Janus: from Workflows to Semantic Provenance and Linked Open Data

Paolo Missier[1], Satya S. Sahoo[3]*, Jun Zhao[2] **, Carole Goble[1], and Amit Sheth[3]

[1] School of Computer Science, University of Manchester, UK
{pmissier,carole}@cs.man.ac.uk
[2] Department of Zoology, University of Oxford, UK
jun.zhao@zoo.ox.ac.uk
[3] The Kno.e.sis Center, Wright State University, Dayton, OH, USA
{sahoo.2,amit.sheth@wright.edu}

Abstract. Data provenance graphs are form of metadata that can be used to establish a variety of properties of data products that undergo sequences of transformations, typically specified as workflows. Their usefulness for answering user provenance queries is limited, however, unless the graphs are enhanced with domain-specific annotations. In this paper we propose a model and architecture for semantic, domain-aware provenance, and demonstrate its usefulness in answering typical user queries. Furthermore, we discuss the additional benefits and the technical implications of publishing provenance graphs as a form of Linked Data. A prototype implementation of the model is available for data produced by the Taverna workflow system.

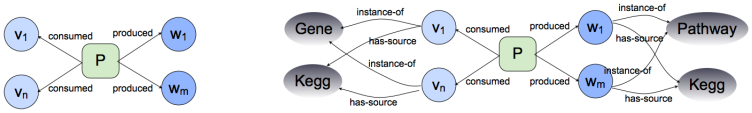
1 Introduction

Experimental science increasingly relies upon computational techniques and large-scale data management to achieve its goals. As with any experimental method, either manual or automated, an important step of the scientific process is the validation of its results. In the case of automated, high-throughput data generation and transformation pipelines, implemented for example as workflows, the complexity of the processes and the volumes of data call for validation procedures to be automated, too. One of the prominent approaches involves the analysis of detailed traces of the data transformations that are recorded during the execution of the data pipeline. These traces are a form of metadata, relative to the data involved in the process, known as data provenance. The growing realisation of the importance of this type of metadata for experimental science has in recent years spurred a wealth of research in provenance acquisition and analysis [17, 1, 5, 7].

Provenance metadata is structured as a causal graph amongst data elements as they undergo several transformations through some composition of processes.

* This work was partly funded by NIH RO1 Grant# 1R01HL087795-01A1
** This work was partly funded by EPSRC Grant# EP/G049327/1

The two main strains of research in this area concentrate on (i) provenance modelling, with the goal of supporting the users' data validation tasks; and (ii) data architectures for provenance management. The work presented in this paper falls in the former of these two categories. Most of the provenance models proposed so far, including those just cited, have been focusing on describing the causal relationships amongst data products, without specific concern for the semantic characterisation of those products. We refer to these graphs as *domain-agnostic*, as they do not include any reference to domain-specific terms. In contrast, we propose a new semantic model of provenance, embodied by *domain-aware* graphs, designed to support data derivation questions that are formulated by user-scientists using domain-specific terminology. Fig. 1 clarifies the distinction between the two types of graphs[4]. The main differences between Fig. 1(a) and Fig. 1(b) are the additional semantic annotations shown in the latter. In this limited example, these are of the form V `instance-of` C or V `has-source` C', where V is a value, and C, C' are terms in some domain vocabulary, for biology concepts and biological database resources, respectively. We expect that, regardless of the specific formalism chosen to specify these annotations, domain-aware graphs be useful to answer a broader class of user questions than their domain-agnostic counterparts (namely those questions that rely upon domain terms).



(a) Domain-agnostic provenance graph

(b) Domain-aware provenance graph

Fig. 1. Adding simple annotations to provenance graphs

Taking this idea further, we also note that grounding a provenance model in the Semantic Web framework presents additionally opportunities for supporting an even broader class of user questions. In particular, we explore the idea of making semantic provenance graphs a part of the broad Web of Data, an increasingly rich source of interconnected data that is uniformly represented according to the principles and conventions of Linked Open Data (LOD) [4]. In practice, we show how mapping data elements in the graph to equivalent data that is published elsewhere in the Web of Data, makes it possible for queries to retrieve properties of data, which are not explicitly represented in the provenance graph or its annotations, but are instead associated with their equivalent external representations.

