Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 100 (2016) 547 – 556

Procedia Computer Science

Conference on ENTERprise Information Systems / International Conference on Project MANagement / Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2016, October 5-7, 2016

E-SECO ProVersion: An Approach for Scientific Workflows Maintenance and Evolution

Tássio F. M. Sirqueira[a,b*], Humberto L. O. Dalpra[a], Regina Braga[a], Marco Antônio P. Araújo[a,b], José Maria N. David[a], Fernanda Campos[a]

[a]*Master Program in Computer Science, Federal University of Juiz de Fora, Juiz de Fora (MG), Brazil*
[b]*Federal Institute of Education, Science and Technology of Southeast Minas Gerais, Juiz de Fora (MG), Brazil*

Abstract

This paper discusses the use of evolution and maintenance techniques in scientific workflows context. Using version control and data provenance techniques, strategical information can be collected and analyzed using ontological rules. We present the E-SECO ProVersion approach, specified in the context of a Scientific Software Ecosystem, named E-SECO. E-SECO ProVersion is aimed of acting as a mediator to extract maintenance and evolution information from different workflows repositories. Therefore, it address the maintenance and evolution of scientific workflows, adapting concepts of software evolution and maintenance for the scientific experimentation context, through versioning and similarity degree obtained from data provenance support.

Peer-review under responsibility of the organizing committee of CENTERIS 2016

Keywords: Workflow Maintenance, Workflow Evolution, Data Provenance, E-SECO;

1. Introduction

The use of computing resources by scientists is becoming a widespread practice, which generates a large volume of data to be considered in further experiments. It is fundamental that these data can be shared with the aim of

* Corresponding author. Tel.: +55-322102-3387; fax: +55-322102-3387.
E-mail address: tassio.sirqueira@ice.ufjf.edu.br

Peer-review under responsibility of the organizing committee of CENTERIS 2016
doi:10.1016/j.procs.2016.09.194

developing high performance solutions based, for example, in reuse, data management and distributed processing concepts. Isolated experimental processes concentrate the knowledge only on the scientist that is conducting the analysis, which is not the proposal of the so-called collaborative laboratories17, the current tendency in scientific experimentation.

A scientific experiment can be defined as the use of interconnected scientific applications, using or not scientific workflows[5]. A scientific workflow is a model or template that represents a sequence of activities, implemented by tools (programs and services)[2], and that can be interpreted and executed through a Scientific Workflow Management System (SWMS).

Thus, experimental processes that utilize workflows often suffer from the problem of loss of knowledge, considering that the processed tasks are not generally adequately documented[11]. The tasks are specified in SWMSs that, in most cases, are limited to manage the execution of the scientific workflow isolated from the experiment. However, to represent and support the development of scientific experiment adequately, SWMS needs to register the changes in the associated workflows, as these workflows can be modified during the research[12]. The loss of information about changes can also cause a lack of control in conducting the experiment, due to the absence of knowledge about the origin of the data generated in the execution life cycle. The inexistence or the failure in capturing this information may also culminate in the difficulty in identifying the workflow that generated the experiment, or more specifically, the version that would enable obtaining more concise information about it.

According to [7], the use of independent tools to manage the experiment and to analyze its data is necessary, considering that the major part of SWMSs do not have such functionality. In this context, workflows version control, based on the maintenance and evolution techniques, can be integrated into the experiment life cycle, as shown in Figure 1. In this case, the workflow life cycle becomes part of the experiment life cycle which contributes to the decision-making and control by the researcher.

