Volume 2 Issue 2

DataONE NEWS

DataONE: Changing the Cyberinfrastructure Landscape

As we move into 2014 it is important to take stock of the significant progress that has been made with respect to building DataONE. During Phase I (2009-2014), more than 300 participants designed, developed and deployed innovative and robust cyberinfrastructure (CI) and services, and directly engaged and educated a broad stakeholder community. DataONE now provides a robust, scalable infrastructure using Member Nodes (data repositories), Coordinating Nodes (global metadata catalogs), and an Investigator Toolkit to support the data access and data management needs of biological, Earth, and environmental science researchers in the U.S. and across the globe.

Member Nodes (MNs) are located worldwide and fill a crucial role in the federation by contributing and curating data and metadata, providing key domain expertise, and enhancing best practices. By participating in DataONE and using a common set of service interfaces, MNs gain a simple, consistent programming interface for data access to better serve a broad range of users. MNs can participate in the federation at different levels of capability. These levels start at Tier 1 (read-only access to public content) and progress through Tier 4 (full capabilities including content replication).

The Coordinating Nodes (CNs) are a mirrored and fault-tolerant service hub, established at three institutions: the University of California at Santa Barbara, the University of New Mexico, and the University of Tennessee/DOE Oak Ridge National Laboratory collaboration (aka Oak Ridge Campus). The CNs facilitate MN operations by ensuring uniqueness of identifiers across the system, providing a common infrastructure for identifying users across institutions, managing replication of metadata and data, hosting the metadata search tools, and providing services to ensure consistent and reliable network operations.

The Investigator Toolkit is the third key CI component. Investigator tools are analysis or data management tools (e.g., R, VisTrails, Kepler) developed or adapted to take advantage of the common service interfaces exposed by the DataONE federation. Data producers and consumers use the Investigator Toolkit via secure protocols to locate, analyze, interpret, and submit relevant data and metadata. Because DataONE uses common service interfaces, users gain access to all data and services in the federation. DataONE has developed open source libraries in Python and Java to assist developers to integrate widely used tools, such as Morpho, Kepler, and Zotero, into the Investigator Toolkit (Fig. 1). Other standalone tools have been developed in conjunction with community partners to support critical elements of the data life cycle such as data management planning (e.g., the widely adopted DMPTool that has been used to create thousands of data management plans for NSF and other research sponsors) and DataUp, a new tool that is used to create well-documented, preservation-ready Excel spreadsheets and deposit them in a Member Node.

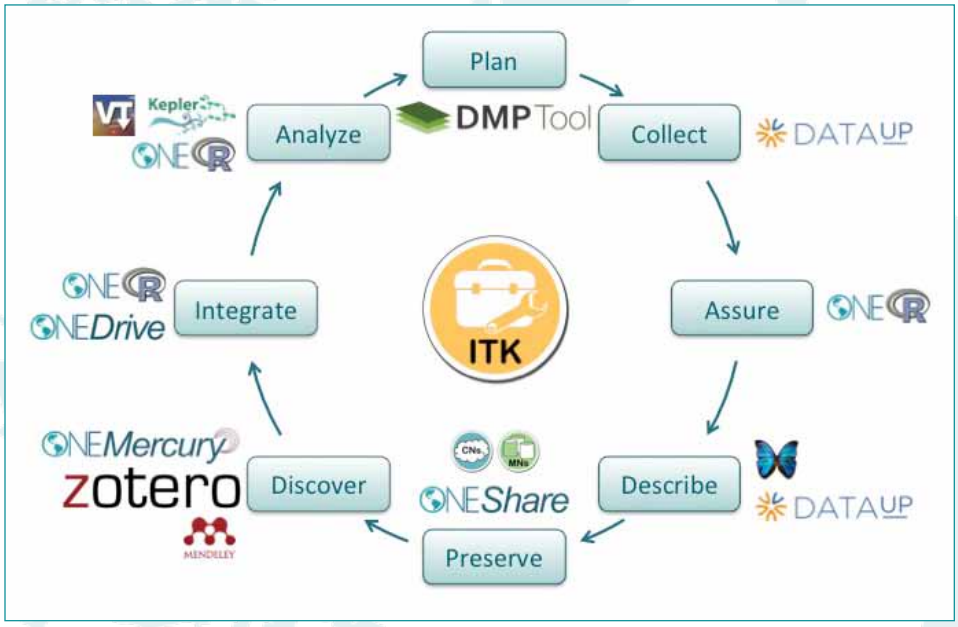


Fig 1: Tools associated with the Investigator Toolkit (ITK) supporting the research data life cycle

Underpinning the DataONE CI are the products of three Working Groups. First, the Provenance in Scientific Workflows Working Group is developing and prototyping innovative provenance management technologies for integration into DataONE during 2014. DataONE views provenance metadata as a "first-class citizen" and foresees its importance to grow as more applications become provenance-enabled to support more reproducible and open science.

