A Workflow PROV-Corpus based on Taverna and Wings

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

1. SUMMARY

We describe a corpus[1] of provenance traces that we have collected by executing 120 real world scientific workflows. The workflows are from two different workflow systems: Taverna [5] and Wings [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 [4]. Terms from other vocabularies, including the Research Object Model[2] [1] and the Open Provenance Model for Workflows (OPMW)[3] [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 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.

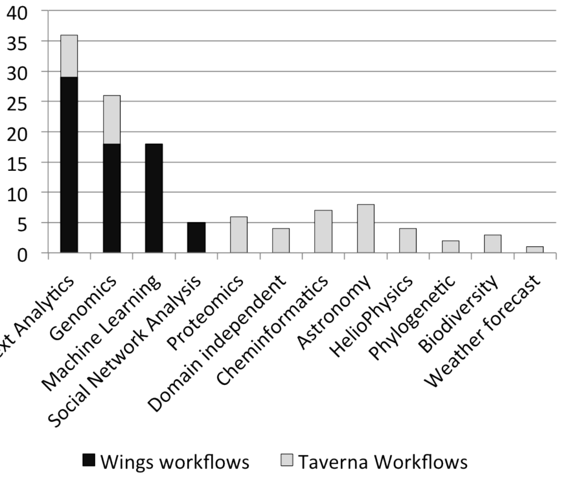

Figure 1: Domains of workflows.

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] `https://github.com/provbench/Wf4Ever-PROV`
[2] `http://wf4ever.github.com/ro-primer`
[3] `http://www.opmw.org/model/OPMW/`

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-O ontology and an extensions of it, namely the wfprov ontology [1] and OPMW, for expressing workflow-specficic provenance information. Table 2 and Table 3 show the coverage of the PROV terms by both workflow systems. As illustrated in Table 2, most of the starting-point PROV-O terms [4] are covered by the two systems, 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 ongoin work).

Table 2: Coverage of Starting-point PROV Terms.

PROV Terms	Support by the Systems	Comments
prov:Activity	Taverna and Wings	
prov:Agent	Taverna and Wings	
prov:Entity	Taverna and Wings	
prov:actedOnBehalfOf	-	
prov:endedAtTime	Taverna	Activity start and end not recorded in Wings provenance traces
prov:startedAtTime	Taverna	Same as above
prov:used	Taverna and Wings	
prov:wasAssociatedWith	Taverna and Wings	
prov:wasAttributedTo	Wings	No direct attribution is recorded in Taverna provenance traces
prov:wasDerivedFrom	-	
prov:wasGeneratedBy	Taverna and Wings	
prov:wasInformedBy	Taverna	Used to express the connection between sub-workflows

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.

Table 3: Coverage of Additional PROV Terms.

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

[4] `http://www.w3.org/TR/prov-o/#description-starting-point-terms`

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 Wf4Ever (http://www.wf4ever-project.org) EU project (FP7-270192) and the Wings project (http://www.wings-workflows.org).

7. REFERENCES

[1] Khalid Belhajjame, Oscar Corcho, Daniel Garijo, Jun Zhao, Paolo Missier, David Newman, Raul Palma, Sean Bechhofer, Esteban Garcia-Cuesta, Jose-Manuel Gomez-Perez, Graham Klyne, Kevin Page, Marco Roos, Jose Enrique Ruiz, Stian Soiland-Reyes, Lourdes Verdes-Montenegro, David De Roure, and Carole Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceedings of Sepublica2012*, pages 1–12, Hersonissos, 2012.

[2] D. Garijo and Y. Gil. A new approach for publishing workflows: Abstractions, standards, and linked data. In *Proceedings of the 6th workshop on Workflows in support of large-scale science*, pages 47–56, Seattle, 2011. ACM.

[3] Y. Gil, V. Ratnakar, J. Kim, et al. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1):62–72, 2011.

[4] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. Technical report, 2012.

[5] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna, reloaded. In M Gertz, T Hey, and B Ludaescher, editors, *Procs. SSDBM 2010*, Heidelberg, Germany, 2010.