DataONE: A Data Federation with Provenance Support

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

[1] Library and Information Science, University of Illinois, Urbana-Champaign, IL
[2] National Center for Ecological Analysis and Synthesis, UCSB, CA
[3] Universidad Popular Autónoma del Estado de Puebla, Mexico
[4] UCAR and U.S. Global Change Research Program
[5] School of Computing Science, Newcastle University, UK
[6] Woods Hole Research Center, Falmouth, MA
[7] Environmental Sciences Division, Oak Ridge National Laboratory, TN

Abstract. DataONE is a federated data network focusing on earth and environmental science data. We demonstrate new provenance capabilities in the DataONE toolkit to facilitate reproducible research. A user "Alice", can annotate a (Matlab, R, etc.) script using the YesWorkflow (YW) tool to describe the underlying workflow or *prospective* provenance. After Alice has run the script, the result files, script, prospective provenance, and *retrospective* provenance, represented in the ProvONE provenance model, are bundled into an OAI-ORE compliant data package and uploaded to the DataONE network. A second user ("Bob") discovers Alice's package and uses her data in his own analysis. 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 discovers the full provenance of Bob's results, all the way back to Alice's original contributions. DataONE provenance systems enable reproducible research and facilitate proper attribution of scientific results transitively across generations of derived data products.

1 Introduction

Scientific workflow *provenance* is valuable in computational science. Provenance can help scientists to understand their own work and share their work with others while maintaining attribution. We refer to two types of provenance: *prospective* and *retrospective* provenance, where the former refers to a specification of a data transformation process [FKSS08], and the latter refers to the derivations that account for the actual outcomes of one execution of the process.

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 [data]. The primary goals of DataONE are: (i) data discovery, access, integration and synthesis; (ii) education and training, building community; and (iii) data sharing. The DataONE infrastructure consists of three principal components:

(1) Member Nodes are existing or new data repositories that support the DataONE Member Node API. Tier-1 Member Nodes (MN) implement anonymous, read-only access to data and metadata and enable discovery of all objects available on the MN, with low level descriptions of each object, object retrieval by identifier, and activity reporting. MNs also provide science metadata and relationships between metadata and data using resource maps [oai] to facilitate discovery. MNs can also implement higher tiers of the DataONE API, including authenticated and restricted data access (Tier-2), write access to storage resources (Tier-3), and replication of data from other nodes (Tier-4).

(2) Coordinating Nodes (CN) serve the coordination and discovery needs of the network. Services of a CN include network-wide indexing of data objects, coordination of data replication between MNs, mirrored content of *science metadata* (detailed descriptions of science data objects and collections) and *system metadata* (low level metadata, e.g., data type, size, ownership, location) that are extracted from data files in the MNs.

(3) Investigator Toolkit. The Investigator Toolkit (ITK) contains developer tools that enable programmatic interaction with DataONE infrastructure through a REST service API exposed by the CNs and MNs. Python and Java libraries are available 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, an infrastructure for federated identity management, and replication service for both data and metadata.

DataONE Search is a web-based application that lets users seamlessly and efficiently discover publicly accessible data packages within the DataONE federated network of Member Node repositories. It allows users to search across space (geographical region) and time, using keywords and other facets and to refine the results using further parameters. Users sign in to DataONE Search using ORCID credentials, Google accounts, or institutional accounts.

> **[PM]** I suggest explaining why we are introducing Search here – saying for instance that DataONE enables new user features like provenance-based browsing as part of its search facility.
>
> Also " other facets and to refine the results using further parameters. " is a bit vague, I would remove this part.

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 new DataONE provenance tools and the visualization of provenance with DataONE Search.

2 Demonstration Description

In this demonstration we describe two features related to provenance. The first is an API for capturing *retrospective* provenance from R [**?**] and Matlab [JCSJ] script executions, called *Run Manager*. The second is a script annotation tool, which we call YesWorkflow [MSK+15], designed to help developers and users better understand the structure and intent of a script.

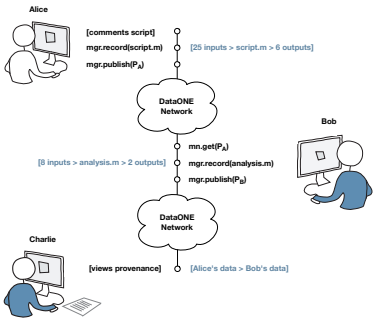

Fig. 1: Run Manager Demonstration: (1) Alice runs `script.m` with the DataONE Run Manager to create data package P_A, which she publishes to the DataONE network; (2) Bob later finds and downloads Alice's data, uses it in his `analysis.m`, creating and then publishing package P_B; (3) Charlie searches DataONE, finds Bob's P_B, and recognizes its dependence on Alice's P_A.

[PM] The automatic capture of the latter is transparent to scientists and provides medium-grained[a] provenance by focusing on file I/O events.

orphan sentence, but it seems to try and say too much – is it necessary?