Second, the Scientific Exploration Visualization and Analysis (EVA) Working Group informed DataONE design and development, especially with respect to the tools and services needed by domain scientists to address challenging research questions. The first EVA Working Group project developed the eBird Gulf Spill Bird Tracker (ebird.org/tools/oilspill) and also focused significant effort on understanding the dynamics of continental-scale bird migration. This latter work was featured in Nature and required 1,088,259 SU's (cpu-core hours) of TeraGrid/XSEDE computing resources. The EVA results provided the foundation for the analyses that underpin the 2011 and 2013 State of the Birds reports. These reports provide the basis for new science and conservation efforts by numerous federal agencies and conservation organizations. The EVA Working Group is now focused on the development of model analysis, visualization, and benchmarking CI for managing output from terrestrial biosphere models. Working with international

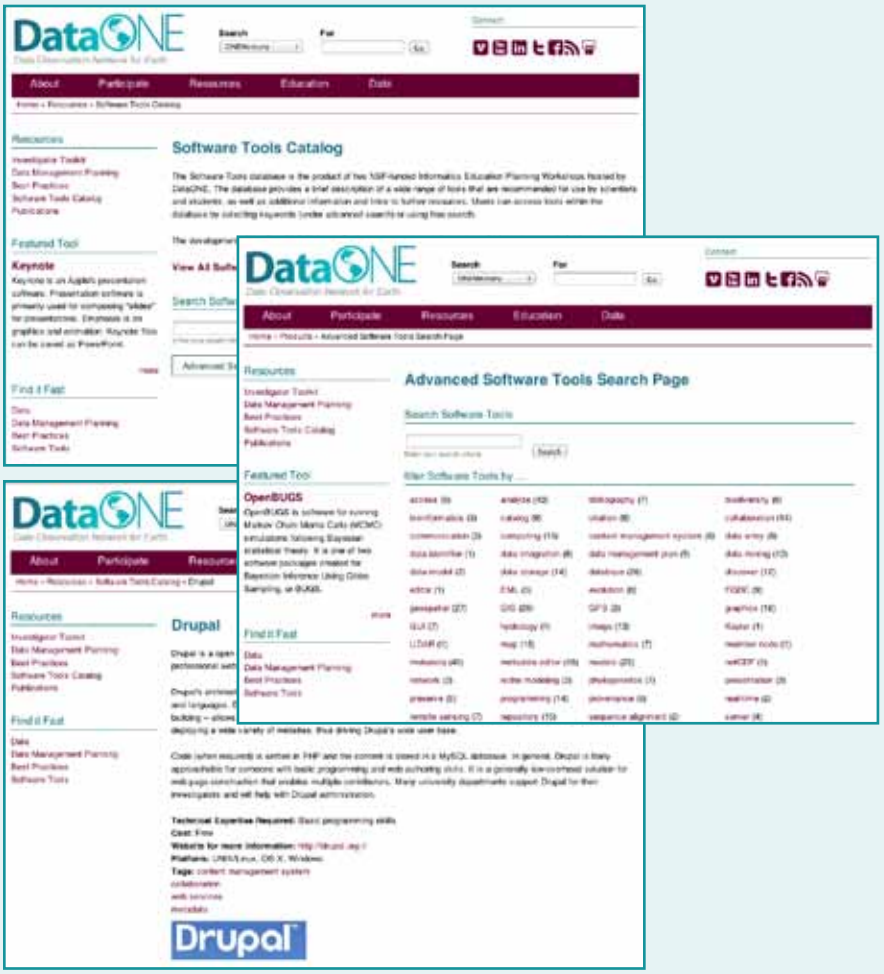
cont'd page 2 >>>

DataONE NEWS

FeaturedRESOURCE

Software Tools Catalog

The DataONE Software Tools Catalog is one of two resources designed to aid users in their need for information and software to support good data management practices; the other resource is the Best Practices Database. The catalog is the product of a collaborative effort across many individuals and provides a brief description of a wide range of tools that are recommended for use by scientists and students. Other relevant information includes required expertise, platform, cost and additional links to relevant information. With over 200 tools listed, tags and free text search make tools within the catalog easily discoverable.



