TWPL: A PROTOCOL FOR MANAGING RDF-ENCODED PROVENANCE RECORDS

By

James Roller Michaelis

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: COMPUTER SCIENCE

Approved:

Deborah L. McGuinness, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

April 2011
(For Graduation May 2011)

ABSTRACT

The term provenance is defined by the Oxford English Dictionary as deriving from the french verb *provenir*, meaning to come forth, originate. Accordingly, the following definitions of provenance are provided: (i) The fact of coming from some particular source or quarter, origin, or derivation; and (ii) The history of the ownership of a work of art or an antique, used as a guide to authenticity or quality.

Currently, digital systems are adopting provenance tracking to supplement their data products. These digital systems are becoming increasingly complex and distributed, sometimes forming networks with a requirement of intersystem communication. To achieve the goal of provenance interoperability, domain independent models of provenance, such as the Open Provenance Model (OPM) and Proof Markup Language (PML), are being developed and in turn adopted by digital systems. In this thesis, two challenges are identified, based on observations of current application and evaluation of provenance models:

The Usability Challenge: Many kinds of useful information, such as system-specific metadata or domain knowledge (termed *supplemental information*), cannot be easily represented with a provenance model. Often, supplemental information will be needed for both humans and computers to effectively interpret provenance records.

The Multiple Models Challenge: Many provenance models have emerged under slightly different contexts. While sharing a common goal of promoting interoperability, these models also tend to have slight design variations. For this reason, comparing and integrating provenance records based on different provenance models becomes difficult.

To address these challenges, a provenance logging protocol - called Tetherless World Provenance Logging (TWPL) - is presented and discussed. TWPL is defined to leverage Semantic Web technologies for managing provenance records, encoded based on the Resource Description Framework (RDF) specification. Two types of logging are supported by TWPL:

Workflow Logging: This generates a provenance record based on the execution of a specified workflow (i.e., an ordered sequence of processing activities). In turn, the provenance record is encoded using a both a specified provenance model and accompanying supplemental information.

Mapper Logging: This takes a provenance record, encoded using one provenance model, and converts it to a record based on an alternate model - based on a conceptual mapping between the two provenance models.

Development of the TWPL protocol was based in part on earlier work conducted as part of RPI's participation in The Third Provenance Challenge (PC3) - a workshop aimed at testing OPM's ability to enable the exchange of provenance records across digital systems. During PC3, participating teams used varying systems to attempt the following tasks:

1. Load and execute a common scientific workflow, based on the Panoramic Sky Survey and Rapid Response System (PAN-STAARS) - a sky survey system designed to scan the visible sky once per week for evidence of near earth objects.

2. During (1), track the provenance of any corresponding data products generated.

3. Following (1), answer a common set of assigned questions based on the provenance generated in (2).

4. Export the provenance generated in (2), encoded based on OPM.

5. Import provenance from other systems, and attempt to answer the assigned questions from (3) over them - as though the provenance were generated natively.

To demonstrate TWPL in the context of prior work for PC3, a Java-based implementation is provided - including components both Workflow and Mapper Logging. The Workflow Logging component is applied as follows:

1. Encoding an RDF-based record of the PAN-STAARS workflow's execution, based on both OPM and supplemental information from the workflow's documentation.

2. Answering the assigned questions over the created OPM record using the RDF querying language SPARQL.

Likewise, the Mapper Logging component is applied as follows:

1. Using the OPM-based provenance record generated via Workflow Logging, creating a corresponding record based on the Proof Markup Language (PML) provenance model - using a conceptual mapping for converting OPM-based records to PML.

2. Answering the assigned questions from PC3 over the created PML record through SPARQL querying.

Based on the TWPL protocol and corresponding implementation, challenges presently unaddressed (such as trust in conceptual mappings used by TWPL's Mapper Logging) are highlighted. In turn, future work for refining TWPL is discussed, based on these challenges.