DataONE: A Data Federation with Provenance Support

Yang Cao[1], Christopher Jones[2], Víctor Cuevas-Vicenttín[3], Steve Aulenbach[4], Matthew Jones[2], Bertram Ludäscher[1], Timothy McPhillips[1], Paolo Missier[5], Christopher Schwalm[6], Peter Slaughter[2], Dave Vieglais[2], Lauren Walker[2], and Yaxing Wei[7]

[1] Graduate School for Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC),
[2] National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara,
[3] TBD,
[4] University Corporation for Atmospheric Research (UCAR) and U.S. Global Change Research Program (USGCRP),
[5] School of Computer Science, University of Newcastle, UK,
[6] Woods Hole Research Center, Falmouth MA 02540, USA,
[7] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Abstract. DataONE (Data Observation Network for Earth) is a federated data network focusing on earth and environmental science data. In this demonstration we show the new provenance capabilities that have been added to the DataONE toolkit: a user (say, "Alice") can annotate a script (e.g., in Matlab or R) using the YesWorkflow (YW) tool to model the script's prospective provenance. After Alice has run the script, the result files, the script and its YW prospective provenance, along with the runtime (retrospective) provenance are bundled into an OAI-ORE compliant data package and uploaded to the DataONE network. A second user ("Bob") can discover and access Alice's package and use her data for his own script-based computations. We show that Bob's results, once published through DataONE link back to Alice's outputs via unique identifiers. Thus, a third user ("Charlie") who browses DataONE will be able to discover the full provenance of Bob's results, all the way back to Alice's original contributions.

Keywords: We would like to encourage you to list your keywords within the abstract section

1 Introduction

DataONE (Data Observation Network for Earth) is a federated data network and a sustainable cyberinfrastructure for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data [3]. There are five primary goals to DataONE: discovery and access, data integration and synthesis, education and training, building community, and data sharing.

1.1 DataONE Architecture

There are three principal components in the DataONE infrastructure:

1. *Member Nodes* represent existing or new data repositories that support the DataONE Member Node APIs. A Member Node functions completely when it can support the required services interfaces for Tier 1 participation (i.e. public access, read only content): discovery of all objects available on the Member Node, low level description of each object, retrieval of the object given its identifier, and reporting activity. Member Nodes also provide science metadata and relationships between metadata and data using resource maps [11] to facilitate discovery through the DataONE Search interfaces.
2. *Coordinating Nodes* serve data management and discovery needs of the network. These services include network-wide indexing of scientific data objects, data replication between Member Nodes, mirrored content of *science metadata* (detailed descriptions of science data objects and collections) and *system metadata* (low level metadata describing the type, size, ownership, and locations of data) present at Member Nodes.
3. The *Investigator Toolkit* contains a set of user tools that enables interaction with DataONE infrastructure through the REST service APIs exposed by the Coordinating and Member Nodes. Low level libraries are initially available in Python and Java for application developers, and DataONE support is present in scientific analysis tools such as Matlab and R.

The DataONE infrastructure was released for public use in 2012 and supports identifier resolution, content search and retrieval, the federated identity management infrastructure, and the replication service.

1.2 DataONE Search

DataONE Search is a web-based application allowing users to seamlessly and efficiently discover publicly accessible data packages within the DataONE federated network of Member Node repositories. It was released in November 2015, and allows users to search by map, dates, keywords and other facets and to refine the results using further parameters. Users can sign in to DataONE Search using ORCID credentials, Google accounts, and institutional accounts.

The software has been designed to facilitate rapid iteration and deployment of new features and to take full advantage of future capabilities offered by upcoming versions of the core cyberinfrastructure. Notable among these is the *provenance information* within search and discovery. In the next section, we will demonstrate the DataONE provenance tools and the visualization of provenance with DataONE Search.

2 Demonstration Description

DataONE have designed a tool (named “run manager”) for capturing the prospective [ ] and the retrospective provenance [ ] traces of R and Matlab script executions [5][10] to provide a complete understanding of a scientific workflow script in terms of script structure and script execution. The automated provenance collection is transparent to scientists and provides a medium-grained data dependencies by focusing on file inputs and outputs.

