Findable and Reusable Workflow Data Products: A Genomic Workflow Case Study

Editor(s): Tomi Kauppinen, Aalto University, Finland ; Daniel Garijo, University of Southern California, USA ; Natalia Villanueva, The University of Texas at El Paso, USA
Solicited review(s): Tomi Kauppinen, Aalto University, Finland ; Angelo Salatino, The Open University, UK ; Monika Akbar, The University of Texas at El Paso, USA

Alban Gaignard [a,*], Hala Skaf-Molli [b] and Khalid Belhajjame [c]
[a] *l'institut du thorax, INSERM, CNRS, University of Nantes, Nantes, France*
E-mail: alban.gaignard@univ-nantes.fr
[b] *LS2N, University of Nantes, Nantes, France*
E-mail: hala.skaf@univ-nantes.fr
[c] *PSL, Université Paris-Dauphine, LAMSADE, Paris, France*
E-mail: kbelhajj@googlemail.com

Abstract.
While workflow systems have improved the repeatability of scientific experiments, the value of the processed (intermediate) data have been overlooked so far. In this paper, we argue that the intermediate data products of workflow executions should be seen as first-class objects that need to be curated and published. Not only will this be exploited to save time and resources needed when re-executing workflows, but more importantly, it will improve the reuse of data products by the same or peer scientists in the context of new hypotheses and experiments. To assist curator in annotating (intermediate) workflow data, we exploit in this work multiple sources of information, namely: i) the provenance information captured by the workflow system, and ii) domain annotations that are provided by tools registries, such as Bio.Tools. Furthermore, we show, on a concrete bioinformatics scenario, how summarising techniques can be used to reduce the machine-generated provenance information of such data products into concise human- and machine-readable annotations.

Keywords: FAIR, Linked Data, Scientific Workflows, Provenance, Bioinformatics, Data Summaries

1. Introduction

We have witnessed in the last decade a paradigm shift in the way scientists conduct their experiments, which are increasingly data-driven. Given a hypothesis that the scientist seeks to test, verify or confirm, s/he processes given input datasets using an experiment which is modelled as a series of scripts written in languages such as R, Python and Perl, or pipelines composed of connected modules (also known as workflows [1,2]). For example, the recent progress in sequencing technologies, combined with the reduction of their cost has led to massive production of genomic data with growth rates that exceed major manufacturers' expectations [3]. A single research lab that is using the last generation sequencer can currently generate in one year[1] the equivalent of the world-wide collaborative sequencing capacity in 2012 [4].

The datasets obtained as a result of the experiment are analyzed by the scientist who then reports on the finding s/he obtained by analyzing the results [5]. As a response to the reproducibility movement [6], which has gained great momentum recently, scientists were

[*] Corresponding author.

[1] Theoretically around 2500 whole genomes per year with an Illumina NovaSeq technology

encouraged to not only report on their findings, but also document the experiment (method) they used, the datasets they used as inputs, and eventually, the datasets obtained a result. To assist scientist in the task of reporting, a number of methods and tools have been proposed (see e.g., [7,8,9]). In [10] Gil *et al.* propose data narratives to automatically generate text to describe computational analyses that can be presented to users and ultimately included in papers or reports.

While we recognize that such proposals are of great help to the scientists and can be instrumental to a certain extent to check the repeatability of experiments, they are missing opportunities when it comes to the reuse of the intermediate data products that are generated by their experiments. Indeed, the focus in the reports generated by the scientist is put on their scientific findings, documenting the hypothesis and experiment they used, and in certain cases, the datasets obtained as a result of their experiment. The intermediate datasets, which are by-products of the internal steps of the experiment, are in most cases buried in the provenance of the experiment if not reported at all. The availability of such intermediate datasets can be of value to third-party scientists to run their own experiment. This does not only save time for those scientists in that they can use readily available datasets but also save time and resources since some intermediate datasets are generated using large-scale resource- and compute-intensive scripts or modules.

We argue that intermediate datasets generated by the steps of an experiment should be promoted as first-class objects on their own right, to be findable, accessible and ultimately reusable by the members of the scientific community. We focus, in this paper, on datasets that are generated by experiments that are specified and enacted using workflows. There has been recently initiatives, notably FAIR [11], which specify the guidelines and criteria that need to be met when sharing data in general. Meeting such criteria remains challenging, however.

In this paper, we show how we can combine provenance metadata with external knowledge associated with workflows and tools to promote processed data sharing and reuse. More specifically, we present FRESH an approach to associate the intermediate, as well as the final, datasets generated by the workflows with annotations specifying their retrospective provenance and their prospective provenance (i.e., the part of the workflow that was enacted for their generation). Both prospective and retrospective provenance can be overwhelming for a user to understand them. Because of this, we associate datasets with a summary of their prospective provenance. Moreover, we annotate the datasets with information about the experiment that they were used in, e.g., hypothesis, contributors, as well as with semantic domain annotations that we automatically harvest from third-party resources, in particular, Bio.Tools[2] [12]. Our ultimate objective is to promote processed data reuse in order to limit the duplication of computing and storage efforts associated to workflow re-execution.

The contributions of this paper are the following:

- Definition of workflow data products reuse in the bioinformatics domain.
- A knowledge-graph based approach aimed at annotating raw processed data with domain-specific concepts, while limiting domain experts overwhelming at the time of sharing their data.
- An experiment based on a real-life bioinformatics workflow, that can be reproduced through an interactive notebook.

This paper is organised as follows. Section 2 presents motivation and defines the problem statement. Section 3 details the proposed FRESH approach. Section 4 presents our experimental results. Section 5 summarises related works. Finally, conclusions and future work are outlined in Section 6.

2. Motivations and Problem Statement

We motivate our proposal through an exome-sequencing bioinformatics workflow. This workflow aims at (1) aligning sample exome data (the protein-coding region of genes) to a reference genome and (2) identifying genetic mutations for each of the biological samples. Figure 1 drafts a summary of the bioinformatics analysis tasks required to identify and annotate genetic variants from exome-sequencing data. For a matter of clarity, we hide in this scenario some of the minor processing steps such as the sorting of DNA bases, but they are still required in practice. The real workflow will be described in detail in the experimental results section.

This workflow consumes as inputs two sequenced biological samples `sample_1` and `sample_2`. For each sample, sequencers produce multiple files that need to be merged later on (`Sequence merging`

[2] http://bio.tools/

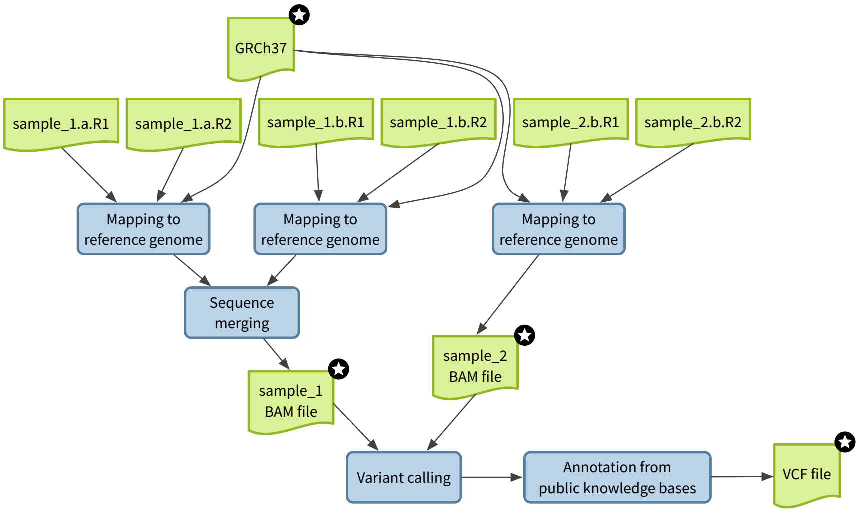

Fig. 1. A typical bioinformatics workflow aimed at identifying and annotating genomic variations from a reference genome. Green waved boxes represent data files, and blue rounded boxes represent processing steps.

step). The first processing step consists in aligning [13] (`Mapping to reference genome`) the short sequence reads to the human reference genome (`GRCh37`). Then, for each sample, data are merged and post-processed [14,15] and result in binary (`BAM`) files representing the aligned sequences with their quality metrics. Finally, from these aligned sequences, the genetic variants are identified [16] and enriched with annotations [17] gathered from public knowledge bases such as DBsnp [18] or gnomAD[3]. This last processing step results in a VCF file listing, for all processed sample sequences, all known genomic variations compared to the `GCRh37` reference genome.

Performing these analyses in real-life conditions is computation intensive. They require a lot of CPU time and storage capacity. As an example, similar workflows are run in production in the CNRGH french national sequencing facility. For a typical exome-sequencing sample (9.7GB compressed), it has been measured that 18.6GB was necessary to store the input and output compressed data. In addition, 2 hours and 27 minutes were necessary to produced an annotated VCF variant file, taking advantage of parallelism in a dedicated high-performance computing infrastructure (7 nodes with 28 CPU Intel Broadwell cores each), which corresponds to 158 cumulative hours for a single sample, i.e. 6 days of computation on a single CPU.

Considering the computational cost of these analyses, we claim that the secondary use of data is critical to speed-up Research addressing similar or related topics. In this workflow, all processing steps produce data but they do not provide the same level of reusability. We tagged reusable data with a white star in Figure 1. More precisely, (`GRCh37`) is by nature highly reusable since it is a reference “atlas” for genomic human sequences, and results from state-of-the-art scientific knowledge at a given time. Then, `BAM` files can also be considered as more reusable than the raw input data since they have been aligned to this atlas and thus benefit from consensual knowledge on this genome. As an example, they provide the relationship between sequences and known genes, they can be visualized in a genome viewer, they can also be reused to regenerate raw unmapped sequences.

From the scientist perspective, answering questions such as *"can or should I reuse these files in the context of my research study"* is still challenging. To reuse the final VCF variant file, it is of major importance to know the version of the reference genome as well as to clearly understand the scientific context of the study, the phenotypes associated to the samples, as well as the possible relations between samples. Finally, having precise information on the variant calling algorithm is also critical due to application-specific detection thresholds [19]. More generally, not only fine-grained provenance information regarding data and tools lineage are required but also domain-specific annotations based on community agreed vocabularies (Issue 1). These vocabularies exist but annotating processed data

[3] `https://gnomad.broadinstitute.org`

with domain-specific concepts requires a lot of time and expertise (Issue 2).

In this work, we show how we can improve the findability and reusability of workflow (intermediate) data by leveraging (1) community efforts aimed at semantically cataloguing bioinformatics processing tools to reduce the solicitation of domain experts, and (2) the automation and provenance capabilities of workflow management systems to automate the annotation of processed data, towards more reusable workflow results.

3. FRESH Approach

FRESH is an approach to improve the *Findability* and the *Reusability* of genomic workflow data. FAIR [11,20] and Linked data[21,22] principles constitute the conceptual and technological backbones in this direction.

FRESH partially tackles FAIR requirements for better *findability* and *reusability*. We address *findability*, by relying on Linked Data best practices, namely associating a URI to each dataset, linking these datasets in the form of RDF knowledge graphs with controlled vocabularies for the naming of concepts and relations.

Being tightly coupled to scientific context, *reusability* is more challenging to achieve. Guidelines have been proposed for FAIR sharing of genomic data [23], however, proposing and evaluating *reusability* is still a challenging and work in progress [24]. In this work, we focus on reusable data as *annotated with sufficiently complete information allowing, without needs for external resources: traceability, interpretability, understandability, and usage by humans or machines*.

To be traceable, provenance traces are mandatory for tracking the data generation process.

To be interpretable, contextual data [11] are mandatory, this includes: i) Scientific context in terms of Claims, Research lab, Experimental conditions, previous evidence (academic papers). ii) The technical context in terms of material and methods, data sources, used software (algorithm, queries) and hardware.

To be understandable by itself, data must be annotated with domain-specific vocabularies. For instance, to capture knowledge associated with the data processing steps, we can rely on EDAM[4] which is actively developed and used in the context of the Bio.Tools registry, and which organizes common terms used in the field of bioinformatics. However, these annotations on processing tools do not capture the scientific context in which a workflow takes place. To address this issue, we rely on the *Micropublications* [25] ontology which has been proposed to formally represent scientific approaches, hypothesis, claims, or pieces of evidence, in the direction of machine-tractable academic papers.

