Becoming DataONE Tier-4 Member Node: Steps Taken by the Nevada Research Data Center

Moinul Hossain, Hannah Muñoz, Rui Wu, Eric Fritzinger, Sergiu M. Dascalu and Frederick C. Harris Jr

Department of Computer Science and Engineering
University of Nevada
1664 N. Virginia St., Reno, Nevada 89557, USA

{hossain, hannahmunoz, raywu1990}@nevada.unr.edu {ericf, dascalus, fredh}@cse.unr.edu

***Abstract*—The Nevada Research Data Center (NRDC) joined up with DataONE by becoming a Member Node to help with their goal of preserving and sharing scientific data. The NRDC is an NSF funded data management center for several climate based research groups across Nevada. DataONE is a collaboration that seeks to conserve scientific data and make researching global ecological issues easier. We brought our Member Node up to Tier 4, which allows us to make replication of other Member Nodes' data for safekeeping. By becoming a Member Node with DataONE, we expect that the projects supported by NRDC will have higher visibility, for the benefit of interested researchers and data users everywhere, who will have straightforward access to the many existing NRDC datasets (currently totaling over 2.1 billion of environmental data measurements). In this paper, we first describe the DataONE model of linked data repositories and then provide details of the configuration and development steps performed to fulfill the requirements of becoming Tier 4 (maximum possible) Member Node with DataONE. The significance of this work is also briefly discussed in the paper, and planned directions of future work are outlined.**

I. Introduction

The Nevada Research Data Center is a scientific data management facility used to store and share research data and share research results for Nevada-based scientific research projects, such as the Solar Energy-Water-Environment Nexus (Nexus) and Walker Basin Hydroclimate projects [1]. Originally created as a data portal for the NSF Experimental Program to Stimulate Competitive Research (EPSCoR) project called Nevada Climate Change Portal (NCCP), the NRDC stores the climate data from remote locations in Nevada. The NCCP is now known as the Nevada Climate-Ecohydrological Assessment Network (NevCAN), and the expended NRDC integrates and maintains the project's data for sustainability purposes [2, 3]. The NCCP aims to build infrastucture for research purposes as well as to support the research on climate change in Nevada [4, 5, 6].

Projects such as the Walker Basin, and NevCAN are local NSF-funded research activities in Nevada that leverage modern sensor technology to record continuous observations of climate and ecohydrology [4]. These remote sites are connected to the NRDC using a private research network that forms the backbone of emerging field cyberinfrastructure in the region. Such networks will play a significant role in the future of environmental science [7]. Sensor data from these sites are generated by a variety of dataloggers, cameras, and other IP-enabled devices, which then stream the data back to storage servers within NRDC. Variables observed include common observations such as air temperature, wind speed, atmospheric pressure, relative humidity, and precipitation, as well as more unique measurements including vegetation sap flow, normalized difference vegetation index (NDVI), and soil column drainage. Weather and climate data are captured at resolutions from 1 minute to 1 hour, with typical camera imagery being archived at the 30 minute or 1 hour interval during daylight hours. Vegetation and soils monitoring data are typically captured at 30 minute or 1 hour intervals. A total of 16 stations between these three projects are connected to NRDC, streaming hundreds of thousands of individual data points and images per day into system servers, amounting to approximately 1 GB of daily data influx from these very remote parts of the state. Currently, NRDC stores over 2.1 billion of data measurements.

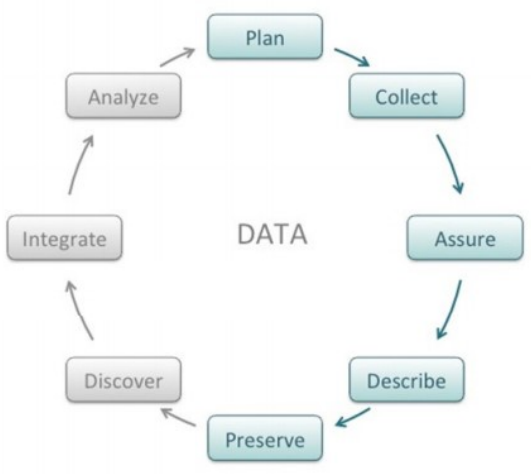

Fig. 1. The lifecycle of data in DataONE [10].

