Making Open Science a Reality

The past two weeks have been marked by two events that directly or indirectly focus on making open science a reality. First, the Organization for Economic Co-Operation and Development (OECD) released a seminal report entitled “Making Open Science a Reality”. The OECD consists of 34 countries including the United States, United Kingdom, Australia, Canada, France, Germany, Japan and New Zealand. The report is notable for at least three reasons. First, it eloquently summarizes the evidence for the benefits of open science (something I have touched upon in numerous prior newsletters). Second, it identifies the key actors in the open science space (e.g., researchers, government ministries, research funding agencies and private non-profit organizations and foundations, universities and public research institutes; libraries and repositories and data centers; and publishers and the business community) and explains their roles in promoting and enabling open science. Third, and most importantly, it details a number of key findings and policy messages to further open science that are highlighted below:

- **Open science is a means and not an end**. Open science strategies and policies are a means to support better quality science, increased collaboration, and engagement between research and society that can lead to higher social and economic impacts of public research.  
- **Open science is more than open access to publications or data**: it includes many aspects and stages of research processes.  
- **Policies to promote open data are less mature than those to promote open access to scientific publications.**  
- **Open science policies should be principle-based but adapted to local realities.**  
- **Better incentive mechanisms to promote data-sharing practices among researchers are needed.**  
- **Data-related skill development is essential.**  
- **Training of and awareness-raising among researchers is important for the development of an open science culture.**  
- **Repositories and online platforms will not have impact if the information they contain is not of good quality.**  
- **The long-term preservation costs of openly available research output need to be considered.**  
- **Clear legal frameworks for the sharing of publications and reuse of data sets are needed at the national and international levels.**  
- **Consultative approaches that involve all relevant actors for open science are a key component of successful open science strategies.**  
- **International collaboration in the area of open science is necessary to address global challenges.**  
- **Policy makers need to promote openness in science while at the same time preserving competition.”**

It is reassuring to note that DataONE and its affiliated data repositories are making great strides in promoting open science by building infrastructure for discovery, use and long term preservation of data as well as solutions for improving data quality and metadata, and training the current and next generation of researchers. The second notable recent event was the Research Data Alliance (RDA) Sixth Plenary Meeting and e-Infrastructures for Data Intensive Science Workshop that was held September 22-25, 2015 in Paris, France. DataONE was well represented at both meetings and played a key role in leading Working Groups and Interest Groups focused on metadata and brokering, and participating in other Groups focused on semantics, provenance and libraries. Over 500 individuals attended the meetings and participated in more than 50 Working and Interest Group activities. RDA is open to the broad community and if you have an interest in participating or tracking progress, I encourage you to visit https://rd-alliance.org.

In addition to the two events noted above, I also want to bring two other publications to your attention. First, the Journal of Librarianship and Scholarly Communication issued Volume 3, Issue 22 which includes numerous papers relevant to open science including institutional data management and data sharing practices, institutional data policies, and training programs and data workshops. Data scientists and librarians should find several articles worthy of their attention. Second, Carol Tenopir, Suzie Allard, Mike Frame and colleagues published a valuable article looking at changes in perceptions and practices related to data sharing that occurred from 2009/10 to 2013/143. In particular, they noted many modest improvements in data sharing attitudes and practices that varied geographically and across age groups. They further emphasized the need for “organizations such as DataONE [that] will continue to assess, monitor, educate, and provide the infrastructure necessary to support such complex grand science challenges.”

In the next newsletter issue, I will switch gears and look at how one can create an effective data management plan. In the meantime, please visit the RDA website and peruse the articles cited below for more information on data sharing and open science.

---

Bill Michener  
Principal Investigator

---

1. OECD. 2015. Making Open Science a Reality. https://www.innovationpolicyplatform.org/content/open-science
Managing Natural History Collections Data for Global Discoverability

DataONE had the pleasure of joining thirty-one folks from all over the USA for 3 biodiversity-informatics-information-filled days at Arizona State University (ASU) for the iDigBio Workshop: Managing NHC Data for Global Discoverability.

Nico Franz, Director of the ASU Biodiversity Knowledge Integration Center (BioKIC) at ASU, supplied extra funding that made it possible for us to have a 3-day (instead of 2-day) workshop. All the participants will tell you they needed 3 days and we packed a lot of information into the time allotted.