Figure 2 illustrates our approach to provide more reusable data. The first step consists in capturing provenance for all workflow runs. PROV[5] is the *de facto* standard for describing and exchanging provenance graphs. Although capturing provenance can be easily managed in workflow engines, there is no systematic way to link a PROV *Activity* (the actual execution of a tool) to the relevant software *Agent* (*i.e.* the software responsible for the actual data processing). To address this issue we propose to provide, at workflow design time, the tool's identifier in the tool catalogue. This allows to generate a provenance trace which associates (*prov:wasAssociatedWith*) each execution, and thus each consumed and produced data to the software identifier.

Then, we assemble a bioinformatics knowledge graph which links together (1) the tools annotations, gathered from the Bio.Tools registry, and providing information on the functions of the tools (bioinformatics EDAM *operations*) and which kind of data they consume and produce, (2) the complete EDAM ontology, to gather for instance the community-agreed definitions and synonyms for bioinformatics concepts, (3) the PROV graph resulting from a workflow execution which provides fine-grained technical and domain-agnostic provenance metadata, and (4) the experimental context using Micro-publication for scientific claims and hypothesis associated to the experiment.

Finally, based on domain-specific provenance queries, the last step consists in extracting few and meaningful data from the knowledge graph, to provide scientist with more reusable intermediate or final results, and to provide machines findable and query-able data stories.

In the remainder of this section, we rely on the SPARQL query language to interact with the knowledge graph in terms of knowledge extraction and knowledge enrichment.

[4] `http://edamontology.org`

[5] https://www.w3.org/TR/prov-o/

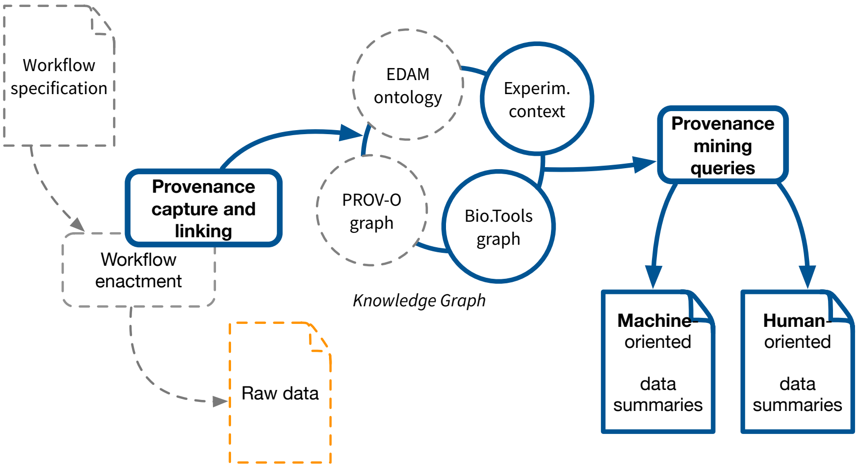

Fig. 2. Knowledge graph based on workflow provenance and tool annotations to automate the production of human- and machine- oriented data summarises.

```
SELECT ?d_label ?title ?f_def ?st WHERE {
    ?d rdf:type prov:Entity ;
        prov:wasGeneratedBy ?x ;
        prov:wasAssociatedWith ?tool ;
        rdfs:label ?d_label .

    ?tool dc:title ?title ;
        biotools:has_function ?f .

    ?f rdfs:label ?f_label ;
        oboInOwl:hasDefinition ?f_def .

    ?c rdf:type mp:Claim ;
        mp:statement ?st .
}
```

Query 1: SPARQL query aimed at linking processed data to the processing tool and the definition of what is done on data.

```
CONSTRUCT {
    ?x2 p-plan:wasPreceededBy ?x1 .
    ?x2 prov:wasAssociatedWith ?t2 .
    ?x1 prov:wasAssociatedWith ?t1 .
    ?t1 biotools:has_function ?f1 .
    ?f1 rdfs:label ?f1_label .
    ?t2 biotools:has_function ?f2 .
    ?f2 rdfs:label ?f2_label .
} WHERE {
    ?d2 prov:wasDerivedFrom ?d1 .

    ?d2 prov:wasGeneratedBy ?x2 ;
        prov:wasAssociatedWith ?t2 ;
        rdfs:label ?d2_label .

    ?d1 prov:wasGeneratedBy ?x1 ;
        prov:wasAssociatedWith ?t1 ;
        rdfs:label ?d1_label .

    ?t1 biotools:has_function ?f1 .
    ?f1 rdfs:label ?f1_label .

    ?t2 biotools:has_function ?f2 .
    ?f2 rdfs:label ?f2_label .
}
```

Query 2: SPARQL query aimed at assembling an abstract workflow based on what happened (provenance) and how data were processed (domain-specific EDAM annotations).

Query 1 aims at extracting and linking together data artefacts with the definition of the bioinformatics process they result from. In this SPARQL query, we first identify data (*prov:Entity*), the tool execution they result from (*prov:wasGeneratedBy*), and the used software (*prov:wasAssociatedWith*). Then we retrieve from the Bio.Tools sub-graph the EDAM annotation which specify the function of the tool (*biotools:has_function*). The definition of the function of the tool is retrieved from the EDAM ontology (*oboInOwl:hasDefinition*). Finally, we retrieve the scientific context of the experiment by matching statements expressed in natural language (*mp:Claim*, *mp:statement*).

The Query 2 shows how a specific provenance pattern can be matched and reshaped to provide a summary of the main processing steps, in terms of domain-specific concepts. The idea consists in identifying all data derivation links (*prov:wasDerivedFrom*). From the identified data, the tool executions are then matched, as well as the corresponding software agents. Similarly, as in the previous query, the last piece of information to be identified is the functionality of the tools. This is done by exploiting the *biotools:has_function* predicate. Once this graph pattern is matched, a new graph is created using a CONSTRUCT query clause, to represent an ordered chain of processing steps (*p-plan:wasPreceededBy*).

4. Experimental results and Discussion

4.1. *Raw provenance traces from a bioinformatics workflow execution*

We experimented our approach on a production-level exome-sequencing workflow[6], designed and operated by the GenoBird genomic and bioinformatics core facility. It implements the motivating scenario we introduced in section 2. We assume that, based on the approach beforehand presented, the workflow has been run, the associated provenance has been captured and the knowledge graph has been assembled.

The resulting provenance graph consists in an RDF graph with 555 triples leveraging the PROV-O ontology. The following two tables show the distribution of PROV classes and properties.

