Imagine a world where scientists and all their data and intellectual contributions are seamlessly interconnected and easily discoverable by all. In such a place, a search for a specific “Jane Smith” would lead to metrics that highlight her impact factor as a scientist—pointing to her publications (and the linked data upon which they were based), blog posts, tweets (and number of re-tweets), patents, presentations, as well as her students and the colleagues that are part of her socio-science network. In this world, Jane would have more time to do actual research as she would be able to readily access the data and information that she needs; for example, to acquire data needed as part of a synthesis project or to acquire an existing science workflow that she needs; for example, to acquire data or existing workflows, and sharing the new information creation and collection efforts.

The picture painted above may sound somewhat farfetched now, but several steps are currently being taken that will make this new world a reality—most likely within the next five years. Some of the most recent actions include:

- The US Office of Management and Budget in its May 9, 2013 Memorandum “requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine-readable and open formats, data standards, and common core and extensible metadata for all new scientific workflows that they have created, allowing others to re-use and enhance existing workflows, and sharing the new version with the broader community.
- DataONE provides the federated data network that enables researchers to easily discover and use data, no matter where the data are located, and simultaneously providing credit to the repositories and providing tools and education resources that are supporting this brave new world of science.

Increasingly, we will see repositories, publishers, and institutions adopting identifiers such as those available through ORCiD as a means to link people and the products of their research and other intellectual endeavors. It is easy to envision many benefits, especially in regards to scientists more readily receiving recognition and attribution for their work.

As we imagine this new interconnected world of science, we must, however, identify and rectify any potential stumbling blocks that might hinder science by under-funded countries, institutions and individuals. In particular, some Federal mandates no matter how well intended, may be more easily accommodated at research universities than at smaller institutions, including those serving under-represented minorities. Clearly, it is not in the best interest of the country if we create a new digital divide. Thus, we need free and open repositories that are open to all scientists as well as access to education resources that allow scientists and students to quickly get-up-to-speed. Likewise, it is important to envision the end game so that we are truly enabling more and better science, not just creating more administrative overhead. Will new Federal mandates alone lead to more and better science?

Join us!

DUG2013

July 7-8th 2013
Courtyard Marriott Chapel Hill, NC
bit.ly/DUGHome

---

—Bill Michener
Project Director, DataONE
Dear DUG Members -

Once again the annual users group is fast approaching. We’ve been working to tweak the agenda in recent days and weeks, and are pleased to have a slightly new format this year, incorporating more opportunities for member participation through roundtables and a poster session. In addition, we will still have the old favorites of DataONE updates and presentations, including status reports from Member Node representatives, tool development, and outreach efforts.

This year’s annual meeting will also be a time for voting on new Chair and Vice Chair DUG leadership. Please make nominations in advance by sending me an email at als9q@virginia.edu, or bring your nomination along. You are welcome to nominate yourself, or a member colleague who you think would serve the community well. Leadership positions are presently for a two-year term, and duties entail communication with DataONE leadership and the DUG membership throughout the year; planning and facilitating the DUG annual membership meeting, and enthusiasm to carry the community forward to reach it’s full potential. It is a great opportunity to engage with a broad spectrum of people interested in research and to refine your leadership skills. Please consider and be in touch.

—Andrew Sallans
Vice-Chair, DataONE Users Group
Head, Strategic Data Initiatives, UVa Libraries
CyberInfrastructure Update

DataONE continues to see steady growth in the amount of content synchronized from Member Nodes and discoverable through the ONEMercury search interface and the programmatic service interfaces. There are currently some 344,000 objects which together provide a total of more than 170,000 publicly accessible data sets. The Taiwan Forestry Research Institute Member Node is the latest to join the federation and represents the second international node after the South African National Parks Member Node. The list of DataONE Member Nodes (besides a couple of replication target nodes operated by DataONE) now includes:

- Knowledge Network for Biocomplexity
- ESA Data Registry
- SANParks Data Repository
- USGS Core Sciences Clearinghouse
- ORNL DAAC
- LTER Network Member Node
- Merritt Repository
- PISCO MN
- Cornell Lab of Ornithology Avian Knowledge Network
- ONE Share Repository
- TFRI Data Catalog

Ensuring service availability for our users is a high priority for DataONE, and since going public in July 2012, we have been able to maintain 100% uptime by balancing maintenance and upgrade activity across the Coordinating Node centers located at Oak Ridge Campus, The University of New Mexico, and the University of California, Santa Barbara. The infrastructure is currently running on version 1.1.3 with a migration to infrastructure version 1.2 anticipated after two more patch releases to the 1.1 series.

Currently under active development for the core infrastructure are two significant features meant to reduce the functional requirements for Member Node participation. The first provides support for “serial identifiers” that enable assignment of unique identifiers to potentially mutable content (e.g., metadata which might be updated to correct spelling) while still ensuring the previous revision(s) can be retrieved through their respective persistent unique identifiers. The second moves much of the responsibility for system metadata (low level information describing the type, size, and so forth of all objects in DataONE) management back to the Member Nodes, and action that makes corrections to system metadata (e.g., changing the format type of the described object) by Member Nodes much simpler and is expected to also reduce latency for propagation of changes to access control rules for example.

The development team is also busy working on the Investigator Toolkit with emphasis on ONEDrive, ONE-R, and the Morpho metadata editor. ONE/Drive is a cross-platform file system driver that enables users to access the entire set of content available through DataONE (or a portion thereof) as a “network drive”. We expect such capability to significantly simplify access to content held in DataONE by many existing analysis tools.

Members of the DataONE Team will be at the following events.

Full information on training activities can be found at bit.ly/D1Training and our calendar is available at bit.ly/D1Events.

- Jul. 7-8 DataONE Users Group Meeting Chapel Hill, NC
  [http://dataone.org/dataone-users-group](http://dataone.org/dataone-users-group)

- Jul. 9-12 Federation of Earth Science Information Partners Summer Meeting Chapel Hill, NC
  [http://esipfed.org/meetings](http://esipfed.org/meetings)

- Aug. 4-9 Ecological Society of America Summer Meeting Minneapolis, MN
  [http://esa.org/minneapolis/](http://esa.org/minneapolis/)

- Aug. 18-23 Intecol 2013 London, UK

- Sep. 6 Camp-4-Data Workshop Lisbon, Portugal

- Sep. 16-18 Research Data Alliance Minneapolis, MN
  [https://rd-alliance.org/future-events](https://rd-alliance.org/future-events)

- Oct. 28-30 EUDAT Rome, Italy
  [http://www.eudat.eu/2nd-conference](http://www.eudat.eu/2nd-conference)
Working Group FOCUS

The Data Integration and Semantics Working Group is currently working on multiple efforts towards DataONE’s vision of helping anyone find, gather, interpret, and use data from the biological, ecological, environmental, and Earth sciences. One such effort is grounded in SemantEco, a semantically-driven framework and portal that has the overarching vision of enabling integration, retrieval, and visualization of environment-relevant data, where anyone can assume the role of data provider or user. The immediate goal of SemantEco, in support of DataONE, is to provide a window into the future of applications that enhance discoverability and search leveraging emerging semantic technologies and practices.

SemantEco is a framework that leverages semantic technologies, in particular the Web Ontology Language (OWL)\(^1\), relevant logical proof reasoners (e.g. Pellet\(^2\)), query languages (e.g. SPARQL\(^3\)), and provenance languages (e.g., PML\(^4\), PROV\(^5\)). The framework ingests data in the Resource Description Framework (RDF) format, explicitly linked to OWL ontologies, to provide search features, data interpretation, and visualization tools to end users. SemantEco includes a modular architecture to address extensibility and debugging issues in a flexible, scalable way to suit the needs of different stakeholders.

Use cases are the driving force behind integrating new data and generating new search capabilities within SemantEco. The initial use case enabled concerned citizens to identify water quality in nearby regions with respect to regulations and health effects. SemantEco has been extended to allow a user to search water or air quality alongside organism (e.g., fish, bird) occurrence data, in an effort to help understand how various chemicals in bodies of water affect nearly populations of various organisms. SemantEco primarily has interfaces for restricting a search for scientific measurement data along geospatial locations (e.g., zip codes), dates, and domains of interest, including air, water, bird, and fish. When water (or air) domains are selected, formal reasoning is used to identify at what sites water (or air) quality measurements violate state or EPA regulations, and correspondingly identify them with the appropriate plotted icons.