The Data Observation Network for Earth (DataONE) is an NSF-supported international collaboration used to gather and store data for researching biological, ecological, or environ-

978-1-5090-4489-4/17/$31.00 ©2017 IEEE

mental details [8]. DataONE tries to conserve datasets from multiple institutions and disciplines and distribute them to the public. Researchers can use this data to help research global environmental problems [7]. DataONE's vision statement includes "be[ing] commonly used by researchers, educators, and the public to better understand and conserve life on Earth and the environment that sustains it" [7]. DataONE uses "a distributed framework and sustainable cyberinfrastructure" to dispense access to scientific data across its affiliated institutions [9]. Through a network of Member Nodes, Coordinating Nodes and with the help of an Investigator Toolkit, DataONE collects data for public distribution. Researchers can use this data to help with studies in their respective fields. Currently, there are 36 Member Nodes and 3 Coordinating Nodes associated with DataONE.

DataONE is heavily dependent on the eight-step lifecycle model shown Fig. 1. This model is based on a lifecycle similar to one developed by the NSF [9]. The cycle follows the collection of data from the beginning of a project, to data preservation, onwards to other researchers using the data in their projects. Any stages may be skipped and researchers have access to the data at any point in the lifecycle [7].

The NRDC chose to work together with DataONE by becoming the maximum Tier 4 Member Node. This will help us support the data collected from the NevCAN and Walker Basin Hydroclimate projects. By aligning ourselves with DataONE, we intend to boost the visibility of NRDC affiliated projects. By joining a national data repository, we hope the data collected from the NevCAN and Walker Basin Hydroclimate projects disseminate more comprehensively to other institutions who can make use of the data. Achieving higher data discovery rates means increased visibility of NRDC supported projects and increased data usage by a wide variety of research works.

When we undertook the work for making NRDC a DataONE Member Node, there was little documentation and guidelines available on how exactly to achieve this, except from the technical support kindly provided by the DataONE team. In this paper, we aim to provide an experience report that not only highlights the significance of sharing data through DataONE but also offers some guidance on how to approach making one's data storage center or facility a member with this NSF-supported large international organization. Although heavily focused on software development and configuration, the work presented in this paper ultimately contributes to supporting scientific research through making data and related resources more easily and comprehensively available to the interested researchers and other data users.

In this paper, we describe the DataONE system architecture, the requirements for membership as a DataONE node, the means we used to meet those requirements and the results of this undertaking. This paper is structured as follows: Section II presents the architecture of DataONE and how to implement its services; Section III details our process of being a Member Node; Section IV presents a discussion of benefits and results of becoming a Tier 4 Member Node; and Section V contains our conclusions and goals for future development.

II. DataONE Model

Establishing a research institution as a DataONE member node can be a major undertaking with numerous steps and requirements to be met. However, we can simplify the explanation of the process by dividing the process into three major parts. The first step in this process is understanding the complex architecture of DataONE. From there, we move to a process of configuring our own infrastructure and connecting to the larger DataONE network. Finally, we end with a local implementation of the DataONE API to take our Member Node status from Tier 1 to the maximum possible Tier 4.

The DataONE infrastructure has three major components.