Our catalog is constantly being reviewed and edited to keep it current and community feedback is critical in this regard. We are also looking at enhancing usability of the tool following discussion and feedback during our DataONE Users Group. If you know of a tool that you feel would be valuable to the Earth and environmental science research community and that is not currently listed, please consider contributing to the resource. The catalog submission form can be download here and you can send this, or other comments and suggestions to info@dataone.org.

CoverSTORY cont'd

modeling teams, the EVA Working Group has added model inter-comparison and model-observation functionality to existing community tools. The EVA Working Group has been collaborating closely with the EarthCube Brokering project to use Web services for discovery, access, and analysis of terrestrial biosphere model output and observations.

Third, the Semantics Working Group, with a goal of improving search and discovery, and the integration of heterogeneous data resources, has assessed and developed practical solutions for incorporating semantic technologies into DataONE. Approaches to challenges such as keyword normalization and alignment of semi-structured metadata with more formal ontologies and other controlled structures have been prototyped by the Semantics Working Group.

Building on our progress in Phase I, DataONE can now target ambitious goals that enable new scientific discoveries and that require massively increasing the scope, interoperability, and accessibility of data. We will do so in Phase II (2014-2019) by: (1) supporting an even wider variety of common repository software than we have already provided, supporting hundreds of community data repositories; (2) enabling user-initiated data processing at local Member Node repositories to support data extraction, quality assurance annotation, and other services on large data sets; (3) augmenting our data discovery system with semantic knowledge about measurements to allow highly precise searches; and (4) supporting emerging provenance frameworks within both our data discovery system and within tools targeting researchers.

We look forward to a very productive 2014 and the beginning of the second phase of DataONE. In addition to the significant technical accomplishments, DataONE has been a major force in engaging the broad community and in training an informatics-literate workforce—a topic for the next newsletter. In the meantime, Happy Holidays and best wishes for the New Year to all our participants, collaborators, and users. ■

—Bill Michener
Project Director, DataONE

DataONE NEWS

CyberSPOT

CyberInfrastructure Update

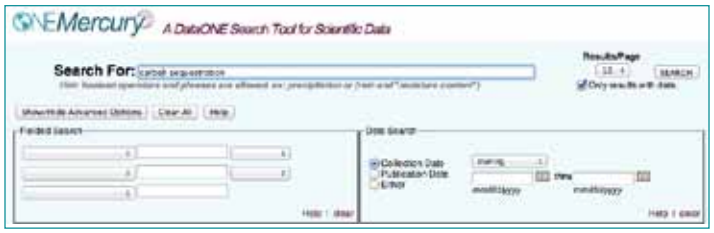
Fall was a period of design work as the DataONE development team prepares to build and deploy some significant changes to the DataONE service infrastructure. As mentioned in the previous newsletter, there are two main areas of active design development. The first is the intent to move authoritative control over system metadata to Member Nodes rather than Coordinating Nodes. This is necessary to reduce latency for client applications using the DataONE service interfaces to manage content on a Member Node.

The other design topic relates to the process for supporting mutable content, a topic which received further detailed analysis and design during the All Hands Meeting and subsequently. It turns out that there are several viable options for supporting mutable content, and the development team is in the process of finalizing design for one such strategy that we expect to be testing early in 2014.

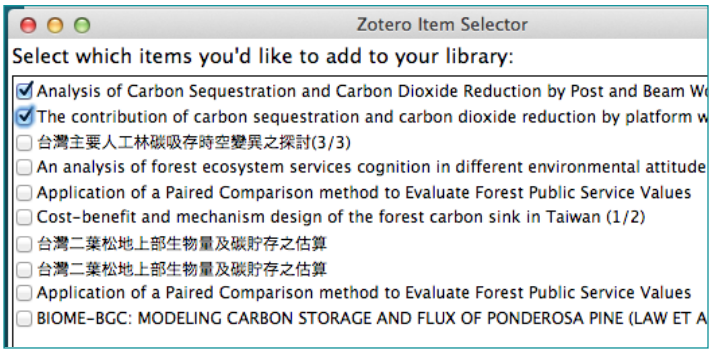
The development team has also been making great progress on a system wide dashboard to display real-time status of the DataONE infrastructure and participating Member Nodes. The initial version of the Dashboard is currently in testing, expected to be released for general use sometime in January.