Table 1

Number of instances per PROV class, resulting from the execution of the exome-sequencing workflow.

Classes	Number of instances
prov:Entity	40
prov:Activity	26
prov:Bundle	1
prov:Agent	1
prov:Person	1

Table 2

Number of predicates per PROV and RDF(S) property, resulting from the execution of the exome-sequencing workflow.

Properties	Number of predicates
prov:wasDerivedFrom	167
prov:used	100
rdf:type	69
prov:wasAssociatedWith	65
prov:wasGeneratedBy	39
rdfs:label	39
prov:endedAtTime	26
prov:startedAtTime	26
rdfs:comment	22
prov:wasAttributedTo	1
prov:generatedAtTime	1

Interpreting this provenance graph is challenging from a human perspective due to the number of nodes and edges and, more importantly, due to the lack of domain-specific terms.

4.2. *Human-oriented data summaries*

Based on query 1 and a textual template, we show in Figure 3 sentences which have been automatically generated from the knowledge graph. They intend to provide scientists with self-explainable information on how data were produced, and in which scientific context, leveraging domain-specific terms.

```
[...]
The file <Samples/Sample1/BAM/Sample1.final.bam>
results from tool <gatk2_print_reads-IP> which
<Counting and summarising the number of short
sequence reads that map to genomic features.>
It was produced in the context of <Rare Coding
Variants in ANGPTL6 Are Associated with Familial
Forms of Intracranial Aneurysm>
[...]
The file <VCF/hapcaller.recal.combined.annot.
gnomad.vcf.gz> results from tool
<gatk2_variant_annotator-IP> which <Predict the
effect or function of an individual single
nucleotide polymorphism (SNP).>
It was produced in the context of <Rare Coding
Variants in ANGPTL6 Are Associated with Familial
Forms of Intracranial Aneurysm>
[...]
```

Fig. 3. Sentence-based data summaries providing, for a given file, information on the tool the data originates from, and the definition of what does the tool, based on the EDAM ontology.

Complex data analysis procedures would require a long text and many logical articulations for being understandable. Visual diagrams provide a compact representation for complex data processing and constitute thus an interesting mean to assemble human-oriented data summaries.

Figure 4 shows a summary diagram automatically compiled from the bioinformatics knowledge graph previously described in section 3. Black arrows represent the logical flow of data processing, and black ellipses represent the nature of data processing, in terms of EDAM operations. The diagram highlights in blue the `Sample1.final.bam`. It shows that this file results from a *Read Summarisation* step and is followed by a *Variant Calling* step.

Another example for summary diagrams is provided in Figure 5 which highlights the final VCF file and its binary index. The diagram shows that these files result from a processing step performing a *SNP annotation*, as defined in the EDAM ontology.

[6] `https://gitlab.univ-nantes.fr/bird_pipeline_registry/exome-pipeline`

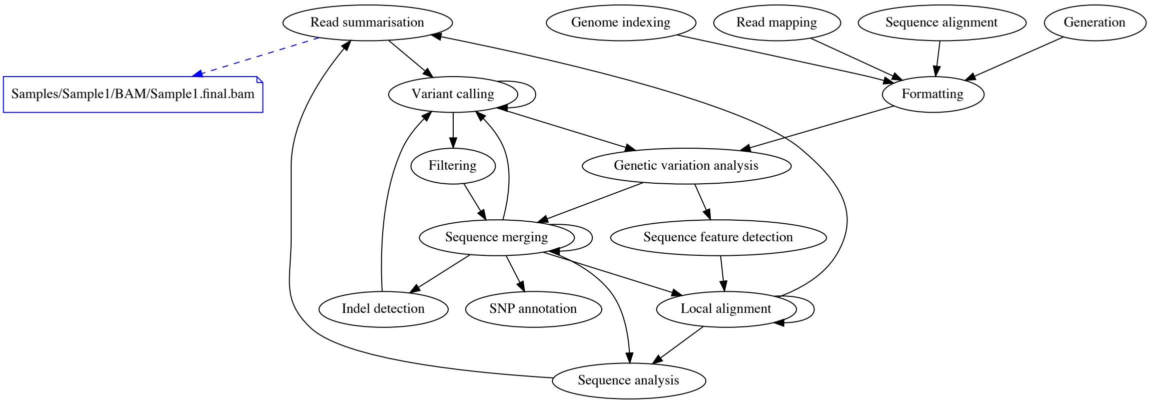

Fig. 4. The `Sample1.final.bam` file results from a *Read Summarisation* step and is followed by a *Variant Calling* step.

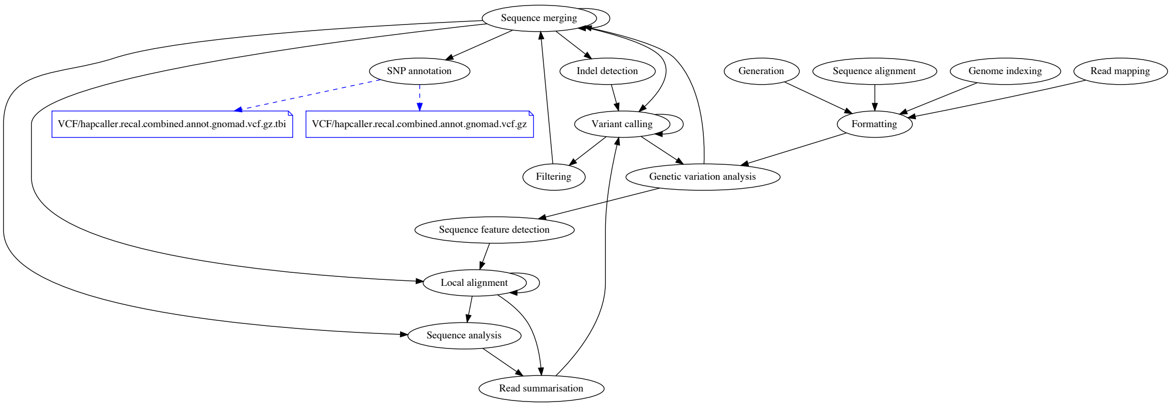

Fig. 5. Human-oriented diagram automatically compiled from the provenance and domain-specific knowledge graph.