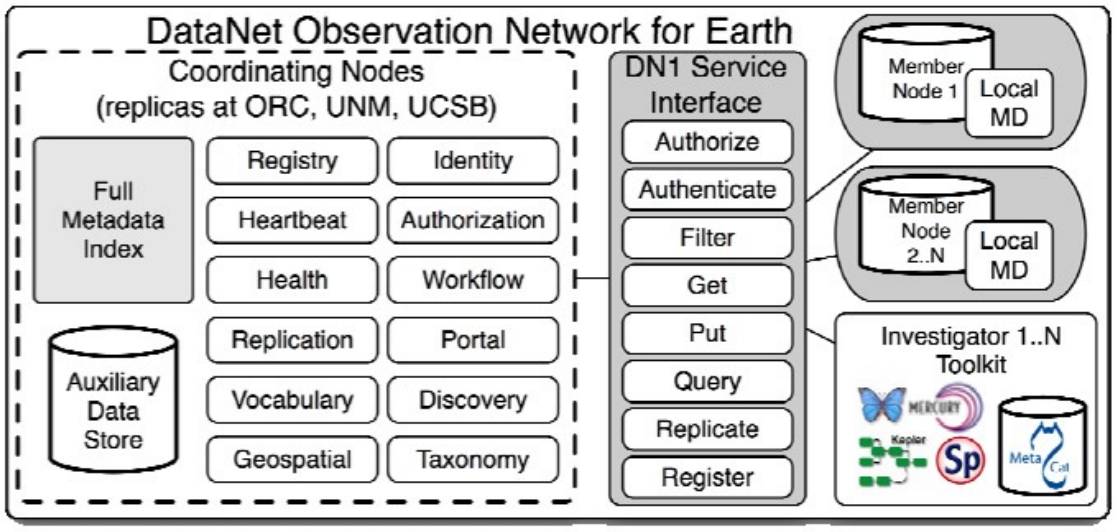

Fig. 2. The connections of the Member Nodes, Coordinating Nodes, and Investigator Toolkit [9].

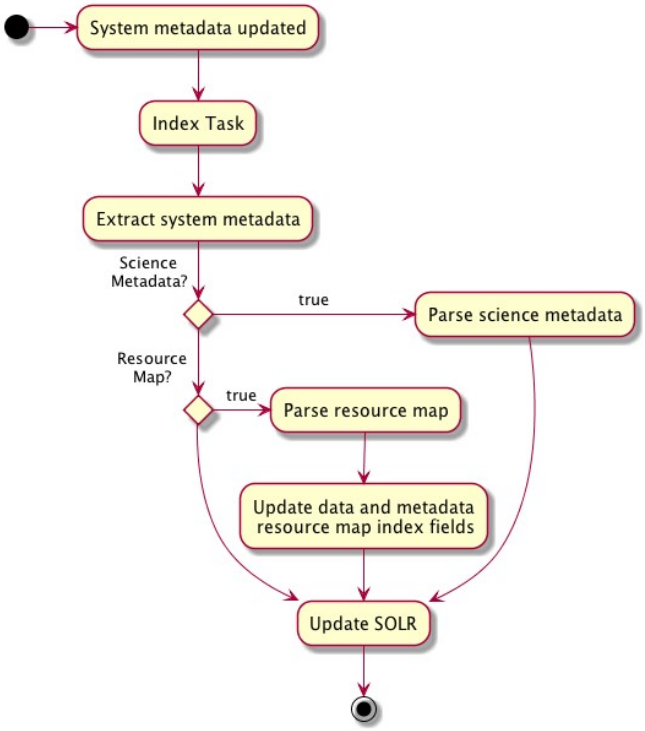

Fig. 3. How the data and system metadata relate to one another [9].

These are detailed in Fig. 2 as Member Nodes (MN), Coordinating Nodes (CN), and the Investigator Toolkit [9, 10]. Member Nodes, where data is gathered and stored, are installed by users with data repositories. These nodes use DataONE's own API to connect to Coordinating Nodes to promote consistency among Member Nodes connections. Coordinating Nodes are a collection of data replication and indexing services that discover and manage data across Member Nodes [9]. The DataONE Investigator Toolkit, provided to each Member Node, formats local data as a Python object and uploads it to DataONE. The Investigator Toolkit is a collection of custom and pre-existing software used to analyze data from and contribute data to DataONE. Some examples of the software in this toolkit are: Morpho, Mercury, and Zotero [10].

