A Workflow PROV-Corpus based on Taverna and Wings

Khalid Belhajjame\textsuperscript{1}, Jun Zhao\textsuperscript{2}, Daniel Garijo\textsuperscript{3}, Aleix Garrido\textsuperscript{4}  
Stian Soiland-Reyes\textsuperscript{1}, Pinar Alper\textsuperscript{1}, Oscar Corcho\textsuperscript{3}  
\textsuperscript{1}School of Computer Science, University of Manchester, UK. \{khalidb,soiland,alper\}@cs.manchester.ac.uk  
\textsuperscript{2}Department of Zoology, University of Oxford, UK. jun.zhao@zoo.ox.ac.uk  
\textsuperscript{3}Facultad de Informática, Universidad Politécnica de Madrid, Spain. \{dgarijo, ocorcho\}@fi.upm.es  
\textsuperscript{4}iSOCO, Spain. agarrido@isoco.com

1. SUMMARY  
We describe a corpus\textsuperscript{1} of provenance traces that we have collected by executing 120 real world scientific workflows. The workflows are from two different workflow systems: Taverna \textsuperscript{5} and Wings \textsuperscript{3}, and 12 different application domains (see Figure 1). Table 1 provides a summary of this PROV-corpus.  
The information in the provenance traces is mostly specified using the PROV-O ontology \textsuperscript{4}. Terms from other vocabularies, including the Research Object Model\textsuperscript{2} \textsuperscript{1} and the Open Provenance Model for Workflows (OPMW)\textsuperscript{3} \textsuperscript{2}, are also used to associate the provenance traces with descriptions about the corresponding workflows.

Table 1: Information about the PROV-corpus.  
| Data format | RDF |  
| Data model | PROV-O |  
| Size | 360 Megabytes |  
| Tools used for generating provenance | Taverna and Wings provenance plug-ins |  
| Domain | see Figure 1 |  
| Submission group | Wf4Ever-Wings |  
| License | Creative Commons Attribution 3.0 Unported |  

2. CORPUS CREATION SETUP  
The selected workflows were executed within the system where they were designed (either Taverna or Wings) and the provenance traces of these runs were exported using the native plugins of the two workflow systems, respectively. All workflows were executed at least one time. In total, we collected the provenance traces of 198 workflow runs. It is also worth mentioning that not all the workflows finished their execution successfully. 30 workflow runs out of 198 failed for different reasons: unavailability of third party resources, illegal input values, etc. We chose to include the traces of failed workflow runs, since they can be particularly relevant to researchers who are investigating topics such as incomplete provenance, workflow decay, etc.

3. APPLICATION  
We expect a wide range of applications to be built from our provenance traces. Some exemplars are listed below:  
i) Identification of dependencies between data products and processes: provenance traces can be used to identify the process that generated a given data product, and how it was derived from other data products in order to identify dependencies.  
ii) Debug workflow executions: the PROV-corpus can be used to identify the processes that are responsible for workflow failure and detect the steps in the workflow that were affected.  
iii) Detection of workflow decay: provenance traces captured over time can be used to monitor and compare the results generated by the same workflow template. Such traces can be used to detect changes in workflow results and/or to repair a failed workflow by using results from previous runs.

4. EXEMPLAR PROVENANCE QUERIES  
We provide a list of exemplar provenance queries that can be issued against our provenance corpus:
1. What are the workflow runs available, and what is their start and end time?
2. What are the workflow runs associated with a given workflow template, and how many of them failed?
3. What are the workflow runs of a given workflow template, and what are the inputs they used and the outputs they generated?
4. How many process runs are associated with a given workflow run, what is the start and end time of each one of them (only available in Taverna provenance logs), and what are the inputs they used and the outputs they generated?
5. Who executed a given workflow run?
6. What are the services executed as a result of the execution of a given workflow run? (only available in Wings provenance logs).

5. COVERAGE OF PROV TERMS

Our provenance traces have been specified using mostly the prov:Activity and its subproperties, e.g., prov:used, etc. in Taverna, and翅系统, except for prov:actedOnBehalfOf (the only chain of responsibility we have is between the user and the software executing the workflow, and it is not recorded) and prov:wasDerivedFrom (data derivation relationships cannot be asserted easily without a proper understanding of the exact function of each process of a workflow run. This is part of our ongoing work).

Table 2 shows additional coverage of PROV terms for each workflow system. Those entries marked with a star imply that the PROV statement is not directly asserted in the traces, but it can be inferred.

Table 3: Coverage of Additional PROV Terms.

<table>
<thead>
<tr>
<th>PROV Terms</th>
<th>Support by the Systems</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>prov:Bundle</td>
<td>Wings</td>
<td></td>
</tr>
<tr>
<td>prov:Plan</td>
<td>Taverna* and Wings</td>
<td>prov:hadPlan is used in Taverna, instead of prov:Plan</td>
</tr>
<tr>
<td>prov:wasInfluencedBy</td>
<td>Taverna* and Wings</td>
<td>No explicit influence relationship is expressed in Taverna, but only its subproperties, e.g., prov:used, etc.</td>
</tr>
<tr>
<td>prov:hadPrimarySource</td>
<td>Wings</td>
<td></td>
</tr>
<tr>
<td>prov:atLocation</td>
<td>Wings</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows additional coverage of PROV terms for each workflow system. Those entries marked with a star imply that the PROV statement is not directly asserted in the traces, but it can be inferred.

6. MAINTENANCE AND FUTURE WORK

We expect new provenance traces will continue to be added to this corpus. Future work includes providing access to the corpus via a SPARQL endpoint and web interfaces, maintaining the corpus to keep it aligned with possible changes in PROV-O, Research Object and OPMW ontologies. We also intend to investigate further interoperable queries to retrieve provenance results from both workflows systems, and improve the corpus in the light of community feedback.

Acknowledgments

This work is supported by the myGrid platform grant, the Wf4Ever (http://www.wf4ever-project.org) EU project (FP7-270192), and the Wings project (http://www.wings-workflows.org).

7. REFERENCES