



# From scientific workflow patterns to 5-star linked open data

Alban Gaignard, Hala Skaf-Molli, Audrey Bihouée

## ► To cite this version:

Alban Gaignard, Hala Skaf-Molli, Audrey Bihouée. From scientific workflow patterns to 5-star linked open data. TaPP 2016: 8th USENIX Workshop on the Theory and Practice of Provenance, Jun 2016, Washington D.C., United States. hal-01768449

**HAL Id: hal-01768449**

**<https://hal.science/hal-01768449>**

Submitted on 17 Apr 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# From scientific workflow patterns to 5-star linked open data

Alban Gaignard

Nantes Academic Hospital, France  
alban.gaignard@univ-nantes.fr

Hala Skaf-Molli

LINA – Nantes University, France  
hala.skaf@univ-nantes.fr

Audrey Bihouée

INSERM, UMR1087, L'Institut du  
Thorax, Nantes, France  
CNRS, UMR 6291, Nantes, France  
Université de Nantes, France  
audrey.bihouee@univ-nantes.fr

## Abstract

Scientific Workflow management systems have been largely adopted by data-intensive science communities. Many efforts have been dedicated to the representation and exploitation of provenance to improve reproducibility in data-intensive sciences. However, few works address the mining of provenance graphs to annotate the produced data with domain-specific context for better interpretation and sharing of results. In this paper, we propose *PoeM*, a lightweight framework for mining provenance in scientific workflows. *PoeM* allows to produce linked *in silico* experiment reports based on workflow runs. *PoeM* leverages semantic web technologies and reference vocabularies (*PROV-O*, *P-Plan*) to generate provenance mining rules and finally assemble linked scientific experiment reports (*Micropublications*, *Experimental Factor Ontology*). Preliminary experiments demonstrate that *PoeM* enables the querying and sharing of Galaxy<sup>1</sup>-processed genomic data as 5-star linked datasets.

**Keywords** Scientific Workflows, Provenance, Rules, Linked Data

## 1. Introduction

Life scientists generate tremendous amounts of biological data, especially in the field of genomics. The availability of next generation sequencing equipments led to an unprecedented growth of sequenced human genomes. The number of data has been doubled every 7 months (Stephens et al. 2015), which even exceeds manufacturers own predictions. Not only the volume of acquired raw genomic sequences is rapidly growing, but also the volume of high value-added processed data. The question is: are the underlying infrastructures ready to preserve both the raw and the processed data ? **Sharing and reusing high value-added biomedical data becomes crucial to limit the duplication of computing and storage efforts.**

Many guidelines to publish Findable, Accessible, Interoperable and Reusable datasets have been proposed<sup>2</sup>. Tim Berners Lee pro-

poses the 5-star model<sup>3</sup> to enhance data sharing through Linked Open Data principles. These principles have been used to assemble reference Linked Open Dataset and Ontologies in Life Sciences (Noy et al. 2009; Callahan et al. 2013; Jupp et al. 2014). **Although life scientists benefit from these curated and trusted open databases in their daily practice, it is often not feasible for small research groups to publish their processed data through the 5-star recommendations.** This is mainly due to the human cost and the technicality of data curation activities.

In Life Sciences, scientific workflows systems such as Galaxy, Taverna, Vistrails, or Wings/Pegasus have gained a large adoption because they define explicitly the main parameters and processing steps, and enhance, therefore, trust in the produced results. A lot of approaches address provenance capture and management towards better reproducibility of *in silico* experiments. Several provenance models have been proposed (OPM, PROV, ProvOne, PAV). However, few works (Alper et al. 2014) address provenance exploitation towards better sharing of massively produced data.

In this paper, we address the issue of sharing data produced by scientific workflow engines by reusing Linked Open Vocabularies. Our in progress work relies on manually annotated workflow patterns, and rules generator. The rules mine generic provenance metadata and produce domain-specific linked experiment reports. **We propose with *PoeM*, a method for populating Linked Data repositories (Bizer et al. 2009) with experiment reports, at a reduced data curation cost.**

The paper is organized as follows. Section 2 presents a motivating scenario in the field of bioinformatics. Section 3 describes our approach for mining generic provenance metadata. Section 4 reports preliminary results in a real-life experiment. Section 5 summarizes related works. Finally, conclusion and future works are outlined in Section 6.

## 2. Motivating example

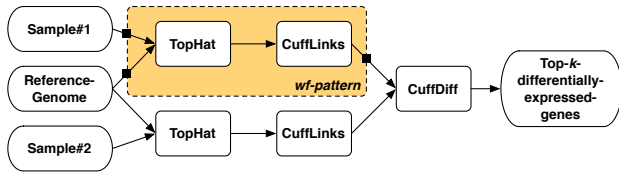
Our work is motivated by data management issues raised in the field of bioinformatics and genomics. *RNAseq* is a high throughput sequencing technology aimed at measuring gene expression levels from multiple experimental conditions. The goal is to identify genetic markers involved in biological or pathological processes.

