The State of Open Data 2017

The State of Open Data 2017 was released in October 2017 to much fanfare. The report includes numerous analyses and commentaries summarizing the current state of open data and changes since 2016.

Not surprisingly, the report noted many positive outcomes as well as several ingrained concerns from survey respondents. For instance, Jon Treadway and Briony Fane from Digital Science noted that “we can see strong signals that open data is becoming more embedded; the trends are positive.”

Notably, “respondents have become more aware of open data sets (82% up from 73%) than in 2016” and, somewhat surprisingly, there were no differences among age groups. Likewise, 80% of respondents across all age groups indicated a “willingness to reuse open data sets in their own research” and 60% of the researchers routinely share their data. Data sharing is accomplished equally through published appendices to articles and submitted data products to data repositories and, to a lesser extent, through data journal publications. On the positive side, researchers increasingly are aware of open data mandates (from funders, institutions and publishers) and licenses that cover the data they share. On the other hand, almost half of the respondents believed that “they own the data they produce before publication, with 40% believing they own it after publication.” It was also noted that most (64%) of the respondents were not very confident in citing secondary research data.

In a separate commentary in the report, Grace Baynes from Springer Nature noted that “50 funders now mandate or encourage data sharing, compared to 28 in 2015” but “as yet, only a few funders have requirements for data management plans or data availability statements, or explicitly make funds available for data management, storage and curation.” She recognized that researchers “need clear information, simple policies and advice” and argued that “concerted efforts by governments, funders, research institutions, publishers and researchers themselves are needed to make widespread open data a reality.”

Robert Kiley (Open Research) and David Carr (Wellcome Trust) described the new policy that the Wellcome Trust rolled out earlier this year with respect to data, software and materials management and sharing. Applicants for “funding are asked to provide an outputs management plan that sets out how they plan to manage and share significant data, software or materials to ensure the greatest benefit to health and research.” Importantly, the Wellcome Trust monitors plan implementation to ensure that outputs have been made available and, furthermore, “support[s] the development of resources to facilitate output sharing.”

Overall, the report continues to highlight what those of us involved in DataONE for almost a decade have long recognized—that is, open data sharing like any other positive, major sociocultural-research change takes time [often, generational in nature], requires concerted collaboration and significant resources and, once a tipping point is reached, can become a self-fulfilling prophecy as key stakeholders recognize that the scientific and societal benefits outweigh the associated costs. DataONE has and will continue to play a pivotal role in the open data movement by federating data repositories, replicating and protecting repository holdings, promoting easy discovery of well-documented data, and providing access to best practices and educational resources. As one example, I draw your attention to a landmark database on lake ecology and water quality for thousands of U.S. lakes that is now discoverable through DataONE and the Environmental Data Initiative (both funded by NSF) and published as a Data Note in GigaScience. The database covers the northeastern-most 17 U.S. states, was compiled by Patricia Soranno and 80 colleagues, and represents one of the largest and most significant databases of its kind, providing a foundational resource for increased understanding of lakes and freshwater ecology. Concerted efforts by researchers such as Soranno and colleagues, publishers like GigaScience, funders including Wellcome Trust and the U.S. National Science Foundation, repositories and data federations (e.g., Environmental Data Initiative and DataONE) are breaking new ground and leading the way to more open, reproducible, efficient and cost-effective science and decision-making.

—William Michener
Principal Investigator, DataONE

---


DataONE Brochures

We are excited to launch our new, audience specific, marketing materials available via the dataone.org website. These 2-sided brochures communicate the value of DataONE tools and services to four distinct community groups and highlight relevant accomplishments and activities. Designed for researchers, repositories, libraries and funders, the brochures contain a range of up-to-date project metrics and will be reviewed and renewed on a quarterly basis.

The materials are the result of a collaboration across team members from all areas of DataONE (Community Engagement, Sustainability and Governance, Usability and Assessment and Cyberinfrastructure) and Megan Mach, a DataONE summer intern working with the Community Engagement and Outreach Working Group. Already presented in hard copy at recent meetings we are pleased to make them available in PDF format for the community to download and circulate. Moving forward, we plan to create an aligned info-sheet providing an explicit focus on DataONE metrics in a set of easy-to-read, well designed visualizations.

To access these resources, please go to our public relations page at: https://www.dataone.org/public_relations.
How can the DataONE Users Group best serve the interests of Member Node operators, researchers, scientists, data managers, and librarians as full NSF funding for DataONE comes to an end in 2019? This is a critical question for the DataONE community as the entire DataONE federation moves toward self-sustainability over the next two years.

These topics were discussed at both the DataONE Users Group meeting held at the University of Indiana in Bloomington, IN on July 24-25, 2017 and again at the DataONE All Hands meeting held in late August. The DUG is working with DataONE leadership to define and implement changes to help the DUG transform itself as a driver of the DataONE cyberinfrastructure and community engagement agendas for 2019 and beyond.

Activities currently on the DUG agenda for 2018 include:

- Increase participation by Member Node operators in DUG activities
- Integrate the DUG with planning the DataONE Webinar Series (https://www.dataone.org/webinars)
- Integrate the DUG with the ongoing Member Node Forum (https://www.dataone.org/member-node-forum)
- Provide feedback for the DataONE Sustainability and Governance Working Group

Two new co-chairs were selected for two-year terms at the 2017 DUG Meeting to lead the DUG through this transitional period. Karl Benedict, Ph.D., is Associate Professor and Director Research Data Services in the College of University Libraries & Learning Sciences at the University of New Mexico. He also serves as the subject liaison for the departments of Geography and Environmental Studies, Earth and Planetary Sciences, and Biology. Robert J. Sandusky, Ph.D., is Associate Professor and Associate Dean for Information Technology in the University Library at the University of Illinois at Chicago.

Bob was a co-investigator on the original DataONE award and served on the Core Cyberinfrastructure Team and the Usability and Assessment Working Group. Since 2013 he has served on the DUG Steering Committee and has continued as a member of the U&A Working Group. He also is a charter member of the DUG and served as its first chair from 2010-2011.

Karl’s involvement with DataONE dates to December 2010 when he attended the inaugural DataONE User’s Group meeting in Chicago and has continued his engagement with the community through participation in a number of the DUG meetings held in conjunction with the Summer ESIP Federation meetings. While serving as Director of the Earth Data Analysis Center at the University of New Mexico Karl managed the development team that performed one of the first implementations of the DataONE Tier 1 Member Node API from the specifications developed by DataONE and has subsequently worked with the EDAC team in deploying Tier 4 Member Nodes based on DataONE’s software stack.

The DUG Steering Committee also welcomes new members Devan Donaldson (Indiana University), Gail Thornburg (OCLC) and Stephen Richard (Lamont-Doherty Earth Observatory). Amber Budden (DataONE), Dave Vieglais (DataONE), Debora Drucker (EMBRAPA, the Brazilian Agricultural Research Corporation), Rebecca Koskela (DataONE), Laura Movers (University of Tennessee, Knoxville), Shannon Rauch (Woods Hole Oceanographic Institute), Stephanie Simms (California Digital Library) and Trisha Cruse (DataCite) are continuing to serve on the Steering Committee.

The DUG Steering Committee welcomes expressions of interest from those who would like to participate on the committee. If you’d like to join the DUG Steering Committee, send a message to dugchairs@dataone.org indicating your interest.

— Robert Sandusky
Co-Chair, DataONE Users Group; University of Illinois, Chicago

— Karl Benedict
Co-Chair, DataONE Users Group; University of New Mexico
Status Update

The DataONE production environment has 43 Member Nodes participating in the production environment. Combined, these Member Nodes provide access to more than 185,000 publicly readable, current version data sets comprised of 363,000 metadata and 1,113,000 data objects. A total of 1,662,122 individual objects are resolvable and retrievable through DataONE and the participating Member Nodes.

Recent updates to the infrastructure have been focused on tuning and improving the throughput of the core Coordinating Node services, particularly under periods where bursts of activity on Member Nodes can induce an unusually high load on the Coordinating Nodes. This in turn was causing some latency in responsiveness. The recent updates address these issues and also help ensure overall system performance will not be impacted by the ever-increasing activity and volume of content available through the infrastructure. Other patches and updates have been included to ensure third party components and tools upon which the Coordinating nodes depend are up-to-date and include all the latest security and performance patches.

The DataONE search interface has seen numerous updates that progressively refine and improve the user experience. Updates include further optimization of queries and interactions with the Coordinating Nodes as well as many updates to rendering of search results including display of provenance information when available for data packages. For example, the dataset with persistent identifier “urn:uuid:3249ada0-afe3-4dd6-875e-0f7928a4c171” includes provenance information documenting the data sources, scripts, and products for its constituent components. This intuitive interface may be explored by visiting the DataONE search interface.

The Generic Member Node software has been updated to version 2.4.0, and includes numerous performance and other improvements. In particular version 2.4.0 now supports the data packaging API that returns BagIt packages when requested by users through the getPackage API. The upgrade process has also been streamlined, so deploying new versions of the software should be seamless. The GMN software and associated Python libraries are hosted from the DataONE GitHub d1-python repository at https://github.com/DataONEorg/d1_python.

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

Jan. 9-11
Federation of Earth Science Information Partners (ESIP)
Bethesda, Maryland
http://www.esipfed.org/meetings/upcoming-meetings/esip-winter-meeting-2018

Jan. 23-24
PIDapalooza
Girona, Spain
https://pidapalooza.org/

Feb. 15-19
AAAS Annual Meeting
Austin, TX
https://www.aaas.org/annual-meeting/future

Feb. 19-22
International Digital Curation Conference (IDCC)
Barcelona, Spain
http://www.dcc.ac.uk/events/idcc18

Mar. 21-23
Research Data Alliance 11th Plenary
Berlin, Germany
https://www.rd-alliance.org/plenaries/rda-eleventh-plenary-meeting-berlin-germany

Jul. 16-17
DataONE Users Group Meeting (DUG)
San Diego, CA
Details to be announced at: https://www.dataone.org/dataone-users-group

The DataONE federation. Metacat has also seen numerous updates and is currently at version 2.8.5. Metacat is written in Java, fully supports the DataONE APIs, and is available for download from https://knb.ecoinformatics.org/knb/docs/.

Figure 1: Counts of data/metadata/resource maps uploaded to DataONE since release in July 2012
For our Outreach Update we are highlighting Make Data Count, a collaborative project across DataONE, the California Digital Library and DataCite to measure and expose data level metrics. Below is a cross-posting of a blog article written by Daniella Lowenberg from the Make Data Count blog.

The Make Data Count (MDC) project is moving ahead with full force and the team wanted to take a moment to update the research stakeholder community on our project resources and roadmap.

In September, the MDC team sat down and mapped out the project plan for our two-year grant. Working in an agile method, we defined a “minimum viable product” (mvp) that comprises a full ecosystem of data usage and citation metrics flowing in and out of the technical hub and displayed on the DataONE repositories, Dash (California Digital Library Data Publishing Platform), and DataCite by summer of 2018.

This fall the MDC team also spent time traveling to several conferences to gather early adopters and gauge interest in data usage metrics. Many energetic and thoughtful discussions occurred regarding what the MDC-envisioned full ecosystem of data usage metrics will look like and how various stakeholders can contribute. The main takeaway: there is a need for a comprehensive and standardized way to count and display data level metrics.

So, what is MDC working on?
All of the MDC project work can be tracked on Github, and we encourage you to follow along.

• MDC and COUNTER are gathering community feedback from the COUNTER Code of Practice for Research Data Draft and turning this outline into a full narrative to be posted as a preprint in December.
• DataCite is working to build out a Data Level Metrics Hub that will ingest data citations and data usage metrics, use the COUNTER recommendation as a standard to log crunch, and push out standardized usage metrics for display on repository interfaces.
• Our first repositories, listed above, will be working to log process usage metrics against the COUNTER recommendation and technical hub for implementation.
• Designs for displayed data metrics on repository interfaces will be created and tested.
• Conversations with any groups that may want to be involved will continue- the more community feedback & support the better!

How can you help?
Everyone: We put out a COUNTER Code of Practice for Data Usage Draft and would appreciate community feedback. As stated above, this recommendation is what the usage metrics ecosystem will be standardized against. We also need help with mass outreach about our project, so please
In each newsletter issue we will highlight one of our current Member Nodes. The full list of Member Nodes and summary metrics can be found on the DataONE.org site at bit.ly/D1CMNs.

**The Forest Ecosystem Monitoring Cooperative**
https://www.uvm.edu/femc/

DataONE welcomed the Forest Ecosystem Monitoring Cooperative (FEMC) as a Member Node to the federation in October 2017.

The cooperative archives, synthesizes, and utilizes forest ecosystem information, as well as facilitating networks and providing tools to understand and monitor the health and management of forested ecosystems across the region. FEMC serves the northeast temperate forest region through improved understanding of long-term trends, annual conditions, and interdisciplinary relationships of the physical, chemical, and biological components of forested ecosystems. The cooperative also promotes the efficient coordination of multi-disciplinary environmental monitoring and research activities among federal, state, university, and private-sector agencies with common interests in the long-term health, management, and protection of forested ecosystems.

FEMC has developed many products and services over the years. In addition to a data archive, collections and tools, FEMC has created long-term monitoring reports which are updated annually, and also reports from partner projects with cooperators from New York, New Hampshire, Massachusetts, New Hampshire, and Maine. A number of other publications are also available, such as annual conference proceedings. FEMC also provides a wide variety of education and outreach programs to the surrounding community. More information can be found at https://www.uvm.edu/femc/.