DataONE for Librarians

Carly Strasser, Stephanie Wright, Gail Steinhart
DataONE Community Engagement and Education Working Group

Objective of This Document

To introduce DataONE to the library community, especially the tools and resources provided by DataONE that help support institutional data management needs.

What is DataONE?

DataONE is a project funded by the National Science Foundation that is focused on federating existing earth and environmental sciences data repositories. The infrastructure being built to perform this task is complemented by educational and outreach activities, which inform the community about data stewardship. To that end, the DataONE project has two main tasks:

- Build the cyberinfrastructure to link together existing data and facilitate the search, discovery and management of these data sets, and
- Build the community of stakeholders around data. This includes researchers, librarians, data managers, policy makers, citizen scientists, and others.

To link together existing data, DataONE is building up a network of existing data repositories. This network is made up of “Member Nodes” and “Coordinating Nodes”. A Member Node is any repository that exposes its data or services through the DataONE service specification. There are three Coordinating Nodes that provide network-wide services to enhance interoperability of the Member Nodes and to support indexing and replication services.

The DataONE organization involves individuals from many different communities including research, administration, computer science, libraries, software development, citizen science, and information science. Many of the tasks accomplished by DataONE are undertaken by one of the 11 working groups, and are informed by the DataONE Users Group. The working groups focus on identifying, describing, and implementing the DataONE cyber-infrastructure, governance, and sustainability models. Education is integral to DataONE and spans formal graduate-level training in research and cyber-infrastructure development, to developing informal inquiry-based education modules that allow students of all ages to ask their own specific questions.

What does data management have to do with libraries and librarians?

Libraries are custodians of the scholarly record. This role of custodian is logically extended to include research data as widely accessible data and cyberinfrastructure make possible new opportunities for scientific discovery, and as research funders increasingly require scientists to share the products of their research. Librarians are well positioned to work in this arena because they bring relevant skills and knowledge to the table, including expertise in information

management, metadata and discovery, digital preservation, and intellectual property concerns. Librarians often have well-established relationships with the researchers in their institutions, and many have discipline-specific expertise to contribute. Libraries, as potential providers of research data management services, have a stake in ensuring effective data management in order to make the best possible use of limited resources and to protect the institution's intellectual assets.

Data management education presents an important opportunity for librarians to engage with researchers and with DataONE. Librarians can encourage and assist their institutions in integrating data management best practices into introductory biology, ecology, and environmental science courses, and providing stand-alone graduate courses on data management.

DataONE's Librarian Outreach Toolkit provides libraries with the tools, resources, and expertise to fill a relevant and timely need for their communities. By participating in data management education, libraries and librarians can extend their historical mission of service and preservation for the academic community, and also promote best practices in data management to accomplish the aims of sharing, providing access to, and preserving data as effectively as possible.

DataONE Resources

Primer on Data Management

The DataONE Primer on Data Management highlights the basics of data management. It provides guidance for researchers on organizing, managing, and preserving their data. Links are provided to best practices and software tools for data management on the DataONE website (described below); these links point to more in-depth descriptions, examples and rationale. Although many of the best practices were created with tabular (i.e. spreadsheet) data in mind, many of the concepts are applicable to other types of data produced by scientists, including databases, images, gridded data, or shape files.

The Primer provides a guide to data management practices that investigators could perform during the course of data collection, processing, and analysis (i.e. components of the data life cycle, Fig. 1) to improve the chances of their data being used effectively by others. These practices could be performed at any time during the preparation of the data set, but we suggest that researchers consider them in the data management planning stage, before the first measurements are taken. In addition, sometimes steps of the life cycle (and data management in general) can and should occur simultaneously; for instance, describing your collection methods is easier during the collection phase, rather than trying to reconstruct methods later to add to your data documentation.

Best Practices Database

The DataONE Best Practices Database provides individuals with recommendations on how to effectively work with their data through all stages of the data life cycle. Users can access best practices within the database by either clicking on a stage of the life cycle, selecting keywords (under advanced search) or using free search.

Software Tools Catalog

The Software Tools Catalog provides a brief description of a wide range of tools that are recommended for use by researchers throughout the data life cycle. Tools entries also include information about level of difficulty, cost, and links to further resources. Users can access tools within the database by selecting keywords via advanced search, or by browsing.

