DataONE—Enabling Cyberinfrastructure for the Biological, Environmental and Earth Sciences

William K. Michener[1,2], Rebecca Koskela[1,2], Matthew B. Jones[2,3], Robert B. Cook[2,4], Mike Frame[2,5], Bruce Wilson[2,4], and David A. Vieglais[2,6]

[1]University Libraries, MSC04 2815, University of New Mexico, Albuquerque, NM 87131-0001
[2]DataONE, University of New Mexico, Albuquerque, NM 87131-0001
[3]NCEAS, University of California Santa Barbara, Santa Barbara, CA 93101
[4]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301
[5]U.S. Geological Survey, Core Science Systems, Biological Informatics Program, Building 1916T2, 230 Warehouse Road: Mailing Address: P. O. Box 6015, Oak Ridge, TN 37831
[6]Biodiversity Research Center, University of Kansas, Lawrence KS 66045-7593

Keywords

Data centers, Federated data systems, User-centered design, Analysis, Data integration, Data life cycle

Introduction

The scope and nature of biological, environmental and Earth science research are evolving in response to environmental challenges such as global climate change, invasive species, and emergent diseases. In particular, scientific studies are increasingly focusing on long-term, broad-scale, and complex questions that require massive amounts of diverse data collected by remote sensing platforms and embedded environmental sensor networks; collaborative, interdisciplinary science teams; and new approaches for managing, preserving, analyzing, and sharing data. Here, we describe the design of DataONE (Data Observation Network for Earth)—a cyberinfrastructure platform developed to support rapid data discovery and access across diverse data centers distributed worldwide and designed to provide scientists with an integrated set of familiar tools that support all elements of the data life cycle (e.g., from planning and acquisition through data integration, analysis, and visualization). Ongoing evolution of the DataONE architecture is based on participatory, user-centered design processes including: (1) identification and prioritization of stakeholder communities; (2) developing an understanding of their perceptions, attitudes, and user requirements; (3) usability analysis and assessment; and (4) engaging science teams in grand challenge science exemplars.

DataONE as an integrative platform for the biological, environmental, and earth sciences

DataONE is designed to provide an underlying infrastructure that facilitates data preservation and re-use for research with an initial focus on the biological, environmental, and earth sciences. DataONE is unique in that it: (1) builds on existing data centers, leveraging the global investment in scientific data preservation; (2) creates a global, federated data network by focusing on interoperability of systems, providing tools and services to enable data access and preservation across institutions using unique software systems; and (3) facilitates evolving communities of practice enabled by the DataONE cyberinfrastructure (CI) and informed by best

practices, exemplary data management plans, and tools that support all aspects of the data life cycle.

The cyberinfrastructure implemented by DataONE is composed of three principal components (Figure 1): *Member Nodes* which are existing or new data repositories that install the DataONE Member Node application programming interfaces (APIs); *Coordinating Nodes* that are responsible for cataloging content, managing replication of content, and providing search and discovery mechanisms; and an *Investigator Toolkit* which is a modular set of software and plug-ins that enables interaction with DataONE infrastructure through commonly used analysis and data management tools. Multiple instances of these three components operate together to provide a reliable fabric from where data may be retrieved by persistent identifiers, and contributions are guaranteed to be available indefinitely.

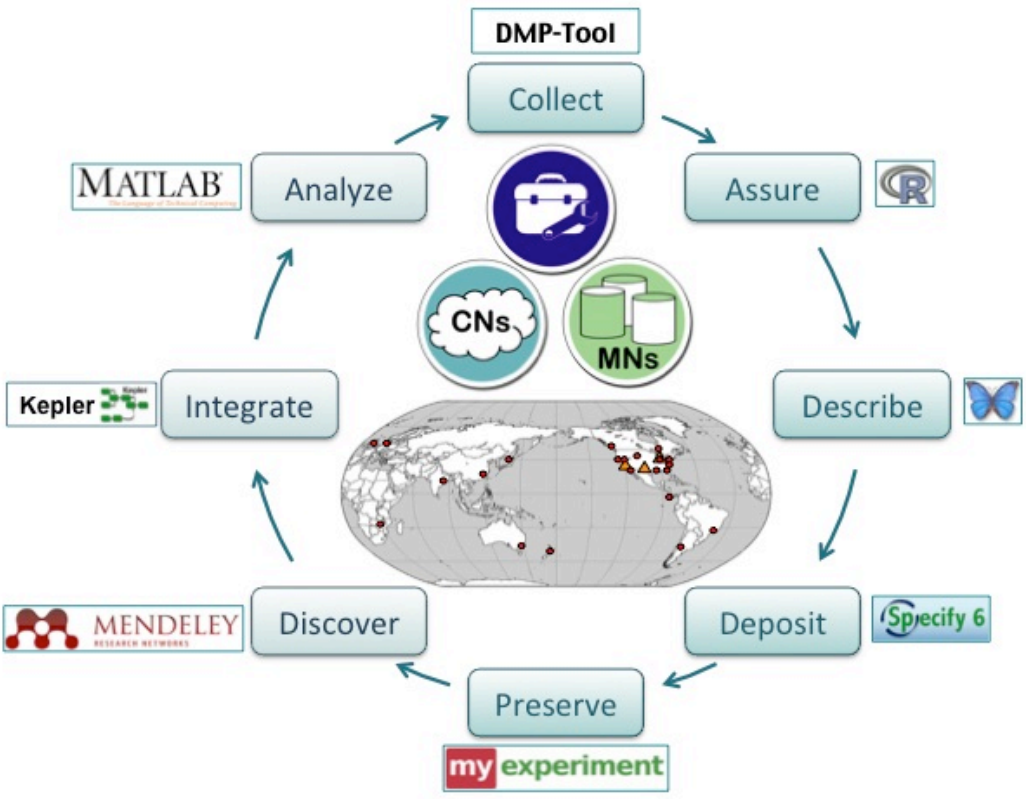

Figure 1. *Member Nodes* (red dots) are existing or new data repositories that install the DataONE Member Node application programming interfaces (APIs); *Coordinating Nodes* (orange triangles) are responsible for cataloging content, managing replication of content, and providing search and discovery mechanisms; and the *Investigator Toolkit* is a modular set of software and plug-ins that enables interaction with DataONE infrastructure through commonly used analysis and data management tools. Multiple instances of these three components operate together to provide a reliable fabric from where data may be retrieved by persistent identifiers, and contributions are guaranteed to be available indefinitely.

Data are principally acquired and maintained by Member Nodes that are located throughout the world. Member Nodes are envisioned to include a wide variety of institutions and organizations including natural history collections, Earth observing institutions, research projects and networks, libraries, universities, and governmental and non-governmental organizations. Each Member Node supports a specific constituency through its own implementation and often provides value-added support services (e.g., user help desk, visualization services). It is, therefore, expected that DataONE will accommodate highly geographically distributed and diverse Member Node implementations. Member Nodes extend the functionality of existing repository software systems by adding a standard set of web services that enable standardized communication between Member Nodes, Coordinating Nodes, and client tools. By implementing a common web service layer that covers the entire data lifecycle, Member Nodes are able to more effectively interact with one another (e.g., providing backup services), client tools can be written against a single service interface but still work with multiple Member Nodes, and a global index is created at Coordinating Nodes that allows for rapid discovery of the distributed data resources. Adding functionality to any service requires resources for implementation and ongoing maintenance, and so the web services are kept as lightweight as possible. Additional feedback from current and potential Member Node operators indicated a preference for a staged approach for implementing the DataONE services, and as a result the web service APIs are now grouped into four tiers with increasing services. Tier 1 APIs provide the minimum necessary functionality for a Member Node to participate in the DataONE network as a read-only Member Node with little authentication or access control for users. Member Nodes conforming to the Tier 4 APIs offer full support for DataONE services including replication and access control. While any repository software system can implement the DataONE service interfaces, we have identified commonly used repository software systems as high-priority for adaption to work with DataONE, including Metacat, Mercury, CUASHI HIS, Dryad, Merritt, Fedora, DSpace, iRODS, GeoNetwork, and others.

Exemplars of Member Nodes offering a wide range of data and existing functionality are summarized in Table 1. Member Nodes are selected based on evaluation criteria that include factors such as diversity of data holdings, readiness to participate, community leadership, and resource availability. These early participants in the DataONE federation cover different information domains of relevance to integrative biological, environmental, and Earth sciences research. Participation in DataONE simplifies user access to these valuable resources through a common set of service interfaces and helps to ensure long-term access to the information through the use of persistent unique identifiers for all data and metadata.

Coordinating Nodes are designed to be tightly coordinated, stable platforms providing network-wide services to Member Nodes. These services include network-wide indexing of digital objects, data replication services across Member Nodes, mirrored content of science metadata present at Member Nodes, management of access control rules, and mapping of identities among different identity providers. The three initial Coordinating Nodes are located at Oak Ridge Campus (a consortium comprised of Oak Ridge National Laboratory and the University of Tennessee), the University of California Santa Barbara, and the University of New Mexico. Coordinating Nodes maintain the integrity of the DataONE federation ensuring sufficient replicas are made of digital objects to facilitate long-term preservation, and tracking those replicas to enable resolution of persistent identifiers to Member Nodes where the content can be retrieved. The Coordinating Node indexing services provide a system-wide search mechanism

enabling users to discover relevant content from all participating Member Nodes. Figure 2 illustrates the values derived by Member Nodes from participation in the DataONE federation.

