NSF DataNet Partners Update

by Dharma Akmon

EDITOR'S SUMMARY

Attendees at the 2014 Research Data Access and Preservation Summit (RDAP14) heard an update on the DataNet project, funded by the National Science Foundation (NSF) and designed to bring together data research infrastructure organizations to support digital preservation, access, integration and analysis. Five project partners have received funding, each convening collaborating institutions and pursuing group goals. DataONE focuses on data preservation and metadata, distributed storage, usability and assessment, educational outreach, data discovery and interoperability. SEAD (Sustainable Environment Actionable Data) seeks to build a cyberinfrastructure for sustainability science, providing a repository, researcher network and virtual archive. Tools developed by Terra Populus will enable users to extract data from different sources to create custom combinations. The goal of the DataNet Federation Consortium is to support collaborative research and federated collections by putting together a national data infrastructure with client-friendly metadata templates and bulk uploading to support long-term management. The DataNet partners continue separate lines of research to build an effective cyberinfrastructure that will support sharing and preserving diverse scientific studies.

KEYWORDS

data curation
access to resources
digital object preservation
research data sets
digital repositories
information infrastructure
scientific and technical information
collaboration

Dharma Akmon is the manager of education and outreach at SEAD (Sustainable Environment Actionable Data) Project based at the University of Michigan. Her work with SEAD focuses on expanding SEAD's user base, evaluating SEAD's services and usability and developing recommendations for improvements. She can be reached at dharmrae<at>umich.edu.

In 2007, responding to a scientific research context that was increasingly data-intensive, the National Science Foundation (NSF) Division of Advanced Cyberinfrastructure created the DataNet funding program. The purpose of this program was to create a set of "exemplar national and global data research infrastructure organizations (dubbed DataNet Partners) that provide unique opportunities to communities of researchers to advance science and/or engineering research and learning" [1]. The original plan called for a $100 million initiative. The NSF would make five awards of $20 million each, to be distributed over five years, with the possibility of continued funding after the initial five years. In the first review cycle (proposals due in March 2008), the NSF awarded two DataNets: DataONE and the Data Conservancy. The second round of DataNet funding was delayed as a result of the 2008 global financial crisis. Ultimately, awards in the second cycle were reduced from $20 million down to $8 million per project, and three additional projects were awarded DataNet funding in 2011: SEAD, DataNet Federation Consortium and Terra Populus.

The DataNet Partners share the goal of developing cyberinfrastructure that advances science; however, they differ in the specific communities and needs they aim to serve as well as in the particular problems they seek to address. Representatives from each of four DataNet Partners came together on an RDAP14 panel to present an overview of their projects and to update RDAP participants on important developments and plans going forward.

DataONE

As a product of the first round of DataNet funding, DataONE – a collaboration of multiple institutions based at the University of New Mexico – is the most established of the projects represented on the panel. In her talk, Amber Budden (DataONE's director for community engagement and outreach) highlighted DataONE's mission to enable universal access to

CONTENTS
< PREVIOUS PAGE
NEXT PAGE >
NEXT ARTICLE >

environmental and earth science data. DataONE (www.dataone.org) carries out this mission through three main sets of activities: community building, data discovery and interoperability solutions, and development of scientific tools and services.

Budden highlighted a number of DataONE's community-building efforts, in particular the establishment of several working groups and the development of a rich set of educational outreach activities and resources. DataONE's working groups serve as the foundation for the project's research and education activities. These groups include data preservation and metadata, distributed storage, and usability and assessment working groups (among several others). Educational outreach programs are another DataONE venue for building community around data curation issues. Specific education resources and programs include a best-practices database that provides users with recommendations for working with data at each stage of the data lifecycle; a set of education modules on data management that users can download and incorporate in their own lessons; and a librarian outreach kit aimed at alerting librarians to DataONE's most relevant products. DataONE actively enlists the participation of multiple stakeholder groups – through the DataONE Users Group and other outreach mechanisms – in disseminating and using these resources.

DataONE has also worked to enable data discovery and interoperability, developing a network of what it terms *member nodes* and *coordinating nodes*. Member nodes – of which SEAD, Dryad and USGS (U.S. Geological Survey) are three examples – are preservation-oriented repositories that have agreed to expose their data through the DataONE API, allowing the data to reach a larger audience. The three coordinating nodes at the University of New Mexico, the University of Tennessee Oak Ridge Campus and the University of California, Santa Barbara retain a complete metadata catalogue for the data, index those metadata for search and ensure content availability.

At the time of the presentation, DataONE had over 20,000 users, 34 production and in-development member nodes, 462,000 data objects, 13 tools and an active and diverse community of collaborators and partnering projects. Looking forward, Budden reported that the DataONE team is particularly focused on organizational sustainability planning to ensure that the project continues beyond the current NSF award.

SEAD

SEAD (Sustainable Environment Actionable Data) is a collaborative project between the University of Michigan, Indiana University and the University of Illinois to build cyberinfrastructure for sustainability science – a multidisciplinary area of research that addresses human impacts on the environment.

SEAD manager of outreach and education, Dharma Akmon, emphasized that SEAD supports active and social curation of data by embedding data curation within tools that support data creators' early work with data. SEAD (sead-data.net) allows scientists to more easily manage data as they work and then leverages scientists' early data curation work to facilitate the data's long-term access.

SEAD is made up of three sets of connected tools and services: the Active Content Repository (ACR), a secure area where project teams can manage their data; the Research Network and the Virtual Archive (VA), which is the data preservation layer for SEAD.

The ACR is intended to support data creators' work with data. Using the ACR, scientists can preview data in a web browser; annotate and describe datasets; automatically extract metadata; and create collections. Additionally, researchers are able to add geographic metadata to their datasets and view map overlays of geospatial data – features that are particularly welcome in sustainability science.

The SEAD Research Network contains over 1,800 profiles for researchers working in sustainability science. This network allows researchers to link their data and publications to a public profile that provides a record of their output. Users can also view visualizations of a researcher's co-author network and disciplinary affiliations.

Lastly, the VA leverages existing institutional repositories to ensure long-term access to data. Researchers can deposit data by marking them "ready for publication" from the ACR, or they can bypass the ACR and submit a collection directly to the VA. The VA matches the collection to an appropriate repository and hands it off for curator review. Published data are assigned a DOI, and the metadata are indexed to facilitate discovery.

SEAD is expanding outreach efforts through workshops that introduce scientists to SEAD's tools and collect feedback for future developments. SEAD is also currently making its user interfaces more intuitive and refining a research-object lifecycle model that will aid in future iterations of the VA and ACR.

Terra Populus

Terra Populus (known as "TerraPop") is based at the Minnesota Population Center and focuses on enabling the integration of data on population and the environment. David Van Riper, director of spatial data at TerraPop, highlighted the project's overarching goal to lower barriers that make it challenging to conduct interdisciplinary human-environment interactions research. To that end, the TerraPop team is building tools that allow users to combine seemingly disparate data from multiple sources into new, customized data extracts.

Scientists' understanding of population-environment dynamics could be significantly enhanced by the combination of different types of data. For example, a researcher might want to combine census microdata that describe socioeconomic/demographic information about individuals, raster data that characterize rainfall and vector data that depict land use. Normally a cumbersome and time-consuming process to carry out, with the tools TerraPop is developing, researchers will be able to select the particular variables from the different datasets of interest, choose the format they would like their data to be in, and create a new data extract.

TerraPop is at a fairly early stage of development. Going forward, the team will expand TerraPop collections to include more census, survey and global environmental data as well as GIS mapping files. This expansion will be accomplished, in part, by developing data curation protocols and mechanisms that will make it easier for scholars to deposit their data and metadata with TerraPop. Shortly after our panel, TerraPop (www.terrapop.org) released a beta version of its customized data extraction and creation service. In addition to collecting feedback from users about the beta version, TerraPop is working on becoming a DataONE member node, ramping up outreach efforts, and forming collaborations with the other DataNet Partners.

DataNet Federation Consortium

The DataNet Federation Consortium (DFC) – based at the University of North Carolina at Chapel Hill – aims to assemble a national data infrastructure in support of collaborative scientific research. Mary Whitton, DFC project manager, underlined three main areas of DFC activity: federating iRODS (Integrated Rule-Oriented Data System open-source data management software)-based grids and making those grids interoperable with other systems; enabling reproducible science through iRODS workflow data objects and extending the iRODS policy-based data system to better support creation, use and management of federated collections. In her presentation, Whitten focused primarily on describing the DFC's work to facilitate interoperability and long-term data management.

The DFC (datafed.org) serves data producers, data users, curators/archivists and data center managers from any discipline. Targeting data producers and users, the DFC is developing client-side tools for creating metadata templates and bulk-uploading data to iRODS collections. The project team is also implementing iRODS policy-based access controls to enforce access rules – such as IRB (Institutional Review Board) restrictions – and allow a researcher to enable data access for a defined group of users in one step. Plans include providing integrated access to data analysis tools (for example, R and MATLAB) from iRODS workflow objects. With data curators and archivists in mind, the DFC is writing iRODS rules to implement archiving best practices such as ISO 16363 (repository trustworthiness) and ISO 14721 (open archival information system). The DFC's development of data-center collection management tools promises to streamline functions such as integrity checking and replication. Future iRODS rules and policies for machine-verifiable best practices and standards will help establish iRODS grids as trustworthy repositories. Additionally, iRODS can now automate execution of important repository tasks such as copy, backup and checksum.

Whitton also highlighted iRODS version 4, which was released shortly after RDAP14. This version embodies current software engineering practices, easier installation and a plug-in model for future extensions. Significant interoperability efforts are ongoing between iRODS grids and DataONE, and between iRODS grids and the Dataverse Network. As with the other DataNet Partners, outreach will be an important component of the DFC's work going forward.

Conclusion

The DataNet Partners have developed and are continuing to develop infrastructure to support 21st century science. Each of the four panelists emphasized the vital role that outreach will play as they cultivate communities, solicit feedback, promote their offerings and work together. The RDAP14 panel was a timely opportunity to update data librarians, curators and archivists on DataNet Partner activities and to invite feedback on the ways in which the partners can support these constituencies' work with scientists and their data.

Acknowledgements

I would like to thank Amber Budden, David Van Riper and Mary Whitton for their participation on the panel and for their review of relevant sections of this article. The projects discussed are funded by the NSF under cooperative agreements #OCI-0830944 (DataONE), #OCI-0940824 (SEAD), OCI-0940818 (Terra Populus) and OCI-0940841 (DFC). ■

Resource Mentioned in the Article

[1] National Science Foundation. (2007). *Sustainable Digital Data Preservation and Access Network Partners (DataNet) program solicitation.* Retrieved from www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm