Context-SE: Conceptual Framework to Analyse Context and Provenance in Scientific Experiments

Lenita M. Ambrósio[1], José Maria N. David[1], Regina Braga[1], Fernanda Campos[1], Victor Ströele[1], Marco Antônio Araújo[1]

[1]Postgraduate Program in Computer Science
Department of Computer Science
Federal University of Juiz de Fora (UFJF)
Juiz de Fora, MG – Brazil

{lenita.martins, marco.araujo, victor.stroele}@ice.ufjf.br,
{jose.david,regina.braga,fernanda.campos}@ufjf.edu.br

Abstract. *Managing contextual and provenance information plays a key role in the scientific domain. Activities which are carried out in this domain are often collaborative and distributed. Thus, aiming to examine and audit results already obtained, researchers need to be aware of the actions taken by other members of the group. Contextual and provenance information are essential to enhance the reproducibility and reuse of experiment. The goal of this work is to present a conceptual framework that provides guidelines capable of supporting the modeling of provenance and context in a software ecosystem platform to support scientific experimentation. Preliminary results are also presented when the proposed solution is used to design software ecosystem platform components.*

1. Introduction

In scientific experimentation domain, a strong computational tendency has arisen the possibility of sophisticated simulations of complex phenomena. The collection and analysis of a large amount of data is now possible using computational resources. This possibility has enabled new ways of doing science. Science is progressively evolving into e-Science – which unifies theory, experiments and simulation, while dealing with a huge amount of information [Hey et al., 2009].

Scientific experiments can now be simulated by supercomputers, through computational tools like Scientific Workflow Management Systems (SWMSs) (for example: Kepler[1], Taverna[2] and VisTrails[3] that model and execute series of operations through scientific workflows. These tools involve steps of data analysis from various sources and large-scale computing, and require collaboration between geographically distributed scientists [Deelman et al., 2009]. In addition, the experiments undergo changes and evolve over time. As new results emerge, research follows new paths. As a result, planning, modification or adaptation of the form of execution are required, or even new external resources, such as a web service or pre / post processing sub-workflows, are needed [Sirqueira et al., 2016].

[1] https://kepler-project.org

[2] http://www.taverna.org.uk

[3] https://www.vistrails.org

Other important aspects are related to the social and organizational dimensions that affect the conduction of the experiments. Often, knowledge about how the experiments are performed is tacit and remains with the involved researchers. Storing and retrieving this knowledge may be critical to the activities of an experiment succeed. To that end, supporting collaborative aspects, especially of larger experiments, may contribute to the verification, reproducibility and reuse of scientific experiments [Mayer et al., 2014].

Considering the previous 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, they are fundamental so that researchers can understand, reproduce, examine and audit the results previously obtained by the experiments, as well as reuse the experiment or parts of them.

The management of provenance from scientific experiments has been widely discussed in scientific community [Simmhan et al., 2005, Davidson and Freire, 2008, Lim et al., 2010]. ProvSearch [Costa et al., 2014] and PBase [Cuevas-Vicenttín et al., 2014] approaches, for example, allow the management of provenance information in scientific experiments. However, in each of these approaches, source information may be only available at a specific level of abstraction, which may or may not be appropriate for the type of analysis required [Missier, 2016].

On the other hand, the use of contextual information in scientific experimentation is still a incipient topic. Brézillon [2011] presents a contextual approach to support researchers to find the correct scientific workflows in the repository. Mayer et al. [2014] present a model based on ontologies to describe the scientific experiments facilitating their reuse and reproducibility. While recognizing the importance of addressing contextual information in e-Science domain, these researches do not provide generic guidelines capable of supporting provenance and context management in a collaborative and distributed experimentation environment, such as scientific software ecosystem platforms. When we consider these platforms a set of variables need to be related and analyzed. Manikas [2016] define a software ecosystem as the software and actor interaction in relation to a common technological infrastructure, that results in a set of contributions and influences directly or indirectly the ecosystem. The activity of each actor is motivated by value creation both towards the actor and the ecosystem.

In addition, Brézillon [2011] and Mayer et al. [2014] handle context and source information in isolation. They do not associate the source information with those obtained through the contextual elements that support the collaborative activities. In our research, a software ecosystem can be considered as a set of actors who collaborate and interact with a common market by focusing on software and services, along with the relationships between these actors. These relationships are often underpinned by a common technological platform which operates through the exchange of information, resources and artifacts.

This article aims to propose a conceptual framework to support the analysis of contextual information and provenance of scientific experiments, assisting in verification, reproduction and reuse of scientific experiments. In addition, as a secondary objective we aimed to present a correlation between the context framework for knowledge processing in group work proposed by Brézillon et al. [2004] with the provenance life cycle framework proposed by Missier [2016]. Through these frameworks, we expect to better identify what activities are needed for context and provenance management support.

To achieve this goal, a conceptual framework for context analysis in collaborative systems proposed by Rosa et al. [2003] was extended considering the particularities of the field of scientific experimentation. The resulting framework aims to provide guidelines for the management of contextual elements and provenance information from experiments.

The contributions of this work includes (i) the analysis of key stages in the life cycle of contextual information and provenance management, and (ii) the specification of a conceptual framework for the analysis of the context in applications designed for scientific experiments management, named Context-SE. An example of instantiation of the framework is presented considering an ecosystem platform for scientific experimentation.

This article is organized as follows. The next section presents a background including the concepts about context and provenance. Section 3 describes the related work. Section 4 presents the approach proposed in this paper. Section 5 exemplifies the use of the proposed framework. Finally, Section 6 concludes this work and presents future work.

2. Background

Scientific experiment is characterized by a series of interrelated analysis operations, which are modeled and executed through scientific workflows [Goble et al., 2010]. A scientific workflow is a model, or template, composed of services, scripts or other workflows. It represents a sequence of scientific activities implemented by tools to reach a certain goal [Deelman et al., 2009]. Aiming to support researchers during the modeling and execution of scientific workflows, Scientific Workflow Management Systems (SWMSs) have emerged. SWMSs explicitly model the dependency between processes within an experiment and coordinate the behavior of processes at run-time [Belloum et al., 2011].

Considering the current scenario of scientific experimentation and the increasing use of large-scale applications, the management of experimental data is becoming increasingly complex. Metadata describing the data products used and generated by such applications are essential to disambiguate the data and enable its reuse and reproducibility.

Context is a broad concept and applicable in many areas, so it has many definitions relative to the area of knowledge to which it belongs [Bazire and Brézillon, 2005]. . Aimed to fully understand many activities or events which are accomplished, it is necessary to have access to relevant contextual information. Another definition for this concept in context sensitive applications domain is that context is any information that characterizes a situation related to the interaction between humans, applications and the surrounding environment Dey et al. [2001].

Context in a work process has a dynamic nature, where new events arise and new decisions are made. Thus, an organization that does not associate context information to the activities it performs and artifacts it generates has in its organizational memory an huge set of documents with little or no connection between them. Since this memory has no associated context, it is often ignored as an information resource [Nunes et al., 2007].

Considering that the explicit representation of the context, in several dimensions, such as: individual, task and team, brings benefits to support the interaction between group members, Brézillon et al. [2004] proposed a framework containing mechanisms associated to the explicit context representation in collaborative systems. This framework is not domain specific, and provides a representation of the context and awareness aiming to promote an adequate treatment of these concepts when developing collaborative systems used in different domains. It encompasses the following phases: **Generation**, **Capture**, **Storage**, **Awareness**, **Visualization** and **Interpretation**. These phases do not necessarily occur in this order, and together they form a cycle of transformation of the context data into knowledge.

Simmhan et al. [2005] define data provenance as a type of metadata that brings the derivation history of a data artifact from its sources. In scientific experimentation, metadata about the history of derivation of data is essential to ensure the reuse of results obtained through the execution of scientific workflows. In addition, the provenance provides greater understanding and verification of the accuracy and timeliness of the data. In this way, provenance management has been considered a key point in the architecture of SWMSs, and widely recognized in the scientific community. In this context, Lim et al. [2010] consider two types of provenance: a) Prospective, which refers to an abstract workflow specification as a recipe for future derivations of the data; and b) Retrospective, which is related to the capture of information on the execution of workflows and on data derivations.

Currently, the main provenance model is PROV[4], which is the default model recommended by the W3C. PROV defines a model, corresponding serializations and other supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the Web. The PROV makes it possible to represent knowledge about provenance centered on processes, entities or agents. This model consists of a family of twelve documents. Among these, the PROV-O is the most important for our work. This document describes an ontology which expresses the PROV data model (PROV-DM) using OWL2[5]. In data provenance domain, ontologies express precisely the concepts and relationships and provide contextual information.

In order to express specific ontological rules related to the context of scientific workflows, Cuevas-Vicenttín et al. [2014] extended the PROV model and created the ProvONE[6]. This ontology describes the structure of the experimentation process together with its data dependencies, which originate from the execution of the process, covering both prospective and retrospective provenance.

Based on the PROV standard, Missier [2016] proposes the provenance life cycle framework. In this model, the main phases of the provenance life-cycle are: **Capture**, **Store**, **Query**, **Sharing**, **Preserve association to data**, **Visualize** and **Analyze**. These phases occur sequentially to treat the raw data which will produce information that facilitates the analysis by the user. It does not consider collaborative and distributed issues in

[4] `https://www.w3.org/TR/prov-overview/`

[5] OWL is a language for defining and instantiating ontologies

[6] Updated in 2016: `http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html`

this framework, so this model alone is not able to model all phases of provenance in a software ecosystem platform.

