6-1-2012

An Evaluation Approach for Interactions between Abstract Workflows and Provenance Traces

Leonardo Salayandia
University of Texas at El Paso, leonardo@utep.edu

Ann Q. Gates
University of Texas at El Paso, agates@utep.edu

Paulo Pinheiro
Pacific Northwest National Laboratory

Comments:
Technical Report: UTEP-CS-12-21

An Evaluation Approach for Interactions between Abstract Workflows and Provenance Traces

Leonardo Salayandia[1] **Ann Q. Gates**[1] **Paulo Pinheiro**[2]

[1]University of Texas at El Paso
500 W. University Drive
El Paso, Texas 79968

[2]Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352

Abstract

In the context of science, abstract workflows bridge the gap between scientists and technologists towards using computer systems to carry out scientific processes. Provenance traces provide evidence required to validate scientific products and support their secondary use. Assuming abstract workflows and provenance traces are based on formal semantics, a knowledge-based system that consistently merges both technologies allows scientists to document their processes of data collection and transformation; it also allows for secondary users of data to assess scientific processes and resulting data products. This paper presents an evaluation approach for interactions between abstract workflows and provenance traces. The claim is that both technologies should complement each other and align consistently to a scientist's perspective to effectively support science. The evaluation approach uses criteria that are derived from tasks performed by scientists using both technologies.

Introduction

Abstract workflows document planned processes to create scientific products. Provenance traces document actual ways in which data were processed to create scientific products. The documentation of abstract workflows and provenance traces are effective ways of capturing knowledge about scientific processes. A framework, which unifies an abstract workflow language and a provenance language, provides the means for capturing knowledge about scientific processes that is richer than any individual abstract workflow language or provenance language (Freire et al., 2006, Garijo and Gil, 2011, Salayandia and Pinheiro da Silva, 2010, Mandal et al., 2007). With a framework of this type in mind, this paper identifies the roles that scientists assume with respect to collecting, transforming and using data. We classify the efforts of scientists, when interacting with data throughout the data life cycle, as *data producers* and *secondary data users*. Data producers are responsible (or at least involved) in the collection and transformation of data to create data products. Secondary data users are interested in using data products that were not created by them (Zimmerman, 2003). Tasks relevant to each role are supported by a framework of this type.

The next section describes the type of framework being addressed here in more detail and the tasks that data producers and secondary data users are able to carry out with them. The analysis criteria section introduces criteria to evaluate these frameworks with respect to how well they support the tasks. The discussion section introduces an example to exercise the criteria. Finally, conclusions are presented in the last section.

Abstract Workflow and Provenance Framework

Frameworks addressed in this paper are those that use two languages based on formal semantics: an abstract workflow language and a provenance language. The abstract workflow language is intended for scientists to document their understanding of processes of collection and transformation of data. Abstract workflow languages are typically graphical; however, they are assumed to be grounded on a formal meta-model. A distinction is made between abstract workflows and concrete workflows, in particular with respect to the level of support of a computer execution model and the level of detail included in a workflow. In this sense, some workflow languages support multiple levels of abstraction, and the user is able to navigate between views with more or less detail. Abstract workflows, as presented in this paper, intend to describe languages that support the specification of processes from the point of view of scientists. Note that there are no stated assumptions about the level of technical expertise that a scientist may have. Hence, communities of scientists that are culturally accustomed to work with specific technical platforms may consider abstract works to be specifications that are in fact executable by a computer. However, abstract workflows typically are documented processes expressed in terms relevant to a scientific discipline and independent of technical platforms used to carry out processes.

The provenance language is intended to document traces of execution of processes that collect and transform data. The provenance research community offers various alternatives for provenance languages, and efforts are underway to establish a standard provenance language for the Web (Gil et al., 2010).

Frameworks that combine an abstract workflow language and a provenance language can support a scientist in documenting planned processes that collect and transform data into scientific products, and they can capture provenance traces of scientific products as those planned processes are carried out. Note that the use of the workflow and provenance technologies on their own may result in alternate applications not addressed by the type of framework described here. For example, provenance languages may be used to capture provenance traces of *ad-hoc* activities, i.e., not following a planned process.