These visualisations provide scientists with means to situate an intermediate result, genomic sequences aligned to a reference genome (BAM file), or genomic variants (VCF file) in the context of a complex data analysis process. While an expert bioinformatician won't need these summaries, we consider that expliciting and visualizing these summaries is of major interest to better reuse/repurpose scientific data, or even provide a first level of explanation in terms of domain-specific concepts.

4.3. Machine-oriented data summaries

Linked Data principles advocate the use of controlled vocabularies and ontologies to provide both human- and machine-readable knowledge. We show in Figure 6 how domain-specific statements on data, typically their annotation with EDAM bioinformatics concepts, can be aggregated and shared between machines by leveraging the NanoPublication vocabulary. Published as Linked Data, these data summaries can be semantically indexed and searched, in line with the Findability of FAIR principles.

4.4. Implementation

Provenance capture. We slightly extended the Snakemake [26] workflow engine with a provenance capture module[7]. This module, written in Python, is a wrapper

[7] `https://bitbucket.org/agaignar/snakemake-provenance/src/provenance-capture`

Table 3

Enhancing reuse of processed data with FRESH

What ?	How?	Results
Traceable	Provenance	PROV traces
Interpertable	Scientific and technical Context	Micropublications vocabulary
Understandable	Domain-Specific Context ontologies	EDAM terms
For human	Human-oriented data summaries	Text and diagrams
For machine	Machine-oriented data summaries	NanoPublications

```
[...]
:head {
    _:np1 a np:Nanopublication .
    _:np1 np:hasAssertion :assertion .
    _:np1 np:hasProvenance :provenance .
    _:np1 np:hasPublicationInfo :pubInfo .
}

:assertion {
  <http://snakemake-provenance/Samples/Sample1/
  BAM/Sample1.merged.bai> rdfs:seeAlso
  <http://edamontology.org/operation_3197> .

  <http://snakemake-provenance/VCF/hapcaller.
  indel.recal.filter.vcf.gz> rdfs:seeAlso
  <http://edamontology.org/operation_3695> .
}
[...]
```

Fig. 6. An extract of a machine-oriented NanoPublication aggregating domain-specific assertions, provenance and publication information.

Table 4

Time for producing data summaries

RDF Graph	load time	Text-based	NanoPub.	Graph-based
218 906 triples	22.7s	1.2s	61ms	1.5s

for the AbstractExecutor class. The same source code is used to produce PROV RDF metadata when locally running a workflow, or when exploiting parallelism in an HPC environment, or when simulating a workflow. Simulating a workflow is an interesting feature since all data processing steps are generated by the workflow engine but not concretely executed. Nevertheless, the capture of simulated provenance information is still possible without paying for the generally required long CPU-intensive tasks. This extension is under revision for being integrated in the main development branch of the SnakeMake workflow engine.

Knowledge graph assembly. We developed a Python crawler[8] that consumes the JSON representation of the Bio.Tools bioinformatics registry and produces an RDF data dump focusing on domain annotations (EDAM ontology) and links to the reference papers. RDF dumps are nightly built and pushed to a dedicated source code repository[9].

Experimental setup. The results shown in section 4 were obtained by running a Jupyter Notebook. RDF data loading and SPARQL query execution were achieved through the Python RDFlib library. Python string templates were used to assemble the NanoPublication while NetworkX, PyDot and GraphViz were used for basic graph visualisations.

We simulated the production-level exome-sequencing workflow to evaluate the computational cost of producing data summaries from an RDF knowledge graph. The simulation of the workflow execution allowed to not being impacted by the actual computing cost of performing raw genomic data analysis. Table 4 shows the cost using a 16GB, 2.9GHz Core i5 MacBook Pro desktop computer. We measured 22.7s to load in memory the full knowledge graph (218 906 triples) covering the workflow claims and its provenance graph, the Bio.Tools RDF dump, and the EDAM ontology. The sentence-based data summaries have been obtained in 1.2s, the machine-oriented NanoPublication has been generated in 61ms, and finally 1.5s to reshape and display the graph-based data summary. This overhead can be considered as negligible compared to the computing resources required to analyse exome-sequencing data as shown in section 2.

To reproduce the human- and machine-oriented data summaries, this Jupyter Notebook is available through a source code repository[10]. To go beyond the provided experimental results, and to apply more generally the FRESH approach, the following requirements should be satisfied:

- the overall data analysis process should be formalised into a computational workflow,
- the running workflow management system should be able to dynamically capture generic prove-

[8] https://github.com/bio-tools/biotoolsShim/tree/master/json2rdf

[9] https://github.com/bio-tools/biotoolsRdf

[10] https://github.com/albangaignard/fresh-toolbox

nance metadata as Linked Data, following the PROV-O ontology,

- the run tools should be semantically annotated with domain-specific concepts. These descriptions should be accessible in a machine-actionable registry through a SPARQL endpoint,
- mappings between workflow steps and the identifiers of the semantically annotated tools should be provided in the workflow specification so that provenance traces refer to semantically annotated tools.

4.5. Discussion

The validation we reported has shown that it is possible to generate data summaries that provide valuable information about workflow data. In doing so, we focus on domain-specific annotations to promote the findability and reuse of data processed by scientific workflows with particular attention to genomics workflows. This is justified by the fact that FRESH meets the findability and reusability criteria set up by the FAIR community[11].

Regarding findability, FRESH partly meets requirements F1 ((Meta)data are assigned a globally unique and persistent identifier), F2 (Data are described with rich metadata) and F3 (Metadata clearly and explicitly include the identifier of the data they describe) since (i) we assign Universal Unique Identifiers (UUIDs) to provenance artefacts and (ii) we reuse the NanoPublication framework, and the EDAM bioinformatics ontology to share and reuse intermediate data results based on rich metadata. Although the generated nanopublications are not yet indexed in a searchable resource, they could be published either through a SPARQL endpoint, or through the network of peer NanoPublication servers.

Regarding reusability, Table 3 points out the reusability aspects of FRESH in line with the FAIR community requirements. In particular, we note that FRESH is aligned with R1.2 ((meta)data are associated with detailed provenance) and R1.3. ((meta)data meet domain-relevant community standards). As illustrated in the previous sections, FRESH can be used to generate human-oriented data summaries or machine-oriented data summaries.

