An Architecture to Enhance Collaboration in Scientific Software Product Line

Anrafel F. Pereira, José Maria N. David, Regina Braga, Fernanda Campos
Federal University of Juiz de Fora, MG, Brazil.
anrafel@ice.ufjf.br, {jose.david, regina.braga, fernanda.campos}@ufjf.edu.br

Abstract

Different approaches have been used to support scientific workflows, including Software Product Line (SPL). A scientific workflow is a mechanism used to represent and perform activities in scientific experiments. However, the design of this mechanism is not a trivial task, and ad-hoc approaches usually create barriers in the domain of e-Science. Lack of support for collaboration issues among scientists can be regarded as one of the main challenges to the scientific endeavour. To tackle this issue Collaborative PL-Science was designed to enhance collaboration among geographically distributed scientists. This paper describes the use of collaborative elements in a Scientific Software Product Line. These elements are designed to create opportunities for collaboration among scientists and contextualize them on their own activities. Case studies were carried out focusing on collaboration elements in order to evaluate the proposed solution.

1. Introduction

Modern Science is increasingly seeking support from computation to conduct experiments. A scientific workflow is understood as a mechanism to represent and perform activities in scientific experiments [2] [3]. However, scientific workflow design and its representation are not trivial tasks, and ad-hoc approaches can create barriers that hinder research activities in this scenario. The lack of support for the reuse of workflow elements that were previously designed by other scientists as well as the absence of control over the evolution of the different versions of the generated products are some examples [2] [8].

To minimize this and assist scientists in solving problems in e-Science domain, the Software Product Line (SPL) approach has been used [3] [14]. Its purpose is to increase quality and productivity in experiments while reducing costs through the reuse of artifacts.

Collaboration among scientists is essential to accomplish experiments [4] [10] [11] and not a new issue. Wulf [16] presents some classical problems in this scenario, such as the lack of services to foster collaboration among researchers. Thus, associated with SPL approach, the use of elements that create opportunities for collaboration in scientific area can be useful for decision support. For example, a history of activities that were performed by scientists can be created. This history can be understood as the steps taken by the researchers when a specific activity was carried out, as well as the decisions that were taken and notes made about the artifacts or features used. This historical information therefore forms a group memory that builds knowledge about the scientific domain. The group memory is useful as it enables scientists to reuse decisions and experiences already carried out by other researchers.

An example is the PL-Science approach, which was proposed by Castro et al [3]. In this approach, authors aimed to support scientists in the selection and use of applications based on scientific workflows. Feature models were associated with ontologies in order to facilitate the development of workflows. However, some questions remain open regarding the need for effective collaboration among scientists in the development of new workflows through the support of a SPL. These include (i) the lack of information about the context of the activities that scientists are accomplishing (ii) the lack of a set of services that can foster interaction among scientists (iii) the difficulty to support systematic reuse of artifacts in e-Science domain (iv) the loss of semantic information when the use of an artifact is finished, among others. The research problem addressed in this paper is the lack of collaboration support between scientists in a Scientific Software Product Line (SSPL) context. The proposed solution aims to provide context information about scientists when designing workflows. As a result, additional knowledge about the considered domain can be built.

From the stated problem, the research hypothesis is the following: *if we offer scientists an architecture that comprises collaboration elements (awareness information, context representation and mechanisms to support communication) in a SSPL, we can create opportunities to enhance and to support collaboration, contextualizing scientists about activities in the shared workspace.*

This article presents an architecture to support scientists during the activity of composing scientific workflows. The workflows presented and used in this paper are related to the bioinformatics area (sequencing/genetic alignment). In this context, DNA sequencing consists of a series of biochemical methods to determine the order of nitrogenous

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.49

bases of the DNA molecule [3][19]. The elements of the architecture are capable of capturing the history of the activities performed by scientists. It also enables queries and new discoveries about these data. It supports the development of awareness elements and the provision of context information to support scientists in their activities while enhancing interaction. In order to fulfill these requirements, the PL-Science [3] approach was extended. A new approach entitled Collaborative PL-Science, was designed and is presented.

This paper is organized as follows. Section 2 presents some background on DNA Sequences. Section 3 presents related works. Section 4 discusses the proposed solution: the Collaborative PL-Science approach. Section 5 presents an evaluation of the Collaborative PL-Science approach in the domain of Bioinformatics, as well as additional contributions. Finally, in Section 6 the conclusions and future works are presented.

2. DNA Sequences and Related Tasks

Our research group works a long time in partnership with bioinformatics research institutions, such as FioCruz and EMBRAPA [3] [12] [13]. Most of our projects are related with projects carried out in these institutions. EMBRAPA, for example, has researches related to Bioinformatics Genomics, with the focus in dairy cattle and its resistance to parasites, in particular, the cattle tick. Therefore, to illustrate our approach, we use examples in the area of bioinformatics. For this purpose, some background in DNA sequencing and related tasks are necessary.

DNA sequencing consists of a series of biochemical methods to determine the order of nitrogenous bases (adenine - A, guanine - G, cytosine - C, and thymine - T) of the DNA molecule. There are several methods available or platforms for performing DNA sequencing, and each of them with advantages and disadvantages. In the Sanger method DNA is used as template to generate a set of fragments that differ in length from each other by a single nitrogenous base. These fragments are separated by size and the bases of the edge are identified in order to recreate the original DNA sequence. The sequencing process using the Sanger method, besides being widely used by research centers, has been used in the execution of various existing genome projects.