The functionality of a Member Node is divided into a four-tier system. A Tier 1 node is a node that allows for public read-only access and indexing by the Coordinating Node [11]. Tier 2 Member Nodes are authorized via either a DataONE password, a 3rd party password, or a CILogon with an X.509 certificate. Tier 3 Member Nodes allow DataONE to handle creates, updates, and archives of their data. In lower tiers, the users must manage their own data. A Tier 4 Member Node implements the MNReplication API. Futhermore, Tier 4 Member Nodes create replications of its data and accept replications for storage [10]. According to DataONE specifications, a "fully functional" node need only be Tier 1.

DataONE uses a collection of data objects and their related science metadata to create resource maps. A resource map uses the science data, science metadata and DataONE objects to map the relations between the objects, as presented in Fig. 3. Data is defined by DataONE as being "a discrete unit of digital content that is expected to represent information obtained from some experiment or scientific study [11]." Its related science metadata, "describes the properties of the data". The Coordinating Node stores science metadata for use in user searches for data. DataONE also has system metadata, which "contains attributes that describe the digital object it accompanies [11]." Persistent Identifiers (PID) identify all science data, science metadata, and system metadata and relates to a static set of data regardless of whether it has been modified or not.

III. Making NRDC a DataONE Member Node

Our DataONE Member Node uses Ubuntu 14.04.5 with kernel 3.13.0-66-generic on a full virtual environment using Microsoft Hyper-V. The virtual machine uses an Intel Xeon-CPU E5-2670 @ 2.60 GHz with 10 GiB of RAM. We are currently running as a version 1.2.10 Member node.

We started our Member Node by making a user account called "gmn", which is short for Generic Member Node, on the virtual machine which houses our DataONE software environment. On this account, we installed the base software stack provided by DataONE. This basic stack downloads and installs all the software packages and dependencies that the Member Node relies on [12]. The GMN package consists of software used for building PostgreSQL databases, DataONE certificates and parsing for resource mapping. Once the installation was finished, we added the GMN's root directory to .bashrc [13].

With the base environment installed, we were able to set up the server. Our MN uses Apache 2.4.7. We enabled the wsgi and ssl modules within Apache to handle and validate the connections and certificates between our MN and the CN and to use the Django framework, as specified by the DataONE documentation [13]. DataONE provides a custom apache2.conf file, which we used without any modifications. Finally, we enabled port forwarding.

Next, we installed PostgreSQL 9.3.15 with default values [10]. The gmn account was made superuser and an empty database, named nrdc_dataone, was created.

We then set up two cron jobs that check the replication and system metadata queues sent by the CNs. The GMN stack provided both these processes. The replication request queue was configured to check for new requests every minute while the system metadata queue was set up to run every 30 minutes.

We then generated client side certificates using OpenSSL as per the Tier 2 requirements. The certificate created for communication with the Coordinating Node was moved to a special location for local certificates. Another certificate, created for communication with other Member Nodes, was moved to the same directory. Next, we deleted and replaced these certificates with the actual ones from DataONE once we were fully registered [13].

We then updated our site settings from the default with information about the NRDC. We created a copy for editing of the site setting template and named the file setting_site.py.

Tier 1

ping () getLogRecords() getCapabilities() get()
getSystemMetadata() describe() getChecksum() listObjects()
synchronizationFailed() getReplica() query()
getQueryEngineDecription() listQueryEngines()

Tier 2

isAuthorized() systemMetadataChanged()

Tier 3

create() update() generateIdentifier() delete() archive()

Tier 4

replicate()

Fig. 4. The methods of the DataONE Common and Client Libraries separated by tier.

We first generated a secret key for Django and added it to the setting_site.py file [13]. Then, we changed the default values to our unique values. The Node Identifier was changed to urn:node:NRDC, a name decided for us after a meeting between the NRDC group and DataONE. We changed Node Base URL to point to our Member Node. At this point we had successfully met all the requirements of a "fully functioning node" so our Tier status was set to 1.

