DataONE NEWS

Volume 7 Issue 1

©2018 DataONE 1312 Basehart SE University of New Mexico Albuquerque NM 87106

Designing Open Science

The news has been full of data- and publishing-related headlines of late from Elsevier and Google announcing their versions of data search to Dryad's partnership with California Digital Library to many universities that are protesting and unbundling from high-cost publication packages. We will touch upon many of these topics in future newsletters and Dave Vieglais briefly discusses the relationship of Google Search to DataONE later in this newsletter (page 3).

One of the most anticipated headlines was the announcement that the National Academies Press had released its seminal report[1] that offers perspectives on the value of and obstacles to open science as well as five recommendations for accelerating progress in open science. The authors of the report note that "openness and sharing of information are fundamental to the progress of science and to the effective functioning of the research enterprise. The advent of scientific journals in the 17th century helped power the Scientific Revolution by allowing researchers to communicate across time and space, using the technologies of that era to generate reliable knowledge more quickly and efficiently. Harnessing today's stunning, ongoing advances in information technologies, the global research enterprise and its stakeholders are moving toward a new open science ecosystem. Open science aims to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms, that were used to generate those data."

The report highlights the many benefits of open science which should come as no surprise to DataONE users, including "increased rigor and reliability, the ability to address new questions, faster and more inclusive dissemination of knowledge, broader participation in research, effective use of resources, improved performance of research tasks, and open publication for public benefit." The report also notes that there are many existing challenges to open science with respect to "costs and infrastructure; structure of scholarly communication; lack of supportive culture, incentives and training; and privacy, security, and proprietary barriers to sharing."

The report postulates that the open science movement is at an important inflection point due, in part, to the plethora of new information technology tools and services that can revolutionize science practice, as well as the array of new policies and initiatives from organizations, funders and publishers. The report convincingly argues that "all phases of the research process provide opportunities for assessing and improving the reliability and efficacy of scientific research" including exploration of research resources, developing and sharing research plans, collecting data and generating knowledge, study reproducibility and replication, and preserving and disseminating research outputs.

Importantly, the report presents five recommendations for accelerating progress in open science:

1. "Research institutions should work to create a culture that actively supports Open Science by Design by better rewarding and supporting researchers engaged in open science practices. Research funders should provide explicit and consistent support for practices and approaches that facilitate this shift in culture and incentives.
2. Research institutions and professional societies should train students and other researchers to implement open science practices effectively and should support the development of educational programs that foster Open Science by Design.
3. Research funders and research institutions should develop the policies and procedures to identify the data, code, specimens, and other research products that should be preserved for long-term public availability, and they should provide the resources necessary for the long-term preservation and stewardship of those research products.
4. Funders that support the development of research archives should work to ensure that these are designed and implemented according to the FAIR data principles. Researchers should seek to ensure that their research products are made available according to the FAIR principles and state with specificity any exceptions based on legal and ethical considerations.
5. The research community should work together to realize Open Science by Design to advance science and help science better serve the needs of society."

There is clearly much work to be done in order to realize the open science environment envisioned in the report and that is shared by many, if not most, DataONE users and stakeholders. In this and subsequent newsletters, we will be highlighting packages of DataONE tools and services that address many of the existing limitations to open science. We will also be seeking your input on how to more deeply engage users and other stakeholders in the DataONE community and in its decision-making so that key tools and services can continue to accelerate open science. ■

— William Michener
Principal Investigator, DataONE

[1] National Academies of Sciences, Engineering, and Medicine. 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/25116.

MemberNodeDESCRIPTION

In each newsletter issue we highlight one of our current Member Nodes. The full list of Member Nodes and summary metrics can be found on the DataONE.org site at bit.ly/D1CMNs.