There are currently more than 230,000 data sets comprised of more than 426,000 individual objects available through the DataONE infrastructure. The ONE*Mercury* web search interface (https://cn.dataone.org) provides the main mechanism for discovering content held by Member Nodes when using a web browser. Keeping track of interesting results can be challenging though, and this is one reason why since early on the ONE*Mercury* interface has supported interaction with online citation managers such as Zotero (www.zotero.org) and Mendeley (www.mendeley.com) through standard COinS tags (Context Objects in Spans, http://ocoins.info). With the appropriate plugin installed, tracking search results in a citation manager is a straight-forward process as shown in the Zotero example with the Google Chrome browser below:

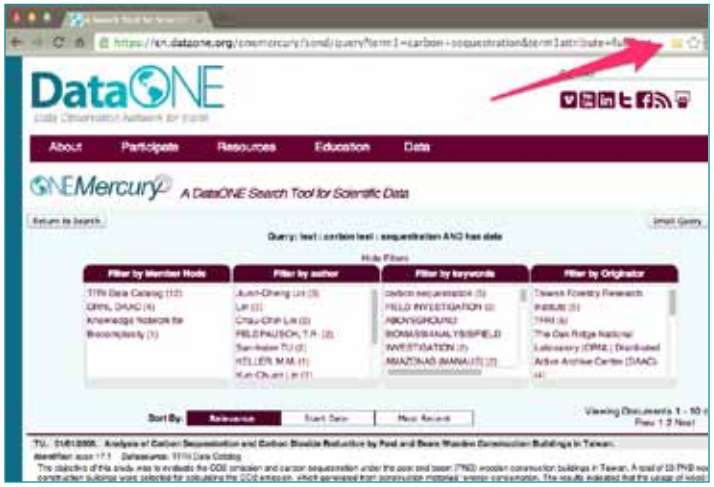
1. Navigate to the ONE*Mercury* search interface, and enter a search term such as "carbon sequestration".



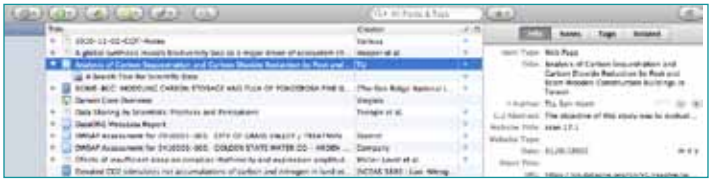
2. On the results page, a folder icon should appear in the URL box of the browser.



3. Clicking on the folder icon will open a list of citable content appearing on that page. Selecting one or more entries here will then add them to your Zotero reference collection.



4. Navigating to your reference collection will show the recently added results, conveniently stored for later use and sharing with your colleagues.



DataONE NEWS

WorkingGroup FOCUS

Preservation and Metadata Working Group

The DataONE Preservation and Metadata Working Group is a multi-tasked working group addressing two key areas, 'preservation' and 'metadata.' We've been pursuing our goals via two cleverly named subgroups: the Preservation subgroup and the Metadata subgroup.

Specific to preservation, our chief accomplishment has been to draft a Preservation strategy statement and a set of guidelines that were adopted by DataONE in 2010. These contributions covered three main areas of digital object preservation: (a) keep the bits safe, (b) protect the form, meaning, and behavior of the bits, and (c) safeguard the guardians, i.e., attend to the sustainability of the federation, the tools, and the partners. This work represents an important milestone, and updates are anticipated in 2014.

During the last year-and-a-half, metadata activities have dominated our work, and they are the focus of this newsletter piece. Briefly, in reaction to a shared dissatisfaction with traditional metadata standardization processes, we embarked on an unusual path to create a very open, crowd-sourced metadata dictionary. The aim is provide, at low cost, a high quality, trustworthy, relevant, cross-domain set of vocabulary terms, all ready to go with URIs for linked open data applications.

