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

Motivation

- Provenance – information about the origin, context, derivation, ownership, or history of an artifact – plays a key role to examine and audit the results of scientific experiments
- Since science is collaborative, the need arises for a repository that facilitates the sharing of scientific workflows and their corresponding provenance traces, also enabling querying and visualization
- Besides, this should be done is an interoperable manner, as many different scientific workflow management systems may be used
- Such functionality should also be supported while taking performance and scalability into account

The PBase Repository

- PBase is a repository for scientific workflows and their associated provenance
- It establishes a functionality that would be incorporated into Member and Coordinating Nodes, providing provenance support to the DataONE Cyber Infrastructure
- The main goal is to address three main challenges:
  1. Facilitate the ***sharing*** of scientific workflows and their provenance traces among the scientific community
  2. Allow ***user interaction*** so that scientists can further explore the repository data
  3. Provide both sharing and interaction in an ***interoperable*** and ***scalable*** manner

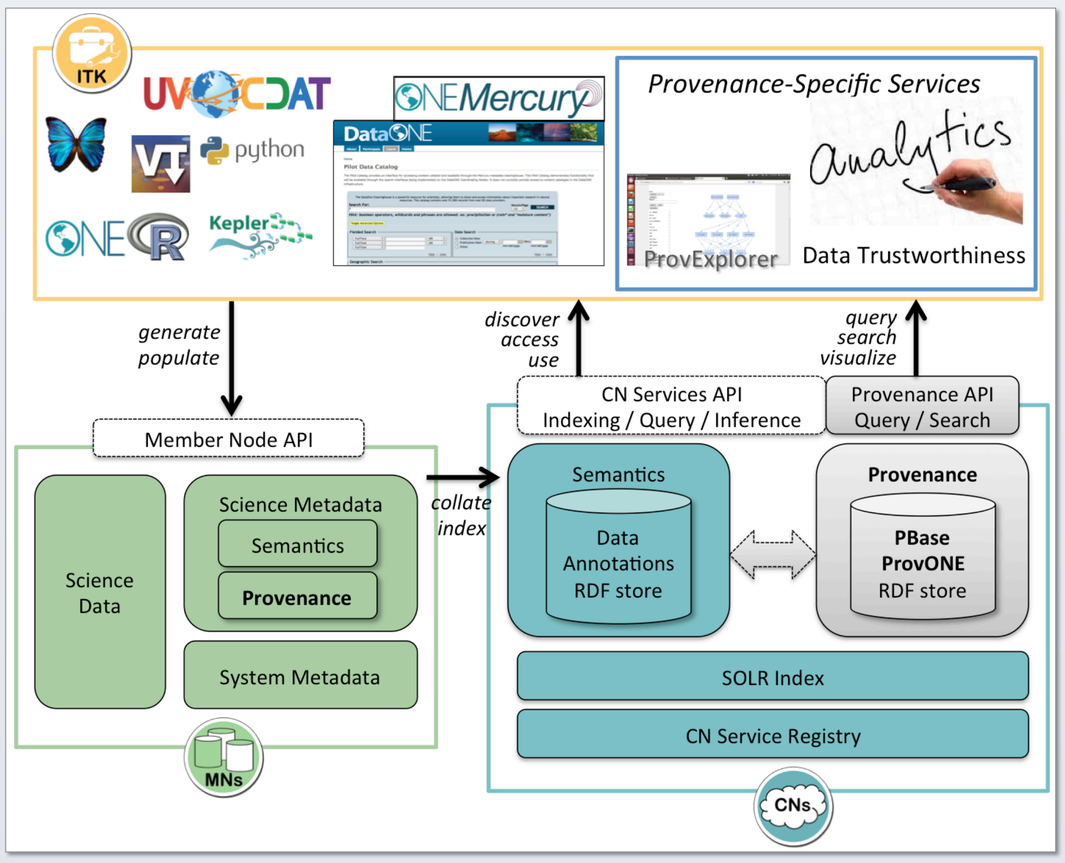

Figure 1. PBase and provenance in the DataONE infrastructure

Interoperability: The ProvONE Model

- PBase uses ProvONE [1,2] to represent both prospective (i.e.: workflow definitions) and retrospective (i.e.: execution traces) provenances
- ProvONE is an extension of the emerging W3C PROV [3] standard that aims to be expressive enough to cover most of the workflow models in use by the leading scientific workflow management systems
- It is specified through an ontology serialized in OWL-2

