Provenance-Based Retrieval: Fostering Reuse and Reproducibility Across Scientific Disciplines

Lucas A. M. C. Carvalho[1], Rodrigo L. Silveira[2], Caroline S. Pereira[2], Munir S. Skaf[2], and Claudia Bauzer Medeiros[1]

[1] Institute of Computing, University of Campinas (UNICAMP), Campinas, Brazil
lucas.carvalho,cmbm@ic.unicamp.br
[2] Institute of Chemistry, University of Campinas (UNICAMP), Campinas, Brazil
rodrigolsilveira@gmail.com, caroline013@gmail.com, skaf@iqm.unicamp.br

Abstract. When computational researchers from several domains cooperate, one recurrent problem is finding tools, methods and approaches that can be used across disciplines, to enhance collaboration through reuse. The paper presents our ongoing work to meet the challenges posed by provenance-based retrieval, proposed as a solution for transdisciplinary scientific collaboration via reuse of scientific workflows. Our work is based upon a case study in molecular dynamics experiments, as part of a larger multi-scale experimental scenario.

1 Introduction

Scientific workflows play an important role in data-centric scientific experiments [1] to speed up the construction of new experiments, and foster collaboration through reuse of workflow fragments. This is specially complicated when scientists work in distinct domains, due to heterogeneity in vocabularies, methodologies, perspectives of solving a problem and granularity of objects of interest.

Our work is concerned with meeting the needs of such a heterogeneous research environment, and is based on our ongoing experience with the CCES[3] (Center for Computational Engineering and Science). CCES congregates experts from 6 different domains – Computer Science, Chemistry, Physics, Biology, Applied Mathematics and Mechanical Engineering.

We are helping these scientists to work together via construction and sharing of workflow fragments. However, this is complicated because of the intense heterogeneity of the domains involved.

To meet reusability and transdisciplinary challenges we designed a provenance-centric software architecture to support workflow reuse. We will implement a prototype of the architecture to validate our proposal, running a case study from Molecular Dynamics Simulation [2] (involving both chemists and physicists working each at distinct aspects of the problem).

In our approach, provenance, provided by a scientific workflow system, is semantically enhanced with domain ontologies. This enriched information is then

[3] http://www.escience.org.br

used to support flexibility in workflow retrieval and adaptation across collaborating teams. As discussed further in the paper, provenance information serves as a basis for a wide (new) range of workflow retrieval parameters; furthermore, it allows scientists to assess quality of a workflow fragment.

2 Related Work

Most of the work related to workflow repositories relies on keyword-based retrieval where a user-provided keyword is matched against terms in a workflow's title, workflow's tags or textual description, e.g., myExperiment[4].

Alternatively, semantics-based retrieval mechanisms rely upon semantic annotations which is the process of annotating resources with semantic metadata, using ontologies. The main problem is that annotations require high user effort to describe a workflow, e.g., [3] by augmenting workflow specification, this approach supports workflow retrieval.

Provenance-based retrieval is found in [4] which adopts the ProvONE[5] model. The work of [5] adopts OPM (Open Provenance Model)[6] and takes advantage of keeping the trace of how abstract workflows are instantiated into workflow instances, to assist users in designing new workflows. In Janus [6], domain-specific ontologies are used to annotate the more traditional ”domain agnostic” provenance representation of Taverna workflows.

3 W2SHARE: Architecture and Prototype

The architecture of our framework is shown in figure 1. It is composed of three main layers - interface, provenance-based management, and persistence. Through the interface, scientists can design, semantically annotate and search for (sub)workflows using multiple modes. The persistence layer is responsible for ensuring independence between the middle layer and several repositories. The core of the architecture is the middle layer (Provenance-based Management) and

[4] http://myexperiment.org

[5] http://vcvcomputing.com/provone/provone.html

[6] http://openprovenance.org/model/opmo

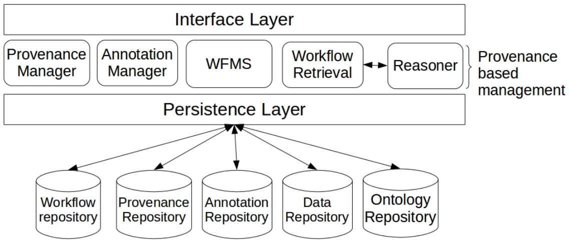

Fig. 1. Architecture of W2SHARE

its semantics retrieval capabilities. This is supported by semantic annotations of: (1) the workflows and their components; and (2) the provenance traces generated by the WFMS. The cross disciplinary search of workflows of interest is based on combining these annotations, emphasizing provenance aspects.

