Provenance of Publications: A PROV style for Latex[1]

Luc Moreau
University of Southampton
l.moreau@ecs.soton.ac.uk

Paul Groth
Elsevier Labs
p.groth@elsevier.com

Abstract

In general, the task of generating provenance is still tedious, and the community still lacks tools to generate provenance easily. In particular, when writing papers, researchers should be able to produce the provenance of their papers, make it available online, and embed provenance metadata directly in their PDF files. To address this goal, we introduce `prov.sty`, a PROV style for LaTeX, allowing LaTeX source to be marked up, and associated provenance to be generated automatically. Provenance captured by this style currently includes: authors, organisations, funders, bibliographic citations, and embedded images. PROV provenance is automatically generated and exported as a Turtle file; further, a link to a provenance resource can be embedded in PDF using the XMP metadata format.

Keywords provenance, PROV, latex style, pdf embed, xmp format, tool

1. Introduction

Provenance, defined as a record that describes how entities, activities, and agents have influenced a piece of data [5], can help users make trust judgements about data. PROV is a set of W3C specifications aiming to facilitate the representation and exchange of provenance on the Web. PROV is domain-agnostic and is been applied to a wide range of applications, including climate assessment[2], legal notices[3].

While the provenance community has made substantial progress in terms of understanding and standardising provenance, it is an unfortunate reality that, due to the lack of easy tools, provenance still remains beyond the reach of the general public. For this paper's authors, who are willing to work with leading-edge, non-mature technology, it is a great frustation that provenance of their papers cannot be generated automatically, and that best practice cannot be demonstrated to the research community. No more!

Publications already contain a lot of provenance information in textual form, but it is not exposed in PROV format: authorship, institutions, sponsoring projects, included graphics, and bibliography. The purpose is this work is to demonstrate how this textual information can be marked up, so that PROV can be generated automatically and provenance metadata embedded in the PDF. The approach is implemented by a style `prov.sty` for LaTeX. The interest of the approach is that provenance is generated systematically, as part of the typesetting process, for each version of the document produced.

The purpose of this short paper is to outline the approach, to describe the LaTeX annotation macros and the PROV-O-compliant [4] provenance they generate, to explain the actual provenance generated for this document, and to explain how provenance metadata can be embedded in the PDF file, by exploiting the XMP metadata standard [1]. At the same time, we are releasing `prov.sty` on GitHub at https://github.com/prov-suite/prov-sty.

2. Author Guide

The style was designed with a view to minimize author's work. The adopted approach is to import the package, annotate the LaTeX source without having to duplicate information, generate the provenance on the fly every time LaTeX is run, and embed provenance information in the PDF using XMP automatically.

In practice, the provenance file in Turtle format [4] has to be deployed and made web accessible. (This process is not handled by `prov.sty`.) Footnote 1 embeds in the text the location of the provenance, for user consumption.

2.1 Preamble

The preamble must import the package `prov.sty`.

```
\usepackage{prov}
```

Whenever LaTeX is run, (see Section 2.3), a unique identifier is created for the current document, and provenance is generated in a separate file. This is being referred to as "author" mode.

Alternatively, using the optional "publisher" mode, provenance is no longer computed, and annotations are simply ignored.

```
\usepackage[publisher]{prov}
```

2.2 Document Annotations

The `prov.sty` package offers a series of macros that the author can use to annotate LaTeX documents, with a view of generating its PROV provenance. This section introduces the `prov.sty` macros. When `prov.sty` is used in publisher mode, these annotations have no effect. In the rest of the section, we describe the macros intuitively, and we illustrate the Turtle statements they generate.

The following key LaTeX macros are discussed in the respective sections.

- `\provAuthor` (see Section 2.2.2)
- `\provBibliography` (see Section 2.2.7)

[1] This document's provenance is embedded in the pdf and can also be found at http://eprints.soton.ac.uk/378019/3/provenance.ttl using <http://openprovenance.org/documents#b2958af5-c5e3-4cb0-b107-bddd65093e96> as `prov:has_anchor`.