Data Management Teaching Modules

The DataONE community has created a set of education modules in Microsoft PowerPoint format. These slide decks are appropriate for a wide range of groups (including students, scientists, librarians, and citizen scientists) and provide a broad overview of the various topics listed. You can download, modify, and uses since the modules are licensed under Creative Commons Zero (CC0), i.e., no rights reserved.

- Lesson 01: Why Data Management
- Lesson 02: Data Sharing
- Lesson 03: Data Management Planning
- Lesson 04: Data Entry and Manipulation
- Lesson 05: Data Quality Control and Assurance
- Lesson 06: Data Protection and Backups
- Lesson 07: Metadata
- Lesson 08: How to Write Good Quality Metadata
- Lesson 09: Data Citation
- Lesson 10: Analysis and Workflows

Researcher Data Management Needs Survey and Assessment Bibliography

The Researcher Data Management Needs Survey and Assessment Bibliography is a CiteULike compilation of surveys and assessments conducted to determine the data management needs of researchers, some of which were generated by DataONE activities. Suggestions for additions to the bibliography can be submitted directly via the CiteULike group.

Investigator Toolkit

The Investigator Toolkit is a collection of software tools for finding, using, and contributing data in DataONE. Some of these tools have been custom written for DataONE, some are existing tools that have been modified to use the DataONE Application Programming Interface (API), and some are tools that have well defined interfaces of their own which can be called by DataONE tools. The toolkit currently includes

- ONE*Mercury*: Web-based tool for searching data held by DataONE member nodes.
- DMPTool: Web application that helps researchers develop practical data management plans consistent with agency requirements and available resources.
- DataUp: Open-source tool that helps researchers in creating metadata, checking for best practices, obtaining a unique identifier for their data set, and depositing their data into a repository. The ONE*Share* repository is set up as a free, open data archive for DataUp users.
- ONE*R*: DataONE R Client provides the ability to access open ecological, environmental, and earth science data from the DataONE network of repositories, and to save data from within R to DataONE repositories that support write access.
- Morpho: is an open source metadata editor for Ecological Metadata Language (EML).

- ONE*Drive*: allows users and developers to access DataONE content like a remote file system (under development)

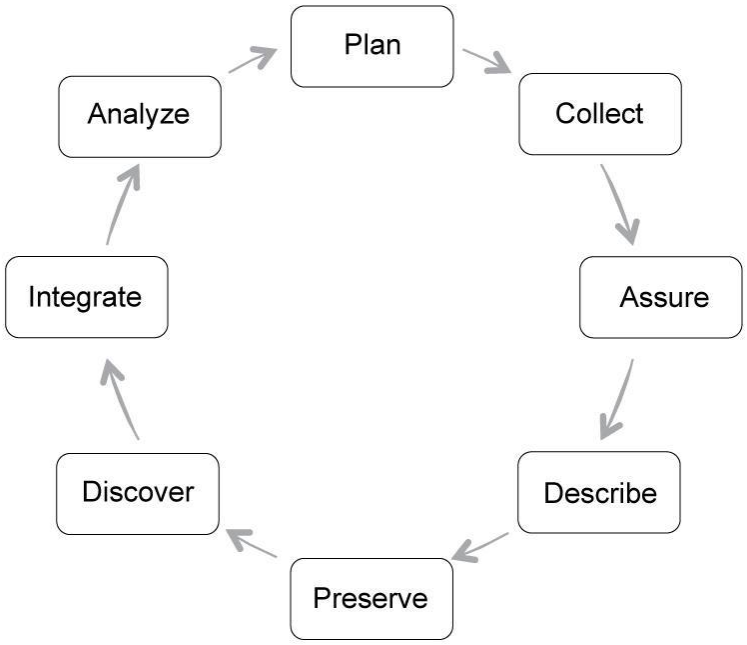

Figure 1: The data lifecycle.

The Data Life Cycle: An Overview

When discussing data management needs and the services librarians can provide to support them, it is helpful to think of them in terms of the data management life cycle. The data life cycle has eight components:

- **Plan**: description of the data that will be compiled, and how the data will be managed and made accessible throughout its lifetime
- **Collect**: observations are made either by hand or with sensors or other instruments and the data are placed a into digital form
- **Assure**: the quality of the data are assured through checks and inspections
- **Describe**: data are accurately and thoroughly described using the appropriate metadata standards)
- **Preserve**: data are submitted to an appropriate long-term archive (i.e. data center
- **Discover**: potentially useful data are located and obtained, along with the relevant information about the data (metadata)
- **Integrate**: data from disparate sources are combined to form one homogeneous set of data that can be readily analyzed
- **Analyze**: data are analyzed

Some projects might use only part of the life cycle and other projects might not follow the linear path depicted in the diagram, or multiple revolutions of the cycle might be necessary.

