DataONE: Enabling Data-Intensive Biological and Environmental Research through Cyberinfrastructure

Leadership Team: William Michener[1], Suzie Allard[2], John Cobb[3], Robert Cook[3], Patricia Cruse[4], Mike Frame[5], Stephanie Hampton[6], Vivian Hutchison[5], Matthew Jones[6], Steve Kelling[7], Rebecca Koskela[1], Carol Tenopir[2], Dave Vieglais[8], Todd Vision[9], Bruce Wilson[2]

Co-Investigators: Paul Allen[7], Peter Buneman[10], Randy Butler[11], Ewa Deelman[12], David DeRoure[13], Cliff Duke[14], Carole Goble[15], Donald Hobern[16], Peter Honeyman[17], Jeffery Horsburgh[18], John Kunze[4], Bertram Ludaescher[19], Maribeth Manoff[2], Line Pouchard[3], Robert Sandusky[20], Ryan Scherle[9], Mark Servilla[1], Jake Weltzin[5]

[1]University of New Mexico; [2]University of Tennessee; [3]Oak Ridge National Laboratory; [4]University of California - California Digital Library; [5]U.S. Geological Survey; [6]National Center for Ecological Analysis and Synthesis - University of California - Santa Barbara; [7]Cornell University; [8]University of Kansas; [9]National Evolutionary Synthesis Center, University of North Carolina; [10]University of Edinburgh; [11]University of Illinois - Urbana Champaign; [12]University of Southern California; [13]University of Southampton; [14]Ecological Society of America; [15]University of Manchester; [16]Atlas of Living Australia; [17]University of Michigan; [18]Utah State University; [19]University of California – Davis; [20]University of Illinois – Chicago

Abstract:

Addressing the Earth's environmental problems requires that we change the ways that we do science; harness the enormity of existing data; develop new methods to combine, analyze, and visualize diverse data resources; create new, long-lasting cyberinfrastructure; and re-envision many of our longstanding institutions. DataONE (Observation Network for Earth) represents a new virtual organization whose goal is to enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it.

DataONE is designed to be the foundation for new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

Supported by the U.S. National Science Foundation, DataONE will ensure the preservation and access to multi-scale, multi-discipline, and multi-national science data. DataONE is transdisciplinary, making biological data available from the genome to the ecosystem; making environmental data available from atmospheric, ecological, hydrological, and oceanographic sources; providing secure and long-term preservation and access; and engaging scientists, land-managers, policy makers, students, educators, and the public through logical access and intuitive visualizations. Most importantly, DataONE will serve a broader range of science domains both directly and through the interoperability with the DataONE distributed network. DataONE is a five year project that began in Fall 2009 (William Michener, PI, *University of New Mexico*).

The Vision: "*DataONE will be commonly used by researchers, educators, and the public to better understand and conserve life on earth and the environment that sustains it.*"

By creating an infrastructure of technology and standards, people, and institutions to support the full life cycle of biological, ecological, and environmental data and tools that enable universal access, DataONE will accelerate use of earth observational data in research, education and decision-making. In so doing, DataONE will transform our understanding of ecological processes and conserve life on earth and the environment that sustains it.

Why do we need DataONE?:

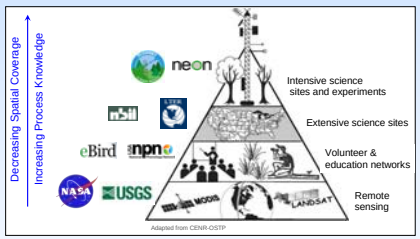
Societal and Environmental challenges

Popular press and results from the International Geosphere Biosphere Program show that environmental challenges are of increasing concern for us all.

Science Challenges

To understand, we need easy access to different types of data

- Wide range of spatial and temporal scales (plot data to remote sensing data)
- Breadth of science domains (biological, environmental, social, and economic)
- Citizen science networks, of increasing importance



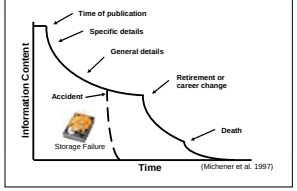
Data Challenges

Scientists need access to the data generated by research to verify findings and test new hypotheses.