Constraints of the Metadata Standards Process

Many of us are veterans of passionate, perhaps highly spirited metadata standards meetings. These are meetings that attract some people, and make others want to run for the doors. There's a lot that happens when trying to achieve consensus on metadata elements: properties, labels, definitions, name spaces, refinements, and the whole package that is ultimately pulled together an as a metadata standard. There are added layers when seeking official endorsement from various agencies, such as the National Standards Organization (NISO), International Standards Organization (ISO), or Institute of Electrical and Electronics Engineers (IEEE).

The standards process often yields rewarding results, and endorsed standards have been tremendously beneficial for advancing cyberinfrastructure. There is, however, another reality, and that is that the metadata vetting process is seen as exclusive, unfriendly, and, at times, even "brutal." These obstacles, coupled with the digital data deluge, have led to a glut metadata committees, and hence a glut of metadata standards that are labeled as discipline-specific, even though they often have considerable overlap in function (Figure 1). The cost of bringing the committees together and achieving consensus reduces the amount of timely review, which means that many of these standards are out of date. Moreover, the quality of these efforts is often unsatisfactory, consistent with "design by committee". We also face a "silo situation" where the most current version of a metadata standard is not widely discoverable by others, residing in a corner of an agency Web site, or somewhere buried in the deep web.

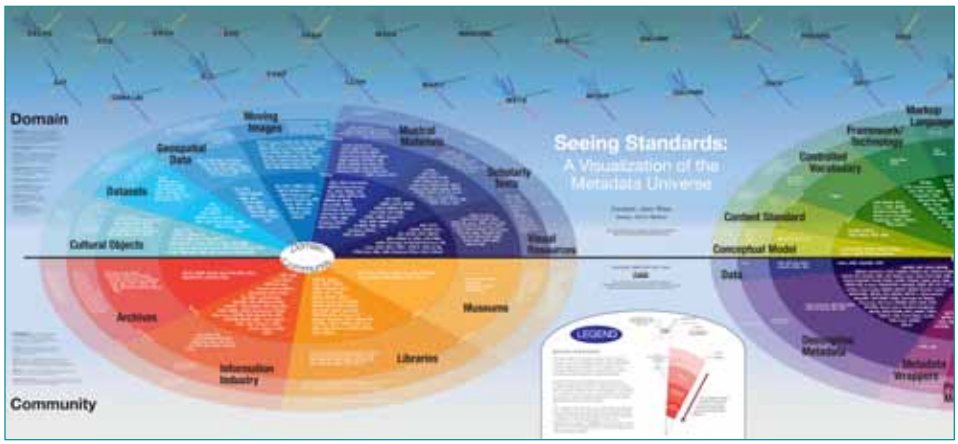


Fig 1: A universe of difference, but overlapping metadata vocabularies has sprung up around traditional, siloed standardization processes. This is expensive and confusing to users [1]

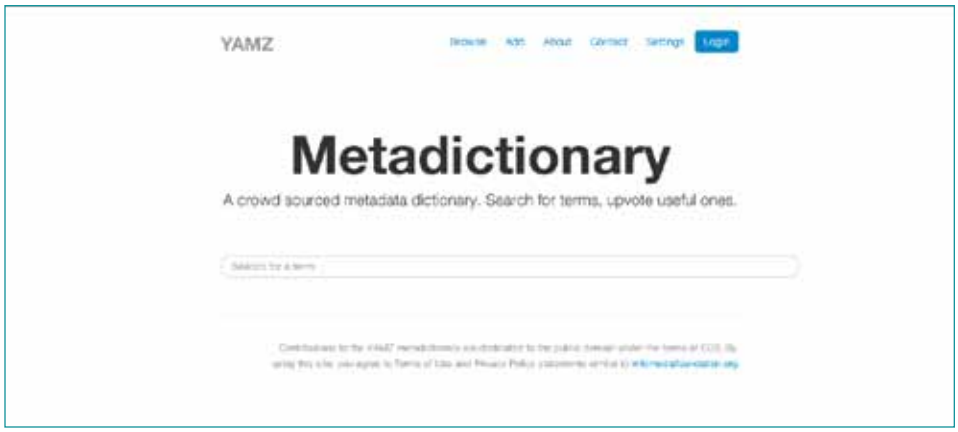


Fig 2: The home page of the YAMZ metadata dictionary (yamz.net)

Innovating via YAMZ (Yet Another Metadata Zoo)