We will use an example to introduce the core provenance and search features of DataONE. Suppose that Alice is an earth scientist whose work involves generating global grass fraction maps from water, air temperature, and other input data. Bob is a scientist who runs a grassland water use efficiency analysis built upon Alice's output. Charlie is a user of DataONE.

2.1 User Alice's view: Local Run Manager View

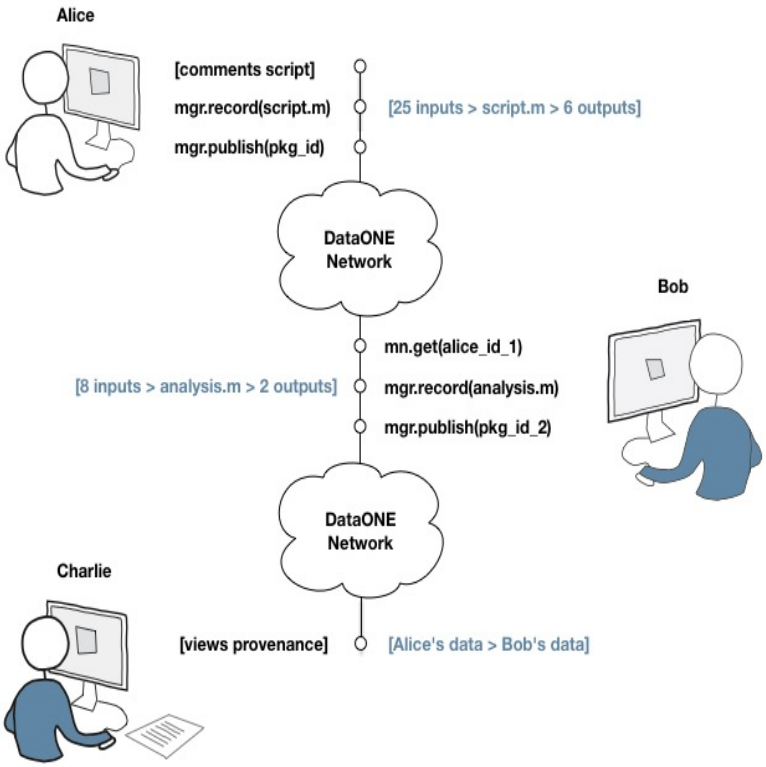

Fig. 1: Example use of DataONE Provenance Tool (Run Manager): (1) Alice runs her script with Run Manager to create a data package (P_A) and publish to DataONE network; (2) Bob downloads Alice's data, uses them in his script and creates another data package (P_B) using Run Manager; (3) Charlie browses data packages P_A and P_B and is able to navigate back to P_A via P_B.

In Figure 1, Alice has developed a script for producing Carbon3/Carbon4 soil maps. She uses YW to mark-up the code and expose the underlying workflow view (prospective provenance). Alice then ran her script using the DataONE provenance capture tool, or Run Manager, in Matlab. After a couple of runs, Alice is happy with the results and publishes them to the DataONE network. To do so, she bundles up the data results

along with the runtime (retrospective) provenance captured by the DataONE Matlab provenance recorder, the script itself and its YW-generated workflow view.

When searching DataONE for grass fraction datasets, Bob finds Alice's data package [2]. Bob uses outputs from Alice for his own study, which he also shares via DataONE. Since he used the unique identifiers from Alice's outputs, the provenance traces produced by Alice's and Bob's runs are naturally connected, through the nodes that carry those data identifiers. Charlie discovers Bob's data packages on DataONE and is able to navigate back to the data that Bob used, i.e., Alice's data package depicted in Figure 3 [6][9].

2.2 User Bob's view: Local Run Manager View

When Bob searches for data in DataONE and finds Alice's data and code, he uses the Matlab Toolbox to fetch the data in a way that is identifier-aware, rather than having to manually maintain the provenance information. For example, he calls the **MemberNode.get**(session, identifier) function in his analysis script because this function has provenance tracking capabilities. Then, the Toolbox downloads the data and uses the same identifier to store the data locally. Additionally, a *prov:used* statement that Bob is using Alice's data is recorded.

