Data Curation: Why Does It Matter?

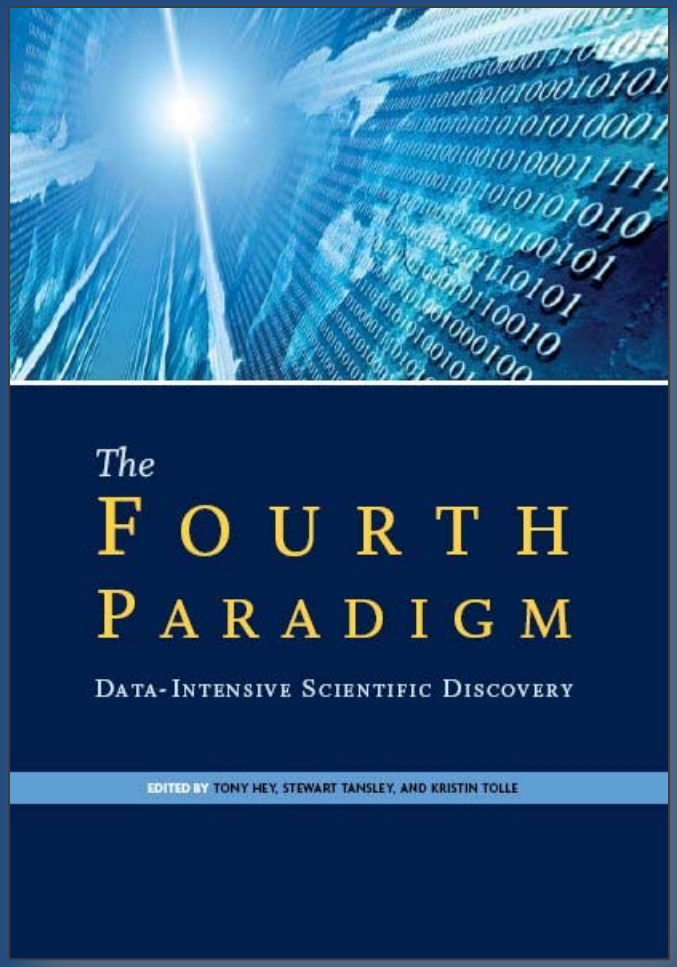
Karon Kelly
University Corporation for Atmospheric Research

Colorado State University
Research Data Management Workshop
March 24, 2011

Overview

- eScience & data landscape
- Importance of data curation
- NSF DataNet program and goals
- Data Conservancy goals & activities