	ORNL DAAC	Dryad	KNB
Community	Agency repository	Journal consortium	Research network
Data	Earth sciences, ecology, and biogeochemical dynamics	Biosciences	Biodiversity, ecology, environment
Size	~ 1,000 data products, ~ 1 TB	~ 1,000 data products, ~ 5 GB	> 25,000 data products, 100s GBs
Services	Tools for data preservation, replication, discovery, access, sub-setting and visualization	Tools for data preservation, replication, discovery and access	Tools for data preservation, replication, discovery, access, management, and visualization
Metadata standards	FGDC subset	Dublin Core application profile	EML, FGDC
Degree of curation	High	Medium	Low
Data submission	Staff-assisted submission and curation of final data product	Web-based data submission at time of journal article submission	Self-submission via desktop tool at any time
Sponsor	NASA	NSF/JISC, societies, publishers	NSF

Table 1. Characteristics of three DataONE Member Nodes (ORNL DAAC, Oak Ridge National Laboratory Distributed Active Archive Center; KNB, Knowledge Network for Biocomplexity; FGDC, Federal Geographic Data Committee; EML, Ecological Metadata Language; NASA, National Aeronautics and Space Administration; NSF, National Science Foundation; JISC, Joint Information Systems Committee).

The primary purpose for developing the DataONE Investigator Toolkit is to adapt existing software tools commonly used and supported by the research community to provide seamless interaction with the DataONE cyberinfrastructure for storing, retrieving, discovering, and visualizing data. Components in the Investigator Toolkit include low-level libraries intended for developers and more technically inclined investigators, plugins for desktop applications like the R analytical environment (http://www.r-project.org/), and operating system extensions such as file system drivers that expose DataONE as essentially a large network drive.

Given the limited resources available for building and deploying the DataONE infrastructure, prioritization of development targets is critical. Early development efforts focus on providing well-documented and well-tested libraries in two widely used languages (Java (http://www.java.com/en/) and Python (http://www.python.org/)) to facilitate development of tools, plugins, and applications that reliably interact with the DataONE infrastructure. In addition, development of extensions for analytical tools representing different roles in the data life cycle is a high priority. Tools initially targeted include the Morpho metadata editor (http://knb.ecoinformatics.org/morphoportal.jsp), the Mercury metadata catalog and search system (http://mercury.ornl.gov/), citation managers such as Zotero (http://www.zotero.org/) and Mendeley (http://www.mendeley.com/), the R statistical programming environment,

(http://www.r-project.org/), workflow tools such as VisTrails (http://www.vistrails.org/index.php/Main_Page) and Kepler (https://kepler-project.org/), and general-purpose applications such as Microsoft Excel (http://office.microsoft.com/en-us/excel/). These applications were identified early on as being important through community surveys, and so are appropriately prioritized for development.

