PROV-Template: A Templating System to Generate Provenance
— Descriptions of the datasets —

Luc Moreau, Belfrit Victor Batlajery, Trung Dong Huynh,
Danius Michaelides, Heather Packer

November 30, 2016

Abstract

PROV-TEMPLATE is a declarative approach that enables designers and programmers to design and generate provenance compatible with the PROV standard of the World Wide Web Consortium. Designers specify the topology of the provenance to be generated by composing templates, which are provenance graphs containing variables, acting as placeholders for values. Programmers write programs that log values and package them up in sets of bindings, a data structure associating variables and values. An expansion algorithm generates instantiated provenance from templates and sets of bindings in any of the serialisation formats supported by PROV. A quantitative evaluation shows that sets of bindings have a size that is typically 40% of that of expanded provenance templates and that the expansion algorithm is suitably tractable, operating in fractions of milliseconds for the type of templates surveyed in the article. Furthermore, the approach shows four significant software engineering benefits: distributed development, provenance maintenance, potential runtime checks and static analysis, and provenance consumption. The article gathers quantitative data and qualitative benefits descriptions from four different applications making use of PROV-TEMPLATE. The system is implemented and released in the open-source library ProvToolbox for provenance processing.

This document overviews the application involved in the empirical evaluation of PROV-TEMPLATE, outlines their use of provenance, and describes the structure of the data sets involved in the evaluation.

1 Overview

Each application's directory is structured as follows, where ⟨*application*⟩ can be foodprovenance, ebook, smartshare, or picaso.

/R/	R analysis
/⟨*application*⟩/⟨*snapshot*⟩	application data
/⟨*application*⟩/⟨*snapshot*⟩/raw/	raw files
/⟨*application*⟩/⟨*snapshot*⟩/raw/templates	raw templates
/⟨*application*⟩/⟨*snapshot*⟩/raw/bindings	raw bindings (original format 2014)
/⟨*application*⟩/⟨*snapshot*⟩/raw/bindings2	raw bindings (new format 2016)
/⟨*application*⟩/⟨*snapshot*⟩/raw/expansions	raw expansions
/⟨*application*⟩/⟨*snapshot*⟩/normalized/	normalized files
/⟨*application*⟩/⟨*snapshot*⟩/normalized/templates	normalized templates
/⟨*application*⟩/⟨*snapshot*⟩/normalized/bindings	normalized bindings
/⟨*application*⟩/⟨*snapshot*⟩/normalized/expansions	normalized expansions
/⟨*application*⟩/⟨*snapshot*⟩/outputs/	analysis output files

Within each directory, the toplevel `Makefile` contains the following useful targets:

- `do.normalize`: normalize files, with following subtargets:
  - `do.normalize.templates`: normalize templates
  - `do.normalize.bindings`: normalize bindings (old format)
  - `do.normalize.bindings2`: normalize bindings in new format
  - `do.normalize.expansions`: normalize expansions
- `do.stats`: run statistics
- `do.R`: produce scatter plot
- `do.time`: produce time box plot

The plots included in the paper have been generated by the R files box.R and time.R. For convenience, R books have been provided in /R/box.{html,pdf} and /R/time.{html,pdf}

2 Smartshare

Smartshare is a ride-sharing application that allows drivers and commuters to offer and request rides. Ride offers and requests include details about required travels, timing, locations, capacity, prices, and other details relevant to car sharing. The application automatically matches commuters to available cars. The application has three components, a user interface (UI), an orchestrator which matches drivers and riders, and a reputation manager which stores ratings and generates rating aggregations. Smartshare is provenance-enabled, capturing the provenance of any user decision, matching or rating managed by the system.

In Smartshare, provenance is used to make the application accountable: specifically, some components of the application, such as the rating service, are using the provenance to provide explanations to users, as to what rating was assigned to participants.

Table 1: Smartshare Templates.

ID	Template Name	Description
1	Login	UI - Records the login of users
2	Change_Page	UI - Records the change of pages in the app
3	Use_Response	UI - Records the use of responses from other components
4	Send_Request	UI- Records which requests are sent from the UI
5	Receive_Request	Orchestrator - Records which requests are recieved
6	Composition	Orchestrator - Records possible compositions of drivers and riders for potential rides
7	Negotiation_Type_1	Orchestrator - Records the first negotiation offer
8	Negotiation_Type_2	Orchestrator - Records a negative negotiation offer
9	Negotiation_Type_3	Orchestrator - Records a positive negotiation offer
10	Gen_Reputation	Reputation Manager - Records the generation of aggregated reviews and ratings
11	Receive_API_Call	Reputation Manager - Records when an API call is made
12	Receive_Feedback	Reputation Manager - Records when a feedback is submitted to the manager

The templates presented in Table 1 are the final version of templates that were refined over several iterations, because of the application's iterative design and distributed development.

