University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

DataONE Sociocultural and Usability & Assessment Working Groups

Communication and Information

1-2019

WHITE PAPER: Environmental Scan for DataONE

Amy Louise Forrester
University of Tennessee, Knoxville, aforres4@utk.edu

Suzie Allard
University of Tennessee - Knoxville, sallard@utk.edu

Leah Cannon
University of Tennessee, Knoxville, lcannon8@vols.utk.edu

Danielle Pollack

Alison Specht
Centre for the Synthesis and Analysis of Biodiversity

Follow this and additional works at: https://trace.tennessee.edu/utk_dataone

 Part of the Library and Information Science Commons

Recommended Citation

Forrester, Amy Louise; Allard, Suzie; Cannon, Leah; Pollack, Danielle; and Specht, Alison, "WHITE PAPER: Environmental Scan for DataONE" (2019). *DataONE Sociocultural and Usability & Assessment Working Groups.*
https://trace.tennessee.edu/utk_dataone/150

This Creative Written Work is brought to you for free and open access by the Communication and Information at Trace: Tennessee Research and Creative Exchange. It has been accepted for inclusion in DataONE Sociocultural and Usability & Assessment Working Groups by an authorized administrator of Trace: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

January 2019

WHITE PAPER: Environmental Scan for DataONE

This environmental scan (conducted by the U&AWG in fall 2018) features a multi-faceted analysis of projects/initiatives in the DataONE space. This analysis will help DataONE leadership better understand the existing competitive ecosystem. Assessing DataONE's place in this broader environment can provide valuable information and insight to inform the transition from a project to a sustainable organization.

This environmental scan identified 27 organizations in the DataONE space whose missions align loosely with the DataONE mission (Table 1). The authors used data and information mined from searches of the Internet to capture details of these cases. The information and data collected is limited by access and availability of publicly available documentation.

This report (1) provides context by identifying organizations in the data space; (2) analyzes those organizations most similar to DataONE regarding key services and products; and (3) explores the data training/education environment. As appropriate, the report offers key insights derived from the analysis.

Table 1. Alphabetical list of organizations scanned

4tu.ResearchData	Figshare	PANGAEA
ANDS	GBIF	SciVerse
Apollo	Globus	Scopus
Center for Open Science	Google Data Search beta	Sead 2.0
Chorus	ICPSR	Springer Nature Research Data Services
Data Conservancy	IEDA	TRY
Dataverse Network	iRODS	Zenodo
Digital Preservation Network	Mendeley data search	
Dryad	NCEI	
Elsevier DataSearch beta	OpenAIRE	

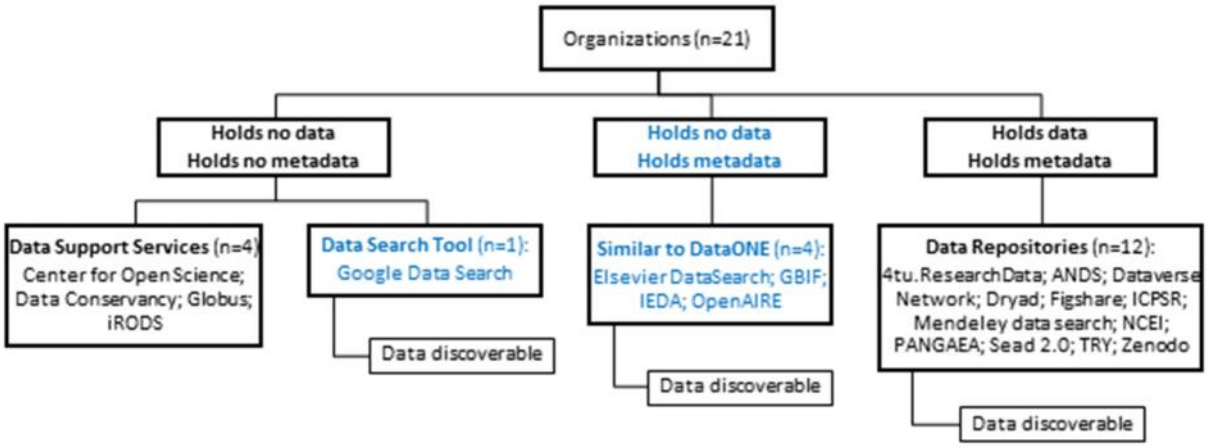
The sample was culled to remove the following six organizations for a total of 21 organizations to analyze in the DataONE space.

