Using Context Elements and Data Provenance to Support Reuse in Scientific Software Ecosystem Platform

Lenita M. Ambrósio, José Maria N. David, Regina Braga,
Fernanda Campos, Victor Ströele and Marco Antônio Araújo
Department of Computer Science, Federal University of Juiz de Fora (UFJF), Juiz de Fora, MG, Brazil

Keywords: Contextual Elements, Provenance, Scientific Experiments, E-Science.

Abstract: [Background] Managing contextual elements and provenance information plays a key role in the context of scientific experiments. Currently the scientific experimentation process requires support for collaborative and distributed activities. Detailed logging of the steps to produce results, as well as the environment context information could allow scientists to reuse these results in future experiments and reuse the experiment or parts of it in another context. [Objectives] The goal of this paper is to present a provenance and context metadata management approach that support researchers to reuse experiments in a collaborative and distributed platform. [Method] First, the context and provenance management life cycle phases were analyzed, considering existing models. Then it was proposed a conceptual framework to support the analysis of contextual elements and provenance data of scientific experiments. An ontology capable of extracting implicit knowledge in this domain was specified. This approach was implemented in a scientific ecosystem platform. [Results] An initial evaluation shown evidences that this architecture is able to help researchers during the reuse and reproduction of scientific experiments. [Conclusions] Context elements and data provenance, associated with inference mechanisms, can be used to support the reuse in scientific experimentation process.

1 INTRODUCTION

Over the last decades, science has explored new possibilities for scientific experimentation. Complex phenomena have been simulated by supercomputers using computational tools. Nowadays, a new paradigm arises to support scientific experiments, which are focused on the open and data-intensive science. In this paradigm, the information is the main product. It is also critical that scientists can share information with other members of the community, as well as reuse data from their colleagues. However, the reuse of knowledge gained in experiments, produced by third parties, is still a challenge. Federer et al. (2015) list several reasons why researchers still do not share data. In addition, even if the data is shared, it needs to be interpreted appropriately by different research groups. The proper interpretation of the shared data set requires that this set be complemented by descriptive metadata (Missier et al., 2010).

This collaborative and distributed scientific experimentation scenario also requires that social and organizational aspects be considered, since the knowledge about the way experiments are performed can be tacit and often remains with the researcher. Thus, storing and retrieving contextual information during the experimentation process may be critical if its activities are performed in order to be reproducible and reusable (Mayer et al., 2014).

Considering these challenges, information about the context and provenance of scientific experiments plays a key role. Provenance information describes the origin, derivation, ownership, and history of the data (Lim et al., 2010). Context is a complex description of shared knowledge about physical, social, historical or other circumstances within which an action or an event occurs (Rittenbruch, 2002). In scientific experimentation domain, we consider provenance information as a kind of contextual element that describes information in the past. Thus, this information is fundamental so that researchers can understand, reproduce, examine, and audit the results previously obtained by experiments, as well as reuse an experiment or parts of it.

Provenance management has been widely discussed in the scientific community (Simmhan et al.,

Ambrósio, L., David, J., Braga, R., Campos, F., Ströele, V. and Araújo, M.
Using Context Elements and Data Provenance to Support Reuse in Scientific Software Ecosystem Platform.
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018) - Volume 2*, pages 255-262
ISBN: 978-989-758-298-1

2005; Lim et al., 2010). In some Scientific Workflow Management Systems (SWfMS), such as Kepler[1], Taverna[2] and VisTrails[3], provenance information is automatically captured. However, in general, their proprietary models make it difficult to share information. Other provenance approaches (Costa et al., 2014; Cuevas-Vicenttín et al., 2014) cover only the capture of information from scientific workflows in isolation to the experiment. In each of these approaches, information may be available at a specific abstraction level, which may or may not be appropriate for the type of analysis required in the experiment context (Missier, 2016).

On the other hand, the use of contextual information in scientific experimentation is an incipient topic. Brézillon (2011), for example, presents a contextual approach to support researchers finding correct scientific workflows in the repository. Mayer et al. (2014) suggest an ontology-based model to describe scientific experiments facilitating their reuse and reproducibility. These researchers describe specific approaches, but do not provide guidelines that can support context management in a collaborative and distributed platform.

In the software development scenario, one of the approaches used to deal with the need of collaboration and distribution in a heterogeneous environment is Software Ecosystems (Manikas, 2016). In the context of this work, a software ecosystem consists of relationships among suppliers of scientific software, research institutes, development agencies, funding institutions and stakeholders to provide and reuse research result, supported by a technological infrastructure (Freitas et al., 2015; Manikas, 2016).

In this vein, Freitas et al. (2015) created the E-SECO (E-Science Software ECOsystem) platform. This platform can manage and support all stages of scientific experiment life cycle. However, it does not support provenance and context management to help in the reuse of scientific experiments.

Therefore, considering the E-SECO platform this work proposes an architecture for the management of provenance and context information that helps researchers to understand scientific experiments and reuse them. In order to reach this objective, this work is based on provenance and context models proposed by Missier (2016) and Brézillon et al. (2004) respectively. The main contribution of this work is the specification of a metadata management architecture, named ContextProv, which aims to manage provenance and context information of scientific experiments in a software ecosystem platform.

This article is organized in five sections besides Introduction. Section 2 presents Related Works. Section 3describes the proposed architecture. Section 4 presents an initial evaluation of the solution considering the ContextProv architecture use in the E-SECO Platform. Section 5 presents the Final Considerations and Future Works.

2 RELATED WORKS

The management of provenance data as well as context information is not recent in the literature. However, the existing works do not associate context and provenance concepts to experiments reuse in scientific software ecosystems platforms.

ProvSearch (Costa et al., 2014) proposes a provenance management architecture for experiments in distributed environments. It combines distributed workflow management techniques with provenance data management. It also allows provenance data to be captured, stored and queried at run-time. In this architecture, data is fragmented into multiple repositories of provenance in the cloud which can be accessed by different SWfMSs. PBase (Cuevas-Vicenttín et al., 2014) is a scientific workflow provenance repository that uses the ProvONE ontology (Cuevas-Vicenttín et al., 2014), allowing the storage, analysis and replication of scientific experiments.

Brézillon (2011) presents an approach that uses contextual graphs to support researchers when reusing scientific workflows. This approach helps researchers to find a workflow through a long process of contextualization (identifying the published workflow that has a context close to the desired one). In addition, it supports the decontextualization, which extracts parts of the workflow that can be reused in a relatively generic way, and the recontextualization, which develops workflow instances adapted to new contexts. TIMBUS context model (Mayer et al., 2014) is a model for the description of scientific experiments focused specifically on the technical infrastructure used as the basis for the experiment. It aims to preserve the processes, the architectural principles, and the core ontologies to extend the experiment, allowing its reuse and reproducibility.

There are other approaches that deal with provenance or context management in scientific experiments. However, these approaches deal with contextual or provenance information in isolation. They do not consider both concepts in a distributed and collaborative context, as we do in our work. In addition, they treat specific problems, and consider provenance

[1]https://kepler-project.org
[2]http://www.taverna.org.uk
[3]https://www.vistrails.org

and context concepts focusing only on the workflow and its results. They do not address the whole process of planning, design, and execution of an experiment and its related workflows. As a result, these approaches are not able to support activities throughout the scientific experimentation life cycle, using a scientific software ecosystem platform.

3 PROVENANCE AND CONTEXT IN SCIENTIFIC EXPERIMENTS

In this section, E-SECO platform and the proposed approach for provenance and context management in scientific experiments are presented.

3.1 E-SECO Ecosystem Platform

E-SECO is a software ecosystem platform developed to support activities carried out during the life cycle of scientific experiments. The key modules of this platform have already been developed and evaluated in e-Science domain, and are illustrated in Figure 1. *E-SECO Development Environment* is a web component where E-SECO code is available, as open source[4]. As a result, the developer community can contribute through software maintenance and evolution. E-SECO relies on a Peer-to-Peer network where different E-SECO nodes can communicate. The ecosystem is made up of artifacts provided by different nodes situated in different institutions, APIs that help the scientific workflow development in its different steps and the open source development environment.