When Bob then calls **publish**() to upload his scripts and data, data in his package that are not present in the DataONE network will be uploaded. When all data are within the DataONE network, their respective provenance statements are indexed on the **Coordinating Node**, and are available to be used in the web provenance display.

2.3 User Charlie's View: DataONE Search Website

When Charlie searches the keyword "grass" in the DataONE Search, two data packages are found and shown in Figure 2. One data package is created by Yaxing Wei (Alice) [2] and the other data package is created by Christopher Schwalm (Bob) [1]. Both data packages show that provenance information is associated with each dataset (via the icon in the search record) and can be seen at DataONE Search demo site [4].

Prospective and retrospective provenance information can be explored from the DataONE Search site. Provenance details for any input or output in the provenance graph can be viewed by clicking on the icon shown on Figure 3. DataONE Search provides human language descriptions of how data are used or generated via the script and models, and provides navigation. Figure 3 shows two data packages produced by the execution of Alice and Bob's workflow scripts, and the provenance lineage between these two data packages.

Alice's data package [2] is shown on the top layer of Figure 3, created when Alice ran her script using the DataONE provenance capture tool, or Run Manager, in Matlab. Her script (C3_C4_map_present_with_comments.m) takes twenty-five input files and produces six outputs, which are shown on the left side and right side of Alice's data package in Figure 3, respectively. The bottom three outputs in Alice's data package are the NetCDF data files that represent three different world map grids of percentage of grass types (C3 grass fraction, C4 grass fraction, and total grass fraction). In addition, a model graph is displayed at the intermediate layer [7]. Alice's workflow script used

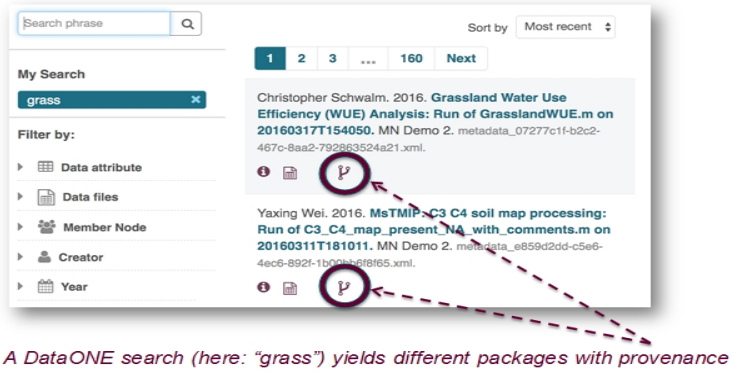

Fig. 2: The search result page for keyword "grass" on a DataONE Search demo site

embedded YesWorkflow annotations to document her script. These annotations declare step by step how data are used and derived in the script.

Bob's data package [1] created by DataONE provenance tool (Run Manager) is displayed at the bottom layer. When Bob browses Alice's data package on the DataONE Search site, he decides to use three NetCDF output data produced by Alice's work in his Grassland Water Use Efficiency Analysis script (GrasslandWUE.m). From the provenance information associated with Bob's data package in Figure 3, we see that it takes eight inputs, and produces two visualizations. By viewing the details for each input, we can see that three of them are the outputs produced by Alice's data package.

In order to maintain the link between the outputs from Alice's data package to the inputs to Bob's package, Bob needs to use the same identifier in his script as Alice does. An alternative approach is that Bob creates a new identifier when he downloads Alice's data. However, the link between two data packages will be broken which has been discussed in [8].

3 Discussion and future work

This paper demonstrates a notable new feature (provenance capture and search) of DataONE. DataONE has released the DataONE Search for public use in November 2015, and the R and Matlab provenance tool to the public in 2016.

