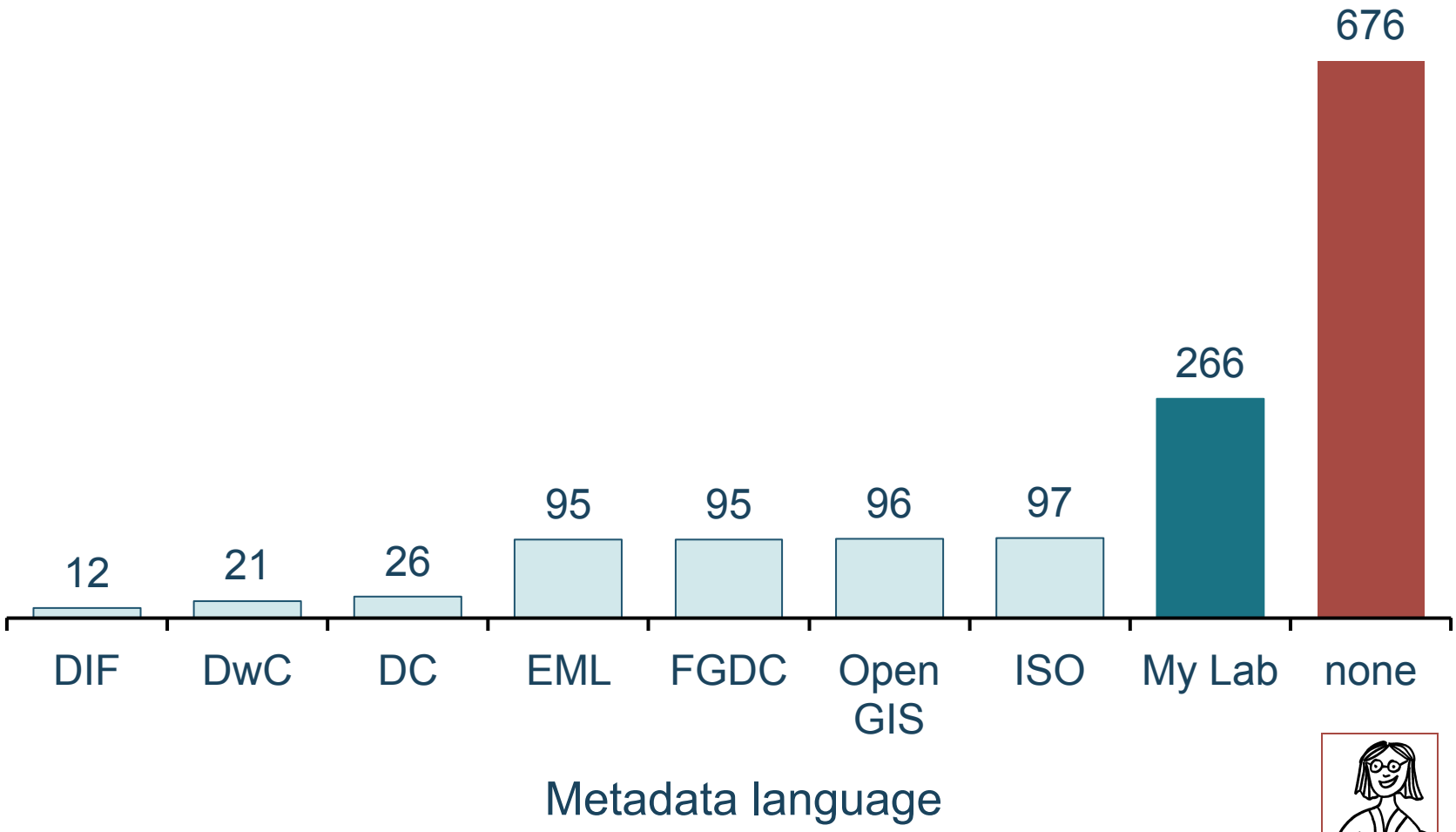


n=1315

Many are interested in sharing data



What standard do you currently use?



Answer: Yes!

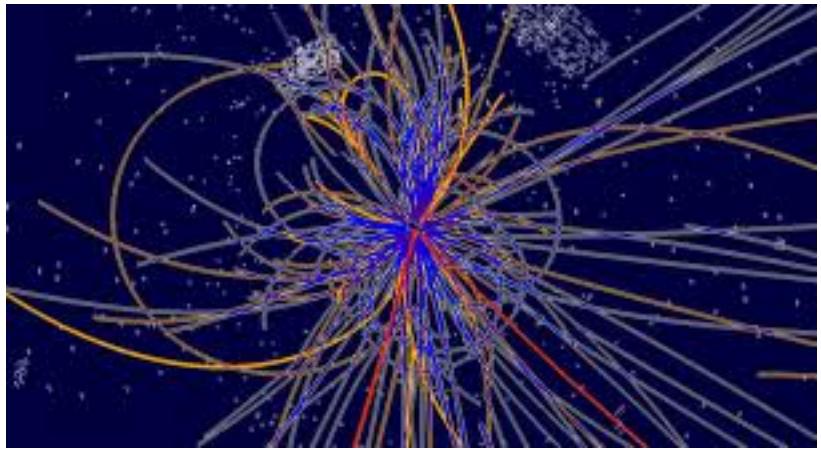
But: There is a gap between desire and practice.

This indicates an opportunity to improve practice and improve science outcomes

"The spirit is willing but the flesh is weak"

How big is big data?

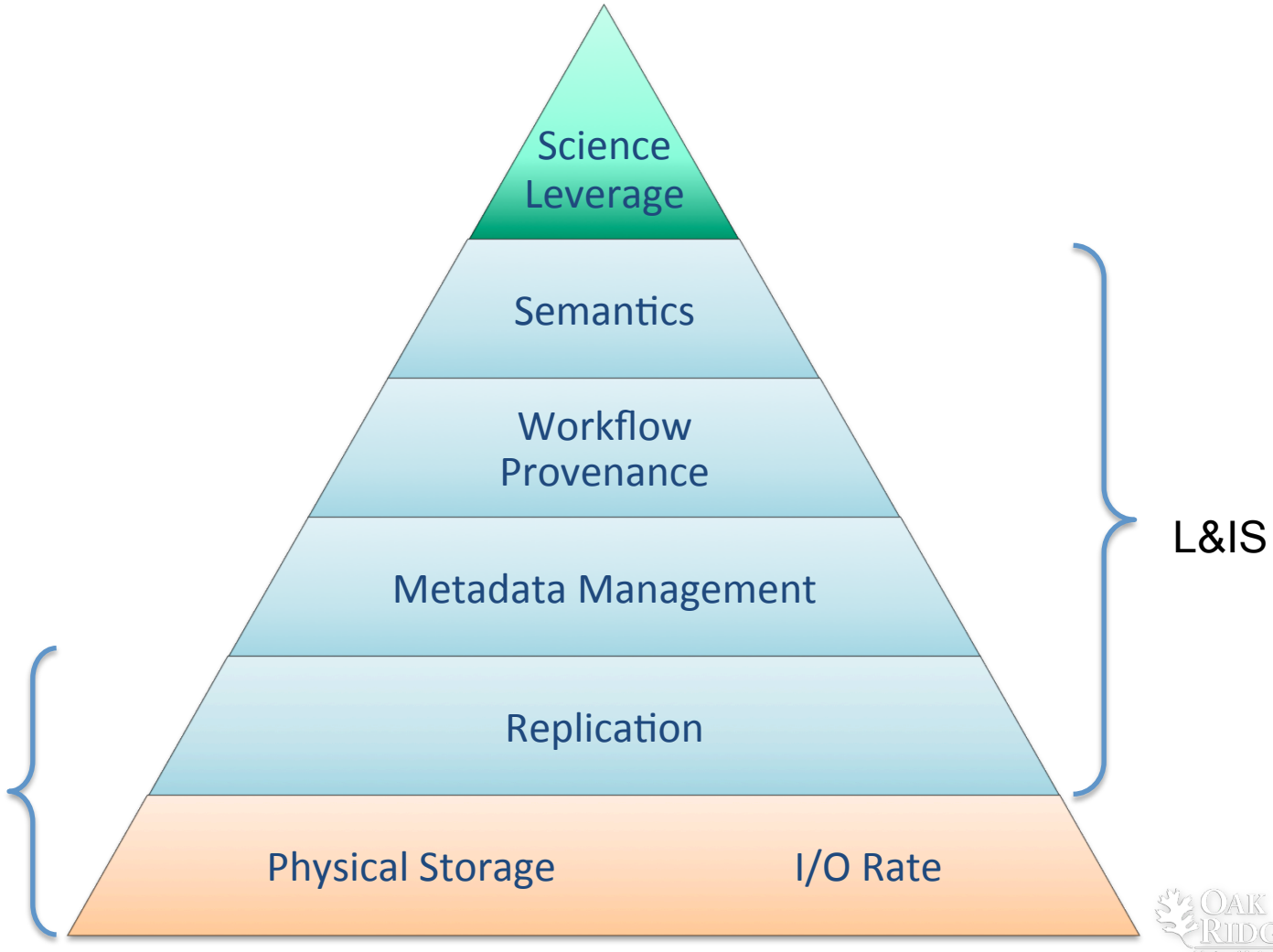
- Possible answers:
  - the largest of all datasets ever created (>10 PB)
  - The largest of all datasets ever created in each discipline
  - larger than we are comfortable managing
  - larger than what we dealt with last week/year/decade