[4] We use an abstract notation that is close to the one adopted in the Open Provenance Model http://www.openprovenance.org, where data values (the circles) are either produced or consumed by processes (the squares).

1.1 Paper scope and contributions

The idea of semantic provenance was first proposed in [15], but few concrete examples exist to date of its realisation beyond, for example, [6]. In this paper we take a concrete step towards the implementation of a semantic provenance model, code-named *Janus*, cast specifically in the context of provenance for data processed by Life Sciences workflows. We describe a practical implementation of *Janus*[5] , which is grounded in the Taverna workflow model [10] and (domain-agnostic) provenance model [12], and demonstrate its technical feasibility as well as its benefits to users in terms of enhanced query answering capabilities. The paper offers the following specific contributions. Firstly, we set *Janus* in the Semantic Web framework, where we define its model as an extension of the *Provenir* upper ontology for workflow-based data provenance [16]. In this setting, *Janus* consists of a domain-agnostic part, which models essentially the same entities as the existing Taverna provenance model, and a domain-aware part, obtained by extending the ontology to include properties and classes like those shown earlier in Fig 1(b). Secondly, we describe the prototype implementation of an extension to the current Taverna provenance architecture, which produces semantic, RDF-based provenance graphs for workflow runs, that conform to the *Janus* ontology.

Thirdly, we show how the RDF provenance graph can be domain-enhanced by associating semantic types from a variety of public ontologies to some of its elements. We also discuss how existing semantic annotations on the workflow and its composing services, when available, can be automatically propagated to the provenance graph. We then show how, in this setting, we can answer a class of user queries that predicate on the domain annotations. Finally, we show on a practical example how the provenance graph can "blend in" as part of the Web of Data, and exemplify our approach by mapping data identifiers in the graph to those in the Bio2RDF project [2], resulting in extended semantic provenance queries.

1.2 Related work

While provenance data model is a well studied topic [17], the challenge of associating domain semantics to it has received relatively little attention. The Open Provenance Model (OPM) [13] provides the annotation framework to support the need for adding extra information to provenance entities. However, this framework is not defined in the current OPM OWL ontology [6]. Previous work by Cao et al. [6] and Zhao et al. [18] experimented with providing semantic annotations to provenance logs by post-processing, but without a clear data model for accommodating domain-semantics. Such a data model is essential for building a domain-aware provenance collection architecture that could scale beyond case studies. In this paper, we extend the Provenir ontology to create the domain-aware Janus provenance model to address the challenge.

Query frameworks and user-facing visualizations to support a user-oriented view of provenance can be found in the work by Biton et al. [3] and Howe et

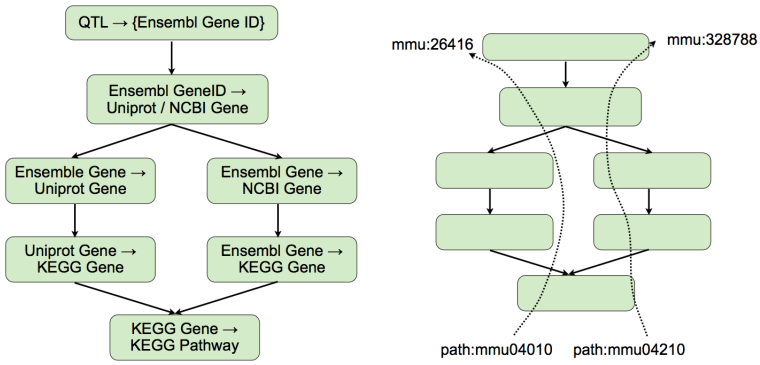
[5] *Janus* is publicly available at: http://purl.org/net/taverna/janus#

[6] http://openprovenance.org/model/opm.owl

al. [9]. Provenance queries that present information in a more meaningful way to the domain scientists have been implemented by Cao et al. [6] and McGuinness et al. [11]. This work takes it further by connecting domain-enhanced provenance graphs created locally with the global Web of Data in order to expand the possible semantic provenance queries.

2 A concrete example

Our running example consists of a bioinformatics workflow designed to find all known relationships between a specific region in the mouse genome, known as a QTL (Quantitative Trait Loci), and the metabolic pathways involving genes that are present in that region. A schematic representation of the workflow is given in Fig. 2(a).[7] The workflow starts by retrieving all the genes known to the Ensembl public database for a given input region, using the Biomart service. It then retrieves all metabolic pathways from the KEGG pathways database, such that at least one of those genes are involved.[8] Note that the schematic representation does not include the many adapter scripts that are required in reality to accomplish this composite task.



(a) Schematic representation of a Taverna workflow

(b) Schematic representation of a provenance graph for a workflow run

Fig. 2.

A scientist may want to ask a number of high-level questions regarding the relationship between the outputs and some of the inputs of a workflow execution ("run"). Amongst these, we are going to consider the following two, which can be expressed in terms of queries on a provenance graph:

[7] The actual workflow, too large to be reproduced here, can be found on the myExperiment Web site: http://www.myexperiment.org/workflows/931.

[8] Ensembl: www.ensembl.org,
Biomart: www.biomart.org/,
KEGG: www.genome.jp/kegg/pathway.html

1. for each *Kegg pathway* observed in the workflow output (or for a specific one), find all *genes* that are within the input *QTL* and are involved in that pathway;
2. amongst all *genes* that are known to *perform a certain biological function*, list those that are involved in a certain *pathway*.

The terms in italics refer to concepts in the bioinformatics domain, similar to those in Fig. 1(b). Intuitively, one can answer (1) for a particular run, by traversing a domain-aware provenance graph for that run, like the one sketched in Fig. 2(b). An output value o for the workflow depends on some input or intermediate value i, if and only if there is a path from i to o in the graph. Thus, (1) can be reduced to a query that finds all pairs (i, o) such that o is of type *pathway*, i is of type *gene*, and there is a path from i to o. In Sec. 3.3 we show how our proposed semantic provenance framework supports this query.

The graph, however, is not sufficient to answer question (2), which refers to the *biological function* of a gene, a concept that is not included in the semantic annotations. Our approach in this case is based upon the idea that the genes that appear in the graph may also be published elsewhere in the broad Web of Data, where the missing annotations can potentially be found. When this is the case, one can formulate a hybrid query that (i) retrieves the biological functions of all the *genes* that appear in the graph, using a Linked Data query, and (ii) for those genes that satisfy the condition, find all paths to the corresponding *pathways* in the graph[9]. We elaborate on this strategy and on its limitations in Sec. 4, showing in particular how that the gene IDs in the graph can be mapped to Bio2RDF genes.

3 The *Janus* Semantic Provenance Infrastructure

The examples from the previous section highlight the need for incorporating domain semantics as part of the provenance model, to bridge the gap between the domain-agnostic provenance produced during workflow execution, and the users' domain-oriented view of provenance. An expressive provenance model with well-defined formal semantics not only enables complex domain-specific information to be modeled, but also facilitates provenance interoperability and supports reasoning over large sets of provenance information. As mentioned in the introduction, formally *Janus* is an extension of *Provenir*, an upper-level reference OWL DL ontology for provenance modeling designed to be extended to represent provenance in multiple domains. In turn, Provenir extends concepts from the well-known Basic Formal Ontology (BFO)[10] to define a set of provenance terms, including the three fundamental concepts of *data*, *process*, and *agent*. Provenir also defines a set of 11 *named relationships* amongst classes, including partonomy relations, temporal information, precedence, and causal relationships, providing

[9] Note however, that there is no guarantee that the gene will be found in the Web of Data, or that the condition on its external annotations can be evaluated there.

[10] http://ontology.buffalo.edu/bfo/

a foundation for the semantic modelling of provenance. As an upper-level reference model for provenance, Provenir ensures a common modeling approach, conceptual clarity of provenance terms, and use of design patterns for consistent provenance modeling.