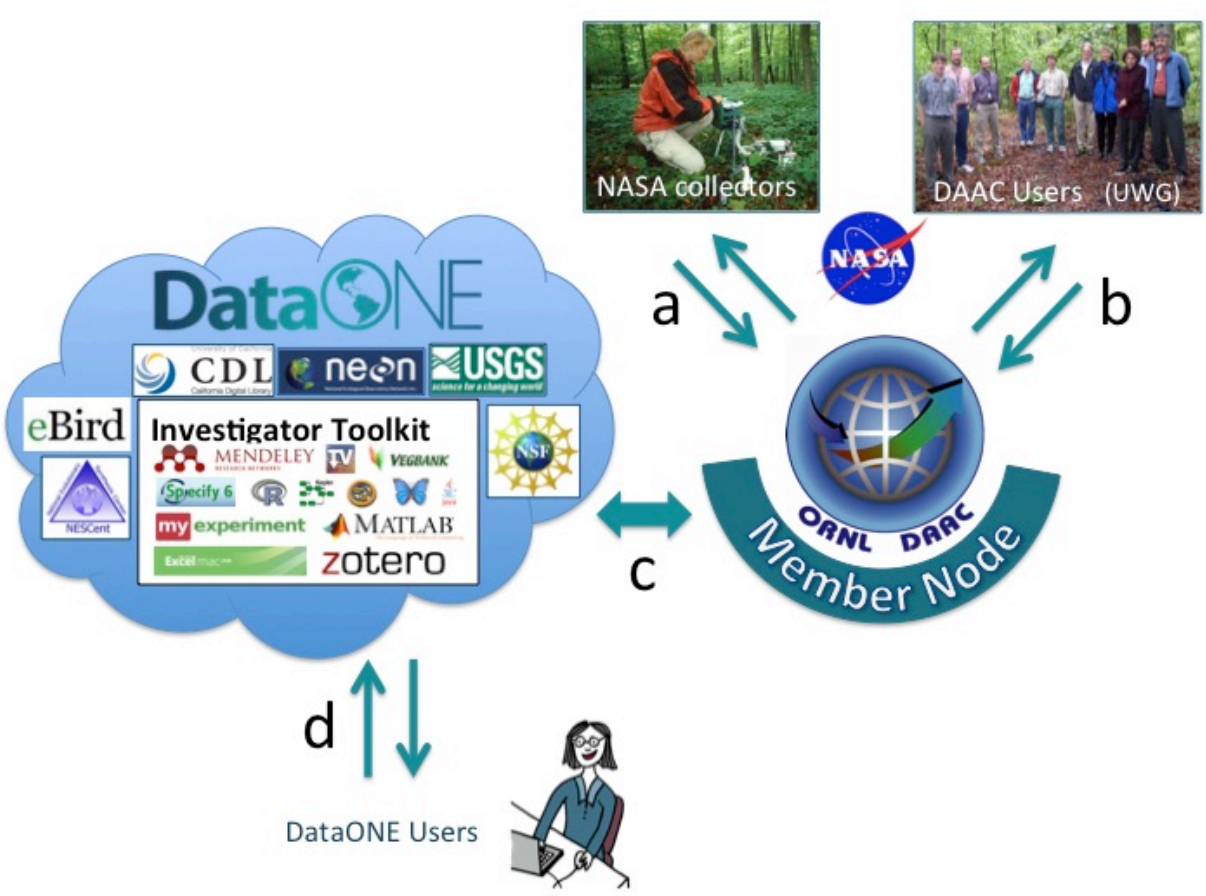

Figure 2. Overview of DataONE cyberinfrastructure elements illustrating value derived to users associated with a Member Node (i.e., the Oak Ridge National Laboratory Distributed Active Archive for Biogeochemical Dynamics; ORNL DAAC). NASA data collectors (a) deposit data into the ORNL DAAC, which is a Member Node of DataONE. Researchers (i.e., users) (b) can access these data directly from the ORNL DAAC. The crescent around the repository represents the software stack that enables the Member Node functionality for the repository. This software stack is developed and installed by DataONE staff, making use of the characteristics of the ORNL DAAC repository system and metadata. ORNL DAAC users can continue to access data as they did before (b). However, by linking up to the DataONE cyberinfrastructure (c), these data are accessible to a broader community of DataONE users (d) through DataONE enabled search and retrieval. The cyberinfrastructure components include Coordinating Nodes that maintain a metadata catalog of data held in the ORNL DAAC and other Member Nodes that support data replication and preservation services as well as integrated data search and retrieval, and that provide access to the Investigator Toolkit (which provides increased functionality through integrated software elements such as Mendeley, Zotero, VisTrails, Kepler, Excel, R, and others).

Concluding remarks

New cyberinfrastructure platforms are needed to resolve the numerous challenges faced by scientists. In particular, science progress is hindered to a large extent by the lack of interoperability of data and metadata with the result that many scientists spend a majority of their time integrating the diverse and heterogeneous data required to address today's scientific questions. Platforms like DataONE and new tools that reduce the amount of time that scientists focus on more mundane data transformation and translation activities are expected to significantly advance the nature and pace of science.

Many of the challenges are primarily technical in nature and can be resolved through continued investment of time, energy, and resources. A greater challenge, however, may lie in the socio-cultural realm—that is, in specifically identifying, understanding, and prioritizing the technical challenges to be overcome, as well as in educating scientists and others about the solutions. The cyberinfrastructure landscape is littered with unused hardware and software solutions that were created without understanding user requirements or by ignoring ease-of-use issues that are critical for tool adoption.

Because science is a human enterprise, it is critical that stakeholders be involved throughout the lifespan of community cyberinfrastructure development efforts. DataONE has used four different, complementary approaches (stakeholder assessment, personas and user scenarios, usability testing, and engagement with domain scientists in use cases) to generate different types of information that aid the development and implementation process. Use case scenarios have been especially effective in helping scientists imagine how their research can be enhanced through cyberinfrastructure and identifying gaps in existing infrastructure.

These results highlight the importance of engaging stakeholder communities early in the planning and building interoperable, international infrastructure to support science. Such efforts will require interdisciplinary teams that encompass both computer scientists and domain scientists (e.g., environmental scientists, library and information scientists, and social scientists). It will be especially critical to educate new generations of scientists in the use of cyberinfrastructure that will enable them to continue to expand the spatial, temporal, and disciplinary scales in which they work.