[a] unlike the fine-grained provenance captured, e.g., by noWorkflow [MBC+15]

We introduce the provenance and search features of DataONE by means of an example involving three earth scientists who interact through a DataONE MN, as shown in Fig. 1. First Alice generates and publishes new global grass fraction maps from observational data, eg water, air temperature. Then Bob uses Alice's data to run further analysis, i.e., on grassland water use efficiency and publishes it back to a DataONE MN. Finally, Charlie, a DataONE user, examines the provenance of Bob's published data to determine its suitability for his own use.

2.1 Alice's data generation

As shown at the top of Fig. 1, Alice has developed a script for producing Carbon3/Carbon4 soil maps. She uses the YesWorkflow (YW) tool to mark-up the script and expose the underlying workflow view (i.e., prospective provenance) that is inherent in her soil mapping code.

[PM] need a ref here? and a figure with the graph produced by YW on this script would be really useful to bring this part to life!

By using the *Run Manager* to run her script, Alice not only obtains the expected results, but she also captures their provenance, compliant with DataONE's ProvONE data model. ProvONE [proa] is an extension of the W3C PROV-O [prob] standard for representing provenance, and includes specializations for representing both retrospective provenance about the runtime execution and prospective provenance about the structure and flow of the analytical script or workflow. At the end of the experimentation phase, Alice is ready to publish her results to a DataONE MN. To do so, she uses the

DataONE Matlab tool to automatically generate a DataONE-compliant data package in OAI-ORE format, including the ProvONE provenance document, the script itself, and its YW-generated workflow view.

2.2 Bob's data reuse

Bob's interaction with DataONE begins with a user interface search, i.e., using the keyword "grass", he discovers Alice's data package [yax], amongst others.

> **[PM]** Why is [yax] cited here?

He decides to use three NetCDF output data files which are part of the package, as input to his Grassland Water Use Efficiency Analysis script (GrasslandWUE.m). Having identified the data of interest in the MN, Bob uses its public identifier *id* to retrieve it and use it in his own code **[put bob's script name here]**. Specifically, the `MemberNode.get`(session, id) call, available from the Matlab Toolbox, not only correctly retrieves Alice's data package, but it also ensures that the download event is recorded as part of a new provenance document, associated to Bob's analysis. Note that, should Bob create a new identifier when he downloads Alice's data, the link between two data packages would be broken, leading to a disconnect in provenance and requiring additional "stitching" operations [MLB+10].

Instead, by retaining the same identifier throughout, the tool implicitly establishes a connection between Alice's work and Bob's, namely by adding a provenance statement of the form *prov:used*

> **[PM]** can you be completely explicit here? it should look like used([**bob's script name as above**], id)

. Bob then proceeds to operate on the data using the DatAONE Matlab just like Alice did, eventually publishing a new data package with his own results and their provenance. At this point, the two provenance documents are physically disjoint, as they reside in different data packages, but they are logically connected, namely through the used() statement mentioned above. As they are both indexed by the CN upon publication of the data package, this logical connection emerges automatically when a third party, such as Charlie in our example, explores one of the two data packages.

2.3 Charlie's visual data navigation

> **[PM]** 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 [KS14,Mis16].

When Charlie searches the DataONE network using the same keyword "grass" from the web search interface, two data packages are displayed as shown in Figure 2, namely Alice's and Bob's. One data package is created by "Alice" (Yaxing Wei in the real world) [yax], the other is created by "Bob" (Christopher Schwalm) [chr].

> **[PM]** I'm not quite sure how to fit Yaxing and Chris names here, it looks odd... why not replace Alice and Bob with Yaxing and Chris from the beginning?!? and explain upfront that they are real scientists.

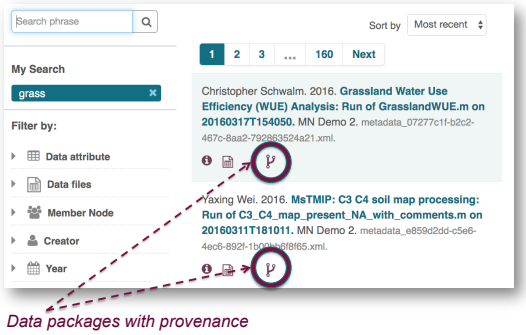

Fig. 2: DataONE Search result for keyword "grass" (DataONE Search demonstration site). The "fork" icon indicates data packages having embedded provenance.

Crucially, the provenance of the two datasets is now manifested visually along with their logical connection, as shown in the DataONE Search web UI for our demo site [datb] (Fig. 3) and is available to Charlie. Specifically, Charlie can not only visualise the two data packages (Alice's is at the top and Bob's at the bottom), but he is also aware of the derivation of Alice's data through Bob's script.

Provenance details for any input or output in the provenance graph can be viewed by clicking on the icon shown in the figure. DataONE Search also provides human language descriptions of how data are used or generated via the script and models, and provides navigation to ancestors and descendants in the data derivation chain. In this example, Charlie quickly learns that Alice's script (C3_C4_map_present_with_comments.m) takes twenty-five input files and produces six outputs, shown on the left and right side of Alice's data package, 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 that was generated by the YesWorkflow tool [MSK+15]. Alice's used embedded YesWorkflow annotations to document her script. These annotations declare step by step how data are used and derived in the script.

Similarly, 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.

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 discussion. Currently, Bob must use certain functions provided by DataONE to ensure

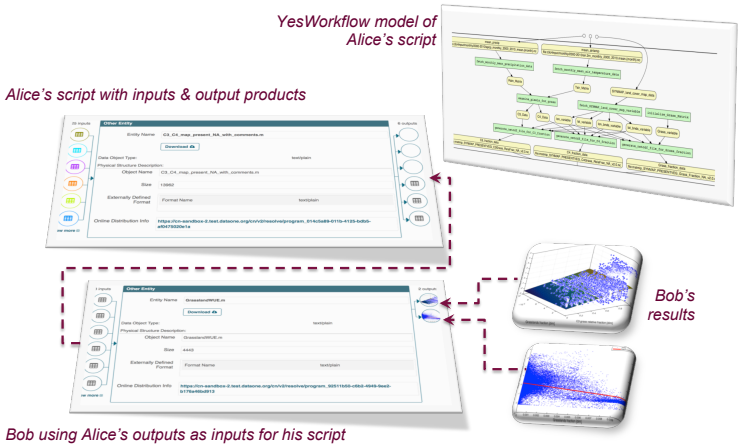

Fig. 3: Charlie's view on the DataONE demo site: (1) A YesWorkflow model for Alice's soil processing script; (2) Data lineage from Bob's results back through his script inputs to Alice's data package; (3) Two visualizations produced by Bob's water use efficiency analysis script.

the provenance link to Alice's work is correctly maintained. There are other possible ways to achieve the same goal. For example, Bob might 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 could easily add, especially for historical data for which provenance was not captured at the time data were generated. The new provenance tools described in this demonstration allow efficient capture of machine-parseable provenance information using analytical tools commonly used in the environmental sciences (e.g., Matlab), thereby significantly improving the replicability of research in the environmental sciences.

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

References

chr. Christopher Schwalm's DataPackage. https://search-sandbox-2.test.dataone.org/#view/metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml/. [last accessed 10-April-2016].

data. Data Observation Network for Earth (DataONE). https://www.dataone.org. [last accessed 10-April-2016].

datb. DataONE Search Demo Site. https://search-sandbox-2.test.dataone.org. [last accessed 10-April-2016].

FKSS08. Juliana Freire, David Koop, Emanuele Santos, and Cláudio T Silva. Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

JCSJ. Christopher Jones, Yang Cao, Peter Slaughter, and Matthew B. Jones. Matlab DataONE Toolbox. https://github.com/DataONEorg/matlab-dataone. [last accessed 10-April-2016].

KS14. Daniel S. Katz and Arfon M. Smith. Implementing transitive credit with JSON-LD. *CoRR*, abs/1407.5117, 2014.

MBC+15. Leonardo Murta, Vanessa Braganholo, Fernando Chirigati, David Koop, and Juliana Freire. *No Workflow: Capturing and analyzing provenance of scripts*, volume 8628 of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 71–83. Springer Verlag, 2015.

Mis16. Paolo Missier. Data trajectories: tracking reuse of published data for transitive credit attribution. In *Procs. 11th International Data Curation Conference*. DCC, 2016.

MLB+10. Paolo Missier, Bertram Ludäscher, Shawn Bowers, Manish Kumar Anand, Ilkay Altintas, Saumen Dey, Anandarup Sarkar, Biva Shrestha, and Carole Goble. Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science. In *Proc.s 5th Workshop on Workflows in Support of Large-Scale Science (WORKS)*, 2010.

MSK+15. Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, R. Kyle Bocinsky, Yang Cao, James Cheney, Fernando Chirigati, Saumen Dey, Juliana Freire, Christopher Jones, James Hanken, Keith W. Kintigh, Timothy A. Kohler, David Koop, James A. Macklin, Paolo Missier, Mark Schildhauer, Christopher Schwalm, Yaxing Wei, Mark Bieda, and Bertram Ludäscher. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. *International Journal of Digital Curation*, 10(1):298–313, 2015.

oai. Open Archives Initiative Object Reuse and Exchange. https://www.openarchives.org/ore/. [last accessed 10-April-2016].

proa. ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance. https://purl.dataone.org/provone-v1-dev. [last accessed 10-April-2016].

prob. W3C PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/. [last accessed 10-April-2016].

yax. YaXing Wei's DataPackage. https://search-sandbox-2.test.dataone.org/#view/metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml. [last accessed 10-April-2016].