

ProtoProv: A provenance mapping system

RPI/TWC Team

5/28/2009

Outline

- Motivation
- Overview of ProtoProv Architecture
- Strategies for recording provenance data from Provenance Challenge workflow

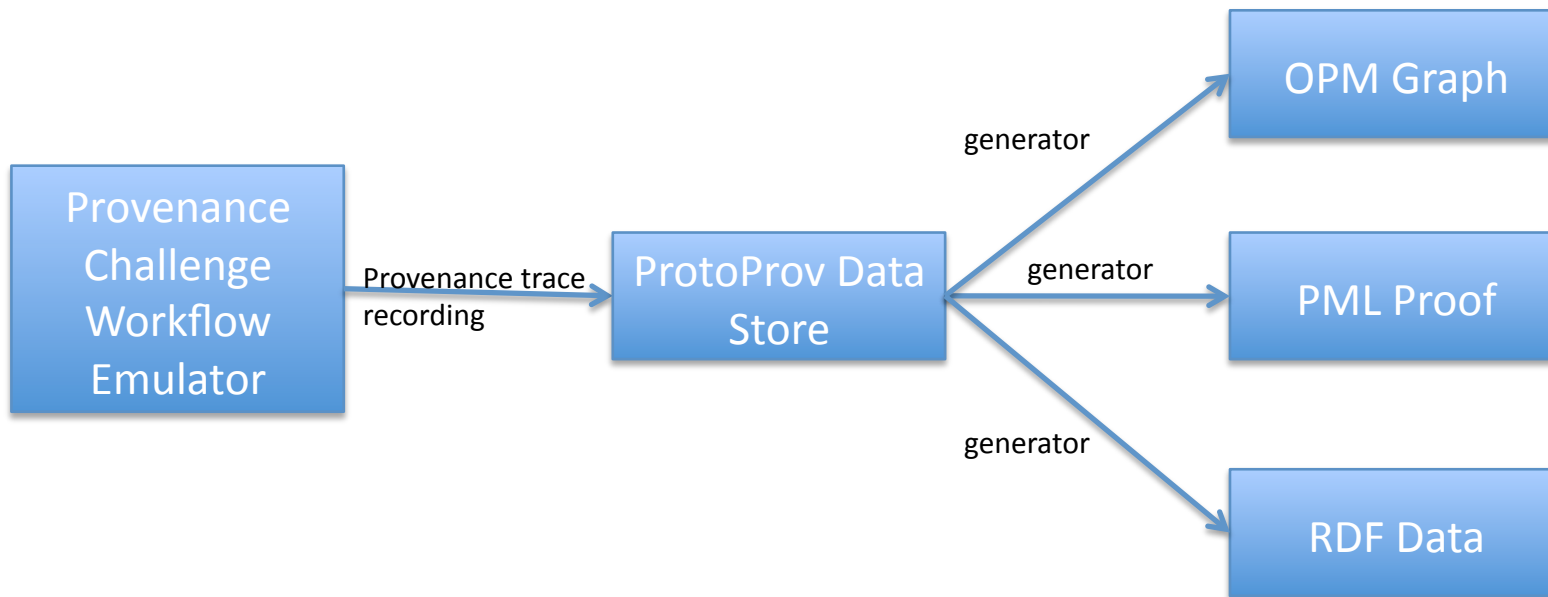
Motivation

- Both PML and OPM have emerged as viable provenance standards
- Useful technologies have been developed for both standards, which are not designed for alternate provenance standards
- Furthermore, there are provenance relations that can be expressed in OPM but not PML, and vice versa

What is ProtoProv?

- A system designed to facilitate mapping between the PML2 and OPM provenance standards
 - Based on the ProtoProv ontology, at: <http://www.cs.rpi.edu/~michaj6/PPV2.owl>
- Semantic Web technologies provide mechanism for mapping between standards, as well as running queries on imported provenance data

Architecture



Our generation of OPM

- In the ProtoProv ontology, a direct mapping has been established to all OPM concepts
- In the Provenance Challenge workflow, we follow certain procedures for capturing:
 - Halting states
 - Data dependencies
 - Control flow

Halting State

- Denoted by a process endState
- endState produces an artifact endValue, which takes one of two values:
 - True: for a successful workflow completion
 - False: for a failed control flow check

Data Dependencies

- When a process P in the workflow directly uses an artifact A , we capture this through the Used relation
- P Used A

Control Flow Dependencies

- Each control flow check denoted by a process C
- Links between C and following states handled through the `wasTriggeredBy` relationship
- For a passed control flow check: (`nextProcess wasTriggeredBy C`)
- For a failed control flow check: (`endState wasTriggeredBy C`)

Generation of PML

- The ProtoProv ontology was designed to capture all OPM relations used by alternate teams in the Provenance Challenge
- In turn, this ontology is being used for constructing mappings back to PML
 - Not all OPM concepts have been successfully mapped back to PML yet

Issues with PML Generation

- No effective way to capture wasTriggeredBy relationships present in OPM data
- Therefore, our strategy for capturing control flow in the workflow won't map to PML

Running Semantic Queries

- The ProtoProv ontology helps facilitate the generation of RDF data from an OPM graph or PML proof.
- In turn, this RDF data can be stored and queried using mainstream Semantic Web technologies