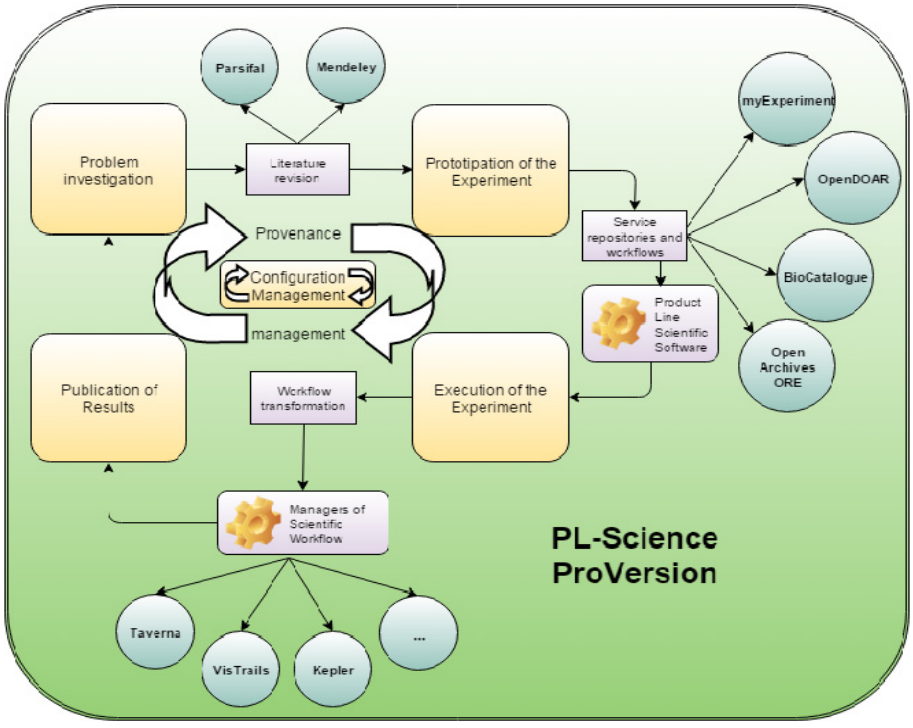

Fig. 1. E-SECO ProVersion Experiment cycle

Therefore, this study aims to address the maintenance and evolution of scientific workflows, adapting concepts of software evolution and maintenance for the scientific experimentation context, through versioning and similarity

degree obtained with data provenance support[18]. The paper proposes the development of an architecture for scientific workflows evolution and maintenance, automatically providing strategic information for scientists to make decisions related to experiment or to obtain additional knowledge about it. This architecture is named E-SECO ProVersion, which extends the E-SECO architecture[4].

The paper has five sections, besides the introduction. Section 2 presents the research context. Section 3 presents the related work. In section 4, it is presented E-SECO ProVersion. Section 5 discuss E-SECO ProVersion use. Lastly, Section 6 presents conclusions and future works.

2. Research Background

The software maintenance activity is characterized by the revision of a software product, already delivered to the client, for errors correction, improvement in performance or product adaptation to a new environment[16]. In this context,[10] presents the "Laws of Software Evolution", which addresses eight aspects, normally observed at a software life cycle. Some of these aspects can be applied in the context of maintenance and evolution of workflows, such as: i) the "Law of Continuing Change", which states that all software has to be modified continuously or will become less useful. Applying to the workflow context, as the research advances, workflows tend to require evolution to keep up useful for research; ii) the "Law of Increasing Complexity", which defines that, with the process of evolution, the structure tends to become more complex and requires more attention. Thus, as a workflow is expanded, it should be optimized so that there is no degradation with time[8]; iii) The "Law of Continuing Growth" defines that during software life cycle, incremental changes are constant. In the context of workflows, it should evolve as the research evolves, with new functionalities.

Considering an experiment life cycle[12,2,8] proposed the creation of a new stage, called "the optimization of the scientific workflow", to identify faults or weak points in the workflow in order to optimize the adaptation and facilitate its reuse. In this context, workflows can be adapted along the experimentation cycle, being necessary the control of changes to enable the replication of the study. In this scenario, workflow data versions should be stored in a repository along with the required parameters and resources necessary for the execution. The capture and storage of workflow versions and their related data may be performed using data provenance techniques[18] aiming to get the description of the origins of a piece of data and the method by which it is passed.