The smartshare dataset contains 12 templates, 1608 bindings, 1608 expanded templates, in both raw and normalized representations.

3 EBook

The EBook data set is generated by the Stat-JR workflow tool called LEAF. Stat-JR (`http://www.bristol.ac.uk/cmm/software/statjr/`) is a software package that acts as a universal gateway to specialised statistic packages and enables users to learn about statistical methods. It consists of a number of interfaces to enable users to run statistical computations. The LEAF component allows complex statistical analyses to be constructed into a workflow using a visual programming language called Blockly[1]. Blocks representing both programming constructs and statistical operations are assembled into an analysis, which LEAF can then execute. The computation model is that a block execution consists of a number of named inputs generating a number of named outputs. A block has a parent and it will be a parent to any blocks it triggers the execution of. LEAF creates a provenance record for an execution of an analysis by creating a binding for each block execution. A single template is used by all the bindings; its structure closely matches the computation model.

Table 2: EBook Template.

ID	Template Name	Description
1	template_block_run	Describes the execution of a block

Provenance is used in EBook to provide explanations of how data products are derived and to reconstruct workflows, which are editable and executable.

The EBook dataset contains 1 template, 235 bindings, 235 expanded templates, in both raw and normalized representations.

[1] `https://developers.google.com/blockly/`

4 PICASO

PICASO (Provenance Interlinking and Collective Authoring for Scientific Objects) is a live web application at `https://provenance.ecs.soton.ac.uk/picaso`. It is an online platform that crowdsources the links between related scientific objects identified by uniform 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 (`http://dblp.uni-trier.de/`) or CiteSeer (`http://citeseerx.ist.psu.edu/`), the data gathered by PICASO are not restricted to bibliographic information and citations, but also include 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 (`http://www.slideshare.net/`) 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.

Table 3: PICASO Templates.

ID	Template Name	Description
1	Attribution	Attributing publications to authors, editors, etc.
2	Citation	Links to cited work
3	Conference-Session	Sessions in a conference's program
4	Dataset-Usage	Dataset used in the preparation of a paper
5	Derived-Material	Posters/slides derived from a published work
6	Work-Element	Figures and tables in a paper
7	Tweet	Tweets mentioning a scientific work
8	Presentation	Presenter presenting slides
9	Presentation v2	The paper presented is added to the Presentation template
10	Derivation (1 $\leftarrow$ n)	Multiple entities derived from one
11	Derivation (n $\leftarrow$ 1)	One entity derived from many
12	Attribution v2	Adding the publication year attribute to the Attribution template and allowing for publications of type Tweet, Book, Slide, Dataset to be described by this same template
13	Project	People involved in a research project

There are 11 original templates used by PICASO. However, two of those were later revised, giving a total of 13 templates provided in the dataset; they are listed in Table 3. In addition, there are 4,019 bindings and 4,415 expansions in the PICASO dataset.

5 Food

The Food application captures the details of food products and tracks their orders and deliveries in the Hampshire county, in England. It aims to develop "due diligence" methods in collaboration with local scientific authorities, through the use of provenance and risk models. Provenance is used to describe characteristics of a food product such as its origin, ingredients, cooking instruction, microorganisms details, etc. This information is recorded in a standardized format that allows any provenance-aware application to make use of it. Moreover, the provenance of food products also helps the application to build a representation of the food supply chain, so that risk models can be overlaid on top of it, in order to identify potential areas of contamination.

There are 3 templates used to generate food provenance, namely the "foodspec" template to capture the specification of food product, the "invoiceitems" template to capture the orders and deliveries of food, andthe "analysis" template to capture the result of microbial sampling at schools (Table 4).

Table 4: Food Templates.

ID	Template Name	Description
1	foodspec	The specification and description of food (e.g, ingredients, origin, allergen, cooking instruction, etc)
2	invoiceitems	Orders and deliveries of food product
3	analysis	Result of microbial sampling at schools

In total, 1031 provenance expansions are generated from 1031 bindings in this application. However, the bindings and expansions are commercial sensitive and are, therefore, not made available in the public dataset.