How big is big data?

- Possible answers:
  - the largest of all datasets ever created (>10 PB)
  - The largest of all datasets ever created in each discipline
  - larger than we are comfortable managing
  - larger than what we dealt with last week/year/decade
- But larger question: what is the measure of data size?

Data Ecosystem

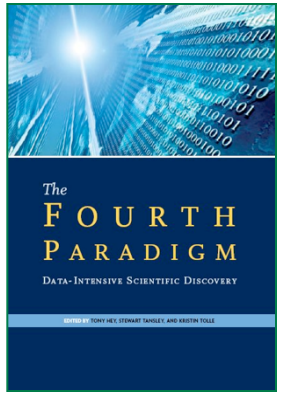


Where are the opportunities?

- **Integrating storage management and information management**
- **Integrating data from different data activities**

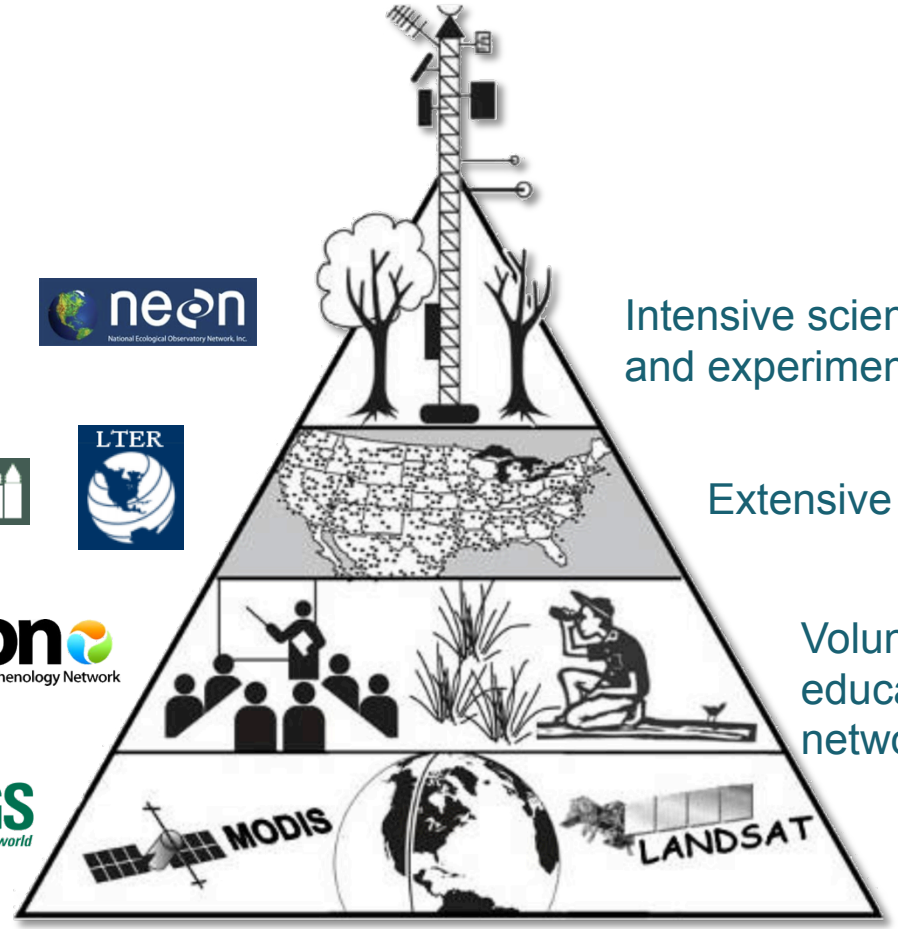
"Building the Knowledge Pyramid"

90:10 → 10:90



Decreasing Spatial Coverage

Increasing Process Knowledge



Intensive science sites and experiments

Extensive science sites

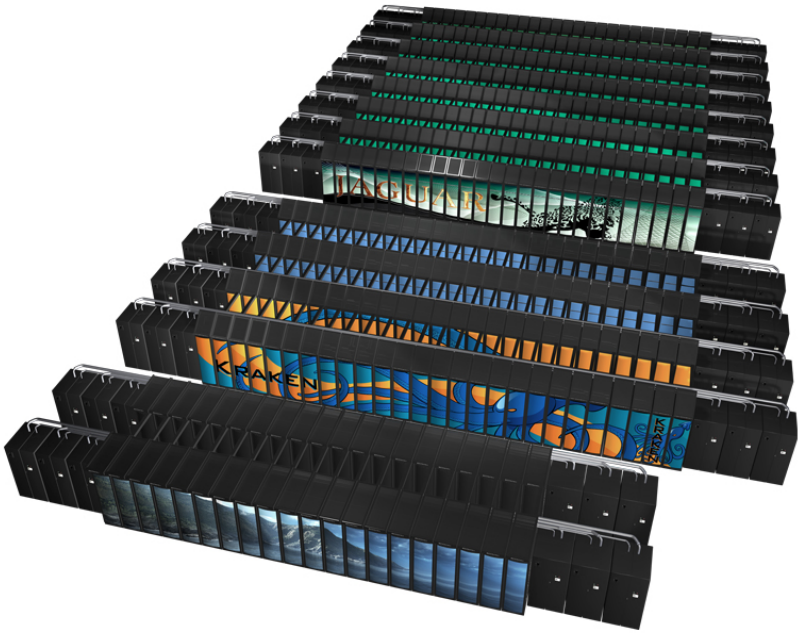
Volunteer & education networks

Remote sensing

Adapted from CENR-OSTP

HPC at scale – example Titan at OLCF

- **Physical plant challenges:**
  - **Size: 40,000 sq-ft (2 floors)**
  - **Power: 10's of MW**
  - **Cooling: dual loops chilled water**
  - **Raised floor high-load capacity (36" , 250 lbs/sq-ft)**

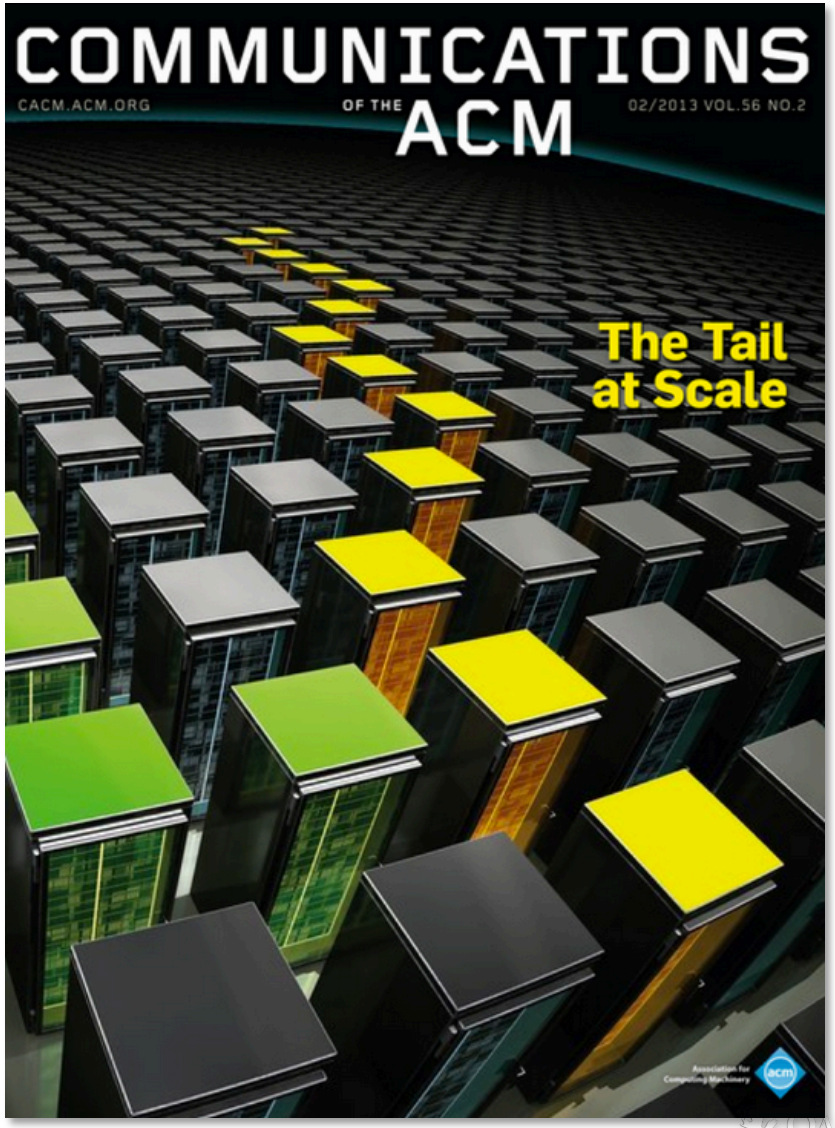


HPC at scale – example Titan at OLCF

- Named Titan
- 27 Petaflops, 710 TB memory
- Spider storage > 10 PB, 250 GB/s
- 8972 GPU-enabled nodes (Kepler) in 200 cabinets
- Each node contains: One AMD 16-core intelagos CPU, one Nvidia K20x Kepler, 32 GB memory
- Note: NVIDIA offers K20x for desktop

Data and the Long Tail of Science

- As data gets larger, the data tail is now quantifiable*: flocks of black swans*
- Extraordinary events are often the most interesting
  - "500 year storms"
  - Best candidate materials (second place is first loser)
  - Very non-uniform utility functions.
- Conclusion: applying large data analysis can create new breakthroughs



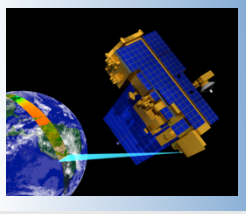
eBird pilot project exploration and visualization

eBird

Land Cover

Meteorology

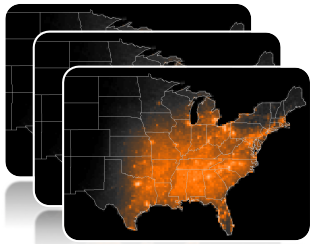
MODIS – Remote sensing data



TheCornellLab of Ornithology

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

TeraGrid



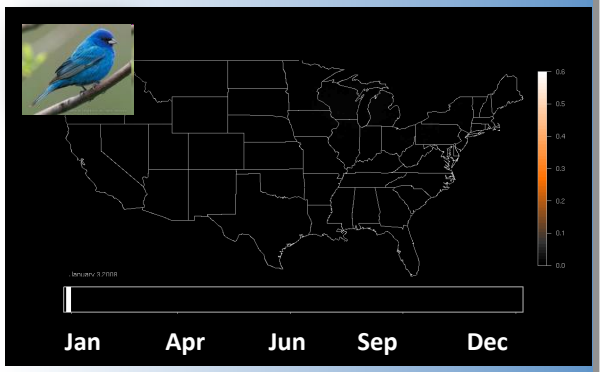
$$F(X,s,t) = \frac{1}{n(s,t)} \sum_{i=1}^{m} f_i(X,s,t) I(s,t \in \theta_i)$$

Spatio-Temporal Exploratory Model identifies factors affecting patterns of migration

DataONE

Model results

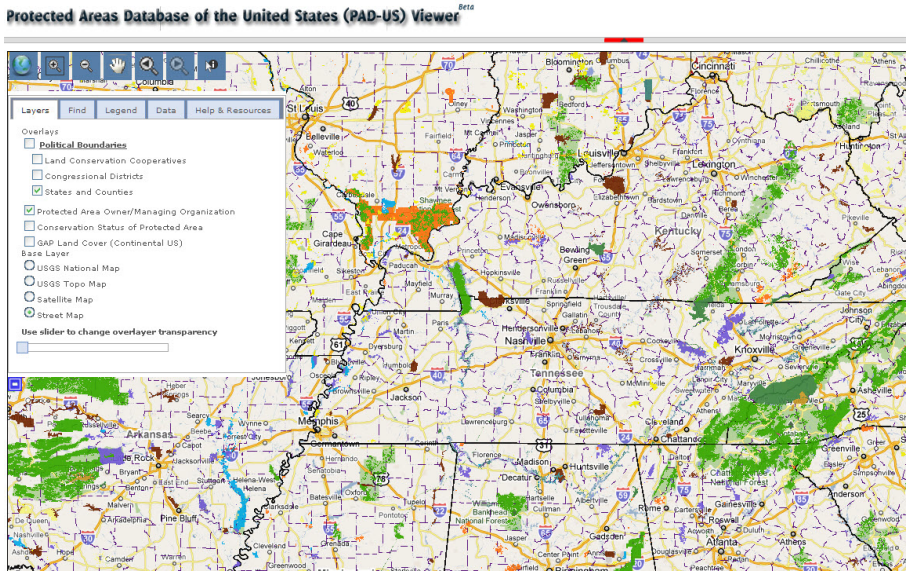
Occurrence of Indigo Bunting (2008)

