The Long and Winding Road ... to Open Science

This space in the DataONE newsletter has frequently focused on the benefits of more open and transparent science as well as some of the tools that enable researchers to arrive at that destination. Most of us recognize, and the political headlines remind us everyday, that saying we are open and transparent “don’t necessarily make it so.”

So, what is openness in the scientific sense? Stephanie Hampton and colleagues have, perhaps, published the clearest definition of open science as “the concept of transparency at all stages of the research process, coupled with free and open access to data, code, and papers”. Along this same vein, publishers and research funders have recently developed definitions and guides of openness as applied to journal publications and research outputs. Importantly, both guides highlight that there is a continuum that exists between fully open and fully closed with respect to such characteristics as article access, data and code access, reuse, costs, compliance and machine readability. Many of us have watched and even taken part in the evolution of journal, professional society and research funder attempts to define and implement appropriate openness standards, although published accounts of the successes, failures, and behind-the-scenes workings are rare.

Individuals, groups and institutions learn from watching others and, presumably, emulating successful efforts that are “told in stories”. Although many publications highlight tools and best practices for promoting open science, few stories have been told that dive into the process of opening up science. A notable exception is the recent publication in Nature Ecology & Evolution by Julia Stewart Lowndes and her colleagues who document the transition of their team’s work on the Ocean Health Index project to a more open, transparent and reproducible process. The article starts with a brief history lesson that begins in 2012 when “we thought we were doing reproducible science” which entailed meticulously documenting data management steps and models in spreadsheets, emails and documents—all at great expense in time and efficiency. The article then proceeds from the recognition that this level of manual labor was unsustainable to “actually doing reproducible science.” Notably, the authors fully detail the transition to openness, documenting past practices and the adoption of new tools and practices (e.g., R, RStudio, Git, GitHub), and showing the time-line when changes were made. Furthermore, they illustrated how each change in their workflow transition improved the ease of reproducibility and collaboration.

Lowndes and her colleagues offer several key lessons from my perspective. First, the transition to open and reproducible science is not an overnight venture; it takes time and effort, especially to learn new tools and methods, but the openness process can successfully occur incrementally over several years as in their case. Second, there are many tools, training materials and intentional learning strategies which they describe and reference that can greatly expedite the process. Third, collaboration is key; the team necessarily had to be flexible in experimenting with new principles, tools, and skills, and in changing the ways they communicated and worked internally. Last, the authors conclude that “the benefits of openness can be a by-product of time-saving efficiencies, because tools that reduce data headaches also result in science that is more transparent, reproducible, collaborative and freely accessible to others.”

For anyone interested in both improving the efficiency of their research and promoting transparency and reproducibility, I heartily recommend reading the Lowndes et al. paper as well as reviewing the papers and learning resources cited therein. If we are to truly solve science and society’s greatest challenges by working within and at the boundaries of disciplines, then we collectively need to learn from ground-breaking efforts (and stories) such as those exemplified in the Ocean Health Index project. The path to more open science need not be quite so long and winding.

—William Michener
Principal Investigator, DataONE


Status Update

1,000,000

July 2017 marks the addition of the one MILLIONTH data object to DataONE. We projected hitting this target in advance of our annual DataONE Users Group (DUG) meeting and were eager to onboard our next Member Node to bring the number of data objects to 1,000,000. However, it was the continued growth of data within existing Member Nodes and community commitment to open and accessible data that has enabled us to reach this milestone. This is a significant accomplishment for DataONE and for all Member Node partners that contribute to the network. Congratulations all!

If you are interested in joining DataONE as a Member Node, read the Working Group Focus on “Becoming a Member Node”.

Working Group Focus

Becoming a Member Node

What is a Member Node?

A Member Node is any preservation-oriented repository that exposes its data products through the DataONE service specification, or Member Node API (purl.dataone.org/architecture). DataONE’s architecture and software products provide the flexibility to allow practically any organization wishing to maintain a Member Node repository to do so.

There are currently 38 Member Nodes participating in the DataONE federation, providing access to more than 330,000 datasets with more than a million data resources (https://www.dataone.org/current-member-nodes).

There are several benefits for a repository joining the DataONE federation as a Member Node. Data collections exposed through the DataONE API are discoverable by a wider audience through the search interface (search.dataone.org). The common service interfaces eases cross-repository data access through the DataONE investigator toolkit. Published data products are easily citable, and with a wider exposure, the value of your organization’s work is increased by receiving credit through data citations in published literature. DataONE Coordinating Nodes facilitate replication of content across Member Nodes, and so helps ensure reliable access to resources in the face of disruptions. Scientists discovering your work can also lead to opportunities for collaboration in the future. By simply limiting access to project data to collaboration partners, DataONE members can share project data without the need for case-by-case data-sharing arrangements.

Requirements for Member Nodes

DataONE’s mission and funding orient it currently towards certain types of organizations, but this profile is not meant to be limiting. Characteristics of potential DataONE Member Node organizations include:

- Existing earth science domain repositories
- Existing Data Centers within the earth science domain
- Long-term projects and research activities wishing to preserve the their data products
- Organizations/Institutions establishing data infrastructures to manage scientific research data
- Organizations with computing and storage capabilities to augment the DataONE cyberinfrastructure
- Organizations/Institutions with expertise in eco-informatics research and applications

DataONE endeavors to maintain a flexible, broadly applicable approach to representing content in a consistent manner across repositories. The following technical requirements arise from the need to ensure long term access to resources, even beyond
Deploying a Member Node

It is recommended that new repositories make use of existing Member Node software such as Metacat (https://www.dataone.org/software-tools/metacat) or the Generic Member Node (GMN, https://www.dataone.org/software-tools/generic-member-node). These web server applications fully support the DataONE APIs and offer a straightforward installation path.

Existing repositories may participate in the DataONE federation by adding the necessary Member Node API implementations to their existing repository application or alternatively, by using an instance of GMN as a proxy to the repository. DataONE provides software libraries in Java and Python that can significantly reduce the implementation time for adapting an existing repository solution. Experience has shown however, that adapting an existing repository requires careful planning because of the extensive test iterations required.

A typically more streamlined approach for existing Member Nodes to participate in DataONE is to make use of the proxy capabilities of the GMN. In this deployment pattern, it is only necessary to create or modify an existing adaptor written in Python that ensures the GMN instance is able to access to resources of the existing repository. GMN will then periodically update an internal catalog of resources which is in turn enables synchronization with the DataONE Coordinating Nodes and replication to other Member Nodes when necessary. This pattern of leveraging the proxy capabilities of GMN for Member Node deployment has been termed a “Slender Node” deployment.

DataONE has currently deployed Slender Node adapters for standard repository services including the Open Geospatial Consortium Catalog Services for the Web and the Open Archives Initiative Protocol for Metadata Harvesting. Other protocols can be readily supported by developing an adapter that enables the GMN application to retrieve the contents from the existing repository and reliably detect when content has been added or changed.

One important consideration for a Slender Node deployment is where the instance of the GMN service should be deployed. Ideally, this should be close to the existing data repository so the instance may take advantage of typically more reliable and performant local networks and may be more readily incorporated into an existing administrative and maintenance schedule.

More resources on becoming a DataONE Member Node:

Benefits of becoming a Member Node

How to become a Member Node

Member Node Forum
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 Environmental Data Initiative
https://environmentaldatainitiative.org/

The Environmental Data Initiative, or EDI, is an National Science Foundation (NSF) project intended to accelerate the curation and archival of environmental data, with a particular focus on data from projects funded by the NSF Division of Environmental Biology (DEB). Programs included are the Long Term Research in Environmental Biology (LTREB), Organization for Biological Field Stations (OBFS), Macrosystems Biology (MSB), and Long Term Ecological Research (LTER). LTER as a mature program with a robust repository which has had a presence in DataONE since its inception. EDI leverages the expertise of the LTER community to improve data management for all its component partners. The current LTER DataONE Member Node will be managed by EDI moving forward in a mutually beneficial relationship.

EDI addresses many of the environmental community’s data management needs, as described in the NSF award announcement. “Global-scale environmental issues such as food security, the spread of disease, and the availability of clean water emphasize the importance of environmental data that can address specific problems while also providing predictions of future conditions. The increasing availability of large volumes of different kinds of data offers new opportunities to address these issues.”

EDI provides “the environmental research community with efficient and reliable means for data management, storage, and sharing. The facilities developed (by EDI) will allow researchers, policy makers, managers, and other stakeholders to bring relevant data to bear on complex environmental questions.” Data that are archived in a stable, accessible repository and that are accompanied by appropriate metadata benefit both data producers and consumers through improved discoverability and reliability. These data, derived from publicly funded research, are one of the most valuable products of EDI.

EDI also has a training component to “overcome technical and social barriers to collaboration, thereby enhancing infrastructure to address ecological questions over broad spatial and temporal scales.” Training activities are being developed that range from the basics of metadata creation to the adoption of standardized best practices for specific types of data. Workshop participants will be trained in developing workflow technology, re-using existing workflows, and archiving and sharing their developments. Through these workshops, together with community-level centers of expertise, and individual-based skill exchanges, EDI will increase the volume of data available along with data discoverability and reuse.

Data sets archived in the EDI Data Repository may be searched for via the EDI portal https://portal.edirepository.org/nis/home.jsp or through the DataONE catalog.

Outreach UPDATE

Summer is always a busy time for DataONE Community Engagement and Outreach. We host our annual DataONE Users Group Meeting (https://www.dataone.org/dug-2017-agendal), are engaged in preparing workshops, presentations and exhibitions for the Ecological Society of America (ESA) meeting and coordinate the DataONE Summer Internship program.

Through collaboration with the Arctic Data Center and The Whole Tale, this year DataONE was able to support six interns working on diverse and compelling projects ranging from semantic annotation of workflows to provenance using YesWorkflow, DataONE messaging and exploration of DataONE search logs. Progress of the projects can be followed in their online notebooks at: https://notebooks.dataone.org and full intern bios can be found at: https://www.dataone.org/2017_interns.

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 nine years the DataONE internship program has attracted highly qualified candidates from a broad range of disciplines and backgrounds. 2017 is no exception.

With the ESA only weeks away, we are busy preparing for our talks, workshops and exhibition booth. As always, the ESA represents a great opportunity to connect with many of you within the ecological research community and we encourage you to find us during sessions or in the exhibit hall at Booth 406. This year we are excited to be co-hosting a workshop with the Arctic Data Center as well as coordinating an ignite session on tips and techniques for data management and a special session on data management planning. Details of all DataONE related talks and sessions can be found at: https://www.dataone.org/training-activities.
Finally, if you have been following the DataONE webinar series you will know that we continue to provide webinars that engage participants in relevant and cutting-edge topics concerning data management within Earth and environmental sciences. Topics may be broad conceptual themes or more specific instructional webinars focussed on open science, stages of the data life cycle or community tools for data management. Although on break over the summer months, the DataONE webinar series will relaunch this September. We are always interested in soliciting topics and speakers so please don’t hesitate to reach out via @DataONEorg #DWS2017 or via webinars@dataone.org with suggestions.