3. Related Work

ProvSearch [Costa et al., 2014] proposes a provenance management architecture for experiments in distributed environments. It combines distributed workflow management techniques with distributed 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 and can be accessed by different SWMSs. The provenance data is treated using a standard model called PROV-Wf, an extension of the PROV model for the domain of scientific workflows. However, this approach does not have a solution for data visualization, and is not capable of extracting implicit provenance information.

The PBase [Cuevas-Vicenttín et al., 2014] is a scientific workflow provenance repository that enable scientists to use provenance for the discovery of experiments, programs, and data of interest. This approach supports declarative graph queries and keyword-based graph searching, complemented with ranking capabilities taking into consideration authority and quality of service criteria. It uses the ProvONE model, treating provenance information in a standardized and interoperable way. In addition, this approach has query and visualization capabilities making exploration of this repository easier. Despite using ProvONE, PBase does not use the ontology of this model to make inferences. In this way, this approach does not provide the extraction of implicit information in the captured data.

Brézillon [2011] presents an approach to support researchers in the reuse of scientific workflows. The context is explained through Contextual Charts (CxGs), which are formalisms to represent uniformly all the components of a collaborative process of scientific workflow design. According to the author, scientific workflow repositories contain workflows successfully applied in specific contexts. Thus, no workflow can be reused directly, because a new experiment involves a new context. 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 the part of the workflow that can be reused in a relatively generic way, and the recontextualization, which develops workflow instances adapted to new contexts.

TIMBUS [Mayer et al., 2014] is a context model for the description of scientific experiments. It focuses specifically on the technical infrastructure used as the basis for the experiment. This model was based on the digital preservation of processes, whose objective is to allow the redistribution (re-staging) of a process when the technical environment has changed. Its main objective is the preservation of the processes, the architectural principles and the core ontologies and the extension of the experiment, thus allowing the reuse and reproducibility of the experiments.

As aforementioned, there are already some approaches that deal with provenance or context management in scientific experiments. However, these approaches deal with contextual or provenance information in isolation, not contemplating both concepts in a process of experimentation. In addition, these solutions address specific problems, and thus do not provide generic guidelines capable of supporting development activities for the scientific experimentation process on a software ecosystem platform.

4. The Context-SE Framework

To introduce the conceptual framework of provenance and context in the scientific experimentation domain, the correlations between the phases of the context and provenance life cycle proposed by Brézillon [2011] and Missier [2016], respectively, will be presented. This correlation is important to establish a model that contemplates both the context and the provenance for the development of collaborative systems in a scientific software ecosystem platform.

4.1. Context and provenance in collaborative systems

Context management is a key activity for collaborative activities. The result of the individual work needs to be known to the group participants, otherwise there will be no real joint work, but an incoherent set of isolated activities. In this way, working in a group assumes explicitly managing the context. For the development of collaborative systems, there are several dimensions of contexts in different granularities that need to be considered, such as: the context of the group (e.g. why this group is constituted), the individual contexts of the members (e.g. their origins and known techniques) and the context of the project (e.g., which products to build) [Brézillon et al., 2004].

However, it is not enough to capture data from contextual elements, these data need to be transformed into useful knowledge to the participants of the group. As aforementioned, Brézillon et al. [2004] proposed a framework based on contextual elements for the processing of knowledge in group work. According to the authors, this framework supports the transformation of contextual elements into some functional knowledge. This framework considers different context dimensions, in a cycle that involves several steps, from the data generation to its interpretation by the participants of the group.

As well as contextual information, provenance information has a diversity of application areas. They are intended to describe the steps needed to manage provenance in a generic and domain-free way. To deal with provenance Missier [2016] proposed a framework for the provenance life cycle. This framework illustrates the main phases of a source document until it can be viewed or analyzed.

Considering the provenance as a type of context, we can consider that these two frameworks have similar phases. However, Missier [2016] does not address collaboration issues during the provenance life cycle. To find a model capable of handling both context and provenance in a collaborative and distributed work environment, such as a software ecosystem platform, we established a correlation between the phases of these two frameworks. Next, we describe each of the contextual framework phases proposed by [Brézillon, 2011] and its correspondence with the provenance framework proposed by [Missier, 2016].

Generation: This phase considers that a member contributes to some content to the group. It is considered a user's task, so information about the individual context of this user is collected at this moment. In the provenance model, this phase corresponds to the **Capture** phase, also called **Production**. It consists of observing the execution of a data transformation process, including human-made or partially automated processes.

Capture: This step consists of procedures to collect some physical data from the generation stage. Thus, this step is performed by the system through sensors. The provenance framework does not separate this step from the previous one, thus, both the information generated by the user and those captured by sensors are treated in the **Capture** phase.

Storage: It consists of storing information from the generation phase, according to pre-established conditions. The model of Missier [2016] has also a phase, called **Store**, for the same purpose, however, in this model the information coming from sensors are also stored.

Awareness: This is the phase in which the data collected in the previous steps are processed to be provided to the other participants of the group. In this process the data is transformed, in a summarized or filtered way, aiming to facilitate its interpretation. In the provenance model, this step corresponds to two different steps, such as: **Query** and **Sharing**.

Visualization: At this stage, the information is arranged in the user interface, providing a physical representation of the processed knowledge. The provenance framework also includes a stage, called **Visualize**, for this purpose.

Interpretation: This is a human processing step. It occurs when, the user assimilates the information presented as knowledge, from the information displayed and its individual context. This knowledge is important to generate new contributions, and thus to close the processing cycle of the context. In the provenance framework, this step can be related to the **Analyze** step which encompasses all forms of consumption and exploitation of provenance data that have been captured and made available through data engineering solutions.

Preserve association to data: Besides the previous mentioned phases, the provenance model of Missier [2016] also has this phase that has no correspondence in the context model. However, this is an important phase so that source information is not lost from the original data to which it belongs.

4.2. Framework Overview

The contextual elements in some situations are unstable and unpredictable, which has a negative influence on the identification and the representation of the contextual elements related to group interactions. In order to reduce this impact, Rosa et al. [2003] proposes the use of a conceptual framework aimed to identify and classify the contextual elements most common in groupware tools.

In order to support the selection of relevant contextual information in a collaborative environment of scientific experimentation, we propose a conceptual framework to identify and classify the most common contextual elements in this domain. The objective of the framework is to provide guidelines for the development of collaborative systems considering the context focusing on scientific experimentation in software ecosystem platforms.

For this purpose, we extend the conceptual framework proposed by Rosa et al. [2003]. In this framework contextual information is grouped into five main categories: (i) information about scientists and groups, (ii) information about scheduled tasks, (iii) information about the relationship between scientists and tasks, (iv) information about the environment where Interaction occurs (v) information about tasks and activities already completed. In synchronous environments, group members need to work simultaneously, but in asynchronous environments, there may be a time lag between interactions. The needs of each type of environment are different, so this framework analyzes these situations accordingly.

For each category, context aspects and the provenance of data in the field of scientific experimentation that influence collaborative activities were identified. These new elements were based on information found in scientific platforms such as Lattes[7] and ResearchGate[8], and mainly based on the Prov-SE-O ontology [Ambrósio et al., 2017]. This ontology is an extension of the ProvONE ontology [Cuevas-Vicenttín et al., 2014]. Figure 1 presents the conceptual model of the ontology, highlighting the classes which represent the implemented extension.