The data provenance can be divided into prospective and retrospective types. According to[3], prospective provenance captures the steps that must be followed in order to generate a product, while the retrospective provenance captures the steps performed by a computational task, as well as the information about the environment. Through a model that supports the exchange of provenance information in heterogeneous environments, it is possible to use provenance information about entities, activities and people involved in producing a piece of data, enabling the certification of quality or reliability of an experiment. Our work uses the PROV provenance model[6], which provides several forms of knowledge representation, which is detailed in Section 4.

3. Related Works

The work presented in[14] proposes the versioning of workflows using the cosine distance measure[9] focusing on supporting the analysis of processes repositories where workflows are grouped, according to the similarity between the activities. An approach independent from the comparison model, through SiDiff application, is presented by[15]. MyExperiment [5] proposes the separation of scientific workflows by organizing them into groups in order to allow the reuse in other experiments. It also allows the interaction among researchers, through the exchange of personal messages. However, it does not allow the analysis of the results of an experiment, to download results or to access provenance repositories directly[13]. Another repository, called CrowLabs, was developed by the VisTrails group[1]. It supports the execution of workflows, the import of data input, the analysis and reuse by third parties. However, it does not allow the analysis of data provenance, being available to researcher only the data in XML format[14].

An architecture for comparison and grouping of pre-existing workflows, by similarity, for the construction of many experiments by means of bottom up approaches, is presented in SimiFlow. A portal, called CollabCumulus, allows access to the content of provenance repository, enabling data analysis that is generated or consumed. The scientists have access to a set of provenance repositories and they can comment, create discussion topics and analyze

the results collaboratively with its research partners. However, despite having workflow information, there are no information about its maintenance and evolution. Besides, data storage format follows its own pattern and does not use a standard model to support data exchanging, such as PROV model and does not use ontologies for performing inferences over workflow information, providing strategical information to researchers.

4. E-SECO ProVersion

The E-SECO ProVersion has as its main objective to adapt maintenance and evolution foundations to scientific workflows context, by storing the modeling and executing data of a workflow in a provenance repository, based on PROV model[6], allowing analysis, versioning, grouping and derivation of scientific workflows. The data provenance assist in evaluating the results of an experiment, enabling the analysis of workflow historical data throughout its life cycle. The use of PROV provenance model enables the knowledge extraction through their causal relationships, and associated with the use of ontology, it allows the inference of new knowledge. The data collected have information that relates the services used by the workflow, its input parameters and the results obtained from its execution.

Among the maintenance types that can be used in a workflow, we can mention: (i) Corrective: corrective maintenance in the workflow are used to adjustments during the experiment creation. This type of maintenance occurs until the workflow process is stable, allowing the use by researchers; (ii) Adaptive: adaptive maintenances are used when there is the necessity to create a workflow similar to an existing one, being used as a base model available in the repository; (iii) Evolutionary: evolutionary maintenance occurs by the need for a new resource for execution of the workflow that is not available, such as web service. This maintenance type has the objective to modify the workflow in order to allow that its use continues active. The replacement of a service with a new one with the same functionalities can be cited as an example; (iv) Reengineering: the reengineering maintenance can be used to workflow process optimization. This maintenance type depicts a constant step in the workflow life cycle, as proposed by[8]. It is important because it enables the evaluation of workflow execution efficiency.

The adoption of a provenance model, coupled with the use of ontology, which uses inference rules (Property Chains† in OWL 2.0), eases the discovery of information about evolution and maintenance of a workflow. Information such as (i) workflow evolutionary line; (ii) with which other workflows the same tasks are shared; (iii) what modifications occurred during the experimental cycle; (iv) which type of experiment is linked and who are its developers, among others, are useful for new researchers, or even researchers who had been using the original workflow. As a result, they can learn about changes and, depending on the context, migrate experiment to the new version of workflow. In addition, based on information obtained by the inferences held by the ontology, one can identify faults or changes in the tasks, and view affected workflows, increasing the need for maintenance to keep the workflows useful for a research.

The architecture of E-SECO ProVersion is an evolution of E-SECO architecture[4] with the addition of two new modules, as presented in Figure 2. The provenance module, is responsible for collecting data from the modeling and execution of the workflow. These data contribute (i) to detect the need for maintenance or evolution of the workflow; (ii) to indicate runtime errors, performance problems, and the need for new functionalities or data failures at third-party services. Data is captured from a web service, coupled with a SWMS, and has the support of an ontology, integrated to the module, which assists in the generation of knowledge through inferences. The other module is the workflow version management where data related to tasks, input values, output and information from experiments are associated with respective workflows and analyzed to allow the identification of similar workflows, evolution characteristics and / or maintenance needs. It uses as a basis the data captured in provenance module and inferred by ontology. Subsequently, researchers can query this data.

Database model‡ extends E-SECO relational database and PROV provenance model[6]. Information about the number of tasks and which of them are used as indicative of workflow evolution or maintenance, are obtained through inferences in the ontology that extends the PROV original ontology, named PROV-O[6]. PROV-O provides a

† Property Chains appeared in the OWL 2 and runs classifying objects, using transitivity between multiple properties.
‡ Database model is available in http://goo.gl/QiiCKb.

set of classes, properties and constraints that can be used to represent and exchange provenance information generated in different systems and contexts. However, PROV-O does not express all required knowledge to support workflow maintenance and evolution. The extension of PROV-O, allowed the discovery of new information for capture prospective and retrospective provenance (original PROV-O only captures retrospective provenance). This new ontology, developed in E-SECO ProVersion context, is named PROV-OEXT. The use of PROV-OEXT, aligned with the database model, allows the extraction and inference of new knowledge related to scientific workflows concerning evolution and maintenance. The causal relationships of PRO-OEXT ontology are presented in Figure 3.

