Towards Interoperable Provenance Publication on the Linked Data Web

Jun Zhao
Department of Zoology
University of Oxford
South Parks Road, Oxford
OX1 3PS, United Kingdom
jun.zhao@zoo.ox.ac.uk

Olaf Hartig
Institut für Informatik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin, Germany
hartig@informatik.hu-berlin.de

ABSTRACT

Provenance provides vital information for evaluating quality and trustworthiness of information on the Web. To achieve this we must have access to semantically interchangeable provenance information and an agreement on where and how this information is to be located. The ongoing W3C Provenance Working Group provides a promise towards leveraging these problems. In this position paper, we provide an overview of how the upcoming standards and the existing vocabularies and publication approaches could fit together so that we achieve an optimal interoperability now and in the near future. Because the standardization is an ongoing effort, any analysis results presented in this paper are positional and are aimed at communicating the latest development of the working group to the community.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: General

General Terms

Linked Data, Interoperability

Keywords

Provenance, Linked Data, Semantic Web, RDF

1. INTRODUCTION

Provenance information about a resource provides information about its origin, such as who created it, when it was modified, or how it was created. It has been widely accepted that this kind of information is vital for evaluating quality and trustworthiness of information on the Web [5, 6]. Interoperability of provenance information is essential for creating a trustworthy Web of Data. Given the nature of distributed data publication and access on the Linked Data Web, provenance information about data can be published by any parties, according to any provenance vocabularies or publication approaches. To evaluate quality of data on the Web, applications must be able to access information through different channels and make sense out of the diverse information described using languages of varied semantics. The ongoing standardization effort from the W3C Provenance Working Group provides a family of standards to leverage this problem. However, before these standards are eventually published and universally adopted, we must understand them in the context of existing provenance vocabularies and publication approaches in order to achieve the optimal interoperability now and in the near future.

There has been a sea of interest in providing provenance-related vocabularies, a summary of which can be found by the group report of the late W3C Provenance Incubator Group [10]. This position paper chose two of these vocabularies to compare their semantic interoperability with the PROV-O ontology [2], being standardized by the working group. The two chosen vocabularies are the OPMV (Open Provenance Model Vocabulary) [12], a lightweight implementation of the community Open Provenance Model [8], and the Provenance Vocabulary [6], another lightweight vocabulary targetted at Linked Data use cases. These two vocabularies were chosen because: 1) both of them were created with the needs of Semantic Web users in minds, 2) they were designed to cover a similar scope of motivation use cases as PROV-O; and 3) they share a largely similar modeling pattern as PROV-O.

Interoperability of provenance data requires not only an agreement on how provenance is represented but also a shared understanding about "what" is described. Researchers from the provenance community emphasize that provenance should provide a precise history of what happened that have led to *the particular state* of an object [8]. The state of an object can be characterised by a set of its attribute values. Resources on the Web are dynamic in nature and their attribute values can be changed at a volatile rate. The definition of the state of an entity should be driven by actual context, and it is hard to reach a universal agreement. For example, an Ajax web page reporting weather forecast of London can be updated regularly with its latest forecast data. Over this time the state of this web page can be regarded as fixed because its key features are not changed: at the same URL and always about London weather. It is sufficient to track who created this document without referring to the document at any specific time instant. However, in another context, changes to the forecast value could be regarded as a change to the state of the web page. Its provenance must include information about when the Ajax page was updated, how and etc.

If the definition of the state of an entity does not match the needs in hand, then we will not access sufficient provenance to recreate its historical record. For example, if a new state is not defined when the forecast data was up-

LDOW2012, April 16, 2012, Lyon, France.

dated then we cannot know how the document was updated with this data. Provenance is less "precise" in this context, even though its precision is sufficient for other context, e.g. knowing the creator of the document. Without an awareness of the co-existence of this "precise" v.s. "imprecise" provenance information on the Web, provenance data consumers could misinterpret the semantics of this information and make incorrect judgement. Hence, our analysis also highlights how the three vocabularies allow users to express provenance in a "state-ful" and "state-less" manner.

Another question that must be addressed towards achieving interoperable provenance on the Web is how to make this information accessible on the Web. Hartig and Zhao [6] have analyzed different possible ways of publishing provenance information onto the Web. But how can this information be discovered in the first place? The Provenance Access and Query (PAQ) working draft [9] from the W3C Provenance Working Group proposes a set of best practices for making provenance information discoverable. The second part of the position paper presents some recommended ways of publishing provenance information according to this specification in order to achieve interoperable provenance access on the Web.

Because these working drafts from the provenance working group are still work in progress, this position paper only provides an analysis as per the state-of-the-art. This is not an advocate of the working group deliverables, but rather a communication of the latest developments of the working group by positioning them in the context of existing work.

2. TERMINOLOGIES

Provenance-related terminologies are very diverse; for example, each of the three selected provenance vocabularies uses different terminology for modeling and describing provenance. To remove ambiguities this paper uses the set of terms introduced in the latest PROV Model Primer [3] and the PAQ working draft [9] released by the W3C provenance working group. The definitions and semantics of these terminologies are still subject to changes, and we are using them in a way as they were available by the time of writing.

- Entities, are the things "that one may ask the provenance of" [3].
- Activities, are "how entities come into existence and how their attributes change" [3] in a way that lead to existence of a new entity.
- Agents, are entities that take "an active role in an activity" by taking "some degree of responsibility" in that activity [3].
- Resources, refer to "whatever might be identified by a URI" as described by the Architecture of the World Wide Web [11].

3. THE PROVENANCE VOCABULARIES

Provenance vocabularies/ontologies provide the building blocks for describing provenance information on the Semantic Web. To achieve interoperable provenance descriptions we must understand the semantic interoperability of these building blocks. Previously the W3C Provenance Incubator group has conducted a thorough survey of the state-of-the-art provenance vocabularies and a mapping between them [10]. To align the chosen vocabularies, this survey used a list of terms from the Open Provenance Model (OPM) [8], a community provenance model. The analysis showed that there is a considerable correspondence among the vocabularies along the core concepts of agents, entities, and activities. It also identified some gaps in OPM for representing things like versions, containment between entities, etc.

For this position paper we picked two of these vocabularies, OPMV (Open Provenance Model Vocabulary) [12] and the Provenance Vocabulary [6], to compare their similarity with the PROV-O ontology that is being proposed and standardized by the W3C Provenance Working Group. Our analysis shows that the three vocabularies employ a common pattern for describing provenance, but have different perceptions with respect to entities whose provenance being described.

3.1 Describing Provenance

The W3C PROV Model Primer [3] points out that provenance could be viewed from three different perspectives:

- **Agent-oriented** provenance focuses on information describing the entities "involved in generating or manipulating the information in question".
- **Object-oriented** provenance focuses on tracing the entities contributing to the existence of another entity.
- **Process-oriented** provenance focuses on tracking the "actions and steps taken to generate" an entity whose provenance information is being described.

Together, through these three perspectives, we capture the 'who', 'what', 'when' and 'how' information, as shown in Figure 1. And this pattern of using the three core concepts of agent, entity and activity is repeatedly applied in the three selected provenance vocabularies, i.e. PROV-O, OPMV, and the Provenance Vocabulary. This forms a so-called process-centric modeling pattern, i.e. an activity class is always introduced to describe the creation or modification of an entity. A relationship between an entity and an agent must be stated by explicitly describing the activity in which the agent is involved that leads to a modification of the entity. There is an exception for stating the relationship between entities, which can be directly stated without having to introduce an activity. This is sometimes regarded as a shortcut or as a data-centric view on top of the process-centric logs. In other provenance-related vocabularies, such as Dublin Core, such a process-centric pattern is not employed. Any statements can be directly associated with an object (be an entity or an agent) without having to make explicit the activities involved in their creation.

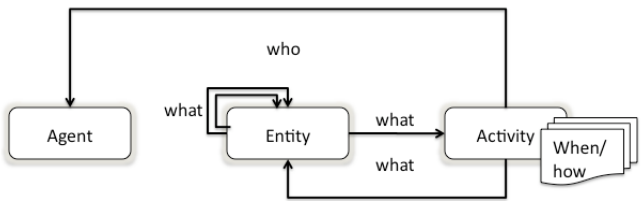

Figure 1: Describing provenance information from three perspectives.

Table 1: Definitions of agents and activities/processes in PROV-O and OPMV.

	PROV-O	OPMV
Agent	a type of entity that "takes an active role in an activity" by taking "some degree of responsibility" in that activity	a contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, or affecting its execution
Activity	"how entities come into existence and how their attributes change" in a way that lead to existence of a new entity	an action or series of actions performed on or caused by artifacts, and resulting in new artifacts.

Table 2: Properties for describing the provenance of an entity.

Descriptions of key properties	PROV-O	OPMV	The Provenance Vocabulary
represents the active involvement of agent in modifying the characteristics of the instance of an activity	wasAssociatedWith	wasControlledBy	performedBy/accessedService
express that an entity was used or consumed during an activity	used	used	usedData/usedGuideline
express that an entity was generated or created by an activity	wasGeneratedBy	wasGeneratedBy	retrievedBy/createdBy
express that the existence of one entity is (at least partly) due to another entity	wasDerivedFrom	wasDerivedFrom	

A further analysis shows that the three vocabularies also share a very similar semantics for their definitions of agents, activities, and related properties. With the latest revision the Provenance Vocabulary even positions itself as a specialization of PROV-O[1]. Tables 1 and 2 summarize correspondences of related concepts and properties from the three vocabularies. Apart from these commonalities, the vocabularies show a key difference in their notion about provenance *entities*, which directly impact on the expression of "precise" and "imprecise" provenance using these vocabularies.

3.2 State-ful v.s. State-less Provenance

Provenance metadata is expected to provide a faithful historical record of what happened. The metadata itself should be immutable and the entities whose provenance being described should be persistent to a particular state. The state of an object can be characterised by a set of its attribute values. If attributes charaterising the "state-ful" entity changed, it should be regarded as a new entity.

However, attributes that *characterise* a resource are subject to the context under which provenance is generated, and the application for which provenance is collected. For example, Listing 1 uses URI `<http://example.org/forecast/london>` to identify the daily weather forecast for London. For applications that are interested in understanding who provides this forecast, even though the forecast data is updated day by day, this URI is regarded as identifying the same entity. It is a "state-less" entity whose state, i.e. being accessible via a specific URI, remains unchanged over time. However, for applications that need to understand how the forecast data was generated everyday, the forecast data of each day needs to be treated as a different entity. From the example in Listing 1, applications are unable to access historical information that records exactly what happened everyday. To fix this, we need to refer to a "state-ful" entity that represents forecast of each particular day.

[1] `http://purl.org/net/provenance/ns-20120314`

Defining clear-cut states for resources on the Web is a challenging task, due to varied interpretation and context under which the data were published. As a standard for the Semantic Web community, PROV-O therefore allows the expression of provenance in both a state-ful and state-less manner, in order to provide a practical solution for a wider range of users in the community. OPMV and the Provenance Vocabulary, however, emphasize more explicitly the immutable nature of entities or artifacts. In OPMV, an `Artifact` is a general concept that represents an immutable piece of state; and it is impossible to express provenance metadata in Listing 1 using this concept. The Provenance Vocabulary ex-

```
1
2  @prefix prov: <http://www.w3.org/ns/prov-o/>
3  @prefix ex2: <http://example.org/2>
4
5  # provenance of London forecast on two different
       days
6
7  <http://example.org/forecast/london>
8    ex2:degree "-6"^^xsd:Integer ;
9    prov:wasAttributedTo  <http://bbc.co.uk> ;
10   prov:wasGeneratedBy [
11       rdf:type    prov:Activity ;
12       prov:used    <http://satellite_a> ;
13       prov:startedAtTime
             "2012-02-06T00:00:00"^^xsd:dateTime ] .
14
15 <http://example.org/forecast/london>
16   ex2:degree "0"^^xsd:Integer ;
17   prov:wasAttributedTo  <http://bbc.co.uk> ;
18   prov:wasGeneratedBy [
19       rdf:type    prov:Activity ;
20       prov:used    <http://satellite_b> ;
21       prov:startedAtTime
             "2012-02-07T00:00:00"^^xsd:dateTime ] .
```

Listing 1: Express provenance of the *state-less* London forecast entity using PROV-O.

```
@prefix prov: <http://www.w3.org/ns/prov-o/>
@prefix prv: <http://purl.org/net/provenance/ns#>
@prefix ex2: <http://example.org/2>

# provenance of London forecast on Feb. 6, 2012

<http://example.org/forecast_0602>
  ex2:degree "-6"^^xsd:Integer ;
  prov:wasAttributedTo  <http://bbc.co.uk> ;
  rdf:type prv:Immutable, prv:DataItem ;
  prv:retrievedBy [
      rdf:type    prv:DataAccess ;
      prv:accessedResource
           <http://example.org/id/forecast_0602> ;
      prv:completedAt
           "2012-02-06T00:00:00"^^xsd:dateTime ] ;
  prv:createdBy [
      rdf:type    prv:DataCreation ;
      prv:usedData    <http://satellite_a> ;
      prv:completedAt
           "2012-02-06T00:00:00"^^xsd:dateTime ] .

# provenance of London forecast on Feb. 7, 2012

<http://example.org/forecast_0702>
  ex2:degree "0"^^xsd:Integer ;
  prov:wasAttributedTo  <http://bbc.co.uk> ;
  rdf:type prv:Immutable, prv:DataItem ;
  prv:retrievedBy [
      rdf:type    prv:DataAccess ;
      prv:accessedResource
           <http://example.org/id/forecast_0602> ;
      prv:completedAt
           "2012-02-07T00:00:00"^^xsd:dateTime ] ;
  prv:createdBy [
      rdf:type    prv:DataCreation ;
      prv:usedData    <http://satellite_b> ;
      prv:completedAt
           "2012-02-07T00:00:00"^^xsd:dateTime ] .
```

Listing 2: Express provenance of *state-ful* London forecast using the Provenance Vocabulary.

tends PROV-O by introducing a concept `prv:Immutable`, that allows users to explicitly mark the immutable nature of an entity at a particular state. Using this concept, Listing 2 rewrites provenance of London forecast data by regarding daily forecast as a state-ful entity. Two separate URIs are created to identify London forecast from two separate days in order to provide a static record for each entity.

This subtlety must be considered when publishing provenance information for resources on the Web. These provenance for "state-ful" v.s. "state-less" entities are not two distinctive types of provenance. They are simply historical statements collected in different context, under different conditions. When a resource is state-ful instead of state-less is all relative speaking. What is indeed needed is an interoperable way to refer to these static, state-ful entities, such as the forecast of each individual day, and their dynamic counterpart (i.e. the daily forecast data as a general concept), to retrieve their provenance information.

4. PROVENANCE PUBLICATION FOR LINKED DATA RESOURCES

To make provenance information accessible on the Linked Data Web in an interoperable way we must have an agreement on how provenance is made available (e.g. embedded in an RDF graph or retrievable via links), and where to look for this provenance information.

Hartig and Zhao [6] propose several choices on where to make provenance available for Linked Data, such as including provenance information in the voiD (Vocabulary of Interlinked Datasets) [1] description about a linked dataset, or in the RDF graph that is served in response to an HTTP GET operation. All these proposed ways are embedding approaches. Although locating provenance information in these cases is made easy, it can however introduce a performance problem if the number of provenance triples is large or even outnumbers the actual triples that describe the resource itself. We should have an alternative choice that allows us to *link* resources to provenance descriptions through a URI identifying these descriptions. Such a URI is called a *provenance URI* in the PAQ document [9].

The PAQ working draft [9] from the provenance working group aims to specify best practices for enabling provenance information to be located in an agreed way. It recommends at least two ways to link provenance descriptions with entities: one is to use HTTP header to indicate the provenance URI, and the other is to use pre-defined properties to express links to provenance URIs in RDF.

The following snippet shows how to indicate provenance information of a specific entity using the HTTP `Link` header field. The `Link` header field can be included in the HTTP response to a `GET` or `HEAD` operation [9]. This approach is very convenient in the Linked Data context where the "following-your-nose" approach is widely appreciated and adopted. In an HTTP response, several `provenance` link header fields could be included, so that a data publisher may indicate provenance information for each separate entity URI.

```
Link: provenance-URI; rel="provenance";
     anchor="entity-URI"
```

Some existing work like Memento [4] and duri [7] have proposed solutions to navigating between a dynamic web resource and different versions of this resource. The PAQ document proposes the use of a property like `ex1:hasAnchor`[2], to link a web resource URI with the entity URIs that represent a particular state of that dynamic web resource. As illustrated in Listing 3, we use `ex1:hasAnchor` to refer the dynamic resource (`<http://example.org/forecast/london>`) to two URIs, each of which represents London forecast taken on a specific day. These entity URIs can then be used to provide provenance information for a particular version of a state-less resource, as previously shown in our example in Listing 2.

All the approaches presented so far are targetted at data owners who will publish provenance along with their data. Provenance information about data can also be published by third-parties. The PAQ document also includes some more complex mechanisms to achieve this, which are not covered here but can be referred to in the PAQ document.

5. CONCLUSIONS AND DISCUSSION

Making interoperable provenance information accessible on the Web is crucial towards achieving a trustworthy web of data/documents. To achieve this we require a language that allows us to interchange provenance information represented using different languages and a mechanism to dis-

[2]Note that the namespace of these properties were not yet defined by the time of writing. This is scheduled to be finalized in according to the PROV-O ontology.

```
@prefix ex1: <http://example.org/t.b.d.> .

<http://example.org/forecast/london>
    ex1:hasAnchor
         <http://example.org/forecast_0602> ,
         <http://example.org/forecast_0702> ;
    ex1:hasProvenance
         <http://example.org/forecast_0602/prvnc> ,
         <http://example.org/forecast_0702/prvnc> .

## Retrieve provenance of each state-ful entity

C: GET /forecast_0602/prvnc HTTP/1.1
C: Host: example.org
C: Accept: text/turtle

S: HTTP/1.1 200 OK
S:
S: <http://example.org/forecast_0602>
S:   prv:createdBy [
S:     rdf:type prv:DataCreation;
S:     prv:completedAt
    "2012-02-06T00:00:00"^^xsd:dateTime ] .

C: GET /forecast_0702/prvnc HTTP/1.1
C: Host: example.org
C: Accept: text/turtle

S: HTTP/1.1 200 OK
S:
S: <http://example.org/forecast_0702>
S:   prv:createdBy [
S:     rdf:type prv:DataCreation;
S:     prv:completedAt
    "2012-02-07T00:00:00"^^xsd:dateTime ] .
```

Listing 3: Linking a state-less resource to state-ful entities and their provenance.

cover and access this metadata unambiguously. The family of standards from the W3C Provenance Working Group are currently geared towards these goals. And our analysis of the interoperability between two widely accepted provenance vocabularies and PROV-O has concluded a promising result.

What is not described here is that PROV-O also provides constructs for expressing some more complicated provenance patterns, such as describing additional attributes of relationships between entities and activities. For example, it can explicitly express recipes used by an activity to `generate` an entity in a reification kind of pattern.

Deciding "what", be state-ful or state-less, is described in provenance information is another longstanding issue to achieve interoperable understanding about this information. Provenance vocabularies largely enforce a strong state-ful mindset; if attributes of an entity changed, it becomes a new, different entity. However, on a open world such as the Web, provenance information is generated and published for applications of varied purposes, from varied perspectives. The representation of a web resource may change over time, for example, the daily forecast of London weather, and it might continually be regarded as the same entity, regardless of its change of "state". If the states of an entity are defined in a very fine-grained manner, e.g. an hourly state for the forecast page, we will have more detailed, or "precise", provenance information. However, too fine-grained distinction between the states of an entity might be impractical and lead to overwhelming provenance data. The trade-off should be considered based on actual context and needs. OPMV only allows more state-ful provenance statements and the Provenance Vocabulary explicitly defines immutable entities, to encourage the publication of more precise provenance. PROV-O provides a relaxed definition of an entity, permitting expression of provenance in both a state-ful and state-less manner, which can hopefully address these subtle differences as a bridging vocabulary.

6. REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets. In *Proceedings of the Linked Data on the Web Workshop (LDOW) at WWW*, 2009.

[2] K. Belhajjame, J. Cheney, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. The PROV Ontology: Model and Formal Semantics. Technical report, 2011. `http://www.w3.org/TR/2011/WD-prov-o-20111213/`, Accessed on February 14, 2012.

[3] K. Belhajjame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, and S. Zednik. PROV Model Primer. Technical report, 2012. `http://www.w3.org/TR/2012/WD-prov-primer-20120110/`, Accessed on February 14, 2012.

[4] H. V. de Sompel, R. Sanderson, M. L. Nelson, L. Balakireva, H. Shankar, and S. Ainsworth. An http-based versioning mechanism for linked data. In *Proceedings of LDOW2010*, 2010.

[5] J. Golbeck. Weaving a web of trust. *Science*, 321(5896):1640–1641, 2008.

[6] O. Hartig and J. Zhao. Publishing and consuming provenance metadata on the web of linked data. In *Proceedings of IPAW 2010*, 2010.

[7] L. Masinter. The 'tdb' and 'duri' URI schemes, based on dated URIs draft-masinter-dated-uri-10. Technical report, 2012. `http://tools.ietf.org/html/draft-masinter-dated-uri-10`, Accessed on February 16, 2012.

[8] L. Moreau, B. Clifford, J. Freire, Y. Gil, P. Groth, J. Futrelle, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, Y. Simmhan, E. Stephan, and J. Van den Bussche. The Open Provenance Model – Core Specification (v1.1), Dec. 2009.

[9] L. Moreau, O. Hartig, Y. Simmhan, J. Myers, T. Lebo, K. Belhajjame, and S. Miles. PROV-AQ: Provenance Access and Query. Technical report, 2012. `http://www.w3.org/TR/2012/WD-prov-aq-20120110/`, Accessed on February 14, 2012.

[10] W3C Provenance Incubator Group. Provenance Vocabulary Mappings. Technical report, 2010. `http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings`, Released on August 06, 2010.

[11] N. Walsh and I. Jacobs. Architecture of the World Wide Web, Volume One. Technical report, 2004. `http://www.w3.org/TR/2004/REC-webarch-20041215/`, W3C Recommendation.

[12] J. Zhao. The Open Provenance Model Vocabulary. Technical report, 2010. `http://purl.org/net/opmv/ns`, Accessed on March 16, 2012.