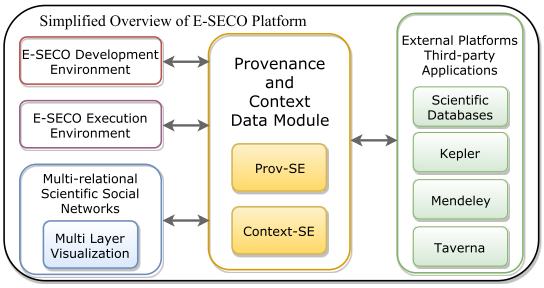

Figure 1: E-SECO Platform Architecture Overview.

The visualization module of E-SECO platform, named *Multi-Layer Visualization*, supports the extraction and analysis of the relationships that are established in scientific social networks. E-SECO platform is integrated with *External Applications* through an Integration Layer. Due to space restrictions, E-SECO platform is not discussed in depth. A detailed presentation of this platform was done by Freitas et al. (2015), Sirqueira et al. (2016).

In order to support reuse during the experimentation process, the ContextProv architecture extends provenance management and adds the management of contextual elements.

3.2 ContextProv Architecture

The verification, reproduction, and reuse of scientific experiments are key activities to support researchers conducting their experiments in a shorten time, and with higher quality. However, these activities are not trivial. Experiments change and evolve over time according to their contexts. As new results emerge, research may follow different approaches and new tasks may arise. As a consequence, planning, modification, or adaptation of the execution process, or even new third-parties' resources are required. Moreover, in a software ecosystem platform, a set of new requirements emerge that need to be fulfilled. As example we can mention, service composition, interoperability and extensibility support. E-SECO platform also aims to fulfill these requirements.

In order to reuse input data produced by a researcher in other context, it is not enough that these data are published on a shared platform. The adequate interpretation requires that these data are complemented by descriptive metadata (Missier et al., 2010). ContextProv architecture is intended to perform metadata management of provenance and context over the experiment life cycle, provinding relevant information to support researchers during the experimentation process.

In previous works, the experiment life cycle has already been extended (Freitas et al., 2015; Sirqueira et al., 2016), but they do not consider contextual and provenance elements. Figure 2 illustrates the experiment life cycle, and highlights some of the contextual elements and provenance information captured throughout this cycle. During the Investigation of the Problem and Prototyping phases, the definition of the scope of the research takes place. In addition, services and workflows to support the experiment are developed and/or reused. In these phases, the elements of the development context and the prospective provenance information are captured. This information represents an abstract specification of the experiment as a guideline for the derivation of future data.

During the Execution of the Experiment and Publication of Results phases, the experiment is carried out in a controlled manner and the results as well as data related to the experimentation process are stored and published. In these phases, context elements

[4]http://pgcc.github.io/plscience/

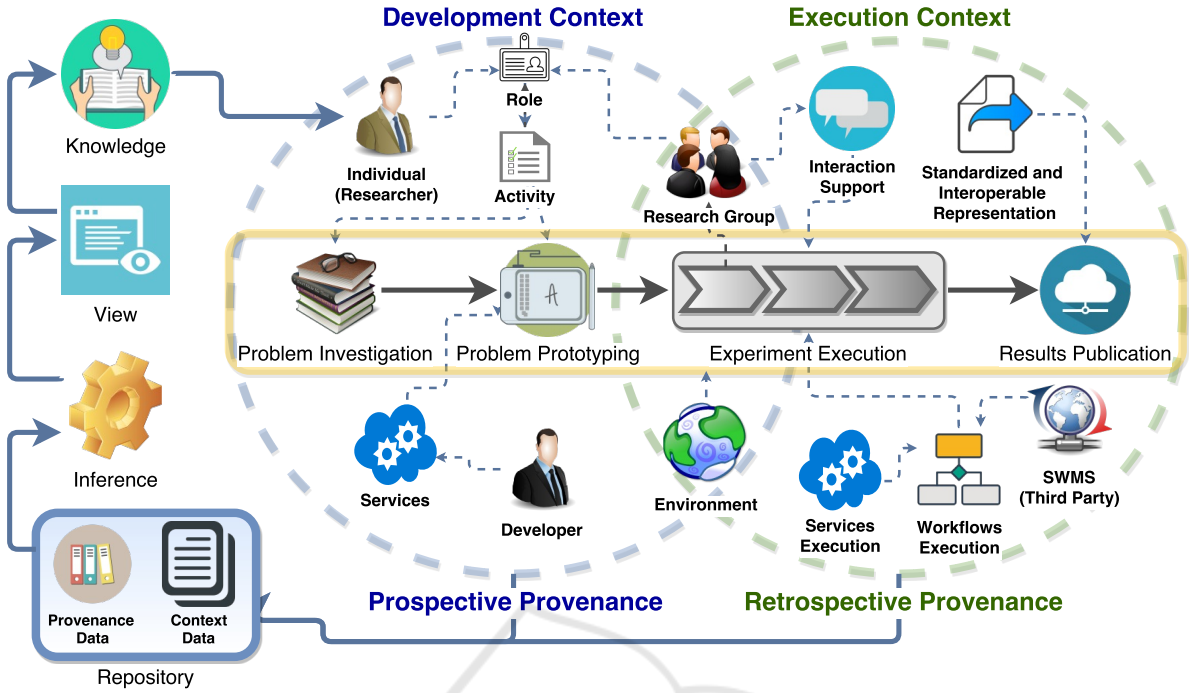

Figure 2: ContextProv Overview.

of the execution phase and the retrospective provenance are captured. They represent, for example, which tasks were performed and how data artifacts were derived. This information helps to verify the results obtained by the experiment.

Throughout the experiment life cycle, contextual and provenance information is captured and stored in a repository. Subsequently, data in this repository are processed by inference machines aimed to extract implicit knowledge about the experiment. Through a visualization mechanism, this information also generates knowledge to support the researchers in future experiment processes. This process of capture, storage, inference and visualization of provenance and context data is done by Prov-SE and Context-SE modules.

3.2.1 Prov-SE Module

It has as its main goal the management of provenance metadata over the experiment life cycle. It also allows the capture and query of provenance data to support the researcher during the experimentation process. In order to describe the steps to manage provenance, Missier (2016) proposed a framework for the provenance life cycle. This framework illustrates the main phases of data provenance until it can be viewed or analyzed. Prov-SE was built based on Missier's model, and includes the capture, storage, inference, query, share and visualization phases.

Initially, provenance information is generated from the experiment specification, during the planning and modeling of the associated workflows. In the execution phase of workflows, data is collected by a Web Service, which is part of the E-SECO ProVersion architecture (Freitas et al., 2015). This service captures information about the experiments execution, such as: start and end time of the execution, input and output data of each task performed by the workflows, and final results.

Some SWfMS, such as Kepler, exports provenance information. Then this data can be imported into E-SECO platform, enriching the information already captured. The data captured are stored in distributed repositories, according to the remote execution of the workflows. They are modeled according to the ProvONE conceptual data model (Cuevas-Vicenttín et al., 2014). Thus, data are stored according to a standard format, aiming to facilitate their interpretation and interoperability across other systems. To deal with a large volume of data, E-SECO platform implements a peer-to-peer (P2P) network. Each node of the network has an E-SECO data repository, which allows the storage of these data in a decentralized but uniform way.

In order to extract implicit knowledge from the provenance data, an ontology was developed, named Prov-SE-O, which is based on the ProvONE model. The main purpose of the original ProvONE ontology is to capture data from scientific experiments related to a specific workflow, its derivations and sub-

workflows. An explicit concern with the distributed nature of the scientific experimentation process is not part of the original ProvONE ontology. So, the distributed nature of scientific experimentation process is considered by Prov-SE-O ontology. Prov-SE-O allows the modeling of scientific experiment as a whole, including the related distributed workflows, capturing important information related to its distributed and collaborative nature. This ontology is detailed in (Ambrósio et al., 2017).