Other newer sequencing technologies can be cited, e.g. Next Generation Sequencing (NGS), which began in 2005. Among them the FLX 454 Roche, the Illumina Solexa of the Applied Biosystems platform (called SOLiD System), Single Molecule Sequencing and HeliscopeTrue (tSMS) of Helicos should be mentioned. The aforementioned platforms have greater power to generate information when compared to sequencing using the Sanger method as well as time and cost savings. This efficiency arises from the use of cloning and in vitro systems solid support for sequencing units. Therefore the intensive laboratory labor for the production of bacterial clones, mounting plates and the separation of sequencing fragments in gels is eliminated.

While being different sequencing platforms, both the Sanger and NGS methods have advantages and limitations, such as cost/benefit to obtain and generate the necessary data. The choice of method will primarily depend on the purpose of the work to be developed by the scientist. For example, the major advantage of the Sanger method is the generation of 'reads' higher than in NGS as well as the high precision of the generated base in the base calling process (this process is described below). The main advantage of the next generation technologies (NGS) is the in vitro construction of genomic libraries without amplification of DNA fragments as well as without cloning. Moreover, these sequencing techniques require equipment and reaction kits which are more expensive than in the Sanger method.

3. Related Works

Various authors have discussed collaboration and SPL support in the e-Science context. Mattoso et al. [8], for example, explore the main challenges when composing scientific experiments. The authors highlight the limited support that is usually offered to scientists in the design and subsequent instantiation of scientific workflows and they propose the use of Software Engineering and Database techniques. The key concept of this research is the use of a "Line of Experiments", which is inspired by the concept of SPL. Nevertheless, its main purpose is to explore the possibility of formalizing a standard workflow in order to perform a particular type of experimentation. The work of Mattoso et al. [8] is related to the Collaborative PL-Science approach as it shows the relevance of exploring the challenges of a line of experiments for research enhancement in e-Science. However, this work does not address the use of collaborative elements to support scientific experiments. Unlike Mattoso et al. [8], the Collaborative PL-Science approach helps support scientists to design scientific workflows in order to create opportunities for interaction among them. To do this, it makes use of collaboration elements, such as awareness and communication mechanisms, and context elements associated with a SSPL.

Ivanov et al. [7] describe a way to enhance decision support based on CLAVIRE platform in the e-Science domain. They emphasize the importance of mechanisms to enhance interaction among domain experts and decision makers to solve complex interdisciplinary problems usually arising in scientific workflows simulations. This work is related to Collaborative PL-Science approach when it aims to encourage interactions among scientists in the shared workspace. However, the authors do not address the use of Software Product Line and collaborative elements in the domain in which they are working. Unlike the work of Ivanov et al. [7], one of the main interests of Collaborative

PL-Science is to support scientists, creating opportunities for interaction among them when designing new scientific workflow and not only during their execution.

Miranda et al. [9] describe a tool that allows scientists to share data about the execution of experiments and analyze them together with other scientists, or geographically dispersed research group. The tool presented is called CollabCumulus, and the focus of the working on the execution of scientific workflows. The tool accesses third-party data repositories so that researchers can perform the analysis of data origin collaboratively. The prime focus of Collaborative PL-Science is not on the execution of workflows, but the capture of the decisions taken by scientists during their composition. These steps may be used later as a way to support scientists in developing new workflows. As a result, they also create opportunities for collaboration between them with the services offered by the application.

Other studies have specifically addressed the use of Software Product Line [3] [5]. Filho et al. [5], for example, present an approach for the semantic enrichment of SPL using a top ontology that specifies generic concepts and their corresponding relationships in a SPL. The authors relate this ontology with feature models by conducting a semi-automatic mapping in order to obtain a semantic enrichment of the application. However, this work is not directly related to a specific domain. The authors do not use any historical information in order to support researchers to know the context where artifacts were deployed, making it difficult to reuse the approach later. Thus, when the use of these artifacts is finished, the entire context in which they were used is lost. As a result, both the presentation and use of information about the research are hindered.

As mentioned earlier, Castro et al. [3] present an approach to connect ontology models to features models in SPL in order to provide additional semantics for the design of scientific workflows. The authors show that through the association of these models, additional information for the domain of Scientific Software Product Line can be provided. They present an approach, named PL-Science, which aims to support the specification and execution of scientific experiments in Bioinformatics. In this research, we found that by using features models associated with ontologies much relevant information about the domain could be lost or not adequately delivered to the scientists. Actions performed in the design of a scientific workflow, for example, are made based only on the knowledge and experience of the scientists involved. The lack of awareness support hinders interaction opportunities and effective collaboration among researchers. Additionally, we can mention the difficulty of reusing core artifacts of the SSPL as Castro et al. [3] do not consider information about how the artifacts were used and in what circumstances they were deployed.

The Collaborative PL-Science approach proposed in this paper aims to enhance reuse of artifacts in the SPL core assets using collaborative elements (awareness and context information, and communication mechanisms). As a result, it is expected to provide additional semantics to the scientific field and generate opportunities for interaction among scientists. The lack of a service to capture and provide information about experiments means that scientists' experience can be lost because the knowledge associated with it has not been kept in a group memory, for example.

Zhang et al. [18] present techniques to support the management of provenance and reproducibility in scientific workflows. These techniques are related to collaboration aspects. Based on scientific collaboration ontology, the authors propose a service-oriented model supported by a set of protocols for collaboration. The key challenge is that these protocols are applied to support scientists in the collaborative composition of workflows. They present a tool to support the design and development of these products, named CONFUCIUS.

Our approach proposes an ontology that describes the context in which the experiment is carried out. It is part of the group memory. The primary goal of this ontology is to enable scientists to perform knowledge discovery about the domain. It also aims to support the accomplishment of the design activity of workflows through a Scientific Software Product Line. Furthermore, we believe that through the discovery of new knowledge from the ontology semantic relations and rules, opportunities for collaboration and interactions among researchers can be enhanced. Artifacts of Collaborative PL-Science, such as communication and awareness mechanisms, and context elements, support collaboration.