Providing an open, cross-domain approach for metadata standardization may eliminate many of the negative aspects of the troubled, traditional "panel of experts" approach. Clearly, metadata is crucial to a full range information activities and workflows underlying DataONE, and that provided a perfect setting for advancing our approach. In fact, it's challenging to find a DataONE working group that actually has no connection with metadata. Moreover, it's hard to draw strict lines around the domains that should be served, especially as DataONE is looking to expand coverage beyond Earth science. These factors, and the above noted constraints inspired the Metadata subgroup to focus on infrastructure design that embraces social technology, and we've

cont'd page 7 >>>

DataONE NEWS

The DUG out

Hello DUG Members:

We are pleased to bring you another update. This update focuses on the DataONE All Hands Meeting (AHM), which Andrew and I attended in October in beautiful Albuquerque, New Mexico. Since we did not have any official responsibilities at the meeting, we floated from group to group to observe and get an overview of what they were working on. It was quite informative and exciting to see the entire DataONE organization working on so many separate projects towards a common goal. Here are some of the highlights from the AHM:

1. The DataONE leadership updated us on the plans for the future. They are working on the proposal for the next five-year renewal from the National Science Foundation. A goal of the next phase will be to develop a plan for sustainability.
2. The Sociocultural Issues Working Group studied a variety of organizations to learn some lessons from successes and failures to apply to DataONE's eventual transition from a funded organization to a sustainable business model.
3. The Usability and Assessments Working Group worked on developing questionnaires for additional surveys, such as the follow up survey for academic libraries and librarians. The usability sub-group tested usability on the new DMPTool version 2.
4. The Community Education and Engagement Working Group developed some hands-on activities to accompany the data management training modules. They also developed some "data stories" to demonstrate the need for sound data management practices.

There were many other groups who worked diligently at the AHM, but we were unable to attend all their meetings. As you can see, there is much going on and many avenues for involvement with DataONE, so we encourage you to get involved. We also encourage DUG members to hold sessions on behalf of the DUG at conferences. This will help to publicize the DUG and broaden its membership.

We are looking forward now to our summer meeting in 2014. We hope to soon be able to give you more details, but at this point, the plans are still being made. We will be in contact with the DUG Steering Committee to plan the 2014 DUG meeting agenda. If you have suggestions for the agenda, or would like to participate in the Steering Committee, please let us know at dugchairs@dataone.org. We look forward to seeing you at the meeting. ■

—Chris Eaker
Vice-Chair, DataONE Users Group
University of Tennessee Library
—Andrew Sallans
Chair, DataONE Users Group
University of Virginia Library

Member Node DESCRIPTION

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.

Sustainable Environment Actionable Data (SEAD) Virtual Archive

http://sead-data.net

One of DataONE's newest Member Nodes is the SEAD (Sustainable Environment, Actionable Data) Virtual Archive. The SEAD Virtual Archive (VA) is a thin virtualization layer on top of multiple university institutional repositories focusing on preservation of long-tail scientific data. SEAD provides lightweight data services specifically designed to support multi-disciplinary research and to meet the needs of small-team and single investigator projects.

The Virtual Archive is one of three interacting components: an Active Content Repository (ACR), a community research profile and analytic service, and the Virtual Archive (VA).

- The ACR provides secure project spaces where data can be collected, shared, annotated, analyzed, used to create new data products, and ultimately published.
- Powered by VIVO, the community research profile and analytic service tracks information about real-world entities (e.g. people, projects, centers) and provides links and citation information about papers, presentations, and data.
- The Virtual Archive (VA) packages fixed, bounded versions of the data and information from ACR spaces and VIVO into new data collections, generates Digital Object Identifiers (DOI), matches collections to appropriate long-term repositories working with SEAD, and registers the data across SEAD and external data discovery services.

DataONE and SEAD are both funded as part of the innovative National Science Foundation (NSF) DataNet program. The SEAD Virtual Archive is the first DataNet project to join the DataONE network as a Member Node, strengthening the collaborative relationships that exist across the DataNet Federation. Other DataNet projects include the Data Conservancy, Terra Populus and the DataNet Federation Consortium.

DataONE NEWS

Outreach UPDATE

Winter marks the beginning of preparations for our 2014 season of conference events and the DataONE Summer Internship Program.