Frameworks that combine abstract workflows and provenance traces support the following tasks, which data producers and secondary data users typically carry out:

- *Process authorship:* For data producers, process authorship refers to documenting processes to collect and transform data. Regardless of the level of technical expertise or technical involvement of the scientist in the data process, process documentation commonly focus on scientifically relevant aspects and ignore technical nuances. A scientist's understanding of a process or a scientist's intended use of a process should guide the identification of relevant aspects.
- *Process analysis:* For secondary data users, process analysis refers to understanding the components and structure of the process used to collect and transform data in order to extract relevant information.
- *Process interoperability:* For secondary data users, process interoperability refers to reusing workflows in other contexts. For example, scientists may be interested in replicating published findings, they may be interested in reusing a workflow to process their own data, or they may want to use portions of the workflow in their own workflows (Garijo and Gil, 2011, Goderis, 2008).
- *Provenance capture*: For data producers, provenance capture refers to documenting a provenance trace that records their account and understanding of how, what, and who was involved in creating a data product.
- *Provenance analysis*: For secondary data users, provenance analysis refers to understanding the components and structure of a provenance trace in order to extract relevant information.
- *Provenance interoperability*: For secondary data users, provenance interoperability refers to using and extending provenance in other contexts. For example, a scientist interested in using a data product may want to extend its provenance trace as he or she manipulates the data product.

Analysis Criteria

Criteria are defined next to evaluate frameworks that use abstract workflows and provenance with respect to their support of the scientist's tasks described in the previous section. The relation between criteria and scientist tasks is summarized in Table 1.

The criteria are used by analyzing the languages used in the framework, i.e., the workflow language with which a data transformation process is documented, and the provenance language with which the data transformation process is documented once it is carried out. With respect to usability, the criteria address the workflow language only because user interaction is mainly through the graphical representation of the abstract workflow language. The provenance language, however, is assumed to be a back-end language, where software tools are used to generate and interpret it.

A situation in which inspection of the workflow and provenance languages is difficult may require applying the framework to a project in order to collect data to support the analysis with respect to the criteria.

Table1: Mapping of scientist tasks to criteria

Task	Criteria			
	Provenance Granularity	Workflow Notation Diversity	Workflow Terminology	Workflow/Prov. Vocabulary Coupling
Proc. authorship		x	x	
Proc. analysis		x	x	
Proc. interop		x	x	
Prov. capture	x			x
Prov. analysis	x			x
Prov. interop	x			x

C1: Provenance Granularity

This criterion is defined as the (number of process steps) / (number of provenance steps) ratio. A ratio of one means that for every process step introduced by the user in the workflow specification, there is one provenance step recorded when the process executes. In this case, the provenance granularity level is classified as *user-determined*. In the opposite situation where the ratio tends to zero, the provenance granularity level is classified as *system-determined*. There is also the situation where the ratio is greater than one, and although this situation is not expected to be common, it reflects that provenance is recorded at a coarser granularity than the workflow specified by the scientist. In the case where this criterion is used by applying the framework, it is assumed that the process specification does not contain loops, or that the number of provenance steps is normalized to remove loop execution steps. Workflow pipelines, i.e., sequential workflows without alternate paths or loops, should be the best case for this criterion, since all process steps in the workflow pipeline would contribute to the provenance trace when the process is carried out. What counts as a process step and as a provenance step is necessarily dependent on the workflow language and the provenance language used. The ratio of steps between both languages, however, is intended to eliminate specific language implementation concerns. This criterion addresses the following scientist tasks:

Provenance capture: For data producers, abstract workflows represent a process description from their perspective. Congruent levels of detail between an abstract workflow and corresponding provenance traces are expected to highlight the data producer's account of how, what, when, and who was involved in generating data products, i.e., a user-determined provenance granularity. On the other hand, provenance traces that include more details than those included in the abstract workflow are expected to capture provenance from the perspective of how the process is being carried out, i.e., a system-determined provenance granularity.

