Challenges in managing implicit and abstract provenance data: experiences with ProvManager

Anderson Marinho, Marta Mattoso, Cláudia Werner, Vanessa Braganholo, Leonardo Murta

Federal University of Rio de Janeiro (UFRJ), Brazil
Fluminense Federal University (UFF), Brazil
Problem/Motivation

• Some challenges in managing provenance
  • Which provenance data should be gathered?
    • Open Provenance Model is a possible solution

• How to capture provenance data?
  • Three levels: Workflow, Operating System, Activity

• How to manage provenance in workflows that are executed by different execution environments (distributed environments)?
ProvManager - Overview of provenance gathering strategy

1. Collect the workflow specifications from SWfMS
2. Publish the workflow specifications
3. Obtain the adapted workflow specifications
4. Load the adapted workflow specifications in the SWfMS
5. Run
6. Provenance data sent by the activities

[SWF 2009; IPAW, 2010]

Scientist ➔ Experiment

Taverna ➔ Keplar

VisTrails

ProvManager
Workflow instrumentation

Original workflow

A

B
Workflow instrumentation

Original workflow

Instrumenting workflow

PGA

PGA

PGA
Workflow instrumentation

Original workflow

Instrumenting workflow

Wrapping activities
Workflow instrumentation

Original workflow

Instrumenting workflow

Wrapping activities

Adapted workflow
There are still some problems…

- Implicit provenance data
  - Difficulty in gathering provenance data when these are not explicitly declared in the workflow specification

- Lack of higher provenance abstraction levels
  - Concrete workflow related provenance data are not enough to help scientists in the experiment analysis
    - Some scientists may not be used to such information
Implicit provenance data

SWfMS domain

OS domain
Lack of higher provenance abstraction levels

**Problem:** scientist does not easily relate data from one abstraction level to the other

Questions such as “what is the result data of the ‘analysis of platform movements’ activity?” can not be easily answered.
Some ideas…

- Implicit data
  - Adopt a OS level provenance gathering mechanism to work together with the PGA in the ProvManager
  - There is a similar initiative in VisTrails (Koop et al. (2010))

- Lack of higher provenance abstraction levels
  - Create a “conceptual provenance data” model
    - Salayandia and da Silva (2010) propose something similar
  - Map this model to the existing “concrete provenance data”
Challenges in managing implicit and abstract provenance data: experiences with ProvManager

Anderson Marinho, Marta Mattoso, Cláudia Werner, Vanessa Braganholo, Leonardo Murta

Federal University of Rio de Janeiro (UFRJ), Brazil
Fluminense Federal University (UFF), Brazil