[2] http://nca2014.globalchange.gov/report

[3] https://www.thegazette.co.uk/

TAPP '15, Month 8–9, 2015, Edinburgh, Scotland, UK.

- \provCitation (see Section 2.2.7)
- \provInclude (see Section 2.2.6)
- \provOrganization (see Section 2.2.3)
- \provProject (see Section 2.2.5)
- \provResource (see Section 2.2.6)
- \provThis (see Section 2.2.1)
- \provTitle (see Section 2.2.4)
- \thisresource (see Section 2.2.1)

2.2.1 \provThis and \thisresource

The macro \provThis generates an RDF description of the current document as a prov:Entity. The macro makes use of the internal macro \thisresource to obtain a URI Reference for this document. The macro \thisresource expands into a string of characters. This macro would typically not be used by authors, unless they program other prov.sty related macros, or they want to include the document's identifier in the text, as we did in Footnote 1.

At the beginning of the document, we expect the following annotations to be inserted.

```
\provThis
```

In response, the following Turtle statement is generated. For every run of LaTeX, a new identifier is generated, since the resulting PDF is a new prov:Entity. To this end, we use a UUID generator.

```
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96   a prov:Entity .
```

2.2.2 \provAuthor

The macro \provAuthor allows the author of the current document to be declared. The macro takes two arguments: the first is the author's name, which is associated with the author resource by the property foaf:name, whereas the second is a URI for the author. In publisher's mode, this macro expands to the author string, ignoring the second argument. Typically, this annotation occurs inside the \author declaration for the paper.

```
\provAuthor{Luc Moreau}%
           {http://orcid.org/0000-0002-3494-120X}
```

In response, the following Turtle statements are generated:

```
<http://orcid.org/0000-0002-3494-120X>
    a prov:Agent, prov:Person;
    foaf:name "Luc Moreau" .
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
    prov:wasAttributedTo
    <http://orcid.org/0000-0002-3494-120X> .
```

2.2.3 \provOrganization

The macro \provOrganization allows an author's organization to be declared. The macro takes two arguments, the first is the organization's name, which is linked with property foaf:name, whereas the second is the organization's URI. In publisher's mode, this macro expands to the organization string, ignoring the second argument.

```
\provOrganization{University of Southampton}%
                 {http://www.soton.ac.uk/}
```

The following Turtle statements are generated:

```
<http://www.soton.ac.uk/>
   a prov:Agent, prov:Organization;
   foaf:name "University of Southampton" .
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
   prov:wasAttributedTo <http://www.soton.ac.uk/> .
```

2.2.4 \provTitle

The macro \provTitle allows a document title to be declared.

The macro takes a single argument: the title itself, which is linked to this resource with the property schema:headline. The macro expands to the title string. Typically, this annotation occurs inside the \title declaration for the paper.

```
\provTitle{A PROV style for latex}
```

In response, the following triple is generated:

```
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
  schema:headline "A PROV style for latex" .
```

2.2.5 \provProject

The macro \provProject allows a sponsoring project to be declared.

The macro takes three arguments, the first is the project name, the second is a URI for this resource, and the third is the funding agency. In publisher's mode, this macro expands to the project string, ignoring the remaining two arguments.

```
\provProject{SOCIAM}%
            {http://www.sociam.org/}%
            {http://www.epsrc.ac.uk/}
```

In response, the following Turtle statements are generated:

```
<http://www.sociam.org/> a prov:Agent;
   foaf:name "SOCIAM" ;
   prov:actedOnBehalfOf <http://www.epsrc.ac.uk/> .
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
     prov:wasAttributedTo <http://www.sociam.org/> .
```

2.2.6 \includegraphics, \provInclude and \provResource

The macro \includegraphics (from package graphicx) can be used to include graphics in the current document. Package prov.sty redefines the macro \includegraphics, so as to call \provInclude, a macro in charge of recording the provenance of this inclusion: the current document is said to be derived from the included resource.

The included resource is a file on the file system, so a third party would typically not be able to access it directly. For this reason, the macro \provResource allows for an online resource, copy of the included file, to be declared. Thus, the provenance of this inclusion is modelled as follows: the current document was derived from the included resource, itself an alternate of the online resource. For a third party to be able to check that the online resource is a copy of the included one, prov.sty computes the md5 hash of the included file.

```
\provResource{http://example.org/myfig.pdf}
\includegraphics{myfig.pdf}
```

In response, the following Turtle statements are generated. A new resource is introduced to represent the included file. Its md5 is associated with property crypto:md5. Its local path on the filesystem is asserted using property schema:contentLocation.

```
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
  prov:wasDerivedFrom
  ex:b2958af5-c5e3-4cb0-b107-bddd65093e96 -1  .

ex:b2958af5-c5e3-4cb0-b107-bddd65093e96 -1
       a prov:Entity ;
       schema:contentLocation <myfig.pdf> ;
       prov:alternateOf <http://example.org/myfig.pdf> ;
       crypto:md5 "6ad7419b3eec7a7ad52931eeb579dba3" .
```

```
<http://example.org/myfig.pdf>
       a prov:Entity .
```

2.2.7 \provBibliography and \provCitation

The macro \provBibliography allows for provenance to be generated for the bibliography. No further annotation is required, but prov.sty requires bibliographic entries to contain URIs or DOIs. The corresponding macros \uri and \doi are overriden, to call \provCitation.

The following declaration is expected to be placed just before the LaTeX \bibliography macro.

```
\provBibliography
```

For instance, this document cites PROV-DM [5], which leads to the following Turtle statement, describing the dependency of this document on the PROV-DM resource.

```
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
   prov:wasDerivedFrom
  <http://www.w3.org/TR/2013/REC-prov-dm-20130430/> .
```

2.2.8 \provSpecialization

The macro \provSpecialization allows for a more general resource to be identified, representing all the variants of the current document. Indeed, a given document may have multiple variants. Not only we have various versions, but also there may be a pre-print version in an institutional repository, an editor-compiled version for the proceedings, and the final version published by the publisher. With \provSpecialization generic version of the document can be hard-coded in the paper.

```
\provSpecialization{C4384149-0B34-4360-B2DA-A1AFFBB90188}
```

In response, the following Turtle statement is generated, making use of the prov:specializationOf property.

```
ex:b2958af5-c5e3-4cb0-b107-bddd65093e96
   prov:specializationOf
   ex:C4384149-0B34-4360-B2DA-A1AFFBB90188   .
```

2.2.9 \provEmbed and \provLocation

The macro \provEmbed allows for metadata about the provenance to be inserted in the PDF document, using the XMP metadata format [1]. This command is expected to be called as the last macro before the end of the document. XMP supports a subset of RDF/XML that does not appear to be expressive enough to embed PROV provenance directly. Instead, using the approach recommended by PROV-AQ [3], a pointer to the provenance is expressed, using the XMP format.

```
\provLocation{http://example.org/provenance.ttl}
\provEmbed
```

In response, the following Turtle statements are generated, and embedded as XMP metadata. The current resource has some provenance prov:has_provenance that can be found at the location provided by macro \provLocation; this resource is known in that provenance file as the resource object of prov:has_anchor.

```
<>
 prov:has_anchor ex:b2958af5-c5e3-4cb0-b107-bddd65093e96  ;
 prov:alternateOf ex:b2958af5-c5e3-4cb0-b107-bddd65093e96  ;
 prov:has_provenance <http://example.org/provenance.ttl>.
```

As noted before, it is the author's responsibility to make the provenance available online at the declared URI.

2.3 Invocation

To run LaTeX, one needs the option --shell-escape to allow for a UUID generator and the ProvToolbox's provconvert to be called during typesetting. (Note that this is not required when prov.sty is used in publisher mode.)

```
pdflatex  --shell-escape prov-sty-tapp15.tex
```

3. Provenance Modelling

Section 2.2 lists the prov.sty annotations and includes snippets of RDF generated by these. For this document, the full provenance is displayed in Figure 1. (Concretely, this image was generated by converting the provenance of a previous version of the document into PDF.)