- Digital Preservation Network – disbanding
- Literature/publisher databases - (Chorus, Sciverse, Scopus, Springer Nature Research Data Services)
- Apollo – Institutional Repository

To identify those organizations most similar to DataONE, the data was parsed by holdings/storage of data and metadata (Figure 1). Occurrence of organizations with data search portals was also assessed.

Four organizations are comparable to DataONE in that they hold metadata, but *not* data. With the exception of Google Data Search, which can be categorized as a data search tool, the organizations that hold neither data nor metadata can be categorized as support services, providing infrastructure or tools for data management to further open science. The twelve organizations that hold both data and metadata can be categorized as data repositories.

Figure 1. Parsed by (meta)data holdings



INSIGHT #1:
DataONE is in a space with a limited number of competitors but some are quite formidable in terms of funding and longevity. While there may be direct competition, DataONE has already established collaborations and there is potential to establish additional relationships.

Profiles of Organizations Most Similar to DataONE

The five organizations (Table 2) most similar to DataONE are not primary data repositories, but provide discoverability and access to data hosted elsewhere and serve as federated indexes to external data. Three operate as non-profit, non-governmental organizations. IEDA and OpenAIRE are funded by National Science Foundation and European Commission respectively. GBIF receives financial support through their voting membership fees.

Elsevier DataSearch and Google Data Search are both commercial operations in beta mode. Both are self-funded via their parent enterprises and Elsevier DataSearch is exploring how to integrate into their other products, e.g., Mendeley Data.

The key products and services outlined at the 2018 All Hands Meeting (AHM) by Matt Jones and Dave Vieglas were the basis for the analysis of these organizations. These products and services were discussed for their capacity and potential as revenue streams.

Table 2. Profile details of similar organizations

Organization	Active Date	Organization Model	Funding Model	Usage Reporting	Data Replication	Data Quality
DataONE	**2008**	**Academic project**	**NSF**	**√**	**√**	
IEDA	2017	NP-NGO	NSF	√		*
GBIF	2001	NP-NGO	Voting Membership fees	√		
OpenAIRE	2010	NP-NGO	European Commission	√		*
Elsevier DataSearch beta	2016	Commercial	Not yet determined			
Google Data Search beta	2018	Commercial	Private			

Organization model, a key issue of sustainability, identifies the fiduciary responsibility and subsequently informs how finances are managed. Academic projects operate as a cost center under the responsibility of an academic institution. NP-NGO is an independent of government, not-for-profit organization, although may be funded by governmental organizations. A commercial organization sells goods and services for a profit.

I. **Usage Reporting:** The ability to provide and view data usage. Metrics are important at the DataONE portal level, the Member Node repository level, and the individual dataset (or data object) level.

DataONE is Counter compliant and provides dataset level metics on citations, downloads, and views for all time and by month. View and download data is also available for individual data objects within a dataset.

IEDA provides data on the usage of their individual portals/repositories to users (https://www.iedadata.org/community/usage/). This includes monthly totals for unique IP address visits and unique data downloads. Individual dataset usage is not publicly visible.

OpenAIRE gathers data on metadata views and downloads. OpenAIRE follows Release 4 of the COUNTER Code of Practice and uses the Matomo Open Source Analytics platform (http://matomo.org) to track usage activity. Repository level usage statistics are provided as a service for participating repositories. Usage metrics are collected on the individual item level and displayed in the data record. However, no record containing data could be found.

GBIF provides metrics regarding species occurrences within their datasets. Dataset downloads are recorded as "activity." GBIF also links literature citations to their datasets.

II. **Data Replication**: The capacity to store data in more than one site. It is useful in improving the availability of data.

DataONE currently provides replication services for its Member Nodes.

None of the five organizations discuss data replication within their products or publicly available documentation.

III. **Data Quality:** An assessment of data properties as a way of measuring a perceived set of values, e.g., fitness, quality, completeness

DataONE does not provide any measure of data quality.

IEDA has one repository, MGDS (Marine Geoscience Data System), that provides a numerical indication of individual dataset quality, e.g.,

Quality 0: Data have not been processed or modified since acquisition
Quality 1: A level of processing has been undertaken, ensuring quality control

OpenAIRE does not provide any public indication of data quality. However, for the repository or data manager, services are provided to "run compatibility tests against the OpenAIRE Guidelines for Data Archives."

INSIGHT #2:
DataONE is well positioned to concentrate on these service areas that currently exist or are extensions of on-going work. While several of these services are present in the competitive landscape, they do not appear to be well-established

Training

Interest in data science training was also highlighted at the AHM as a potential service/product for DataONE's sustainability. Analysis was performed on all 21 organizations to fully understand the training landscape across a full spectrum of the ecosystem.

Only eight organizations had an identifiable section or link to "training" on their website (Table 3). As language differs between UIs, for a multitude of reasons, a description of the resource was collected—verbatim where possible. Further information was collected on training topics and/or audience when the description was lacking the detail.

Table 3. Training Resources

Organization	Scan Category	Website Section/Link Title	Resource Description	Topics - Audience
4tu.ResearchData	Data repository	training & events	provides trainings, sessions and presentations on various aspects of data management.	Data management -- data-support staff PhD students
ANDS	Data repository	Skills section	● A collection of pages and resources for training and teaching data and research data management skills o Domain specific o Data trainers o Technical skills ● Five-day Data in the Scholarly Communications Life Cycle Course	
ICPSR	Data repository	Teaching & learning	● Undergraduate data-driven learning materials ● Summer Program in Quantitative Methods of Social Research	
NCEI*	Data repository			1. "data management"
IEDA	Similar to DataONE	Educator Resources	Tutorials and workshops to earth science learning modules for K-12 and undergraduate students	
OpenAIRE	Similar to DataONE	Training	● Webinars: current topics categorized by audience (e.g., content providers, funders, research librarians) ● Workshops: various topics	Data management Data curation
Center for Open Science	Support services	Training Services	● Workshops: increase the reproducibility and transparency of their work	

			• Webinars: related to open, reproducible research	
Data Conservancy	Support services	Education	• Webinars & workshops: data management • Classes: Data Curation	

*Further data collection needed for NCEI. NCEI website not available at this time due to a lapse in appropriation.

DataONE has both proximity to the data and the data creators through its activities with metadata and search. There are not many training organizations that are as close to the data (i.e., similar to DataONE or data repository) that are providing training services to their users. OpenAIRE stands out as being most similar to DataONE and also offers an abundance of webinar trainings. Only two of the four organizations identified as support services deliver outreach training.

INSIGHT #3:
"Data proximity" is DataONE's competitive advantage. Unlike training-only focused organizations (e.g., Data Carpentry, Lynda.com) and support services, DataONE can leverage the insight and foresight gained by being connected to the data through rich cyberinfrastructure. Thus enabling DataONE to deliver training that is cutting edge and more meaningful than offered by the competition.

Submitted by:

Amy Forrester (Report lead)
Suzie Allard
Leah Cannon
Danielle Pollack
Alison Specht
+ input from U&AWG

Data availability: Google - https://docs.google.com/spreadsheets/d/1D2D2PLt24vsbaclCO3kHFc-yIQUYiYLPer6mBQANXHg/edit?usp=sharing