Environmental System Science Data Infrastructure for a Virtual Ecosystem

https://ess-dive.lbl.gov/

ESS-DIVE (Environmental System Science Data Infrastructure for a Virtual Ecosystem) launched in April 2018 to serve as a repository for hundreds of U.S. Department of Energy (DOE)-funded research projects focusing on terrestrial and subsurface ecosystems. Under the DOE's Environmental System Science umbrella, which includes the Subsurface Biogeochemical Research and Terrestrial Ecosystem Sciences programs, the digital library also serves datasets that were previously stored in DOE's Carbon Dioxide Information Analysis Center archive.

Environmental system scientists study vital processes like nutrient and water cycling, or carbon and energy fluxes which rely on high-quality, unreproducible and diverse observational datasets collected over years. ESS-DIVE stores the critical data generated by environmental field, experimental, and modeling activities with their metadata to allow researchers to examine and predict long-term ecosystem and watershed behaviors. These data require intensive fieldwork to collect and no field season is ever going to be the same because the environmental conditions are always changing. As a result, every single data point is incredibly valuable and has the potential to be reused for purposes that go way beyond the intent for which it was originally collected.

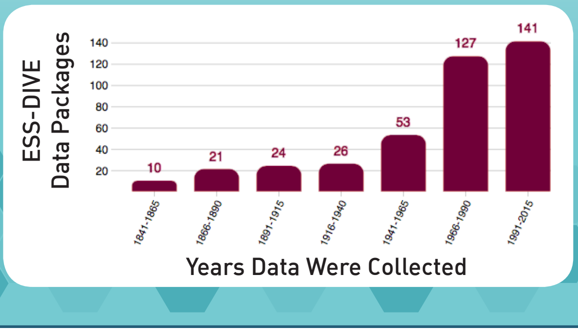
Designed as a scalable framework, ESS-DIVE incentivizes data providers to contribute well-structured, high-quality data and enables the user community to easily build data processing, synthesis, and analysis capabilities using those data. ESS-DIVE includes functionalities that allow researchers to upload their data and keep it proprietary as they work on the paper. Once the paper is published, researchers can go into the library and easily make their data public. ESS-DIVE also provides tools that allow users to access data, contribute and publish data and track data downloads. Each data package is assigned a unique Digital Object Identifier (DOI) when it is uploaded allowing researchers to cite the dataset that they used. All basic components of the architecture run inside Docker containers so that users can upload and interact with data in a controlled environment, in addition to facilitating multiple redundant instances of ESS-DIVE.

By joining DataONE, ESS-DIVE reinforces its mission to preserve, expand access to, and improve usability of critical data. Deb Agarwal, a scientist in Berkeley Lab's CRD and lead of the ESS-DIVE project, explains that becoming a DataONE Member Node will make ESS-DIVE "an even more powerful tool, as the library's DOE-funded data contents will be discoverable in cross-catalogue searches."

ESS-Dive was built by a collaboration of scientists in Lawrence Berkeley National Laboratory's (Berkeley Lab's) Computational Research Division (CRD) and Earth & Environmental Sciences Area (EESA), the National Energy Research Scientific Computing (NERSC) and digital librarians at the National Center for Ecological Analysis and Synthesis (NCEAS)—a research center based at UC Santa Barbara.

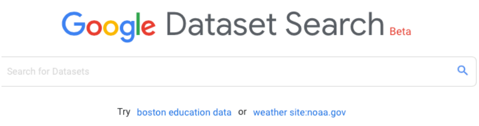
ESS-DIVE Brief

233 Datasets
28 GB of content
Over **150** years of data



CyberSPOT

Where does Google Dataset Search take us?


Google Dataset Search Beta
Search for Datasets
Try boston education data or weather site:noaa.gov

On 5 September 2018, Natasha Noy of Google AI wrote they had "launched Dataset Search, so that scientists, data journalists, data geeks, or anyone else can find the data required for their work"[1].

This commitment by Google to support and promote open sharing of structured data represents another step in the continually evolving path towards widely adopted principles of findable, accessible, interoperable and reusable data.

