A Provenance-Based Infrastructure to Support Reuse and Reproducibility of Computational Experiments

Lucas Augusto M. C. Carvalho[1], **Claudia Bauzer Medeiros**[1] **(advisor)**

[1] Programa de pós-graduação em Ciência da Computação
Instituto de Computação da Universidade Estadual de Campinas (Unicamp)

lucas.carvalho@ic.unicamp.br, cmbm@ic.unicamp.br

Level: Doctoral Degree

Admission: September 2014

Qualifying exam: October 2016 (Expected)

Expected Conclusion: August 2018

Concluded stages: Literature Review; Preliminary definition of the software architecture; Implementation of the initial software prototype; Methodology for conversion of script-based experiments to reproducible workflow-based experiments.

Future stages: Improve the software architecture and prototype; Develop ontology structures; Investigate similarity metrics to allow semantic provenance-based ranking; Validation of the architecture.

Abstract. *One recurrent problem in multidisciplinary research is finding reusable objects (e.g., scripts, code, documents, workflows) that can be used across disciplines to enhance collaboration. This paper presents our ongoing work taking advantage of provenance information, combined with scientific workflows, to help find such objects. We also present challenges posed by provenance-based retrieval, which we propose as a solution for transdisciplinary scientific collaboration via reuse. Our case study in molecular dynamics simulations is part of a larger multi-scale experimental scenario that requires cooperation involving scientists from different disciplines.*

Resumo. *Um problema recorrente em pesquisas multidisciplinares é encontrar objetos reutilizáveis (e.g. scripts, códigos, documentos, workflows) que possam ser usados por diferentes disciplinas para facilitar colaboração entre elas. Este artigo apresenta nosso trabalho em andamento que considera informações de proveniência, combinadas com workflows científicos, para ajudar a encontrar esses objetos. Apresentamos também desafios encontrados na recuperação baseada em proveniência, que propomos como solução para a colaboração científica transdisciplinar através de reuso. Nosso estudo de caso em simulações de dinâmica molecular é parte de um experimento maior em um cenário de multi-escala que demanda cooperação envolvendo cientistas de diferentes disciplinas.*

Keywords: Scientific workflows, workflow retrieval, reuse, provenance, semantic web

1. Introduction and Motivation

Scientific workflows play an important role in data-centric scientific experiments [Cohen-Boulakia and Leser 2011]. As such, they have been often pointed out as a means to speed up the construction of new experiments, and foster collaboration through reuse of (sub)workflows, and/or adaptation or repurposing of entire workflows.

Retrieval for reuse of workflows 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[1] (Center for Computational Engineering and Science), established at University of Campinas (Unicamp).

An ongoing project within CCES involves cooperation between experts in Physics, Chemistry and Mechanical Engineering, to work across different spatial and temporal scales towards developing new materials using nanotechnology. In each of these disciplines, experts have developed data intensive simulations, which are domain-specific, and rely on customized scripts, which are not easily understood nor shareable. Considerable human effort is continuously spent to transfer and reuse results. Thus, we began working towards a more scientist-friendly collaboration environment, based on scientific workflows – see [Carvalho et al. 2016a], for more details.

In [Goderis et al. 2005] it is argued that designing new workflows by reusing and re-purposing previous workflows or workflow patterns has the advantages of reducing workflow authoring time, improving quality through shared workflow development, improving experimental provenance through reuse of established and validated workflows and avoiding workflow redundancy.

Given this scenario, the thesis aims to design a software architecture, called W2SHARE, which provides a flexible semantic provenance-based retrieval mechanism to support workflow reuse in a transdisciplinary research environment. In W2SHARE, provenance information (a.k.a. the history of the origins and transformations applied to a given data product), provided by a scientific workflow system, is semantically enhanced with domain ontologies. This enriched information is then used to support flexibility in workflow retrieval and adaptation across collaborating teams. Here, provenance information serves as a basis for a wide (new) range of workflow retrieval parameters. To validate our proposal, we are implementing and enhancing an initial prototype of the architecture, running a case study from Molecular Dynamics Simulation [Silveira and Skaf 2014].