The overall provenance graph is rooted at this resource, appearing at the bottom of the figure. Annotations in the figure indicate which prov.sty macro was used to generate which portion of the graph.

We have refrained from designing our own ontology for expressing this provenance. Instead, we relied on existing vocabularies, such as schema.org, foaf, and a cryptography ontology crypto.

While most of the modelling in PROV is straightforward, the bibliography raises an interesting issue. Currently, a citation is modeled by a PROV derivation, to express that the current document was derived from the cited document: derivation is to be understood as building on, improving over, or addressing a problem differently than previous work. In the day-to-day practice of the scientific community, it is possible for two documents to cite each other. This would result in a cycle of derivation, which is regarded as invalid provenance.

4. Related Work

PDF allows general metadata in the form of key-value pairs to be embedded (see XMP metadata format [1]). LaTeX offers some style to embed metadata in PDF documents, including xmpincl.sty used by prov.sty and hyperref.sty. A variety of tools allow for direct PDF manipulation including Adobe's Acrobat and the command line pdfinfo.

Sumatra[4] is a tool for managing numerical processing projects. It relies of a database indexing all generated artifacts, e.g. figures, data, etc. It also offers a LaTeXfile allowing their inclusion in documents as well as their provenance information.

Beyond LaTeXand PDF, Vistrails [2] offers a mechanism to track provenance of the figures of a paper.

5. Discussion

With this paper, we have showed that it is possible to lower PROV's barrier of adoption, by adapting tools to generate provenance automatically. For those tools to be useful, they need to generate provenance systematically, for every created artifact. Over time, as similar tools get developed, their provenance should be linked up. For instance, the git2prov converter is capable of exporting PROV from GIT. It should be possible for users to seamlessly navigate the provenance generated by both tools.

While the prov.sty style is still a proof of concept, we feel that it is time to release it, and have others to use it. Improving usability, enhancing the quality of provenance, and strengthening of LaTeX integration are all desirable. prov.sty is available at https://github.com/prov-suite/prov-styunder the MIT Open Source license.

While it is great for metadata about the provenance to be embedded in the PDF using XMP, it would have been nice to embed the

[4] Sumatra: https://pythonhosted.org/Sumatra/publishing.html

provenance itself. However, despite supporting RDF/XML, XMP imposes limitations on the metadata content, and does not allow arbitrary PROV graphs to be embedded.

Acknowledgments

This work is funded in part by the EPSRC SOCIAM (EP/J017728/1) and ORCHID (EP/I011587/1) projects, the FP7 SmartSociety (600854) project, and the ESRC eBook (ES/K007246/1) project.

References

[1] Xmp specification part 1 (data model, serialization, and core properties), July 2010. URL http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf.

[2] *The Architecture Of Open Source Applications*. June 2011. ISBN 1257638017. URL http://www.aosabook.org/en/.

[3] G. Klyne, P. Groth (eds.), L. Moreau, O. Hartig, Y. Simmhan, J. Myers, T. Lebo, K. Belhajjame, and S. Miles. PROV-AQ: Provenance Access and Query. W3C Working Group Note NOTE-prov-aq-20130430, World Wide Web Consortium, Apr. 2013. URL http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/.

[4] T. Lebo, S. Sahoo, D. McGuinness (eds.), K. Behajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. PROV-O: The PROV Ontology. W3C Recommendation REC-prov-o-20130430, World Wide Web Consortium, Oct. 2013. URL http://www.w3.org/TR/2013/REC-prov-o-20130430/.

