**UF Data Management/Curation Task Force, Review of the Dataverse Network with Report from Testing:**

**SURA Research Data Management (SURA RDM) Dataverse Pilot Project**

**Please note:[[1]](#footnote-1)**

The University of Florida participated in the Dataverse Network (DVN) pilot looking for DVN as a possible tool to be one of many tools in our garage of tools for data management. UF is always interested in leveraging capacity where it exists and optimizing for scale (ensuring solutions meet immediate needs, are scalable for full needs or build us towards scalability and support building capacity at UF and across/with our research communities at many institutions).

Since the initial testing time and report, UF’S Research Computing planned and began implementation on an alternate solution (as part of a statewide solution with SSERCA).[[2]](#footnote-2) With the new solution meeting all needs expected to have been supported by DVN, UF is not pursuing DVN further at this time. The new solution is in queue for production availability in Q1 2014. Representatives from the Libraries, Research Computing, and the Office of Research will continue collaborating to promote and provide support for data needs using the new tool.

**UF Data Management/Curation Task Force, Review of the Dataverse Network with Report from Testing:**

**SURA Research Data Management (SURA RDM) Dataverse Pilot Project**

In early 2013, the SURA (Southeastern Universities Research Association) Research Data Management (RDM) group issued a call for participation for a multiple institution pilot implementation of the Dataverse Network (DVN, <http://thedata.org>). The goal of the SURA pilot is to work through a DVN implementation with several SURA RDM members. The pilot project for SURA is hosted by UNC’s Odum Institute. The evaluative results from the pilot will be used to inform SURA RDM’s next steps (direction, enhancements, gaps, integrated approaches, etc.) for enhancing best practices and services. UF elected to participate in the pilot.

**UF Evaluation Process**

On April 24 and 29, members of the Data Management/Curation Task Force took part in two hands-on evaluation sessions of the Dataverse Network for possible use at UF.

From the two evaluation sessions, attendees identified several features of interest:

* Overall feel as a system designed and intended for “data”
* Support for versioning
* Opportunities with specific fields and types of research, as with GIS-related research, and opportunities for specific use cases with research labs and graduate students for DVNs as a possible part of the data curation lifecycle

Because the Dataverse Network (DVN) requires submission of minimal metadata for the data to be uploaded, it ensures capture of some metadata when data is stored in the DVN. This is in many ways parallel to the IR@UF and other subject repositories. The DVN differs from the IR@UF and other subject repositories in that researchers can use the DVN space as a sort of processing or lab space where data can be uploaded and all of the metadata and data files can remain in a protected status where limited or no information is accessible. This could be very useful for individual researchers, as with graduate students working on data collection and analysis for their dissertations to have an organized, backed up, secure, and protected space to store their data.

DVNs allow for different individuals and groups to have access to files, even files in a protected status using different pre-set levels of permissions. Thus, a graduate student could use the DVN as a data workspace for their immediate needs which could include sole access as well as access by the student and committee members. For a research lab or team, all members could have access to data that would be protected to limit access to only members of the group. Access and authority levels can be further defined with only certain people with higher level roles and permissions to be able to “release” data. This could allow for a full research team to use the DVN as the working space for data, with that working space also being part of the larger data curation lifecycle infrastructure at UF.

The working-space aspect of the DVN could prove beneficial as a replacement for current structures where data is stored in simple directories on local or network drives. While entering data into the DVN requires only minimal metadata, the DVN system immediately structures the data into “studies” and automatically captures additional data (person submitting files and making changes, dates of entry and changes, etc.). The structuring of the data and additional metadata, even if relatively minimal, is a dramatic improvement over the current workspace for data on drives that are only organized by folder and where, at best, a README file provides information on contents and organization.

In addition to possibly serving as a working space for data, DVNs offer additional functionality that could be of high interest and added value for research across campus. For instance, DVNs allow researchers or organizational units to group their research with similar research, providing opportunities for customized research portals. This allows researchers to create a research environment that is useful for the full research process individually and for research teams. For instance, a researcher collecting GIS data could use the DVN to store, version, and share that data. If the researcher was also utilizing US Census data and environmental data from another outside source in concert with the researcher’s GIS data for analysis and findings, all of these materials could be available within the same DVN using the collection functionality to add US Census data, environmental data, and other resources available on other DVNs. The DVN collection functionality could allow for researchers to create research portals for specific research needs.