Our audience.

Some people in the collections community wear multiple hats, so these numbers don’t add up to 31, but when asked to self-identify as 1) researcher, 2) collection manager, 3) data manager, we had 16, 26, 20, respectively.

Highlights.

Fourth in-a-series of biodiversity informatics workshops, the material covered combined digitization specifics (how to choose collection management software), with biodiversity informatics skills and knowledge needed to give collections data a longer life through the robust use of appropriate standards and sharing strategies.

For example, participants learned about the world of data re/use and data management planning from Amber Budden from DataONE. One of her most compelling slides shows the need for the kind of information presented in this workshop (see below).

If you use a spreadsheet to manage data and sometimes find yourself wishing you knew a bit more about how to manipulate data in a spreadsheet, you’ll want to check out one of the course sections on better spreadsheet skills: http://idigbio.github.io/spreadsheet-skills/00-intro.html You can walk through the same course materials as the participants.

Our workshop participants enjoyed two remote presentations. Anne Thessen (http://datadetektiv.com), presented Data Publishing: in the context of the data life cycle. David Bloom (from VertNet.org) demystified the timely and sometimes confusing topics of Copyright and Intellectual Property Rights (IPR) and encouraged all of us to support the adoption of community norms.

Some other compelling topics covered in this workshop included:

- more on data cleaning and potential tools to make this task easier and more efficient
- a hands-on experience with the GBIF IPT tool that makes NHC data-sharing easier
- how to map data to a standard
- best practices for data entry (from DataONE) and
- iDigBio Data Management and Recordset Data Quality

You can find out more, and see the materials for yourself (and hear the recordings) on the workshop wiki.

Thoughts.

We are eagerly awaiting the post-workshop survey results and thinking about future workshops, webinars, and materials. From workshop conversations, several participants shared that this was “just what they needed, just when they needed it” which of course, is wonderful to hear. The workshop team is pleased to know that the materials and skills being taught seem to be filling a need in the community - and that these skills, and knowledge in turn, support better digitization. That is, digitization and in-turn, better data, are supported by enhanced community knowledge of data standards, sharing protocols, data standardization tools, and increased knowledge of where to find help. We also believe that this increased knowledge supports better communication in the future in collaborative efforts between Information Technology staff, researchers, collection and data managers, and students.

— Deb Paul
iDigBio Technology Specialist
(reprinted from the iDigBio blog with permission)

This workshop is the fourth in a series of four biodiversity informatics workshops, collaboratively developed by the broader scientific and research community. You can check out the first three from the following links:

Data Carpentry at iDigBio
Data Sharing, Data Standards, and Demystifying the IPT
Field to Database
Replacing content with a new version is a common case, and within DataONE the process is described as obsoleting an object. This operation may be performed by a user that has write access to the object, and is performed using the update method on the Member Node. During the update process, DataONE records that the old object is being “obsoleted by” a new object, and that the new object “obsoletes” the previous version. In this situation, both the old and the new objects remain resolvable and discoverable in DataONE, however the new object must also have a new persistent identifier. In order to retrieve the most recent version of the object, the user must examine the obsolescence information, and traverse the obsoleted by links to locate the newest version of the object. In version 2.0 of the DataONE infrastructure, a “Series Identifier” (SID) may be assigned to a series of object revisions. A SID will resolve to the latest version of an object, making retrieval of the latest version of an object much simpler while also retaining support for retrieving older versions of an object using the PID. In some cases it is desirable to prevent further discovery of an object through the search interfaces. In DataONE, this is achieved through the archive method which is available on both Member and Coordinating Nodes, and may be invoked by a user with write permission on the object. Setting the archive flag on an object is an irreversible process and indicates to DataONE that the object should be removed from the search indexes. An object that has been archived is no longer discoverable, though remains resolvable and retrievable using its persistent identifier.

Deleting content from DataONE is an unusual process that completely removes an object from DataONE, and is generally only performed in circumstances such as in response to legal requirements or presence of inappropriate content. Deletion can only be performed by DataONE administrators. After an object has been deleted, it is no longer discoverable, resolvable, or retrievable. The infrastructure will instead return “Not Found” errors indicating the object is no longer available. However, identifiers of deleted content remain in the system to ensure that these are never reused for different content.

Through these three types of operation, the long term integrity and availability of resources made available through DataONE Member Nodes is assured. Content creators have the facilities to ensure that the best quality information is available to consumers and investigators can continue to reliably retrieve information cited in publications or other studies.

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

<table>
<thead>
<tr>
<th>Date Uploaded to DataONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>150</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>250</td>
</tr>
<tr>
<td>300</td>
</tr>
<tr>
<td>350</td>
</tr>
<tr>
<td>400</td>
</tr>
</tbody>
</table>

Spotlight: Obsolete, Archive and Delete

An important goal of the DataONE infrastructure is to ensure long term access to the content provided by participating Member Nodes. At the most atomic level, every object that has been assigned a persistent, unique identifier (PID) is resolvable to one or more URLs from which the object can be retrieved. Content in DataONE is also indexed and made discoverable through the query service, and at a higher level through user interfaces such as ONE Mercury.

Practical realities of content management dictate that any particular object may be replaced with a newer version (obsoleting older content), the content owner may desire to reduce the discoverability of a data set (archiving content), or in rare cases, circumstances may require that some content is no longer downloadable (deleting content).
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.

UC3 Merritt and ONEShare


UC3 Merritt

The California Digital Library (CDL) has been a long-time partner with DataONE, involved from the beginning on various facets of the project, including participation on the leadership team and the cyberinfrastructure, preservation and metadata, and sustainability working groups, and the collaborative development of the Data Management Planning Tool (DMPTool). The University of California Curation Center (UC3) at the CDL also manages two DataONE Member Nodes, one provided as a service component of the Merritt repository, and the other being the ONEShare repository that sits behind UC3’s Dash open research data curation portal. Both Merritt and ONEShare are Tier 1 Member Nodes.

Merritt

Merritt is a general purpose curation repository service supporting both preservation and access functions for the entire UC community. Community members from all 10 UC campuses use Merritt to manage, preserve, and share their valuable digital content. Merritt currently hosts over 1.8 million objects totaling 37.8 TB representing all forms and genres of scholarly resources. The Merritt Member Node manages a small but growing subset of these resources falling under the environmental sciences domain. The Member Node provides access to over 31,000 datasets submitted by UC researchers, most contributed by the UC Davis Information Center for the Environment (ICE), as part of its California Drinking Water Source Assessment Program. These datasets are discoverable through DataONE’s ONEMercury search tool as well as Merritt’s native search interface. The ability to discover this content in aggregated form via the DataONE network enhances the global visibility of UC researchers’ efforts, facilitates the discovery of complementary datasets managed in other repositories, and enables collaboration with colleagues and other institutions.

Merritt currently exposes approximately 95,000 data files and 31,000 metadata files through ONEMercury.

ONEShare/Dash

ONEShare is an open data repository for researchers who do not have access to local institutional or disciplinary repositories, or whose data do not meet prescriptive eligibility requirements. Technically, ONEShare is a public collection in the Merritt repository integrated with its own DataONE Member Node that is co-sponsored by and hosted at the University of New Mexico. While data can be submitted to and accessed from ONEShare through Merritt’s native interface, the primary mode of interaction with ONEShare is through the DataONE Dash research data curation portal. Since ONEShare is also a DataONE Member Node, all ONEShare datasets are also accessible through ONEMercury. For information on DataONE Dash, see our “Featured Resource” section.
Dear DUG Members

Following a successful 2015 DUG Meeting in Pacific Grove, CA this summer, it is my pleasure to introduce to you our new co-chairs; Felimon Gayanilo and Plato Smith.

Felimon and Plato will serve as co-chairs for a two year term, bringing together the DataONE user community and project leads for discussion of planned future developments and opportunities.

The DUG also welcomed three new members to the DUG Steering Committee; Stephanie Simms, Steve Aulenbach and Shannon Rauch. If you are interested in volunteering on the DUG Steering Committee, or have other comments or suggestions, please contact us at dugchairs@dataone.org.

Felimon Gayanilo is the System Architect for the Gulf of Mexico Coastal Ocean Observing System (GCOOS; data.gcoos.org) Data Portal that aggregates near real-time data from over 1,900 sensors in the Gulf of Mexico designed to provide timely information about the environment of the United States portion of the Gulf of Mexico and its estuaries for use by decision-makers, including researchers, government managers, industry, the military, educators, emergency responders, and the general public. Mr. Gayanilo is also the System Architect that initiated the design and development activities of the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIC; https://data.gulfresearchinitiative.org/) that is responsible in aggregating and managing scientific data from over 220 scientific research project and over 2,700 researchers that investigates the impacts of dispersed oil and dispersants in the ecosystem, and affected coastal States in a broad context of improving fundamental understanding of the dynamics of such events and their environmental stresses and public health implication. Most recently, he is also involved as a co-PI to the NSF/EarthCube project, ‘Cross-Domain Observational Metadata Environmental Sensing Network [X-DOMES].’ Mr. Gayanilo has over three decades of experience in information and computer sciences in international, regional and national settings.

Plato Smith is in the second year of a two-year CLIR/DLF Postdoctoral Fellowship within the Research Data Services [RDS] program in the UNM Libraries. He co-authored UNM Libraries’ ACRL Assessment in Action [AIA] 2015 proposal and is working very closely with Karl Benedict in defining the project objectives and charting the course of the project throughout its lifecycle. He participates in delivery of the research data management training for graduate students and is active in the DLF eResearch Network and EarthCube Technology and Architecture Gap Analysis Working Group. He has experience in academic research libraries, digital libraries, and data management.

DataONE Dash

Dash is a University of California service that has been adapted to connect to the DataONE ONESHare Member Node. DataONE Dash is a free web based application that allows researchers to easily describe, deposit and share their research data publicly. Any dataset type can be deposited and users are able to log in with their Google ID.

DataONE Dash has been optimized for use by individual researchers and provides simple, intuitive interfaces for self-service curation of research data. Open to all researchers and research data, DataONE Dash supports users with the following important curation functions:

- Prepare data for curation by reviewing best practice guidance for the creation or acquisition of digital research data.
- Select datasets for curation through local file browse or drag-and-drop operation.
- Describe data in terms of the DataCite metadata schema.
- Identify datasets with persistent DataCite Digital Object Identifiers (DOI) for permanent citation and discovery.
- Preserve, manage, publish, and share data by uploading to the ONESHare repository.
- Discover and retrieve data through faceted search and browse.

UC3 operates six UC campus-branded Dash portals with object hosting in Merritt, in addition to DataONE Dash with data hosting in ONESHare. DataONE Dash is a collaborative project between DataONE and the UC Curation Center (UC3) at the California Digital Library.
Outreach UPDATE

The summer has been busy with the conference and workshop season. As always, the Ecological Society of America meeting was a great opportunity for us to connect with end users of environmental data. We held a couple of great workshops and special sessions, and were involved in multiple ignite talks. These were all well attended and the insightful questions and quality discussion continued at the DataONE booth in the exhibit hall. DataONE also had the opportunity to meet with data managers from the Natural History Collections community as part of an iDigBio funded workshop in Arizona this month. It was a great opportunity to share stories, identify common challenges and learn from each others’ experiences and a summary of the 3-day workshop can be found on page 2 of this issue.

With Fall comes the launch of our 2015-2016 webinar series. Our opening presentation, by Dr Margaret Leinen, was held in September and covered the topic of data stewardship in the culture of discovery. If you weren’t able to make it, don’t worry. All our webinars are recorded and posted online for later viewing. You can catch Dr Leinen’s webinar here: https://www.dataone.org/previous-webinars. While you’re there, take a look at some of the webinars from the last series and register for our upcoming presentation on the Open Science Framework. Registration is free and we enjoy the questions posed by our community. Also, suggestions for future webinar topics and speakers are always welcomed so don’t hesitate to get in touch.

Next DataONE Webinar:
Tuesday October 13th

The Open Science Framework: Increasing Reproducibility Across the Entire Research Lifecycle

Join us at 12 noon Eastern Time. Free to attend, register now at: https://www.dataone.org/webinars

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.

Nov. 15-20
Supercomputing Austin, TX
http://sc15.supercomputing.org/

Dec. 14-15
Coalition for Networked Information Membership Meeting Washington DC
https://www.cni.org/events/membership-meetings/upcoming-meeting/fall-2015/

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

Jan. 6-8
Federation of Earth Science Information Partners Washington, DC
http://commons.esipfed.org/2016WinterMeeting

Mar. 20-23
iConference Philadelphia, PA
http://ischools.org/the-iconference/about-the-iconference/iconference-preview/