In a change from previous years, we have decided not to run data management workshops at the Ecological Society of America meeting in Sacramento, August 2014. While our workshops have always been well attended and well received, we were limited by the number of people that could participate in half-day, hands-on sessions. In contrast to our 1.5 hr special sessions that attracted in the region of 200 people per event, workshops could only accommodate 30 or so individuals. So, while we will not be running lengthy workshops, we will continue our training and education activities through special sessions. Additionally, we had so much fun participating in the ESA 2013 Ignite sessions that we have submitted a couple of proposals in that format. In collaboration with Sandra Chung from NEON and others, DataONE plans to bring you a series of 5 minute ignite presentations on Best Practices for Data Management, Tools for Data Management, and Science Communication.

We are also preparing for the International Digital Curation Conference in February 2014. IDCC will be hosting two DataONE affiliated workshops. On Monday Feb. 24th the DMPTool group will be running a morning session titled "Data Management Planning: Get up to date with DMPTool and DMPonline". Later in the week on Thursday Feb 27th, DataONE will be running a Member Node Implementation Workshop. This one-day, hands-on workshop will:

- Provide a brief overview of the Data Observation Network for Earth (DataONE) project,
- Explain the benefits for groups and institutions of collaborating with DataONE as a Member Nodes,
- Present different ways to participate as a Member Node,
- Provide hands-on time in how to establish a Member Node using available software systems, and
- Demonstrate how to use the DataONE web services to access content from client applications.

By the end of the workshop, participants will understand the design of DataONE, the services that DataONE provides to its Member Nodes, and the technical details needed to establish a Member Node at their organization. Free to attend, the target audience includes information managers and technical staff at organizations that are interested in becoming DataONE Member Nodes or that have started the process of becoming a DataONE Member Node. Members of the DataONE Core Cyberinfrastructure Team and Community Engagement and Outreach Team will be there to provide support and information. And if that weren't enough, we now have DataONE stickers for visitors to pick up at our workshops, exhibition stands and other events.

For more details on the IDCC meeting and to register for either of the workshops visit: http://www.dcc.ac.uk/events/idcc14.

DataONE stickers, coming to an event near you (bit.ly/D1Events)

By the next issue of DataONE NEWS we will have announced our 2014 Summer Internship Program. Planning is underway and as in previous years, we are looking to support up to eight summer interns to work on projects that are related to either Cyberinfrastructure or Community Engagement and Education. The specific nature of individual projects has not yet been finalized but information will be released via the DataONE website and mailing lists in mid-February. Applications will be due in March 2014 and if interested, you can browse through previous projects on the DataONE internship page and through the open notebooks kept by the summer interns. Details on eligibility requirements, time frame and stipends are already posted and we will update the site with project information in the new year.

Finally, having mentioned the mailing list, we've now made it even easier for you to subscribe and to join our Users Group. Just go to DataONE.org, click on the desktop mouse and provide your name, email and institution. Simple! ■

Upcoming EVENTS

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*.

Jan. 8-10
Federation of Earth Science Information Partners (ESIP) Winter Meeting Washington, DC
http://commons.esipfed.org/taxonomy/term/464

Feb. 24-27
International Digital Curation Conference San Francisco, CA
http://www.dcc.ac.uk/events/idcc14

Mar. 26-28
Research Data Alliance Dublin, Ireland
https://rd-alliance.org/rda-third-plenary-meeting.html

WorkingGroupFOCUS cont'd

developed a prototype metadata dictionary known as YAMZ (formerly SeaIce).

YAMZ (Yet Another Metadata Zoo) can be seen as a diverse, living collection of terms that have been contributed, refined, and approved by the community that directly uses them. YAMZ supports a community-driven approach to establishing metadata term semantics and definitions. Chief goals include reducing duplicative metadata activity; reducing costs associated with the metadata standards process; enabling an open, community driven approach for all metadata stakeholders; and unifying metadata practices across disciplines. Figure 2 shows the present home page (with the old name not yet changed unfortunately).

The YAMZ design was inspired by the kind of reputation-based voting found in Stack Overflow (http://stackoverflow.com/). Imagine an unencumbered, trustworthy, community-driven meritocracy bringing you the best metadata terms and definitions. Further, consider an environment where any stakeholder could engage in discussion about the various facets of a metadata term (e.g., name, definition and so forth).