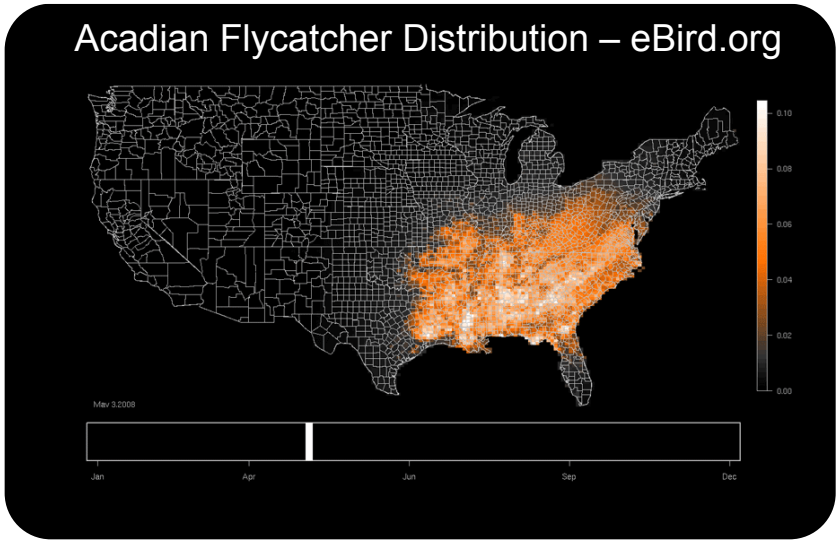


- Examine patterns of migration
- Infer how climate change may affect bird migration

Secretary Salazar on Birds (May 3, 2011):

"The State of the Birds report is a measurable indicator of how well we are fulfilling our shared role as stewards of our nation's public lands and waters."





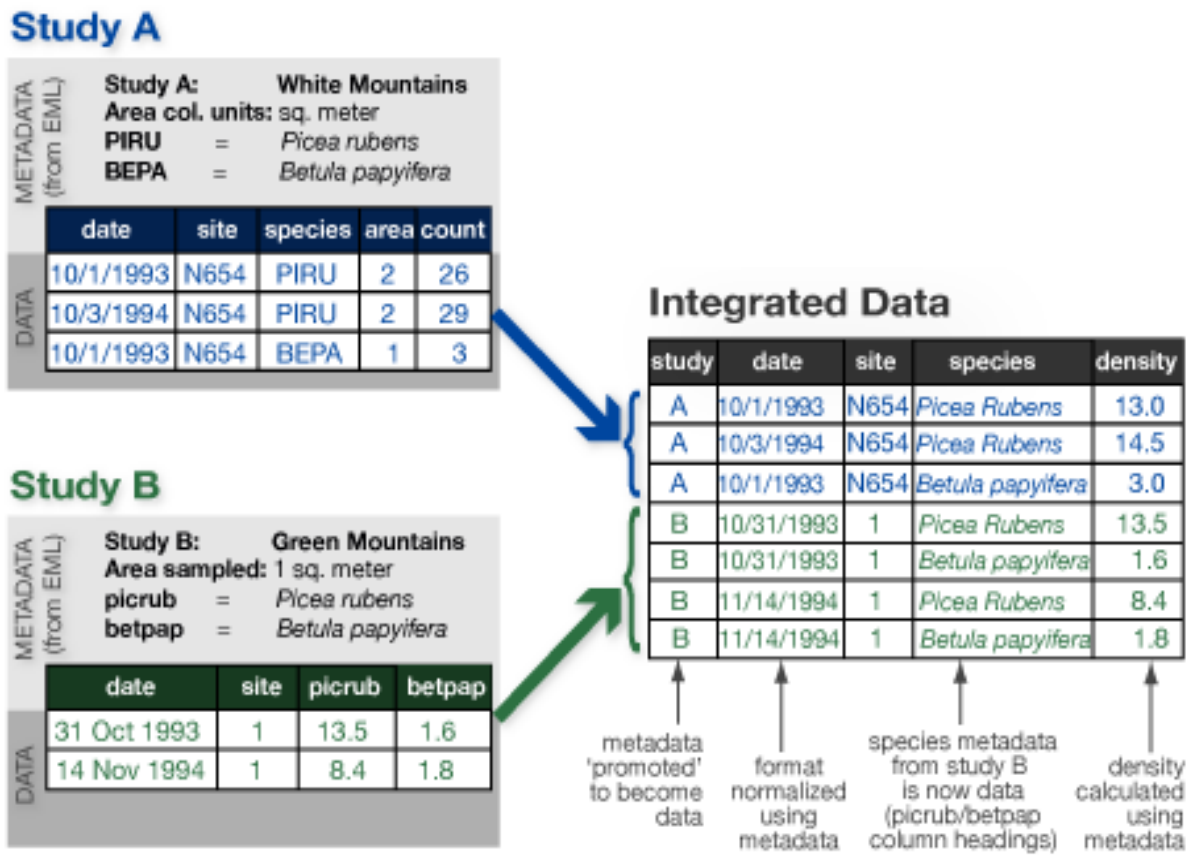
HPC centers and data management

- Often HPC focused – cycles (and storage)
- Data and information management may be a foreign culture
- HPC can enable extreme scalability: *"What would you do if you had unlimited computing/storage/badnwidth?"*
- Bottlenecks:
  - Data management issues
  - Metadata creation and harmonization
  - Data preservation
  - Items not scaling with Moore's law: metadata, human effort

Data deluge and interoperability

"the flood of increasingly heterogeneous data"

- Data are heterogeneous
  - Syntax
    - (format)
  - Schema
    - (model)
  - Semantics
    - (meaning)