While ontologies have been proposed to enrich provenance data (see [6]), this has not yet been exploited to support the selection/retrieval of appropriate workflow fragments. The use of provenance information to help workflow retrieval appears in [5,4], but these solutions do not fully meet our needs.

The Provenance Manager module is based on extending the work of [7]. It extracts information from provenance traces provided by the WFMS, storing their metadata in the Provenance Repository. It interacts with the Annotation Manager to support annotation of these traces. Annotated provenance is subsequently used by Retrieval mechanisms.

The Annotation Manager is responsible for generating semantic annotations of workflow components (interacting with the WFMS and the Persistence layer) and of provenance information (interacting with the Provenance Manager and the Persistence layer). It also manages the Ontology Repository and feeds the Reasoner. This module is also responsible for connections to other Linked Open Data repositories. This makes it possible to retrieve properties of data which are not explicitly represented in annotated data.

Workflow retrieval combines several kinds of semantics-based mechanisms, taking advantage of annotations managed by the Annotation Manager. The approach to be used to rank the results is still under investigation. However our idea is to use data quality assessment to provide information to the ranking algorithm.

The Inference Reasoner expands knowledge of workflow and provenance annotations through Linked Open Data principles. Moreover, it allows additional relationships among annotated items, this offers possibility to search for concepts which are not explicit in annotation.

4 Case Study - Molecular Dynamics

Our case study concerns molecular dynamics (MD), where simulations are used in material sciences, computational engineering, physics and chemistry.

A typical MD simulation experiment receives as input the structure, topology and force fields of the molecular system and produces molecular trajectories as output. Simulations are subject to a suite of parameters, including thermodynamic variables.

Simulations involve both the atomistic modeling, employed by computational physicists and chemists, and the modeling techniques mostly adopted by engineers to treat problems at the macroscopic scales.

To implement a MD [2] simulation, first, we manually analyzed a suite of scripts designed by physiochemists to translate them into Taverna workflows. Its inputs are the protein structure (PDB: 8CEL), the simulation parameters and force field files. Next, we executed the workflow in Taverna. Then, we will use the annotation facilities provided by our future prototype to annotate workflow components used and provenance data generated by Taverna. To perform annotations, we also have to create an ontology with help of experts, since no such ontology exists.

Once all these (annotated) items are stored, we could then proceed with workflow retrieval. Examples of future search requests include: workflows that uses a protein or a liquid solution; or that are derived from a specific and more abstract workflow; or that involve a specific module; or that were designed by groups based in a certain region or workflow authors.

5 Conclusions and Ongoing Work

This paper presented a provenance-based software infrastructure to enable scientists to reuse and repurpose experiments, modeled as workflows, across different disciplines. We show how we are meeting the challenges faced by CCES to convert script-based experiments into scientific workflows, and subsequently navigate through the workflow repository to find the "most appropriate" workflow fragment. There are many challenges in taking advantage of workflows to support transdisciplinary collaboration. We have chosen semantically enriched provenance information as a basis for workflow retrieval in this context, given the many benefits that can be gained from exploring such information. Our prototype implementation and ontology modeling are ongoing work.

Acknowledgments. Work partially financed by FAPESP (2014/23861-4) and FAPESP/CEPID CCES (2013/08293-7).

References

1. Cohen-Boulakia, S., Leser, U.: Search, adapt, and reuse: the future of scientific workflows. ACM SIGMOD Record **40**(2) (2011) 6–16
2. Silveira, R.L., Skaf, M.S.: Molecular dynamics simulations of family 7 cellobiohydrolase mutants aimed at reducing product inhibition. The Journal of Physical Chemistry B **119**(29) (2014) 9295–9303
3. Gil, Y., Kim, J., Florez, G., Ratnakar, V., González-Calero, P.A.: Workflow matching using semantic metadata. In: the 5th K-CAP, ACM (2009) 121–128
4. Cuevas-Vicenttín, V., Ludäscher, B., Missier, P.: Provenance-based searching and ranking for scientific workflows. In: IPAW. Springer (2014) 209–214
5. Zhai, G., Lu, T., Huang, X., Chen, Z., Ding, X., Gu, N.: Pwmds: A system supporting provenance-based matching and discovery of workflows in proteomics data analysis. In: the IEEE 16th CSCWD, IEEE (2012) 456–463
6. Missier, P., Sahoo, S.S., Zhao, J., Goble, C., Sheth, A.: Janus: From workflows to semantic provenance and linked open data. In: IPAW. Springer (2010) 129–141
7. Malaverri, J., Santanche, A., Medeiros, C.B.: A provenance-based approach to evaluate data quality in eScience. IJMSO **9**(5) (2014) 15–28