The final steps involved in creating a MN began with the initialization and start-up of the programs we had previously downloaded and configured. We then set all the files in the GMN stack to be owned by gmn. The PostgreSQL database was also initialized. We opened the firewall on port 443 and restarted Apache [13]. Finally, STAND_ALONE mode was set to false in setting_sites.py, leaving the MN now ready to run as part of the DataONE network.

To prepare for the transfer of data to DataONE, we first downloaded the DataONE Common Library and the DataONE Client Library whose methods are shown in Figure 4. The DataONE Common Library provides connection functionally between Member Nodes and Coordinating Nodes. Essentially, by downloading the DataONE Common Library, a Member Node possesses Tier 1 capabilities. The library also handles XML file errors and the DataONE Client Library provides the API that lets us further develop the capabilities of our MN [13].

We began our code by connecting the Member Node to DataONE using the security certificates we created while setting up the Member Node. We then opened the PostgreSQL database. Upon finding data files associated with projects, we generated the data's metadata.

The system metadata entity holds the science object's metadata: the PID, object size, encryption type, encrypted checksum, access policy, and creation and modification dates. The access policy defines the public use of the data. We established default configurations for our access policy, except for permissions, which were set to read only. We then searched the database for our local science data. If it is in the database, all rows are returned. Otherwise, nothing is returned.

When no data object is returned, this indicates that the data is not yet in the Member Node and needs to be updated. The data, metadata, and package are all generated and assigned a PID. System Metadata is created from the data, metadata settings, and a MD5 hash of the science data.

Three new objects were created on the Member Node. The first is comprised of the data PID, science object and system metadata. The second holds the metadata PID, the science metadata, and the system metadata. And, the final object holds the package's PID, package itself, and system metadata. Using the PIDs of the science object, the metadata, the package with the file location, hash and date variables, we could create associated values in the database and update the table.

If a data object was returned, the object already existed within DataONE and was checked for any changes made to

Node Id:	urn:node:NRDC
Type:	mn
Name:	NRDC DataONE member node
Description:	Production Member Node for Nevada Research Data Center (NRDC)
Base URL:	https://dataone.sensor.nevada.edu/mn
Subject[s]:	CN=urn:node:NRDC,DC=dataone,DC=org

Services:	Name	Version	Available
	MNCore	v1	true
	MNRead	v1	true
	MNAuthorization	v1	true
	MNStorage	v1	true
	MNReplication	v1	true

Status:	up
Replicate:	true
Synchronize:	true

Synchronization schedule:	Year	Month	Day of Month	Day of Week	Hour	Minute	Second
	*	*	*	?	*	0/3	0

Last Harvested:	2016-12-12T23:23:46.099+00:00
Last Complete Harvest:	1900-01-01T00:00:00.000+00:00

Fig. 5. Descriptors of the final NRDC Tier 4 DataONE Member Node.

the metadata. To accomplish this we used an MD5 hash to check for possible changes to the metadata. If changes were found, we rehashed the edited data and system metadata using the MD5 algorithm. After we generated and assigned new PIDs for the data, metadata, and package, we updated the Member Node and the database.

Finally, to achieve the maximum Tier 4 Member Node status, we needed to enable replication. This was accomplished by adding a replication policy to the system metadata and adjusting the tier level in settings_site.py. We currently do not allow replications to be made of our data objects, so we set replication permissions to "false". In settings_site.py we adjusted our Member Node status to Tier 4 allowing replications of other Member Nodes to be made on our server. We pushed the changes to DataONE and restarted Apache so the tier changes could take effect. Our finished node is shown in Figure 5, where all the DataONE libraries have been implemented [14]. In the above, one can notice that moving from Tier 1 to Tier 4 does not require prescribed "stationing" periods in Tiers 2 and 3, which were reached in the process.