DataONE is committed to supporting long term access to interoperable and reusable data and is continually striving to simplify and enhance the capabilities of the infrastructure it provides. A recent comparison of searches on earth science topics against both Google Dataset Search and the DataONE search interface at search.dataone.orgshowed that both mechanisms provided benefits to users in terms of results found and the precision of the matches, though additional information such as dataset provenance support is indicative of enhancements supporting earth science research available through the DataONE federation.

DataONE has provided discovery of structured earth science data since 2012 for repositories participating as Member Nodes in the federation[2]. Other efforts have provided similar capabilities to varying degrees of success, and in most cases have served the important goal of enabling researchers to do better science by helping find and access relevant data resources. One of the key challenges faced by these efforts is ensuring data repositories have a consistent, reliable way to expose the data resources as well as describe those resources. Protocols such as OAI-PMH[3], CSW[4] and the DataONE Member Node API[5] ensure consistent programmatic access. Metadata standards such as EML[6], ISO-19115[7], Dublin Core[8], and the widely used venerable FGDC[9] provide community adopted mechanisms for describing data resources. Scientific search portals such as DataONE Search build upon infrastructure that leverages these protocols and standards to enable high quality search and retrieval systems.

The underlying patterns and standards utilized by Google Dataset Search[10] leverage the same general patterns that have been used by web crawlers and indexers for populating internet search engines for many years. These approaches generally utilize "robots.txt" files[11] to guide crawlers to XML "sitemap" documents[12], that in turn identify web pages and other resources for indexing. This approach to locating resources for indexing has proven successful and generally applicable across Internet accessible resources.

The indexers would then employ sophisticated algorithms and information mining approaches to extract and index the web page content, an enormous challenge given the diversity of web accessible resources[13]. Indexing the rich, diverse, unstructured content of the World Wide Web has been achieved to varying degrees of success, however is generally unsatisfying when trying to locate highly specific resources desired for scientific analyses.

The emergence of schema.org[14] has provided broadly applicable guidance for including consistent, structured markup in web accessible resources. This in turn enables indexers to reliably index and categorize content, and so provide search interfaces to users that can yield high the precision and recall required for scientific research purposes. Schema.org provides some direct support for describing scientific resources and may be extended to include additional properties as desired by the content publisher, or alternatively the data set web landing page can include a reference to the full dataset

UpcomingEVENTS

Members of the DataONE Team will be at the following events.
Full information on training activities can be found at *bit.ly/D1Training* and our calendar is available at *bit.ly/D1Events*.

Oct. 11-12
FORCE2018
Montreal, Canada
https://www.force11.org/meetings/force2018

Nov. 5-8
RDA 12th Plenary
Botswana, Africa
https://www.rd-alliance.org/plenaries

Nov. 5-9
International Data Week
Botswana, Africa
http://internationaldataweek.org

Dec. 10-11
CNI
Washington, DC
https://www.cni.org/mm/fall-2018

Dec. 10-14
AGU
Washington, DC
https://fallmeeting.agu.org/2018/

Feb. 4-7
IDCC
Melbourne, Australia
http://www.dcc.ac.uk/drupal/events/idcc19

Apr. 2-4
RDA 13th Plenary
Philadelphia, Pennsylvania
https://www.rd-alliance.org/plenaries

