Volume 3 Issue 4

DataONE NEWS

©2015 DataONE 1312 Basehart SE University of New Mexico Albuquerque NM 87106

DataONE Through the Looking Glass: How does it work and does it work well?

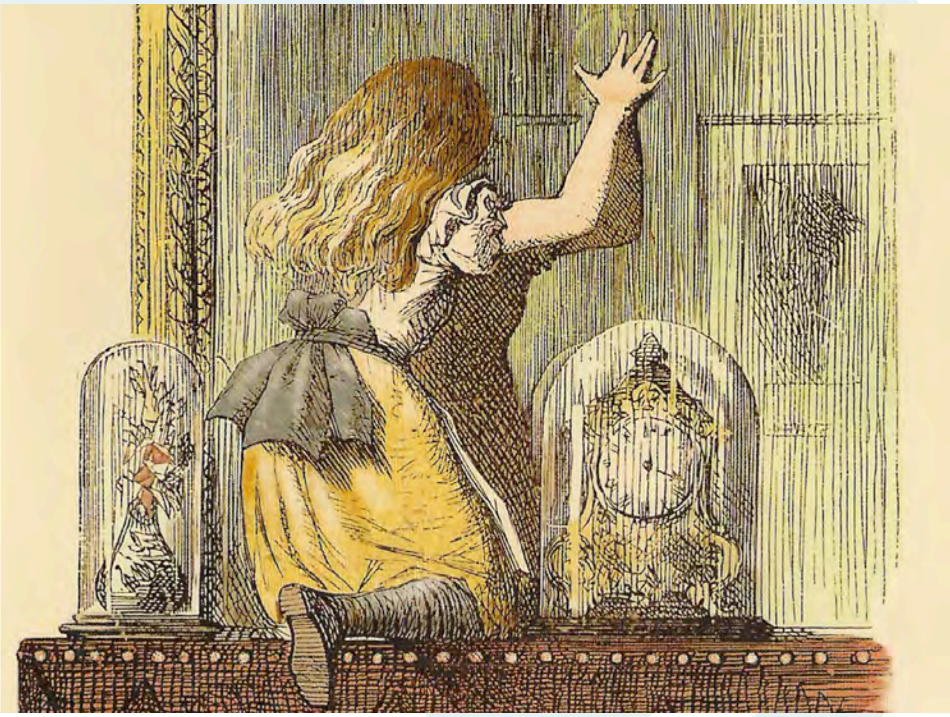
Scientists, engineers, and schoolchildren all recognize that the structure and function of "things" are closely related. This relationship also applies to organizations, including DataONE. I am often asked about how DataONE is organized and how does it work. With respect to the first question, DataONE has adopted a hierarchical structure that is designed to support high productivity and communication among Working Groups—usually 6-16 individuals that focus their efforts on a particular topic or set of related topics (e.g., cyberinfrastructure (CI), citizen science, usability & assessment (U&A)). The Working Group concept was adapted from the approach that was more or less pioneered at the National Center for Ecological Analysis and Synthesis (NCEAS) and that has now been adopted in synthesis centers worldwide. Working Groups provide a loosely structured framework that enables small groups of scientists and educators to intensely focus on a challenging topic(s) at meetings (usually 3-5 days in length) that are repeated two or more times a year for one to two years and sometimes longer. The current (starting in 2014) DataONE organizational structure includes four Working Groups (i.e. two that were mentioned previously, CI and U&A, plus Community Engagement and Outreach (CEO) and Sustainability and Governance (S&G)). Previously in Phase I of DataONE (2009-2014), there were up to ten Working Groups that were operational at any given time covering the topics previously mentioned plus Citizen Science, Sociocultural, and numerous CI-related topics.

