



## RDFProv: A relational RDF store for querying and managing scientific workflow provenance

Artem Chebotko <sup>a,\*</sup>, Shiyong Lu <sup>b</sup>, Xubo Fei <sup>b</sup>, Farshad Fotouhi <sup>b</sup>

<sup>a</sup> Department of Computer Science, University of Texas-Pan American, 1201 West University Drive, Edinburg, TX 78539, USA

<sup>b</sup> Department of Computer Science, Wayne State University, 431 State Hall, 5143 Cass Avenue, Detroit, MI 48202, USA

### ARTICLE INFO

#### Article history:

Received 12 October 2008

Received in revised form 8 March 2010

Accepted 11 March 2010

Available online 23 March 2010

#### Keywords:

Provenance

Scientific workflow

Metadata management

Ontology

RDF

OWL

SPARQL-to-SQL translation

Query optimization

RDF store

RDBMS

### ABSTRACT

Provenance metadata has become increasingly important to support scientific discovery reproducibility, result interpretation, and problem diagnosis in scientific workflow environments. The provenance management problem concerns the efficiency and effectiveness of the modeling, recording, representation, integration, storage, and querying of provenance metadata. Our approach to provenance management seamlessly integrates the interoperability, extensibility, and inference advantages of Semantic Web technologies with the storage and querying power of an RDBMS to meet the emerging requirements of scientific workflow provenance management. In this paper, we elaborate on the design of a relational RDF store, called RDFProv, which is optimized for scientific workflow provenance querying and management. Specifically, we propose: i) two schema mapping algorithms to map an OWL provenance ontology to a relational database schema that is optimized for common provenance queries; ii) three efficient data mapping algorithms to map provenance RDF metadata to relational data according to the generated relational database schema, and iii) a schema-independent SPARQL-to-SQL translation algorithm that is optimized on-the-fly by using the type information of an instance available from the input provenance ontology and the statistics of the sizes of the tables in the database. Experimental results are presented to show that our algorithms are efficient and scalable. The comparison with two popular relational RDF stores, Jena and Sesame, and two commercial native RDF stores, AllegroGraph and BigOWLIM, showed that our optimizations result in improved performance and scalability for provenance metadata management. Finally, our case study for provenance management in a real-life biological simulation workflow showed the production quality and capability of the RDFProv system. Although presented in the context of scientific workflow provenance management, many of our proposed techniques apply to general RDF data management as well.

© 2010 Elsevier B.V. All rights reserved.

## 1. Introduction

With recent advances in the development of Scientific Workflow Management Systems [34,39,46,69,70,80,123], scientists from various domains are able to automate their experiments using scientific workflows to achieve significant scientific discoveries via complex and distributed scientific computations. As a result, scientific workflow has emerged as a new field to address the new requirements from scientists [70,75]. One such important requirement is provenance management which is essential for scientific workflows to support scientific discovery reproducibility, result interpretation, and problem diagnosis [3,20,96]. This support is enabled via provenance metadata that captures the origin and derivation history of a data product, including the original data sources, intermediate data products, and the steps that were applied to produce the data product. The provenance management

\* Corresponding author. Tel.: +1 956 381 2577; fax: +1 956 384 5099.

E-mail addresses: [artem@cs.panam.edu](mailto:artem@cs.panam.edu) (A. Chebotko), [shiyong@wayne.edu](mailto:shiyong@wayne.edu) (S. Lu), [xubo@wayne.edu](mailto:xubo@wayne.edu) (X. Fei), [fotouhi@wayne.edu](mailto:fotouhi@wayne.edu) (F. Fotouhi).

problem concerns the efficiency and effectiveness of the modeling, recording, representation, integration, storage, and querying of provenance metadata.

