

Semantic Workflow Management

Edoardo Pignotti
Dept. of Computing Science
University of Aberdeen
Aberdeen, AB24 3UE, Scotland
epignott@csd.abdn.ac.uk

ABSTRACT

In the e-Science context, workflow technologies provide a problem-solving environment for scientists by facilitating the creation and execution of experiments from a pool of available data and computation services. We argue that in order to characterise scientific analysis we need to go beyond low-level service composition and execution details by capturing higher-level description of the scientific process. The aim here is to make the experimental conditions and goals of the experiment transparent. Current workflow technologies do not incorporate any representation of these goals and conditions, which we call the *scientist's intent*. Our hypothesis is that by extending workflow representation in this way, scientists (including social scientists) would be able to analyse, verify, execute, monitor and re-use workflows more efficiently.

Categories and Subject Descriptors

H.4.1 [Office Automation]: Workflow management. D.2.6 [Programming Environments]: Integrated environments.

General Terms

Management, Design, Experimentation, Human Factors.

Keywords

Semantic workflow, scientist's intent.

1. OVERVIEW

In the e-Science context, workflow technologies provide a problem-solving environment for researchers by facilitating the creation and execution of experiments from a pool of available data and computation services. We argue that in order to characterize such analysis we need to go beyond low-level service composition and execution by capturing a higher-level description of the experimental process. The aim here is to make the conditions and goals of the experiment transparent. Current workflow technologies do not incorporate any representation of these goals and conditions, which we describe as the scientist's *intent*.

Early in our work we identified a number of scenarios through interactions with collaborators from the social simulation community. We now present a simulation case study using a virus model developed in NetLogo¹; the model is an agent-based simulation of the transmission and perpetuation of a virus in a human population. An experiment using this model might involve studying the differences between different types of virus in a specific environment. A researcher wishing to test the hypothesis

'Smallpox is more infectious than Bird Flu in environment A' might run a set of simulations using different random seeds. If in this set of simulations, *Smallpox* outperformed *Bird Flu* in a significant number of simulation runs, the experimental results could be used to support the hypothesis.

Figure 1 shows a workflow built using the Kepler editor tool (Ludäscher *et al.*, 2005) that uses available services to perform the experiment described above. The *VirusSimulationModel* generates simulation results based on a set of parameters loaded as input from a data repository; the experiment definition is selected by *Experiment ID*. These simulation results are aggregated and fed into the *Significance Test* component which outputs the results of the test. The hypothesis is tested by looking at the result of the significance test; if the virus that we are considering (e.g. *Smallpox*) outperforms others in a significant way, we can use this result to support our hypothesis.

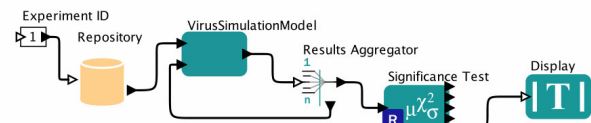


Figure 1 - Simulation Workflow Example.

However, the experimental workflow defined in Figure 1 has some limitations as it is not able to capture the scientist's goals and conditions (scientist's *intent*). For example, the goal of this experiment is to obtain a significant number of simulation results that support the hypothesis. Imagine that the scientist knows that the simulation model could generate out-of-bound results and these results cannot be used in the significance test as they will affect the experiment. For this reason, we don't know *a priori* how many simulation runs per comparison we need to do in order to have a significant number of results. There may also be constraints associated with the workflow (or specific activities within the workflow) depending upon the intent of the scientist. For example, a researcher may be concerned about floating point support on different operating systems; if the *Significance Test* activity runs on a platform not compatible with IEEE 754 specifications, the results of the simulation could be compromised. Existing workflow languages are unable to explicitly associate such information with their workflow descriptions.

2. RESEARCH QUESTION

We argue that by extending workflow representations to include the scientist's *intent*, researchers would be able to analyse, verify, execute, monitor and re-use workflows more efficiently. The main challenges are to represent the intent in such a way that:

¹ <http://ccl.northwestern.edu/netlogo/>

- it is meaningful to the researcher, e.g. providing information about the context in which the experiment has been conducted so that the results can be interpreted more precisely;
- it can be reasoned about by a software application, e.g. an application can make use of the intent information to control, monitor or annotate the execution of a workflow;
- it can be re-used across different workflows, e.g. the same high-level intent may apply to different workflows;
- it can be used as provenance (documenting the process that led to a specific piece of data).

3. WORK IN PROGRESS

We are developing a model of *scientist's intent* based upon rules which operate on metadata generated by the workflow. Details of the intent are kept separate from the operational workflow as embedding constraints and goals directly into the workflow representation would make it overly complex (e.g. with a large number of conditionals) and would limit potential for sharing and re-use. Such a workflow would be fit for only one purpose and addition of new constraints would require it to be substantially re-engineered. Using the support for *scientist's intent*, a new experiment might be created just by changing the rules but not the underlying operational workflow. We have identified SWRL (Semantic Web Rule Language²) as a language for capturing such rules. SWRL enables Horn-like rules to be combined with metadata. The rules are of the form of an implication between an antecedent (body) and consequent (head). The intended meaning can be read as: whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold. The *scientist's intent* reflected in the example in Figure 1 can be therefore be represented as a combination of goals and constraints:

Goal: Run enough simulations to provide valid results to support (**significance test < significance level**).

Constraints: Simulation has to run on a platform compatible with IEEE 754, (**platform = IEEE 754**).

Central to our idea of capturing intent is the concept of the Semantic Grid. Where Grid technologies (Foster *et al.*, 2001) provide an infrastructure to manage distributed computational resources (e.g. the FEARLUS Grid Service Pignotti, *et al.*, 2005), the Semantic Grid aims to provide a "human-centred" Grid which integrates Semantic Web and Grid technologies. To enable such rules to be applied, the workflow must have supporting ontologies and should produce metadata that can be used against scientific intent (rules) to document the execution of the workflow. We have identified the following possible sources of metadata: (i) metadata about the result(s) generated upon completion of the workflow; (ii) metadata about the data generated at the end of an activity within the workflow or sub-workflow; (iii) metadata about the status of an activity over time, e.g. while the workflow is running.

We plan to evaluate our approach by assessing the impact of the enhanced workflow representation from two perspectives: (i) the utility of the *intent* construct as additional metadata to facilitate

interpretation of experimental results and workflow re-use; (ii) how the real-time monitoring of experiments guided by *intent* affects use of Grid resources.

4. EXPECTED CONTRIBUTIONS

We will contribute three (or more) case studies based on our approach with related workflows and scientist's *intent* representation. The case studies will include but will be not limited to the following:

- Social simulation experiment involving a land use model.
- Ecology simulation experiment.
- Data-driven simulation where the simulation depends on external data sources.

We will contribute the following ontologies:

- Simulation ontology to support the classification of simulation models based on the characteristic of the simulation environment, e.g. type of simulation, behaviour, space model, execution type, etc.
- Framework ontology, developed as part as a modelling framework at the Macaulay Institute to describe the components of a simulation model (Polhill & Gotts, 2006). Input from us on the design of the ontology will provide the necessary links to integrate this ontology with our scientist's intent framework. This will allow us to make use of real-time metadata in our workflow during the execution of a simulation.
- Workflow ontology, to describe workflow components and their relationship so they can be used as part as the scientist's *intent*.

We will produce a stand-alone software architecture to manage *scientist's intent* creation and use. We will also provide a set of APIs to interface our architecture with existing workflow tools.

We will provide an extension of the existing Kepler workflow tool to make use of the scientist's intent framework.

5. References

- [1] Foster, I. and Kesselman, C. (1998): 'Globus: A Toolkit-Based Grid Architecture.' In: The Grid: Blueprint for a Future Computing Infrastructure. Morgan- Kaufmann, 1998, pp. 259-278.
- [2] Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J. and Zhao, Y., (2005): 'Scientific Workflow Management and the Kepler System' In: Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows.
- [3] Pignotti, E., Edwards, P., Preece, A., Polhill, G. and Gotts, N. (2005) 'Semantic Support for Computational Land-Use Modelling' Proceedings of the Fifth IEEE International Symposium on Cluster Computing and Grid (CCGrid)2005, IEEE Press, 2005.
- [4] Polhill, J. G. and Gotts, N. M. (2006): 'A New Approach to Modelling Frameworks' *Proceedings of the First World Congress on Social Simulation*, Kyoto, Japan.

² <http://www.w3.org/Submission/SWRL/>