Working Groups operate within a hierarchical structure that is designed to optimize communication with all participants. Each Working Group is run by two co-chairs. One or both of the co-chairs also serve on the DataONE Leadership Team which meets weekly for one hour to review and report on progress, plan next steps, and address emerging issues. Four senior staff members serve on both the Leadership Team and an Executive Team, which also meets weekly (and more often via email) to manage the overall project. Several developers (i.e., programmers) report to the co-chairs of the CI Working Group. Each Working Group also engages and mentors one or more graduate students, post-docs, or staff member(s).

So, how does it all work and does it work well? DataONE has been fortunate in that several members of one of its Working Groups (Kevin Crowston, Alison Specht, Carol Hoover, Katherine Chuboda and Mary Beth Watson-Manheim) set about to answer that question in a study that culminated in a paper entitled "perceived discontinuities and continuities in transdisciplinary scientific working groups"[1]. Discontinuities refer to disruptions in the communication flow within an organization whereas continuities refer to those actions that can be taken to mitigate the disruptions. The paper summarizes the results of surveys of the members of the DataONE Working Groups.

Key findings included:

- Working Groups in the first phase of DataONE included ten or more major disciplinary fields (e.g. biologists, computer scientists, librarians, engineers, educators) and eight different professional positions (e.g., student, tenured faculty, librarian).
- Communication challenges for Working Group participants were most pronounced at the start of the project and included: difficulty in learning one another's discipline-specific jargon; lack of time, on one hand, versus the desire to have more frequent communication (especially between different Working Groups), on the other; maintaining momentum between face-to-face meetings; and uncertainty about roles.
- Numerous continuities, or mitigation approaches, emerged within the Working Groups and DataONE. First, Working Group



Tenniel sketch from "Alice through the looking glass" 1871

Outreach UPDATE

2015 Summer Interns

Booma Sowkarthiga Balasubramani
Making a Robust and Useful Earth Science Ontology Repository

Michael Yoo
A Tool for Live, Interactive Workflow Views over Programming Scripts

Mark Anthony Freeman
Evaluating the impact of data access: The role of metrics

Yue Zhang
Network and social media communication analysis of the DataONE user community

I am pleased to announce the 2015 DataONE Summer Interns. Working on projects ranging from Network and Social Media Communication Analyses to Developing Live Interactive Workflow Views to Evaluating the Impact of Data Access and Development of an Earth Science Repository, the 2015 Summer Interns are just over mid way through their projects.

Prospective interns apply for the opportunity to work on one or more advertised opportunities late Winter and begin their nine week internship early in Summer. In most cases, interns work remotely from their primary and secondary mentors, communicating via email and teleconferences following a face-to-face meeting at the beginning of the internship. Competition is high for the internship positions and we are pleased that over the last seven years the DataONE internship program has attracted highly qualified candidates from a broad range of disciplines and backgrounds. 2015 is no exception.

This year, four internship opportunities were offered and the progress of the projects can be followed in their online notebooks at: https://notebooks.dataone.org. Intern bios can be found at: https://www.dataone.org/2015_interns.

If you've been following our DataONE Webinar Series you'll know that after 4 excellent and well attended webinars, we are now taking a break until the Fall. Summer is a time for re-energizing, vacationing and conferencing! We have been busy preparing our own DataONE Users Group Meeting that will precede the Summer ESIP meeting in Pacific Grove, CA Jul 12-13. We have a full agenda of exciting presentations and discussions planned, as well as opportunity to enjoy the natural beauty of the area. Can't attend? Not a problem! Much of the content from the two days will be open to remote participation. The full agenda, registration information for both in-person and remote participation can be found at https://www.dataone.org/dataone-users-group/2015-meeting.

We are also preparing for our annual set of talks, workshops and exhibition booth at the Ecological Society of America meeting in Baltimore this August. As always, the ESA represents a great opportunity to connect with many of you within our ecological research community and we encourage you to find us during sessions or in the exhibit hall. This year we are excited to be co-hosting workshops with Terra Populus (another NSF DataNet) and the Center for Open Science. As the ESA draws closer, details of all DataONE related talks and sessions can be found at: https://www.dataone.org/training-activities. ■

››› cont'd from page 1

members reported that they actively listened to their colleagues, learned the disciplinary terminologies, and viewed the overall Working Group experience as "open and collegial." Second, Working Groups instituted regular virtual meetings in between the face-to-face meetings. Third, the annual All Hands Meeting was valuable for reducing uncertainty about roles, setting goals and agreeing on priority actions, and communicating across Working Groups. Fourth, librarians and information scientists may have played a key "translator" role in bridging disciplinary divides.

The authors concluded that challenges (i.e. discontinuities) are bound to exist whenever transdisciplinary teams are formed. These communication challenges may best be resolved by creating an environment where people feel safe in presenting diverse opinions, where there is a willingness to share and adapt communication practices, and where bridge builders such as librarians are actively engaged.

In addition to the article's recommendations, I offer my own observations about how Working Groups can be most successful. First, spend sufficient time in the selection process to ensure that you are bringing in knowledgeable experts that, importantly, are open-minded and enjoy working with others. Second, respect everyone's time by creating conditions where individuals can be optimally productive (e.g., setting group expectations, allowing time for innovation and discovery). Third, seek and act upon suggestions from the Working Group participants (including allowing your project to be studied). Last and perhaps most importantly, feed people well and celebrate group and individual successes. ■

—Bill Michener
Principal Investigator

[1] Crowston, K., et al., Perceived discontinuities and continuities in transdisciplinary scientific working groups, Sci Total Environ (2015), http://dx.doi.org/10.1016/j.scitotenv.2015.04.121

CyberSPOT

Status

The DataONE production environment now has 30 participating Member Nodes providing access to more than 93,000 publicly readable, current version data sets comprised of 136,000 metadata and 312,000 data objects. A total of 831,953 individual objects are being tracked and can be resolved and retrieved through the DataONE infrastructure.

New Member Nodes since the last report include:

- IARC Data Archive
- NM EPSCoR Tier 4 Node
- TERN Australia
- Northwest Knowledge Network

Infrastructure updates since the last DataONE Newsletter include several bug fixes, improved log aggregation services with COUNTER compliance, and about to be released additional indexing support to facilitate science metadata conforming to implementation of the ISO 19115 specification for geographic information and services.

Development activities continue on several topics including support for Member Node service registration (through ISO 19119), light weight Member Nodes ("Slender Nodes"), implementation of the Version 2.0 service interfaces, enabling provenance tracking and search, and improved search through semantics

Spotlight: Logging

Content creation, modification, replication, and access is logged across all Member and Coordinating nodes and is available through the log service end point. Since data and metadata in the DataONE federation may be replicated across multiple servers, a complete view of content access is only possible by aggregating the log information from all Member Nodes that may have a copy of a particular object. The Coordinating Nodes perform the log aggregation process, periodically harvesting the raw log records from all Nodes and storing them in a high performance search index that allows reporting across different analysis dimensions. Summary log results are available through the Coordinating Node log retrieval interface, though detail information (such as IP addresses) is not available unless appropriately authorized.

While unfiltered log records are useful for some system monitoring and related activities, scientifically-meaningful analysis of log records requires that we correct log records for common events that would otherwise artificially inflate the statistics, such as access by web-indexing robots and multiple accesses from the same individual. Within the publishing community, the COUNTER standard has been used to provide a consistent set of guidelines as to how resource access statistics should be reported.

The Coordinating Node log aggregation post processing has been recently updated to provide COUNTER compliant metrics by applying the following filters to ingested logs:

- Only allow HTTP status codes of 200 and 304 on READ requests. This ensures that redirects are only counted once and that incomplete or unsuccessful requests are ignored.
- Known robots (including DataONE services) are excluded. This ensures that automated harvesting or indexing systems do not artificially inflate the results.
- Repeat visits within a short time frame are excluded. This helps to ensure that accidental double-clicks or repeated requests from a client tool in a short time period are only counted once.