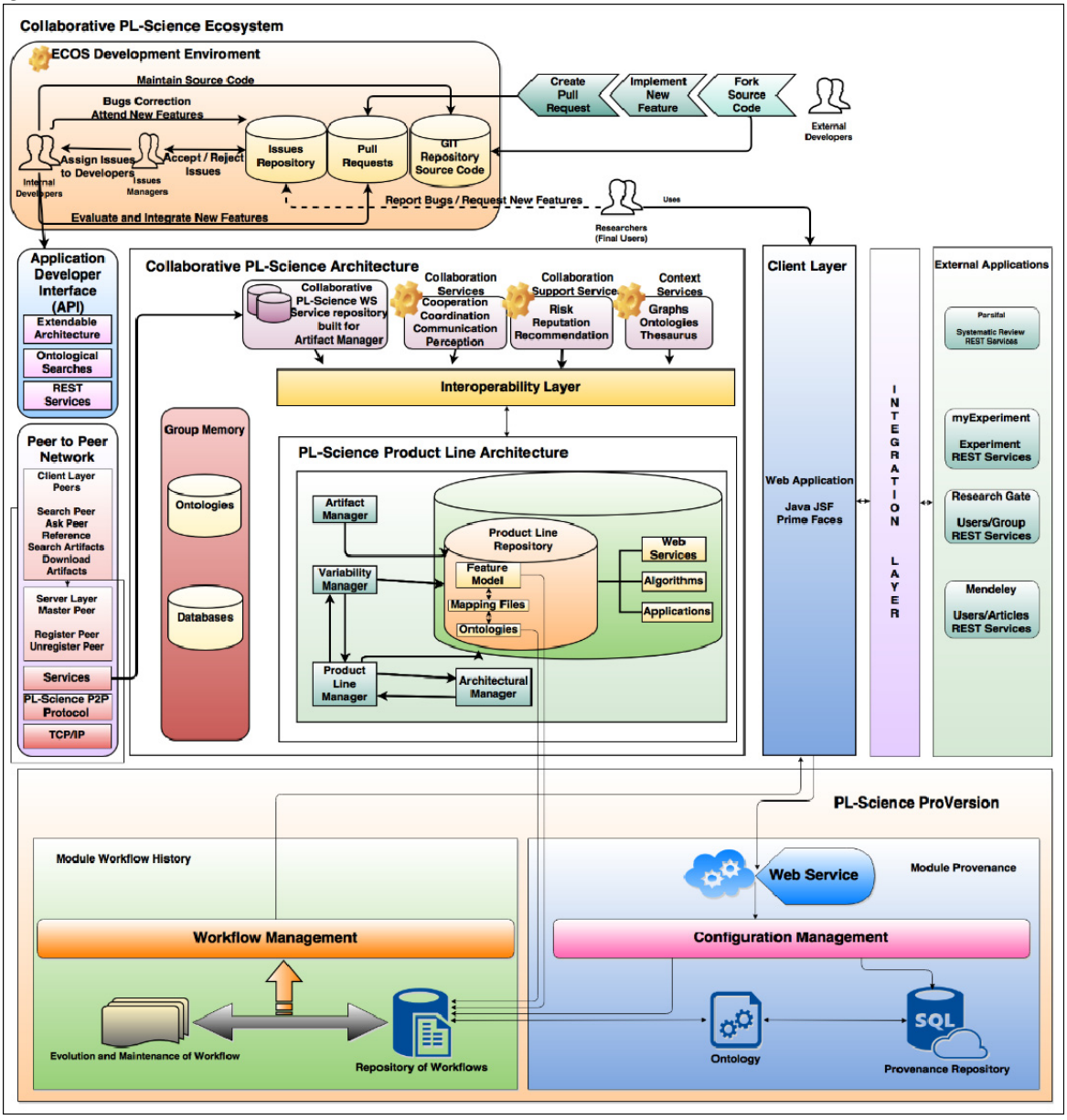

Fig. 2. E-SECO ProVersion Architecture (adapted from [4])

5. E-SECO ProVersion in Action

As mentioned, the E-SECO ProVersion primarily operates in modeling and execution of the experiment, being possible to access workflows and services repositories such as MyExperiment, BioCatalogue, among others. Considering the number of tasks and the verification of the tasks that are used in a specific workflow, it is possible to detect other similar workflows. A scientist can also get information about failures in tasks through the data captured during execution and information of changes that occurred between versions of a given workflow, providing to the researchers strategical knowledge about workflows versions used in an experiment. For this, the data collected during workflows modeling and execution are analyzed using the PROV-OEXT ontology. It allows the identification of the evolution history and indicates which workflows are predecessors and successors of a given workflow, as well as the tasks and SWMS used by them, and the experiment to which it belongs, as shown in Figure 4.