While ontologies have been proposed to enrich provenance data (see [Missier et al. 2010]), this has not yet been exploited to support the selection/retrieval of appropriate (sub)workflows. The use of provenance information to help workflow retrieval appears in [Da Cruz et al. 2009, Zhai et al. 2012, Cuevas-Vicenttín et al. 2014], but these solutions do not fully meet our needs.

The two main expected contributions of this research are thus: (1) a software infrastructure supporting semantic provenance-based workflow retrieval; (2) "ontological bridges" between multidisciplinary sciences to facilitate collaboration through reuse data

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

and model sharing.

As result of this research we already have two papers published or accepted [Carvalho et al. 2016b, Carvalho et al. 2016a]: The first article [Carvalho et al. 2016b] describes the initial architecture of W2SHARE. The second article [Carvalho et al. 2016a] describes our methodology for conversion of script-based experiments to reusable and reproducible workflow-based experiments. At the final step of our proposed methodology we have a bundle called Workflow-centric Research Object [Belhajjame et al. 2012] aggregating the resources used or produced in the experiment.

This paper focus on the main contributions described in [Carvalho et al. 2016b]. Extensions to this paper include longer related work section (Section 2), more details about the software architecture (Section 3), information about an initial implementation of our prototype (Section 4) and a longer list of ongoing work (Section 5).

2. Related Work

There are many workflows repositories available in the Web. With the growing amount of workflows in these repositories, workflow retrieval mechanisms are gaining increasing attention to help users to find the workflow of interest [Cohen-Boulakia and Leser 2011].

Workflow Retrieval mechanisms can be roughly divided into four main categories: keyword-based, structure (or topology)-based, semantics-based and provenance-based. Some mechanisms combine features from more than one category.

In keyword-based retrieval, a user-provided keyword is matched against terms in a workflow's title, workflow's tags or textual description. For instance, myExperiment[2] looks for keywords to search on workflow title, descriptions and tags. The drawbacks of this approach are the ambiguity of keywords use and lack of semantics in descriptions.

In structure or topology-based approaches, retrieval mechanisms consider workflows as graph structures. Retrieval is based on graph pattern matching, and does not require any human-provided textual information. For example, [Goderis et al. 2006] has investigated retrieval based on component orderings, proposing retrieval techniques and methods for ranking workflows based on graph-subisomorphism matching. Such work has the limitations of high processing costs and lack of semantics of tasks, limiting the similarity to the topology of the labelled graphs.

Semantic-based approaches use semantic annotations. The main problem is that annotations require high user effort to describe a workflow. On the other hand, they are independent of a workflow's internal representation and can be used to compare workflows both across different WfMS (Workflow Management System) and across multiple repositories. In order to attack the manual effort issues, [Gil et al. 2009] augments data and workflow descriptions data with constraints derived from properties obtained from catalogs external to the workflow system. This approach supports workflow retrieval given data-centered queries (e.g. input or output data types) and their combination with other constraints on workflow specification.

Provenance-based retrieval uses provenance information to help search for scientific workflows. Despite most WfMS collect provenance information from the work-

[2] http://www.myexperiment.org

flow execution or during workflow design, provenance is considered only by a few approaches [Zhai et al. 2012, Cuevas-Vicenttín et al. 2014, Da Cruz et al. 2009] to retrieve workflows.