While there is an ongoing community effort on standardizing provenance modeling via the Open Provenance Model (OPM) [3], it is still not clear which storage and query model is most suitable for provenance management. Recently, the Semantic Web [16,94] technologies have been increasingly used for provenance management due to their flexibility and semantics support [48,50,65,88,121], such that provenance metadata is represented and captured via Resource Description Framework (RDF) [111,114], RDF Schema (RDFS) [113], and Web Ontology Language (OWL) [110], and queried using the SPARQL [115] query language. This technological suite, enhanced with the Semantic Web inference support, was shown to address [88] the four functional requirements for provenance identified by the Open Provenance Model: (1) provenance information interoperability, (2) ease of application development, (3) precise description of provenance information, and (4) inference capability and digital representation of provenance. In addition, in our work, we choose a Semantic Web approach for provenance management due to its several advantages. First, a *flexible and extensible* data model is needed for provenance representation as what provenance information should be recorded can differ from one system to another and from one domain to another domain and can evolve over time; the RDF data model satisfies such a requirement. Second, it is important to *interpret and reason* about provenance using domain knowledge via domain-specific provenance ontologies; therefore, an inference engine with support of user-defined inference rules is needed as domain-specific provenance ontologies can contain various inference rules (such as “a peptide is derived from a protein”) that cannot be known in advance, and domain-specific provenance ontologies can evolve rapidly over time. Third, provenance interoperability becomes more and more important due to the need of integrating provenance across different provenance models, domains, and organizations in collaborative scientific projects. The RDF model facilitates such *integration and interoperability*. Finally, as RDF serializes graphs, it is naturally suitable for *representation* of provenance graphs with no further adaptation, even though the mapping does not have to be one-to-one (e.g., the OPM implementation as RDF/OWL by the Tupelo project [6]).

In this paper, we propose an approach to provenance management that seamlessly integrates the interoperability, extensibility, and inference advantages of Semantic Web technologies with the storage and querying power of an RDBMS to meet the emerging requirements of scientific workflow provenance management. Our motivation of using the mature relational database technology is provided by the fact that provenance metadata growth rate is potentially very high since provenance is generated automatically for every scientific experiment. On the Semantic Web, large volumes of RDF data are managed with the so called *RDF stores*, and majority of them, including Jena [118,119], Sesame [23], 3store [56,57], KAON [107], RStar [71], OpenLink Virtuoso [42], DLDB [81], RDFSuite [9,105], DBOWL [77], PARKA [101], and RDFBroker [100], use an RDBMS as a backend to manage RDF data. Although a general-purpose relational RDF store (see [15] for a survey) can be used for provenance metadata management, the following provenance-specific requirements bring about several optimization strategies for schema design, data mapping, and query mapping, enabling us to develop a provenance metadata management system that is more efficient and flexible than one that is simply based on an existing RDF store.

- As provenance metadata is generated incrementally, each time a scientific workflow executes, provenance systems should emphasize optimizations for efficient incremental data mapping. As we show in this work, one of such optimizations, a join-elimination optimization strategy, can be developed for provenance based on the property that workflow definition metadata is generated before workflow execution metadata.
- As the performance for provenance storage and that for provenance querying are often conflicting, it may be preferable for a provenance management system to trade data ingest performance for query performance. For example, for long-running scientific workflows, trading data ingest performance for query performance might be a good strategy.
- The identification of common provenance queries has the potential to lead to an optimized database schema design to support efficient provenance browsing, visualization, and analysis.
- Update and delete are not the concern of provenance management since it works in an append fashion, similarly to log management. Therefore, we can apply some denormalization and redundancy strategies for database schema design, leading to improved query performance.

These provenance-specific metadata properties cannot be assumed by a general-purpose RDF store, hampering several interesting data management optimizations to gain better performance for data ingest and querying. While conducting a case study for a real-life scientific workflow in the biological simulation field (see Section 7 for detailed information) to illustrate and verify the validity of our research, we observed that two popular general-purpose RDF stores, Jena and Sesame, could not completely satisfy the provenance management requirements of the workflow. While Sesame could not keep up with the data ingest rate, Jena could not do as good as Sesame on query performance. Both systems lacked support for some provenance queries.

Therefore, by exploiting the above provenance characteristics, we design a relational RDF store, called RDFProv, which is optimized for scientific workflow provenance querying and management. RDFProv has a three-layer architecture (see Fig. 1) that complies with the architectural requirements defined for the reference architecture for scientific workflow management systems [68]. The provenance model layer is responsible for managing provenance ontologies and rule-based inference to augment to-be-stored RDF datasets with new triples. The model mapping layer employs three mappings: (1) schema mapping to generate a relational database schema based on a provenance ontology, (2) data mapping to map RDF triples to relational tuples, and (3) query mapping to translate RDF queries expressed in the SPARQL language into relational queries expressed in the SQL language. These mappings bridge the provenance model layer and the relational model layer, where the latter is represented by a

متن کامل مقاله

دریافت فوری ←

**ISI**Articles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات