A WORKFLOW LANGUAGE FOR THE EXPERIMENTAL SCIENCES

Yuan Lin, Thérèse Libourel and Isabelle Mougenot
LIRMM, UMR 5506 - CC 447, 161 Rue Ada, 34392 Montpellier Cedex 5 - France

Keywords: Scientific workflow, Meta-model.

Abstract: Scientists in the environmental domains (biology, geographical information, etc.) need to capitalize, distribute and validate their experimentations of varying complexities. The concept of the scientific workflow is increasingly being considered to fulfill this requirement. This article presents the first phase of the establishment of a workflow environment corresponding to the static part, i.e., a meta-model and a language dedicated to the design of process-chain models. We illustrate our proposal with a simple example from the spatial domain and conclude with perspectives that open up with the establishment of a workflow environment.

1 CONTEXT

Environmental applications are undergoing considerable growth, entailing the establishment of an efficient mutualisation infrastructure because the data involved is often voluminous and complex to acquire. All these experimental domains in which information is often spatialized share common characteristics: data and processes exist. However even if data exists in bulk and is often perennial in nature, the processes associated with it change over time. Moreover, the experiments are rarely simple and most often correspond to a more or less sophisticated combination of processes. Finally, in critical situations (as, for example, those of natural or anthropogenic risk), perennial data has to be correlated with data acquired in real-time, i.e., experimentation has to be conducted (predefined process chains) on the new data batches.

This first observation has led us to focus our research on the concepts of collaborative work and workflows, as introduced in (Khoshafian and Buckiewicz, 1998). Workflow is the automatization of a process (partially or completely) during which documents, information and tasks pass from one participant to another within a working group, in conformity with a set of predefined rules. A workflow system defines, creates and manages the execution of such processes.

The workflow concept is, of course, very much present in traditional organizations (industrial or financial management). But in the environmental context, even if the idea of sequencing and monitoring different tasks as part of a complex process is normal for scientists, the functionality of automatizing these processes in existing distributed infrastructures which manage heterogeneous resources remains a basic challenge. Furthermore, natural or anthropogenic phenomena necessitate modelling in which the sequencing of processes also resembles workflows.

In this article, we will first cover the general issues involved by invoking an example. Then, in section 3, we will introduce our proposal. We present there the first phase of our work, i.e., the meta-model of the scientific workflow language and the method of using it. Finally, the last part shall consist of perspectives and a general conclusion.

2 ISSUES INVOLVED

2.1 Example

The example that we have chosen arises from a simplified analysis of a case of a natural hazard: we would like to identify, on a map of the area, the buildings at risk in the Mauguio commune if a nearby dyke fails.

The scientist in charge of the project knows:

- The data that he has available which will constitute the input to his process chain and also the type of information that he wants as a result.

Lin Y., Libourel T. and Mougenot I. (2009).
A WORKFLOW LANGUAGE FOR THE EXPERIMENTAL SCIENCES.
In *Proceedings of the 11th International Conference on Enterprise Information Systems - Information Systems Analysis and Specification*, pages 372-375
DOI: 10.5220/0002000803720375

1. Input: A data layer[1] relating to dykes (linear) in the area.
2. Input: A data layer of the buildings in the concerned area (polygon).
3. Result: A map showing the buildings at risk in the flood zone in case of a failure of the dyke (which will be identified by an expert on the ground).

- Methods and Processes Adopted:
  1. Positioning the coordinates on a layer. Different geocoding techniques can be used to do so.
  2. Constructing a buffer zone from a geolocalization.
  3. Marking up a data layer. This method adds a detailed legend to the underlying data layer.
- The triggering event which is transmitted to it by the ground operator (or possibly a sensor).

2.2 Analysis of Scientific Stumbling Blocks

Our objective is thus to give to this scientist an environment, as simple as possible, to describe the analysis to be conducted and launch the execution of the underlying processes.
This workflow environment will be integrated into an already existing mutualisation platform. The underlying community (bio-diversity, ecology, environment) has already shared data via the platform and a metadata-based localization engine (Barde et al., 2005).

This platform offers a search engine based on the metadata of description of resources (data and processes) as well as on the shared knowledge of the underlying domains.

Initial analysis leads us to highlight two underlying aspects of the workflow:

1. the static aspect devoted to the management of process chains (definition, saving);
2. the dynamic aspect devoted to possible executions.

Component and model engineering (Tamzalit and Aniorté, 2005; OMG, 2003) is the basis for the proposal. The environment must be as simple as possible on the one hand, and the most adaptable and reusable on the other.

We can provide an overview of such architecture environment which we divide into:

- A Static Part: it consists of a meta-model from which experts can design a descriptive business model of the desired process chain, conforming to the domain of expertise;
- A Dynamic Part: the models to be executed will be instantiations of business models created from resources (components, services, data, etc.) that are available before the actual execution.

To arrive at our goal, we have to:

1. produce a simple meta-model for rapid adoption by the experts (this is topic of our current proposal), and
2. demarcate instantiation techniques and validation.

3 OUR PROPOSAL

As mentioned in the introduction, our long-term goal is to offer a complete environment for describing process chains and their execution.

Based on the summary, above, we initially propose a workflow description language. This language is defined by a meta-model which is inspired by the existing meta-models we have analyzed and also by the general meta-model relating to graphs and ontologies.

3.1 The Simple Workflow Meta-model (SWM)

3.1.1 General Introduction

The aim is to define the minimum number of elements necessary to illustrate a maximum number of possible situations (Fürst, 2002).

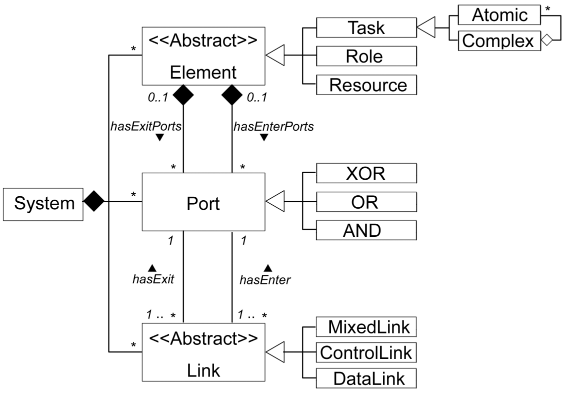

Figure 1: Our meta-model.

[1] The term data layer or layer is used by geographical information systems to designate a set of geometrically homogeneous spatialized data.

The meta-model was designed from the point of view of the workflow software environment. It is thus perceived, at the most abstract level, as a composition of *elements* and *links* between elements. The concept of the *port* allows connections between elements and links.

The elements can be divided into:

- *Tasks* predefined to be one-time or reusable [2],
- Existing *roles* (which will be involved during the execution phase),
- *Resources* available to be mobilized.

The concept of *tasks* corresponds to those of Activity, Process, etc. generally used by the other workflow meta-models. We break up this concept into a composite: a task can be complex or atomic, with the possibility of reusing a complex task concatenated into an atomic task.
The elements are connected by unidirectional links [3] by the intermediary of ports. We distinguish between:

- The *DataLinks* which serve, on the one hand, for transferring data between elements and, on the other, for ensuring the sequence of processes in the plan.
- The *ControlLinks* and *MixedLinks* which are included mainly for controlling the authorization of executions and/or temporal scheduling.

Links connect elements by way of *ports* (normal ports by default) which are attached to them. Each element has input/output ports (the I/O type is connected in the direction of the corresponding link).

In addition, to be able to process more complex examples such as data fusion, synchronization, etc., of elements (ports and links), specific ports are introduced: AND, OR, XOR.

To facilitate the manipulation of processes, an associated graphical language is proposed, cf. figure 2.

3.2 Implementation

We have constructed a first prototype of the meta-model and of the graphical language[4]. We present the illustration made from the original example.
The model obtained (cf. figure 3) from the current prototype shows instantiated elements conforming to the meta-model and represented by the symbolic language introduced earlier, in figure 2.

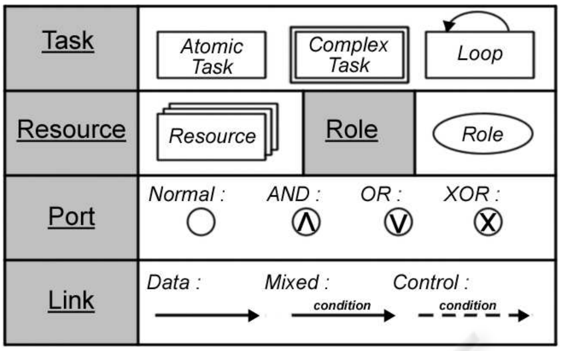

Figure 2: An associated graphical language.