4. Collaborative PL-Science

The scenario encompassed by Collaborative PL-Science is the domain of Bioinformatics, more specifically the subdomain of genetic sequencing[1]. The approach presented in this paper is an extension of PL-Science [3] considering its architecture and the inclusion of mechanisms that can enhance collaboration among scientists.

In the context of PL-Science, scientists only access artifacts, persisted in the core assets of the SSPL (feature models, ontology mapping files, and so on), to specify a given workflow. As a result, in order to design a workflow, for example, the **PhredPhrap Pipeline** workflow, a scientist could recover the feature model SequenceAligningFeatureModel.xml from the core assets, which can be found in (*http://plscience.superdignus.com/SequenceAligningFeatureModelv1b.xml*). Variation points on this model can be

[1] UML models and scenarios can be reached at *http://plscience.superdignus.com/iaplicacao.html*

chosen based only in the experience of scientist, which was acquired during the composition of similar products. Therefore, the scientist develops the scientific workflow based solely on his/her previous knowledge.

Collaborative PL-Science aims to minimize the constraints found in PL-Science regarding collaborative aspects. To do this the PL-Science architecture was redesigned and some services were developed. They support scientists in carrying out the composition of scientific workflows. Figure 1 illustrates Collaborative PL-Science architecture. The PL-Science approach is presented in the third layer of the architecture. The purpose of PL-Science is to assist scientists in the choice and definition of scientific applications to compose workflows. The layers 1 and 2 are also part of PL-Science approach. However the layer 2 had to be redesigned to meet the new approach, i.e., Collaborative PL-Science. In Layer 3 were added the Artifact and Group Memory Managers so that it could address the gaps not fullfilled by PL -Science. Layer 4 was totally implemented in Collaborative PL-Science approach, being composed of services that provide collaborative functionalities.

To illustrate this several applications can be generated from the similarities and variabilities. These may consist of various algorithms that together compose the final product of SSPL. Products are built from templates, which are stored in the core assets of SSPL. Among these products, we can mention the conceptual models, especially in the context of this work, feature models, ontologies and mapping files [3] [5]. Each mapping file is associated with a specific feature model and ontology. The ontology, stored in the core repository, is used to describe knowledge about the domain. To illustrate the applicability of this study, sequence alignment ontology is used as an artifact to support the development of new workflows.

We believe that by understanding how a workflow has been created, scientists can be aware of the functional (FR) and non-functional (NFR) requirements that make up the developed application. For example, selected features for the development of the product pipeline PhredPhrap are part of their functional requirements. If a scientist prioritizes performance in this workflow, s/he can choose to use another algorithm that was used by another scientist who also developed this product.

In addition, traces left by scientists throughout their activities can help other developers through the use of a common knowledge base. Moreover, previously obtained results can be reused in order to lower the cost of experiments and increase quality and productivity issues. Traces and these functionalities were implement using historical data.

Considering the **Pipeline PhrepPharp** workflow and the operation of the proposed approach, algorithms or web services can be chosen to compose the workflow. To visualize them, scientist A as a domain expert can understand that there are other algorithms that can also be used to perform a specific task in the workflow. For example, instead of using the web service runPhrap (*http://www.biocatalogue.org/services/2268*) to compose the workflow, s/he can use another algorithm or web service that can provide a better result, considering non-functional requirements. Currently information about algorithms and web services suggested to scientists when designing a product are stored as individuals in the ontology that is being used. In this context, interactions with other scientists who work or have worked, making up the Pipeline PhredPharp product, can be started. Through these interactions, they can discuss the inclusion of a new individual in the ontology in order to support workflow composition.

4.1. Collaborative PL-Science Architecture

To tackle the aforementioned challenges, Collaborative PL-Science was designed with specific services to support scientists to accomplish the composition of scientific workflows activities. The architectural elements, which were previously presented in the PL-Science approach, were also used in the architecture of the proposed approach (Figure 1). As a Service Oriented Architecture (SOA) we aimed to encourage among other things: service discovery, reuse, composition, application interoperability and finally the separation between domain business logic and application.

According to Figure 1, in the first layer of the architecture, we can visualize the users of Collaborative PL-Science, i.e. the scientists. They can be geographically distributed and are identified using pseudonyms: Scientist A, Scientist B, Scientist C and Scientist D. In second layer, the web interface of the application is detailed. This layer, named Visualizations, enables future views to be connected to any display tool, such as data mining tools, in order to facilitate the work of scientists. Through this layer, users can interact with the application. An example of this interface is shown in Figure 2, highlighting awareness information, box [A], which presents the latest activities developed by the scientists.

The third layer shows the core assets of PL-Science approach. In this layer, the modules proposed by Castro et al. [3] are shown, as well as the processes involved in the activity of designing scientific workflows. This layer also comprises the functionalities proposed by Collaborative PL-Science, such as (i) the capture of the history of activities carried out by scientists in the application (ii) the storage of data that were captured in the group memory (iii) the awareness and context information presentation about users through pre-defined queries, which were executed in the database that makes up the group memory. These functions aim to contextualize users about what happens in the shared workspace (iv) the services to support interaction between scientists. From these services, scientists can interact and share experiences, answer

questions, or even suggest the correction of artifacts, include new algorithms for performing a task, among others. Interaction information is stored in the group memory. As a result, scientists can know about the notes taken, the artifacts used during that task as well as the scientists who have interacted in the current activity; and (v) the service to support group memory query. Through this mechanism, complex queries can be performed in the ontology, which is part of the group memory. As a result, scientists can have access to detailed information about the domain of the experiment.

