Modeling Uncertain Provenance and Provenance of Uncertainty in W3C PROV

Tom De Nies Sam Coppens Erik Mannens Rik Van de Walle
{tom.denies,sam.coppens,erik.mannens,rik.vandewalle}@ugent.be

Ghent University - iMinds
Department of Electronics and Information Systems, Multimedia Lab
Gaston Crommenlaan 8 bus 201
B-9050 Ledeberg-Ghent, Belgium

ABSTRACT

This paper describes how to model uncertain provenance and provenance of uncertain things in a flexible and unintrusive manner using PROV, W3C's new standard for provenance. Three new attributes with clearly defined values and semantics are proposed. Modeling this information is an important step towards the modeling and derivation of trust from resources whose provenance is described using PROV.

Categories and Subject Descriptors

H.1.m [**Models and Principles**]: Miscellaneous

Keywords

W3C; provenance; uncertainty; trust; standardization

1. INTRODUCTION

Recently, W3C has published the four main documents of the PROV family of specifications as Proposed Recommendations (PR). These documents include a standard data model, constraints and serializations to facilitate the assertion and interchange of provenance information. To support as many use cases as possible, the W3C Provenance Working Group has kept PROV purposely general, but also easily extensible. Currently, the PROV Data Model [2] does not model uncertainty, which is a good thing, because that would make the model unnecessarily complex for those who do not need to model it. For asserting provenance of provenance (PoP), PROV already has a mechanism in place: bundles[1]. However, in most cases, bundles contain many provenance statements, which makes it very hard, if not impossible, to talk about the provenance of individual statements. Whereas bundles enable coarse-grained PoP, this paper illustrates how to model finer-grained Uncertainty Provenance (UP) using a lightweight approach.

2. UNCERTAINTY ATTRIBUTES

One of the advantages of PROV is its flexibility when it comes to attributes. PROV allows almost all provenance statements to be annotated with optional attributes, with a few exceptions, which are discussed in Sect. 2.3. Therefore, the most straightforward way of modeling UP is to specify one or more optional attributes for the existing constructs that allow them. First, we explain how to model uncertain provenance and how to allow provenance consumers to make trust assessments about the provenance itself. Then, we discuss how to model uncertainty of the content whose provenance is asserted, which is currently not possible in PROV. In total, three attributes are proposed, with predefined values and semantic guidelines.

2.1 Modeling of Uncertain Provenance

According to [1], there are three aspects to uncertainty of provenance:

1. the assigned truth value of the asserter to a statement;
2. the truth value in the eyes of the consumer;
3. the trust relation between asserter and consumer.

When modeling uncertain provenance using PROV, only the first and last aspects apply, since the truth value a provenance consumer assigns to a statement is not meant to be asserted in PROV (and if it is, the consumer becomes an asserter, and we are back to the first aspect). Therefore, we define the following attributes[2], allowing a *degree* and *type* of (un)certainty to be specified for each PROV statement:

1. **up:assertionConfidence** This attribute has a *numerical value between 0 and 1*, and signifies the confidence assigned to a provenance statement by the asserter.

2. **up:assertionType** This attribute describes the type of uncertainty associated with a provenance statement. In our vocabulary, we predefined several values, to interpret as follows. The values *up:HumanAsserted*, *up:MachineGenerated* and *up:MachineCollected* specify whether a provenance statement was generated by a human asserter, or generated or collected by an automated process. *up:Complete* and *up:Incomplete* signify whether all information about this statement is known. For example, this could mean that the statement has missing (optional) arguments, or that a collection has unknown members other than those asserted. *up:Future* signifies that the provenance de-

[1] `http://www.w3.org/TR/prov-dm/#component4`

[2] to improve readability, we use @prefix up: <`http://users.ugent.be/~tdenies/up/`>

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

scribes a process that is yet to be executed, or entities that do not exist yet at the time of assertion. And finally, the values *up:Trusted* and *up:Untrusted* describe whether the provenance comes from a trusted or untrusted source.

2.2 Modeling of Uncertainty

It is important to distinguish the difference between uncertainty of asserted provenance itself, and asserting the uncertainty of information using provenance. This last concept is what we model in this section. Similar to the assertion confidence from the previous section, we define a new attribute:

3. **up:contentConfidence** This attribute specifies a confidence score, denoting how confident a user or application was about the content whose provenance is asserted. It has a *numerical value between 0 and 1.*

This type of provenance is useful in cases where applications or users make fuzzy decisions, and want to assert the provenance of these decisions. Typical examples of such use cases are Named Entity Recognition (NER), Automatic Speech Transcription (ASR), Visual Concept Detection, etc.

