DataONE: A Data Federation with Provenance Support

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

Abstract. DataONE is a federated data network focusing on earth and environmental science data. We present the provenance and search features of DataONE by means of an example involving three earth scientists who interact through a DataONE Member Node. 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 better understand 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 or *workflow* [5], and the latter refers to the derivations that account for the actual outcomes of an execution of the process.

DataONE (Data Observation Network for Earth) is a federated data network for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data [3]. DataONE's primary goals include support for: data discovery, access, integration, and synthesis; education, training, and building community; and data sharing. The DataONE infrastructure consists of three principal components:

Member Nodes (MN) represent existing or new data repositories that support the DataONE Member Node API; *Coordinating Nodes* (CN) serve the coordination and discovery needs of the network; and the *Investigator Toolkit* which contains tools that enable programmatic interaction with DataONE infrastructure through a REST service API exposed by the CNs and MNs.

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), time, and using a set of keywords. Users sign in to DataONE Search using ORCID credentials, Google accounts, or institutional accounts. DataONE enables new user features like provenance-based browsing as part of its search facility. In the next section, we will present new DataONE provenance tools and the visualization of provenance with DataONE Search [3].

* [1]University of Illinois, Urbana-Champaign, [2]National Center for Ecological Analysis and Synthesis, UCSB, [3]Universidad Popular Autónoma del Estado de Puebla, Mexico, [4]School of Computing Science, Newcastle University, UK, [5]Woods Hole Research Center, Falmouth, MA, [6]University of Kansas, Lawrence, [7]Environmental Sciences Division, ORNL, TN.

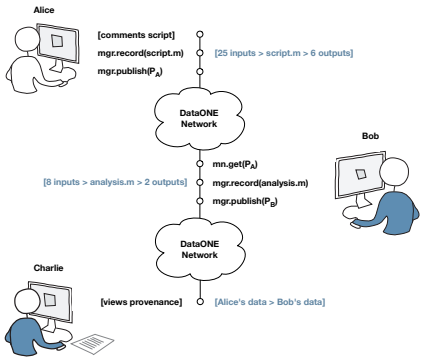

Fig. 1: Provenance Use Case: (1) Alice runs script.m with the DataONE Run Manager to create data package P_A, which she publishes to the DataONE network; (2) Bob later finds and downloads Alice's data, uses it in his analysis.m, creating and then publishing package P_B; (3) Charlie searches DataONE,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

2 Provenance Feature Description

We present two features related to provenance: *Run Manager*, an API for capturing retrospective provenance from R [14] and MATLAB [7] script runs; and *YesWorkflow* [9], a script annotation and provenance querying tool, designed to help users better understand the structure and intent of a script, and to expose and query its provenance.

We introduce the provenance and search features of DataONE by means of an example involving three Earth scientist personas who interact through a DataONE Member Node: In Figure 1, Alice has developed a script for producing C_3/C_4 carbon soil maps [15]. 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 as shown in Figure 2.

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 [2] is an extension of the W3C PROV-O [12] 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 Member Node. 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.

Bob's interaction with DataONE begins with a user interface search, i.e., using the keyword "grass", he discovers Alice's data package, amongst others. He decides to use three NetCDF output data files which are part of her package, as input to his Grassland Water Use Efficiency Analysis script [6]. Having identified the data of interest in the Member Node, Bob uses its public identifier *id* to retrieve it and use it in his own code. Specifically, the MemberNode.get(session, id) call, available from the MATLAB toolbox, not only retrieves Alice's data package, but it also ensures that the download event is recorded as part of a new provenance document, associated with Bob's analysis. If Bob manually downloaded Alice's data (i.e., without using the DataONE tool), then

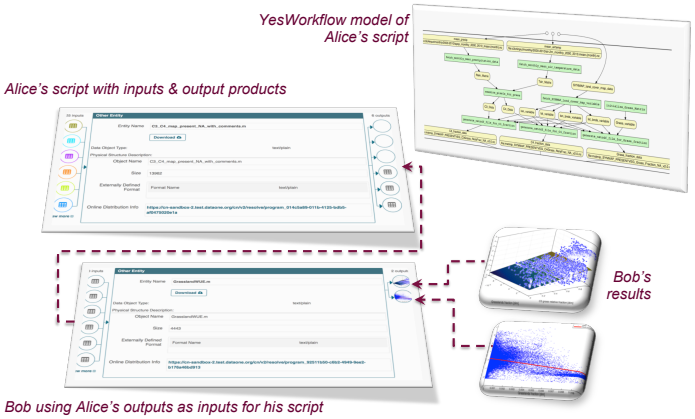

Fig. 2: Charlie's view on the DataONE demo site: (1) A YesWorkflow model for Alice's soil processing script; (2) Data lineage from Bob's results back through his script inputs to Alice's data package; (3) Two visualizations produced by Bob's water use efficiency analysis script.

the link between the data packages would likely be broken, leading to a disconnect in provenance and requiring additional "stitching" operations [11].

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 (*Bob's_execution*, prov:used, *Alice's_data_id*). Bob then proceeds to operate on the data using the DataONE MATLAB toolbox just as 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 prov: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, say Charlie, explores one of the two data packages:

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 2 [8,10]. When he searches the DataONE network using the same keyword "grass" from the web search interface, two data packages are displayed as shown in Figure 2. One data package was created by Alice [16], the other was created by Bob [13].

Crucially, the provenance of the two datasets is now manifested visually along with their logical connection, as shown in the DataONE Search web UI [4] (Fig. 2) and is available to Charlie. Specifically, Charlie can not only visualize 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 icons 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 takes twenty-five input files [15] 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 (C_3 grass fraction, C_4 grass fraction, and total grass fraction) [15]. In addition, a model graph is displayed at the intermediate layer that was generated by the YesWorkflow tool declaring step by step how data are used and derived in the script [9]. Similarly, the provenance information is associated with Bob's data package in Fig. 2 [13,1,6].

Conclusions. As outlined above, we have described new and unique provenance capabilities in the large, scientific data federation network DataONE. The search feature was released to the public in late 2015; the R and MATLAB provenance tools in early 2016.

References

1. Cao, Y., Jones, C., Cuevas-Vicenttín, V., Jones, M.B., Ludäscher, B., McPhillips, T., Missier, P., Schwalm, C., Slaughter, P., Vieglais, D., Walker, L., Wei, Y.: DataONE: A Data Federation with Provenance Support (2016), Demo-Paper (long version)
2. Cuevas-Vicenttín, V., et al.: ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance (2015)
3. Data Observation Network for Earth (DataONE). www.dataone.org and search.dataone.org
4. DataONE Search Demo Site. https://search-sandbox-2.test.dataone.org
5. Freire, J., Koop, D., Santos, E., Silva, C.T.: Provenance for Computational Tasks: A Survey. Computing in Science and Engineering 10(3), 11–21 (2008)
6. Huntzinger, D., Schwalm, C., Wei, Y., Cook, R., Michalak, A., Schaefer, K., Jacobson, A., Arain, M., Ciais, P., Fisher, J., Hayes, D., Huang, M., Huang, S., Ito, A., Jain, A., Lei, H., Lu, C., Maignan, F., Mao, J., Parazoo, N., Peng, C., Peng, S., Poulter, B., Ricciuto, D., Tian, H., Shi, X., Wang, W., Zeng, N., Zhao, F., Zhu, Q.: NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (version 1) in Standard Format
7. Jones, C., Cao, Y., Slaughter, P., Jones, M.B.: MATLAB DataONE Toolbox. https://github.com/DataONEorg/matlab-dataone (2016)
8. Katz, D.S., Smith, A.M.: Implementing Transitive Credit with JSON-LD. CoRR abs/1407.5117 (2014)
9. McPhillips, T., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., Bocinsky, K., Cao, Y., Chirigati, F., Dey, S., Freire, J., Huntzinger, D., Jones, C., Koop, D., Missier, P., Schildhauer, M., Schwalm, C., Wei, Y., Cheney, J., Bieda, M., Ludäscher, B.: YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. International Journal of Digital Curation 10, 298–313 (2015)
10. Missier, P.: Data trajectories: tracking reuse of published data for transitive credit attribution. In: Procs. 11th Intl. Data Curation Conference. DCC (2016)
11. Missier, P., Ludäscher, B., Bowers, S., Anand, M.K., Altintas, I., Dey, S., Sarkar, A., Shrestha, B., Goble, C.: Linking Multiple Workflow Provenance Traces for Interoperable Collaborative Science. In: 5th Workshop on Workflows in Support of Large-Scale Science (WORKS) (2010)
12. W3C PROV-O: The PROV Ontology. https://www.w3.org/TR/prov-o/
13. Schwalm, C.: Data-Package of "Bob" (2016), https://goo.gl/rYOZyh
14. Slaughter, P., Jones, M.B., Jones, C.: recordr: Provenance tracking for R. https://github.com/NCEAS/recordr (2016)
15. Wei, Y., Liu, S., Huntzinger, D., Michalak, A., Viovy, N., Post, W., Schwalm, C., Schaefer, K., Jacobson, A., Lu, C., Tian, H., Ricciuto, D., Cook, R., Mao, J., Shi, X.: NACP MsTMIP: Global and North American Driver Data for Multi-Model Intercomparison (2014)
16. Wei, Y.: Data-Package of "Alice" (2016), https://goo.gl/BsHSuK