The PBase Scientific Workflow Provenance Repository

Víctor Cuevas - UNM, UC Davis
Parisa Kianmajd – UC Davis
Bertram Ludäscher - UC Davis
Paolo Missier - Newcastle U, UK
Fernando Chirigati - NYU Poly
Yaxing Wei - ORNL
David Koop - NYU Poly
Saumen Dey - UC Davis
Provenance

The origin and processing history of an artifact

- **Evidence, Credibility**: How was this data object produced?
- **Data Attribution**: Who is responsible/should be credited for it?
- **Data Discovery**: find data based on its provenance properties
- **Data Relevance**: is the data relevant to me?
- **Reproducibility**: can the experiment be repeated with the same results?
Scientific Workflow Provenance (1)

• Summarize ecological spatio-temporal data
Scientific Workflow Provenance (2)

- ProvONE: standardized model for workflow provenance
# Querying Workflow Provenance

- **Graph queries**: node and edge properties, aggregation, reachability
- **Declarative query language**: specify information needs rather than traversals

<table>
<thead>
<tr>
<th>#</th>
<th>Query</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>START n=node:node_auto_index(name=&quot;e7_Regrid&quot;) MATCH m-[:wasAssociatedWith]-n RETURN count(m)</td>
<td>Statistical</td>
</tr>
<tr>
<td>2</td>
<td>START n=node(*) MATCH (n)-[:used]-(a) WHERE n.name=&quot;e7_Regrid&quot;-[:wasGeneratedBy]-&gt;() AND HAS(n.wfID) AND (n.wfID=&quot;wf1&quot;) RETURN DISTINCT a</td>
<td>Lineage</td>
</tr>
<tr>
<td>3</td>
<td>START n=node(*) MATCH n-[:wasAssociatedWith]-m WHERE HAS(n.vtType) AND HAS(n.wfID) AND HAS(n.runID) AND n.vtType=&quot;vt:module_exec&quot; AND n.completed=-1 AND n.wfID=&quot;wf1&quot; AND n.runID=&quot;ex1&quot; RETURN m</td>
<td>Execution</td>
</tr>
<tr>
<td>4</td>
<td>START n=node(*) WHERE HAS(n.completed) AND n.completed=-1 AND HAS(n.wfID) AND n.wfID=&quot;wf1&quot; RETURN DISTINCT n</td>
<td>Execution</td>
</tr>
<tr>
<td>5</td>
<td>START n = node:node_auto_index(name=&quot;e10_regrid_method&quot;) MATCH n-[:used]-a RETURN distinct a</td>
<td>Search</td>
</tr>
<tr>
<td>6</td>
<td>START n=node(*) MATCH n-[z]-&gt;a WHERE n-[:wasGeneratedBy]-&gt;() AND HAS(n.wfID) AND n.wfID=&quot;wf1&quot; AND HAS(a.module) AND a.module=&quot;e7_Regrid&quot; RETURN DISTINCT n</td>
<td>Search</td>
</tr>
</tbody>
</table>
Graph Reachability Encoding

- Tree cover encoding [Agrawal et al.]
- Efficiently determine ancestors and descendants
PBase Architecture

- PBase functionality implemented as Restful Web Services
- Use of JSON for interoperability
PBase Web GUI
Conclusions

• Repository for querying and visualizing scientific workflow provenance
  • Characterization of information needs as graph queries
  • Customized user interface favoring usability
  • Provenance model similar to other PROV extensions (e.g. Belhajjame et al.)

• Future work
  • Explore the use of RDF/SPARQL as an alternative
  • Extend capabilities to keyword search and ranking
  • Incorporate functionality into DataONE Cyberinfrastructure
Q & A
References


Provenance

The origin and processing history of an artifact

- **Evidence, Credibility:** How was this data object produced?
- **Data Attribution:** Who is responsible / should be credited for it?
- **Data Discovery:** find data based on its provenance properties
- **Data Relevance:** is the data relevant to me?
- **Reproducibility:** can the experiment be repeated with the same results?

**Provenance Repository**

- **Find data and workflows**
- **Repeat experiment**
- **Compare results**
- **Share findings**
Scientific Workflows: ASAP

- Automated execution of computational experiments
- Scalable, fault-tolerant via distributed, parallel platforms
- Adopt, adapt, share, reuse, archive
- Provenance: traceable processing history, data lineage, wf evolution
Graph Reachability Encoding

- Tree cover encoding [Agrawal et al.]
- Efficiently determine ancestors and descendants