Because the DVN software is Open Source, if the DVN model is found useful, further development could be pursued to add advanced features. DVNs already offer advanced visualization and other tools for certain types of data most common to social, business, and economic fields. If the DVNs are useful and implemented at part of the campus-wide support for UF, future steps could include investigating possibilities for developing additional advanced features to serve campus needs.

**Recommendations for Further Testing of DVN**

Based on the evaluation sessions, initially two and then another two volunteered for specific scenarios identified of interest for further pilot testing with a system installed at UF to fully test the configured system with integration with Shibboleth, permissions, support on UF servers, record harvest to the IR@UF for discoverability of released data, user training and support requirements, etc. These pilot tests would be implemented and then evaluated in context with the current level of support and other tools and supports.

* Scenario 1: Research Labs/Teams using the DVN as an integrated part of a research project for the data curation lifecycle
* Scenario 2: Graduate students: using the DVN as an integrated part of their thesis/doctoral research
* Scenario 3: Chemistry department, 2 labs that serve the department, campus, external institutions, and businesses: using the DVN as an integrated part of the workflows with different security constraints
* Scenario 4: IFAS NFREC: 2 extension offices using the DVN to track metadata for data for their in process publications and to track and share completed data following publication
* Possible Scenario 5: Existing archival/record collections from data providers regularly consulted by campus faculty (e.g., water management districts; University Archives)

In addition to the pilot testing scenario considerations with Shibboleth integration, transition from drive to DVN workspace, and researcher needs and feedback, considerations identified as needing further investigation as part of the pilot testing include:

* Interface refinements (metadata forms; main screen long descriptions complicate the item list)
* Relationship to electronic lab notebooks (ELNs) and related considerations
* More testing and evaluation overall for the process
* Considerations for developing guidance and best practices recommendations
* Readiness for inclusion in data management plan, and template plans using the DVN
* Analysis for future opportunities: DVN as a system kernel with additional functionality like visualization and subsetting and what those would be for different fields like chemistry, etc.

**Recommendations for How to Proceed with Pilot Testing**

The UF pilot testing is expected to require 12-24 months for actionable results that go beyond the known evaluation of DVNs as preferable to simply having data files on drives.

The first 1-3 months will be required for set up, including installing and configuring the software and identifying and setting up appropriate test projects. Actionable results will depend on the activity for the research projects.

Research Computing is best positioned to install the DVN software on their servers, providing technical support in collaboration with the libraries. This will best support testing of all aspects of the technology, including needs assessment and integration with other systems.

Subject specialist librarians are best positioned to identify appropriate research projects for testing, where projects match the test scenarios and are in their initial stages or other defined data lifecycle stages for testing the software as it relates to the data lifecycle process needs. The subject specialist librarians are best positioned to ensure that the researcher or researchers are interested and that the testing process will best serve the needs for the researcher project and for wider evaluation.

During the testing process, the Dataverse software should be available for others who are interested in testing it because these users may provide valuable feedback. The Libraries and Research Computing will need to draft standard language for data management plans for researchers when preparing grant proposals.

**Integration with Existing Systems, Sustainability after the Pilot**

The pilot testing process will provide data as to the efficacy and value of supporting DVNs locally. Whether or not the pilot results in adoption and the commitment to support DVNs moving forward, the IR@UF can support the final, released data as record-only items (if the DVN is adopted) or as full resource objects with all data files and metadata. DVNs support OAI-PMH and so the IR@UF could harvest released records from the DVN into the IR@UF for discoverability through the IR@UF. The IR@UF is optimized for search engine indexing and has a library catalog record feed, so all materials would also be widely discoverable in search engines like Google and Bing and through the UF Library Catalog and other library catalogs around the world. This level of integration and support would be in place regardless of the adoption of DVNs. If the DVNs are not adopted, any final data could be added to the IR@UF; however, unreleased data would need support for transition to an alternative with the existing alternative being drive space.

Additional information on the costs, requirements, impacts, value, and opportunities in regards to sustainability and integration with existing systems for the DVN will be developed during the pilot process.

1. Note added in Dec. 2013, after the initial evaluation process for this report, with new information that impacted this report and planned activities. [↑](#footnote-ref-1)
2. More information available in the Research Computing Advisory Committee meeting notes: <http://www.it.ufl.edu/wp-content/uploads/2012/03/research-computing-minutes20131202.pdf> [↑](#footnote-ref-2)