The data captured by the Prov-SE module is loaded into the ontology, and through inference algorithms, implicit knowledge can be derived. For the derivation of this knowledge, Prov-SE-O ontology uses Property Chains[5] and SWRL (Semantic Web Rule Language) rules (Horrocks et al., 2004).

In addition to the P2P network, this module provides a semantically annotated RESTful web service to facilitate data sharing. The P2P network allows the data sharing between different instances of E-SECO platform. In the other hand, the Web Service allows that this data can also be shared with other platforms or external services.

This module also uses a visualization tool based on the PROV model. In this step, the stored data are loaded into the ontology and the visualization is generated from the developed tool (Oliveira et al., 2017).

Information about the provenance of scientific experiments can be considered as a kind of contextual information. However, they are not sufficient to support researchers during the experiment execution. Dealing with a collaborative and distributed activity, researchers need to know the results of the individual work of the group participants. Otherwise, there will be no collaboration, but set of isolated activities (Brézillon et al., 2004). Group work needs explicit context management. Hence, the ContextProv architecture encompasses Context-SE module.

3.2.2 Context-SE Module

This module is based on the framework proposed by Brézillon et al. (2004) which presents groupware mechanisms associated with explicit context representation. Similar to the model used to support provenance, this module supports generation, capture, storage, awareness, interpretation and visualization stages.

Context model was developed in order to determine what part of the contextual information is relevant to scientific experimentation. This model considers the absolute / relative space and time dimensions, context history, subject, and user profile (Bolchini et al., 2007). As a result, it aims to make context management more efficient by storing only relevant information. Annotating all the critical details of an experiment in a laboratory notebook is a standard scientific procedure, especially in the experimental sciences. The key issue in e-Science is that the number and granularity of critical details are high, and identifying them thoroughly is a challenge, and writing all of them is time-consuming. In this way, the Context-SE module has a conceptual framework to identify and classify common contextual elements in a collaborative environment of scientific experimentation.

This framework is an extension of the conceptual framework proposed by Rosa et al. (2003) which considers the relevant elements for context analysis in groupware applications. Context-SE framework modeled these elements in five context categories.

The first category refers to information about group members. This is information about researchers and research groups to which they belong. The second category concerns information about scheduled tasks. In scientific experimentation domain it is related to the planning of the experiment and is characterized by the tasks to be performed by the group until the conclusion of the experiment.

The third category concerns the relationship between group members and scheduled tasks. It relates each researcher or research group to the interactions in which they are involved. This category is divided into two types of contexts: interaction context (information representing the actions that occurred during the experiment execution) and the planning context (information about the project execution plan).

The fourth category brings together information about the environment. It covers both organizational issues and the technological environment, that is, all information outside the experiment, but within the organization that can affect the way the tasks are performed. Finally, the fifth category gathers all the information about the completed tasks. Its purpose is to provide basic information about the lessons learned, whether from the same group or from similar tasks carried out by other groups. It should therefore include all contextual and provenance information about previous experiments.

Ambrósio et al. (2017) present a detailed description of this framework. Following this guidelines, E-SECO platform captures information, such as:

1. **Researchers and Research Groups.** This information helps other researchers identifying who are involved in the experiment. So, it is possi-

[5]Property Chains appeared in OWL 2 and works by sorting objects, where it allows transitivity between multiple properties.

ble to contact these researchers to collaborate on a particular experiment.

2. **Planning the Experiment.** This information refers to the experiment, the workflows and the tasks to be performed, which contribute for the experiment to be reused by other researchers, in a new context.
3. **Relationship between Tasks and Researchers.** This information allows credits to be given to the authors, and that they are responsible or questioned for any errors that occurred during its execution.
4. **Environment.** Information about used technologies, SWfMS, and external services are also essential for the reproducibility of the experiment.
5. **Tasks Completed and Provenance.** Information about the experiments, workflows, and tasks executed are stored, as well as the provenance of them. Thus, it is possible to identify all the processes undergone by an artefact, until the end of the experiment, as well as to identify the workflow and the experiment that gave origin to this artefact, and the involved researchers.

This contextual information can be informed by the researcher, but must be primarily obtained through integration with other platforms, such as Mendeley, which allow the extraction of information about researchers, research groups and institutions.

All contextual information captured by the platform are stored in a repository. In addition, this repository data is connected with related provenance information, which describes the completed tasks, and their associated contexts elements. This information can also be processed by the ontology, allowing the extraction of implicit knowledge.

4 ContextProv IN ACTION

This section presents a feasibility study, with the aim of detailing the reproducibility support, from the use of the ContextProv architecture, in the E-SECO platform. This study is composed of three phases, which detail the activities of experimentation, reuse and reproducibility. During the analysis of this scenario, we tried to answer questions that arise when reusing or reproducing an experiment. These questions are:

Question 1. What is the process responsible for constructing a given result?

Question 2. Is there another similar experiment?

Question 3. Why two similar experiments yield different results?

4.1 Experimentation

An oncologist researching the mutations of the ATRX[6] protein in human tumors wishes to keep up to date with published research related to this protein. His research group uses the E-SECO platform to facilitate collaboration between researchers. Thus, this scientist is part of an E-SECO research community that explores the use of this protein and its effects.

This scientist receives information that the Nuclear Protein Database[7] (NPD), provides a Web Service that searches for information about a given protein, including links to papers published in PubMed[8], Entrez Protein[9] bases, among others. In addition, the E-SECO platform recommends the use of a workflow, available at Kepler website and linked to E-SECO repository, which searches the NPD database and counts the number of published articles on this protein. This workflow, named PapersCount, can be executed occasionally to update the information. The workflow version has specific services that let the use of the ContextProv architecture to allow the capture of contextual information during its execution.

Thus, from the execution of the PapersCount workflow, in the context of the "ATRX protein" experiment, provenance and context information related to each of the workflow executions is generated. This data is stored in one of the E-SECO repositories.

4.2 Reuse

Another scientist, biochemist specialized in studying proteins present in the human body, is part of another research group, related to proteins. This scientist also studies the ATRX protein, and needs to find scientific articles related to this protein. From an initial search conducted by this scientist, the E-SECO platform returns information about the ATRX protein. The biochemist found that there are currently 52 articles on the ATRX protein. Thus, this scientist questioned: (Q1) What process is responsible for this result?

Analyzing the information about result's provenance, available at E-SECO platform he can see that it was obtained through the execution of the PapersCount workflow. Based on this information, the scientist would like to reuse the PapersCount workflow by restructuring only the last task. Instead of returning the number of articles found, would return the links to these articles.

[6]Transcriptional regulator ATRX is a protein that in humans is encoded by the ATRX gene.

[7]http://npd.hgu.mrc.ac.uk

[8]https://www.ncbi.nlm.nih.gov/pubmed

[9]http://www.ncbi.nlm.nih.gov/entrez

However, before reusing this workflow, the biochemist wants to be sure of the reliability of the results. The scientist accessed the provenance information of the PapersCount workflow (Figure 3). Based on this information, he took notice about the origin of the workflow, the qualifications of the researcher who used it and his work institution. Knowing that the authorship of the workflow (Kepler development group) and the scientist who used it as well as his institution, are recognized by the scientific community, the biochemist considered the PapersCount workflow as reliable. So, he decided to reuse it in his research. In this way, a new version of the workflow, called PapersLinks, has been created. In this version, the last task has been modified so that its output is a list of links to the articles found.

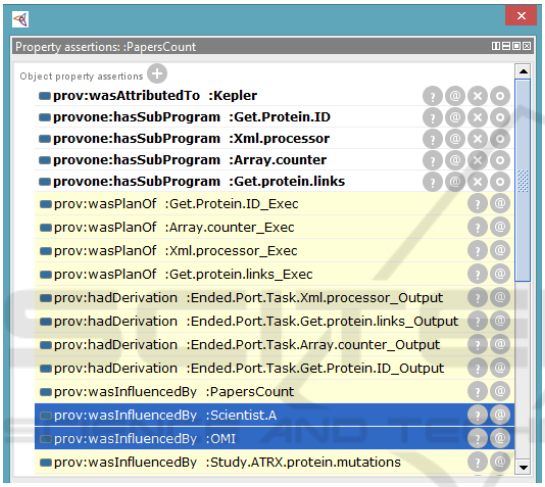

Figure 3: PapersCount Provenance – presented in Protégé.

4.3 Reproducibility

Later, a biologist, a protein researcher connected to the E-SECO platform, needs to make a comparison between the ATRX and ATR proteins. As a first step, he wants to check each one has a greater number of published articles. During the prototyping of his experiment on the E-SECO platform, this scientist conducted a workflow search, where he found the PapersLinks workflow. Analyzing this workflow, he realized that its output was not exactly what he required, as he needed to count the number of articles. Then he wondered: (Q2) Are these experiments similar?

To answer this question, the scientist searched for similar experiments on the E-SECO platform. Through the execution of inferences of the Prov-SEO ontology, the E-SECO platform shown that the PapersCount experiment seems to be similar to PapersLinks because they have common tasks, and use the same input information and the same Web Service. Through the use of the PapersCount workflow, this scientist realized that it returns exactly what he needs. As a result, he decided to replicate the experiment that used PapersLinks, for the two proteins he wants to compare, ATRX and ATR.

When reproducing the experiment for the protein ATRX, the scientist obtained 63 articles. Comparing this result with that obtained in the previous experiment (52 articles), he realized that the results were different. This divergence brings doubt on the reliability of this experiment. So, he need to know: (Q3) Why did the experiments yield different results?

By analyzing the contextual information captured by the ContextProv architecture, this scientist can compare workflow versions, tasks performed, the Web Service used and its version, the SWfMS used, the input and output values of the workflow, the scientist who is responsible for the experiment, and the date of execution. Observing this information, he realized that there was no significant change in the context of the two experiments, except that the first experiment had been executed more than a year. In this way, the scientist realized that there are evidences that the difference in results is due to the new articles published during this period.

So, this scenario shows that the ContextProv architecture can possibly help researchers to answer previously raised questions and others, which usually arise during the reuse and reproduction of scientific experiments. In addition, it is also worth noting that information about provenance and context is essential in this process, and that the use of inference is capable of deriving implicit information that would not be used without the support of ContextProv.

5 CONCLUSIONS

This work presented an architecture to support the management of provenance and context metadata in scientific experiments on E-SECO ecosystem platform. This architecture aims to support researchers in the understanding and reuse of scientific experiments in collaborative and distributed environments. For this purpose, it captures relevant context information and data provenance of the experiments, processes this data in an ontology based on ProvONE model, and thus succeeds in extracting implicit knowledge. This information can be queried and visualized in a standardized way through the ecosystem platform. As a result, it can facilitate its interpretation and provides knowledge to the researchers.

Regarding the limitations of this research, we can point out that the interface of the platform needs to be

evaluated in order to implement improvements for the scientist. For this purpose, we intend to advance experimental studies, evaluating the operation of this platform integrated to third-parties' scientific software ecosystem platform. Furthermore, extensions to other SWfMS databases that have proprietary provenance models also need to be addressed. Results show that the context elements and provenance data could be used to support the reuse of scientific experiments in a collaborative and distributed environment, but they cannot be generalized. Experiments need to be carried out considering the real-world contexts the design decisions of scientists and developers. In the future works, we intend to carry out a formal evaluation of ContextProv architecture through a Case Study at an Agricultural Research Corporation that conducts experiments related to feed efficiency in dairy cattle.

