A Foundational Ontology to Support Scientific Experiments

Sergio Manuel Serra da Cruz[1,3], Maria Luiza Machado Campos[2], Marta Mattoso[1]

[1] Programa de Engenharia de Sistemas e Computação (PESC/COPPE-UFRJ)

[2] Programa de Pós Graduação em Informática (PPGI-UFRJ)

[3] Programa de Pós Graduação em Modelagem Matemática e Computacional (PPGMMC - UFRRJ)

serra@ufrrj.br, mluiza@ufrj.br, marta@cos.ufrj.br

Abstract. *Provenance is a term used to describe the history, lineage or origins of a piece of data. In scientific experiments that are computationally intensive the data resources are produced in large-scale. Thus, as more scientific data are produced the importance of tracking and sharing its metadata grows. Therefore, it is desirable to make it easy to access, share, reuse, integrate and reason. To address these requirements ontologies can be of use to encode expectations and agreements concerning provenance metadata reuse and integration. In this paper, we present a well-founded provenance ontology named Open proVenance Ontology (OvO) which takes inspiration on three theories: the lifecycle of in silico scientific experiments, the Open Provenance Model (OPM) and the Unified Foundational Ontology (UFO). OvO may act as a reference conceptual model that can be used by researchers to explore the semantics of provenance metadata.*

1. Introduction

Currently, researchers are facing a proliferation of a huge volume of scientific data. These data are often shared and further processed and analyzed among collaborators. There is a consensus among science data communities that metadata is the foundation for data discovery, use, and preservation. The importance of managing the provenance metadata of scientific experiments is becoming vital to researchers as they have to share results and also consider the long-term usability of data products generated by their investigations.

Provenance captures a derivation history of data products, and is essential to the long-term preservation, to reuse, and to determine data quality (Freire *et al.*, 2008, Moreau *et al.*, 2011). Provenance provides transparency in data acquisition and processing, managing trustworthiness of data sources, allowing those who use the data to determine its validity and to verify its accuracy. For instance, in scientific domains such as Biology, Chemistry and Physics, the tendency toward collaborative scientific process is increasingly evident (Hey, Tansley, Tolle, 2009). Thus, with the proliferation and sharing of such data, questions such as "where did this data come from?", "who else is using this data?" and "for what purpose was it generated?" are becoming increasingly common in the scientific arena. To ensure that data provided by third-party can be trusted, shared and reused appropriately, it is imperative that the semantics of provenance is clearly and precisely defined and made available to users.

Despite of the amount of research papers and surveys about provenance (Buneman *et al.*, 2001, Oinn *et al.*, 2004, Sahoo *et al.*, 2008, Cruz *et al.*, 2009, Mattoso *et al.*, 2010), each work describes it according to a different and particular perspective. For instance, Buneman *et al.* (2001) and Oinn et al.(2004) describe provenance in terms of common data, while others describe it in terms of metadata (Cruz *et al.*, 2009 and Mattoso *et al.*, 2010)),

i.e., as a secondary piece of information that is complementary in some way to the primary piece of information to which it refers. Unfortunately, despite of all these research efforts, scientific data and provenance integration is a problem not completely solved, especially when it involves semantic issues. With such issues in mind, we are interested in the semantics of provenance at a novel perspective when compared to other related works (Salayandia *et al.,* 2006, McGuinness *et al.,* 2007, Sahoo *et al*., 2008, Zhao, 2010).

The goal of this article is to present and discuss the features of a novel ontology named *Open proVenance Ontology* (OvO) which is inspired in three other theories: the lifecycle of scientific experiments, presented by Mattoso *et al.* (2010), the Open Provenance Model (OPM), discussed by Moreau *et al.* (2011) and the Unified Foundational Ontology (UFO), proposed by Guizzardi (2005). We advocate that when binding higher levels of provenance metadata (regarding organization and knowledge about the experiment and its scientific hypothesis) with fine grained operational provenance (collected during experiments´ execution) and the explicit specification of the conceptualizations of the *in silico* scientific experiments domain; unanswered research questions might be solved by exploring the semantics of provenance metadata. For instance, one can perform reasoning about the history of how hypothesis, models, decisions, annotations evolve during recurrent runs of a workflow that is part of an *in silico* scientific experiment. Furthermore, one can explore how an abstract workflow conceived by a researcher is related to a given data product generated at the laboratory of a research partner.

Differently from related works, in this article we present the conceptual model of a well-founded provenance ontology about *in silico* experiments. The purpose of the ontology is to provide a provenance reference model in order to: (i) support the integration of distinct kinds of provenance produced during the lifecycle of scientific experiments; (ii) address semantic interoperability and data/standard integration between experiments; (iii) allow a novel approach for provenance querying; (iv) convey a knowledge repository about scientific experiments results; and, finally, (v) represent the provenance of scientific experiments as a semantic network, exposing it to automated capture and query mechanisms.

The remainder of this paper is organized as follows: Section 2 provides a brief background about the role of provenance in the lifecycle of a scientific experiment; Section 3 introduces the ontological engineering approach adopted in this work; Section 4 presents the OvO ontology; Section 5 discusses related work; and, finally, Section 6 concludes the paper.

2. Background – The role of provenance on *in silico* scientific experiments

In silico is an expression used to mean "performed on computer or via computer simulation". *In silico* scientific experiments are characterized by the composition and execution of several variations of scientific workflows (Mattoso *et al.,* 2009); they are performed with the aid of computer systems and provide researchers a number of advantages, such as: higher precision and better quality of experimental data; better support for data-intensive research and access to vast sets of experimental data generated by scientific communities; more accurate simulations through scientific workflows and higher productivity. However, *in silico* experiments nowadays suffer from an increasing complexity on setting up, maintaining and making changes to the simulation systems and also the shortcomings of managing large data sets of experimental data and provenance metadata.

Like traditional scientific experiments, *in silico* scientific experiments, independent of their domain, follow some common directions (Jarrad, 2001) such as: (i) they need to be

re-executed, enabling other researchers to conduct similar experiments to confirm (or refute) the results; (ii) the results need to be well documented to be used as a baseline for further experiments, conducted by different researchers in different laboratories; (iii) they must follow a protocol or a methodology and be executed under controlled conditions. However, *in silico* experiments are often specified and materialized as scientific workflows.

Scientific workflows are defined in two levels. In a higher level of abstraction, an abstract workflow is described as conceptual model without binding to specific computational resources. In the lower level, a concrete workflow binds programs and data allowing the structured composition of a sequence of operations aiming its execution to achieve a desired scientific result. To be effectively managed the scientific workflows require a specific set of cardinal facilities, such as experiment specification techniques, workflow derivation heuristics, provenance gathering mechanisms and high performance computing environments ranging from private clouds, commercial clouds such as Amazon, GoGrid, Rackspace to supercomputing centers (Stevens *et al.*, 2007, Mattoso *et al.*, 2010).

Manual analysis of the results data set of *in silico* experiments is commonly unfeasible. This involves, for instance, checking input and output data sets of each activity of the workflow, verifying if computations failed on remote computational resources, and checking all activities that contributed to the creation of a particular data set. Many of these activities can be automatically executed by querying provenance metadata.

The treatment of provenance as a first-class data artifact was primarily introduced at OPM by Moreau et al. (2011). The OPM is a way of recording information about artifacts, agents and processes as they occur which includes constructs for representing causal and dependency relationships between sub-processes and the data items or other artifacts that they use or produce. The OPM is a provenance metamodel which is gaining popularity and for which implementations are becoming available in OWL, RDF and Java. Despite of its increasing popularity OPM has some issues, for example, it neither considers all kinds of provenance of in silico experiments nor expresses an unambiguous semantic model.

There are two types of provenance. *Prospective provenance* describes how these scientific workflows were composed and *Retrospective provenance* describes how they were executed and also the relationships between the input and output data sets (Freire *et al.*, 2008). Another important characteristic of provenance is related with its granularity (Cruz *et al.*, 2009), also referred to as provenance level. It is generally classified as coarse or fine-grained. The desirable level of provenance granularity depends on application domain requirements, and the cost of collection, storage and processing. The finer the granularity of the provenance record, the higher the information entropy and associated cost.

According to Mattoso *et al.* (2010), *in silico* scientific experiments can be described as being part of a lifecycle, which consists of three stages: *Composition, Execution* and *Analysis*. Each stage has an independent cycle, taking place at distinct moments of time and handling different kinds of provenance metadata. At the *Composition* stage, researchers either elicit the requirements to build a new abstract workflow as software or retrieve old ones to reuse for new purposes. They start the composition process building the raw version of the workflow by choosing programs incrementally, backward or forward, along the stage. Then, they refine the successive versions towards a concrete workflow.

At the *Execution* stage researchers make their essays by executing instances of a concrete workflow according to their own needs using real parameters and data sets in a production environment. Researchers can also monitor (local or distributed) experiments and register retrospective provenance metadata that can be further used in debugging or

optimization activities, e.g., researchers may optimize concrete workflows, this can be obtained by usage of parallelism and distributed computation. They are also able to initiate an analytical process undertaken over outcomes of collections of experiment runs,

Finally, at the *Analysis* stage researchers can investigate the data products generated by the experiment and then publish results or share not only its outcomes, but also the whole workflow or its parts to other domain experts. This stage may be further decomposed to support the analysis of results. The researchers may face two different situations when analyzing the results: (i) with the generated data, they may conclude that results are likely to be correct, but decide to continue with the experiment, varying parameters and ingesting others data sets, to prove or refute the hypothesis; (ii) when analyzing the data products, researchers discover that their hypothesis is refuted, but faces a new scenario that may lead to a new discover, thus raising a new hypothesis.

In this article we advocate that *in silico* scientific experiments can be fully described as hierarchical trails of provenance metadata with different kinds of granularity collected during its lifecycle. First, at a higher and abstract level, we have the prospective provenance; it cannot be gathered by the existing Scientific Workflow Management Systems (SWfMS); at a lower level we have fine grained provenance collected during the execution and analysis of the experiments.

3. The Ontological Engineering Approach

The first version of the Open proVenance Ontology (OvO) was initially developed by Cruz (2011). The author adopted a combination of two methodologies found in the literature. Firstly, we have employed the ontology engineering process which includes the phases of conceptual modeling, design and codification (Gomez-Perez *et al.,* 2004 and Guizzardi, 2007). The conceptual modeling of ontologies should strive for expressivity, clarity and truthfulness in representing the subject domain at hand. Due to space restrictions, the methodology and the rationale behind of the OvO´s conceptual modeling and design will not be fully discussed in this paper. A detailed description can be found at Cruz (2011).

As second methodology, we have used the ontologically well-founded UML modeling profile proposed by Guizzardi and Wagner (2005). This profile comprises a number of stereotyped classes and relations implanting a metamodel that reflect the structure and axiomatization of a foundational and domain independent ontology named Unified Foundation Ontology (UFO). The UFO was stratified in three ontological layers (fragments), namely UFO-A, UFO-B and UFO-C. A complete description of UFO falls completely outside the scope of this paper. However, we give an overview of the fragments of this ontology which were used in the instances of the modeling profile employed in this article.

UFO-A is the core of UFO, it defines concepts related to *endurants.* An endurant is an entity that is present as a whole at any given point in time, i.e., it does not have temporal parts and persists in time while keeping its unique identity. The endurants (universals and particulars) can be summarized as follows: *Kinds* are rigid, independent sortals that supply a principle of identity for their instances; *Phases* are independent anti-rigid sortals; *Roles* are anti-rigid and relationally dependent sortals, *RoleMixins* are non-sortals that can be partitioned into disjoint subtypes which are sortals. Examples of endurants are a laboratory, an organization, a researcher, an experiment and its phases. UFO-A also uses relations. Relations are entities that connect other entities together, they are divided as follows: Formal relations hold between two or more entities directly, without any further intervening individual. Examples of formal relations include “a laboratory has projects”. Contrarily, material relations have a material structure of their own and have their *relata* mediated by

individuals called relators. For instance, a workflow_run is a relator that connects a concrete workflow and an executor.

UFO-B builds upon UFO-A and defines concepts related to *perdurants*. A perdurant is an entity composed of temporal parts, i.e., its existence extends in time accumulating temporal parts. Examples of perdurants are the codification of a concrete workflow and the composition of an abstract workflow. It is not always the case that whenever a perdurant is present all of its temporal parts are also present. With a perdurant, if we freeze time we can only see a limited number of parts of the perdurant and not the whole. For instance, in a "coding session" of an concrete workflow, if we take a snapshot at the point in time when the programmer is having its code validated in order to execute a concrete workflow, we cannot determine that this task is part of the "execution" of the scientific experiment.

UFO-C defines social and intention-related concepts (both *endurants* and *perdurants*) and is built on top of UFO-A and UFO-B. One of the main distinctions made in UFO-C is between *Agents* and *Objects*. An Agent is a substantial that creates actions, perceives events and to which we can ascribe mental states (intentional moments). Agents can be physical (a football player) or social (a stadium). An Object is a substantial unable to perceive events or to have intentional moments. Objects can also be further categorized into physical (e.g. a ball) and social objects (e.g., money).

4. Open Provenance Ontology[1]

In this section we discuss the key concepts of Open Provenance Ontology (OvO). OvO´s concepts are depicted as UML class diagrams because of the widespread understanding of UML classes and relations and their suitability for our purposes of illustrating how the presented concepts are organized and how they are related to each other. OvO was developed as a set of three sub-ontologies: (i) *in silico scientific experiment sub-ontology*, (ii) *experiment composition sub-ontology*, (iii) *experiment execution sub-ontology*. The sub-ontologies complement each other; they are connected by relations between their concepts as well as by formal axioms.

The UFO stereotypes of the modeling profile used are signed between the sign << >> and the names of the concepts of the ontology (classes) are typed in *italics*. The classes are colored to facilitate understanding. The concepts colored in yellow map prospective provenance and in green retrospective provenance. Due to space restrictions, solely the key concepts are discussed, the complete description of all concepts of OvO can be found at (Cruz *et al.*, 2009 and Cruz, 2011).

4.1. The *in silico* scientific experiment sub-ontology

This sub-ontology captures the structure of *in silico* scientific experiments at a high abstraction level (Figure 1).

An *in silico* experiment is a research that has the purpose of discovering something unknown by adopting a computational model and by the evaluation of a hypothesis that involves the design and implementation of an *in silico* experiment. *Experiment* and *Project* are labeled as <<Kind>> that is a rigid sortal universal that can be identified in all possible worlds. We consider *in silico* experiments as a composition of three essential parts: the model, the hypothesis and project (all are represented by the *essential=true* tagged value in the part-whole relation notion); this relationship indicates that an experiment is made of parts and inseparable relationally dependent. *Hypothesis* and *Model* are also rigid sortal

[1] Only key concepts and relationships were illustrated and discussed.

universals, represented as disjoint concepts tagged as <<Kind>>. The models do not exist outside the context of the experiment as a whole. They are inseparable parts of the scientific experiment they comprise (represented with the mark *inseparable=true*). The *Model* defines the limits of a scientific investigation, recording the circumstances surrounding an experimental study. The *Hypothesis* is a proposition that is stated in an attempt to verify the validity of a provisional response.

Project refers to the research project in which the *in silico* experiments are conducted. A project defines the location of the research and the involvement of several types of researchers. *Project* is represented as an hierarchy. The *Laboratory* and *Organization* concepts are rigid sortals and are related by aggregation and composition with the underlying concepts. The *Laboratory* uniquely identifies the points in geographical space where a research project is designed and where the *in silico* experiments are conceived.

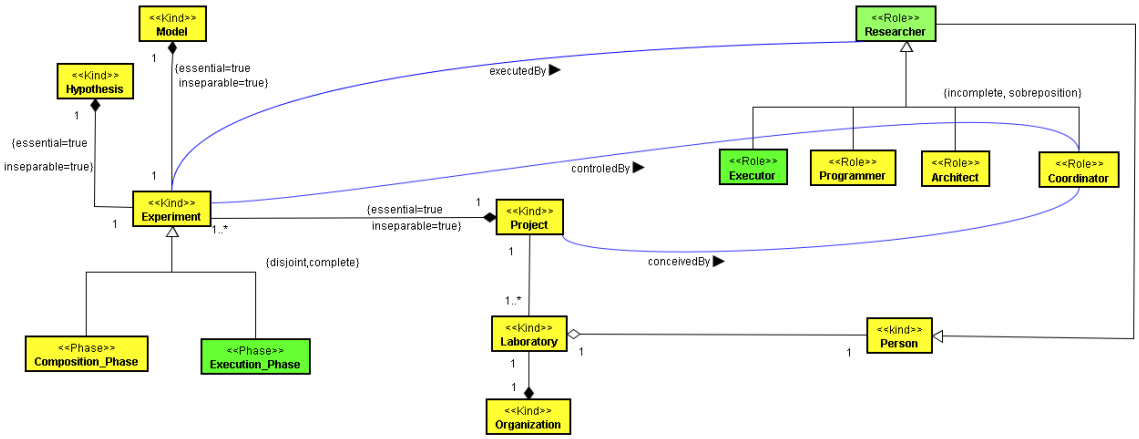

Figure 1. Fragment of the *in silico* scientific experiment sub-ontology

The *Organization* is a rigid sortal universal which plays an important role in the sub-ontology. For example, COPPE/UFRJ is an organization that has several laboratories (*e.g.*, databases, software engineering, among others) and they have material and human resources that can be allocated to conduct a research project. *Laboratory* and *Organization* are independent, but complementary, one can change the members of an organization (university, laboratory) without losing its identity principle. The *Laboratory* also has another important feature, it stresses the relationship with *Person* as an aggregation (tagged as a rigid sortal). Here, we assume that being a researcher is an extrinsic property of a person, *i.e*, there are worlds in which a person is not a researcher and however, he still remains a person. *Person* is regarded to any human being, however, only those who plays any role in conducting a scientific experiment are mapped as *Researcher*. To represent the possible roles played by people whose main attribute is to be a researcher, we adopt the stereotype <<Role>> of the UFO-A, representing them as anti-rigid and relationally dependent sortals. Each instance of researcher must be an instance of person who inherits its identity principle. The concepts *Executor*, *Programmer*, *Architect* and *Coordinator* are sub-types of *Researcher*.

Returning to *Experiment*, it is tagged as rigid sortal and its graphical representation is the same as a generalization of the UML metamodel. However, we can notice different semantics. For instance, in UML, the classes to a generalization are necessarily disjoint and by default do not form a partition. Taking into account the lifecycle of *in silico* experiments (as described in Section 2) and the theory behind UFO-A. An experiment may be represented as a disjoint set (label *{disjoint, complete}*) of stages. It can be decomposed by

two different concepts: *Composition_phase* and *Execution_phase*. Both are tagged as <<Phase>> which means that each stage of the lifecycle is an independent anti-rigid sortal that happens in different points in time and represents a condition that depends solely on its intrinsic properties.

4.2. The experiment composition sub-ontology

This sub-ontology represents *prospective provenance* associated to the composition stage of an *in silico* experiment (Figure 2). The concepts related to prospective provenance are yellow colored while the ones related to retrospective provenance are green.

The composition stage of *in silico* experiments comprises all tasks of specification and modeling of abstract and concrete workflows and their activities. During the modeling task, we capture the knowledge related to the materialization of the experiment and the design of the scientific workflows. At the highest level of the metamodel there is the concept *Composition_Phase*. Such concept is tagged as anti-rigid sortal. By this kind of representation we ensure the unique identification of each cycle of composition in the life cycle of the experiment executed in a given organization.

The association between *Composition_Phase* and *Workflow* has an important semantic explanation; it allows the specialization of the two types of workflows found on *in silico* experiments: abstract and concrete workflows. *Workflow* is tagged as <<Complex event>>, which means that such kind of event is composed of other events by means of event composition operators.

UFO-B events are possible transformations from a portion of reality to another, i.e., they may change the reality by changing the *state of affairs* from one (*pre-state*) situation to a (*poststate*) situation. These are complex entities that are constituted by possibly many endurants. Situations are taken to be synonymous to what is named *state of affairs* in the literature, *i.e.,* a portion of reality which can be comprehended as a whole. According to our metamodel, the concepts *Workflow_Code* and *Workflow_Description* are tagged as <<Kind>>, such approach allow us to uniquely identify the versions from one prior-version of a workflow (*pre-state*) to a novel-version (*poststate*) of a workflow.

Complex events are aggregations of at least two perdurants (either atomic or complex events). Perdurants are ontologically dependent entities, *i.e.,* perdurants existentially depend on their participants in order to exist. The *Concrete_workflow* only exists with the participation of the *Workflow_code*, the *Concrete_activity* and the *Atomic_Concrete_Activity* (the substantial in the model). Similar arguments can be used for *Abstract_workflow*.

Concrete_activity and *Abstract_activity* are also complex events. For instance, *Concrete_activity* is modeled using the weak supplementation pattern (Guizzardi, 2005). The pattern represents a parthood where the hollow diamond is connected to the whole (*Complex_Concrete_Activity*) and the aggregation is supposed to be a irreflexive and anti-symmetric relation.

The specialization of the concept *Researcher* (discussed in section 4.1) in distinct roles, *i.e.* relationally dependent sortals, is crucial to expose the different duties of the members of a research team during the lifecycle of the experiment. For example, the architect of the experiment is a highly trained domain researcher who is in charge of planning, design and oversight/supervision of the experiment, he may not worry about the software packages versions involved, technical and operational details needed to build concrete workflows. His duties are related to the composition of the sequence of activities of an abstract workflow, while the programmer is related to technicalities and the

production of executable codes. He is the person who writes concrete workflow as computer software. The roles represented by *Architect* and *Programmer* are tagged as <<Role>> having two explicit relationships. The first refers to actions that happen in time, the actions *Coding_Session* and *Design_Session* are tagged as <<Action>>, the second relationship involves the rigid sortal *SWfMS*.

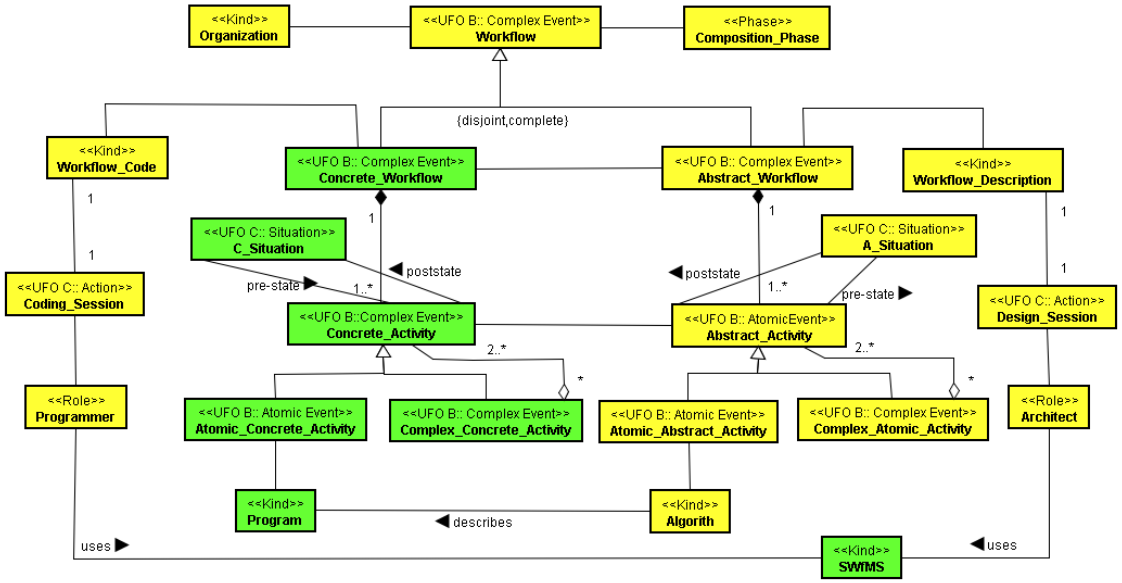

Figure 2. Fragment of the experiment composition sub-ontology

The semantics of *Coding_Session* and *Design_Session* exposes different events that happen in time (the act of coding a concrete workflow or designing an abstract workflow respectively). An *Action* is a UFO-C perdurant that is an individual instance of an *Action Universal*, with the purpose of satisfying the propositional content of an intention. Actions are intentional events, *i.e.*, events which instantiate a plan with the specific purpose of satisfying some internal commitment of entities capable of bearing intentional moments. *C_Situation* and *A-Situation* are tagged as UFO-C <<Situation>>.

4.3. The execution experiment sub-ontology

This sub-ontology represents *retrospective provenance* associated to the execution stage of an *in silico* experiment (Figure 3). The concepts related to retrospective provenance are green colored.

At the highest level there is *Execution_Phase* tagged as a universal sortal <<Phase>>; such representation ensures the unique identification to each execution of a concrete workflow during the lifecycle of the experiment. During the execution stage, the researcher can execute different instances of a concrete workflow. Therefore, the specific workflows of an experiment are associated by composition to the execution stage.

A concrete workflow comprises of one or more concrete activities. The *Concrete_Workflow*[2] represents the association, by composition, to the concept *Concrete_Activity*. A concrete activity can be understood as a codification of an executable scientific application individually (*Program*) which shall be governed by the principle of unique identification.

[2] In order to simplify the diagram of the experiment execution sub-ontology (Figure 3), the weak supplementation pattern and the *C_Situation* on *Concrete_Activity* were omitted.

The *Artifact* is a piece of data represented by a rigid sortal. For instance, a researcher Diogo uses a resource "FastaFile_TripCruzi_20122010.txt" which is owned by researcher Alberto. All artifacts, individually, must have their own principle of identity and also additional essential properties, such as name, type and date of creation, among others. The artifacts (pieces of data) handled by a specific concrete workflow, can be specialized as: *Input_Data*, *Output_Data* and *Parameter* which are shown as <<SubKind>>.

It is worth mentioning that associations between *Artifact* and *Concrete_Activity* have a particular meaning. The association *usedBy*, *generatedBy*, and *triggeredBy*, *derivedBy* (shown in blue color in Figure 3) are inherited from the OPM metamodel (Moreau *et al*., 2011). The association describes the artifacts which were consumed by a given process (concrete activity), while the second describes those that were produced during the processing of a given activity. Finally, the third represents the sequence of the specific activities of a particular concrete workflow. Strictly speaking, the associations originally defined by OPM have semantic meanings that could be better explained. For example, *usedBy* was plotted as a simple association. However, if we consider a semantically rigorous model, the association between *Concrete_Activity* and *Artifact* can be represented through a third-class R´ type material relator to explain all the details of the semantics of this relationship.

The execution of the concrete activities of a workflow is performed in a computational environment, so it is necessary to expose this association. In Figure 3, we use a combination of *Concrete_Activity* and *Environment*. However, an execution environment is not a monolithic entity, by the contrary; it is represented by an aggregation of the rigid sortals Program and Hardware. We have described the succession of resources involved in the execution of a specific workflow. In this case, Resource is tagged as a non-sortal <<Rolemixin>> type of UFO-A. That is, an abstract class that allows for specialization of other classes and subsequent identification of these resources, as *Hardware_Resource*, *Software_Resource*, and *Researcher*. We decided to use the term *Researcher* rather than, *Human Resource*, since the second term represents the entire set of employees of an organization, while the first represents the subset of employees who are trained and qualified to conduct scientific research. *Hardware_Resource* and *Software_Resource* are tagged as <<Role>>.

Finally, it is important to stress the material relations depicted in Figure 3. The explicit association between *Executor* and *Concrete_Workflow* is represented by an n-ary relationship. In this case, such relationship is mapped as *Workflow_Run*, representing sessions of work performed by the executor during the execution of a concrete model of a scientific workflow. In light of the UFO-A, universal relators are represented by the stereotype <<Relator>> and material relations are represented through associations stereotyped as <<Material>>. The dotted line relationship between a material relationship and the universal relator indicates that the relationship is derived from a material relator. The presence of the graphic symbol (●) on the relator (at the end of the dotted line) is used to distinguish the graphical representation of a class of a simple UML association class. The relationship between the relator *Workflow_Run*, *Concrete_Workflow* and the *Executor* tagged as an anti-rigid <<Role>> is an example that represents the above mentioned material relationship. Textually, we have the following: *Workflow_Run* (X, Y) is true if and only if there is an *Executor* X and a *Concrete_Workflow* Y.

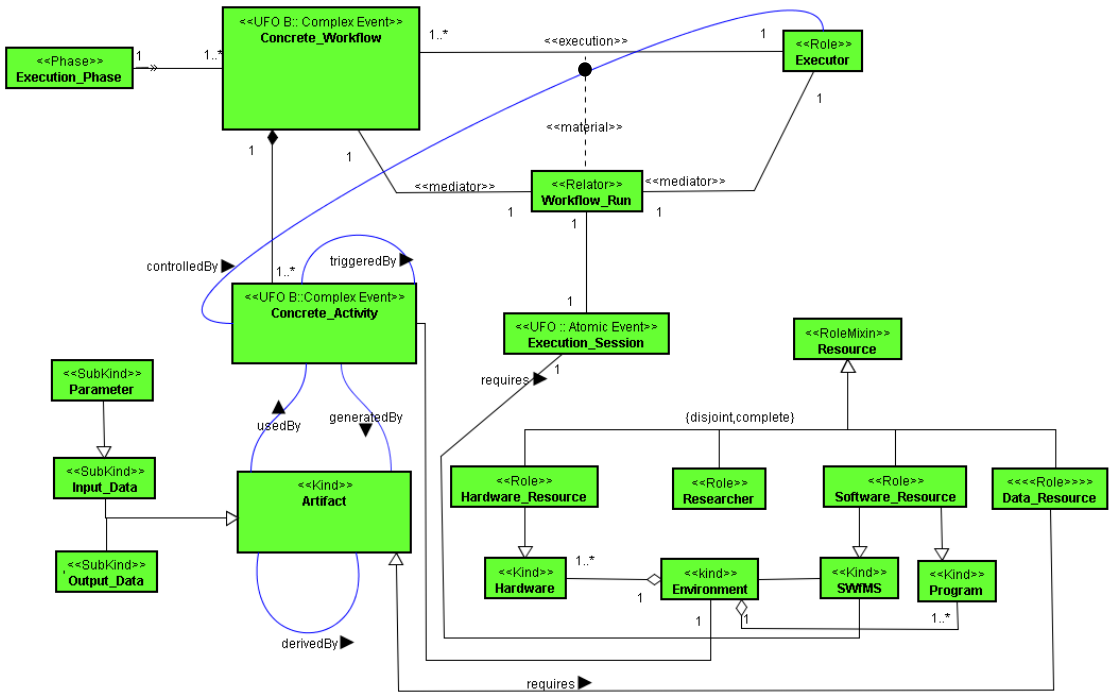

Figure 3. Fragment of the experiment execution sub-ontology

5. Related Work

There are few research initiatives committed with ontology-based models and formalizations of the *in silico* scientific experiments. A number of provenance ontologies have emerged from scratch or through conversion of existing provenance vocabularies. The main features with these provenance ontologies are: (i) they are focused on the provenance of digital objects (especially data) not the experiment; (ii) they leverage only the Web as the underlying infrastructure for data generation and data access; (iii) they encode computational semantics for provenance vocabularies using declarative representations; and finally (iv) they do not take into account a formal method like the one based on foundational ontology principles to validate its development.

The Open Provenance Model Ontology (Zhao, 2010) focuses solely on retrospective provenance of concrete scientific workflows. It defines a small set of core concepts for general entities (*artifacts*, *agents*, and *processes*) and relations in workflows, *e.g.*, "artifact wasGeneratedBy process" and "process wasControledBy agent". With a small number of concepts to represent such a complex domain, it results in concepts being semantically overloaded. It was developed by a community of workflow researchers and has multiple serialization profiles, including XML Schema and OWL. Currently, the OWL profile is still evolving to adapt the OPM specification to be a W3C standard.

The Proof Markup Language (PML) (McGuinness *et al.*, 2007) focuses on information manipulation processes such as logical reasoning, information extraction, and more recently, machine learning. It is modularized into several loosely coupled modules to: (i) annotate provenance metadata for sources of knowledge, (ii) encode information manipulation processes and data dependencies for deriving the conclusions or executing workflows and (iii) annotate trustworthiness assertions about knowledge and sources. PML uses a proof theoretic foundation for its model, its specification is synchronized with its own OWL ontology.

The Workflow Driven Ontology (Salayandia *et al.*, 2006) uses PML to represent abstract workflows, which is a plan for a workflow but not the execution of a workflow. It uses the *Method* concept to represent the class of actions to be executed and uses the *Data* concept to represent the class of data to be operated on by an action in the workflow. The Provenance Vocabulary (Hartig, Zhao, 2010) focuses on information manipulation. It consists of three modules: the core module defines basic concepts for representing data creation and data access processes, and the other two modules extend the core module by (i) adding classifications specific to Web information transfer and (ii) supporting authentication of information. It should be noted that this ontology uses OWL 2 language features. Provenir (Sahoo, Seth, 2009) is an ontology based on information manipulation. It reuses and redefines some provenance relations from the OBO Relation Ontology (Smith *et al.*, 2007), which defines generic binary relations without domain/range specifications.

Provenir, like the previously mentioned works, (re)defines some of the universal OPM classes (*process*, *artifacts* and *agent*) to its own purposes. These ontologies do not consider the role and the granularity of the different kinds of provenance metadata generated during the complete lifecycle of *in silico* experiments. However, as discussed, they tend to be too much driven by both requirements of specific applications and less committed to maximizing expressivity, clarity and truthfulness with regard to the domain of generic *in silico* scientific experiments. Furthermore, as a rule, such initiatives are not committed to the use of top-level ontologies such as UFO and, consequently, do not take advantage of neither its ontological foundation nor its subjacent ontological engineering approach to create clear conceptual models of high expressivity.

6. Conclusion

This article has presented a novel provenance ontology name Open Provenance Ontology. We advocate that the theory behind the OvO provides: (i) a knowledge repository about the different kinds of provenance of *in silico* scientific experiments and (ii) a reference conceptual model that may be used for promoting interoperability between distinct provenance systems.

Our contributions cover a gap in literature with regard to ontological approaches for modeling provenance metadata of scientific experiments. However, there is a need for future research on suitable languages to model phenomena such as the dynamics of the composition and execution of workflows in distributed environments. At this time, OvO is part of a distributed provenance gathering system named Matriohska, its modeling is being reviewed and part of it, implemented as OWL, and is under evaluation with *in silico* bioinformatics experiments at Fiocruz. Further details can be found at Cruz (2001). Additional future work includes: (i) the design and implementation of the OvO ontology in one or more codification languages (*e.g.*, F-Logic) as a proof of concept and (ii) the extension of the ontology for covering as far as possible novel situations.

References

Buneman, P., Khanna, S., W-C Tan, (2001) “Why and Where: A Characterization of Data Provenance”. In LNCS v. 1973 n. 2001 p. 316-330.

Cruz, S. M. S., Campos, M. L. M., Mattoso, M. (2009) “Towards a Taxonomy of Provenance in Scientific Workflow Management Systems”. In: SERVICES '09, p. 259-266.

Cruz, S.M.S., (2011) “Uma Estratégia De Apoio À Gerência De Dados De Proveniência Em Experimentos Científicos” Tese de Doutorado, PESC/COPPE-UFRJ , 234p.

Freire, J., Koop, D., Santos, E., Silva, C. T. (2008) "Provenance for Computational Tasks: A Survey," *Computing in Science and Engineering*, v. 10, n. 3, p. 11-21.

Gomez-Perez, A., Corcho, O., Fernandez-Lopez, M. (2004) “Ontological Engineering : with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. 1 ed. Springer, 2004.

Guizzardi, G. Wagner, G. (2005) “Some applications of a unified foundational ontology in business modeling”. Business Systems Analysis with Ontologies, chapter 13, p. 345–367.

Guizzardi, G. (2005)"Ontological Foundations for Structural Conceptual Models" CTIT PhD.-thesis series, No. 05-74, 441p.

Guizzardi, G. (2007). “On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models”. In Frontiers in Artifical Intelligence and Applications, Databases and Information Systems IV, IOS Press, Amsterdam.

Hartig, O., Zhao, J., Provenance Vocabulary Core Ontology Specification. (2010). Available at: <http://trdf.sourceforge.net/provenance/ns.html>.

Hey, T., Tansley, S., Tolle, K., (2009) “The Fourth Paradigm: Data-Intensive Scientific Discovery”. 1. ed. Redmond, Microsoft Research.

Jarrard, R. D. Scientific Methods, an online book. 1. ed. Dept. of Geology and Geophysics, University of Utah, 2001.

Mcguinness, D., Ding, L., Silva, P. P. (2007) “PML 2: A Modular Explanation Interlingua”. In: Proc of the 2007 Workshop on Explanation-aware Computing, pp. 49-55.

Mattoso, M., *et al.*. (2010) "Towards Supporting the Life Cycle of Large Scale Scientific Experiments". International Journal of Business Process Integration and Management, v. 5, p. 79-92.

Moreau, L., *et al.*., (2011) "The Open Provenance Model core specification (v1.1)". *Future Generation Computer Systems*, v. 27, n. 6, pp. 743-756.

Oinn, T. M., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, R.M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. P. Li. (2004) “Taverna: a tool for the composition and enactment of bioinformatics workflows”. *Bioinformatics*, 20(17):3045– 3054.

Sahoo, S. S., Sheth, A., Corey, H. (2008) “Semantic Provenance for eScience: Managing the Deluge of Scientific Data” *Internet Computing*, IEEE, v. 12 n. 4, p. 46-54.

Sahoo, S. S., Shet, A., (2009) "Provenir ontology: Towards a Framework for eScience Provenance Management". In: Microsoft eScience Workshop.

Salayandia, L., da Silva, P.P. Gates, A. Q., Salcedo, G. F. (2006) “Workflow-Driven Ontologies: An Earth Sciences Case Study”. In E-SCIENCE '06 Proc. of the 2nd IEEE International Conference on e-Science and Grid Computing.

Stevens, R., Zhao, J., Goble, C, (2007) “Using provenance to manage knowledge of in silico experiments”, Brief Bioinformatics, v.8. n. 3, p. 183-194.

Smith, B., *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, 1251–1255.

Zhao, J. Open Provenance Model Vocabulary Specification. (2010). Available at <http://open-biomed.sourceforge.net/opmv/ns.html>.