Still in the context of genomic data analysis, a typical reuse scenario would consists in exploiting as inputs, the annotated genomic variants (in blue in Figure 4), to conduct a rare variant statistical analysis. If we consider that no semantics is attached to the names of files or tools, domain-agnostic provenance would fail in providing information on the nature of data processing. By looking on the human-oriented diagram, or by letting an algorithm query the machine-oriented nanopublication produced by FRESH, scientists would be able to understand that the file results from an annotation of single nucleotide polymorphisms (SNPs) which was preceded by a variant calling step itself preceded by an insertion/deletion (Indel) detection step.

We focused in this work on the bioinformatics domain and leveraged Bio.Tools, a large-scale community effort aimed at semantically cataloguing available algorithms/tools. As soon as semantic tools catalogues are available for other domains, FRESH can be applied to enhance the findability and reusability of processed data. Even if more recent, similar efforts address the bioimaging community through the setup of the BISE[12] bioimaging search engine (Neubias EU COST Action). Annotated with a bioimaging-specific EDAM extension, this tool registry could be queried to annotate bioimaging data following the same approach.

In our work, we validated our solution by manually inspecting the usefulness of the summaries that are constructed given a real-life workflow. That said, we believe that there is a need for a benchmark that can be utilized by the community to systematically assess and compare the effectiveness of the proposed solutions. We also think that such a benchmark should be the result of a community-led effort to cater for different needs/requirements and different scientific domains.

5. Related Work

Our work is related to proposals that seek to enable and facilitate the reproducibility and reuse of scientific artefacts and findings. We have seen recently the emergence of a number of solutions that assist scientists in the tasks of packaging resources that are necessary for preserving and reproducing their experiments. For example, OBI (Ontology for Biomedical Investigations) [27] and the ISA (Investigation, Study, Assay) model [28] are two widely used community mod-

[11] https://www.go-fair.org/fair-principles

[12] http://www.biii.eu

els from the life science domain for describing experiments and investigations. OBI provides common terms, like investigations or experiments to describe investigations in the biomedical domain. It also allows the use of domain-specific vocabularies or ontologies to characterize experiment factors involved in the investigation. ISA on the other hand structures the descriptions about an investigation into three levels: Investigation, for describing the overall goals and means used in the experiment, study for documenting information about the subject under study and treatments that it may have undergone, and assay for representing the measurements performed on the subjects. Research Objects [29] is a workflow-friendly solution that provides a suite of ontologies that can be used for aggregating workflow specification together with its executions and annotations informing on the scientific hypothesis and other domain annotations. ReproZip [7] is another solution that helps users create relatively lightweight packages that include all the dependencies required to reproduce a workflow for experiments that are executed using scripts, in particular, Python scripts.

The above solutions are useful in that they help scientists package information they have about the experiment into a single container. However, they do not help scientists in actually annotating or reporting the findings of their experiments. In this respect, Alper *et al.* [9] and Gaignard *et al.* [8] developed solutions that provide the users by the means for deriving annotations for workflow results and for summarizing the provenance information provided by the workflow systems. Such summaries are utilized for reporting purposes.

While we recognize that such proposals are of great help to the scientists and can be instrumental to a certain extent to check the repeatability of experiments, they are missing opportunities when it comes to the reuse of the intermediate data products that are generated by their experiments. Indeed, the focus in the reports generated by the scientist is put on their scientific findings, documenting the hypothesis and experiment they used, and in certain cases, the datasets obtained as a result of their experiment. The intermediate datasets, which are by-products of the internal steps of the experiment, are in most cases buried in the provenance of the experiment if not reported at all. The availability of such intermediate datasets can be of value to third-party scientists to run their own experiment. This does not only save time for those scientists in that they can use readily available datasets but also save time and resources since some intermediate datasets are generated using large-scale resource- and compute-intensive scripts or modules.

Of particular interest to our work are the standards developed by the semantic web community for capturing provenance, notably the W3C PROV-O recommendation, and its workflow-oriented extensions, e.g., ProvONE [13], OPMW [14], Wfprov [15] and P-Plan [30]. The availability of provenance provides the means for the scientist to issues queries on *Why* and *How* data were produced. However, it does not necessarily allow the scientists to examine questions such as "Is this data helpful for my computational experiment ?", or "if potentially useful, does this data has enough quality ?". These queries are particularly challenging since the capture of related meta-data is in general domain-dependent and should be automated. This is partly due to the fact that provenance information can be overwhelming (large graphs), and partly because of a lack of domain annotations. In previous work [8], we proposed *PoeM* an approach to generate human-readable experiment reports for scientific workflows based on provenance and users annotations. *SHARP* [31,32] extends *PoeM* for workflows running in different systems and producing heterogeneous PROV traces. In this work, we capitalize in our previous work to annotate and summarize provenance information. In doing so, we focus on Workflow data products re-usability as opposed to the workflow itself. As data re-usability require to meet domain-relevant community standards (R1.3 of FAIR principles). We rely on Bio.tools (https://bio.tools/) registry to discover tools descriptions and automatically generate domain-specific data annotations.

The proposal by Garijo and Gil [10] is perhaps the closest to ours in the sense that it focuses on data (as opposed to the experiment as a whole), and generate textual narratives from provenance information that is human-readable. The key idea of data narratives is to keep detailed provenance records of how an analysis was done, and to automatically generate human-readable description of those records that can be presented to users and ultimately included in papers or reports. The objective that we set out in this paper is different from that by Garijo and Gil in that we do not aim to generate narratives. Instead, we focus on annotating intermediate workflow data. The scientific communi-

[13] https://purl.dataone.org/provone-v1-dev
[14] https://www.opmw.org
[15] http://purl.org/wf4ever/wfprov#

ties have already investigated solutions for summarizing and reusing workflows (see e.g., [33,34]).

The solution proposed by Starlinger et al. [33] aims at identifying similarities between workflows. The authors exploit three sources of information, namely the labels used to describe the modules that compose the workflow, the structure (i.e., dataflow) of the workflow, and authorship information. In doing so, the authors do not tackle the problem that the human user faces when trying to understand a potentially complex workflow. Such a solution can be envisaged when the aim is to effectively search similar workflows in a repository given an initial input workflow. Our objective is different in that we aim to promote the reuse not only of workflows but also of the data products that the execution of such workflows produce, and we do so by leveraging summarisation techniques to produce human-friendly account on the data products.