The work of [Cuevas-Vicenttín et al. 2014] uses provenance information and keywords as search features. Additional criteria are used to improve ranking of results, such as authority and quality of service. The infrastructure is composed of a workflow provenance repository, PBase, which adopts the ProvONE[3] model, an extension of PROV-O to support provenance information from workflow specification and execution. The repository also supports declarative graph queries. The work of [Zhai et al. 2012] takes advantage of keeping the trace of how abstract workflows are instantiated into executable workflows to assist scientists in designing new workflows. This approach is limited to the composition stage of the workflow life cycle. The work of [Da Cruz et al. 2009] uses provenance in a resource discovery approach – the process of identifying, locating and accessing resources that have a particular property to implement a single task. Here, resources are elements such as workflows, web services and data sets. However, this work does not consider semantic information associated to the provenance information during the resource discovery process.

In Janus [Missier et al. 2010], domain-specific ontologies are used to annotate the more traditional "domain agnostic" provenance representation of Taverna workflows. Janus also extends an "agnostic" provenance ontology to support annotation of provenance graphs and semantic annotations linked to the Web of Data. However, annotations are not used to retrieve or explore workflows.

3. Contributions

Our interest is to consider provenance information (i.e., the history of the origins and transformation processes applied to a given data product) provided by WfMS [Cruz et al. 2009]. Thus, in this context, provenance includes prospective information (i.e., the computational steps documented by the workflow specification) and retrospective (i.e., the exact steps followed during execution and the data used and produced).

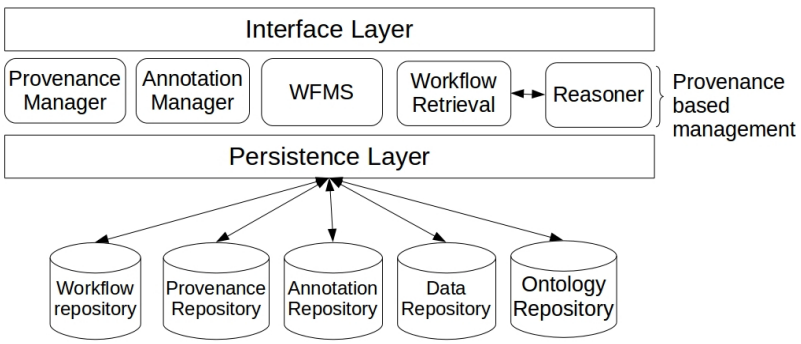

Figure 1. Architecture of W2SHARE as previously defined in [Carvalho et al. 2016b].

The architecture of our framework (W2SHARE), as previously defined in [Carvalho et al. 2016b], is shown in figure 1. It is composed of three main layers - interface,

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

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, as well as managing the links across those repositories.

The main functionality of W2SHARE is 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. The core of the architecture is the middle layer (Provenance-based Management) composed of the following interconnected modules:

- **Provenance Manager**: Responsible for collecting and managing provenance data from workflow execution traces.
- **Annotation Manager**: Responsible for managing semantic annotations of workflow and provenance.
- **Workflow Retrieval**: Responsible for implementing the retrieval mechanisms.
- **WfMS**: In W2SHARE, the main interests are support to design, execution, and provenance generation.

The Provenance Manager is based on extending the work of [Malaverri et al. 2014]. 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.

Typically, a WfMS generates traces that contain information such as input, output and intermediate data and processes. However, provenance data itself is insufficient to answer typical provenance queries. For this reason, [Missier et al. 2010] proposes to semantically annotate provenance data. We follow the same approach, extending it to support provenance-based workflow retrieval.

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.

Annotated provenance is key to reusability. First, by annotating provenance data, it is possible to check quality metrics. In most scientific domains, provenance is key to checking quality criteria such as reliability and soundness of process and completeness of data sources. Search annotated provenance supports more advanced queries wrt data-centric properties, and abstract and concrete tasks.

The semantic annotation of a workflow creates new possibilities for collaboration across domains. Our hypothesis is that the abstract model can be derived from the concrete model annotated using domain-aware ontologies. By annotating the workflow specification, it is possible to construct a more abstract model of the workflow which is more reusable and offer higher level concepts to be used on the retrieval mechanism.