CyberSPOT cont'd

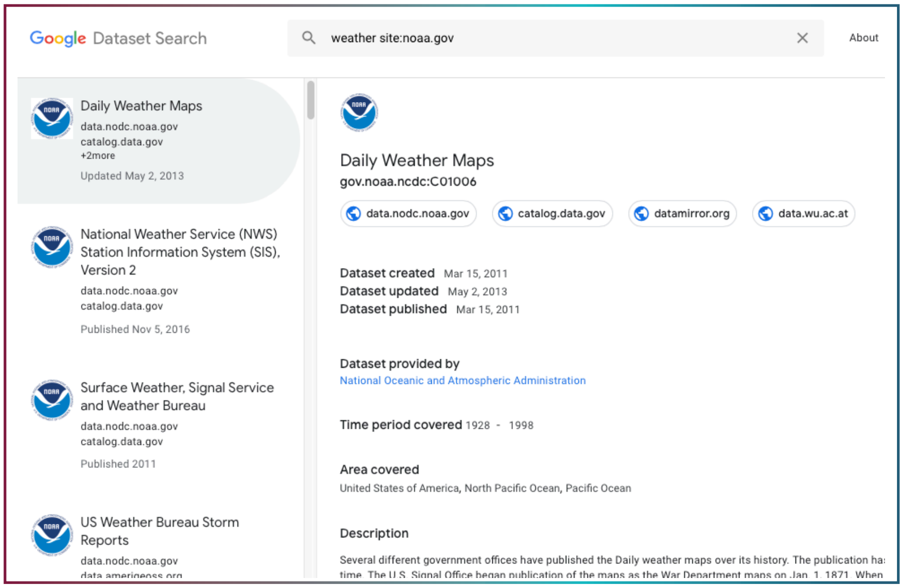
metadata expressed for example in the richer ISO-19115 or EML standard.

The sitemap + schema.org approach for exposing and describing structured data resources is an important evolutionary step towards open science data sharing and the EarthCube Project 418[15, 16] has clearly demonstrated how this approach can be leveraged for sharing sophisticated resources. This approach is being incorporated into DataONE at multiple levels for both data indexing and as a mechanism for more broadly exposing resources indexed from repositories not already supporting schema.org. The DataONE infrastructure expands on the information available through schema.org by also supporting indexing and discovery of related datasets described with web provenance standards PROV-O[17] and the ProvONE enhancements[18]. Further enhancements such as semantic measurement search which enables more precise search against measurement descriptions are being progressively rolled out in the DataONE infrastructure.

The announcement by Google backing schema.org and structured data sharing will provide greater commercial incentive for adoption and so also result in more resources and tools available to work with this approach. While Google and other commercial enterprises continue to promote broad solutions to data discovery and sharing, DataONE will continue to expand, enhance, and refine the rich discovery capabilities offered by the federation. By leveraging the light-w[1] [2] [3] [4] eight schema.org approach while also exposing the richer, standards compliant metadata available from Member Nodes participating in the federation, DataONE provides a continually improving rich discovery mechanism tailored to the demands of earth science researchers.■

—David Vieglais
Director for Development and Operations, DataONE

[1] https://www.blog.google/products/search/making-it-easier-discover-datasets/
[2] https://search.dataone.org
[3] https://www.openarchives.org/pmh/
[4] http://www.opengeospatial.org/standards/cat
[5] https://purl.dataone.org/architecure/apis/MN_APIs.html
[6] https://knb.ecoinformatics.org/external/emlparser/docs/index.html
[7] https://webstore.ansi.org/RecordDetail.aspx?sku=ISO%2019115-1:2014
[8] http://dublincore.org/
[9] https://www.fgdc.gov/
[10] https://toolbox.google.com/datasetsearch
[11] http://www.robotstxt.org/
[12] https://www.sitemaps.org/protocol.html
[13] https://link.springer.com/article/10.1134/S0361768816050078
[14] https://schema.org
[15] https://www.earthcube.org/
[16] https://www.earthcube.org/group/project-418
[17] https://www.w3.org/TR/prov-o/
[18] http://purl.org/provone



weather site:noaa.gov

Daily Weather Maps
data.nodc.noaa.gov
catalog.data.gov
+2more
Updated May 2, 2013

National Weather Service (NWS) Station Information System (SIS), Version 2
data.nodc.noaa.gov
catalog.data.gov
Published Nov 5, 2016

Surface Weather, Signal Service and Weather Bureau
data.nodc.noaa.gov
catalog.data.gov
Published 2011

US Weather Bureau Storm Reports
data.nodc.noaa.gov
data.ameriqeoss.org

Daily Weather Maps
gov.noaa.ncdc:C01006
data.nodc.noaa.gov | catalog.data.gov | datamirror.org | data.wu.ac.at

Dataset created Mar 15, 2011
Dataset updated May 2, 2013
Dataset published Mar 15, 2011

Dataset provided by
National Oceanic and Atmospheric Administration

Time period covered 1928 - 1998

Area covered
United States of America, North Pacific Ocean, Pacific Ocean

Description
Several different government offices have published the Daily weather maps over its history. The publication ha:
time. The U.S. Signal Office began publication of the maps as the War Department maps on Jan. 1, 1871. When

Google Dataset Search interface

1312 Basehart SE
University of New Mexico
Albuquerque, NM 87106
Fax: 505.246.6007

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement.