IV. Results

The steps presented above allowed NRDC to become a Member Node with DataONE at the highest possible level (Tier 4) and enabled our center to join the company of similar other data centers and repositories from the US and worldwide. Several of these data sharing facilities, including NRDC, are shown in the partial list of DataONE member nodes presented in Figure 6. The process described in this paper could be repeated by other data repositories similar to ours, which consequently could more easily and comprehensively share their data and benefit from the existing DataONE resources. Ultimately, the main beneficiary of this work are researchers and other interested users of environmental Earth data.

The steps described in Section III can also serve as encouragement for data repositories similar to NRDC to start undergoing the DataONE membership joining process, which initially might look overwhelming but, according to our experience, proved to be reasonably straightforward. Our team wanted to share this experience as we hope it could facilitate similar data-sharing approaches.

By joining DataONE and raising the member status to the highest Tier 4, NRDC affiliated projects have become more visible. This in turn creates an opportunity for researchers within the NRDC project to share results with other NSF and environment-based institutions. Additionally, through NRDC joining DataONE, this action also supports DataONE's goal of uniting environment-based research through its distributed architecture. Furthermore, joining an ever-growing community of environmental and earth scientists, some of which are seen in Figure 6, can potentially open up new research collaborations and proposals for NRDC.

By bringing the NRDC Member Node up to Tier 4, we will be able to replicate the data of other institutions while allowing others to replicate our own. This action is very significant in that it promotes data preservation. To elaborate, this can potentially serve as redundancy for backing up data in such cases when an incident occurs and NRDC data becomes compromised. Through DataONE, NRDC data would be preserved through DataONE and the replications across other institutions.

With the Member Node set up, should the Walker Basin HydroClimate, NevCAN, and other future projects request the NRDC to replicate their data, we can ensure that there will be

Fig. 6. Excerpt from DataONE Membership list.

external backups of project data outside the main system. This is of high interest to the NRDC because of its goal of keeping the data available to researchers at all times [4]. By joining DataONE, affiliated researchers will now have unrestricted, reliable access to their data at anytime.

V. Conclusion and Future Work

DataONE is a collection of data repositories that help with the dissemination of data. A Tier 4 Member Node allows us to share our data with the others and also replicate our data sets for safekeeping. We hope to encourage other to join DataONE by discussing DataONE's requirements, architecture and implementation of its services. By becoming a Member Node in DataONE, we hope to increase visibility of our supported projects.

Our next step in joining DataONE is to upgrade our Member Node to version 2. Version 2 Member Nodes have updated a new DataONE API [12]. We plan to upgrade to version 2 by May 2017.

The NRDC also plans to join the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) which has data sharing services, Hydrologic Information System and HydroShare, that allow researchers to upload their scientific data and enable others to use it as part of their research [15]. By joining CUASHI, we add more visibility to the research done by our supported projects.

We also plan to join the Community Surface Dynamics Modeling System (CSDMS), a project that helps model the Earth's surface [16]. By contributing our hydrologic data to the CSDMS, we hope to gain further visibility for our supported projects and promote their data sets.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under grant number IIA-1301726. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We are also grateful to the DataONE team, in particular to Laura Moyers and Mark Servilla, for the support they provided during our membership joining process.

References

[1] "Nevada research data center," *Nevada Research Data Center*. [Online]. Available: http://www.sensor.nevada.edu/NRDC/. [Accessed: 7-Feb-2017].

[2] Dascalu, S., Harris, F.C., Jr., McMahon, M., Jr., Fritzinger, E., Strachan, S., and Kelley, R. (2014). An Overview of the Nevada Climate Change Portal. *Proceedings of the 7th Intl. Congress on Environmental Modelling Software (iEMSS-2014)*, San Diego, CA, June 2014, vol. 1, pp. 75-82.

[3] Le, V.D., Neff, M.M., Stewart, R.V., Kelley, R., Fritzinger, E., Dascalu, S., and Harris, F.C., Jr. (2015). Microservice-based Architecture for the NRDC. *Proceedings of the 13th IEEE International Conference on Industrial Informatics (INDIN-2015)*, July 2015, Cambridge, England, pp. 1659-1664.