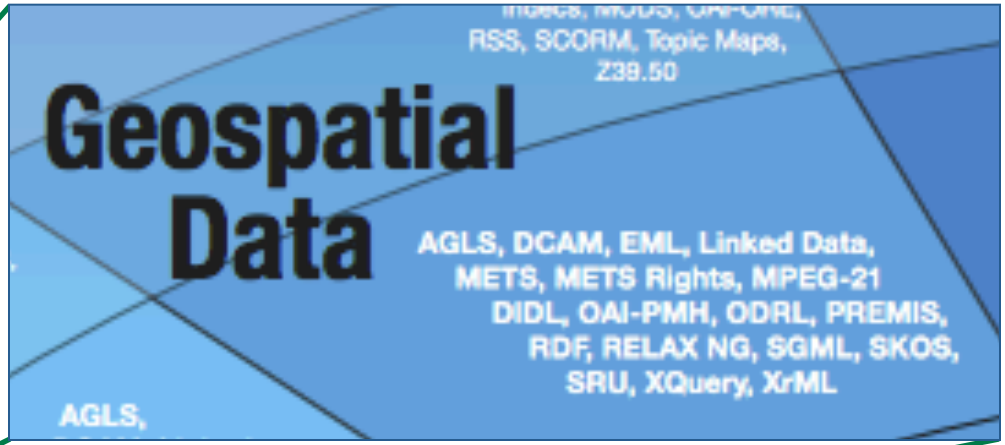


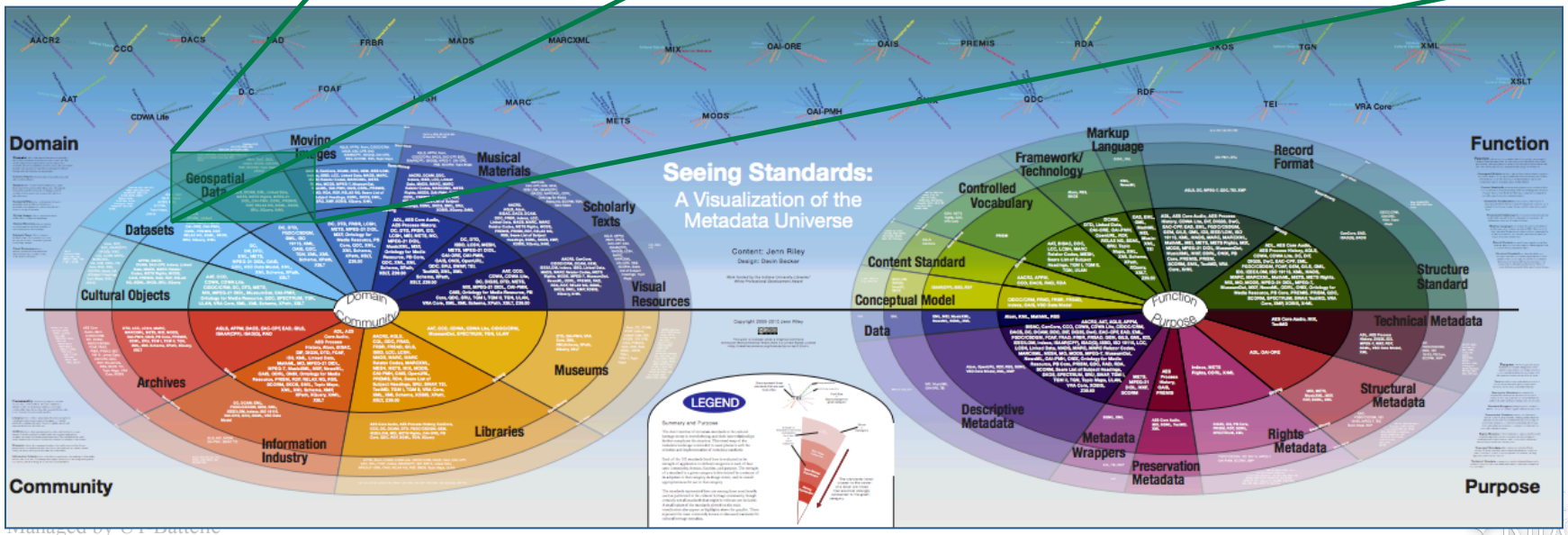
By hand is time-consuming and brittle

Jones et al. 2007

Myriad Metadata Standards

For instance: Metadata Crosswalks

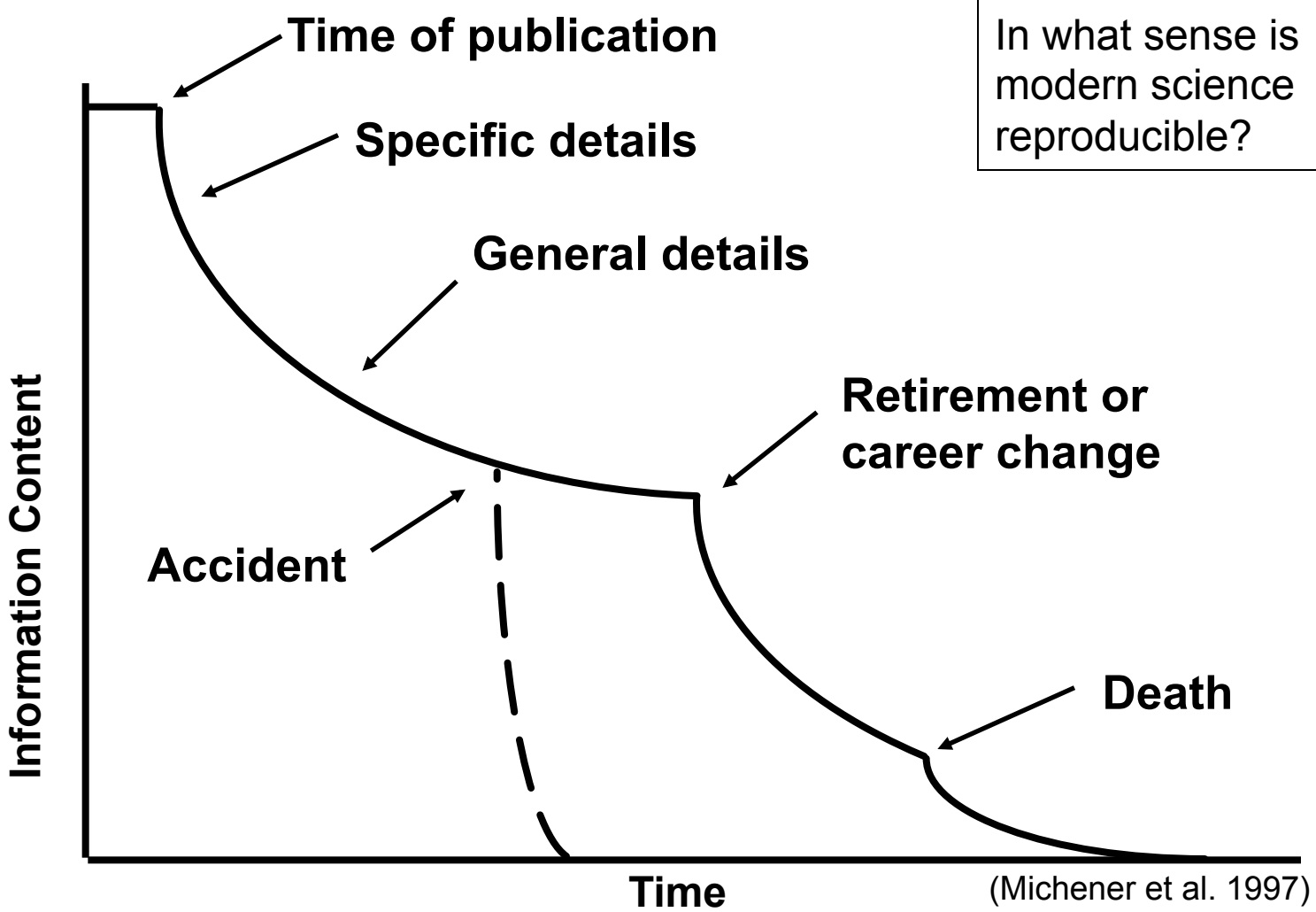




Credit: Jenn Riley Indiana University Digital Library Program 2012

Poor data practice

"data entropy"



(Michener et al. 1997)

In what sense is modern science reproducible?

DataONE project (movie with sound)

http://vimeo.com/36383735

DataONE Component Interdependency

Scientists:

Receive: Access to more data sources and tools

Provide: Scientific progress and acknowledgment

Member Nodes:

Receive: Additional users, replication, communities of best practice, appreciation

Provide: Access to data collections, service interfaces

DataONE:

Receives: MN and scientist appreciation, access to MN data

Provides: "Glue" services
to enable interoperability, communities of best practice, standard interfaces

Funders:

Receive: More efficient science output, chances for breakthrough advances

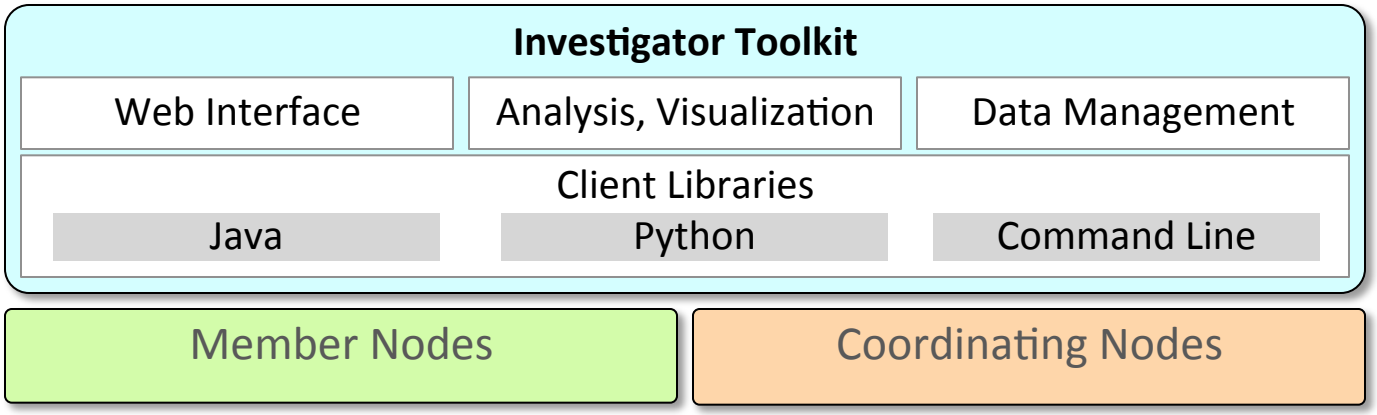
Provide: Resources to facilitate science

Current Operational Member Nodes

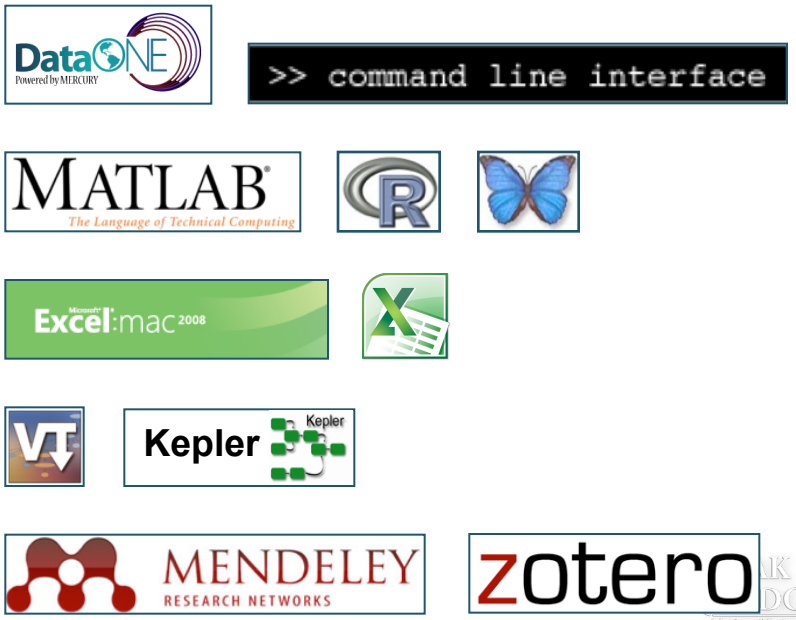
- Released production CI 10 months ago
- Today: 13 production Member Nodes
- 300,000 Data objects represented

- Near-term 15 more candidates

The Investigator Toolkit



- **Developer, end-user tools**
- **Creation, search, retrieval, management**
- **Plugins, extensions for analysis tools**



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

UC3EZID long-term identifiers made easy

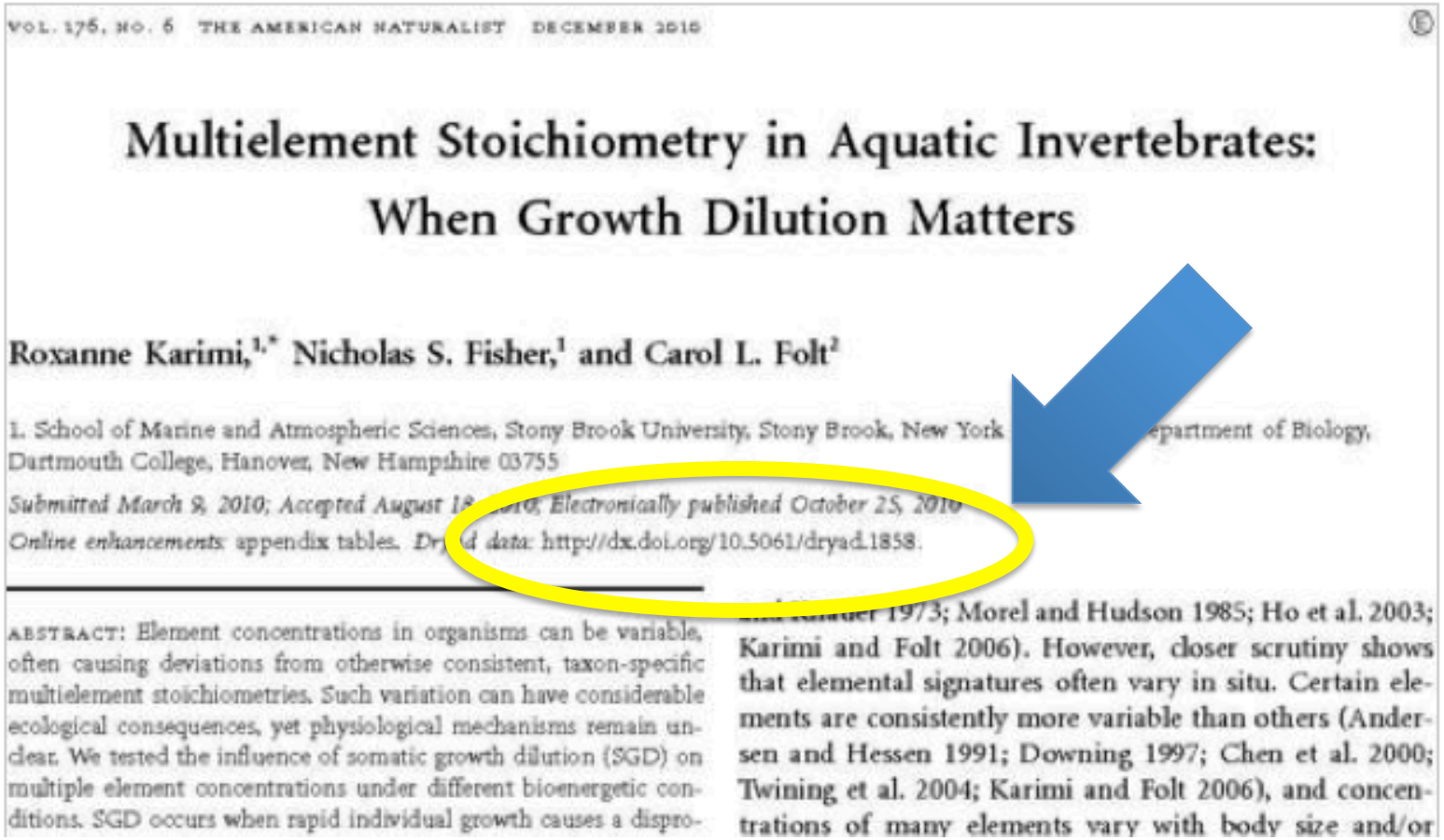
doi>®

URI
URL URN

GUID LSID PURL Handle System®

{3F2504E0-4…

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 epartment of Biology, Dartmouth College, Hanover, New Hampshire 03755

Submitted March 9, 2010; Accepted August 18 010; Electronically published October 25, 2010

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

ABSTRACT: Element concentrations in organisms can be variable, 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-

1973; 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 vary with body size and/or

- Data citation standards and courtesy customs
- Needs to metrics – how often cited
- Socio-cultural change: include data citations in promotion and tenure
- DataONE needs to nurture Member Node needs not work against them

Identify people: federated identity

- Identity provider selected by the user
- Member nodes define access rules
- Rules propagated by Coordinating Nodes
- Identity and access control consistent across entire infrastructure
- (note similarity with Globus Online approach)

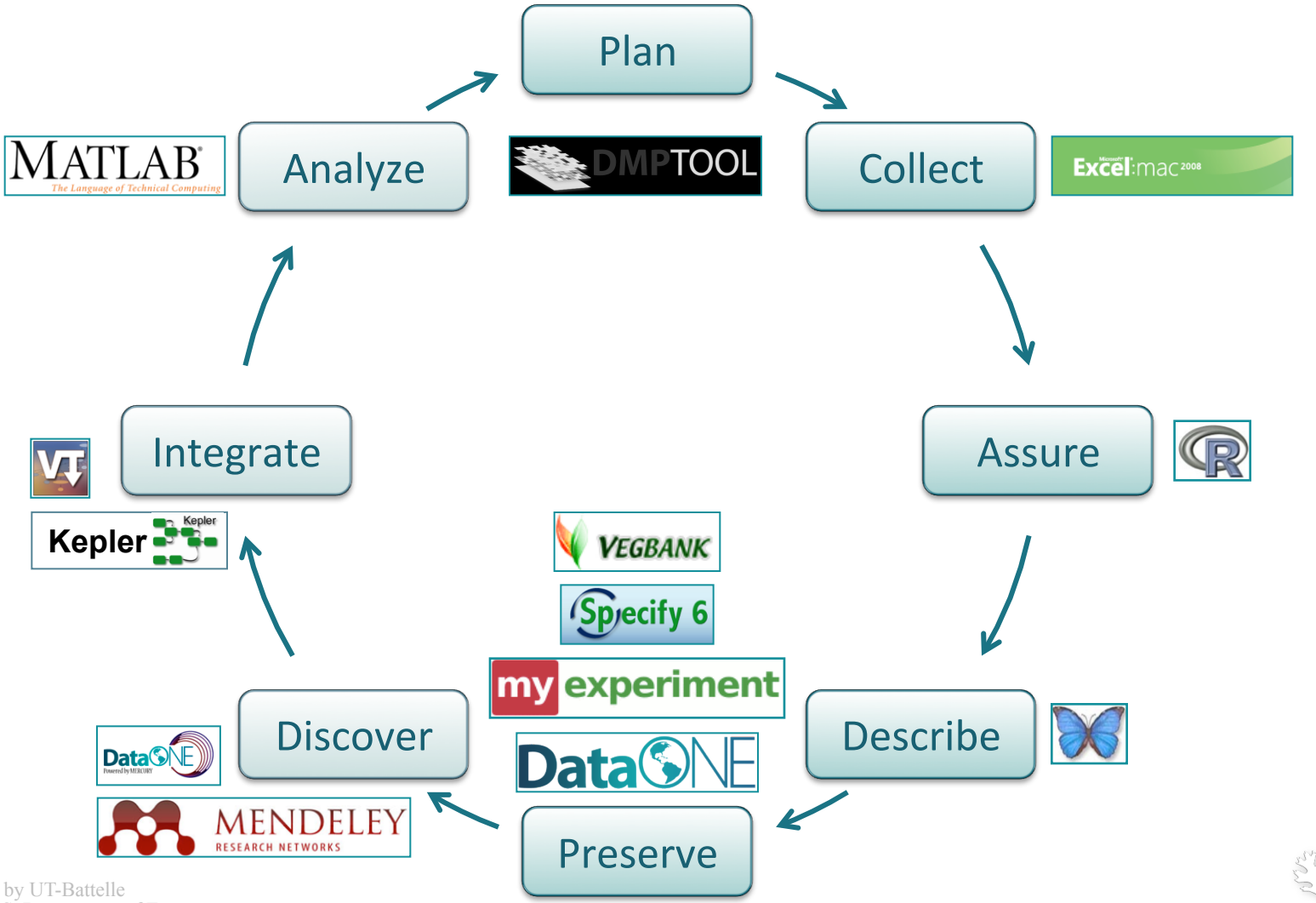
CILogon

InCommon

ProtectNetwork
Cloud based E-Credential Management service

OpenID

Support for Entire Data Lifecycle



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 publically 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
YOUR VOICE IN OUR GOVERNMENT

CREATE A PETITION | OPEN PETITIONS | RESPONSES | HOW & WHY

OFFICIAL OFFICE OF SCIENCE AND TECHNOLOGY POLICY RESPONSE TO
Require free access over the Internet to scientific journal articles arising from taxpayer-funded research.

Increasing Public Access to the Results of Scientific Research

By Dr. John Holdren

Thank you for your participation in the We the People platform. The Obama Administration agrees that citizens deserve easy access to the results of research their tax dollars have paid for. As you may

OAK RIDGE National Laboratory

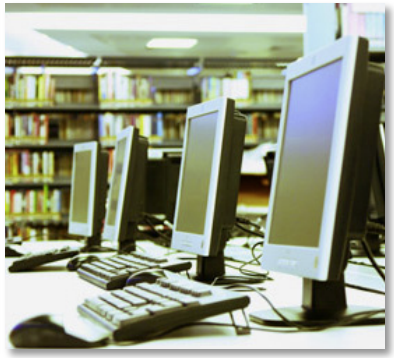
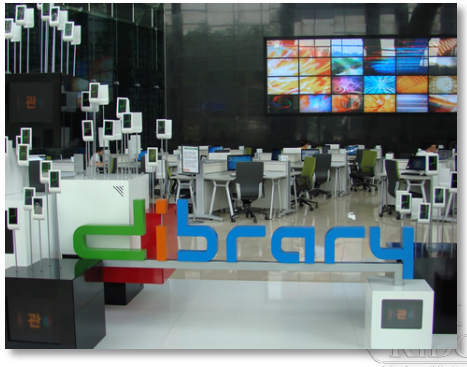
Building global communities of practice: … creating long-lived CI enterprises,

- 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

Libraries and museums: value

- As Member Nodes:
  - Facilitate the teaching and research mission of institution
  - Build data collections for the 21st century
- In support of Data Librarians:
  - Provide access to data management plans
  - Provide best practices for faculty and students
  - Cyberinfrastructure supporting the data lifecycle