Figure 1 illustrates a typical *RNAseq* data analysis workflow (Trapnell et al. 2012). The first step consists in mapping the RNA sequence reads of two biological samples (*Sample#1* and *Sample#2*) to an annotated reference genome (*Reference-Genome*). The resulting mapped reads are counted with *CuffLinks* to quantify the expression of each gene (or transcript) in the corresponding biological sample. Finally, based on the initial experimental conditions, *CuffDiff* is responsible for the analysis and selection of the

<sup>1</sup> <https://galaxyproject.org>

<sup>2</sup> <https://www.force11.org/group/fairgroup>

<sup>3</sup> <http://5stardata.info/en/>



**Figure 1.** A typical RNA-seq bioinformatics workflow aimed at filtering the top- $k$  differentially expressed genes.

top- $k$  most differentially expressed genes. *TopHat* and *Cufflinks* are CPU intensive tasks and their re-computation should be avoided whenever possible. For instance, a typical RNA-seq sample alignment to a reference human genome may involve 2 paired-end input sequences of 17 GB each, and produce, after 170 hours of single-core computing, a 12 GB aligned sequence. These computations are heavy and time consuming. Even if parallelization helps, the computational cost is still challenging, *e.g.*, in a study with hundreds of biological samples. Without descriptive enough metadata, the reuse and sharing of produced data is particularly difficult.

Most of scientific workflow management systems address reproducibility and interoperability issues through the capture of provenance metadata. PROV<sup>4</sup> is the *de facto* standard for describing and exchanging provenance graphs. PROV is a domain-agnostic provenance ontology relying on entities, activities and agents (software or people). In bioinformatics applications, PROV can be used to document data analysis, at a very fine grain (at each tool invocation), as well as attribution and versioning. However, when it comes to share and reuse produced raw data, PROV traces fail in describing required domain-specific informations for life scientists, such as the associated experimental condition, the biological or medical hypothesis or the nature of biological samples and results.

### Need for domain-specific linked experiment reports

TCGA<sup>5</sup> and ICGC<sup>6</sup> are international initiatives aimed at sharing multi-modality clinical, imaging, and omics datasets in the context of cancer research. Providing meaningful domain-specific metadata is the cornerstone of successful data sharing and reuse in cancer research. Existing data processing tools and pipeline should be able to generate these meaningful metadata based on biomedical context.

Several reference domain ontologies are already available to represent these metadata. In our motivating example, EDAM<sup>7</sup> (Ison et al. 2013) could be used as a common terminology to describe the nature of the processing tools involved in bioinformatics workflows, as well as the format and the nature of tools parameters. Regarding the representation of biomedical experimental factors, EFO<sup>8</sup>, the *Experimental Factor Ontology* (Malone et al. 2010), is of particular interest. Finally, the *Micropublications* (Clark et al. 2014) enables to formally represent scientific approaches and evidences towards machine-tractable academic papers. However, documenting datasets produced by scientific workflow managements systems with these domain-specific ontologies is generally a manual task that requires a deep knowledge of semantic web technologies and ontologies.

<sup>4</sup> <https://www.w3.org/TR/prov-o/>

<sup>5</sup> <https://browser.cghub.ucsc.edu>

<sup>6</sup> <https://dcc.icgc.org>

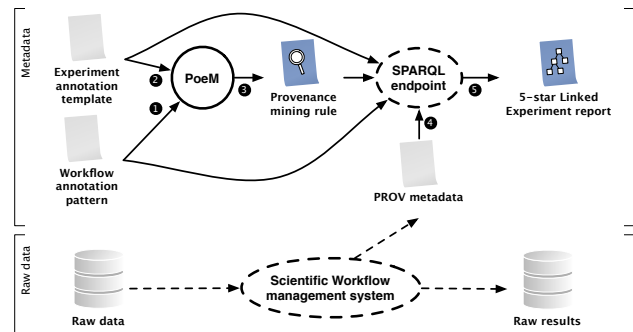
<sup>7</sup> <https://biportal.bioontology.org/ontologies/EDAM>

<sup>8</sup> <https://biportal.bioontology.org/ontologies/EFO>

A coherent set of domain-specific metadata would bridge together biomedical claims and evidences, experimental factors, and produced data. This would definitely accelerate the availability of query-able data repositories in the direction of machine- and human-tractable scientific reports.

### 3. Provenance mining rules

We propose *PoeM*, an approach aimed at rewriting annotated workflow patterns and experiment report templates into provenance mining rules. These rules are used for the annotation of scientific data. They are grounded to the PROV vocabulary to be as much as possible independent from any workflow management system implementation.



