

DataONE News Summer 2011 - News Report No. 4 (7 July 2011)

William Michener, Amber Budden, and Rebecca Koskela – University of New Mexico; Dave Vieglais – University of Kansas

DataONE continues to make significant progress through the spring and summer of 2011 and is on track to release the first public version of its cyberinfrastructure at the end of 2011. In addition, significant effort has focused on developing educational resources, including a Data Management Planning Online Tool that will soon be released by a consortium of partners including the California Digital Library, University of California Los Angeles, University of California San Diego, Smithsonian Institution, University of Virginia Library, University of Illinois Urbana-Champaign, Digital Curation Centre, and DataONE. We have much to report, so this newsletter focuses on the progress related to infrastructure development. The next upcoming newsletter will feature results of the DataONE Users Group meeting and additional activities of the Community Engagement Working Groups.

Cyberinfrastructure Progress

The primary focus for the quarter has been on preparation for the public release of the DataONE infrastructure targeted for later in 2011. This release will include fully operational, stable Coordinating Nodes, Member Nodes (MN) and essential pieces of the Investigator Toolkit. Refactoring of core cyberinfrastructure following outcomes from prototyping is nearing completion and includes authentication and authorization services based on client side certificate-based authentication through the CILogon service and identity providers through InCommon. The authentication and authorization services represented a major uncertainty in the process for building out services suitable for public use, and so having this well defined and functional represents a significant milestone for the DataONE service infrastructure. Other software development activities have focused on addressing software bugs and performance aspects that were identified during the testing phase in early 2011. Some refactoring of the DataONE service interfaces has occurred to help with the process of third-party MN development and participation at different levels of functionality. Some specific activities are summarized below:

Member Nodes. Besides addition of the authentication and authorization interfaces and implementation, perhaps the most significant change to MN implementation has been the re-factoring of the MN Application Programming Interfaces (APIs) to better represent the functionality required for participation of MNs at different tiers of DataONE API conformance. Four tiers of MN functionality have been created:

1. Tier 1: Supports publicly readable content without authentication or more specific access control rules. Tier 1 nodes do not support content creation through the DataONE service interfaces and cannot operate as replication targets.
2. Tier 2: Read-only MNs that extends Tier 1 by supporting access control rules for non-public content.
3. Tier 3: Extends Tier 2 by adding the ability to add content through the DataONE service interfaces, and so provides full support for interaction with DataONE Investigator Toolkit applications and plugins.
4. Tier 4: Support the full set of DataONE APIs and can operate as replication targets, accepting content from compatible (technical and policy) MNs and fully supporting the DataONE content access control rules.

These four tiers provide natural divisions in functionality and are expected to simplify the process of adapting existing data repositories and building new MN software stacks to enable participation in the DataONE federation.

An internal committee was convened to formalize the process for prioritization of DataONE MN deployment. This committee shall review MN proposals and provide recommendations to help ensure optimal allocation of project resources to support MN development and deployment.

Progress continues on actual MN implementations with significant activity on implementing native Dryad MN services, as well as ongoing development of MN capabilities for the Avian Knowledge Network, CUAHSI, and Fedora and the UC3 Merritt system. It is expected this activity will be ongoing through 2011.

Considerable progress has also been made on the deployment of MNs operating within the context of the TeraGrid, where the MNs operate as staging locations where data necessary for experiments to be executed on TeraGrid resources is migrated using the replication capabilities of DataONE. After execution, the replication capabilities are again utilized to easily transfer result data sets to the MNs of the investigator's preference. The main TeraGrid resources we are interacting with include Lonestar at the Texas Advanced Computing Center (though migrating from this resource to the following), the Pittsburgh Supercomputing Center (PSC) (Albedo, Lustre storage services) and Indiana University (Quarry, Data Capacitor). This TeraGrid interaction is still in early stages, and is currently blocked by the refactoring of Coordinating Node services, but is expected to be online in prototype form some time before the end of July 2011.

Coordinating Nodes. Significant revisions are being made to the internal functionality of the DataONE Coordinating Nodes to reduce latency in tracking and updating system metadata associated with content, to support identity mapping and session management required for the authentication and authorization infrastructure, and to increase modularity and where possible, decrease interdependence of services critical to the internal functioning of the Coordinating Nodes.

Investigator Toolkit (ITK). Software tools representative of each stage of the data management lifecycle were presented at the NSF review in February. Little additional progress has been made on components of the ITK as resources have been focused on building out the core Member and Coordinating Node services. Products demonstrated at the February review included 1) the Morpho Metadata editor modified to interact directly with DataONE; 2) the Mercury search interface as modified by DataONE usability assessments and including extensions to support online citation managers such as Zotero and Mendeley; 3) A file system driver that enabled faceted browsing of the entire holdings of the DataONE federation or restriction to specific portions through filters; and 4) a plugin for the R statistical analysis package that enables retrieval of content from DataONE and storing results back to a MN through the DataONE service APIs.