Poor data practices place the scientific record at risk

- Data are massively dispersed
  - Not on the Web
  - Orphaned
- Multiple Semantics
  - Not easily discovered
- Poor data practices
  - Documentation
  - Formats / obsolescence
  - Lack of Standards
- Poor stewardship
  - Media Obsolescence
- Heterogeneous, incompatible formats
  - Difficult to combine data from diverse sources

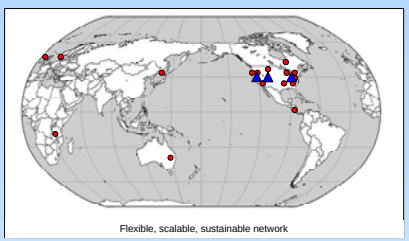
Over time the information content of data products can be lost.



What is DataONE?

Cyberinfrastructure enterprise (tools, ideas, people)

Distributed Framework of Coordinating Nodes and Member Nodes



Member nodes
- Diversity tolerant (less tightly coordinated)
- Freedom to try or build new tools, methods, and leapfrog forward

Coordinating nodes
- Tightly coordinated, stable service platform
- CI based on use cases and prioritized need for functionality
- Support and adapt existing community software efforts

Coordinating Nodes
- retain complete metadata catalog
- perform basic indexing
- provide network-wide services
- ensure data availability (preservation)
- provide replication services

1. University of New Mexico
2. University of California-Santa Barbara
3. Oak Ridge Campus (UT & ORNL)

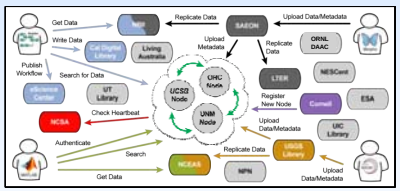
Member Nodes
- diverse institutions
- serve local community
- provide resources for managing their data

Member Nodes Deployed in 2010:
ORNL DAAC, Knowledge Network for Biocomplexity, Dryad, USGS NBII, and California Digital Library

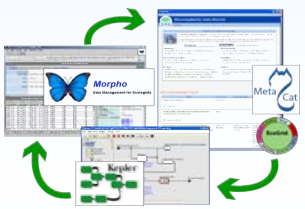
Provide tools to better manage data

Provide support for the entire data life cycle—preparation of data sets, stewardship, tools to access and use the data

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



Metadata creation, management
Search catalog
Work Flows



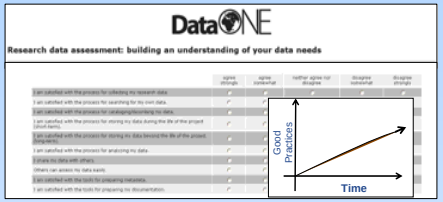
Metadata catalog with >70,000 records

How will we build DataONE?

Understand the community's need, have the community envision solutions

Perform baseline and iterative Community Assessments and Usability Studies

To see where data practices and policies are now, so we can see how practices change over the life of DataONE:



DataONE Assessment: http://vovici.com/wsb.dll/s/aaeg3cfe6

Leverage existing CI whenever possible, build new CI whenever necessary

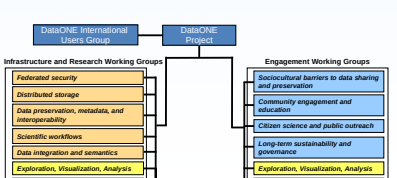
- Many existing open source efforts exist
  - Metadata Editors: *Mercury, Morpho*
  - Data management: *MATT, UDig, Specify*
  - Analysis and modeling: *R, Octave, netCDF*
  - Workflow systems: *Kepler, Taverna, VisTrails*
  - Grid systems: *Condor, Globus, BOINC*
  - Data and workflow portals: *VegBank, myExperiment*
- Commercial tools important too
  - *MATLAB, SAS, ArcGIS, R,*
- DataONE: help communities build their own tools
  - Integrate, interoperate, stabilize
  - Create libraries to DataONE Service Interface

Engage the community, reach out, educate, enable new science, and demonstrate success

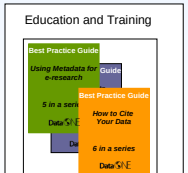
DataONE will use Working Groups

Structure
- 10 – 20 participants
- Deep analysis
- Intensive collaboration
- Neutral territory
- 2 week-long meetings per year

Success
- Community-driven
- Inclusive
- High productivity
- High impact



Education and Training



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