Project Director:
William Michener
wmichene@unm.edu
505.814.7601

Executive Director:
Rebecca Koskela
rkoskela@unm.edu
505.382.0890

Director of Community Engagement and Outreach:
Amber Budden
aebudden@dataone.unm.edu
505.205.7675

Director of Development and Operations
Dave Vieglais
dave.vieglias@gmail.com

Upcoming Webinar

Community Directed Resources for Data Management
Tuesday, October 9, 2018
9 am Pacific / 12 noon Eastern

Information and registration at:
https://www.dataone.org/upcoming-webinar

DataONE NEWS

ByTheNUMBERS

DATA DISCOVERABLE THROUGH DATAONE

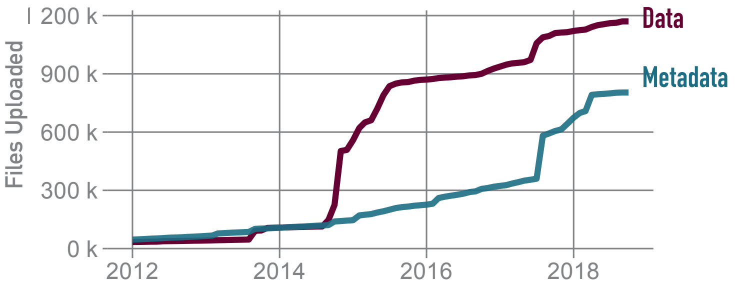
48 TB of content

804 K metadata

1.17 M data

2,359	Visitors to our search page*	+ 66
1,140	Searches conducted*	+ 29
99.99%	Uptime of Coordinating Nodes	

*metrics are running monthly averages; symbol denotes change since last quarter



SOURCE: CN.DATAONE.ORG
Only the first version of each file is counted

OUR COMMUNITY

41 Contributing Members

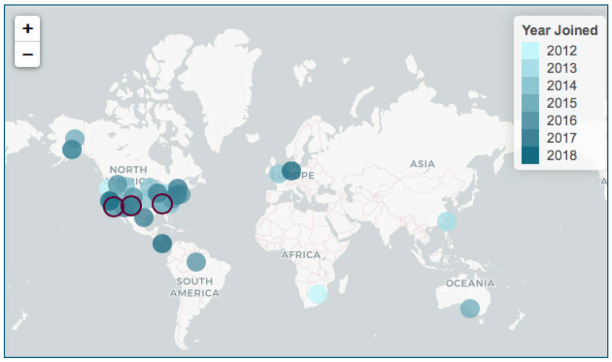
New Member this Quarter

ENVIRONMENTAL SYSTEM SCIENCE DATA INFRASTRUCTURE FOR A VIRTUAL ECOSYSTEM

ESS-DIVE Deep Insight for Earth Science Data

500	DataONE User Group members
5,300+	Users trained

Repositories in the DataONE Federated Network



EDUCATION AND OUTREACH

Webinar Series

32	Webinars
93	Average number of attendees
2977	Unique webinar attendees

Most Downloaded Resources

1 Data Managment Plan Example for NSF

2 Best Practices Primer

3 Data Management Plan Example from Manua Loa

Education Resources

19,087	Visits to the public webpage*	+ 406
217	Education Module downloads*	=

*metrics are runniing monthly averages; symbol denotes change since last quarter

Most Visited Pages

1. Education Module homepage
2. Data Management Planning
3. Best Practices: Create and document data backup policy