Provenance Storage, Querying, and Visualization in PBase*

Víctor Cuevas-Vicenttín[1], Parisa Kianmajd[1], Bertram Ludäscher[1],
Paolo Missier[2], Fernando Chirigati[3], Yaxing Wei[4],
David Koop[3], and Saumen Dey[1]

[1] University of California at Davis, USA
[2] Newcastle University, UK
[3] New York University, USA
[4] Oak Ridge National Laboratory, USA
{victorcuevasv, parisa.kianmajd}@gmail.com,
{ludaesch, scdey}@ucdavis.edu, Paolo.Missier@ncl.ac.uk,
{fchirigati, dakoop}@nyu.edu, weiy@ornl.gov

Abstract. We present PBase, a repository for scientific workflows and their corresponding provenance information that facilitates the sharing of experiments among the scientific community. PBase is interoperable since it uses ProvONE, a standard provenance model for scientific workflows. Workflows and traces are stored in RDF, and with the support of SPARQL and the tree cover encoding, the repository provides a scalable infrastructure for querying the provenance data. Furthermore, through its user interface, it is possible to: visualize workflows and execution traces; visualize reachability relations within these traces; issue SPARQL queries; and visualize query results.

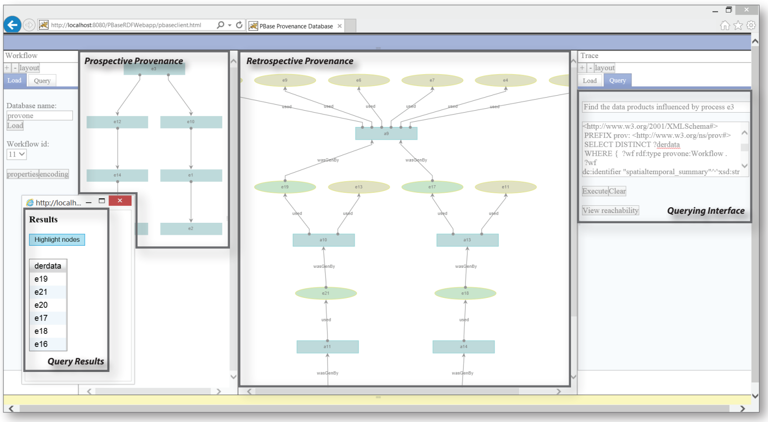
Keywords: PBase, ProvONE, Scientific Workflows, Provenance Repository

1 Introduction

In the past few years, scientific workflows have been often used to define and execute a range of experiments. As science is collaborative, the need arises for a repository that allows multiple users to store and query scientific workflow provenance information. Additionally, such a repository must be interoperable, in the sense that workflow traces may come from different systems, and scalable as the number and the size of traces grow, providing an efficient query evaluation.

This paper presents PBase [CVKL+14], which addresses three main key points: facilitate the *sharing* of scientific workflows and their corresponding execution traces among the scientific community; allow *user interaction* so that users can further explore the repository data; and provide both sharing and interaction in an *interoperable* and *scalable* manner. Our repository achieves these

* The original publication is available at http://www.springerlink.com/



Find the data products influenced by process e3

```
<http://www.w3.org/2001/XMLSchema#>
 PREFIX prov: <http://www.w3.org/ns/prov#>
 SELECT DISTINCT ?derdata
 WHERE {  ?wf rdf:type provone:Workflow .
 ?wf
dc:identifier "spatialtemporal_summary"^^xsd:str
```

derdata
e19
e21
e20
e17
e18
e16

Fig. 1. The PBase Web GUI and its main components.

goals by: (i) making use of ProvONE [Dat14a], a standard provenance model that brings the advantages of the emerging W3C PROV standard [W3C13] and that addresses the interoperability challenge; (ii) defining a representative set of queries, identified in collaboration with climate scientists, that characterizes the required functionality and user interaction; and (iii) providing a scalable infrastructure based on TDB, the RDF triplestore of the Jena Framework[5] that supports SPARQL, an expressive query language, and its efficient evaluation. PBase also incorporates the tree cover encoding proposed by Agrawal et al. [ABJ89] to improve the performance of reachability queries.

To the best of our knowledge, PBase is the first repository to address all the aforementioned challenges.

2 PBase Features

Interoperability. PBase uses ProvONE [Dat14a] to represent both prospective provenance (i.e. workflow specifications) and retrospective provenance (i.e. execution traces). ProvONE is an extension of the W3C PROV [W3C13] standard and it is specified through an ontology serialized in OWL-2. Its goal is to be expressive enough to cover most workflow models used by different scientific workflow management systems, which allows PBase to work in an interoperable manner.

User Interaction. An essential feature for a provenance repository is to *visualize* a workflow and its various execution traces. PBase uses a Web GUI for this purpose (see Figure 1). Furthermore, in collaboration with climate scientists, we have identified a series of queries, specified in SPARQL, that are representative for the functionalities that they require (such queries are available in [Dat14b]). As users may not be familiar with SPARQL, PBase also allows these queries to

[5] http://jena.apache.org/

be issued from the GUI interface through their textual description. When the results of a query are generated, besides presenting them in a text representation, the provenance nodes corresponding to the results are highlighted. To see the lineage of a particular node in a workflow or trace, users can select this node and use the option to highlight its ancestors and descendants.

Scalability. We adopt RDF to store workflows and execution traces—in particular, we use TDB from the Jena Framework. As an example, XML traces from VisTrails[6] can be uploaded through the Web and they are automatically translated into ProvONE RDF and stored in TDB. As mentioned before, PBase uses SPARQL to issue queries in the repository, which allows for an expressive and efficient evaluation. The tree cover encoding [ABJ89] is also implemented: it enables determining reachability relations between nodes by simply comparing integer range intervals, thus avoiding more costly graph explorations and enhancing the performance of PBase.

3 Conclusion

We have presented PBase, a repository for scientific workflows and their corresponding execution traces. It can be regarded as a step towards a repository supporting sophisticated provenance querying and analytics over a large collection of traces. PBase was developed in the context of DataONE[7], a large scale and federated data infrastructure serving the Earth Sciences community, and our ultimate goal is to incorporate it into this infrastructure.

Acknowledgments. The authors thank: members of the DataONE Provenance Working Group, for helping in the specification of PBase; and members of the DataONE EVA Working Group, for their collaboration. This work was supported by NSF Award OCI-0830944 (DataONE).

References

ABJ89. R. Agrawal, A. Borgida, and H. V. Jagadish. Efficient Management of Transitive Relationships in Large Data and Knowledge Bases. In *Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data*, SIGMOD '89, pages 253–262, New York, NY, USA, 1989. ACM.

CVKL+14. Víctor Cuevas-Vicenttín, Parisa Kianmajd, Bertram Ludäescher, Paolo Missier, Fernando Chirigati, Yaxing Wei, David Koop, and Saumen Dey. The PBase Scientific Workflow Provenance Repository. In *Proceedings of the 9th International Digital Curation Conference*, IDCC '14, 2014.

Dat14a. DataONE Provenance Working Group. ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance. http://purl.org/provone, 2014.

Dat14b. DataONE Provenance Working Group. The ProvONE Scientific Workflow Provenance Dataset. http://purl.org/provone/provbench, 2014.

[6] http://www.vistrails.org/

[7] http://www.dataone.org/

W3C13. W3C Provenance Working Group. PROV Overview. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/, 2013.