Experiencing PROV-Wf for Provenance Interoperability in SWfMSs

Wellington Oliveira[1,2], Daniel de Oliveira[1], Vanessa Braganholo[1]

[1]Instituto de Computação, Universidade Federal Fluminense (UFF), Brazil
[2]Departamento Acadêmico de Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais – Rio Pomba Campus, Brazil
{wellmor, danielcmo, vanessa}@ic.uff.br

Abstract. Analyzing disperse and heterogeneous provenance data usually requires using higher-level tools which scientists need to learn. In our view, scientists should be able to analyze provenance in the SWfMS of their choice. In this paper, we propose Géfyra, an architecture based on the PROV-Wf model, which provides a way to capture heterogeneous provenance data from different SWfMSs into a single format. Géfyra exports and imports provenance data to/from different SWfMSs, allowing scientists to use the system of their choice.

1 Introduction

Depending on the size and complexity of the scientific experiment, it can be divided/modelled into two or more workflows (*i.e.* fragments) [1]. This division can ease the management of the experiment, reducing the total execution time, and enables a cooperative work where each research team works on parts of the experiment in an "independent" way [2]. From there, data provenance management becomes a challenge when the workflow (and their fragments) needs to be executed in more than one SWfMS, and each SWfMS has its own associated provenance model.

For scientists to analyze, share, and combine provenance data generated by different systems, it is necessary to ensure the interoperability between these SWfMSs. Some authors propose an additional layer in a higher level of abstraction [3] to perform the mediation between the provenance data items collected in the various SWfMSs. In our view, the provenance data should be "imported" to one of the SWfMSs used (preferably that the scientist is used to) so that the analysis is performed in a single system, taking advantage of the existing analysis infrastructure of these SWfMSs. Thus, in this paper we propose Géfyra, an approach for provenance data interoperability between existing SWfMSs. Géfyra is based on a recently proposed provenance model called PROV-Wf [4].

LNCS, p. 1, 2014.

2 Géfyra: Making Provenance Interoperable

Our main goal in this paper is to provide a bridge between different SWfMSs so that it allows scientists to analyze provenance data generated by other SWfMS. Thus, we named our approach as Géfyra, which means "bridge" in Greek. We designed a representation schema in XML Schema (which we call *Prov-Wf Schema*) to create and/or validate provenance data from heterogeneous data sources. While designing it, we used some elements of PROV-XML [5] and included all entities and relationships of the PROV-Wf conceptual model. The resulting schema is available at `www.ic.uff.br/~vanessa/papers/PROV-Wf.xsd`.

The Géfyra architecture is shown in Fig. 1. To convert provenance data from SWfMS *A* to SWfMS *B*, the *Géfyra Broker* triggers the cartridge of SWfMS *A*, which converts the data stored in SWfMS *A*'s provenance repository to an XML file that follows the *PROV-Wf Schema*. This XML file is then sent to the *Géfyra Broker*, which stores it in the *PROV-Wf Repository* and sends it to the cartridge within SWfMS *B* for conversion. The cartridge of SWfMS *B* then converts the XML file to SWfMS *B*'s provenance repository format, and stores the provenance information in the repository of that SWfMS. Note that each *Cartridge* knows how to convert from a specific SWfMS format to the PROV-Wf XML format, and vice-versa.

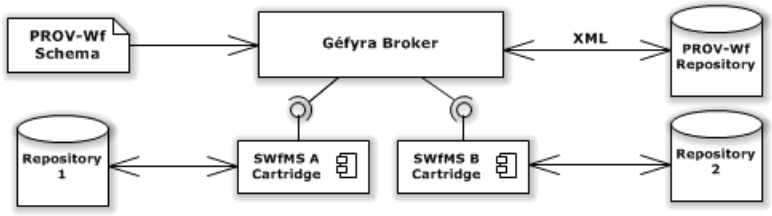

Fig. 1. The Géfyra Conceptual Architecture.