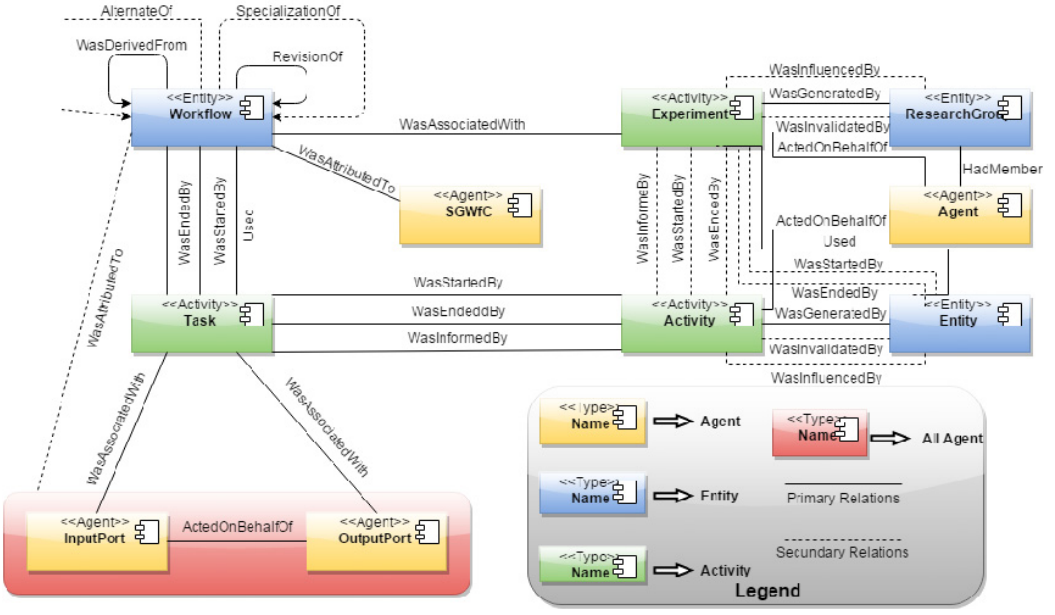

Fig. 3. Causal relations of PROV-OEXT ontology

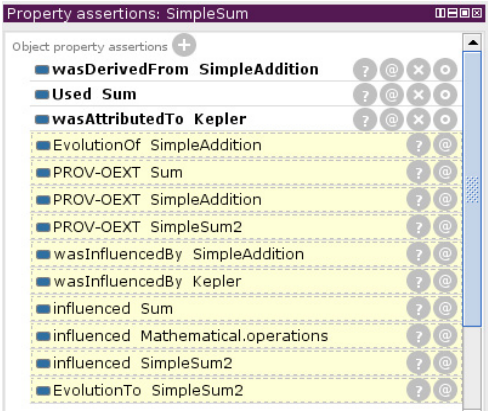

Fig. 4. Inferences using PROV-OEXT and data from workflows executions

(presented in Protégé tool interface).

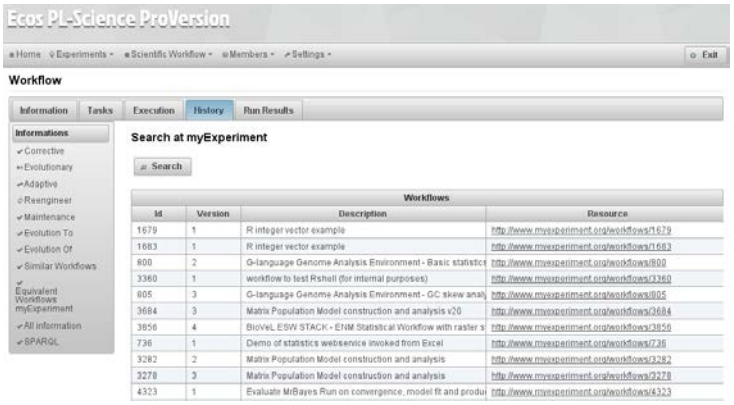

Fig. 5. E-SECO ProVersion Interface.

From this information, the E-SECO ProVersion may suggest strategic changes both related to the modeling and execution of the workflow, helping the scientist, or the group coordinator to improve experiment results and to identify failures or points that need to be corrected later. For this, E-SECO ProVersion architecture provides a web interface so that the information about evolution and maintenance can be easily queried and analyzed, contributing to the evaluation of data produced in the experiment. An example of the interface containing inferred workflows information, using PROV-OEXT ontology, can be seen in Figure 5.

Furthermore, with aim of verifying the use of descent approaches to the construction of new workflows, which generates an evolutionary line, it was analyzed a group of workflows of a scientist, from the myExperiment repository. Considering these data, an example of E-SECO ProVersion use is presented in Figure 6 (A§ and B**). Both workflows were created by the same scientist and are related to the search for information on a medical basis. E-SECO ProVersion detected that workflow B is an evolved version of workflow A, and other researchers, who have no knowledge of the existence of two versions of the same workflow, are informed about it, avoiding rework. Besides, the scientists can access information about their success or failure, problems reported during the execution, and modifications required from one version to another.

