Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data

Daniel Garijo
OEG-DIA, Facultad de Informática, Universidad Politécnica de Madrid
dgarijo@delicias.dia.fi.upm.es

Yolanda Gil
Information Sciences Institute and Department of Computer Science,
University of Southern California
gil@isi.edu

Abstract. Provenance models are crucial for describing experimental results in science. The W3C Provenance Working Group has recently released the PROV family of specifications for provenance on the Web. While provenance focuses on what is executed, it is important in science to publish the general methods that describe scientific processes at a more abstract and general level. In this paper, we propose P-PLAN, an extension of PROV to represent plans that guided the execution and their correspondence to provenance records that describe the execution itself. We motivate and discuss the use of P-PLAN and PROV to publish scientific workflows as Linked Data.

Keywords: provenance, plan, PROV, scientific workflows, Linked Data.

1 Linked Data and Scientific Processes

A crucial element of Linked Data for science is the publication and sharing of scientific processes to document how scientific results are generated. There are many kinds of scientific processes that could be shared as Linked Data. Recent research in this area includes publishing scientific workflows that capture data analysis processes [2] documenting scientific experiments, structuring claims and conclusions in scientific publications (as in SWAN[1]), and more recently organizing Research Objects [1]. For all these efforts, it is important to publish how scientific processes were executed, but it is also important to publish how they were planned. For example, assays are used to describe an experimental procedure in a biology laboratory and are very precise, while protocols are more general descriptions of those procedures. In essence, these protocols represent the plans that are followed in carrying out the assays.

There is now an emerging standard for publishing the provenance of processes on the Web. This standard could be used for the publication of scientific processes, improving interoperability across scientific software as well as interoperability with

[1] http://www.w3.org/TR/hcls-swan/

adfa, p. 1, 2011.

other web provenance. The W3C PROV model[2] describes the provenance of objects (prov:Entities) as a record of assertions about the steps (prov:Activities) that generated them and the entities used in those steps. Provenance describes past execution, but does not offer a vocabulary to express the plan that the execution was supposed to follow. In terms of our example above, provenance vocabularies are appropriate for describing assays once they are executed, but are not designed to describe protocols.

Therefore, in addition to the provenance record, it is often desirable to publish the plan that was followed during the execution. This would allow the provenance record to include what was envisioned would happen prior to the execution. Publishing the plan has several benefits: 1) the plan can provide a higher-level, more abstract description of what was executed, which improves understandability and facilitates reuse in future situations; 2) the plan can describe the expectations for the execution, which can then be contrasted with the provenance to detect deviations and correct abnormalities. Acknowledging this need, PROV includes the term "prov:Plan". However, it does not elaborate any further how plans can be described or related to other provenance elements of the execution.

Several vocabularies have been proposed to represent different aspects of scientific processes, including SWAN and OBI[3]. Ideally, the PROV standard would be adopted for all provenance aspects of these vocabularies, enabling interoperability of their records. However PROV will not address aspects concerned with methods and abstract plans, which would be useful for interoperability of linked science data.

In this paper we propose to address this necessity by extending PROV with P-PLAN, a vocabulary for describing abstract scientific workflows as plans. This proposal builds on our previous work on OPMW[4] where we published scientific workflows compliant with OPM as Linked Data [2].

2 Representing Plans and their executions

Developing a vocabulary for plans is a daunting task. Plan representations vary widely in formalism and complexity, from simple graph-based plan representations to dynamic logics with quantification and temporal reasoning. Some plan representations encompass meta-planning (e.g., to decide what goals to take on), scheduling (to allocate resources to steps), interoperability (with special focus on reuse[5]) and failure handling. The DOLCE [3] ontology includes a representation of plans that is a superb synthesis of this representational diversity.

In order to expose the relation between the plan and the execution, we wanted to have plans aligned closely with provenance records. Because provenance assertions in PROV can be seen as a direct acyclic graph of steps and entities used and generated by them, we set out to develop a simple plan vocabulary for plans that can be represented also as a directed acyclic graph of steps and relevant entity descriptions. This is

[2] http://www.w3.org/TR/prov-dm/

[3] http://obi-ontology.org/

[4] http://www.opmw.org

[5] http://www.shiwa-workflow.eu/web/guest

a limitation of our approach, but we believe that this model can capture a significant amount of workflows (which are often described as simple pipelines) and other types of processes. The initial model that we propose here could be extended later on with more complex plan (and workflow) constructs.

2.1 Plan execution as PROV

PROV describes the usage and generation of entities through two main properties: prov:wasGeneratedBy (an Entity wasGeneratedBy an Activity) and prov:used (an Activity used an Entity for the execution). The agents responsible for the execution are linked to the activity as prov:Agent with the property prov:wasAssociatedWith. All properties can be qualified with prov:Roles. Provenance assertions can be grouped in prov:Bundles, so that provenance can be asserted for the bundle.

In PROV plans are defined as entities associated with an agent and an activity. PROV does not specify anything further about plans and how they correspond to parts of the execution, as it is considered out of the scope of the model for provenance.

2.2 Extending PROV to represent plans

Figure 1 shows an overview of P-PLAN and how it relates to PROV assertions, showing plans at the top and plan executions at the bottom. The provenance of the execution is entirely captured with PROV. Entity, activity and bundle concepts are subclasses of PROV classes (p-plan:Bundle, p-plan:Entity and p-plan:Activity) to be able to represent their relationship to the parts of the plan (p-plan:correspondsTo property for activities and entities and prov:wasInfluencedBy to connect the bundle representing the execution to the plan).

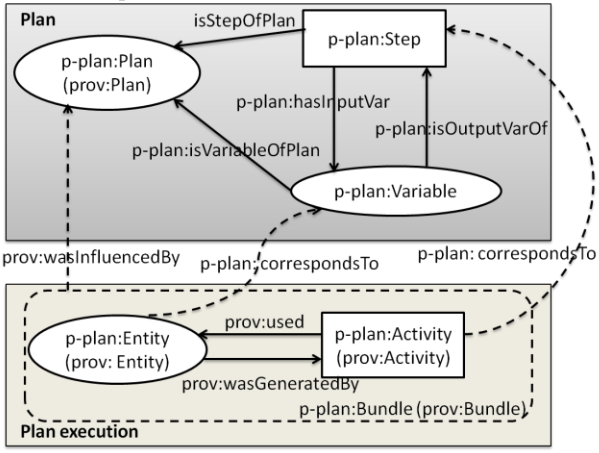

Fig. 1. P-PLAN as an extension of PROV to describe plans.

p-plan:Plan is a subclass of prov:Plan. p-plan:Steps represent the planned execution activities. Plan steps may be bound to a specific executable step or refer to a class of steps, providing an abstraction layer over the execution. As a result, a plan step could be carried out in different ways in different executions of the same plan. A step may not have a corresponding activity, (as in an execution failure). p-plan:Variables represent the inputs of the steps and can have properties (i.e., type, restrictions, metadata, etc.). p-plan:Steps have p-plan:Variables as input and p-plan:Variables are output of p-plan:Steps. Both of them are associated to a p-plan: Plan. The relation of the plan with agents is not specified P-PLAN, since it can be modeled with PROV.

3 Publishing scientific workflows

Plans can be used to represent abstract workflows that describe reusable templates of computations. p-plan:Step can be used to describe workflow steps, and p-plan:Variable can be used to represent input and output datasets of the step as well as parameters of the computation. PROV can be used to represent the execution, including each step as a prov:Activity and each dataset as a prov:Entity.

PROV also allows defining chains of responsibility for agents. This is very useful when publishing a workflow execution, since the user triggering the submission delegates to the workflow system(s) and in turn to the execution engine.

An example of what this modeling would enable is linking across different executions of workflow templates from different platforms and different domains published as Linked Data. For example, the query below would return all abstract workflows (plans) in which a given entity (?entity) has been used when executing them. This helps to understand the usage of a dataset across workflow executions and how different workflow templates relate to each other.

```
SELECT DISTINCT ?plan WHERE {
  ?entity a p-plan:Entity,prov:Entity;
         p-plan:correspondsTo ?templVariable.
   ?templVariable a p-plan:Variable;
         p-plan:isVariableOfPlan ?plan.}
```

4 Conclusions

We propose P-PLAN as a vocabulary for publishing plans and linking executions to them as a proposed extension to the PROV standard. P-PLAN is generic and could be extended to represent more complex plans, (e.g., plans with hierarchical decompositions). We consider it a necessary step in scientific workflow publication as Linked Data, and crucial to understand the method performed in an experiment.

We are currently exporting the provenance of workflow templates and executions with both OPMW and PROV, with a mapping to P-PLAN. In future work, we plan to map P-PLAN to broadly used ontologies like DOLCE and other vocabularies that describe scientific processes, such as SWAN and OBI.

References

1. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delder_eld, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Su_, S., Goble, C.: Why linked data is not enough for scientists. Future Generation Computer Systems, 2011.
2. Garijo, D., and Gil, Y. “A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data”. In Proc. WORKS’11, Seatle, WA, 2011.
3. Oberle, D., Lapmarter, Grimm, S., Vrandecic, D., Staab, S, Gangemi, A. “Towards Ontologies for Formalizing Modularization and Communication in Large Software Systems”. Applied Ontology 1,163-202