Provenance analysis: For secondary data users, user-determined provenance granularity should be more intuitive and less voluminous than system-determined provenance granularity.

Provenance interoperability: For secondary data users, provenance that is recorded at a user-determined granularity should be easier to reuse in other contexts, especially where the operational environment is different.

C2: Workflow Notation Diversity

This criterion is defined as the number of symbols used in the workflow graphical language. Although it is impossible to determine a specific value as the ideal for a given application, the literature suggests that graphical languages with diverse notation and secondary notation have a high learning curve (Petre, 1995). On the other hand, an over simplistic graphical language may lack expressivity to document processes from the perspective of data producers. This criterion uses one factor of language complexity that is straightforward to determine and that affects both creators of workflow specifications and interpreters, i.e., data producers and secondary data users. This criterion addresses the following scientist tasks:

Process authorship: For data producers, a minimal graphical language with reduced notation diversity is assumed to favor process authorship since the language would be easier to learn and would be potentially more intuitive.

Process analysis: For secondary data users, a minimal graphical language with reduced notation diversity is assumed to favor process readership for similar reasons as in the previous item.

Process interoperability: For secondary data users, it is assumed that reduced notation diversity in the workflow graphical language would result in a language with fewer restrictions to be imposed on the executing environment, hence, favoring the adaptability of the workflow language.

C3: Workflow Terminology

The intention of this criterion is to evaluate the abstract workflow language with respect to its flexibility to support terminology from users, e.g., scientists. If an abstract workflow is described using terminology introduced by a scientist, then potentially the abstract workflow is meaningful to a community of users with a similar disciplinary background. If, on the other hand, the scientist is forced to choose among technical terms suggested by software tools, then understanding the abstract workflow is more likely to require technical training on the specific software tools used to create the abstract workflow.

Qualitatively, the evaluation of the framework with respect to this criterion should yield *user-driven* or *system-driven workflow terminology*. Quantitatively, this criterion is defined as the percentage of terms used in an abstract workflow that are introduced by scientists. A percentage of 100 means all terms used in an abstract workflow are introduced by scientists, while a percentage of zero means that scientists choose terminology provided by the technical platform. Notice that the graphical language may implicitly provide technical terms. However, this type of implicit terminology is not considered here and, instead, is addressed by the *notation diversity* criterion. The vocabulary independence criterion also includes only the terms that are visible in the graphical layout of the workflow specification and does not consider other features of development environments, e.g., features to assist scientists in choosing technically-oriented components. The intention is to evaluate the graphical representation of the workflow, not other features of tools used to create them. There is also the case of technical platforms that target specific disciplines or that becomes widely adopted in a community (Oinn et al., 2006). In these cases, the vocabulary provided by the technical platform may in fact be compatible with the vocabulary preference of scientists. This criterion should provide best results in evaluating generic technical platforms that are intended to be used across disciplines and that are flexible with respect to user vocabulary preference. What is more, the intention is to provide a criterion to assess the level of technical expertise that a creator or interpreter of a workflow specification must have to use it. For example, it is assumed that a scientist reading a workflow will better understand the workflow if it includes terms from his/her background discipline. This criterion addresses the following scientist tasks:

Process authorship: For data producers, flexibility to choose terminology from a familiar domain of expertise should facilitate process authorship, making the exercise more intuitive for data producers.

Process analysis: For secondary data users, workflows that use vocabulary common to their discipline should be easier to analyze. Ideally, the graphical representation of the workflow would be enough for scientists to interpret the process of data collection and transformation, minimizing the need to understand the technical platform in order to analyze the process.

Process interoperability: For secondary data users, workflows that use vocabulary that is independent of a specific platform should be easier to transfer and reuse in other operational environments, i.e., assuming that scientists have to understand the workflow as a requirement to adopt it in their operational environments. However, there may also be the case where software tools are available to automate the conversion of workflows from one platform to another; even in these cases a scientist's interpretation of the workflow is still necessary to validate that automatic conversions are sound.

C4: Workflow/Provenance Vocabulary Coupling

The intention of this criterion is to evaluate the level of vocabulary commonality between a workflow specification expressed in the abstract workflow language and a corresponding provenance trace expressed in the provenance language. The abstract workflow language and the provenance language are naturally different, having different design goals and intended uses. However, given that abstract workflows represent processes of collection and transformation of data from the perspective of scientists, data provenance should be easier for scientists to understand and use if there is a clear correspondence between the abstract workflow and the provenance trace. While the *provenance granularity* criterion evaluates correspondence between abstract workflows and provenance traces from a structural stand point, this criterion evaluates correspondence from a terminology stand point.

Qualitatively, the evaluation of the framework with respect to this criterion should yield *high* or *low vocabulary coupling*. Quantitatively, the level of vocabulary coupling can be defined as the percentage of terms in the workflow specification that are used in the provenance trace; a percentage of 100 means that all terms used in the abstract workflow are used in the provenance trace and would be qualified as *high vocabulary coupling*. A percentage of zero means that the provenance trace is independent of the abstract workflow and would be qualified as *low vocabulary coupling*. Notice that the quantification of this criterion measures a percentage with respect to the terms in the abstract workflow, which are potentially introduced by scientists. Quantifying the criterion this way intends to disregard the complexity of the provenance language, i.e., if the criterion was quantified as the percentage of terms used in the provenance trace that were common in the abstract workflow, the outcome would be susceptible to syntax complexity of the provenance language.

Similar to the *workflow terminology* criterion, this criterion considers the terms that are visible in the graphical layout of the abstract workflow. Similar to the *provenance granularity* criterion, this criterion is best employed on workflow pipelines where all process steps contribute to the provenance trace. This criterion addresses the following scientist tasks:

Provenance capture: For data producers that have documented their processes of collection and transformation of data as abstract workflows, capturing provenance in a language that supports high vocabulary coupling should be more intuitive and easier to validate.

Provenance analysis: For secondary data users, assuming that an abstract workflow is specified using vocabulary that is familiar to them, a corresponding provenance trace should be easier to analyze if there is high vocabulary coupling between the abstract workflow and the provenance trace. Low vocabulary coupling, on the other hand, would mean that the provenance trace is expressed using vocabulary that is specific to the provenance language or operational environment, which the scientist would have to understand a priori in order to analyze the provenance trace in detail.

Provenance interoperability: High vocabulary coupling is indicative of provenance traces that are expressed in languages that are less dependent on operational environments. For secondary data users wanting to extend a provenance trace in their own contexts, high vocabulary coupling is desired, since the provenance trace is more likely to be adaptable across operational environments.

Discussion

The Workflow-Driven Ontology (WDO) framework is a framework that combines abstract workflows and provenance (Salayandia and Pinheiro da Silva, 2010, Pinheiro da Silva et al., 2010). The criteria, which is presented in the previous section, is used to evaluate the WDO framework. The abstract workflow language of WDO is based on Data Flow Diagrams (DFD's) (Davis, 1990), chosen for their simplicity as the abstract workflow language is expected to facilitate use by scientists. The modular design of the WDO framework is intended to support the exchange of provenance languages. However, the initial work uses the Proof Markup Language (PML) (McGuiness et al., 2007).

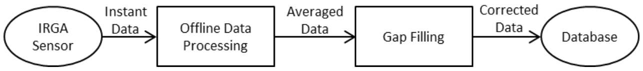

Figure 1: Abstract workflow of eddy covariance process

Figure 1 presents an abstract workflow created with the WDO framework. It corresponds to a data process from the environmental sciences community where the technique of eddy covariance is employed to monitor carbon and water fluxes in the environment (Jaimes et al., 2010). The process starts with an Infrared Gas Analyzer (*IRGA*) sensor deployed in the field of study. Sensed data is stored in a data logger, transmitted over WIFI to a regional field office, and eventually transmitted to a processing server in the main laboratory. The data is referred to as *instant data* at this point, a common term for projects of this nature. Notice that technical details about storing and transmitting the data to a server in the main laboratory are not included in the abstract workflow since they are not relevant from the scientist's perspective. Other frameworks may require the scientist to include such details, depending on the level of abstraction supported by the workflow language and the level of process automation. Instant data is filtered and processed using various specialized algorithms described in more detail as a sub-process, not included here because of space constraints, but generalized as the o*ffline data processing* step depicted in Figure 1. The outcome of this generalized step is *averaged data*, also a common term used in this community. Finally, the nature of the process makes environmentally exposed instrumentation susceptible to failure. Given the dynamic changing conditions of the environment and the high impact on results for missing data, a *gap filling* step is necessary to extrapolate sensed data with specialized algorithms that account for other environmental factors. The resulting dataset is called *corrected data*, which is stored into the project's *database*, eventually to be shared among colleagues.

Table 2: Evaluation of the WDO Framework

Criteria	**Result**
Provenance Granularity	User-determined
Workflow Notation Diversity	Low (3 symbols)
Workflow Terminology	User-driven
Workflow/Provenance Vocabulary Coupling	High

Table 2 summarizes the result of the evaluation for the WDO framework. The results are explained next. In order to determine provenance granularity, it is necessary to define what constitutes a process step and a provenance step in the WDO framework. A process step is counted for each data transformation step in the abstract workflow, i.e., each rectangle. A provenance step is counted for each *NodeSet*, a construct used in the Proof Markup Language (PML) to link antecedents to conclusions and the main mechanism in PML to record data provenance (McGuiness et al., 2007). Figure 2 shows a snippet of the provenance trace for the last part of the abstract workflow of Figure 1, where the NodeSet has *Corrected Data* as conclusion (line 3), uses the *Gap Filling* rule (line 10), and has antecedents represented by another NodeSet (line 13). Hence, it is expected that for each process step there will be a provenance step, making the outcome of this criterion a user-determined provenance granularity.

With respect to provenance interoperability and its relation to provenance granularity, Lebo et al. (2012) provide an approach to normalize the level of detail of provenance traces from multiple sources. The results are derived provenance traces from different sources documented at consistent levels of detail for a given application. While the provenance granularity criterion presented in this paper intends to align provenance level of detail to a scientist's perspective, it is clear that a consistent level of detail across projects is not guaranteed.

With respect to the workflow notation diversity criterion, the framework can be evaluated by inspecting the abstract workflow in Figure 1. The diagram uses 3 symbols: directed edges represent data (and flow of), ovals represent sources and sinks of data, and rectangles represent process steps.

With respect to the workflow terminology criterion, the framework can also be evaluated by inspecting the abstract workflow of Figure 1. All terminology in the figure was introduced by the scientist and is meaningful to colleagues from similar disciplinary backgrounds. Hence, the framework is evaluated to support user-driven terminology.

Finally with respect to the workflow/provenance vocabulary coupling criterion, a comparison is made between the abstract workflow depicted in Figure 1 and the provenance trace of Figure 2. The conclusion of the provenance trace indicates that the type of data being concluded is of type *Corrected Data* (line 3) and that the *Gap Filling* rule is used (line 10). Both of these terms are direct references to the terminology introduced in the abstract workflow. Inspection of NodeSets corresponding to the rest of the abstract workflow is expected to include the remaining terminology introduced by the scientist. Hence, the framework is evaluated to have a high coupling of vocabulary between the abstract workflow and the provenance trace.

```
<NodeSet rdf:about="http://URI-of-this-nodeset">
  <hasConclusion>
    <mywdo:CorrectedData>
      <hasURL rdf:resource="http://../data.csv"/>
    </mywdo:CorrectedData>
  </hasConclusion>
  <isConsequentOf>
    <InferenceStep>
    <hasInferenceEngine rdf:resource="http://..#exec-environ"/>
      <hasInferenceRule rdf:resource="http://../mywdo.owl#GapFilling"/>
      <hasAntecedentList>
        <NodeSetList>
          <ds:first rdf:resource="http://URI-of-another-nodeset"/>
        </NodeSetList>
      </hasAntecedentList>
    </InferenceStep>
  </isConsequentOf>
</NodeSet>
```

Figure 2: Provenance trace in PML of Corrected Data

The WDO framework is specifically designed to align to a scientist's perspective in documenting data processes and capturing provenance traces. This is reflected in the outcome of evaluating the WDO framework with respect to the criteria presented. Workflow/provenance frameworks typically require compromise between supporting a scientist's perspective and other factors, e.g., the expected level of process automation. The criteria should be helpful in assessing the impact of such compromises.

Conclusions

Abstract workflows promote understanding of processes by end users and documentation of processes in early stages. Provenance languages promote understanding of end results in support of secondary use and repeatability. This paper describes criteria to evaluate frameworks that combine both technologies, emphasizing the need to align to a scientist's perspective over a technical perspective in order to support a scientist's tasks.

The use of the criteria was demonstrated by evaluating the WDO framework as it was applied to capture a data process and provenance traces for an environmental sciences project.

References

Davis, A.M. 1990, Software Requirements: *Analysis and Specification*, Upper Saddle River, NJ: Prentice Hall Press.

Freire, J., Silva, C.T., Callahan, S.P., Santos, E., Scheidegger, C.E., and Vo., H.T. *Managing rapidly-evolving scientific workflows*. In Proc. IPAW 2006, Chicago, IL, May 2006.

Garijo, D., and Gil, Y., *A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data.* In Proc. WORKS'11, Seatle, WA, 2011.

Gil, Y., Cheney, J., Groth, P., Hartig, O., Miles, S., Moreau, L., and Pinheiro da Silva, P. *Provenance XG final report*. Technical report, W3C, December 2010.

Gil, Y., Groth, P., Ratnakar, V., and C. Fritz. *Expressive Reusable Workflow Templates*, In Proc. IEEE e-Science Conference, Oxford, UK, pages 244–351. 2009.

Gooderis, A., Workflow Re-use and Discovery in Bioinformatics, Doctoral Thesis, University of Manchester, 2008.

Jaimes, A., Salayandia, L., Gallegos, I., Pennington, D., Gates, A.Q., and Tweedie, C., *Establishing Cyberinfrastructure for Studying Land-Atmosphere Interactions using Eddy Covariance*. AGU Fall Meeting, 2010.

Lebo, T., Wang, P., Graves, A., McGuiness, D., *Towards Unified Provenance Granularities*. In Proc. IPAW 2012, Santa Barbara, CA, June 2012.

Mandal, N., Deelman, E., Mehta, G., Su, M., and Vahi, K. 2007. *Integrating existing scientific workflow systems: the Kepler/Pegasus example*. In Proc. WORKS'07. ACM, New York, NY, USA, 21-28.

McGuiness, D.L., Ding, L., Pinheiro da Silva, P., and Chang, C. *PML 2: A Modular Explanation Interlingua*. In Proc. 2007 Workshop on Explanation-aware Computing, ExaCt-2007, 2007.

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A. and Wroe, C. (2006), *Taverna: lessons in creating a workflow environment for the life sciences*. Concurrency Computat.: Pract. Exper., 18: 1067–1100.

Petre, M., *Why looking isn't always seeing: readership skills and graphical programming*. Commun. ACM, 38:33–44, June 1995.

Pinheiro da Silva, P., Salayandia, L., Del Rio, N., and Gates, A.Q. *On the use of abstract workflows to capture scientific process provenance*. In Proc. TaPP'10, San Jose, CA, February 2010.

Salayandia, L. and Pinheiro da Silva, P. *On the use of semantic abstract workflows rooted on provenance concepts*. In Proc. IPAW 2010, Troy, NY, June 2010.

Zimmerman, A.S, *Data sharing and secondary use of scientific data: experiences of ecologists*. Doctoral Thesis, University of Michigan, 2003.