[4] McMahon, M.J., Jr., Dascalu, S., Harris, F.C., Strachan, S. and F. Biondi (2011). Architecting Climate Change Data Infrastructure for Nevada, in Salinesi, C. and Pastor, O. (eds.), *Advanced Information Systems Engineering Workshops CAISE-2011, Lecture Notes in Business Information Processing*, LNBIP-83, June 2011, Springer, pp. 354-365.

[5] Strachan, S., Kelsey, E. P., Brown, R. F., Dascalu, S.-M., Harris, F., Kent, G., Lyles, B., McCurdy, G., Slater, D., Smith, K. D. (2016). *Filling the Data Gaps in Mountain Climate Observatories Through Advanced Technology, Refined Instrument Siting, and a Focus on Gradients*. Mountain Research and Development, 36(4), 518–527.

[6] McMahon, M.J., Jr., Harris, F.C., Jr., S. Dascalu, and Strachan, S. (2011). S.E.N.S.O.R.- Applying Modern Software and Data Management Practices to Climate Research, *Proceedings of the 2011 Workshop on Sensor Network Applications (SNA-2011)*, Nov. 2011, Honolulu, HI, pp. 147-153.

[7] S. Allard, "DataONE: facilitating eScience through collaboration", *Journal of eScience Librarianship*, 14-Feb-2012. [Online]. Available: http://escholarship.umassmed.edu/jeslib/vol1/iss1/3/. [Accessed: 06-Feb-2017].

[8] Mensing, S., Strachan, S., Arnone, J., Fenstermaker, L., Biondi, F., Devitt, D., Johnson, B., Bird, B. and Fritzinger, E., 2013. A network for observing Great Basin climate change. *Eos, Transactions American Geophysical Union*, 94(11), pp.105-106.

[9] Michener, W.K., Allard, S., Budden, A., Cook, R.B., Douglass, K., Frame, M., Kelling, S., Koskela, R., and Tenopir, C. (2012), Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences, *Ecological Informatics*, vol. 11, September 2012, pp. 11-15.

[10] "DataONE: Data Observation Network for Earth," *DataONE*, 14-Mar-2017. [Online]. Available: https://www.dataone.org/. [Accessed: 7-Feb-2017].

[11] "DataONE Architecture, Version 1.2," *DataONE Architecture, Version 1.2*, 17-Oct-2013. [Online]. Available: https://releases.dataone.org/online/api-documentation-v1.2.0/. [Accessed: 07-Feb-2017].

[12] "DataONE Python Products documentation," *DataONE Python Products*, 2016. [Online]. Available: http://dataone-python.readthedocs.io/en/latest/gmn/setup-local-d1-stack.html. [Accessed: 06-Feb-2017].

[13] "DataONE Python Products," *DataONE Python Products*, 22-Feb-2017. [Online]. Available: https://media.readthedocs.org/pdf/dataone-python/latest/dataone-python.pdf. [Accessed: 07-Feb-2017].

[14] "Registered DataONE nodes. Includes Coordinating Nodes (CN) and Member Nodes (MN).," *Registered DataONE nodes. Includes Coordinating Nodes (CN) and Member Nodes (MN)*. [Online]. Available: https://cn.dataone.org/cn/v1/node. [Accessed: 06-Feb-2017].

[15] "Data Publication," *Universities Allied For Water Research*, 2017. [Online]. Available: https://www.cuahsi.org/data-publication. [Accessed: 07-Feb-2017].

[16] C.S.D.M.S., "About CSDMS," *CSDMS: Community Surface Dynamics Modeling System. Explore Earth's surface with community software, 26-Jan-2017*. [Online]. Available: http://csdms.colorado.edu/mediawiki/index.php?title=About_CSDMS&oldid=119937. [Accessed: 23-Feb-2017].