Virtual Update Notes May 4 & 5, 2009 - Data Centers for Ecology

Published on *LTER Information Management* (http://im.lternet.edu)

Home > Virtual Update Notes May 4 & 5, 2009 - Data Centers for Ecology

Virtual Update Notes May 4 & 5, 2009 - Data Centers for Ecology

Thu, 04/16/2009 - 1:47pm — sremillard

Participants:
Monday (5/4): Bill Michener (moderator, LNO), Barbara Benson (NTL), Dave Balsiger (NTL), Don Henshaw (AND, Hap Garritt (PIE), Jason Downing (BNZ), Suzanne Remillard (AND), Kristin Vanderbilt (SEV), Karen Baker (PAL, CCE), James Conner (CCE), ? Peterson, John Porter (VCR)

Tuesday(5/5): Bill Michener (moderator, LNO), Christine Laney (JRN), Mark Servilla (LNO), Duane Costa (LNO), Nicole Kaplan (SGS), Margaret O'Brien (SBC), Corinna Gries (CAP), Wade Sheldon (GCE), Gastil-Buhl (MCR), Hope Humphries (NWT), John Campbell (HBR), Emery Boose (HFR), Jonathan Walsh (BES)

See attached powerpoint presentation (below).

Monday Notes (by Suzanne Remillard):
Environmental science challenges
Cyberinfrastructure challenges
DataONE: A solution

Using the knowledge pyramid to get at science and human issues. Cyberinfrastructure challenges are many, but some pertinent ones include format obsolescence, and loss of institutional commitment. The available storage capacity is not keeping up with the amount of information being collected. This is manly due to large arrays of sensor networks. Much of this data stored has never been looked at, therefore we're storing a massive amount of garbage. Data entropy is the loss of data value over time. There are ways to mediate this; i.e. good metadata. Data longevity has been tremendously increased with digital libraries which have increased the resource half-life.

DataONE
There are many existing biological data archives; Ecological Archives, DAAC, NBII, Frames, LTER, KNB.
There are also many data holdings although metadata interoperability across data holdings differs.
DataONE will use existing CI (morpho, metacat, ecogrid, kepler).
Building new global CI with DataONE. Green stars are member nodes (usually associated with an institution) and red nodes are coordinating nodes (see slide). Will include both biological and environmental data types as well as data sources. DataONE will include a distributed framework that is flexible, scalable, and sustainable.
Coodinating nodes; ORC (oakridge lab), UCSB (NCEAS and UCSB), UNM. The number of coordinating nodes and member nodes is likely to increase across the globe.
How does it work? It's a one-stop shop for data. Searching is similar to a Google search. Includes both simple and advanced searching mechanisms.
Challenges includes immense heterogeneity.
A use case scenario could be forecasting the spread of an invasive species across North America.
My Experiment - a workflow sharing portal.
Changing science culture and institutions through education and training, engaging citizens in science, and building global communities in practice.
Will need to re-envision academic CI for the 21st century and engage diverse partners
What shapes DataOne? Challenges associated with climate variability; community needs good data.

DataONE does not have funding yet, but await to hear about funding very soon. It has been approved by NSF, but the money hasn't been awarded.

Questions:
How could LTER help the success of DataONE?
Several education and outreach activities, i.e., technical training videos. Having LTER data associated with a global archive can increase the visibility.

Who at the LNO is involved in DataONE?
Bill Michener, Mark Servilla, Duane Costa (partly in associated project).
5 PIs (Bill Michener, Kathleen Smith, Stephanie Hampton, Mike Frame, Bob

Cook) and 30 co-investigators.

Funding timeframe?
5 year project with 5 year renewal from NSF. Datanet programs should last decades to centuries (as dictated by NSF and the white house). First member nodes within 3 years (early adopters) then subsequent member nodes in an orderly fashion to follow. There will be a significant amount of prototyping before breaking into full production.

Should be a sustainable business model for the future.

Tuesday Notes:
by Emery Boose

See power point on IM website for more details.

DataONE = data observation network for the earth. One of two data solutions supported by NSF.

Environmental challenges are receiving public attention. E.g. Inconvenient Truth, hockey diagram of increase in population, temperature, etc. Since 1950, all parameters have exceeded historical limits.

Global warming is expected to be significant and heterogeneous. See map showing concern areas. Many unknowns: e.g. how will Gulf Stream be affected?

Knowledge pyramid is essential to address these issues.

Data loss. Format obsolescence and loss of institutional commitment are critical issues now.

Increasing trend of information exceeding available storage. Good documentation, storage, etc help with data entropy.

Data longevity. Resource half-life is only 1-4 years for many types of data. Digital Library Object (supported by NSF) is an exception with half-life of 25 years.

Existing biological archives. Currently 40-50, including Ecological Archives, DAAC, NBII, FRAMES (fire research and management exchange system), LTER, KNB (Seek, etc). Large number of metadata standards.

Tools include Morpho, Metacat, EcoGrid, Kepler. Seek project just ended.

Metacat standard adoption curve shows rapid increase. Currently about 25,000 data packages.

Kepler supports other open source tools. Data Turbine for real-time data streams. Kepler will likely be adopted by NEON. Supports R as well as MatLab, etc. Kepler can save workflows for reuse. Kepler is domain agnostic and is used by other disciplines. Adoption community is becoming large.

DataONE. See global participant map, list of disciplines, contributors. Member nodes include institutions, networks, libraries. Coordinating nodes on all continents.

Pilot catalog (not yet funded) includes 48,000+ records. Advanced search capabilities.

Ontologies and semantic mediation will enhance data interoperability.

See use case poster (Eurasian collared dove). Collection and integration of data from many sources.

My Experiment project. Workflow sharing portal. Working with Kepler developers.

DataONE includes career-long experiential effort. Best practices guides, exemplary data management plans, podcasts and webcasts, workshops and seminars, downloadable curricula.

Engaging citizen scientists. E.g. Project Budburst (K-12). Citizen science toolkit.

Building global communities of practice. Broad, active community engagement. Includes many library systems. New, transparent governance structures.

Website will go live when NSF gives OK. Build sustainable business model with partners such as IBM.

Re-envision CI at universities. Not just high-performance computers. How does it all fit together? R and D in areas such as ontologies.

Wide diversity of partners: universities, terragrid, libraries, overseas universities, oak ridge national lab, Amazon, IBM, etc.

DataONE shaped by need for good data. Good data needs good infrastructure, organization, and engagement.

Example of new planet discovered in 11-year-old dataset using new algorithm.

Funding from NSF for DataONE expected shortly. Invite community involvement in near future. Create DataONE international users group (comparable to SAS users groups).

Other interoperability projects funded by NSF include Sonet (Mark Schildauer) for semantics and ontologies.

IBM is developing open source tools. Intel is a potential funder and may house a coordinating node.

NSF special competition through office of CI to support 5 Datanet programs. Each would focus on a major discipline at NSF. DataONE for bio-geo. Other partners will represent other areas. Two funded last year. Three, maybe four this year. 5 years, $20M. Successful programs refunded after 5 years.

Only 5 PIs: Bill Michener, Kathleen Smith, Stephanie Hampton, Bob Cook, Mike Frame. Co-PIs (about 30) scattered across globe, include librarians, computer scientists, etc.

How would DataONE interact with participating networks? Centralized metadata holdings and directory, replicated in multiple locations (coordinating nodes). Nodes will have partnership agreements. Member nodes would replicate data holdings of one or more other nodes. LOCKS principle: lots of copies keeps (data) safe.

Many PIs are members of these organizations. Many LTER sites (or host institutions) may wish to become member nodes. A large field station may wish to become a member node.

Rational adoption policy. Early adopters, about 3 member nodes to test. Goal is world-wide network of nodes.

Expectations for quality of data and metadata? Not clear yet what policy will be. An important question.

Lots of good data resources. LTER data are mostly in good shape. Other data are less reliable, less well documented. Focus on good quality datasets but also enable individual scientists to participate. Develop mechanisms to facilitate participation.

Generational change required before end game is reached.

Easier to get buy-in when it's clear what people will gain from participation.

What are needs of policy makers? How to engage stake holders?

Baseline assessment tool to get input from scientists, students, decision makers. Involve usability expert (as in Seek project).

Historically data management has been significantly underfunded. But NSF and other funding agencies are beginning to recognize the importance of data (not just publications). We can help by providing feedback on the real costs of managing data. Approaching a turning point in terms of support and funding.

Many threads coming together here.

Attachment	Size
DataONE_VWC_LTER_4May2009.ppt [1]	17.09 MB

- Virtual Updates [2]

This material is based upon work supported by the National Science Foundation under Cooperative Agreement #DEB-0236154. Any opinions, findings, conclusions, or recommendations expressed in the material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Please contact us with questions, comments, or for technical assistance regarding this web site.

Source URL: http://im.lternet.edu/node/439

Links:
[1] http://im.lternet.edu/sites/im.lternet.edu/files/DataONE_VWC_LTER_4May2009.ppt
[2] http://im.lternet.edu/taxonomy/term/169