Cerezo et al. [34] proposed a conceptual workflow model, close to end-user's domain of expertise, aimed at enhancing the sharing and reuse of scientific workflows. These conceptual workflows are conceived at workflow design-time and are then semi-automatically refined into concrete executable workflows through a set of semantic transformations. Although our approach tackles reuse in data-driven sciences, we focus on the reuse of intermediate produced/consumed data whereas Cerezo *et al.* focus on the reuse of the data transformation process itself. In addition, our approach is bottom-up, based on workflow executions, and tends to limit the solicitation of domain experts, by leveraging already running semantically annotated tools catalogues.

It is worth noting that our work is complementary and compatible with the work by Garijo and Gil. In particular, the annotations and provenance summaries generated by the solution we propose can be used to feed the system developed by Garijo and Gil to generate more concise and informative narratives.

Our work is also related to the efforts of the scientific community to create open repositories for the publication of scientific data. For example, Figshare[16] and Dataverse[17], which help academic institutions store, share and manage all of their research outputs. The data summaries that we produce can be published in such repositories. However, we believe that the summaries that we produce are better suited for repositories that publish knowledge graphs, e.g., the one created by the whyis project[18]. This project proposes a nano-scale knowledge graph infrastructure to support domain-aware management and curation of knowledge from different sources.

6. Conclusion and Future Works

In this paper, we proposed FRESH, an approach for making scientific workflow data more findable and reusable, with a focus on genomic workflows. To do so, we utilized data-summaries, which are generated based on provenance and domain-specific ontologies. FRESH comes in two flavors by providing concise human-oriented and machine-oriented data summaries. Experimentation with a production-level exome-sequencing workflow shows the effectiveness of FRESH in terms of time, the overheads of producing human-oriented and machine-oriented data summaries are negligible compared to the computing resources required to analyze exome-sequencing data. FRESH open several perspectives, which we intend to pursue in our future works.

So far, we have focused in FRESH on the findability (F) and reuse (R) of workflow data products. We intend to extend FRESH to cater for the two remaining FAIR criteria (A, I). To do so, we intend to rethink and redefine interoperability and accessibility when dealing with workflow data products and public catalogues, before proposing solutions to cater for them. We then plan to evaluate the effectiveness of FRESH through a user study when it comes to the reuse of genomic data, and its portability to other domains and communities. Finally, we intend to identify means for the incentivization of scientists to (1) provide tools with high quality domain-specific annotations (2) generate and use domain-specific data summaries to promote reuse.

7. Acknowledgements

We thank our colleagues from "l'Institut du Thorax" and CNRGH who provided their insights and expertise regarding the computational costs of large-scale exome-sequencing data analysis.

We are grateful to the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest) for providing technical support and computing resources.

[16] https://figshare.com/

[17] https://dataverse.org/

[18] http://tetherless-world.github.io/whyis/

References

[1] J. Liu, E. Pacitti, P. Valduriez and M. Mattoso, A survey of data-intensive scientific workflow management, *Journal of Grid Computing* **13**(4) (2015), 457–493. doi:10.1007/s10723-015-9329-8.

[2] S.C. Boulakia, K. Belhajjame, O. Collin, J. Chopard, C. Froidevaux, A. Gaignard, K. Hinsen, P. Larmande, Y.L. Bras, F. Lemoine, F. Mareuil, H. Ménager, C. Pradal and C. Blanchet, Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities, *Future Generation Comp. Syst.* **75** (2017), 284–298. doi:10.1016/j.future.2017.01.012. https://doi.org/10.1016/j.future.2017.01.012.

[3] 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 genomical?, *PLoS Biology* **13**(7) (2015), 1–11, ISSN 15457885. ISBN ISBN 1545-7885. doi:10.1371/journal.pbio.1002195.

[4] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth and G.A. McVean, An integrated map of genetic variation from 1,092 human genomes., *Nature* **491**(7422) (2012), 56–65, ISSN 1476-4687. ISBN ISBN 1476-4687 (Electronic)\r0028-0836 (Linking). doi:10.1038/nature11632.An. http://dx.doi.org/10.1038/nature11632.

[5] P. Alper, Towards harnessing computational workflow provenance for experiment reporting, PhD thesis, University of Manchester, UK, 2016. http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.686781.

[6] V. Stodden, The Scientific Method in Practice: Reproducibility in the Computational Sciences, *SSRN Electronic Journal* (2010). doi:10.2139/ssrn.1550193.

[7] F. Chirigati, R. Rampin, D. Shasha and J. Freire, Reprozip: Computational reproducibility with ease, in: *Proceedings of the 2016 International Conference on Management of Data*, ACM, 2016, pp. 2085–2088. doi:10.1145/2882903.2899401.

[8] A. Gaignard, H. Skaf-Molli and A. Bihouée, From scientific workflow patterns to 5-star linked open data, in: *8th USENIX Workshop on the Theory and Practice of Provenance*, 2016.

[9] P. Alper, K. Belhajjame and C.A. Goble, Automatic Versus Manual Provenance Abstractions: Mind the Gap, in: *8th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2016, Washington, D.C., USA, June 8-9, 2016.*, USENIX, 2016. https://www.usenix.org/conference/tapp16/workshop-program/presentation/alper.

[10] Y. Gil and D. Garijo, Towards Automating Data Narratives, in: *Proceedings of the 22Nd International Conference on Intelligent User Interfaces*, IUI '17, ACM, New York, NY, USA, 2017, pp. 565–576. ISBN ISBN 978-1-4503-4348-0. doi:10.1145/3025171.3025193. http://doi.acm.org/10.1145/3025171.3025193.

[11] W. et al., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3** (2016). doi:10.1038/sdata.2016.18.

[12] J. Ison, K. Rapacki, H. Ménager, M. Kalaš, E. Rydza, P. Chmura, C. Anthon, N. Beard, K. Berka, D. Bolser, T. Booth, A. Bretaudeau, J. Brezovsky, R. Casadio, G. Cesareni, F. Coppens, M. Cornell, G. Cuccuru, K. Davidsen, G. Della Vedova, T. Dogan, O. Doppelt-Azeroual, L. Emery, E. Gasteiger, T. Gatter, T. Goldberg, M. Grosjean, B. Gruüing, M. Helmer-Citterich, H. Ienasescu, V. Ioannidis, M.C. Jespersen, R. Jimenez, N. Juty, P. Juvan, M. Koch, C. Laibe, J.W. Li, L. Licata, F. Mareuil, I. Mičetić, R.M. Friborg, S. Moretti, C. Morris, S. Möller, A. Nenadic, H. Peterson, G. Profiti, P. Rice, P. Romano, P. Roncaglia, R. Saidi, A. Schafferhans, V. Schwämmle, C. Smith, M.M. Sperotto, H. Stockinger, R.S. Varěková, S.C.E. Tosatto, V. De La Torre, P. Uva, A. Via, G. Yachdav, F. Zambelli, G. Vriend, B. Rost, H. Parkinson, P. Løngreen and S. Brunak, Tools and data services registry: A community effort to document bioinformatics resources, *Nucleic Acids Research* (2016), ISSN 13624962. ISBN ISBN 13624962 (Electronic). doi:10.1093/nar/gkv1116.

