PICASO: Provenance Interlinking and Collective Authoring for Scientific Objects

Trung Dong Huynh, Danius T. Michaelides, and Luc Moreau

Electronics and Computer Science, University of Southampton
Southampton, SO17 1BJ, United Kingdom
{tdh,dtm,l.moreau}@ecs.soton.ac.uk

Abstract. PICASO is an online platform that crowdsources the links between related scientific objects identified by unified resource identifiers (URI). Its aim is to collect the provenance of scientific work and to publish it as open data to allow for further analyses and research over this kind of information.

Keywords: data provenance, crowdsourcing, templating, linked data

1 Introduction

PICASO[1] is an online platform that crowdsources the links between related scientific objects identified by unified resource identifiers (URI). Its aim is to collect the provenance of scientific work and to publish it as open data to allow for further analyses and research over this kind of information. Unlike existing open bibliographic databases like DBLP [3] or CiteSeer [1], the data gathered by PICASO are not restricted to bibliographic information and citations, but also includes the links between a piece of work and any other relevant entities and events such as the dataset(s) it used, the poster or slides presenting it, the project that funded its authors, and even the presentation activity of the work in a conference session. PICASO encourages linking to objects residing in their own silos, such as linking a presentation on SlideShare[2] to the digital object identifier (DOI) of the original paper. By so doing, PICASO provides the tool for researchers to publicly document the origins and derivatives of their work, or its *provenance*. In order to facilitate the consumption of such information, we model the data following the PROV data model [5], a recommendation of the World Wide Web Consortium (W3C) for exchanging provenance information over the Web. We used the Seme4 Platform to publish the data collected by PICASO (i.e. the PICASO dataset) as Linked Open Data, available at http://data.provenance.ecs.soton.ac.uk.

[1] PICASO stands for Provenance Interlinking and Collective Authoring for Scientific Objects. The PICASO web application is live at https://provenance.ecs.soton.ac.uk/picaso.

[2] http://www.slideshare.net/

In addition to the extensive scope of the information it contains, the PICASO dataset is different to most bibliographic and citations databases in that its data was crowdsourced from the public. Given the inherently varied knowledge, and even different opinions, of the public, it is crucial that PICASO keeps track of the authorships of its data for quality management purposes and supports different views on a resource or an event. For the former, we provide the PICASO dataset along with the complete provenance of all its assertions — another unique feature of this dataset. For the latter, we propose an approach for explicitly modelling different opinions.

2 A templating approach to provenance generation

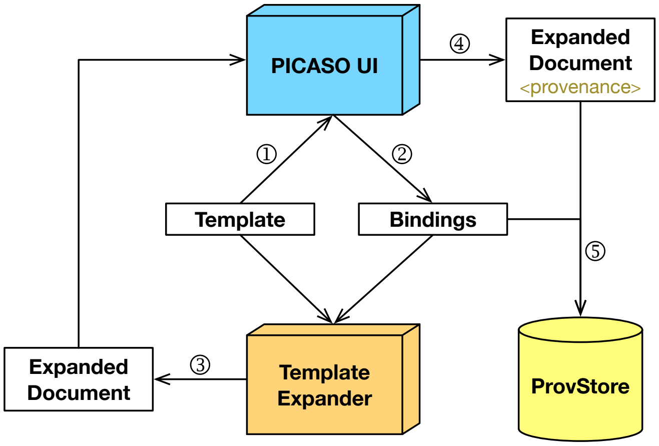


Fig. 1. The data generation process in PICASO involving its three main components: PICASO UI, Template Expander, and ProvStore.

PICASO was designed to provide the academic community with an easy way of asserting links between scientific objects, identified by their URIs, without requiring them to understand the underlying data model. In order to do so, PICASO was equipped with a number of *templates* describing common relations between scientific objects like papers, datasets, slides, etc. From those built-in templates, links between such entities can be asserted by PICASO users in the 5-step process as depicted in Figure 1:

1. **Template selection**: The user first selects one of the provided templates. A template is a PROV document that contains “variables” in any position where URIs are allowed in PROV (see [4] for a complete description of the PROV template system used by PICASO). For instance, the Presentation