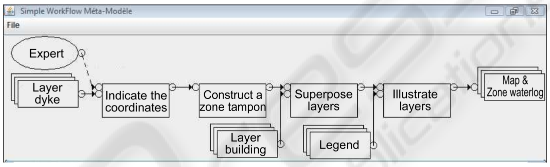

Figure 3: Occupational model.

4 PERSPECTIVES AND CONCLUSIONS

4.1 Perspectives

We can now state our vision of what remains to be done and highlight the essential difficulties and stumbling blocks that we foresee.

4.1.1 Levels of Expertise and Resource Typing

The experimental approaches that interest us are based on a protocol requiring many different levels of expertise. For example, in the context of the illustrative example of section 4, the specialist will describe the scenario as we have presented it. If his expertise in GIS tools is insufficient, he will then have to call upon another geomatics expert who will specify the processes in detail. Finally, if the data or some required feature is not available, a network administrator will have to intervene to provide the resources necessary for the execution.

Model transformations are therefore to be put in place, with each level of expertise needing verifications of compatibility in association with supplementary information on the model elements being used (signatures and resource typing).

[2] In the current context, Web services can, for example, be considered tasks.

[3] There are no direct links between role and resource. In most cases, the links between role and resource can be deduced from role-task and task-resource links.

[4] Prototype written in JAVA

4.1.2 Control and Traceability of Processes

In the current context, distributed executions cannot be ignored. Because of this, more and more work on workflows is focusing on the integration of services available online (such as WS), i.e., on the reuse of existing resources.

However, a whole set of problems of execution still remain: How will the different components of a workflow interact amongst themselves? How can we guarantee correct execution in such environments?

Experts in the experimental domain are interested not only in the final results of their experiments, but also in the way these results are obtained, in the type of type of dependence between different data items, etc. (Bowers et al., 2006; Moreau and Foster, 2006).

The execution modalities should take into account the expression of this requirement: the processes should be traceable. On the one hand, traceability provides the possibility of verifying the results at every stage, even to monitor each stage execution (and therefore to better identify points of error) and, on the other, it allows users to complete data descriptions (for example, by automatizing the entry of the metadata genealogy field).

4.2 Conclusions

The meta-model whose rough draft we have presented here was created after a survey of existing work on the subject. However, our analysis has ignored work on web services and the coordination of services (which we feel correspond more to the dynamic part). We have also not covered component languages and component-assembly languages which address the compatibility problems we refer to.

The targeted users should have a simple language at their disposal and should encounter easily appropriable concepts; this is what has led us to choose a relatively simple meta-model and symbolism.

The perspectives that we can quickly draw are:

- On the short term, we have to complete, refine, even simplify the meta-model and try it out on several diverse examples to judge its suitability;
- The concept of role, which so far has been rather nebulous, could lead to a more modular vision of workflow by including the notions of hierarchy of control and/or collaboration;
- On the longer term, we have to go beyond the functionalities of description to develop the dynamic aspect (execution).

REFERENCES

Barde, J., Libourel, T., and Maurel, P. (2005). A metadata service for integrated management of knowledges related to coastal areas. *Multimedia Tools Appl.*, 25(3):419–429.

Bowers, S., McPhillips, T. M., Ludäscher, B., Cohen, S., and Davidson, S. B. (2006). A model for user-oriented data provenance in pipelined scientific workflows. In *IPAW*, pages 133–147.

Fürst, F. (Octobre 2002). L'ingénierie ontologique. Technical report, IRIN, Université de Nantes.

Khoshafian, S. and Buckiewicz, M. (1998). *Groupware & Workflow*. Masson.

Moreau, L. and Foster, I. T., editors (2006). *Provenance and Annotation of Data, International Provenance and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers*, volume 4145 of *Lecture Notes in Computer Science*. Springer.

OMG (2003). Mda guide version 1.0.1.

Tamzalit, D. and Aniorté, P. (2005). Ingénerie des composants et systèmes d'information. *RSTI - Série L'Objet (RSTI-Objet),vol 13/4, Hermès - Lavoisier*.