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 requirements for appropriate workflow matching and discovery [Goderis et al. 2005, Goderis et al. 2006]: make clear that in order to support them the retrieval mechanism needs more information available than there is in scientific workflow repositories, and that provenance information is required. Thus, the need to support annotating not only provenance data but also workflow specifications with domain-specific ontologies. In particular, a heterogeneous inter-disciplinary collaboration environment requires additional provenance-based retrieval such as based on: (i) relations between abstract task and concrete instances; (ii) semantic aspects of abstract and concrete tasks and (iii) quality criteria.

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. Preliminary Evaluation

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 simulation, first, we manually analyzed a suite of scripts designed by physiochemists and converted them into Taverna workflows using our semi-automatic conversion approach defined in [Carvalho et al. 2016a]. We will use the future annotation facilities provided by our prototype to annotate the workflow components and provenance information exported by the WfMS.

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; that are derived from a specific and more abstract workflow; that involve a specific module; that were designed by groups based in a certain geographic region or workflow authors.

After retrieving a workflow, the challenge is how to support its reuse by experts from other domains (e.g. mechanical engineering). Although they cooperate with physics and chemistry researchers, and also use MD, they diverge issues such as program, input data, system environment and data granularity. These differences may require considerable modifications in the workflow. We are working under the assumption that retrieval of abstract workflows (and the corresponding concrete workflows) may help reuse.

As a proof-of-concept we are developing an web prototype with the main functionalities to support scientists in publishing and searching workflows for reuse or repur-

posing. The prototype is available at `http://w3id.org/w2share/`.

We used Virtuoso triple store database to implement all the repositories specified in figure 1. For workflow design and execution (see figure 1), we adopted Taverna. Thus, at this moment, the prototype only supports Taverna workflows. The provenance model supported is the one exported by Taverna's provenance manager and defined by the Research Objects suite of ontologies [Belhajjame et al. 2015].

The Annotation Manager follows the Linked Data principles to connect the semantic annotation to Linked Open Data (LOD) portals. For Chemistry, our case study, there are some LOD portals such as Worldwide Protein Data Bank[4] in RDF format. The user can enter annotations to the Annotation Manager via a friendly web interface. They are then transformed by the Manager in SPARQL expressions, to be stored in Virtuoso. To perform annotations, we are still considering three options: (i) create an ontology from scratch with help of domain experts; (ii) use an initial ontology designed to Biomolecular Dynamics simulations; or (iii) extend some existing chemistry ontologies involving computational chemistry.

The Workflow Retrieval module enacts the inference reasoner provided by Apache Jena and interacts with Virtuoso. We are considering how we will implement this module.

5. Final remarks

This paper presented W2SHARE, our proposal to create a semantic provenance-based software infrastructure to enable scientists to reuse and repurpose experiments, modeled as workflows, across different disciplines. We have chosen semantically enriched provenance information as a basis for workflow retrieval in this context, given the many advantages that can be gained from exploring such information, as opposed to other techniques. We showed, through our implemented case study, how we are meeting the challenges faced by CCES to share experiments and find the "most appropriate" (sub)workflows.

We are concentrating our activities in the following important aspects, in close cooperation with all domain experts involved: (i) Integrating the methodology developed in [Carvalho et al. 2016a] into the W2SHARE architecture; (ii) Developing the semantic annotation facilities in W2SHARE; (iii) Developing the "ontological bridge" and extending the target domain to other disciplines; (iv) Investigating approaches where semantic annotations originate abstract workflows from the concrete workflows; (v) Investigating similarity metrics to allow ranking of semantic provenance-based workflow retrieval.

Acknowledgments Work partially financed by FAPESP (2014/23861-4), FAPESP/CEPID CCES (2013/08293-7). We thank Prof. Munir Skaf and his group from the Institute of Chemistry at Unicamp for their scripts, data and valuable feedback.