Both unfiltered and COUNTER compliant aggregated log records and summaries are available from the Coordinating Nodes. Details for accessing this information is available in the architecture documentation at: https://purl.dataone.org/architecture-dev/design/UsageStatistics.html. ■

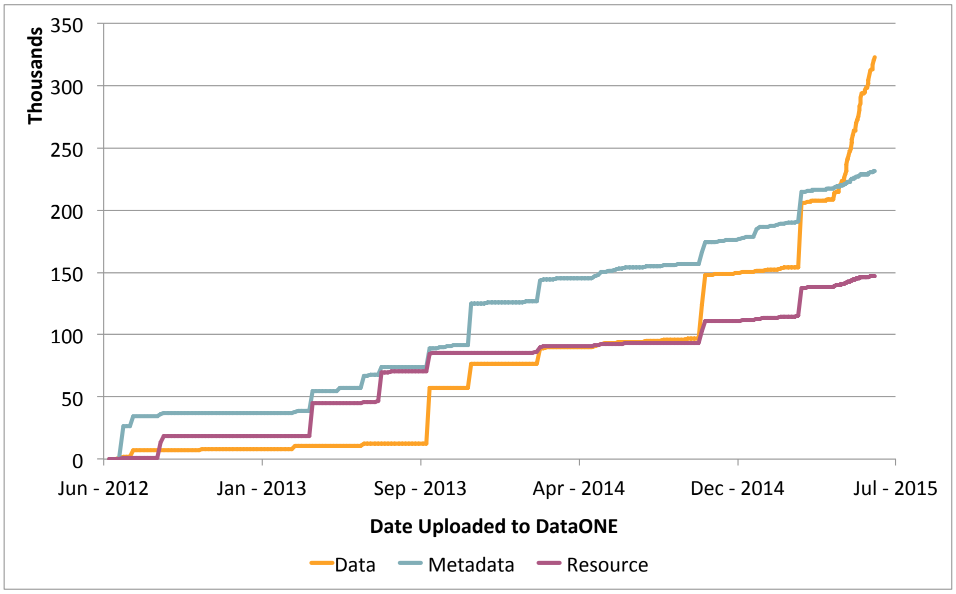


Figure 1: Counts of data/metadata/resource maps uploaded to DataONE since release in July 2012

MemberNodeDESCRIPTION

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.

NM EPSCoR: The New Mexico Experimental Program to Stimulate Competitive Research

http://www.nmepscor.org

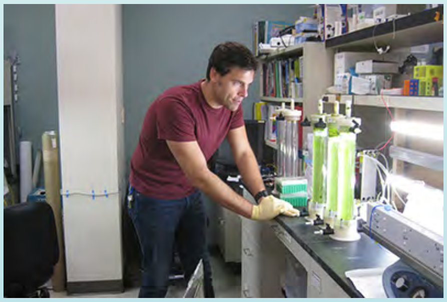
EPSCoR, the Experimental Program to Stimulate Competitive Research, is an NSF-funded program designed to *"assist the National Science Foundation in its statutory function 'to strengthen research and education in science and engineering throughout the United States"*.

EPSCoR goals are:

1. to provide strategic programs and opportunities for EPSCoR participants that stimulate sustainable improvements in their R&D capacity and competitiveness;
2. to advance science and engineering capabilities in EPSCoR jurisdictions for discovery, innovation and overall knowledge-based prosperity."

A primary objective of EPSCoR is *"to broaden participation in science and engineering by institutions, organizations and people within and among EPSCoR jurisdictions"* (generally, states). EPSCoR fosters well-organized programs that emphasize Science, Technology, Engineering and Mathematics (STEM) in the pursuit of activities that will enable jurisdictions to individually and collectively increase knowledge and both environmental and economic sustainability.

Over half of the United States' states and territories currently have an EPSCoR office, including New Mexico. New Mexico's EPSCoR office's "Energize New Mexico" effort focuses on three areas of interest: sustainable energy development, the cultivation of a well-qualified STEM workforce, and the development of a culture of innovation and entrepreneurship in these areas. Data collected as an output of NM EPSCoR research (both the current Energize New Mexico project and earlier work on understanding mountain streams and their relationship to climate, fire and geology) are available via their Data Portal and now discoverable through the DataONE Member Node.



Photos: Natalie Willoughby

New Mexico researchers focus on five energy areas: bioalgal fuels, hydrothermal energy, solar power, Uranium transport and remediation, and osmotic power development, as well as the integrated modeling of the social and natural science nexus. Each research topic brings together researchers from around the state. For instance, bioalgal energy research includes researchers at the University of New Mexico, New Mexico State University, Santa Fe Community College and Los Alamos National Laboratory who are developing and testing new strains of algae and encapsulation methods to improve biofuel production efficiencies as well as researchers at Eastern New Mexico University who are using innovative bioalgal "farming" methods to treat wastewater effluents from agriculture and dairy production.

DataONE Network Expands into New Continents

Australia's Terrestrial Ecosystem Research Network (TERN) has recently joined DataONE as our first Member Node in Australia and the 29th overall to join the federation.

The Terrestrial Ecosystem Research Network (TERN) connects ecosystem scientists and enables them to collect, contribute, store, share and integrate data across disciplines. Collectively this increases the capacity of the Australian ecosystem science community to advance science and contribute to effective management and sustainable use of our ecosystems.

TERN is a network-of-networks, somewhat similar to the Long Term Ecological Research Network based in the United States, which also participates with DataONE as a long-term collaborator. The network provides a data catalogue (TERN Data Discovery Portal) of content held in domain portals by its data partners so the community may share and collaborate on important environmental initiatives.

Each of TERN's data partners serves a specific data community; coming together under the umbrella of TERN, the individual organizations' impact is enhanced by broader visibility via DataONE. By exposing content via its DataONE Member Node, TERN further expands its exposure of Australian ecosystem data to a world-wide audience.

- The AusCover Facility is a national expert network that provides remote sensing data time-series and satellite based biophysical map products for Australia. AusCover also provides associated field calibration and validation data at continental scales.
- AusPlots is a plot-based surveillance monitoring program, undertaking baseline assessments of ecosystems across the country. The aim of AusPlots is to establish and maintain a national network of plots that enables consistent ecological assessment and ongoing monitoring. The AusPlots network collects a range of field data for integration with other existing data sources and current knowledge.
- The Australian Centre for Ecological Analysis and Synthesis (ACEAS) is a virtual and physical Facility within TERN designed to link ecosystem scientists and environmental managers to improve our understanding and management of Australian ecosystems that operated from 2010 to 2014. ACEAS activities supported multi-disciplinary integration, synthesis and modelling of ecosystem data.
- The Australian Coastal Ecosystems Facility (ACEF) collects and distributes key coastal datasets for use in policy and management decisions about the protection and use of Australia's coastal assets, both marine and freshwater. It addresses data collection needs from fine scale to satellite collections of flora, fauna and biophysical properties of coastal ecosystems.
- The Australian SuperSite Network (ASN) is a national network of multidisciplinary ecosystem observatories. The ASN includes ten SuperSites that each represent a significant Australian biome. The network covers all States and Territories and spans a wide range of environmental conditions.

TERN
Terrestrial Ecosystem
Research Network

FeaturedRESOURCE

DataONE Webinar Series

Contribute to the DataONE webinar series development for Fall 2015

During early 2015 DataONE ran a short series of four webinars featuring solo and panel presentations on a number of topics relevant to the open data / informatics community. We received extensive positive feedback for the series and are excited to launch the next set of nine webinars in Fall 2015. To do this, we want your help.

Do you have suggestions for a topic that would be great as a webinar? Do you know of speakers that have expertise in a field you would like covered? Would you like DataONE to facilitate open community discussions in addition to webinars? Take a few moments to answer these simple questions in survey format so that we can ensure the development of the 2015-2016 webinar series meets the needs of our diverse community (http://svy.mk/1CI0Gez).

The DataONE Spring Webinar Series included presentations on the following, all of which case be viewed at https://www.dataone.org/previous-webinars:

- Open Data and Science: Towards Optimizing the Research Process
- Boyle's Laws in a Networked World: How the future of science lies in understanding our past
- Make Data Count: Measuring Data Use and Reach
- Provenance and DataONE: Facilitating Reproducible Science

Over 460 individuals attended these webinars with many attending multiple presentations. We are excited to build on this success and develop the DataONE Webinar Series into a widely regarded program with your help.

Promoting an Open and Transparent Research Culture

Transparency and reproducibility are cornerstones of how science creates knowledge. Evidence for scientific claims should be shared openly so others can evaluate, question, replicate, or extend scientific studies. When evidence cannot be reproduced independently, then it should not be accepted as credible evidence. Despite their importance, transparency and reproducibility are not often rewarded. Lack of transparency reduces the credibility of published results, which in turn undercuts the efficient and effective use of funding to support scientific advancement. To improve research transparency, the scientific community is undertaking a series of reforms.

The Transparency and Openness Promotion (TOP) Committee recently published the TOP Guidelines[1], a set of author guidelines that journals can adopt to enhance the transparency of the research they publish. These guidelines represent a concrete and actionable strategy toward improving research and publishing practices. Already, 111 journals and 33 organizations (incl. DataONE) have expressed their support for the principles of openness, transparency, and reproducibility, and have committed to conducting a review within a year of the standards for potential adoption.

For more information on the Transparency and Openness Guidelines, and for the full press release got to: http://centerforopenscience.org/pr/2015-06-25/

[1] Nosek, BA et al., 2015, Promoting an open research culture. Science, 348: 6242, DOI: 10.1126/science.aab2374

- The Australian Transect Network (ATN) comprises seven major subcontinental transects that span biomes and traverse major rainfall, temperature and land-use gradients from the coast to inland areas.
- The Eco-informatics Facility has cyber-infrastructure experts working with governments, researchers, educators and students to make ecological "plot" data (including quadrats, transects, pitfall traplines, cage trap arrays, and other systematic collection methods) discoverable and freely accessible. A key product from the Eco-Informatics Facility is the Australian Ecological Knowledge and Observation System (ÆKOS), a 'one stop shop' for Australia's ecological plot data comprising a data repository, web data portal and web services enabling data users to store, discover, access and understand the context of the data in detail.
- The Ecosystem Modelling and Scaling Infrastructure Facility (e-MAST) is enhancing the capacity for assimilation and integration of data into modelling applications. By assembling data sets and developing software e-MAST enables testing, updating, and improvement of models used for such applications as future climate scenarios and assessment of primary production, at a wide range of temporal and spatial scales.
- The Long-Term Ecological Research Network (LTERN) facility integrates key established plot networks across Australia to tackle critical questions associated with the impacts of disturbance on Australian ecosystems. In a collaborative arrangement, LTERN brings together some of Australia's leading ecologists, from seven separate institutions. Formally established in 2012 and administered by the LTERN facility at The Australian National University, LTERN draws a range of existing long-term ecological monitoring programs together to establish a new coordinated and collaborative approach.
- The OzFlux Facility is a network of towers around Australia that continuously measures the exchanges (flux) of carbon dioxide, water vapour and energy between the terrestrial ecosystem and atmosphere. It is a national partnership with significant contributions from universities and research agencies around the country, coordinated by CSIRO Marine and Atmospheric Research based in Canberra and is also part of a global network of over 400 flux towers, most of which are located in the northern hemisphere.
- The Soil and Landscape Grid of Australia provides easy access to nationally-consistent and comprehensive soil and landscape attribute data at a finer resolution than ever before in Australia. The Grid is an essential piece of national information infrastructure, required to support informed use and management of our soil and landscapes. The Grid delivers a step-change in the way we can view and access the best available information about Australian soils and landscapes for a wide range of applications and users including urban and regional planners, land managers, farming groups, scientists and engineers.

TERN has built on existing data collection programs and established interconnected facilities spanning the Australian ecosystem science spectrum, from soils to satellites, biota to atmosphere, and data to knowledge synthesis. TERN delivers: (1) conventional "hard" data collection, storage and sharing infrastructure, e.g. instrumented towers, transects and plot networks for flora and fauna surveys, real-time environmental sensors of all kinds, and their resultant data streams; and (2) equally important "soft" infrastructure, e.g. nationally standardised methods, new ways of collecting, managing and discovering data, new multi-disciplinary collaborations and capacities for synthesis and policy translation.

"Being able to access TERN's data via DataONE's ONE*Mercury* platform enables TERN's datasets to increase the breadth of discoverability and provides a direct link to our Facility portals for data access," says Tim Clancy, TERN's Director.

"This partnership increases data sharing and opens up new opportunities for Australian ecosystem data to be used in new global science whilst maintaining data owners' rights to appropriate attribution". ■

To read the accompanying TERN news piece go to: http://tern.org.au/Newsletter-2015-June-DataONE-Member-Node-pg30794.html

Dear DUG Members

We are now approaching the last few days before the annual DUG Meeting, and Chris and I will soon be finishing in our roles as co-chairs. We've had a great experience serving in this capacity and have learned a lot more about DataONE and other parts of the community through the experience. We are looking forward to seeing many members at the meeting and transitioning over to a new leadership co-chair team.

In anticipation of the meeting, we'd like to give you a summary of what to expect. As is the usual tradition, we'll be starting with an introduction by the DUG co-chairs and Bill, Principal Investigator of DataONE, covering activities over the past year and where things are headed in the future. Afterwards, we'll hear from representatives of the Member Node network. That session should be particularly interesting as the network has grown significantly over the past year. In later sessions on Day 1 we'll have a series of breakouts to cover infrastructure log reporting, the investigator toolkit, education/outreach activities and Member Node services. We'll finish the day with the business meeting, which will entail voting on the next round of co-chairs and discussion of future directions for the DUG. On the evening of the first day, we'll have a reception/poster session, so if you'd like to present a poster about something you're working on related in any way to DataONE, please let Amber Budden know via aebudden@dataone.unm.edu.

Day 2 will involve the fan favorite community led roundtable discussions (topics still being finalized) in the morning, and a session in the afternoon on the NSF-funded Data Level Metrics project, led by the Public Library of Science, California Digital Library, and DataONE. That should be very interesting! To top-off the day and meeting, attendees will be invited to a tour of Asilomar, seeing the scenic and historic grounds of the conference location.

We look forward to seeing those who have already registered and hope that more will find time to attend. It's a great opportunity to hear what has been happening with DataONE and how you can benefit. ■

—Andrew Sallans
Chair, DataONE Users Group; Center for Open Science
—Chris Eaker
Vice-Chair, DataONE Users Group; University of Tennessee Library

UpcomingEVENTS

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. 12-13
DataONE Users Group Pacific Grove, CA
http://www.dataone.org/dataone-users-group

Jul. 13-17
Federation of Earth Science Information Partners Pacific Grove, CA
http://commons.esipfed.org/2015SummerMeeting

Jul. 26-30
XSEDE St Louis, MO
https://conferences.xsede.org/

Aug. 9-14
Ecological Society of America Baltimore, MD
http://esa.org/baltimore/

Sep. 15-17
Workshop: Managing Natural History Collections Data Tempe, AZ
http://bit.ly/1C3fqTw

Sep. 23-25
RDA 6th Plenary Meeting Paris, France
http://bit.ly/1NutPjP

Dec. 14-18
American Geological Union Annual Meeting San Francisco, CA
http://fallmeeting.agu.org/2015/

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