SHARP: Harmonizing Galaxy and Taverna workflow provenance

Alban Gaignard, Khalid Belhajjame, Hala Skaf-Molli

SHARP: Harmonizing Galaxy and Taverna workflow provenance

Alban Gaignard[1], Khalid Belhajjame[2], and Hala Skaf-Molli[3]

[1] Nantes Academic Hospital, France
alban.gaignard@univ-nantes.fr
[2] LAMSADE – Paris-Dauphine University, France
kbelhajj@googlemail.com
[3] LS2N – Nantes University, France
hala.skaf@univ-nantes.fr

Abstract. SHARP is a Linked Data approach for harmonizing cross-workflow provenance. In this demo, we demonstrate SHARP through a real-world omic experiment involving workflow traces generated by Taverna and Galaxy systems. SHARP starts by interlinking provenance traces generated by Galaxy and Taverna workflows and then harmonize the interlinked graphs thanks to OWL and PROV inference rules. The resulting provenance graph can be exploited for answering queries across Galaxy and Taverna workflow runs.

Keywords: Reproducibility, Scientific Workflows, Provenance, Prov Constraints

1 Introduction

Imagine a system that allows scientists to answer queries like *Which parameters were used by my colleagues in their workflow that would explain my workflow results !*

Answering such queries requires the exploitation of provenances traces generated by different workflow systems. PROV has been adopted by a number of workflow systems for encoding the traces of workflow executions. However, workflow systems extend PROV differently which yields heterogeneity in the generated provenance traces. This heterogeneity diminishes the value of such traces, *e.g.* when combining and querying provenance traces of different workflow systems.

In this demo, we present SHARP; an approach for interlinking and harmonizing provenance traces of different workflow systems using PROV inferences. We demonstrate SHARP through a real-world omic experiment involving workflow traces generated by the Galaxy and Taverna systems.

2 SHARP: Harmonizing multi-PROV Graphs

SHARP exploits the fact that provenance vocabularies used by workflow systems extend the W3C PROV-O ontology. It uses reasoning for harmonizing provenance traces thanks to **OWL entailment regime** and **PROV constraints** [3]. SHARP identifies three categories of rules of **Prov constraints** with respect to expressiveness (i) rules that contain only universal variables, (ii) rules that contain existential variables, and (iii) rules making use of n-array relations (with $n \geqslant 3$). SHARP formalizes constraints as tuple-generating dependencies TGDs and equality-generating-dependencies EGD[1], and then implements them as Jena rules.

The following example illustrates usage and generation from derivation rule. The following rule states that if an entity $\mathtt{e}_2$ was derived from an entity $\mathtt{e}_1$, then there exists an activity `a`, such that `a` used $\mathtt{e}_1$ and generated $\mathtt{e}_2$.

$\texttt{wasDerivedFrom}(\texttt{e}_2, \texttt{e}_1) \rightarrow \exists\, \texttt{a}\ \texttt{used}(\texttt{a}, \texttt{e}_1),\ \texttt{wasGeneratedFrom}(\texttt{e}_2, \texttt{a})$.

The overall provenance harmonization process consists in i) interlinking data artifacts through *owl:sameAs* property, ii) applying OWL inferences, iii) applying PROV constraints TGDs and EGDs as a set of inference rules. This process terminates because PROV constraints are known to be weakly acyclic [4].

3 SHARP Demo Scenario

We use a real scientific experiment conducted through two bioinformatics workflows.

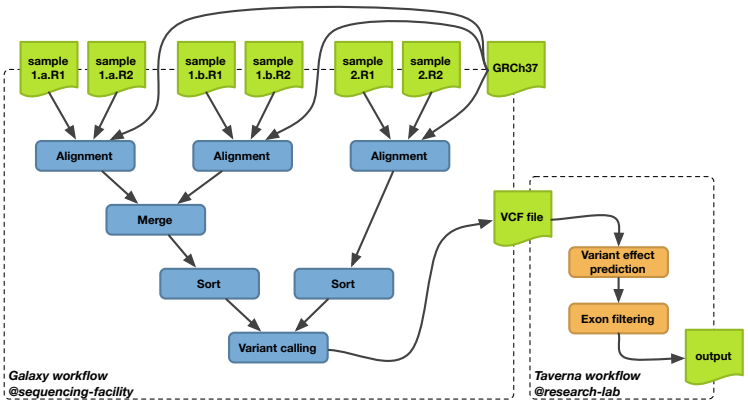

Fig. 3.1: A multi-site genomics workflow, involving Galaxy and Taverna systems.