Besides these three layers, there is the group memory. It comprises both the relational database and the ontology that represents context information. The relational database was modeled so that it can store essential data related to awareness information about the experiments. Considering that there are different types of awareness information, in this first stage we aim to support the following types: social and workspace awareness [6] and activities awareness [15].

The ontology was modeled focusing the support on the development of scientific workflows. For this purpose, Property Chains specified the rules and restrictions in the ontology. It is a feature available in OWL 2.0, which allows the specification of more complex ontological constraints, without the need to use SWRL. Figure 3 shows some Property Chains specified in the ontology. The first Property Chain establishes a relationship between the artifact and the generated product, highlighting that the artifact is part of the product. The second shows the scientist who is associated with the product that has been or is being built. Finally, the third Property Chain shows that a scientist, when performing an activity, handles an artifact.

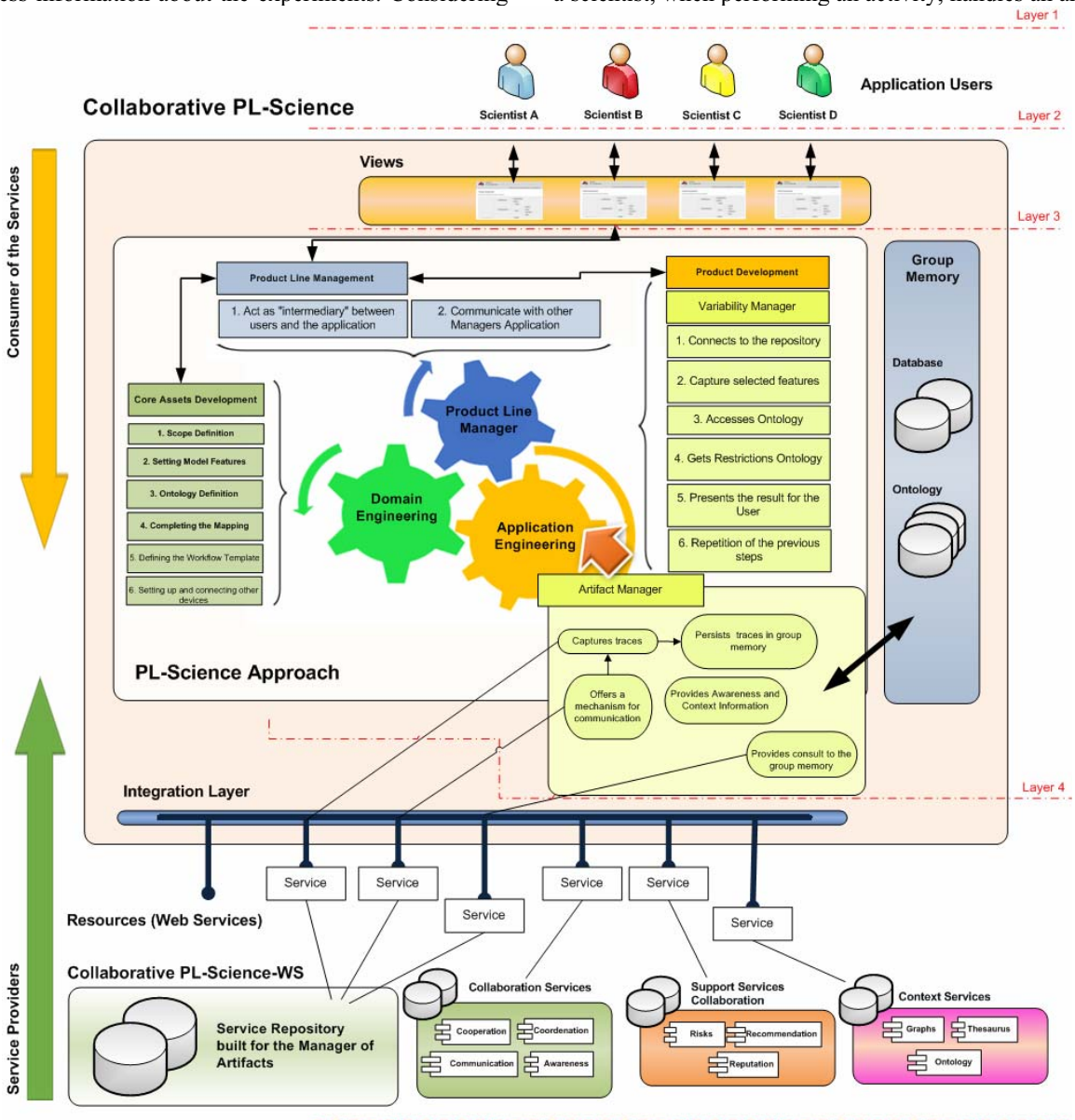

Figure 1. Overview of the Architecture of Collaborative PL-Science Approach.

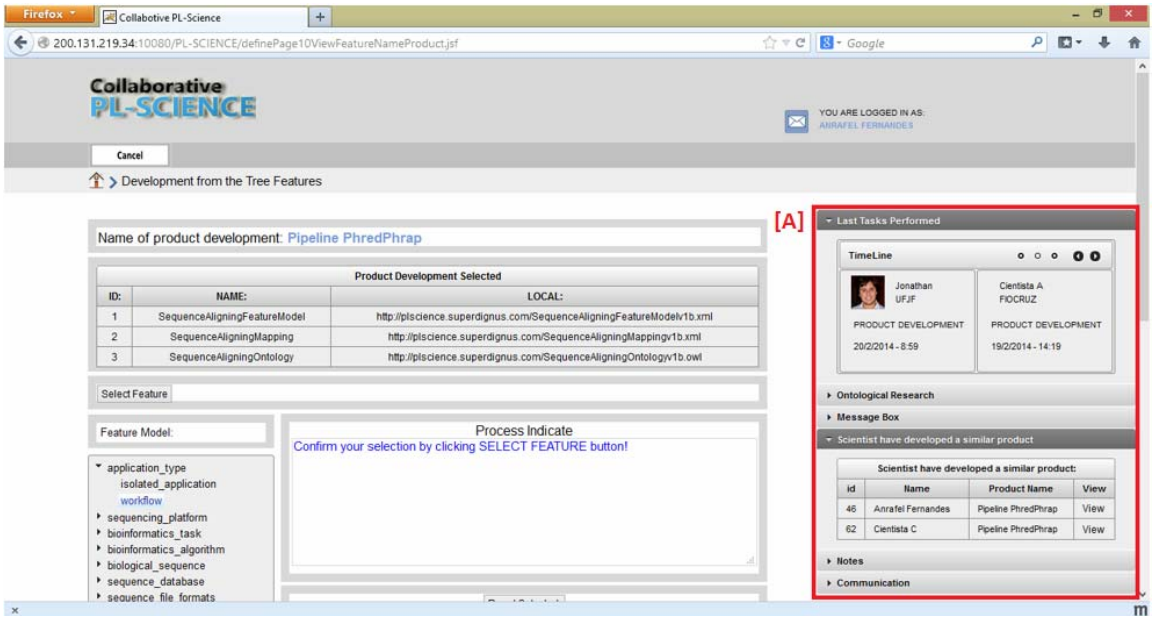

Figure 2. Interface illustrating the Development of Scientific Products.

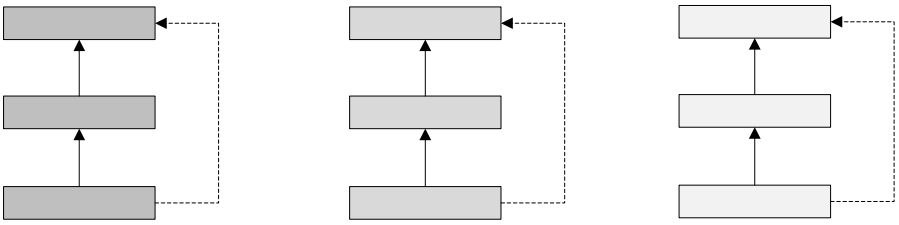

Figure 3. Examples of Property Chains.

Inverse properties of the Properties Chains were also modeled in the ontology. Although not presented in this paper, the inverse properties are also important in this scenario, especially if we need to know the reverse path of a process quickly and formally.

The last layer allows that new services be added to the SPL. Therefore, with an integration layer service, the connection of the processes that are carried out by scientists can be performed. In order to accomplish this task, an Enterprise Service Bus (ESB) was designed. It aims to support SOA governance, security policies implementation, among other activities. Traceability as a requirement is not the focus of our research. However, in each application layer, we have a different abstraction level for the information presented to scientists and captured by the approach. This information is persisted in the group memory mechanism (Figure 4), so that it could support scientists in the design of new workflows. Thus, this feature has the goal to register scientist´ actions at Collaborative PL -Science. However it also helps, indirectly, in the traceability of artifacts used by scientists in the SPL, as well as to products generated by application.

4.2. Collaborative PL-Science Implementation Issues

In this section, some implementation issues related to the approach are presented. The database model is shown in Figure 4. As highlighted in section 4.1 its main purpose is to store the steps taken by scientists during the use of the application, and in addition provide awareness information in order to contextualize scientists about the activities they are performing.

The proposed ontology used in the Collaborative PL-Science architecture is also important to support scientists in knowledge acquisition during the activities of Collaborative PL-Science. The great advantage of its use is the opportunity to process inferences about the context and thus enable the discovery of additional knowledge, or create opportunities for interaction among scientists. For the Collaborative PL-Science Ontology accomplish its role within the application, a data load with the knowledge of relevant information for this context is done in this ontology. This data load includes new individuals so that inferences can be processed and other knowledge about the domain can be presented to scientists. Figure 3 shows the classes modeled for this ontology (indicator 1), as well as the comprised individuals (indicator 2). Using an inference engine (Pellet), the inferences presented on the activity performed to build the Pipeline PhredPhrap product can be obtained (indicator 3). Figure 5 also shows an example of a property chain created in this ontology (4) and the results inferred by Pellet inference machine.

Additional information about the implementation aspects of the Collaborative PL-Science can be obtained at *http://plscience.superdignus.com/*.

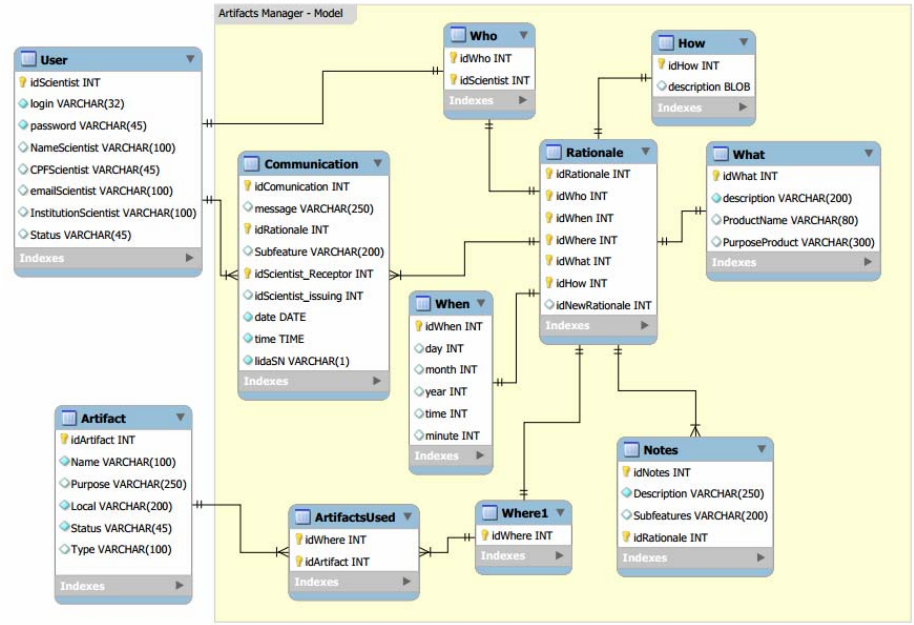

Figure 4. Collaborative PLScience Database

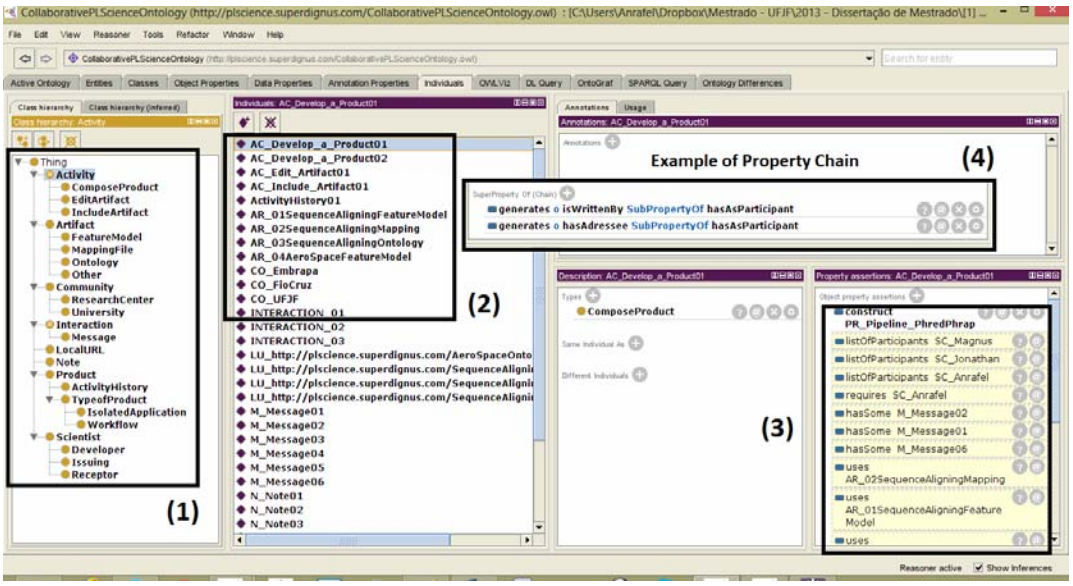

Figure 5. Collaborative PL-Science Ontology

5. Evaluation

In order to verify the feasibility of Collaborative PL-Science some case studies were carried out. For space restrictions, only one of these studies is presented here. It aims to evaluate the benefits offered in supporting scientists during scientific workflows conception as well as when generating opportunities for interaction among them. Other case studies can be acessed on http://plscience.superdignus.com/avaliacao.html.

The case study was considered a suitable choice as a research method, according to Yin [17]. Thus, we evaluate a contemporary phenomenon, considering its 'real-world context' [17].

The evaluation scope was defined based on the GQM method [1] as follows: "*Analyze* the use of awareness and context information in Bioinformatics domain *for the purpose of* characterizing the support offered to scientists and opportunities for interactions generated between them *regarding* the effectiveness of the services offered by the Collaborative PL-Science from *the observer's point of view in the context of* scientific workflows conception to support the alignment/genetic sequencing constructed through a Scientific Software Product Line".

From the defined scope, the hypotheses were derived. The **null** hypotheses were:

(H0): The architecture offered by Collaborative PL-Science does not enhance collaboration among scientists and does not contextualize them about the shared workspace during scientific workflow composition.

The **alternative** hypotheses were defined as follows:

(H1): The architecture offered by Collaborative PL-Science enhances collaboration among scientists during scientific workflow composition.

(H2): The architecture offered by Collaborative PL-Science contextualizes scientists about the shared workspace in which they are interacting during scientific workflow composition.

We used the Collaborative PL-Science approach in the context of one renowned Brazilian research institution, identified as E[2]. The evaluation was applied to a real world context problem when designing scientific workflows in the Bioinformatics Genomics area [12]. This usage scenario was specified, focusing on dairy cattle and resistance to parasites, in particular, the genetic characteristics of the cattle tick. We also aimed to observe the effectiveness of collaborative elements support offered to scientists when using Collaborative PL -Science.

In order to execute the case study, the Collaborative PL-Science application was installed at E[2] institution and a training section was provided. The sciencist W[3], a trainee of E institution and the Collaborative PL-Science developer participated in this training. This last about 1 hour, when all functionalities were explained to W and trainee. After this preeliminary section, Collaborative PL-Science scenario use was scheduled to the next business day.

In the scheduled day, the Collaborative PL-Science application was instantiated for Scientist W[3]. The trainee had a health problem and could not participate in the case study. The Scientist W is a bioinformatics PhD, with a solid formation in computer science. Scientist W is also member of the E research group. He knew that he was engaged in a experiment, but he did not know what was being evaluated. The objective was that scientist W could work on the composition of the scientific workflow and could explore the features offered by the application. The duration of the case study was of approximately 10 minutes because the objective only comprises the conception of a new product and not its execution using a Scientific Workflows Management System.