References

[Belhajjame et al. 2012] Belhajjame, K., Corcho, O., Garijo, D., Zhao, J., Missier, P., Newman, D., Palma, R., Bechhofer, S., García Cuesta, E., Gómez-Pérez, J. M., et al. (2012). Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceedings of Workshop on the Semantic Publishing, (SePublica 2012)*.

[4] `http://rdf.wwpdb.org/`

[Belhajjame et al. 2015] Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., Mina, E., Corcho, O., Gómez-Pérez, J. M., Bechhofer, S., et al. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Web Semantics: Science, Services and Agents on the World Wide Web*.

[Carvalho et al. 2016a] Carvalho, L. A. M. C., Belhajjame, K., and Medeiros, C. B. (2016a). Converting scripts into reproducible workflow research objects. *In Proceedings of the 2016 IEEE 12th International Conference on eScience (eScience 2016)*.

[Carvalho et al. 2016b] Carvalho, L. A. M. C., Silveira, R. L., Pereira, C. S., Skaf, M. S., and Medeiros, C. B. (2016b). Provenance-based retrieval: Fostering reuse and reproducibility across scientific disciplines. *In IPAW 2016*, pages 183–186.

[Cohen-Boulakia and Leser 2011] Cohen-Boulakia, S. and Leser, U. (2011). Search, adapt, and reuse: the future of scientific workflows. *ACM SIGMOD Record*, 40(2):6–16.

[Cruz et al. 2009] Cruz, S. M. S. d., Campos, M. L. M., and Mattoso, M. (2009). Towards a taxonomy of provenance in scientific workflow management systems. In *The 4th IEEE SERVICES-1*, pages 259–266. IEEE.

[Cuevas-Vicenttín et al. 2014] Cuevas-Vicenttín, V., Ludäscher, B., and Missier, P. (2014). Provenance-based searching and ranking for scientific workflows. In *IPAW*, pages 209–214. Springer.

[Da Cruz et al. 2009] Da Cruz, S. M. S., Barros, P. M., Bisch, P. M., Campos, M. L. M., and Mattoso, M. (2009). A provenance-based approach to resource discovery in distributed molecular dynamics workflows. In *Resource Discovery*, pages 66–80. Springer.

[Gil et al. 2009] Gil, Y., Kim, J., Florez, G., Ratnakar, V., and González-Calero, P. A. (2009). Workflow matching using semantic metadata. In *the 5th K-CAP*, pages 121–128. ACM.

[Goderis et al. 2006] Goderis, A., Li, P., and Goble, C. (2006). Workflow discovery: the problem, a case study from e-science and a graph-based solution. In *ICWS*, pages 312–319. IEEE.

[Goderis et al. 2005] Goderis, A., Sattler, U., Lord, P., and Goble, C. (2005). Seven bottlenecks to workflow reuse and repurposing. In *The Semantic Web–ISWC 2005*, pages 323–337. Springer.

[Malaverri et al. 2014] Malaverri, J., Santanche, A., and Medeiros, C. B. (2014). A provenance-based approach to evaluate data quality in eScience. *IJMSO*, 9(5):15–28.

[Missier et al. 2010] Missier, P., Sahoo, S. S., Zhao, J., Goble, C., and Sheth, A. (2010). Janus: From workflows to semantic provenance and linked open data. In *IPAW*, pages 129–141. Springer.

[Silveira and Skaf 2014] Silveira, R. L. and Skaf, M. S. (2014). Molecular dynamics simulations of family 7 cellobiohydrolase mutants aimed at reducing product inhibition. *The Journal of Physical Chemistry B*, 119(29):9295–9303.

[Zhai et al. 2012] Zhai, G., Lu, T., Huang, X., Chen, Z., Ding, X., and Gu, N. (2012). Pwmds: A system supporting provenance-based matching and discovery of workflows in proteomics data analysis. In *the IEEE 16th CSCWD*, pages 456–463. IEEE.