Maintaining the link between Alice's work and Bob's subsequent work is worth to be discussed. Currently Bob must use certain functions provided by DataONE to make the provenance link to Alice's work to be correctly maintained. There are other possible ways to achieve the same goal. For example, Bob can download Alice's data and manually add the links to Alice's work back into his data package before sharing via DataONE. A *prov:was_derived_from* statement that local data copy for Bob is derived from Alice's data is easily to be added.

Future work of the current provenance tools include: (1) solve the broken link use case; (2) support more I/O functions; (3) handle complex scenarios such as multiple

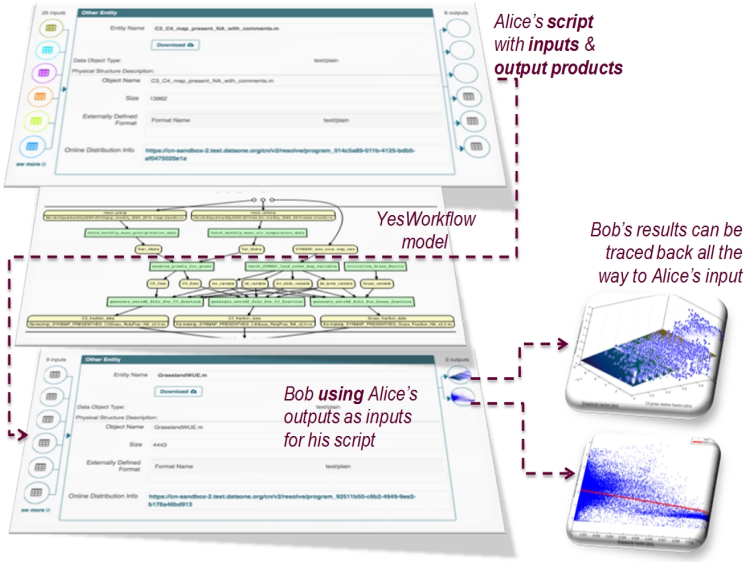

Fig. 3: User Charlie's view on DataONE Search site: (1) A combined-view YesWorkflow model for Alice's soil processing workflow script; (2) Provenance lineage between Alice's data package and Bob's data package; (3) Two visualizations produced by Bob's water use efficiency analysis script.

runs of a script and multiple users; (4) variable-level provenance capture warrants further investigation and efforts.

References

1. Christopher Schwalm's DataPackage, `https://search-sandbox-2.test.dataone.org/#view/metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml`
2. YaXing Wei's DataPackage, `https://search-sandbox-2.test.dataone.org/#view/metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml`
3. Data Observation Network for Earth (DataONE), `https://www.dataone.org`
4. DataONE Search Demo Site, `https://search-sandbox-2.test.dataone.org`
5. Jones C., Cao Y., Slaughter P., Jones M.: Matlab DataONE Toolbox. `https://github.com/DataONEorg/matlab-dataone` (2016)
6. Katz, D.S. & Smith, A.M.: Implementing Transitive Credit with JSON-LD. Journal of Open Research Software. 3(1), p.e7. (2015)
7. McPhillips T., Song T., Kolisnik T., Aulenbach S., Belhajjame K., Bocinsky R.K., Cao Y., Cheney J., Chirigati F., Dey S., Freire J., Jones C., Hanken J., Kintigh K.W., Kohler T.A., Koop D., Macklin J. A., Missier P., Schildhauer M., Schwalm C., Wei Y., Bieda M., and Ludascher B.: YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. International Journal for Digital Curation, vol. 10, 298–313. (2015)

8. Missier, P., Ludascher, B., Bowers, S., Anand, M. K., Altintas, I., Dey, S., Sarkar, A., Shrestha, B. and Goble, C.: Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science. In: 5th Workshop on Workflows in Support of Large-Scale Science. (2010).
9. Missier, P.: Data Trajectories: Tracking Reuse of Published Data for Transitive Credit Attribution. In: 11th International Data Curation Conference. (2016).
10. Slaughter P., Jones M., Jones C.: NCEAS recordr: Provenance tracking for R. `https://github.com/NCEAS/recordr` (2016)
11. Open Archives Initiative Object Reuse and Exchange, `https://www.openarchives.org/ore/`