GoldenTrail: Retrieving the Data History that Matters from a Comprehensive Provenance Repository

Paolo Missier, Newcastle University, UK
Bertram Ludäscher, Saumen Dey, Michael Wang, Tim McPhillips, UC Davis, USA
Shawn Bowers and Michael Agun, Gonzaga University, USA
Ilkay Altintas, UC San Diego, USA

IDCC
Bristol, 6-7 Dec. 2011
Prologue: DCC "REPRISE" workshop, 2009

Full-fledged data-mediated collaborations

1. exp. A
   - workflow A + input A
   - Research Object A
   - result A
   - datasets A
   - result A
   - prohance A

2. exp. B
   - workflow B + input B
   - Research Object B
   - result B
   - datasets B
   - result B
   - prohance B

result A → input B
“Virtual experimental science” (DCC’09)

Full-fledged data-mediated collaborations

Provenance composition accounts for implicit collaboration
A provenance trace is an account of the history of a data item through multiple processing steps.

- Instrumental to verification and reuse of results -- Trustworthiness
- Enabler for “reproducible science” [1]

Prior work on provenance composition

2010: the DataTree Of Life summer project [2]

• Provenance stitching:

• Multiple, independently produced provenance traces expressed using the Open Provenance Model (OPM) can be “joined up” on shared datasets

• provided the data resides in a provenance-aware data repository.

Limitations:

– automated “stitching” requires data ID mapping and provenance-aware data copy operations

– in general, it requires human intervention

A broader vision

- Experimental science is explorative and evolutionary
  - many experiments, few will succeed
  - from parameter sweeps to changes in methods

- E-science infrastructure should be able to capture the exploration process in addition to the “good” results
  - Implicit collaboration becomes “just” a special scenario
A broader vision

- Experimental science is explorative and evolutionary
  - many experiments, few will succeed
  - from parameter sweeps to changes in methods

- E-science infrastructure should be able to capture the exploration process in addition to the “good” results
  - Implicit collaboration becomes “just” a special scenario

- **Golden Data**: the dataset(s) that scientists decide to share/publish
- **Golden Trail**: an account of how the Golden Data was obtained
  - a view over the provenance of the entire experiments history
  - describes a virtual experiment
Approach: a generalized provenance base

**PBase Requirements**

- Account for multiplicity of
  - workflow specifications and runs
  - workflow models
  - users

- Capture details of every execution into a persistent provenance repository

- Let scientists upload new provenance traces

- Support the provenance stitching process interactively

- Support queries on the provenance base to compute Golden Trails
Goal and associated technical challenges

**Goal:**
To offer an extensible framework for building PBases

- The Open Provenance Model is adequate for describing traces of workflow execution: “trace-land”
  - to be superseded by PROV-DM, currently W3C Public Working Draft (*)

- But we also need to record workflow specifications: “workflow-land”
  - by supporting multiple heterogeneous workflow models
  - e.g. ASKALON, Galaxy, Kepler, Taverna, Pegasus, Vistrails, etc.
  - currently only Kepler (UCSD, UC Davis), Taverna (myGrid, UK) supported

- Integration with the DataONE data preservation architecture
  - Provenance base as a new type of Member Node

(*) FPWD as of October, 2001: http://www.w3.org/TR/2011/WD-prov-dm-20111018/
D-OPM - a minimal model

- Trace-land inspired by the OPM
- Workflow-land inspired by Janus [1]

- Actor, a single computational step within a workflow
- Run: a single execution of an entire workflow
- Actor invocations: executions of individual steps that either Use or Generate Data Items
- Attribution: reference to users who run the workflow and thus “own” the traces.

UI: upload a new trace
Trace Parser
  – maps native formats to D-OPM
Graph Visualization
  – displays provenance graphs
Data Store: provenance store
Provenance queries

- Exploit the synergy between workflow-land and trace-land

Data-level and actor-level queries

Ancestor / Descendant queries (backwards / forward traversal)

Find all **Actors** that contributed to / impacted the generation of D

Find all **data** D’ that contributed to / impacted the generation of D

Workflow-level queries

Find all data that flowed through a workflow W during one run R

User-related queries

Find all data items used / generated on behalf of a user

---

**Diagram**

- **d1**: Used
- **i1**: GenBy
- **d2**, **d3**, **i2**, **d4**, **i3**, **d7**, **d8**, **i4**, **d9**: Data and actors

**Nodes**

- **d1**: Data
- **i1**: Information
- **d2**, **d3**, **d4**, **d5**, **d6**, **d7**, **d8**, **d9**: Data
- **i2**, **i3**, **i4**: Information

**Connections**

- **d1** used by **i1**
- **d2** used by **i1**
- **d3** used by **i1**
- **i1** genBy **d3**
- **d4** genBy **i3**
- **d7** genBy **i3**
- **i4** used by **d8**
- **d8** used by **i4**
- **d9** used by **i4**
Provide user name and the workflow name

Workflow system (e.g. Kepler, COMAD, Taverna, etc)

Browse the trace file to be loaded
Select provenance detail level and dependency type

Filter results using conditions

Add additional conditions

Query conditions
In tabular format

<table>
<thead>
<tr>
<th>Start Node</th>
<th>End Node</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>i1:1</td>
<td>Input</td>
</tr>
<tr>
<td>d2</td>
<td>i2:1</td>
<td>Input</td>
</tr>
<tr>
<td>i1:1</td>
<td>d2</td>
<td>Output</td>
</tr>
<tr>
<td>i2:1</td>
<td>d3</td>
<td>Output</td>
</tr>
</tbody>
</table>

In graphical format
• GoldenTrail: a “Provenance Base” for workflow-related datasets
  – across users
  – across workflow models
  – across sessions
  – dedicated provenance model and query layer

• State:
  – early prototype completed (summer 2011) [1]

• Ongoing work within the DataONE project, Provenance Working Group
  – PBase to be integrated into DataONE as Member Node
  – Ongoing engagement with the scientific workflow community
    • get buy-in on the PBase idea
    • collect feedback on current prototype
    • collect additional use cases