3.1 Modeling Domain-agnostic Provenance in *Janus*

The Taverna provenance model defined in [12] includes both a static and a dynamic portion. The static portion describes the graph structure of a workflow specification (processors, processor ports, and data dependencies as links between ports), such as the one in our running example of Fig. 2(a). The dynamic portion accounts for multiple invocations of a processor that occur during workflow execution, as well as for the binding of actual values to the processors ports. In the first step of the design, we model the existing Taverna provenance model as an OWL ontology. As illustrated in Fig. 3, the classes in the static portion (`janus:workflow_spec`, `janus:processor_spec`, and `class janus:port`) extend corresponding *Provenir* classes and are associated through appropriate properties, for example `janus:processor_spec provenir:has_parameter janus:port`. Note that data links in the workflow are modelled using the `link_from` property from `port` onto itself. Individuals in these classes include the work-

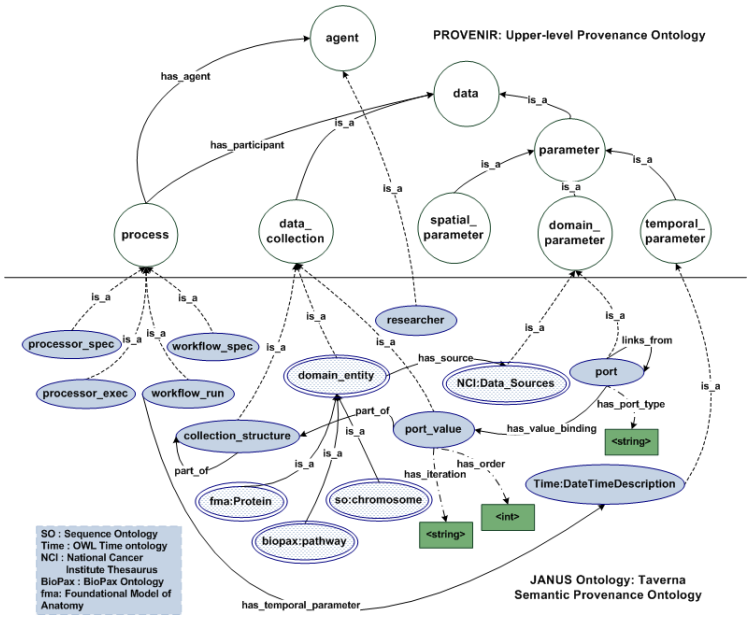


Fig. 3. Domain-aware *Janus* as an extension of *Provenir*

flow itself (`gene_pathway_workflow`), its processors (eg. `genes_in_qtl`), and the

processors' ports (eg. `qtl_end_position`, `chromosome_name`). In turn, these individuals may be related to one or more run-time counterparts in the dynamic portion of the ontology through object property `has_execution`:

$$dom(\texttt{has\_execution}) = \texttt{workflow\_spec or processor\_spec}$$
$$range(\texttt{has\_execution}) = \texttt{workflow\_exec or processor\_exec}$$

and `has_value_binding`:

$$dom(\texttt{has\_value\_binding}) = \texttt{port}, \quad range(\texttt{has\_value\_binding}) = \texttt{port\_value}$$

3.2 Modeling Semantic Provenance in *Janus*

We now describe the *Janus* extension to include domain-specific terms. A variety of scientific communities are creating ontologies to model domain knowledge, for example the National Center for Biomedical Ontologies (NCBO)[11] currently lists 166 publicly available ontologies in the Life Sciences domain. To model semantic provenance in *Janus*, we re-use the classes defined in four public ontologies listed at NCBO, namely the BioPAX, National Cancer Institute Thesaurus, Foundational Model of Anatomy (FMA), and the Sequence ontologies, while the fifth ontology, OWL Time[12] is available from the W3C. This reuse strategy facilitates the interoperability of *Janus*-conformant provenance graphs with large public datasets. For example, these graph can be easily linked to the KEGG, Reactome, and BioCyc databases, which currently make their biological pathway datasets available as BioPAX-conformant RDF datasets.