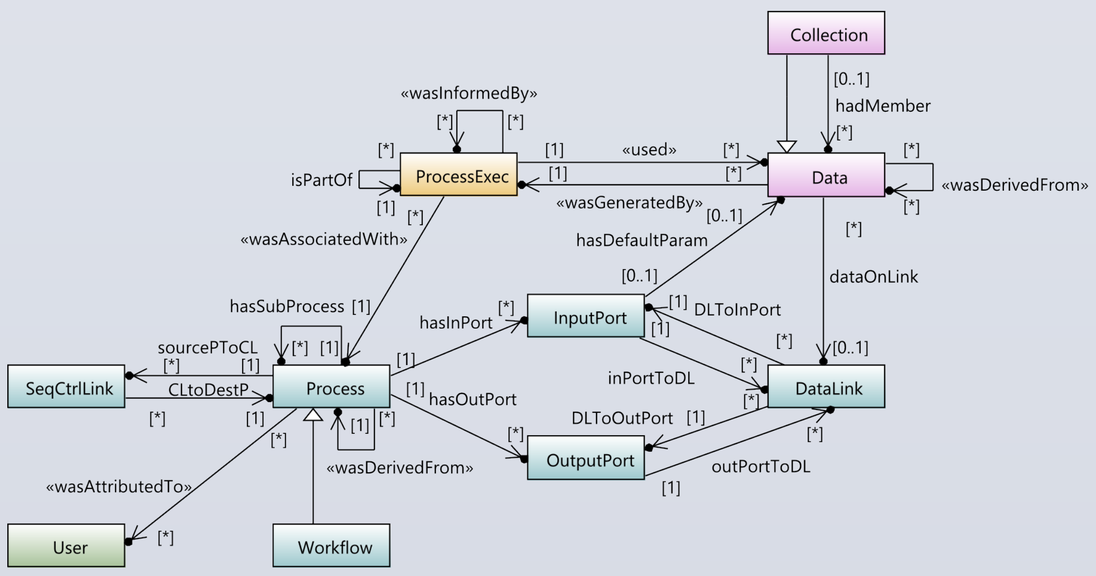

Figure 2. The ProvONE model

User Interaction: Querying and Visualization

- A Web GUI is used to visualize workflows and their corresponding execution traces
- A set of representative queries, defined in collaboration with climate scientists, is used to characterize the querying functionality
- SPARQL is used for the querying interface
- Queries can also be issued from the GUI through their textual description
- Scientists can interactively select nodes in a workflow or trace to highlight important features
  - E.g.: If a user is interested in the lineage of a particular node, she can select the node and highlight ancestors and descendants of that particular node

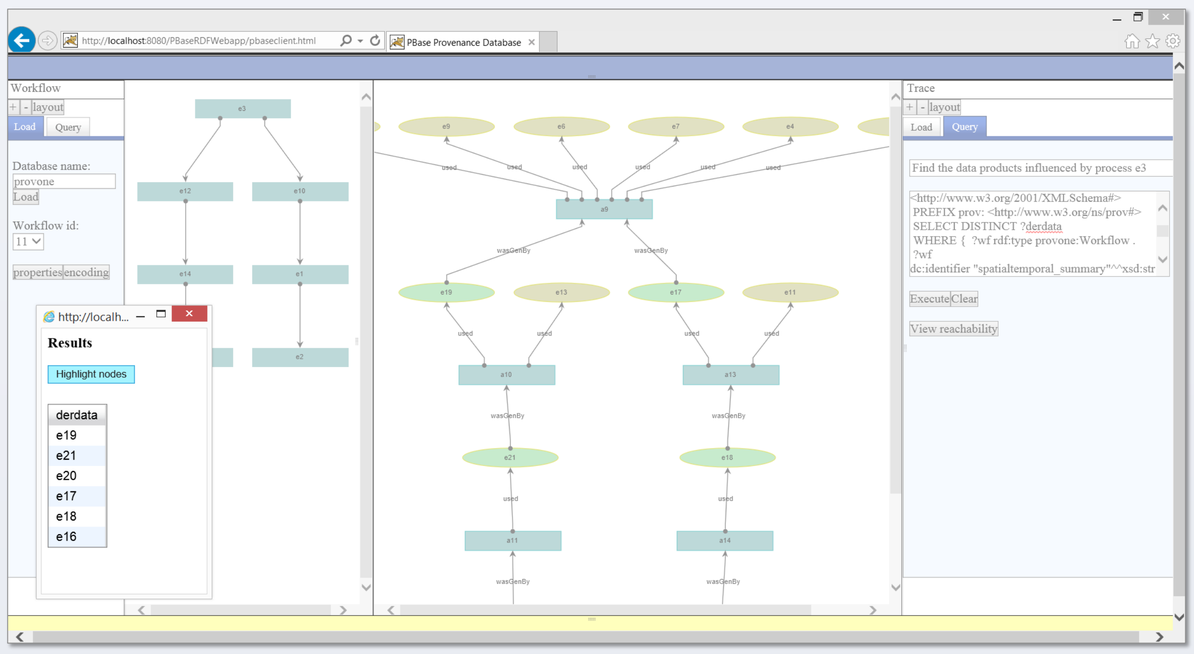

Figure 3. The PBase Web GUI: on the left, a climate analysis workflow; in the middle, its corresponding execution trace; and on the right, the querying interface.

Scalability: RDF and Tree Cover Encoding

- Workflows and provenance traces are stored and serialized in RDF
- In particular, PBase makes use of the TDB component of the Jena framework [4]
- The tree cover encoding [5] is computed on the backend to determine reachability relations – this avoids more costly graph exploration, increases the performance of such queries, and makes the tool more scalable

```
SELECT DISTINCT ?data_id
WHERE {  ?wf rdf:type provone:Workflow .
?wf dc:identifier "spatialtemporal_summary"^^xsd:string .
?wf provone:hasSubProcess ?p .
?pexec prov:wasAssociatedWith ?p .
?pexec prov:used ?data .
?data dc:identifier ?data_id .
FILTER NOT EXISTS { ?data prov:wasGeneratedBy ?other_pexec } }
```

Figure 4. Example of a lineage query, which finds all the inputs of the workflow across multiple executions, in SPARQL

References

[1] ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance. In: http://purl.org/provone (2014)

[2] Cuevas-Vicenttín, V., Kianmajd, P., Ludäescher, B., Missier, P., Chirigati, F., Wei, Y., Koop, D., Dey, S.: The PBase Scientific Workflow Provenance Repository. In: Proceedings of the 9th International Digital Curation Conference. IDCC '14 (2014)

[3] PROV Overview. In: http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/ (2013)

[4] TDB – Jena Framework. In: http://jena.apache.org/documentation/tdb/ (2014)

[5] Agrawal, R., Borgida, A., Jagadish, H.V.: Efficient Management of Transitive Relationships in Large Data and Knowledge Bases. In: Proceedings of the 1989 ACM SIGMOD International Conference on Management of Data. pp. 253–262. SIGMOD '89, ACM, New York, NY, USA (1989)

Acknowledgments

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