Another goal of this case study was to evaluate the feasibility of the proposed group memory as a resource not only to maintain records regarding activities performed by scientists, but also to provide awareness information in order to contextualize them about the shared workspace activities. This evaluation also aimed to verify the opportunities for collaboration between these researchers.

After choosing the option of product development and notifying the artifacts that would be used, Scientist W was able to see that another scientist had already developed a similar product. From this information, he could reuse services. After knowing how similar products were built, Scientist W designed the workflow. The first step comprised the selection of the features to build the product. From the awareness information offered to the user from a table in the design of scientific workflows (Figure 2) screen, Scientist W could choose the variation points in the feature model. This was not only based on his previous experience, but also on the knowledge he obtained in the log from other scientists who developed similar workflows activities. Thus, the features chosen by Scientist W to compose the product were: (i) type of application built: feature workflow; (ii) sequencing platform: *feature Sanger*; and (iii) task to compose the product: *sequence_grouping*, *vector_masking* and *base_calling* features.

Considering the operation of the Collaborative PL-Science, the next step was to choose the algorithms and/or web services to compose the final product. In this case study, Scientist W needed additional performance from the workflow that he was developing. Thus, from the awareness information presented by the application, he chose a web service, which was different from that which was used by scientists who had developed similar products.

[2] The organization name was omitted by confidential reasons.

[3] The scientist name was omitted by confidential reasons

When accessing algorithms and web services used by other researchers, Scientist W, as the domain expert, was able to understand from previous experience that there were web services that could also perform the *sequence_grouping* task. Thus, instead of using the *runPhrap* resource, available at: *http://www.biocatalogue.org/services/2268*, to accomplish this task, he chose another artifact, named *runPhrapService* (*http://www.biocatalogue.org/services/1567*), that could also perform this task. It is worth noting a better result was achieved with this service. This allowed Scientist W to know some of the non-functional requirements of scientific workflows that had already been built. Furthermore, results previously obtained by other researchers could be reused in order to reduce the cost of the experiment. Moreover, Scientist W could use the support of the communication mechanism provided by the application. As a result, he had the opportunity to interact with other researchers in the group to discuss the inclusion of web service used by him/her in the domain ontology. After finishing the product design, the log of the activities performed by Scientist W could also support other scientists in the Collaborative PL-Science application in later experiments.

The evidences observed from this case study were made through direct observations on scientists' activities and throughout the process in which the participant developed proposed activities. Some evidences were (i) the adequate support to scientists to maintain them updated on the shared workspace activities (ii) the importance of maintaining a common core of updated knowledge about the domain so that scientists can exploit the database at any time (iii) the adequate interaction support from the awareness information presented by the Collaborative PL-Science and (iv) the decision support given to scientists to build experiments from artifacts which have already been analyzed, commented on and persisted on group memory by other researchers (reuse of knowledge and artifacts from SSPL).

Additionally, it was observed that some PL-Science gaps, previously mentioned in Castro et al. [3] were addressed: (i) the lack of context information in which the activities were accomplished (ii) the absence of a service to enhance interaction between scientists. Initially, a communication mechanism was implemented which supported scientists to exchange experiences and knowledge when carrying out their activities and (iii) the difficulty of reusing core artifacts could be tackled by querying group memory in order to obtain information about how artifacts were used, in which applications, and which scientists had deployed artifacts.

This evaluation has demonstrated the feasibility of the proposed solution. It has also allowed us to observe that the formulated hypothesis can be exploited by further experimental studies. Additional evaluation and the corresponding usage scenario, which describes the development of a product for Multiple Sequence Alignment and a workflow for Genetic Sequencing and Alignment of certain types of cattle, were described in Pereira et al. [12] and Pereira et al. [13] respectively.

Considering the specification and implementation of the proposed architecture, and its use in the above described scenario, the following contributions of this research can be highlighted: (i) the design of an architecture to provide awareness and context information to scientists to enhance interaction between them (ii) the opportunity to allow scientists to make decisions and develop their experiments from the already analyzed, commented on and persisted artifacts (iii) the use of a Service-Oriented Architecture that enables the development, support and maintenance of Collaborative PL-Science and,at the same time, fulfilled the non-functional requirements and (iv) the reuse support of services, which increased productivity and tackled the challenges usually found in the e-Science domain.

6. Conclusions and Futher Work

This paper presented an approach named Collaborative PL-Science. In this approach, an architecture that uses some collaboration elements (awareness and communication mechanisms) and a Software Product Line (SPL) in the scientific area was presented. It aims to support scientists in carrying out activities of composition in scientific workflows. Among such activities, we can mention those that enhance collaboration between researchers in a shared workspace. However, some limitations can still be found, such as those related to (i) collaboration support. The Collaborative PL -Science in this version does not provide mechanism to recommend artifacts. This mechanism could help, for example to recommend resources that best fit scientists' activities and then facilitate the interaction among researchers; (ii) support tools to assist users during the collaborative execution of scientific workflows built from the proposed approach; (iii) the treatment of data provenance. The Collaborative PL-Science approach does not collect provenance data during composition and execution of experiments. Currently, our group is working on it to provide a new provenance module to Collaborative PL-Science context.