[5] L. Moreau and P. Missier (eds.). PROV-DM: The PROV Data Model. W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium, Oct. 2013. URL http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

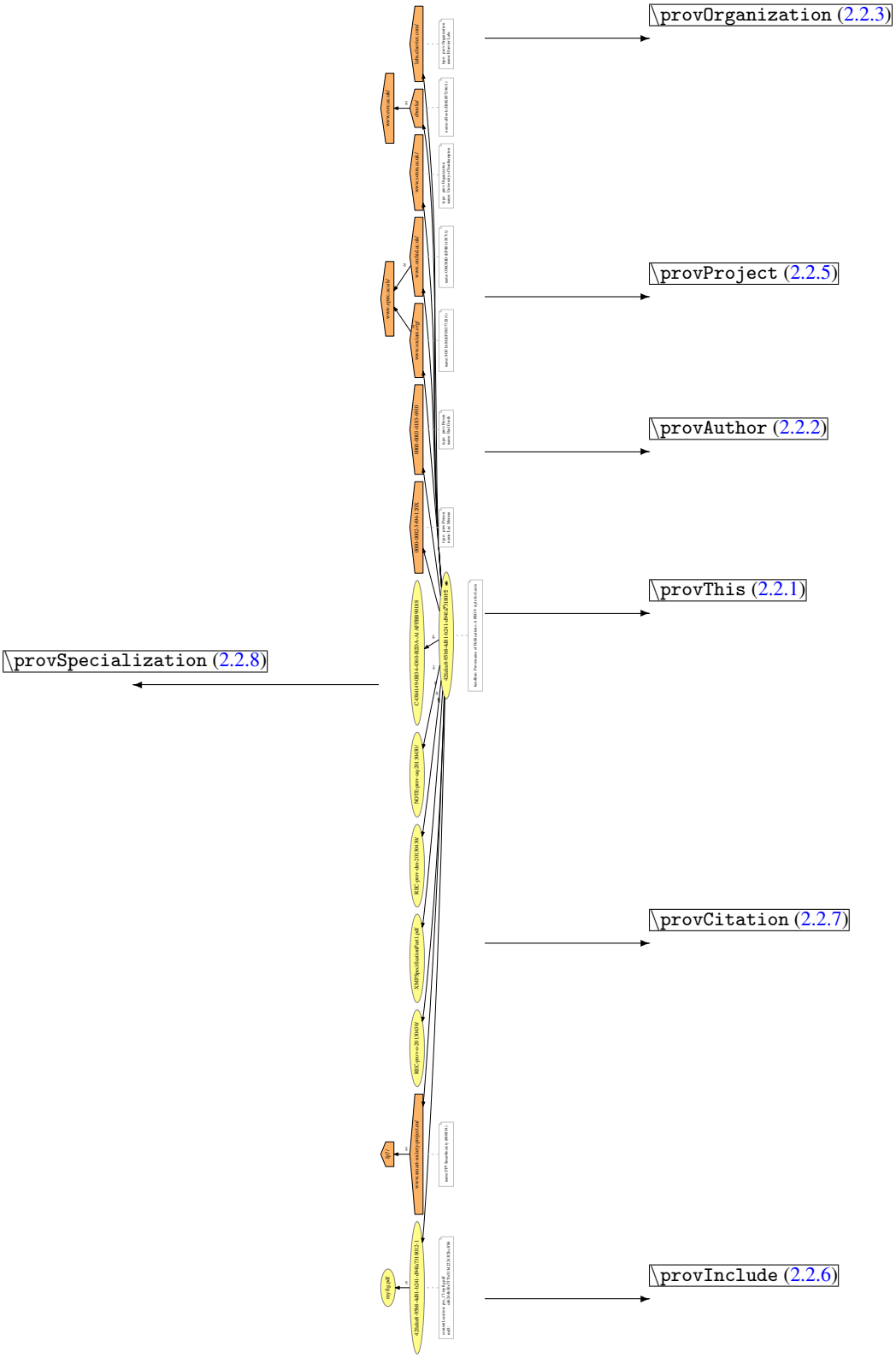

Figure 1. A Graphical Illustration of the Provenance of the Current Document. The annotations indicate which `prov.sty` macro underpinned the generation of which part of the graph. Vector graphics make this figure zoomable. Online version of the figure is available from https://eprints.soton.ac.uk/378019/1/myfig.pdf.