§ http://www.myexperiment.org/workflows/1975.html
** http://www.myexperiment.org/workflows/1976.html

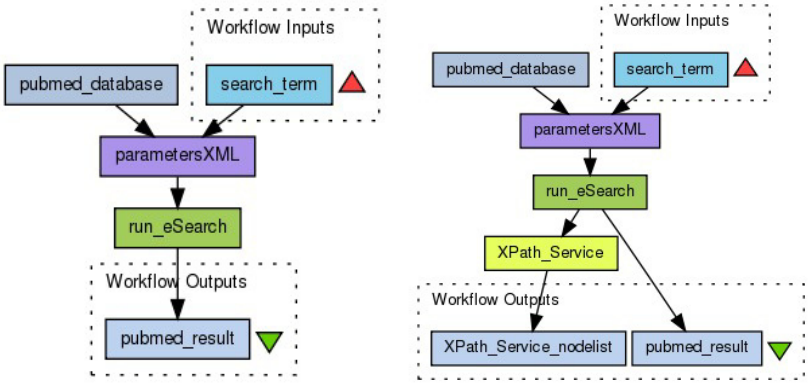

Fig. 6. (A) Workflow A; (B) Workflow evolved (A).

Figure 6 presents Taverna workflows, with tasks registered in E-SECO ProVersion, and generated an identifier for each task. Verifying which tasks use a particular service, it is possible to identify which workflows use the same tasks. The use of common tasks by a workflow defines the similarity between them, been more or less similar, depending on the number of common tasks. It is possible to note that the workflow B is an evolution of A, given that the inputs and, consequently, the parameters are the same. They also use the same task "run_eSearch", but workflow B has the addition of a new service "XPath_Service", which results in a different output . That information underlies the need for corrective maintenance, such as replacing the old service by a similar one, keeping the workflow useful for the experiment and for use by other scientists in similar research context. Another illustrative example of E-SECO ProVersion use, considering the myExperiment repository, is an workflow that has inactive services. This inconsistence is not detectable on myExperiment repository. This issue can contribute to hinder the workflow reuse. However, with help of E-SECO ProVersion, it can be easily detected.

Another important source that can help to improve the results is historical data about workflows. The use of ontologies and inference mechanisms help to derive implicit knowledge for the improvement of the evolution and maintenance of workflows, but important information can also be discovered from historical data. Considering this context, one of the problems encountered in historical analysis of scientific workflows and the proposition of improvements based on these data is the lack of historical databases. In this vein, E-SECO ProVersion is prepared to have an evolution and maintenance repository. Our proposal is prepared to combine historical workflow data with ontology, using a wrapper to existing workflow repositories (such as myExperiment) in order to contribute to the construction of a knowledge base about workflow evolution and maintenance.

The use of provenance, ontologies and inference rules to support evolution and maintenance of workflows were presented in section 4. The specification of a database model based on PROV model, has also been specified and implemented. Thus, the next step in E-SECO ProVersion specification is to provide access to existing workflows repositories in order to capture historical data needed to support the evolution and maintenance of workflows. For this purpose, based on search conducted in the repositories and in the literature, the following information must be captured from these repositories: version information, task information, creation information, maintainability, dependences and usage information.

The extracted historical data can be used in E-SECO ProVersion in order to mine newer evolution and maintenance information related to a given workflow or its services.

6. Conclusion

The E-SECO ProVersion is an extension of E-SECO[4], a web-based software ecosystem designed to support researchers to accomplish activities during the overall scientific workflow life cycle. Through the collection and

analysis of workflow execution data, it is possible to build an evolution and maintenance knowledge base, supporting reuse, adaptation and optimization of workflows.

Currently, E-SECO ProVersion allows scientists to capture workflow data available via a web service. Data collected by the architecture are stored in a database, which was modeled based on the PROV provenance model, allowing querying workflow data. These data feed PROV-OEXT ontology where specific domain rules detect information about the evolution and maintenance of a workflow, and this information is made available to the scientist using a web interface. This information is provided to scientists in an automated manner, facilitating the identification of failures and hence the search for a solution. The approach also contributes to the specification of a knowledge base, which can detect the need for maintenance or evolution points in workflows. Besides, the strategic information can also help scientists when optimizing the process and detecting eventual failure.

However, historical workflows data can bring even more gains, concerning evolution and maintenance. Based on preliminary analysis of available workflows repositories, it was possible to identify a lack of evolution and maintenance information in such repositories. Therefore, it was observed the need to store the workflow history and that this information can become available to other scientists.