Besides that, some tasks need to be undertaken in the domain of Collaborative PL-Science project. For example: (i) to address the use of other mechanisms to support collaboration, such as reputation in collaborative activities, reward, conflict management, among others (ii) enhance awareness mechanism support through the implementation of additional awareness types proposed in the groupware literature, (iii) treat large data volumes, for example, searching alternatives on the use of mechanisms for data summarization and query optimization, (iv) address the connection of the ontology of Collaborative PL-Science with other ontologies. For example, using an ontology network, or algorithms to address the interoperability across ontologies.

Additionaly, future work can be added in this context, specifically to minimize problems in a distributed environment and geographically dispersed work, such as: (i) treating the traceability of the core assets of the SPL, as well as the traceability of services used and added to the architecture; (ii) improve the architecture as regards its scalability (Scale out / Scale up).

Finally, it is important to emphazise that Collaborative PL -Science approach can easily be used in other domains and scenarios. For this to be possible, the artifacts that will be used (feature model, mapping file and the domain ontology) need only to meet the technical specifications. More details on these specifications can be found on the group's research page, or through the members of the research group (*http://plscience.superdignus.com*).

7. Acknowledgments

This research has been partially supported by CAPES, CNPq, FAPEMIG, and Brazil.

8. References

[1] Basili, V. "GQM Approach Has Evolved To Include Models". IEEE Software. V. 11, 1994, pp. 1-8.

[2] Belloum, A., Inda, M.A., Vasunin, D., Korkhov, V., Zhiming Zhao, Rauwerda, H., Breit, T.M., Bubak, M., Hertzberger, L.O. "Collaborative e-Science Experiments and Scientific Workflows", Internet Computing, IEEE, vol.15, no.4, 2011, pp.39-47.

[3] Castro, G.; Braga, Regina; David, J. M. N.; Campos, F. "A Scientific Software Product Line for the Bioinformatics Domain". Journal of Biomedical Informatics, v. 56, p. 239-264, 2015.

[4] Chin, G., Jr., & Lansing, C. S. "Capturing and supporting contexts for scientific data sharing via the biological sciences collaboratory". Proceedings of the 2004 ACM conference on computer supported cooperative work, New York: ACM Press, 2004, pp. 409-418.

[5] Filho, J.B.F., Barais, O., Baudry, B., Viana, W., Andrade, R. M. C. "An Approach for Semantic Enrichment of Software Product Lines". Proceedings of the 16th International Software Product Line Conference - SPLC 2012, Salvador, 2012, pp. 188-195.

[6] Gutwin, C., Stark, G., Greenberg, S. "Support for Workspace Awareness in Educational Groupware". In: Proceedings of Computer Supported Collaborative Learning Conference – CSCL´95, USA, 1995, pp. 1-8.

[7] Ivanov, S. V.; Kovalchuk, S. V.; Boukhanovsky, A. V. "Workflow-Based Collaborative Decision Support for Flood Management Systems". Procedia Computer Science, 2013, pp. 2213-2222.

[8] Liu, J.; Pacitti, Esther; Valduriez, Patrick; Mattoso, M. "A Survey of Data-Intensive Scientific Workflow Management". Journal of Grid Computing, 2015, v. 13, p. 850.

[9] Miranda, G., Souza, J. A., Braganholo, V., Oliveira, D. "CollabCumulus: A tool to support provenance collaborative analyses in Scientific Workflows". Proceedings of the SBSC 2014 Brazilian Symposium on Collaborative Systems. Curitiba, Brazil, 2014, pp. 94-101 (in Portuguese).

[10] Moreira, A., Vieira, V., del Arco, J.C. "Sanar: A Collaborative Environment to Support Knowledge Sharing with Medical Artifacts". SBSC - Brazilian Symposium on Collaborative Systems, São Paulo, 2012, pp. 35-42.

[11] Olson, G.M., Zimmerman, A., & Bos, N. (Eds.). "Scientific collaboration on the Internet. Cambridge", MIT Press, 2008.

[12] Pereira, A. F., David, J. M. N., Braga, R., Campos, F. "An Approach to Integrate Collaborative Elements in the Core Assets of a Scientific Software Product Line". Proceedings of the X Brazilian Symposium in Collaborative Systems. v. 179, Manaus, Brazil, 2013, pp. 16-23 (in portuguese).

[13] Pereira, A. F., David, J. M. N., Braga, R., Campos, F. "CollabPL-Science: Using Collaborative Elements in a Scientific Software Product Line". In e-Science (e-Science), 2014 IEEE 10th International Conference on e-Science. Vol. 2, 2014, pp. 49-54.

[14] Remmel, H., Paech, B., Engwer, C., Bastian, P. "Supporting the Testing Of Scientific Frameworks with Software Product Line Engineering: A Proposed Approach". SECSE '11 Proceedings of the fourth International Workshop on Software Engineering for Computational Science and Engineering, 2011, pp. 10-18.

[15] Dourish, P., Bellotti, V., "Awareness and Coordination in Shared Workspaces". In: Proc. ACM Conference on Computer Supported Cooperative Work (CSCW'92), 1992, pp. 107-114.

[16] Wulf, W. "The national collaboratory". In Towards a National collaboratory. Unpublished report of a National Science Foundation invitational workshop, Rockefeller University, New York, 1989.

[17] Yin, R. K. "Case Study Research Design and Methods", fifth ed., Sage Publications, Beverly Hills, 2014.

[18] Zhang, J., Kuc, D., Lu, S. "CONFUCIUS: A Tool Supporting Collaborative Scientific Workflow Composition", In: IEEE Transactions On Services Computing, Vol. 7, N.1, 2014, pp. 2-17.

[19] Zhanger, F.; Nicklen, S.; Coulson, A.R., "DNA sequencing with chain-terminating inhibitors", Proceedings of the National Academy of Sciences, v.74, p.5463-5467, 1977.