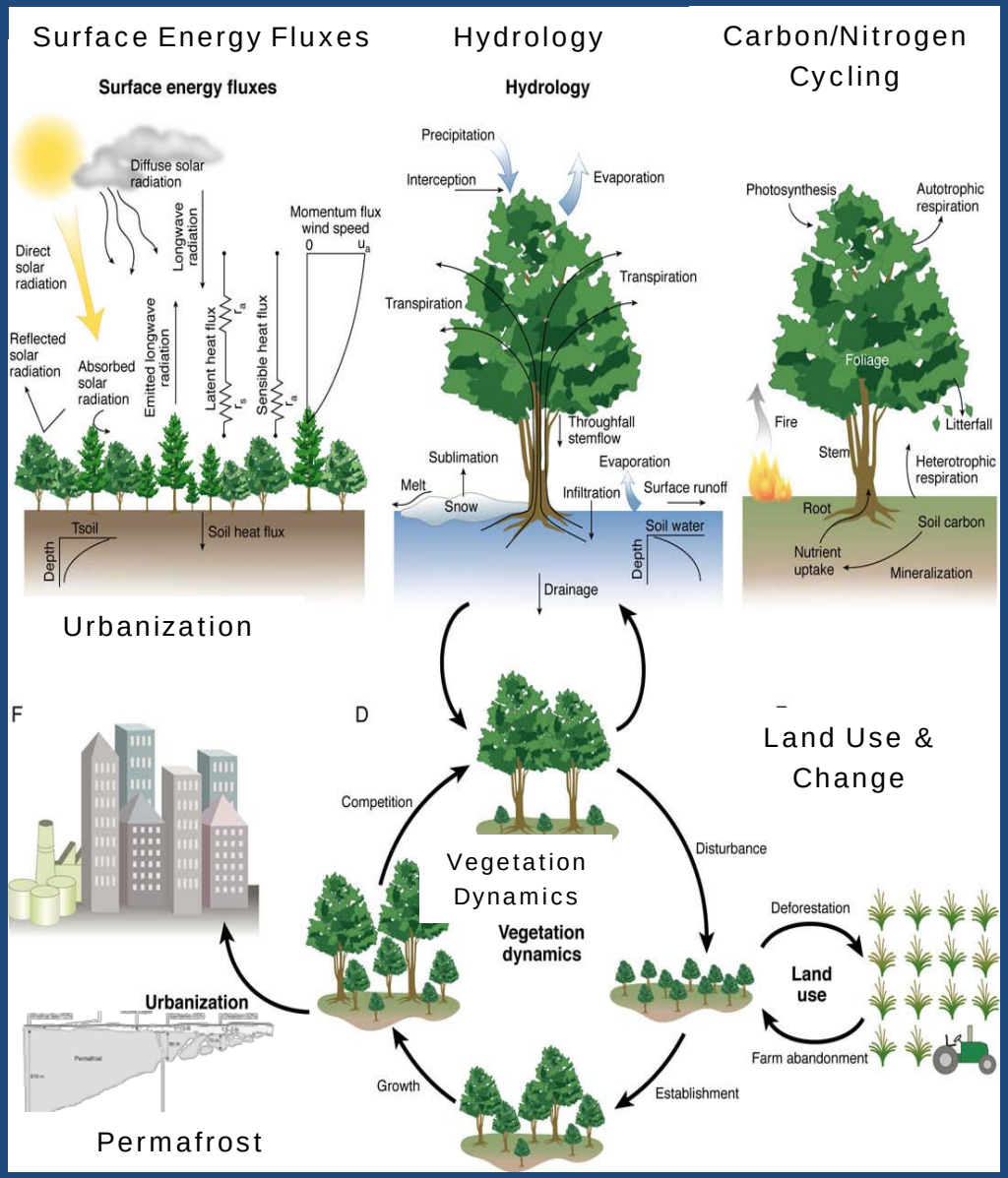
E-Science and Data



- Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets
- The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud computing technologies

Data & Complexity

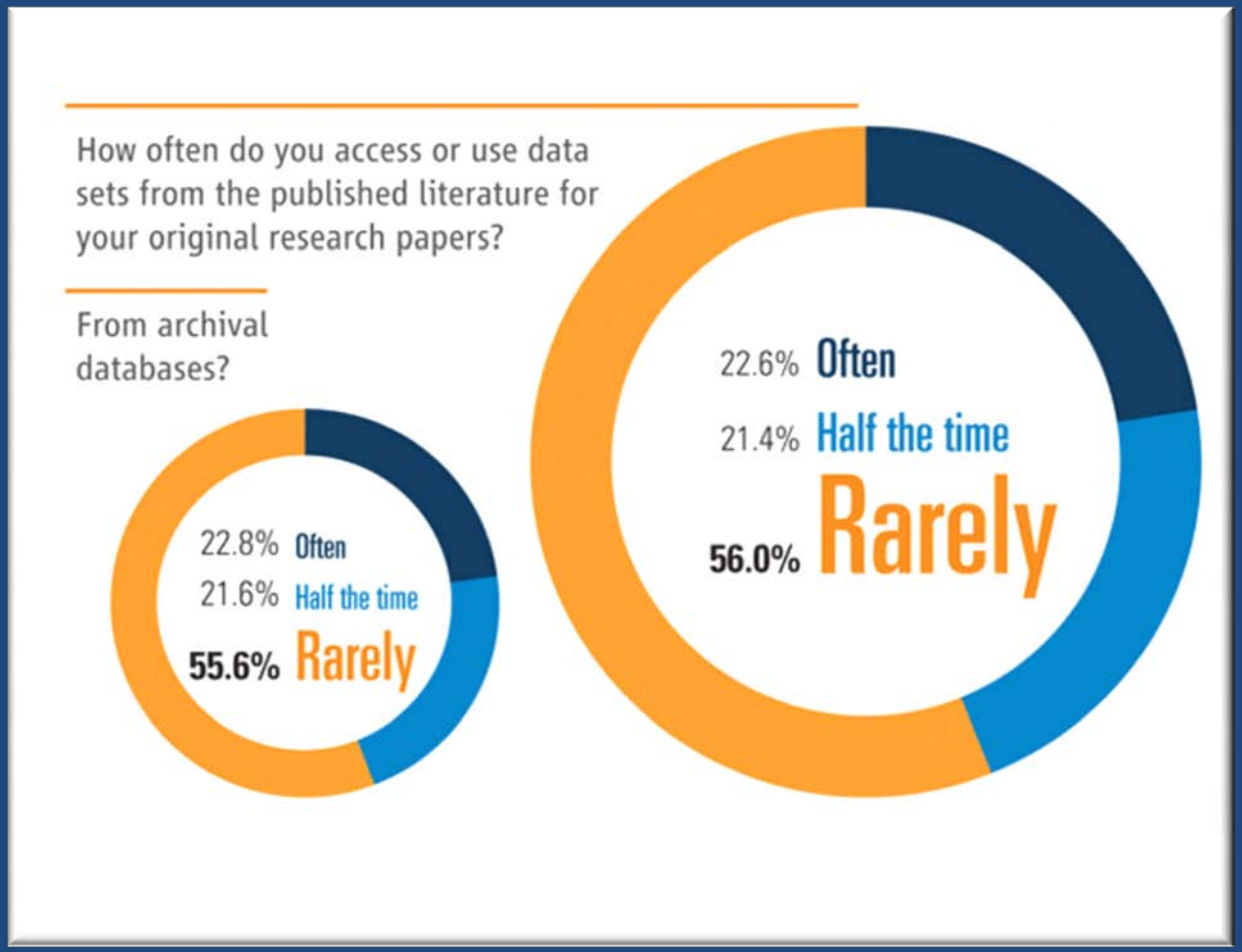
- Research problems increasingly interdisciplinary and complex
- Collaboration requires open sharing of data
- Data are highly heterogeneous and largely incompatible in their native forms
  - The semantics and contexts within which data are gathered and interpreted are important to preserve



Challenge or Opportunity?

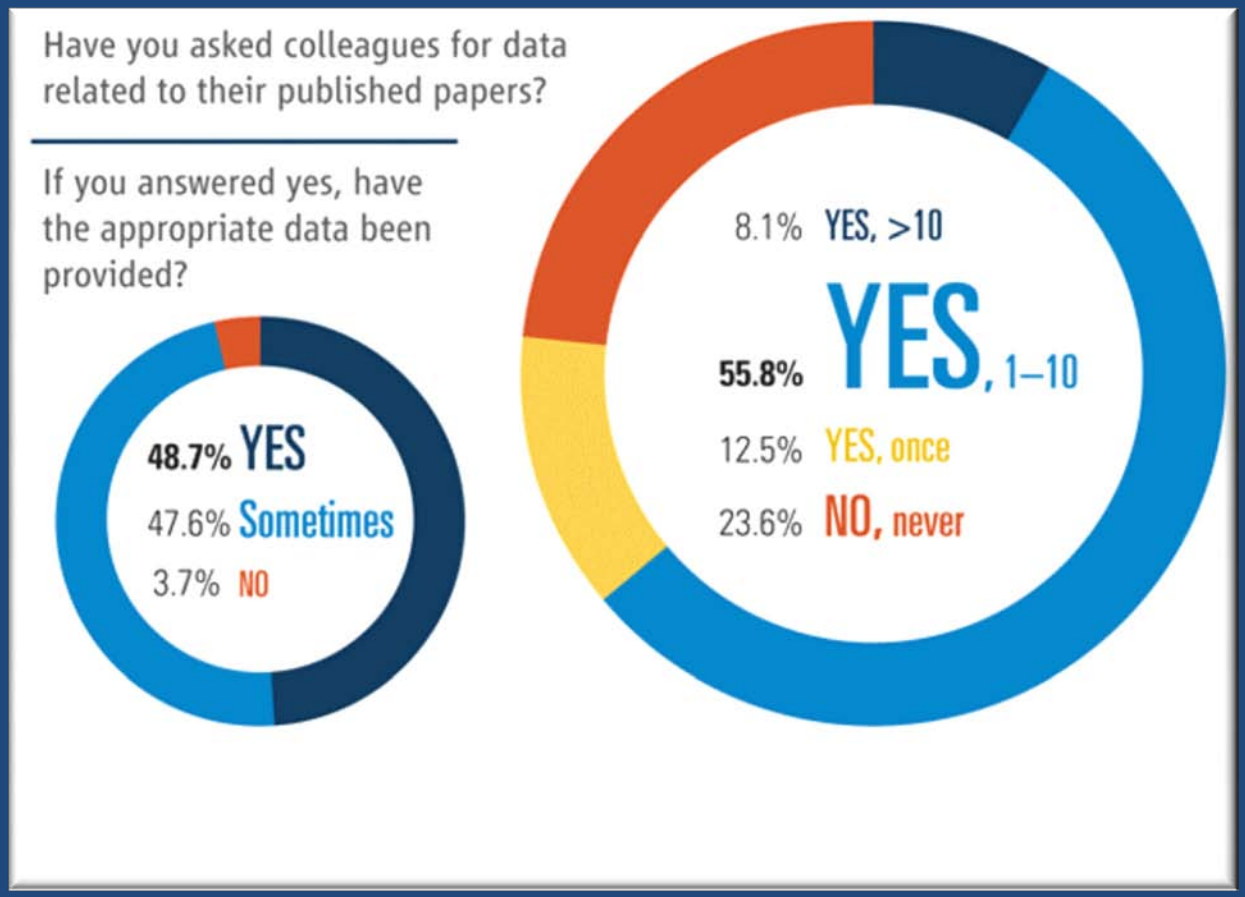
- Scientific innovation will spur economic recovery
- Science and technology are essential to improving public health and welfare and to inform sustainability
- The scientific community has been criticized for not being sufficiently accountable and transparent
- Data Collection, curation, and access are central to all of these issues

Data Access



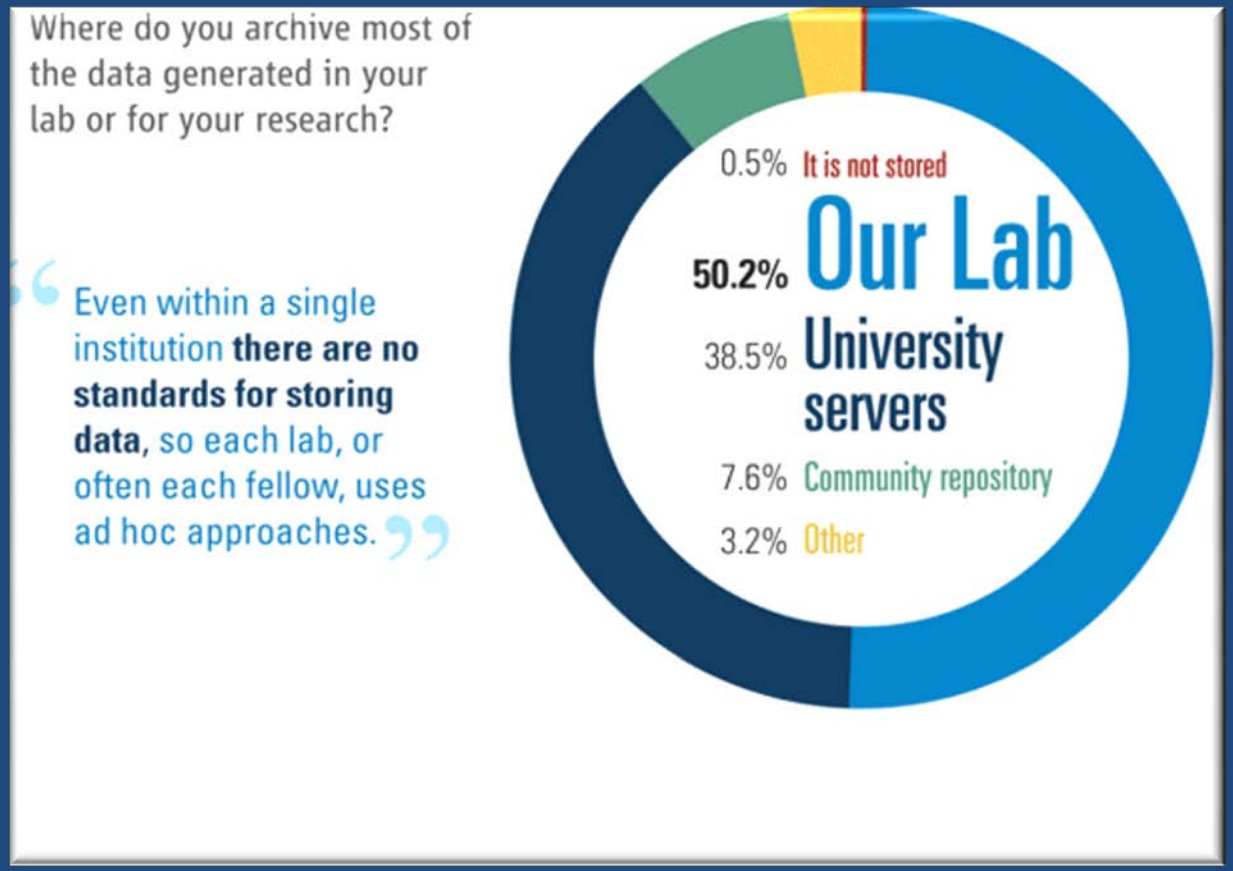
M. Twombly/*Science*; Source: *Science* Online Survey

Data Sharing



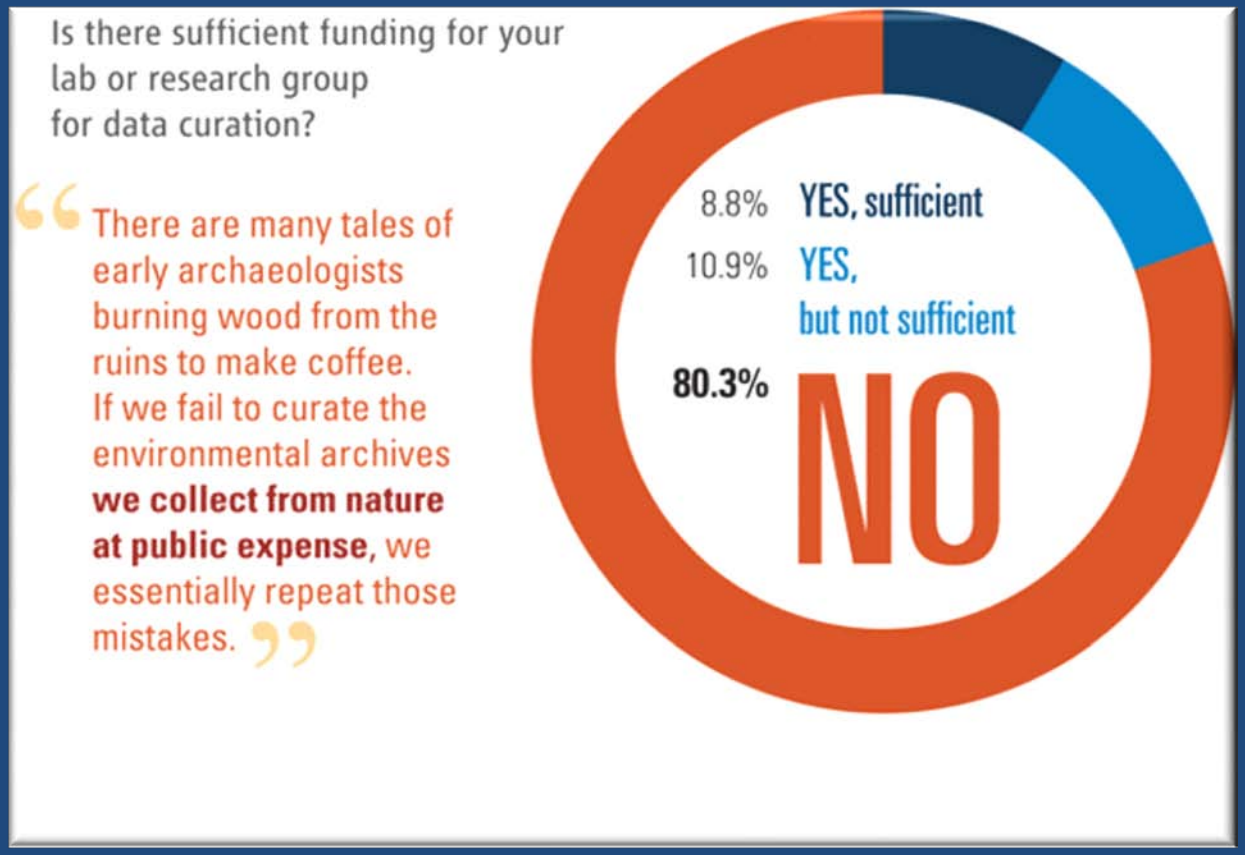
M. Twombly/*Science*; Source: *Science* Online Survey