2.3 Relations without Optional Attributes

Using the three attributes described above, we provide uncertainty information about almost all provenance concepts defined in PROV-DM. However, there are three relations in PROV that do not support optional attributes: *specialization*, *alternate* and *membership*. Here, the solution lies in specifying an additional entity, with the optional attributes, as a **specialization** of the specializing, alternate or member entity. This is illustrated in Example 2 of Sect. 3.

3. USE CASE EXAMPLES

In this section, we will clarify the use of the attributes defined in Sect. 2, by providing a number of use cases where uncertain provenance is asserted.

Example 1: Provenance Reconstruction.

```
entity(ex:document1)
entity(ex:document2)
entity(ex:document3)
wasDerivedFrom('d1'; ex:document3, ex:document1,
  [up:assertionConfidence="0.6",
   up:assertionType="up:MachineGenerated"])
wasDerivedFrom('d2'; ex:document3, ex:document2,
  [up:assertionConfidence="0.9",
   up:assertionType="up:HumanAsserted"])
```

In this example, `ex:document3` is derived from two different documents. While this is technically possible, the derivation `'d1'` was automatically generated with a relatively low confidence score, whereas `'d2'` was asserted with high confidence by a human. Applications consuming these provenance assertions now have the option to accept or reject the automatically generated assertions, if they decide not to trust them.

Example 2: Named Entity Recognition.

```
entity(ex:document)
entity(ex:namedEntities,
  [prov:type="prov:Collection"])
activity(ex:NER)
wasDerivedFrom(ex:namedEntities,ex:document,ex:NER)
entity(dbpedia:New_York)
entity(dbpedia:Joe_Biden)
entity(ex:New_York,
  [up:contentConfidence="0.6"])
entity(ex:Joe_Biden,
  [up:contentConfidence="0.8"])
specializationOf(ex:New_York, dbpedia:New_York)
specializationOf(ex:Joe_Biden, dbpedia:Joe_Biden)
hadMember(ex:namedEntities, ex:New_York)
hadMember(ex:namedEntities, ex:Joe_Biden)
```

Here, we model the confidence the NER algorithm `ex:NER` had when extracting the Named Entities `dbpedia:New_York` and `dbpedia:Joe_Biden` from `ex:document`. Normally, this information is stored with the content, causing overhead for those users that are not interested in the provenance.

Example 3: Automatic Speech Recognition.

```
entity(ex:word1, [prov:value="this"])
entity(ex:word2, [prov:value="it's"])
activity(ex:ASR)
entity(ex:transcript)
used(ex:ASR, ex:word1,
  [up:contentConfidence="0.8"])
used(ex:ASR, ex:word2,
  [up:contentConfidence="0.2"])
wasGeneratedBy(ex:transcript, ex:ASR,
  [up:contentConfidence="0.8"]))
```

In this last example, we model the process of the detection of a spoken word by an ASR algorithm. The two words "this" and "it's" are very similar, and the algorithm had to choose an option based on the likelihood of it being the correct word. Logically, "this" was chosen because it had the highest confidence, but that does mean that the generation of the transcript only has the same confidence score of 0.8.

4. CONCLUSION

The examples in Sect. 3 show that there are plenty of use cases for fine-grained UP. The attributes we introduced provide a flexible means of asserting this kind of provenance, while preserving the validity of the assertions in conformance with the PROV standards. Adaptation of these lightweight attributes opens an array of possibilities regarding trust assessment of both content and provenance information.

5. ACKNOWLEDGMENTS

The research activities in this paper were funded by Ghent University, iMinds (a research institute founded by the Flemish Government), the Institute for Promotion of Innovation by Science and Technology in Flanders (IWT), the FWO-Flanders, and the European Union. We would also like to thank the members of the W3C Provenance WG for their hard work on the PROV family of specifications.

6. REFERENCES

[1] J. Huang and M. Fox. Uncertainty in knowledge provenance. *The Semantic Web: Research and Applications*, pages 372–387, 2004.

[2] W3C Provenance Working Group and others. PROV-DM: The PROV Data Model (L. Moreau & P. Missier, Eds.). W3C, 2012.