**Figure 2.** Automated generation of provenance mining rules to produce 5-star linked experiment reports.

Figure 2 describes the mains steps of *PoeM*. Our contribution is represented with plain arrows, the results produced by *PoeM* are highlighted in blue. We make the hypothesis that the underlying workflow enactment system is capable of producing PROV metadata to document each data consumption and production activities. Step ① consists in manually annotating workflow patterns with domain specific concepts. Step ② consists in manually annotating an experiment template to capture domain-specific knowledge. These annotations cover the nature of the experiment parameters, the expected results, as well as the associated scientific hypothesis or evidences. The core of our contributions is in step ③. We propose a query rewriting algorithm responsible for the generation of semantic web rules, materialized by SPARQL CONSTRUCT queries. Step ④ consists in extracting PROV metadata from workflow engine execution traces. Finally step ⑤ consists in applying the resulting rule on PROV metadata to generate a 5-star linked experiment report. This report can document the raw produced data with meaningful domain concepts, including the associated scientific hypothesis or evidences.

#### Workflow annotation patterns (①)

More than forty Workflow patterns<sup>9</sup> have been proposed to capture dependencies between process activities, *e.g.*, sequence, parallelism, choice, synchronization, *etc.*. These patterns were developed to address business process requirements (van der Aalst et al. 2003). They can also apply to address scientific workflow requirement as detailed in (Yildiz, U. et al 2009). In Figure 1, we use the *Sequence* pattern where an activity is enabled after the completion of another activity in the same workflow. We rely on *Step* class and on *isPrecededBy* property of *P-Plan* (Garijo et al. 2012) ontology to describe the sequence pattern. In addition, we rely on

<sup>9</sup> <http://www.workflowpatterns.com/>



through internal provenance representation and exploitation. These approaches give valuable insight on the nature and parameters of experiments, however, we argue in *PoeM* that Open Linked Data approaches are a step towards machine- and human-tractable experiment reports.

## 6. Conclusion and perspectives

We propose *PoeM*, a provenance mining approach aimed at populating 5-star Linked Open Data repositories with scientific experiment reports. *PoeM* is non-invasive and can adapt to different workflow engines that export PROV metadata. *PoeM* is a declarative lightweight approach based on semantic web standards, and does not require additional programming effort.

*PoeM* is in progress work and presents many opportunities. For the moment the identification of processing steps and variables is based on label matching. We plan to use semantic tagging to improve genericity. We also plan to extend the supported workflow patterns with multi-sequences, split-and-join patterns, or based on common motifs (Garijo et al. 2014). To assess the accuracy of the produced linked experiment reports, we plan to conduct a user evaluation based on competency questions. To assess the versatility of *PoeM*, we also plan to produce Research Objects (Belhajjame et al. 2015) as a machine-tractable way of reporting research. Finally, as a continuation of this work, we will evaluate how *PoeM* can adapt to other domains such as bioimaging, high energy physics workflows, and how it can scale to face large scale, real-life, data-science research studies in the context of the SyMeTRIC collaborative personalized medicine project.

## Acknowledgments

This work has been supported by the SyMeTRIC project, funded by the Région Pays de la Loire Connect Talent research call. We are grateful to the BiRD bioinformatics facility for providing support and computing resources.

## References

- Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomic? *PLoS Biol*, 13(7):e1002195, 07 2015.
- S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, and A. M. Jenkinson. The EBI RDF platform: Linked open data for the life sciences. *Bioinformatics*, 2014.
- A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier. Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In *The semantic web: semantics and big data*, pages 200–212. Springer, 2013.
- N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, et al. Biportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440, 2009.
- J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 04 2010.
- J. Ison, M. Kalaš, I. Jonassen, D. Bolser, M. Uludag, H. McWilliam, J. Malone, R. Lopez, S. Pettifer, and P. Rice. Edam: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, 29(10):1325–1332, 05 2013.
- T. Clark, P. N. Ciccarese, and C. A. Goble. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, 5(1):1–33, 2014.
- C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of rna-seq experiments with topht and cufflinks. *Nat. Protocols*, 7(3):562–578, 03 2012.
- C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.
- W. van der Aalst, A. ter Hofstede, B. Kiepuszewski, and A. Barros. Workflow patterns. *Distributed and Parallel Databases*, 14(1):5–51, 2003.
- U. Yildiz, I. Guabtni, and A.H. Ngu, A. H. Towards scientific workflow patterns. In *the 4th Workshop on Workflows in Support of Large-Scale Science*, (p. 13). ACM. 2009.
- D. Garijo and Y. Gil. Augmenting prov with plans in p-plan: Scientific processes as linked data. In *Second International Workshop on Linked Science: Tackling Big Data (LISC), held in conjunction with the International Semantic Web Conference (ISWC)*, Boston, MA, 2012.
- A. Gaignard, J. Montagnat, B. Gibaud, G. Forestier, and T. Glatard. Domain-specific summarization of life-science e-experiments from provenance traces. *Web Semantics: Science, Services and Agents on the World Wide Web*, 29:19 – 30, 2014. Life Science and e-Science.
- P. Alper, K. Belhajjame, C. A. Goble, and P. Karagoz. *Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014. Revised Selected Papers*, chapter LabelFlow: Exploiting Workflow Provenance to Surface Scientific Data Provenance, pages 84–96. Springer International Publishing, Cham, 2015.
- F. Chirigati, D. Shasha, and J. Freire. Reprozip: Using provenance to support computational reproducibility. In *Presented as part of the 5th USENIX Workshop on the Theory and Practice of Provenance*, Lombard, IL, 2013. USENIX.
- K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. Gómez-Pérez, S. Bechhofer, G. Klyne, and C. Goble. Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:16 – 42, 2015.
- D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. Common motifs in scientific workflows: An empirical analysis. *Future Generation Computer Systems*, 36:338 – 351, 2014.