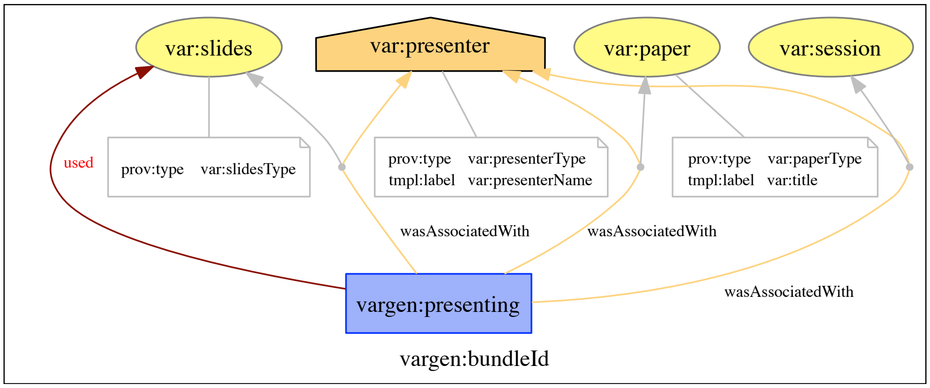


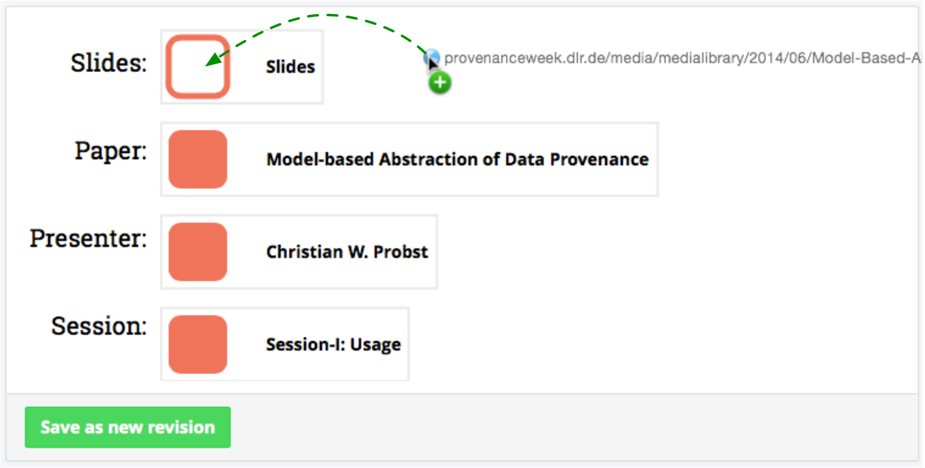
Fig. 2. The Presentation template in PROV graphical representation [?]. `var:x` is a variable to be replaced by a URI provided by the users while that for `vargen:x` will be automatically generated.

template (Figure 2), has the variable `var:slides` for the slides presented by a presenter, denoted by `var:presenter`, and so on.

2. **Bindings creation**: PICASO User Interface (UI) represents the variables of a templates as empty slots when the template is first shown to a user. By dragging a resource[3] into a slot, as shown in Figure 3, the user actually assigns the resource's URI to the corresponding variable. When the user saves their assertions, PICASO UI generates a *binding* document to represent the mappings between variable names and the resources' URIs assigned to them.
3. **Template expansion**: The Template Expander uses the binding document produced in (2) along with the template document in (1) to instantiate the template into a valid PROV document, called an *expanded* document (see [4] for the expansion algorithm), which contains the assertions that link the resources selected by the user according to the selected template.
4. **Provenance annotation**: The PICASO application then annotates the expanded document with its provenance information, detailing how the document was generated.
5. **Storage**: The binding and expanded documents are posted to ProvStore [2], an online repository for PROV documents that provides extra query, visualising, and conversion capabilities. PICASO, however, keeps an index of those documents in order to retrieve them when required.

As an example, Figure 3 shows the screenshot of a template instantiation. In this case, the document describes a presentation that was delivered by a presenter in a conference session, presenting a particular paper, including the URI for the slides used therein. As it happened, the expanded document #1003

[3] A resource's URI can be dragged into a variable slot from any browser window. In addition, users can search for those known to PICASO or those in DBLP database [3] from the same window.



Slides: Slides

provenanceweek.dlr.de/media/medialibrary/2014/06/Model-Based-A

Paper: Model-based Abstraction of Data Provenance

Presenter: Christian W. Probst

Session: Session-I: Usage

Fig. 3. A screenshot of PICASO showing the expanded document #1003 (an instantiation of the Presentation template shown in Figure 2) with an external URI being dragged into the "Slides" slot. The green dotted line was added to show the dragging action.

in fact contains an out-of-date URI for the slides and the user was dragging the correct URI into the document to update it. After having been edited, the updated assertions will be saved in the form of a new binding document, resulting in a new expanded document in the process. In this way, users can enrich existing PICASO data at any time by revising them with new information and the history of such revisions is fully recorded.

Acknowledgements.

We gratefully acknowledge funding from the UK Research Council for project Orchid (grant EP/I011587/1) and the UK Economic and Social Research Council (grant ES/K007246/1).

References

1. Giles, C.L., Bollacker, K.D., Lawrence, S.: CiteSeer: An automatic citation indexing system. In: Proceedings of the third ACM conference on Digital libraries - DL '98. pp. 89–98. ACM Press, New York, New York, USA (1998)
2. Huynh, T.D., Moreau, L.: ProvStore: A public provenance repository. In: Ludäscher, B., Plale, B. (eds.) 5th International Provenance and Annotation Workshop, IPAW 2014, Lecture Notes in Computer Science, vol. 8628, pp. 275–277. Springer International Publishing, Cologne, Germany (2015)
3. Ley, M.: DBLP — some lessons learned. Proceedings of the VLDB Endowment 2(2), 1493–1500 (Aug 2009)

4. Michaelides, D., Huynh, T.D., Moreau, L.: PROV-TEMPLATE: A template system for PROV documents (2014), https://provenance.ecs.soton.ac.uk/prov-template/, [Online; Draft 07-June-2014]
5. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model. W3C Recommendation, World Wide Web Consortium (2013), http://www.w3.org/TR/2013/REC-prov-dm-20130430/