The specification of this body of knowledge in evolution and maintenance activities contributes to the community as a whole, since anyone can make use of provenance information for reusing workflows in their experiments. Thus, it is part of E-SECO ProVersion architecture, the development of a historical workflow data repository, which can take information from the existing repositories, and use the PROV model as a basis for modeling, aligned with ontologies for extracting implicit knowledge.

Acknowledgements

The authors thank CAPES, Fapemig and CNPq for the support and encouragement of research.

References

1. Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. Managing the evolution of dataflows with VisTrails. In Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on. IEEE, 2006. p. 71-71.
2. Deelman, E., Gannon, D., Shields, M., & Taylor, I. Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25(5), 2009. p. 528-540.
3. Freire, J., Koop, D., Santos, E., & Silva, C. T. Provenance for computational tasks: A survey. Computing in Science & Engineering, 10(3), 2008. p. 11-21.
4. Freitas, V.; David, J. M. N.; Braga, R. M.; Campos, Fernanda. An architecture for a Scientific Ecossystem. In: 9th Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems (WDES 2015), Belo Horizonte, Brazil, SBC, 2015. p. 41-48.
5. Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D. & De Roure, D. myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic acids research, 38(suppl 2), 2010. p. 677-682.
6. Groth, P., Moreau, L., PROV-Overview: An Overview of the PROV Family of Documents. Disponível em: <http://goo.gl/tkloV5>. Accessed in nov. 2015.
7. Hasan, R., Sion, R., & Winslett, M. Introducing secure provenance: problems and challenges. In Proceedings of the 2007 ACM workshop on Storage security and survivability, 2007. p. 13-18.
8. Holl, S., Zimmermann, O., Palmblad, M., Mohammed, Y., & Hofmann-Apitius, M. A new optimization phase for scientific workflow management systems. Future generation computer systems, 36, 2014. p. 352-362.
9. Jung, J.-Y., Bae, J., Workflow Clustering Method Based on Process Similarity, Computational Science and Its Applications - ICCSA 2006, 2006. p. 379- 389.
10. Lehman, M., Laws of Software Evolution Revisited. In: 5TH European Workshop on Software Process Technology. 1996. p. 108-124.
11. Marinho, A., Murta, L., Werner, C., Braganholo, V., Cruz, S. M. S. D., Ogasawara, E., & Mattoso, M. ProvManager: a provenance management system for scientific workflows. Concurrency and Computation: Practice and Experience, 24(13), 2012. p. 1513-1530.
12. Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Ogasawara, E., Oliveira, D. & Murta, L. Towards supporting the life cycle of large scale scientific experiments. International Journal of Business Process Integration and Management, 5(1), 2010. p. 79-92.
13. Miranda, G; de Souza, J. A.; Braganholo, V.; Oliveira, D. de. CollabCumulus: Uma Ferramenta de Apoio à Análise Colaborativa de Proveniência em Workflows Científicos. XI Brazilian Symposium of Collaborative Systems (SBSC), 2014. p. 94-101 (in portuguese).
14. Santos, Emanuele et al. Vismashup: Streamlining the creation of custom visualization applications. Visualization and Computer Graphics, IEEE Transactions on, 2009. v. 15, n. 6, p. 1539-1546.

15. Schmidt, M., & Gloetzner, T. Constructing difference tools for models using the SiDiff framework. In Companion of the 30th international conference on Software engineering, 2008. 947-948, ACM.
16. Sommerville, I., Boggs, W., Boggs, M., Bruegge, B., Dutoit, A. H., Boggs, W., & Boggs, M. Software Engineering 7 th ed. 2012.
17. Olson G.M. The next generation of science collaboratories. Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems, CTS '09, IEEE Computer Society:Washington, DC, USA; 2009. xv–xvi, doi:10.1109/CTS.2009.5067429.
18. Buneman, Peter, Sanjeev Khanna, and Tan Wang-Chiew. "Why and where: A characterization of data provenance." Database Theory—ICDT 2001. Springer Berlin Heidelberg, 2001. 316-330.