## A. Detailed algorithms

---

**Algorithm 2: genProvBGP** to generate a BGP matching a provenance path between two processing steps.

---

**Input** :  $W$  the set of RDF triples describing the workflow pattern.  
**Output**:  $provBGP$  the set of triple patterns matching a PROV path,  
 $M$  a HashMap binding workflow pattern variables to SPARQL query variables.

```

1 begin
2    $S_1 \leftarrow getFirstStepTriples(W)$ 
3    $S_2 \leftarrow getLastStepTriples(W)$ 
4    $provBGP \leftarrow "?step1 \text{ rdfs:type } prov:Activity \text{ .}"$ 
5    $provBGP \leftarrow "?step1 \text{ prov:wasAssociatedWith } ?soft1 \text{ .}"$ 
6
7   /* Iterate over pattern INPUT variables */
8    $i \leftarrow 0$ 
9   foreach ( $inVar \in S_1$ ) do
10     $M \leftarrow (inVar, "?in"+i+"S1.")$ 
11     $provBGP \leftarrow "?step1 \text{ prov:used } ?in"+i+"S1."$ 
12     $i \leftarrow i + 1$ 
13
14    $provBGP \leftarrow "?outS1 \text{ prov:wasGeneratedBy } ?step1 \text{ .}"$ 
15    $provBGP \leftarrow "?in2 \text{ (rdfs:label | ^rdfs:label | } \text{ prov:wasDerivedFrom)* } ?outS1 \text{ .}"$ 
16    $provBGP \leftarrow "?step2 \text{ prov:used } ?in2 \text{ .}"$ 
17    $provBGP \leftarrow "?step2 \text{ prov:wasAssociatedWith } ?soft2 \text{ .}"$ 
18
19   /* Iterate over pattern OUTPUT variables */
20    $j \leftarrow 0$ 
21   foreach ( $outVar \in S_2$ ) do
22     $M \leftarrow (outVar, "?out"+j+"S2.")$ 
23     $provBGP \leftarrow "?out"+j+"S2 \text{ prov:wasGeneratedBy } ?step2 \text{ .}"$ 
24     $j \leftarrow j + 1$ 

```

---



---

**Algorithm 3: genReportBGP** to generate a BGP describing the linked experiment report.

---

**Input** :  $M$  a HashMap binding workflow pattern variables to SPARQL variables,  
 $R$  the set of triples describing the linked experiment report template.  
**Output**:  $reportBGP$  the set of triple patterns describing the annotation template.

```

1 begin
2   foreach ( $t \in R$ ) do
3     if ( $t.getPredicate()$  matches "oa:hasTarget") then
4        $o \leftarrow t.getObject()$ 
5        $t.replaceObject(M.get(o))$ 
6        $reportBGP \leftarrow t$ 

```

---



---

**Algorithm 4: genProvMiningRule** to write the resulting SPARQL CONSTRUCT query.

---

**Input** :  $P$  the set of PROV triples resulting from a workflow run,  
 $R$  the set of RDF triples describing the annotation template,  
 $W$  the set of RDF triples describing the workflow pattern.  
**Output**:  $Q$  the SPARQL CONSTRUCT provenance mining rule.

```

1 begin
2    $Q \leftarrow mergePrefixes(W, R, P)$ 
3    $Q \leftarrow "CONSTRUCT \{ " + genReportBGP(W, M) + " \}"$ 
4    $Q \leftarrow "WHERE \{ " + genProvBGP(W, M) + " \}"$ 

```

---