REFERENCES

Ambrósio, L. M., David, J. M. N., Braga, R., Ströele, V., Campos, F., and Araújo, M. A. (2017). Prov-se-o: a provenance ontology to support scientists in scientific experimentation process: Wip. In *Proceedings of the 12th International Workshop on Software Engineering for Science*, pages 15–21. IEEE Press.

Ambrósio, L. M., David, J. M. N., Braga, R., Ströele, V., Campos, F., and Araújo, M. A. (2017). Context-SE: Conceptual framework to analyse context and provenance in scientific experiments. In *14th Brazilian Symposium in Collaborative Systems*, pages 1372–1386. CSBC.

Bolchini, C., Curino, C. A., Quintarelli, E., Schreiber, F. A., and Tanca, L. (2007). A data-oriented survey of context models. *ACM Sigmod Record*, pages 19–26.

Brézillon, P. (2011). Contextualization of scientific workflows. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 40–53. Springer.

Brézillon, P., Borges, M. R., Pino, J. A., and Pomerol, J.-C. (2004). Context-based awareness in group work. In *FLAIRS Conference*, pages 575–580.

Costa, F., de Oliveira, D., and Mattoso, M. (2014). Towards an adaptive and distributed architecture for managing workflow provenance data. In *Proceedings of the 2014 IEEE 10th International Conference on e-Science*, pages 79–82.

Cuevas-Vicenttín, V., Kianmajd, P., Ludäscher, B., Missier, P., Chirigati, F., Wei, Y., Koop, D., and Dey, S. (2014). The PBase scientific workflow provenance repository. *International Journal of Digital Curation*, pages 28–38.

Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J., and Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLOS ONE*, 10(6):1–17.

Freitas, V., David, J. M., Braga, R., and Campos, F. (2015). An architecture for scientific software ecosystem. In *9th Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems (WDES 2015)*, pages 41–48. (in portuguese).

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., et al. (2004). Swrl: A semantic web rule language combining owl and ruleml. *W3C Member submission*, page 79.

Lim, C., Lu, S., Chebotko, A., and Fotouhi, F. (2010). Prospective and retrospective provenance collection in scientific workflow environments. In *Services Computing (SCC), 2010 IEEE International Conference on*, pages 449–456.

Manikas, K. (2016). Revisiting software ecosystems research: A longitudinal literature study. *Journal of Systems and Software*, 117:84–103.

Mayer, R., Miksa, T., and Rauber, A. (2014). Ontologies for describing the context of scientific experiment processes. In *e-Science, 2014 IEEE 10th International Conference on*, pages 153–160. IEEE.

Missier, P. (2016). *The Lifecycle of Provenance Metadata and Its Associated Challenges and Opportunities*, pages 127–137. Springer International Publishing.

Missier, P., Ludäscher, B., Bowers, S., Dey, S., Sarkar, A., Shrestha, B., Altintas, I., Anand, M. K., and Goble, C. (2010). Linking multiple workflow provenance traces for interoperable collaborative science. In *Workflows in Support of Large-Scale Science (WORKS), 2010 5th Workshop on*, pages 1–8.

Oliveira, W., Ambrósio, L. M., Braga, R., Ströele, V., David, J. M., and Campos, F. (2017). A framework for provenance analysis and visualization. *Procedia Computer Science*, 108(Supplement C):1592 – 1601. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Rittenbruch, M. (2002). Atmosphere: a framework for contextual awareness. *International Journal of Human-Computer Interaction*, pages 159–180.

Rosa, M. G., Borges, M. R., and Santoro, F. M. (2003). A conceptual framework for analyzing the use of context in groupware. In *International Conference on Collaboration and Technology*, pages 300–313. Springer.

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *ACM Sigmod Record*, pages 31–36.

Sirqueira, T. F., Dalpra, H. L., Braga, R., Araújo, M. A., David, J. M. N., and Campos, F. (2016). E-SECO proversion: Manutenção e evolução de experimentos científicos. In *BreSci - 10º Brazilian e-Science Workshop*, pages 253–260. CSBC.