The first workflow, in blue in Figure 3.1, is implemented in Galaxy [2] and addresses DNA data pre-processing which is loosely coupled to the scientific hypothesis. This workflow takes as input two DNA sequences from two biological samples `s1` and `s2`, represented in green. The second workflow is implemented with Taverna [5], and highly depends on scientific questions. It is generally conducted by life scientists possibly from different research labs and with less computational needs. In the following section, we explain how SHARP interlinks and harmonizes provenances graphs produced by two workflow systems.

4 Implementation

4.1 Galaxy PROV

For now, it is not possible to export provenance from Galaxy workflow environment[4] through a standard schema and format. To address this issue and to allow the combination of multiple workflow execution traces, we developed a loosely coupled tool aimed at exporting Galaxy user histories into PROV RDF data. This tool results in a command-line interface and a web application which allows users to list and export the content of their Galaxy Workflow histories, as illustrated in Figure 4.1.

Users provide the URL of their Galaxy workflow portal, and their private API key. Then, the tool communicates with the Galaxy API through the HTTP protocol and JSON documents. It can produce as a result RDF PROV triples in the *Turtle* syntax, or a graphical representation of the PROV sub-graph. The command line interface has been implemented in Java and makes use of the Jersey library for connecting to the REST API. The command line interface is available as open-source[5]. The back-end web application was also implemented in Java. Jetty was used as a standalone web server hosting an HTTP

[4] `https://usegalaxy.org`

[5] galaxy-PROV: `https://github.com/albangaignard/galaxy-PROV`

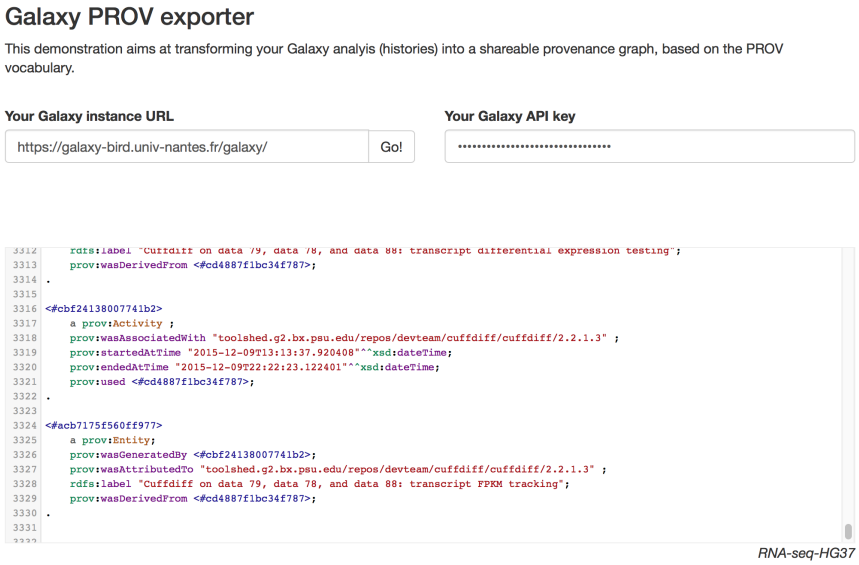

Fig. 4.1: A web application to export provenance graphs from Galaxy user histories in PROV.

REST API, and a mongoDB database was setup to collect basic usage metrics. The front-end was implemented in HTML and JavaScript (BackboneJS, D3.JS). The web application will be available shortly as open-source in the same repository.

4.2 Multi PROV harmonization

We implemented the PROV harmonization process in a command line tool available as open-source[6]. This tool can be used to infer new PROV statements for a single provenance trace by providing an input trace in the *Turtle* syntax. More interestingly, it can be used to interlink and harmonize cross-workflow provenance traces by specifying multiple provenance traces as input, accompanied with *owl:sameAs* statements. Finally, based on inferred *prov:wasInfluencedBy* predicates, cross-workflow data lineage can be visualized.

This tool has been implemented in Java and is supported by Jena[7] for RDF data management and reasoning tasks. PROV Constraints[8] inference rules have been implemented in the Jena syntax[9]. HTML and JavaScript (D3.JS) code templates have been used to generate harmonized provenance visualization.

The automatic generation of *owl:sameAs* statements between files based on hashing techniques, as well as a web interface are still under active development. These features should available soon in the same repository.

5 Results

This demonstration shows that SHARP allows to homogenize PROV graphs and achieve unified SPARQL queries across multiple provenance traces. For instance, the following query assembles a data influence graph between multiple workflow execution traces in which some data artifacts play two roles, i) *output* for a give workflow, and ii) *input* for other workflows.

[6] sharp-prov-toolbox: `https://github.com/albangaignard/sharp-prov-toolbox`
[7] Jena: `https://jena.apache.org`
[8] `https://www.w3.org/TR/prov-constraints/`
[9] `https://github.com/albangaignard/sharp-prov-toolbox/blob/master/SharpProvToolbox/src/main/resources/provRules_all.jena`

```
CONSTRUCT {
        ?x ?p ?y . ?x rdfs:label ?lx . ?y rdfs:label ?ly
} WHERE {
        ?x ?p ?y .
        FILTER (?p IN (prov:wasInfluencedBy)) .
        ?x rdfs:label ?lx . ?y rdfs:label ?ly }
```

This query matches *prov:wasInfluencedBy* properties resulting from harmonization process. These properties were not initially stated neither in the Taverna nor in the Galaxy provenance traces.

Figure 5.1 shows the resulting data lineage graph associated with the two workflow traces of our motivating use case (Figure 3.1). While the left part of the graphs represents the Galaxy workflow invocation, the right part represents the Taverna one.

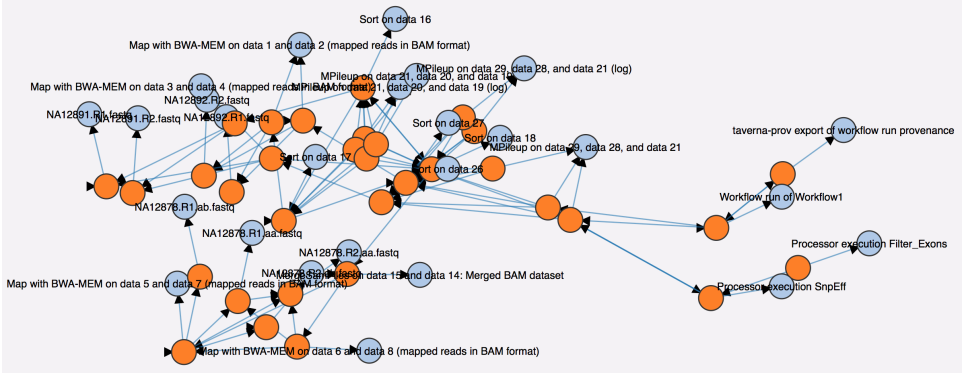

Fig. 5.1: *prov:wasInfluencedBy* properties between Galaxy and Taverna.

6 Conclusions

In this paper, we presented SHARP, a Linked Data approach for harmonizing cross-workflow provenance. The resulting harmonized provenance graph can be exploited to run cross-workflow queries. Our ongoing work includes the automation of *owl:sameAs* property generation as well as providing a unified web interface. As future works, we will address human-centered interpretation of possibly massive PROV graphs through domain-specific summarization techniques.

References

1. S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
2. E. Afgan, D. Baker, van den Beek, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1):W3–W10, 2016.
3. J. Cheney, P. Missier, and L. Moreau. Constraints of the provenance data model. Technical report, 2012.
4. R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. Data exchange: semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.
5. K. Wolstencroft, R. Haines, D. Fellows, et al. The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. *Nucleic Acids Research*, 41(Webserver-Issue):557–561, 2013.