These extensions are used to annotate both workflow processors and their ports. For example, the three input ports for our example workflow: `chromosome_name`, `start_position` and `end_position`, are annotated with concepts `so:chromosome` and `so:base_pair`, respectively, where the `so` prefix denotes the NCBO-listed Sequence Ontology. Similarly, ports that denote proteins and pathways are annotated using terms `fma:protein` and `biopax:pathway`, from the FMA and the BioPax ontology, respectively. In general, semantic types are associated to ports in an extensible way through the generic `has_value_type` property, according to the following pattern:

$$dom(\texttt{has\_value\_type}) = \texttt{port}, \quad range(\texttt{has\_value\_type}) = \texttt{domain\_entity}$$
`BioPax:pathway owl:subClassOf domain_entity`
`FMA:protein owl:subClassOf domain_entity`

For each workflow run, the Taverna provenance component produces domain-agnostic provenance in the form of an RDF graph that conforms to the *Janus* ontology just described, i.e., it contains RDF statements of the form N `rdf:type` C, where N is a node in the provenance graph and C is some *Janus* concept. The semantic annotation of these graphs assumes that the workflow specification is itself semantically annotated, and it involves automatically

[11] http://www.bioontology.org/
[12] http://www.w3.org/TR/owl-time

propagating those annotations, first to the static portion of the provenance graph, and then to the dynamic portion. Statically annotating the workflows prior to their execution is a realistic proposition. While this may involve a manual curation process, typical workflows never include more than a handful of services, and furthermore, in the long run one can assume that these annotations will be available through a registry that describes the services that compose the workflow (Taverna workflows essentially specify Web service compositions). The BioCatalogue registry for Life Science services[13], for example, is set out to provide semantic annotations for hundreds of services, and these annotations carry over to the workflows where the services are invoked.

The propagation of workflow annotations to the provenance graph is fairly straightforward. Firstly, consider a static workflow element, say port $X =$ `chromosome_name`, annotated with concept $C =$ `so:chromosome` in the workflow (using any available formalism). In the provenance graph this is expressed using the pattern:

X `rdf:type` `Port`, $C = \{c\}$, X `has_value_type` c

for example:

`chromosome_name rdf:type Port`, `so:chromosome` $= \{$`singleton_chromosome`$\}$

`chromosome_name has_value_type singleton_chromosome`

Secondly, the annotations on a port carry over to each of the values that are bound to that port[14], using a collection of inference rules like the following:

$$\frac{X\ \texttt{rdf:type}\ \texttt{Port} \quad C = \{c\} \quad X\ \texttt{has\_value\_type}\ c \quad X\ \texttt{has\_value}\ v \quad v\ \texttt{rdf:type}\ \texttt{PortValue}}{v\ \texttt{rdf:type}\ C}$$

This rule asserts the value v as an individual of the *Janus* class C. The set of rules accounts for various annotations, for example the following rule:

$$\frac{X\ \texttt{rdf:type}\ \texttt{Port} \quad X\ \texttt{has\_source}\ S \quad X\ \texttt{has\_value}\ v \quad v\ \texttt{rdf:type}\ \texttt{PortValue}}{v\ \texttt{has\_source}\ S}$$

annotates v with the data source of the port (for instance, the KEGG database).

As a proof of concept, the *Janus* ontology currently models the semantic provenance terms that are adequate for representing the domain semantics of our example workflow, using less than 30 classes and properties with a DL expressivity of *ALCH(D)*. Many of the classes, for example to model collection data structures, have not been described as they are less relevant to our discussion in this paper. In the future, we plan to extend *Janus* with the domain terms used to annotate the default set of services in the Taverna release version.

[13] http://www.biocatalogue.org

[14] This assumes that the ports are strongly typed, i.e., that all values bound to the port have the same type as the port.

3.3 Provenance Query Infrastructure for *Janus*

We now describe the *Janus* query infrastructure that has been implemented to support the example provenance queries discussed in Sec. 2. The query infrastructure is implemented using the open source Jena ARQ tool[15], and supports provenance queries expressed in the SPARQL query language [14]. We composed the SPARQL query pattern corresponding to the example query (1) from Sec. 2: "Find all the QTL genes that are involved in KEGG pathways". The SPARQL query pattern first identifies port values that are individuals of class `biopax:pathway` and are linked to values "KEGG", which are themselves individuals of class `NCI:Data_Sources`, through property `has_source`. In the next step, the query pattern traverses the property `has_value_binding` between a `port` and a `port_value`, followed by traversal of the property `links_from` between individuals of class `port`, until it reaches individuals of class `so:base_pair` that represent the result QTL genes (the second provenance query proposed in Sec. 2: "Find pathways that contain genes with specific functions," is discussed in the next section).

Provenance queries typically involve a recursive traversal of the graph to compute a transitive closure, namely over the `links_from` property. We had two options for implementing the transitive closure function, namely a function that is tightly coupled to the RDF data store implementation, or a generic module that can be used with any RDF data store. We chose a generic implementation using the SPARQL ASK function, which allows the provenance query infrastructure to be used over multiple RDF stores. The SPARQL ASK function allows "application to test whether or not a query pattern has a solution," [14] without returning a result set or graph. The transitive closure functions starts with the port instance linked to the input value and then recursively expands the SPARQL query expression using the ASK function until a *false* value is returned. The SPARQL ASK function, in contrast to the SELECT and CONSTRUCT functions, does not bind the results of the query to variables in the query pattern, and is therefore a low-overhead function for computing transitive closures.

4 Taverna provenance and Linked Data

So far we have shown how the domain-aware extensions to *Janus* enable answering domain-specific semantic provenance queries. In this section we describe how we can, in addition, also use these semantic annotations to link *Janus*-compliant provenance graphs to the open Web of Data in order to expand the range of supported domain provenance queries.

4.1 Publishing Taverna Provenance as Linked Data

Because *Janus* provenance is already available as RDF graphs, we only need to make these graphs Linked data-compliant and accessible on the Web. This means that 1) each Janus entity URI should be derefenceable, and 2) wherever possible, the data URIs under the *Janus* namespace should be mapped to other linked

[15] http://jena.sourceforge.net/ARQ/

data URIs on the Web. We use existing Linked Data publication tools, namely Pubby [16], to implement the first step. In order to connect *Janus* graphs with LOD we create `rdfs:seeAlso` links between *Janus* data URIs and Bio2RDF [2] data URIs. We use Bio2RDF data URIs because Bio2RDF is one of the earliest linked datasets and it is regarded as a nucleus of the Life Science datasets. Using the semantic annotations associated with *Janus* provenance, we define a set of rules for the identity mapping. Given a *Janus* data item d_i with value $value(d_i)$, its mapping Bio2RDF URI $URI(d_i)$ is determined by the type of d_i and the data source where d_i comes from, according to the following rules:

- IF `isType(`d_i`) == Gene` AND `isSource(`d_i`) == Entrez` THEN
  - URI(d_i) = http://bio2rdf.org/geneid: + value(d_i)
- IF `isType(`d_i`) == Gene` AND `isSource(`d_i`) == UniProt` THEN
  - URI(d_i) = http://bio2rdf.org/uniprot: + value(d_i)
- IF `isType(`d_i`) == Gene` AND `isSource(`d_i`) == KEGG` THEN
  - URI(d_i) = http://bio2rdf.org/kegg: + value(d_i)
- IF `isType(`d_i`) == Pathway` AND `isSource(`d_i`) == KEGG` THEN
  - URI(d_i) = http://bio2rdf.org/path: + value(d_i)

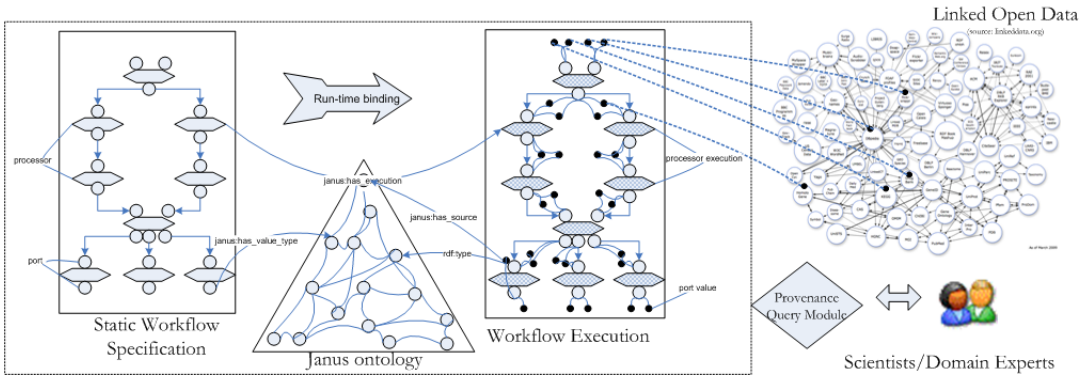


Fig. 4. Semantic provenance for Taverna in the Linked Data context

4.2 Consuming Taverna Provenance as Linked Data

As mentioned, creating Janus Linked Data provenance that is connected to Bio2RDF makes the provenance graphs an integral part of the Web of Life Science data (see Figure 4). This opens the provenance graph to queries that run on the Web of Data. Furthermore, provenance graphs that are created during different workflow runs are now indirectly, and automatically connected through their common external data URIs, thus supporting queries that span across multiple runs.

To demonstrate this, we show how we can support a semantic provenance query that requires access to both *Janus* and the various Bio2RDF repositories, by executing a single SPARQL query against SQUIN [8], a Linked Data query

[16] http://www4.wiwiss.fu-berlin.de/pubby/

engine. Instead of having to write separate SPARQL queries against each individual data source, SQUIN allows us to treat the whole Web of Data as one single data space. It is a query engine that applies the "follow your nose" principle of Linked Data: it traverses the whole Web of Data to retrieve all relevant data sources for a query by taking the URIs in the query and those in the intermediate results, following links of these URIs to other data sources, and applying the querying graph pattern to the intermediate result space in order to obtain relevant results.

Our example query below searches for the functions of those proteins encoded by the Entrez genes that were generated by the executions of the example workflow in Figure 2(b). The domain knowledge about the genes is drawn from two Bio2RDF data repositories and the knowledge about which Taverna data products are Entrez genes comes of the domain-enhanced *Janus* provenance. This simple SPARQL query needs access to at least three linked datasets. SQUIN query engine allows us to write one single query against these multiple data sources. The result will return the biological process related to the data products from any workflow runs that are Entrez genes. We can then use semantic provenance queries similar to the one presented in Sec. 3.3 to search for KEGG pathways that contain these genes.

```
PREFIX uniprot: <http://purl.uniprot.org/core/>
PREFIX ex: <http://purl.org/net/taverna/janus/>
PREFIX : <http://purl.org/net/taverna/janus#>
SELECT distinct ?entrezgene ?function
WHERE {
ex:dataproudct1 rdf:type <http://purl.org/obo/owl/sequence#gene> .
ex:dataproudct1 :has_source :entrez_gene .
ex:dataproudct1 rdfs:seeAlso ?entrezgene .
?entrezgene <http://bio2rdf.org/bio2rdf_resource:xPath> ?protein .
?protein  uniprot:classifiedWith ?function.
}
```

This example shows that drawing on the domain knowledge from the Linked Data cloud enables us to extend the kind of domain-level provenance queries that we can implement that are more meaningful to the scientists. Finding specific KEGG pathways that are related to genes of interesting functions will help scientists quickly identify potential pathways from hundreds of experiment results. The above example query could enable scientists to quickly identify the presence of pathways that consistently exist in different experimentations, including those that were conducted by the scientists themselves.

5 Conclusions and further work

We have presented a semantic provenance model for workflow data, called *Janus*, and a prototype implementation for the Taverna workflow system and provenance model. The implementation demonstrates the benefits of collecting semantic provenance, by showing exemplars semantic provenance queries that can now be answered by the system.