Librarians can assist researchers in their data management needs at various stages along the data life cycle, particularly as described below.

Data Management Throughout the Data Life Cycle

Plan
It is extremely valuable for librarians to engage researchers in thinking about data management issues as early as possible in the research planning process. Librarians can provide guidance for researchers in thinking about and planning for challenges they may encounter in each phase of the life cycle through data management planning consultations.

Many funders (particularly federal funding agencies) require researchers include a data management plan (DMP) in their grant proposal. Librarians can partner with their sponsored programs office to become involved in the data management plan review process and identify specific resources and services available for data management planning at their institution.

Resources
- Section 5.1 of the DataONE Best Practices Primer identifies several questions to help guide researchers in thinking about managing their data at the beginning of the research project.
- The DMPTool is an online resource for helping researchers through the development of a DMP with specific guidance for several funding agencies.

Describe
High quality metadata enables others to discover, understand, and use data, and description is a traditional area of expertise for librarians. Assisting researchers with data set description involves becoming familiar with general and discipline-specific metadata standards and tools.

Resources
- The Digital Curation Centre's (DCC) list of metadata standards
- Metadata editors from the DataONE Software Tools Catalog

Preserve
Researchers have multiple options for ensuring preservation of their data; librarians can help them understand the key characteristics of each: disciplinary repositories provide visibility within the relevant community of practice and support discipline-specific tools and standards. Publishers may accept data sets as supporting materials associated with articles, recommend deposit to a third-party repository, or support the publication of peer-reviewed data papers. Libraries themselves may choose to host data sets within an institutional repository or a purpose-built repository specifically for research data.

Resources
- Consult the list of current DataONE member node repositories that accept data (repository login required for data deposition in some cases)
- Repositories, including institutional repositories, may elect to become DataONE member nodes to promote broader discovery and access to content.
- DataBib and re3data are searchable catalogs of repositories.

Discover
A variety of tools are available to specifically aid in discovery of existing data sets. Researchers may need to find data to use to answer new questions, and also need to ensure that their own data are readily discovered by others. Librarians have well developed (often domain-specific) skills and expertise appropriate for both tasks.

Sound citation practice serves to facilitate discovery, as well as ensuring attribution and credit. Just as with traditional publications, researchers should, at a minimum, provide attribution when reusing existing data sets created by others. Librarians can point researchers to citation style guides to assist users in citing data sets in the correct format. In turn, researchers can make their data sets more easily citable by providing a permanent identifier for their data set.

Resources
- ONE*Mercury* is a web-based tool for searching data held by DataONE Member Nodes.
- DataCite Metadata Search is a search tool for data sets registered with DataCite.
- EZID: a subscription service provided by the California Digital Library that makes it easy to create and manage permanent identifiers.
- The DOI Citation Formatter provides citation formats for DataCite and CrossRef DOIs.

How to Participate in DataONE

Join the Users Group

The DataONE Users Group (DUG) is comprised of the stakeholders from the many communities of DataONE. The primary function of the DUG is to represent the needs and interests of these communities in the activities of the DataONE organization. In particular, the DUG provides guidance that facilitates DataONE in achieving its vision and mission. The DUG meets annually to identify the evolving technical challenges and opportunities that can be applied to advance education, research, and policy through the use of DataONE data products, tools, and services. Anyone can join the Users Group by filling out this brief form.

Contact your favorite repository about becoming a DataONE Member Node

Any repository can become a DataONE “member node”. See the “Preservation” section above for more information, and refer to the DataONE web page: Benefits of Becoming a Member Node.

Contribute to FAQ by asking questions on ask.dataone.org

Ask.dataone.org is a community forum for asking about DataONE products or services that are not answered on the website. By posting questions to the forum, librarians can help inform others with similar questions. Go to http://docs.dataone.org and click “register” in the upper right corner to get an account.

Acknowledgements
This work was supported by the National Science Foundation (grant numbers 0753138 and 0830944).