YAMZ integrates these considerations in its approach to crowd-sourcing, seeking to reduce traditional barriers, particularly the endless heated discussion that seems an inescapable part of traditional metadata consensus gathering. DataONE is the target implementation community, which is ideal in some sense, given the range and variety of disciplines represented (e.g., ecology, biology, geology, astronomy, etc., and the many sub-disciplines) and the diversity of metadata stakeholders.

Uptake and Next Steps

The YAMZ metadata dictionary has already received a good deal of attention, via our initial prototype. Already the DataONE Semantics, the Provenance and Workflows, and the Public Participation In Scientific Research (PPSR) working groups have all expressed interest in using YAMZ. Starting in 2014, we will be running an exploratory study with the PPSR group, as they develop a core set of metadata properties for community efforts that engage the public in the collecting scientific data. On an international level, the Research Data Alliance Data Foundation and Terminology Working Group has been actively working with YAMZ, and contributed over half the terms in it.

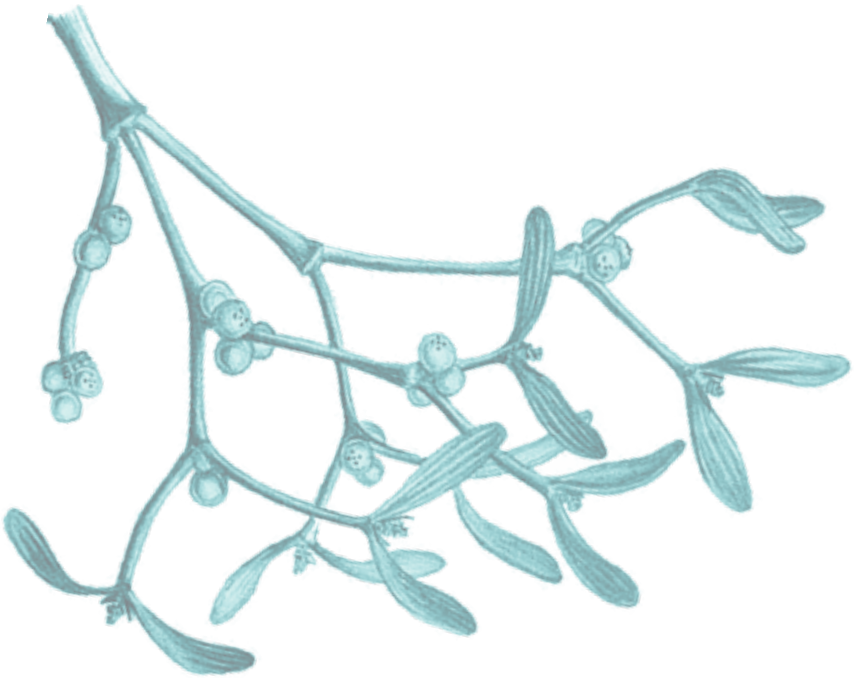
Priorities for the Metadata subgroup for 2014 include further testing; improving our ranking algorithm to allow proposed terms to move from vernacular (where all terms are born and evolve) to canonical (where terms cease to evolve), or to deprecated; and strengthening our identifier strategy[2]. These goals are interwoven with our efforts to spread the word, and encourage all metadata stakeholders to experiment with YAMZ.

Please visit us at yamz.net, but be kind, as it is still a prototype! ■

—John Kunze
California Digital Library
CoChair, Preservation and Metadata Working Group
—Jane Greenberg
University of North Carolina Chapel Hill
CoChair, Preservation and Metadata Working Group

[1] Riely, J. (2009-2010). Seeing Standards: A Visualization of the Metadata Universe: http://www.dlib.indiana.edu/~jenlrile/metadatamap/.

[2] Kunze, J., Janee, G., and Patton, C. (2013). Persistent Identifiers for Terms in a Crowd-Sourced Vocabulary. CAMP-4-DATA, 2013 Dublin Core international metadata conference: http://wiki.dublincore.org/images/3/39/Jak_pids4terms_d1mwg.pptx.pdf.



DataONE
1312 Basehart SE
University of New Mexico
Albuquerque, NM 87106
Fax: 505.246.6007

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement.

Project Director:
William Michener
wmichene@unm.edu
505.814.7601

Executive Director:
Rebecca Koskela
rkoskela@unm.edu
505.382.0890

Director of Community Engagement and Outreach:
Amber Budden
aebudden@dataone.unm.edu
505.205.7675

Director of Development and Operations
Dave Vieglais
dave.vieglias@gmail.com