Data Archiving



M. Twombly/*Science*; Source: *Science* Online Survey

Support for Curation



M. Twombly/*Science*; Source: *Science* Online Survey

Data Curation Is…

The active and on-going management of (research) data through its lifecycle of interest and usefulness to scholarship, science, and education, including:

- Collection development
- Quality assurance
- Enhanced value and enabled discovery and retrieval
- Archiving
- Preservation for re-use over time

Why Publish/Archive/Curate Data?

- Call for accountability and transparency
- Funding agency requirements
- Publisher requirements
  - AGU: “Data sets that are available only from the author, through miscellaneous public network services, or academic, government or commercial institutions not chartered specifically for archiving data, may not be cited in AGU publications”
- Scholarly communication chain—connecting data to publication
- Digital data is inherently fragile and often at risk of loss
- Future access to valuable digital assets depends upon curation/preservation actions taken today

Sharing vs Publication

Many of the issues regarding data availability can be addressed if the principles of “publication” rather than “sharing” are applied. However, online data publication systems also need to develop mechanisms for data citation and indices of data access comparable to those for citation systems in print journals.

Mark J. Costello. Motivating Online Publication of Data. *BioScience,* Vol. 59, No. 5 (May 2009), pp. 418-427.

Data Citation

- Provides credit to creator and stewards, tied to promotion and tenure
- Implies peer review
- Enables tracking impact of data publication and use
- Provides for reproducibility through direct, unambiguous connection to data

What's Required?

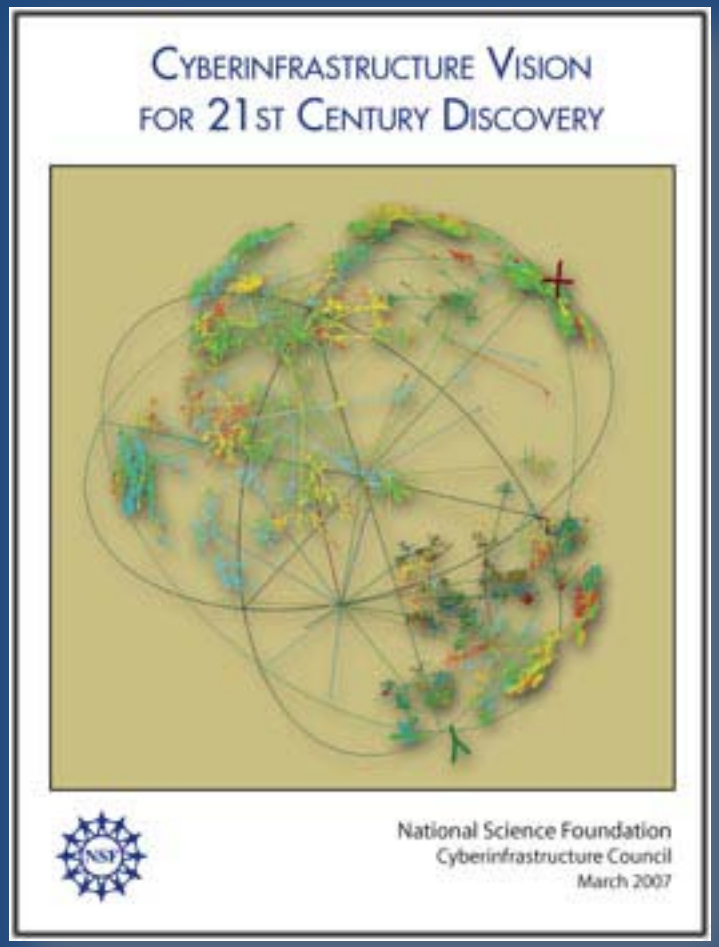
- Understanding of existing and emerging data curation and preservation practices within disciplines
- More accurate understandings of situated data curation and stewardship practices in scientific collaboration
- Incentivizing data sharing/publishing

What's Required (2)

- Make it easy—maximize cyberinfrastructure support to reduce archive/publication overhead
- Make it citable: motivating publication through peer recognition
- Make it useful – moving beyond the data archival to access and reuse

Vincent Smith. Data Publication: Towards a Database of Everything, *BMC Research Notes* 2009, 2:113

NSF DataNet Program



Vision:

"…science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved."