Data Management Planning Tool

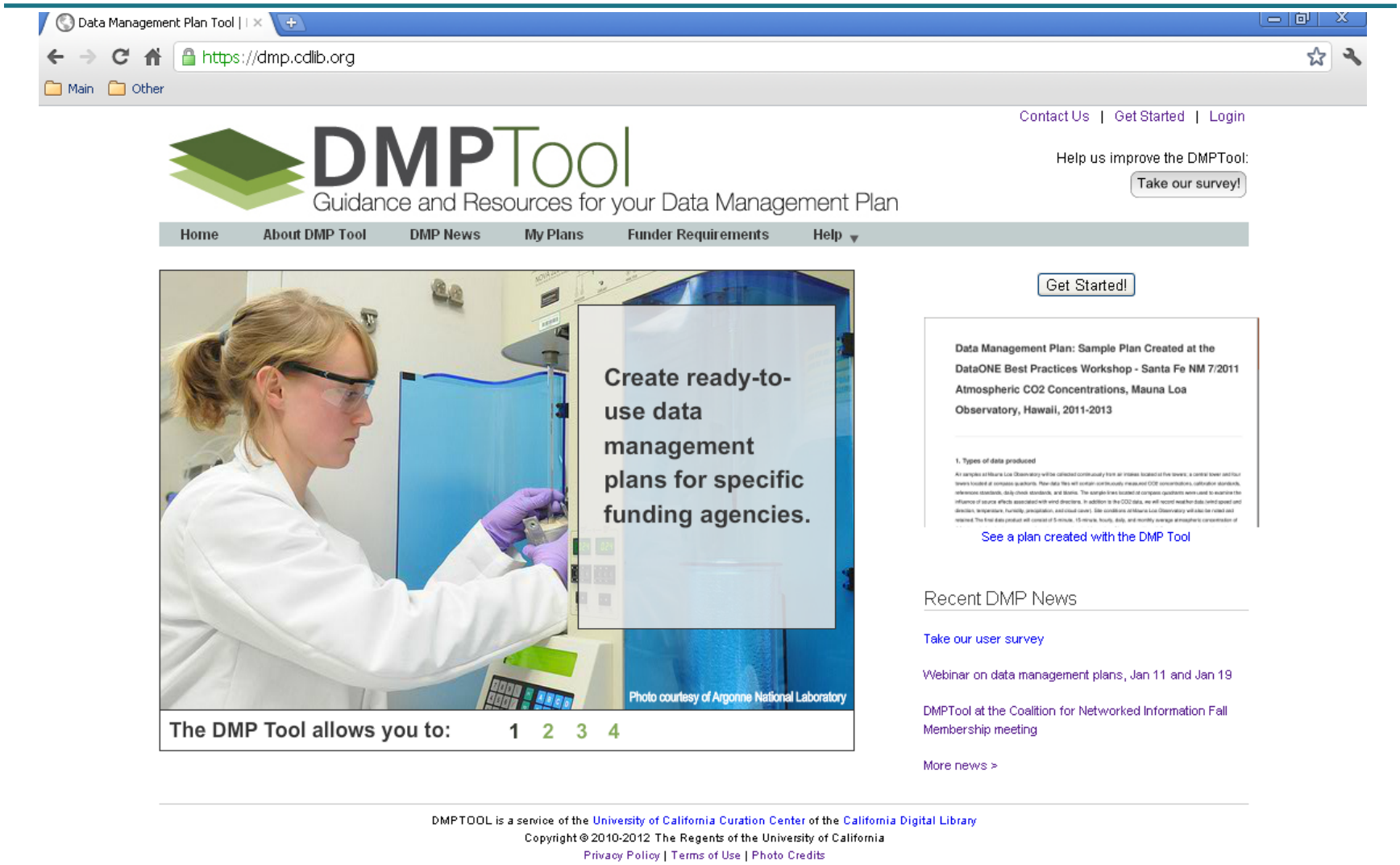
DMPTool

Guidance and Resources for your Data Management Plan

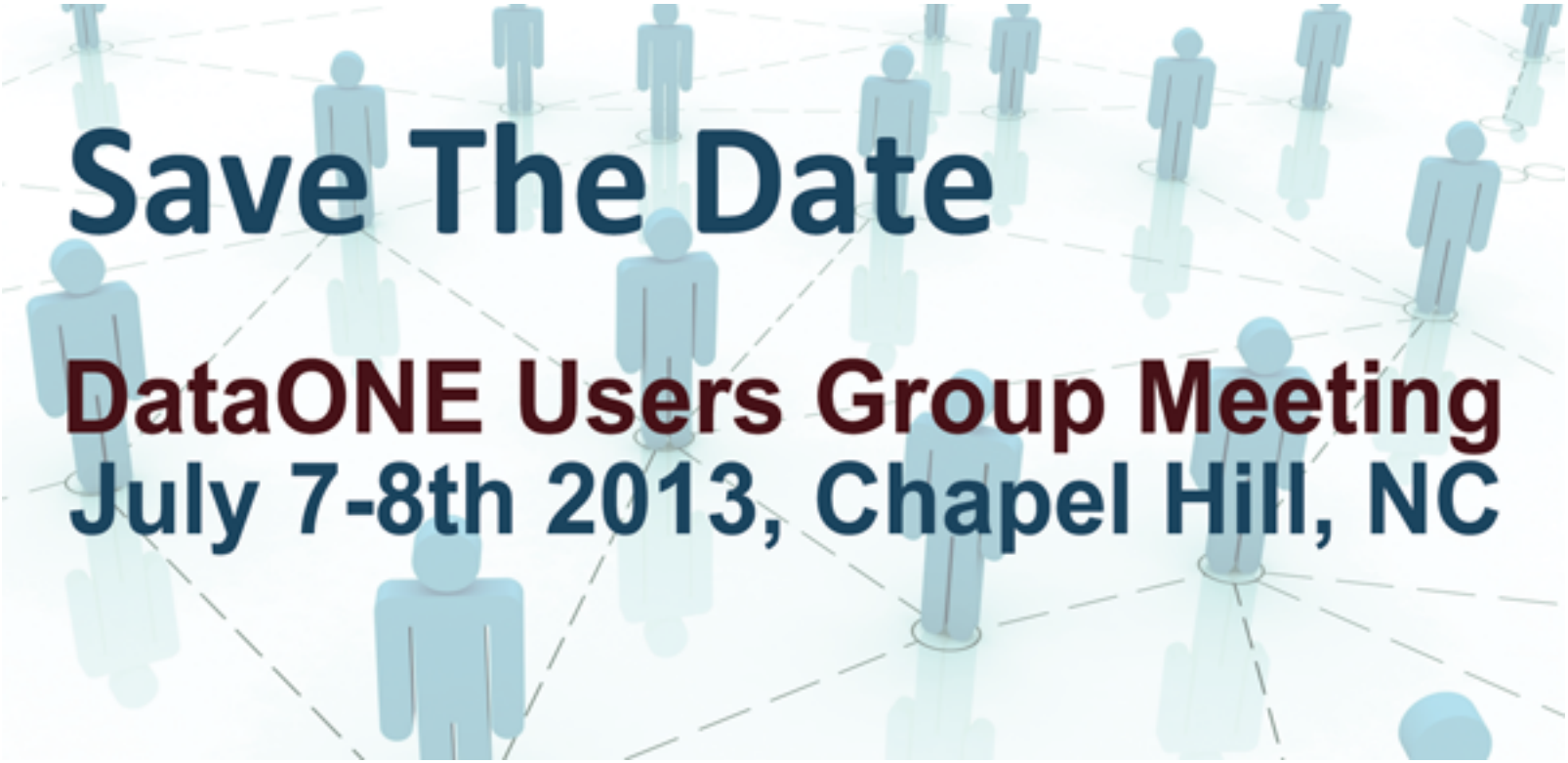
https://dmp.cdlib.org/

- Create ready-to-use data management plans for specific funding agencies
- Meet funder requirements for data management plans
- Get step-by-step instructions and guidance for your data management plan as you build it
- Learn about resources and services available at your institution to help fulfill the data management requirements of your grant
- Released: Oct. 2011
- Support for NIH requirements added 2/22/2012
- Other similar efforts now also underway at institutional levels or with other entities.

Plug: DMPTool next rev upcoming



DataONE DUG July 7-8 Chapel Hill NC



Co-located with ESIP Federation Meeting.

Question & Discussion

John W. Cobb

865.576.5439

cobbjw@ornl.gov

http://www.slideshare.net/johnwcobb/cobb-u-massnealesciencev06