The main objection that is often raised in connection with semantic annotations, is the annotation cost. We have noted, in Sec. 3, that the annotation effort is actually limited to the workflow specification, and indeed, possibly just to the services used in the workflow, when those are annotated once and for all as part of the service registry curation process. In turn, this observation provides additional motivation for the development of registries like Biocatalogue.

Our investigation into the idea of publishing provenance graphs as Linked Data is still preliminary and requires additional insight. For instance, the simple rules used to link *Janus* provenance with the Web of Data do not consider the possibility that the workflow and Bio2RDF refer to different copies of the same database. Also, some of the mapping Bio2RDF URIs might not exist at all or are actually linked to mismatching data entities, and the precision of the mapping between *Janus* and Bio2RDF data URIs needs to be evaluated. Finally, we plan to conduct a user assessment as a way to establish the perceived value of semantic provenance from the users' perspective.

References

1. R S Barga and L A Digiampietri. Automatic capture and efficient storage of e-Science experiment provenance. *Concurrency and Computation: Practice and Experience*, 20:419–429, 2008.
2. F. Belleau, M.A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems. *Journal of Biomedical Informatics*, 41:706–716, 2008.
3. O Biton, S Cohen Boulakia, and S B Davidson. Zoom*UserViews: Querying Relevant Provenance in Workflow Systems. In *VLDB*, pages 1366–1369, 2007.
4. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *Int. Journal on Semantic Web and Information Systems, Special Issue on Linked Data*, 2009. in press.
5. S Bowers, T M McPhillips, and B Ludäscher. Provenance in collection-oriented scientific workflows. *Concurrency and Computation: Practice and Experience*, 20:519–529, 2008.
6. B. Cao, B. Plale, G. Subramanian, P. Missier, C. Goble, and Y. Simmhan. Semantically Annotated Provenance in the Life Science Grid. In Juliana Freire, Paolo Missier, and Satya S. Sahoo, editors, *1st International Workshop on the Role of Semantic Web in Provenance Management. CEUR Proceedings*, 2009.
7. Susan B Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD Conference*, pages 1345–1350, 2008.
8. O. Hartig, C. Bizer, and J.C. Freytag. Executing SPARQL queries over the web of linked data. In *Procs ISWC*, pages 293–309, Washington D.C., USA, 2009.
9. B. Howe, P. Lawson, R. Bellinger, E. Anderson, E. Santos, J. Freire, C. Scheidegger, A. Baptista, and C. Silva. End-to-end escience: Integrating workflow, query, visualization, and provenance at an ocean observatory. In *Procs Fourth IEEE International Conference on eScience*, pages 127–134, 2008.
10. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M.R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34:729–732, 2006.

11. D. L. McGuinness, P. Fox, P. Pinheiro da Silva, S. Zednik, N. Del Rio, Li Ding, P. West, and C. Chang. Annotating and embedding provenance in science data repositories to enable next generation science applications. In *American Geophysical Union, Fall Meeting (AGU2008), Eos Trans. AGU, 89(53), Fall Meet. Suppl., Abstract IN11C-1052*, 2008.
12. P. Missier, N.W. Paton, and K. Belhajjame. Fine-grained and efficient lineage querying of collection-based workflow provenance. In *Procs. of EDBT*, Lausanne, Switzerland, 2010.
13. Luc Moreau. *The Open Provenance Model v 1.1.* 2009.
14. E Prud'ommeaux Seaborne, A. SPARQL Query Language for RDF. *W3C Recommendation*, 2008.
15. S S Sahoo, A Sheth, and C Henson. Semantic provenance for eScience: Managing the deluge of scientific data. *IEEE Internet Computing*, 12:46–54, 2008.
16. S S Sahoo Sheth, A. Provenir ontology: Towards a Framework for eScience Provenance Management, 2009.
17. Y Simmhan, B Plale, and D Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34:31–36, 2005.
18. J Zhao, C Wroe, C Goble, R Stevens, D Quan, and M Greenwood. Using Semantic Web Technologies for Representing e-Science Provenance. In *Third International Semantic Web Conference (ISWC2004)*, LNCS, pages 92–106, Hiroshima, Japan, November 2004. Springer-Verlag.