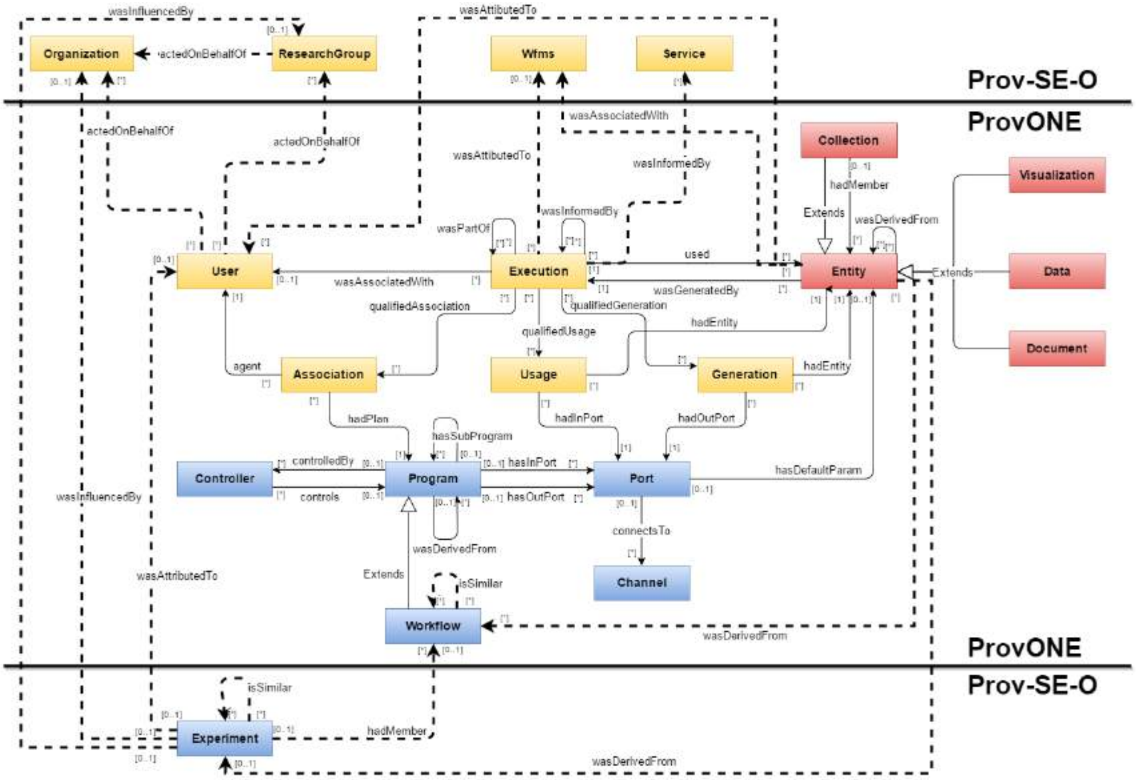


Figure 1. Conceptual model of the Prov-SE-O [Ambrósio et al., 2017]

Prov-SE-O ontology includes new classes, properties and rules in SWRL (Semantic Web Rule Language) [Horrocks et al., 2004]. Thus, it was possible to model not only the workflows, but also scientific experiments. Moreover, we can capture information related to its distributed nature and to support collaboration activities between different agents. Through this ontology, data provenance and context, relevant to the scientific experimentation domain are presented in a standardized way and, at the same time, are capable of being interoperable according to the service to support interoperability avail-

[7] http://lattes.cnpq.br/

[8] https://www.researchgate.net/

able on the platform. In addition, inference of implicit new knowledge can be performed.

Table 1 presents the five context categories modeled by the Context-SE framework and its goals, as well as the associated contextual elements that may influence the group's scientific experimentation activities. The highlighted elements represent the points where the framework was extended.

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 execution of the experiment) 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.

5. Context-SE in Action

Aimed at verifying the applicability of this framework, specifically in the domain of scientific software ecosystems platforms, we decided to analyze the E-SECO platform considering the presented conceptual framework.

5.1. E-SECO Platform

This section describes E-SECO (E-science Software ECOsystem) platform [Freitas et al., 2015] a web-based software ecosystem designed to support researchers' activities during the overall scientific workflow life cycle. The key modules of this platform have already been developed and evaluated in e-Science domain, and are illustrated in Figure 2. During the development process of the collaborative services which support E-SECO activities we have identified the need to enhance this process. So, we consider that it represents an interesting opportunity to enhance the development of services through the proposed framework.

E-SECO Development Environment is a web component where E-SECO code is available, as open source[9]. 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.

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

Information Type	Associated Contexts	Goals	Examples of contextual elements	
Group Members	Individual (Synchronous & Asynchronous)	To identify the participants through the representation of their personal data and profiles.	• Name • Qualifications • Interests • **Degree** • Previous Experience • Location • Working hours • Web page	• **Institution** • **Position (profession)** • **E-mail** • **Awards** • **Skills** • **Languages** • **Publications** • **Research field**
	Group (Synchronous & Asynchronous)	To identify the group through the representation of its characteristics.	• Name • Members • Roles • Abilities • Previous Experience • Geographical Location	• Organization Structure • Working hours • **Institution** • **Web page** • **E-mail** • **Partners**
Scheduled Tasks or **Experiment Plan**	**Experiment** (Synchronous & Asynchronous)	To identify the **experiments** through the representation of its characteristics.	• Name • Description • Goals • Deadlines • Estimated effort • **Tasks** • Restrictions • Workflow ○ **Title** ○ **Version** ○ **SWMS**	○ **Description** ○ **Tasks** • **Similar workflows** • **History** ○ **Evolution To** ○ **Evolution Of** • **Problem Investigation** ○ **Literature Review** ○ **Related Experiments** • **Group in-charge** • **Similar experiments**
Relationship between people and tasks or **Experiment Execution**	Interaction (Synchronous)	To represent in detail the **tasks** performed during the **experiment** completing.	• Group in-charge • Messages exchanged • Presence Awareness • Gesture awareness • **Tasks completed** ○ Author	○ Goal ○ Report ○ **Name** • **Input** • **Output** • **Used services**
	Interaction (Asynchronous)	To represent an overview of the **tasks** performed during the **experiment** completing.	• Group in-charge • Artifacts generated ○ Versions ○ Timestamp ○ **Name** • **Tasks completed**	○ Author ○ Goal ○ Report • **Input** • **Output** • **Used services**
	Planning (Synchronous & Asynchronous)	To represent the Execution Plan of the task to be performed.	• Roles in the interaction • Rules • Aim • Responsibilities	• Strategies • Coordination Procedures • Working Plan ○ **Task Name**
Setting	Environment (Synchronous & Asynchronous)	To represent the Environment where the interaction occurs; i.e., characteristics that influence task execution.	• Quality patterns • Rules • Policies • Institutional deadlines • Organizational structure • **Cultural features**	• Financial constraints • Standard procedures • Standard strategies • **Communication Tool** • **SWMS** • **Geographical Location**
Completed Task and **Provenance**	Historical (Synchronous & Asynchronous)	To provide understanding about tasks completed in the past and their associated contexts.	• **Tasks** ○ Task Name ○ **Group in-charge** ○ Goal ○ Justification ○ Date • Versions of the artifacts • Working Plan	• Contextual elements used to carry out the task • Task Goals • **Input** • **Output** • **Used services** • **SubTasks**

Table 1. Context-SE: Conceptual Framework to Analyse Context and Provenance in Scientific Experiments (extended from Rosa et al. [2003])

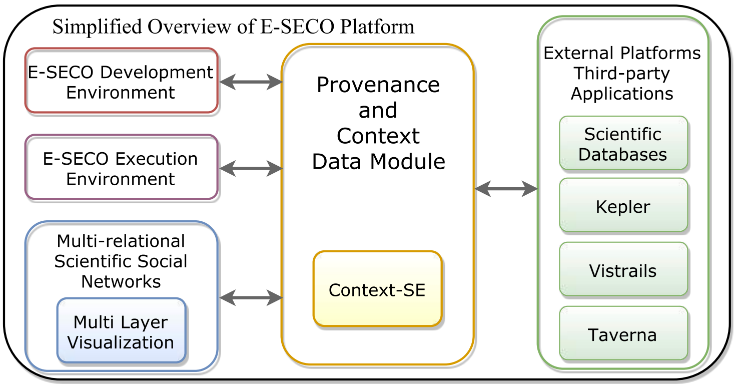

Simplified Overview of E-SECO Platform
E-SECO Development Environment
E-SECO Execution Environment
Multi-relational Scientific Social Networks
Multi Layer Visualization
Provenance and Context Data Module
Context-SE
External Platforms Third-party Applications
Scientific Databases
Kepler
Vistrails
Taverna

Figure 2. Overview of E-SECO Platform

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. 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] and Pereira et al. [2016].

Moreover, E-SECO platform is a collaborative environment to support the development and execution of scientific workflows. It supports systematic literature review processes and it has a repository of existing reports and publications on experiments that are made available to the researcher. In addition, it provides the analysis of data provenance during the execution of workflows in a *Provenance Data Module*.

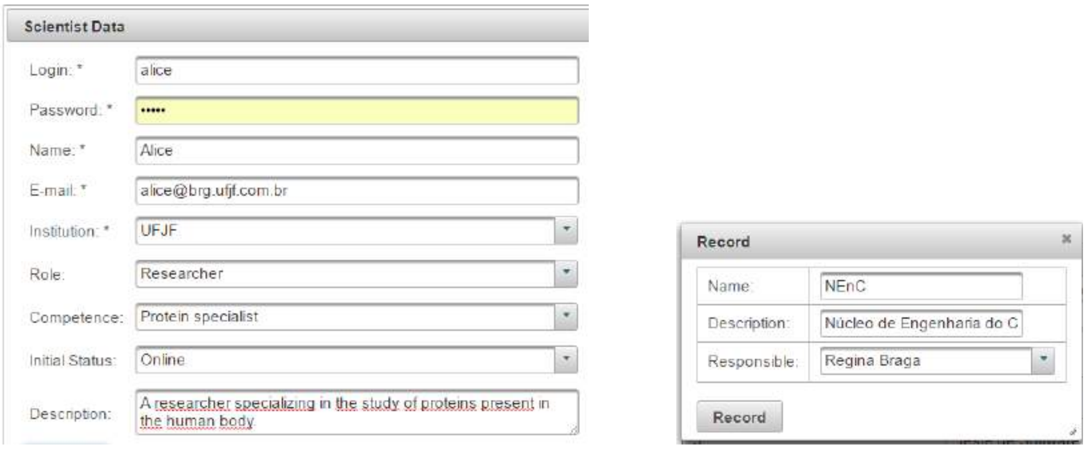
5.2. Context-SE in E-SECO Platform

Considering the characteristics of the frameworks of context and provenance, analyzed in Section 4.2, we can state that they can be applied in the E-SECO platform. At each stage of the scientific experiment life cycle, the contextual and provenance information may be of differentiated importance. However, by space constraint, it is not on the scope of this paper to analyze the effects of using the solution in specific steps of the life cycle. The ways in which the dimensions of the frameworks are contemplated in the E-SECO platform are discussed below.

Generation of data occurs through the user interface, through which the researcher can record information about their profile, group profile or experiment. Data is also acquired through the **Capture** step, implemented by web services run by SWMSs. These services send experiment data to the platform. The E-SECO platform also has a distributed database, implemented in a peer-to-peer network, which is responsible for the **Store's** data. **Awareness** support occurs through the Prov-SE-O ontology, which uses inference mechanisms to make explicit new knowledge about the context and provenance of scientific experiments. **Visualization** is based on the user interface and through a graph visualization tool that follows the conventions of the Prov model. **Interpretation** can be done by the user with the support of the inferences made by the ontology, and the visualizations provided by the platform.

Next, some of the contextual elements identified by the conceptual framework, which are available in this platform, are presented. In addition, the ways in which these elements assist the researcher in the collaborative process of scientific experimentation are described.

Group Members: The register of researchers allows that information such as, name, e-mail, institution, role, skills, and a description of their interests to be made available, as shown in Figure 3(a). Regarding research groups, information such as name, description and responsible researcher, shown in Figure 3(b), is provided. This information helps researchers identify who is involved in an experiment. Thus, researchers can be contacted to collaborate on an experiment or, for example, credits can be given to them, in case of reuse of some artifact.



Scientist Data

Login: *	alice
Password: *	•••••
Name: *	Alice
E-mail: *	alice@brg.ufjf.com.br
Institution: *	UFJF
Role:	Researcher
Competence:	Protein specialist
Initial Status:	Online
Description:	A researcher specializing in the study of proteins present in the human body

(a) User register

Record

Name:	NEnC
Description:	Núcleo de Engenharia do C
Responsible:	Regina Braga

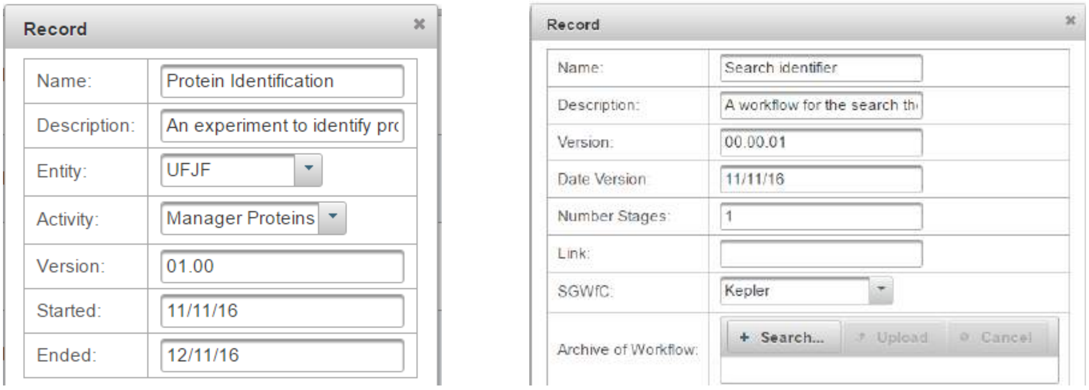
(b) Group register

Figure 3. E-SECO GUI - Member and group registration

Scheduled Tasks: The scheduling of tasks, which in E-SECO corresponds to the planning of the experiment, is done in several steps. First, the register of the experiment is carried out, including information such as: name, expected initial and final dates, version, institution and its description, as shown in Figure 4(a). Next, the workflows involved in the experiment, shown in Figure 4(b), are registered with information such as name, description, version, number of steps, link for download, and the SWMS that will be used. Finally, the tasks planned for each workflow, its name, type and description are registered. In addition, at this stage the system searches for related experiments, and allows the researcher to carry out systematic literature review. This information is important for the experiment to be reproducible in a new context.

Relationship between people and tasks: During the execution of the workflow, a web service is able to capture the information as, the inputs and outputs of each task, the final result, the errors occurred during its execution, as well as the responsible user. In addition, information about exchanged messages and interactions between researchers through the platform is recorded. This information allows credits to be given to the authors, and that they are held responsible or questioned for any errors that occurred during its execution.

Setting: The system allows the identification of SWMS and external services used in the experiment. However, the capture of information about the experimentation environment on the platform is in an initial stage. To tackle this issue, new services,



Record

Name:	Protein Identification
Description:	An experiment to identify pr
Entity:	UFJF
Activity:	Manager Proteins
Version:	01.00
Started:	11/11/16
Ended:	12/11/16

(a) Experiment register

Record

Name:	Search identifier
Description:	A workflow for the search th
Version:	00.00.01
Date Version:	11/11/16
Number Stages:	1
Link:	
SGWfC:	Kepler
Archive of Workflow:	

(b) Workflow register

Figure 4. E-SECO GUI - Member and group registration

or sensors, are necessary to capture this contextual information automatically. They are also necessary for the reproducibility of the experiment to succeed, or for checking the correctness of the results obtained.

Completed Task and Provenance: E-SECO allows to store not only information about the experiments, workflows and tasks performed, but the provenance data of the experiments. Thus, it is possible to identify all the processes of a document, until the end of the experiment, as well as to recognize the workflow and the experiment that gave rise to this document, and the researchers involved. In addition, information based on the Prov-SE-O ontology is stored, such as: the inputs and outputs, similar experiments, or those that were derivatives. The ontology also allows inferences of implicit information to be made on the data origin. This information is essential to ensure the comprehension of this data by researchers as well as the experiment reuse.

Analyzing the E-SECO platform, and considering the proposed conceptual framework, we realize that contextual and provenance information of this framework can be applied in collaborative and distributed platforms of scientific experimentation. Moreover, this information is valuable in order to support the verification, reproducibility and reuse of scientific experiments. It is important to highlight that this platform does not yet have all the elements proposed by the framework, but it is our interest to develop all of them. However, even considering that the proposed solution is based on existing frameworks, and previously evaluated, it is fundamental to carry out additional evaluations. In all these evaluations, we should evaluate the extent in which reproducibility and reuse of scientific experiments are potentialized.

6. Final Considerations

This work presented an analysis of context and provenance frameworks with their specificities in the domain of scientific experimentation. As a result, a conceptual framework was proposed with the aimed at supporting the analysis of context and provenance of data in scientific experiments. This solution also aims to provide guidelines for the modeling of this information in collaborative and distributed scientific experimentation environments. E-SECO platform of scientific experimentation was analyzed from the perspective of the proposed framework. Thus, we could verify the applicability of the solution in the e-

Science domain, and obtain some evidence that the modeled information supports the verification, reproducibility and reuse of scientific experiments. The proposed framework is a first step towards the understanding of the way in which contextual and provenance information can be presented in platforms of scientific software ecosystems.

Regarding the limitations of this research, we can highlight that this framework is still a prototype and some elements still need to be reviewed and improved. As future work, it is important to conduct more complete experimental studies. From these experiments we must evaluate not only the completeness of the proposed solution in a real context of use, but the way in which this framework supported the reproducibility and reuse of scientific experiments in software ecosystem platforms.

Acknowledgements

We would like to thank UFJF, CAPES, FAPEMIG and CNPq for their financial support and encouragement of research.

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. In *Proceedings of the International Workshop on Software Engineering for Science - International Conference on Software Engineering*. ACM.

Bazire, M. and Brézillon, P. (2005). Understanding context before using it. In *International and Interdisciplinary Conference on Modeling and Using Context*, pages 29–40. Springer.

Belloum, A., Inda, M. A., Vasunin, D., Korkhov, V., Zhao, Z., Rauwerda, H., Breit, T. M., Bubak, M., and Hertzberger, L. O. (2011). Collaborative e-science experiments and scientific workflows. *IEEE Internet Computing*, 15(4):39–47.

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., Oliveira, D. 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*, 9(2):28–38.

Davidson, S. B. and Freire, J. (2008). Provenance and scientific workflows: Challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1345–1350.

Deelman, E., Gannon, D., Shields, M., and Taylor, I. (2009). Workflows and e-science: An overview of workflow system features and capabilities. *Future Gener. Comput. Syst.*, 25(5):528–540.

Dey, A. K., Abowd, G. D., and Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-computer interaction*, 16(2):97–166.

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).

Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., et al. (2010). myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(suppl 2):W677–W682.

Hey, T., Tansley, S., Tolle, K. M., et al. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA.

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*, 21:79.

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 (e-Science), 2014 IEEE 10th International Conference on*, volume 1, pages 153–160. IEEE.

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

Nunes, V. T., Santoro, F. M., and Borges, M. R. (2007). Um modelo para gestão de conhecimento baseado em contexto. *XXVII Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*, pages 69–82.

Pereira, A. F., Braga, R., Campos, F., et al. (2016). An architecture to enhance collaboration in scientific software product line. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 338–347. IEEE.

Rittenbruch, M. (2002). Atmosphere: a framework for contextual awareness. *International Journal of Human-Computer Interaction*, 14(2):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. *SIGMOD Rec.*, 34(3):31–36.

Sirqueira, T. F., Dalpra, H. L., Braga, R., Araújo, M. A. P., David, J. M. N., and Campos, F. (2016). E-seco proversion: An approach for scientific workflows maintenance and evolution. *Procedia Computer Science*, 100:547–556.