Hardware Purchase at Coordinating Node Locations. Significant hardware purchases are in process at the three Coordinating Node locations to establish virtualization servers that will operate the critical Coordinating Node services in a high availability configuration. In addition, significant storage space and processing capability is being added to these physical locations to enable the provision of at least three moderately large (several hundred terabytes of storage at each location) Tier 4 MNs that will act as replication targets to help facilitate the early replication of information as the DataONE infrastructure moves into the public release phase. Hardware purchases are expected to be fully executed by the end of July 2011, with installations following shortly. The process at UNM differs slightly in that DataONE will be leveraging existing server and storage capacity available in the Library IT system with full virtual machine failover to other physical servers being brought online as part of the UNM Research Storage Consortium. This arrangement has potential to provide significant cost savings through the sharing of administrative resources and through simple economy of scale made

possible through collaboration with several other groups on campus. This model although still in its infancy, appears to augment the long term sustainability goals of DataONE – as the services provided by DataONE are seen as crucial, core operating services that can be leveraged by the library communities as they support the IT demands of increasingly technical clientele.

Concerns and their Mitigation. Resource availability remains the primary limitation on the rate of implementation of the DataONE infrastructure. Discussions with the USGS NBII resulted in some significant developer contributions, as such activity is considered well aligned with the goals and objectives of that group. For example, one developer has been active porting the DataONE file system driver to Microsoft Windows, and other activity has focused on the addition of DataONE MN service interfaces to data repositories within the USGS for both internal use and to expose appropriate content to the larger community through standardized service interfaces. Other opportunities for additional resources are being continually explored, for example through collaboration with new research projects (such as the recently awarded VertNet project that will be leveraging DataONE to support biodiversity data management services), and through other collaborations such as the recent agreement between DataONE and the UNM library information infrastructure to support the Coordinating and MN installations at UNM.

Another collaboration opportunity has emerged through a recent meeting with the Executive Director and lead staff of the Encyclopedia of Life (EoL). Although still early in the planning stages, there are several key services that could be provided to DataONE by EoL especially in the interests of reducing semantic chaff in concepts such as scientific names and other core organismal metadata. In return, EoL gains access to broader arrays of information pertinent to the goals of that organization, namely to provide rich metadata associated with all species of life on earth.

Overall, development progress is still on target for a 2011 release of infrastructure for public use. Resources remain very tight, resulting in highly focused development activities. Community excitement about the DataONE project though is providing some tangible benefits through contribution of developer resources to assist in the build out of this core infrastructure for earth sciences.

Community Engagement Progress

Overall summary. Community Engagement and Outreach in DataONE has made significant progress over the recent quarter in all four of its major activities: (1) providing responsive governance and management (focusing on the second DataONE Users Group and development of draft marketing materials); (2) engaging the broad community in DataONE and building an extensive data resource (e.g., focusing on the second DataONE Users Group); (3) creating an informatics literate populace (i.e., major updates to the Best Practices and Tools Databases, completion of a Data Management Primer, and collaboration in development of a Data Management Planning Online Tool); and (4) ensuring financial support and sustainability (focusing on the Marketing Plan and briefings with key partnering institutions). Highlights include:

- 1) Preparation for the establishment of the last Community Engagement and Outreach Working Group, Public Participation in Science Research, through the development and approval of the charter and membership list.
- 2) Launch of the DataONE Summer Internship Program for the third year in a row, soliciting applications from 75 well-qualified and motivated students. Eight students are actively working on a range of community engagement and cyberinfrastructure projects.
- 3) Significant progress on the development of the DataONE marketing and business plans by the Sustainability and Governance Working Group during their meeting in April in addition to

initiating planning for the 2nd meeting of the DataONE Users Group (DUG), in conjunction with the Federation of Earth Sciences Information Partners (ESIP) summer meeting.

- 4) The DataONE Users Group meeting is to be held July 11th & 12th in advance of the ESIP meeting, Santa Fe. Tentatively there are: 42 registered attendees. (15 of these are current members, 12 are new members and 12 individuals are ESIP partners who we anticipate will become members during the meeting. Another 4 new members are unable to attend making the DUG membership 44 individuals with the potential to reach 56 at the meeting. Our goal for the end of the year was 50 members.) The agenda for the meeting includes a day of updates on DataONE activities and preparations for public release with specific focus on the ITK, Education Resources, DMP Tool and Data Citation and Preservation. This will include ample opportunity for community feedback. The second day will focus on specific DataONE projects allowing for user engagement in the development of the DataONE marketing plan and products and in developing a mechanism for community feedback following public release.
- 5) Best Practices Workshop held in May 2011 was designed to bring together 40 experts to help augment the current DataONE online resources focused on 'Best Practices' for data management (i.e., 55 new BPs created and reviewed a new DataONE 'Best Practices Primer') and 'Software Tools' for implementation of those best practices (i.e., 150 new Software Tools created and edited for the DataONEpedia. These resources are currently located within our DataONEpedia (<http://www.dataone.org/dataonepedia>) and upgrades will be ongoing throughout summer 2011. In addition to Best Practices and Software Tools, we had a day focused on developing resources for the release of the Data Management Planning Tool (i.e., created 4 new exemplar DMPs and reviewed 2 existing DMPs for publication on the DataONE website; created a comprehensive data repository list, and community standards and metadata standards databases). We look forward to the beta release of the DMP Tool (<http://www.cdlib.org/uc3/datamanagement/dmpo.html>) early next month.

A Final Note

We thank all of the DataONE developers and the numerous Working Group participants as well as the Leadership Team and the External Advisory Board for their hard work over the past 20 months. Stay tuned to the web site, as we will be incrementally rolling out infrastructure and new educational resources throughout the summer and fall, right up to the full public release at the end of the year.