[13] H. Li and R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics* (2010), ISSN 13674803. ISBN ISBN 1367-4811 (Electronic)\r1367-4803 (Linking). doi:10.1093/bioinformatics/btp698.

[14] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* (2009), ISSN 13674803. ISBN ISBN 1367-4803\r1460-2059. doi:10.1093/bioinformatics/btp352.

[15] Broad Institute, Picard tools, 2016, ISSN 1949-2553. ISBN 8438484395. doi:10.1007/s00586-004-0822-1.

[16] C. Alkan, B.P. Coe and E.E. Eichler, GATK toolkit, *Nature reviews. Genetics* (2011), ISSN 1471-0064. ISBN ISBN 1471-0064 (Electronic)\n1471-0056 (Linking). doi:10.1038/nrg2958.

[17] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek and F. Cunningham, The Ensembl Variant Effect Predictor, *Genome Biology* (2016), ISSN 1474760X. ISBN ISBN 1474760X (Electronic). doi:10.1186/s13059-016-0974-4.

[18] S.T. Sherry, dbSNP: the NCBI database of genetic variation, *Nucleic Acids Research* (2001), ISSN 13624962. ISBN ISBN 1362-4962 (Electronic)\r0305-1048 (Linking). doi:10.1093/nar/29.1.308.

[19] N.D. Olson, S.P. Lund, R.E. Colman, J.T. Foster, J.W. Sahl, J.M. Schupp, P. Keim, J.B. Morrow, M.L. Salit and J.M. Zook, Best practices for evaluating single nucleotide variant calling methods for microbial genomics, 2015, ISSN 16648021. ISBN 1664-8021 (Electronic)\r1664-8021 (Linking). doi:10.3389/fgene.2015.00235.

[20] M.D.e.a. Wilkinson, A design framework and exemplar metrics for FAIRness, *Scientific Data* **5** (2018). doi:10.1038/sdata.2018.118.

[21] C. Bizer, M.-E. Vidal and H. Skaf-Molli, Linked Open Data, in: *Encyclopedia of Database Systems*, L. Liu and M.T. Özsu, eds, Springer, 2017. https://link.springer.com/referenceworkentry/10.1007/978-1-4899-7993-3_80603-2.

[22] C. Bizer, T. Heath and T. Berners-Lee, Linked Data - The Story So Far, *Int. J. Semantic Web Inf. Syst.* **5**(3) (2009), 1–22. doi:10.4018/jswis.2009081901. https://doi.org/10.4018/jswis.2009081901.

[23] M. Corpas, N. Kovalevskaya, A. McMurray and F. Nielsen, A FAIR guide for data providers to maximise sharing of human genomic data, *PLoS Computational Biology* **14**(3) (2018). doi:10.1371/journal.pcbi.1005873.

[24] M.D.e.a. Wilkinson, FAIRMetrics/Metrics: Proposed FAIR Metrics and results of the Metrics evaluation questionnaire,

2018. doi:Zenodo https://doi.org/10.5281/zenodo.1065973.

[25] 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* (2014), ISSN 20411480. ISBN ISBN 2041-1480 (Linking). doi:10.1186/2041-1480-5-28.

[26] J. Köster and S. Rahmann, Snakemake—a scalable bioinformatics workflow engine, *Bioinformatics* **28**(19) (2012), 2520–2522. doi:10.1093/bioinformatics/bty350.

[27] R.R. Brinkman, M. Courtot, D. Derom, J.M. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra et al., Modeling biomedical experimental processes with OBI, in: *Journal of biomedical semantics*, Vol. 1, BioMed Central, 2010, p. 7. doi:10.1186/2041-1480-1-S1-S7.

[28] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann et al., ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level, *Bioinformatics* **26**(18) (2010), 2354–2356. doi:10.1093/bioinformatics/btq415.

[29] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K.M. Hettne, R. Palma, E. Mina, Ó. Corcho, J.M. Gómez-Pérez, S. Bechhofer, G. Klyne and C.A. Goble, Using a suite of ontologies for preserving workflow-centric research objects, *J. Web Semant.* **32** (2015), 16–42. doi:10.1016/j.websem.2015.01.003. https://doi.org/10.1016/j.websem.2015.01.003.

[30] D. Garijo and Y. Gil, Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data, in: *Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, Boston, MA, USA, November 12, 2012*, CEUR-WS.org, 2012. http://ceur-ws.org/Vol-951/paper6.pdf.

[31] A. Gaignard, K. Belhajjame and H. Skaf-Molli, SHARP: Harmonizing and Bridging Cross-Workflow Provenance, in: *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia,Revised Selected Papers*, 2017, pp. 219–234. doi:10.1007/978-3-319-70407-4_35.

[32] A. Gaignard, K. Belhajjame and H. Skaf-Molli, SHARP: Harmonizing Cross-workflow Provenance, in: *Proceedings of the Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics co-located with 14th Extended Semantic Web Conference, SeWeBMeDA@ESWC 2017, Portoroz, Slovenia.*, 2017, pp. 50–64. http://ceur-ws.org/Vol-1948/paper5.pdf.

[33] J. Starlinger, S. Cohen-Boulakia and U. Leser, (Re)use in public scientific workflow repositories, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, ISSN 03029743. ISBN ISBN 9783642312342. doi:10.1007/9783-642312359_24.

[34] N. Cerezo and J. Montagnat, Scientific Workflow Reuse Through Conceptual Workflows on the Virtual Imaging Platform, in: *Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science*, WORKS '11, ACM, New York, NY, USA, 2011, pp. 1–10. ISBN ISBN 978-1-4503-1100-7. doi:10.1145/2110497.2110499. http://doi.acm.org/10.1145/2110497.2110499.