3 Experimental Evaluation and Final Remarks

To evaluate Géfyra, we use the SciPhy [6] workflow that is executed in two SWfMSs that can collect and store provenance data in a relational database: SciCumulus and VisTrails. This way, we develop two cartridges (*PROV-Wf_Sci* and *PROV-Wf_Vis*) to map the provenance data from SciCumulus and VisTrails to XML (according the PROV-Wf Schema) and in the opposite direction, from XML to the SWfMS itself. In order to assess the quality of our mapping, we developed a series of queries to evaluate the amount of tuples and fields, types and values of attributes and the compatibility between the databases of the two SWfMSs. We also were inspired by the queries of the *First and Second Provenance Challenges* [7]. Our main goal was to evaluate information loss that might occur in the import process (since there are some attributes that do not exist in both models), and capture mistakes in our mapping.

To evaluate our results we use the concepts of *precision* and *recall*. Thus, we execute the queries in two provenance databases (SciCumulus and VisTrails) to assess

the amount of records and check whether the fields were aligned to the attributes of the respective elements in the PROV-Wf Schema. Tables 3 and 4 show the results.

Table 1. Results for SciCumulus

Query	Precision		Recall	
	Tuples	Fields	Tuples	Fields
1	100%	100%	100%	100%
2	100%	100%	100%	86%
3	100%	100%	100%	100%
4	100%	100%	100%	100%
5	100%	100%	100%	50%

Table 2. Results for VisTrails

Query	Precision		Recall	
	Tuples	Fields	Tuples	Fields
1	100%	100%	100%	100%
2	100%	100%	100%	100%
3	100%	100%	100%	100%
4	100%	100%	100%	100%
5	100%	100%	100%	75%

The Géfyra architecture is flexible and extensible: new cartridges of different SWfMSs can be connected to it at any time (as shown in Fig. 1). Géfyra maps heterogeneous provenance data sources, allowing the data from a SWfMS to be converted to XML and the latter to another SWfMS. This way, it is not necessary to convert provenance data from each SWfMS to all other provenance systems one wants to use, as Géfyra converts the data to a single XML format that can be shared by all SWfMSs. As a limitation, since one data model may contain data that cannot be mapped to Prov-Wf, some data may be lost in the conversion process. This is the tradeoff of being able to use a single system for analysis.

As future work, we intend to implement cartridges to other SWfMS, we intend to further explore the semantic dimension of the provenance data, and the implications of such a dimension in the mapping of different provenance data sources.

References

1. Marinho, A., Murta, L., Werner, C., Braganholo, V., Cruz, S.M.S. da, Ogasawara, E., Mattoso, M.: ProvManager: a provenance management system for scientific workflows. Concurr. Comput. Pract. Exp. 24, 1513–1530 (2012).
2. Altintas, I., Anand, M.K., Crawl, D., Bowers, S., Belloum, A., Missier, P., Ludäscher, B., Goble, C.A., Sloot, P.M.A.: Understanding Collaborative Studies through Interoperable Workflow Provenance. In: McGuinness, D.L., Michaelis, J.R., and Moreau, L. (eds.) Provenance and Annotation of Data and Processes. pp. 42–58. Springer Berlin Heidelberg (2010).
3. Ellqvist, T., Koop, D., Freire, J., Silva, C., Stromback, L.: Using Mediation to Achieve Provenance Interoperability. 2009 World Conference on Services - I. pp. 291–298 (2009).
4. Costa, F., Silva, V., de Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J., Mattoso, M.: Capturing and querying workflow runtime provenance with PROV: a practical approach. Joint EDBT/ICDT 2013 Workshops. pp. 282–289. ACM, New York, NY, USA (2013).
5. Moreau, L.: PROV-XML: The PROV XML Schema, http://www.w3.org/TR/prov-xml/.
6. Ocaña, K.A.C.S., Oliveira, D. de, Ogasawara, E., Dávila, A.M.R., Lima, A.A.B., Mattoso, M.: SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes. In: Souza, O.N. de, Telles, G.P., and Palakal, M. (eds.) Advances in Bioinformatics and Computational Biology. pp. 66–70. Springer Berlin Heidelberg (2011).
7. Moreau, L., Ludäscher, B. et al.: Special Issue: The First Provenance Challenge. Concurr. Comput. Pract. Exp. 20, 409–418 (2008).

Acknowledgments. The authors would like to thank CNPq and FAPERJ for partially supporting this work.