One area for enhancement is extracting...
Each Member Node within the DataONE federation completes a description document summarizing the content, technical characteristics and policies of their resources. These documents can be found on the DataONE.org site at bit.ly/D1CMNs. In each newsletter issue we will highlight one of our current Member Nodes.

United States Geological Survey Core Sciences Clearinghouse

www.usgs.gov/core_science_systems

The Core Science Metadata Clearinghouse is an initiative by the USGS Core Science Analytics and Synthesis (CSAS) program to locate, evaluate, and access Earth Science data and information. It contains standardized metadata descriptions of Earth Science data sets and information products contributed from over 95 research organizations (e.g. National Park Service, US Fish & Wildlife, state of Colorado) to help scientists discover a broad spectrum of data. Records describe geospatial and tabular data, in addition to software tools used in analyzing, integrating, and applying data.

Metadata descriptions follow the Federal Geographic Data Committee’s (FGDC’s) Content Standards for Digital Geospatial Metadata; however, other standards are also accepted, such as Ecological Metadata Language (EML), ISO19115, NASA’s Directory Interchange Format (DIF), and Dublin Core.

Metadata and data are acquired from Federal, State, and other research organizations who primarily want to make their research results freely available to the broad biological/ecological communities. Users acquire data by searching the Clearinghouse and downloading data via a “Get Data” capability or by the links to data contained in the metadata records.

The Clearinghouse is a participating node on the FGDC’s National Spatial Data Clearinghouse (NSDI) and the Data.gov.

Other services from the USGS CSAS Member Node include:

- Training/Outreach expertise in Data Management
- Software Development related to ONEShare
- Metadata Editor/Creation Tool for FGDC BDP Metadata
- Computational Science Capabilities related to scientific data analysis and visualization
- Expertise and cross agency leadership related to Federal organizations data management needs/requirements
- Science Data Management best practices, training, leadership and tools.

Data management is increasingly recognized as a vital element of the science lifecycle. More and more science organizations, including the National Science Foundation, require a data management plan as part of the proposal process, consistent with recommended good practice. The Science Data Management component of CSAS provides expertise in data lifecycle best practices and community and interdisciplinary partnership facilitation.

One challenge for this work, as well as the entire DataONE effort to improve discovery and search, is to “link” datasets and metadata about those datasets to existing, relevant Earth, Ecological, and Environmental Science ontologies. These ontologies can be used to enhance data discovery and access through semantic annotations of each dataset, thus improving overall metadata quality. As DataONE is poised as the primary point of access to earth and environmental data and practices, and is schema agnostic, semantic descriptions of these datasets and practices are crucial to discovery across schemas.
Working Group FOCUS cont’d

However since the quality, curation, and coverage of the Earth, Ecological, and Environmental Science ontologies vary, other internship efforts guided by our working group address ontology mappings and ontology evaluation. The ontology evaluation internship focuses on evaluating the coverage of Earth and Environmental Science ontologies for representing DataONE metadata. Ontologies including the Semantic Web Earth and Environmental Taxonomy and the Environment Ontology are being evaluated for the coverage of DataONE-relevant terms and concepts. The evaluation uses automatic methods such as part-of-speech tagging on the DataONE corpus and will compare classes and relationships thus uncovered with entities in the chosen ontologies.

Another way to ascertain this degree of quality is to locate terms with similar semantics between two or more ontologies and, based on their annotations and surrounding concepts in the ontologies, have domain users assess the comparative quality. The ontology mapping internship is focused on implementing various algorithms for mapping ontologies and evaluating their results in the Earth, Ecological and Environmental Science domains. The intern uses state-of-the-art algorithms from the Ontology Alignment Evaluation Initiative to discover potential mappings between the Semantic Web Earth and Environmental Taxonomy, the Environment Ontology, and others. A previous internship applied a Probabilistic Topic Modeling method to enrich DataONE metadata and thus improve the discoverability of DataONE datasets. The results of this internship are documented[1].

There are other working group efforts aimed at improving the quality of Earth, Ecological, and Environmental Science ontologies and improving existing search capabilities. Within the Text Mining for Ontologies (TMO) project (see next issue), we are developing topic models from collections of Earth, Ecological, and Environmental Sciences corpora. These models can serve as raw material to identify ways to improve domain ontologies by helping to identify gaps in coverage, and as materials to evaluate and improve the quality of semantic annotations at large. These models can also be used for improving search capabilities, including search term suggestions and automated query expansion.[2]