NSF DataNet Program Goals

- Provide systematic, long-term preservation, access and analysis capabilities in an environment of rapid technology advances
- Engage at the frontiers of science and engineering research and education
- Serve as part of an interoperable data network spanning national and international boundaries

NSF Office of Cyberinfrastructure

DataNet Partner Requirements

- Combine expertise in library and archival sciences; computer, computational and information sciences, cyberinfrastructure; domain sciences and engineering
- Develop models for economic and technological sustainability over multiple decades
- Work cooperatively to create a functional data network with revolutionary new capabilities for access, use and integration

The Data Conservancy (DC)

- DC is one of first two awards through the DataNet program
- Led by Sheridan Libraries at Johns Hopkins University
- DataONE: Observation Network for Earth, led by University of New Mexico Libraries
- Next round of DataNet will add up to three more partners into the network

Data Conservancy Partnership

DC is a network of domain scientists, information and computer science researchers, enterprise experts, librarians, and engineers

Sayeed Choudhury, PI—Sheridan Libraries, Johns Hopkins University

Other DC Partners

Australian National Data Service
Australian National University
British Library
Digital Curation Centre
Microsoft Research
Monash University
Nature Publishing Group
Optical Society of America
Sakai Foundation
Space Telescope Science Institute
SPARC
Sun Microsystems (Data Curation Center of Excellence)
University of Queensland
Zoom Intelligence

DC Vision and Goal

- Support new forms of inquiry and learning through the creation, implementation, and sustained management of integrated and comprehensive data curation strategy

- DC embraces a shared vision – data curation is not an end, but rather a means to collect, organize, validate, and preserve data to address grand research challenges that face society

DC Objectives

- Infrastructure research and development
  - Technical requirements
- Information science and computer science research
  - Researcher/user requirements
- Broader impacts
  - Educational requirements for capacity building
- Sustainability
  - Business requirements

Understanding Scientific and User Needs: Data Practices in Disciplines

Multi-site user research methods are a blend of:

- Case study and domain comparisons
- Depth and breadth
- Local and global

	Astronomy	Life Sciences	Earth Sciences	Social Sciences	
UCAR		Action research in social sciences; Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations			**UCAR**
UCLA	Ethnography, virtual ethnography, oral histories ⇒ Use cases, data requirements	Interviews, Surveys, Worksheets, Content analysis ⇒ Curation requirements, taxonomy, metadata/provenance framework			**UIUC**

Research Questions

- **Data practices:** What are the data management, curation, and sharing practices?
- **Networks:** Who uses what data when, with whom, and why?
- **Curation:** What data are most important to curate, how, and for whom?

DC Progress to Date

- Connecting data and publication through arXiv.org
- Access to DC data through Sakai collaboration and learning environment
- Integration of DC data with an existing science research framework through the International Virtual Observatory Alliance
- Interoperability between the DC and National Snow and Ice Data Center glacier photo service

Broader Impacts and Capacity Building

- Ensuring the wider community is involved with and will benefit from the infrastructure being developed

- Data curation outreach and education
  - Professional degree programs, in-service professional development, certification and institutes at Library/Information schools
  - Mentoring and “boot camps”
  - Field work practica and internships
  - Extending programs to educate more diverse set of students
  - Fellowships for students from traditionally underserved populations

- Communications on DC outcomes to university, scientific, and citizen stakeholders

Things to Watch

- Task Force on Data Policies – March 28-29, 2011 – National Science Board
  - Vision of Data-Intensive Science
  - Reproducibility, First Steps and Guiding Principles
  - Impacts
- Measuring the Impacts of Federal Investments in Research – April 18-19, 2011 – Board on Science, Technology, and Economic Policy (STEP) and Committee on Science, Engineering, and Public Policy (COSEPUP)

Acknowledgements

Office of Cyberinfrastructure DataNet Award #0830976

Data Conservancy Partnership

Sayeed Choudhury, Johns Hopkins University

Christine Borgman, UCLA

Carole Palmer and Melissa Cragin, Illinois

Thank You

Contact kkelly@ucar.edu for more information