Mem. S.A.It. Vol. 80, 559

Memorie della

Workflow systems and VO standards

André Schaaff[1], François Bonnarel[1], Mireille Louys[2], Eric Slezak[3], Brice Gassmann[1], Cyril Pestel[1], Omar Benjelloun[1], Grégory Mantelet[1]

[1] Centre de Données astronomiques de Strasbourg, 11 rue de l'Université, F-67000 Strasbourg, France. e-mail: `schaaff@astro.u-strasbg.fr`
[2] Laboratoire des Sciences de l'Images, de l'Informatique et de la Télédétection, F-67412 Illkirch, France
[3] Observatoire de la Côte d'Azur, F-06304 Nice, France

Abstract. After a short introduction to workflow systems, we focus on the use of "Characterization", an IVOA standard, in a workflow test bed architecture. The execution of a workflow may require substantial computing resources and this can take a significant amount of time. Our goal is to go as far as possible in the validation of the workflow process before the "real" execution of its components. This validation needs a knowledge of the components and of the data they consume/produce. The "Characterization" standard introduces a way to satisfy these needs at the data level.

1. Introduction

At the lowest level, a workflow can be composed by a set of tools executed by hand under shell. It is not very efficient but it can be a solution if the tools are interactive. A workflow can also be a script file which constitutes a first step to the formalization (easily reusable, ...). In this case the tools should be black boxes (or converted to) with just I/O parameters. At a highest level, a workflow can also be more sophisticated and based on a description language (often in XML) and interpreted by a workflow engine.

Workflows are not just useful to create an automatic execution of a set of tools. They make possible the capture and the preservation of scientific methodology (formalization of the scientific analysis). They give information about the tools (different algorithms) to use, the data flow, the execution details, The management of the computation is also possible at a large-scale (involving an access to clusters, grids, large DBs,) and the complexity (introduction of synchronisation, conditions, ...) is often hidden. A workflow system provides a collaborative environment for the analysis/design/execution/validation of new use cases.

2. A sophisticated workflow system

It is an architecture involving :

– A graphical design tool allowing the selection of the tasks, the I/O parameters, the controls, etc.
– A description file of the workflow is generated from the graphical representattion. It is often in XML.
– A workflow engine which is able to understand the description file and executes it by dispatching the tasks and managing the controls.

- An execution (often) visible step by step.
- A possible storage of intermediate data : uesful to experiment different scenarios at different steps of the re-execution (change of some parameters without the execution of the whole workflow).
- The result(s) can be linked to external tools depending on the kind of output data (FITS, ...).

3. Brief state of the art

Much projects have defined their own workflow language, engine and/or design tool. We give a non exhaustive list of these different components.

3.1. Workflow languages

SCUFL/XSCUFL is the language used in Taverna. Other languages : AGWL, BPEL4WS, BPML, DGL, DPML, GJobDL, GSFL, GFDL, GWorkflowDL, MoML, SWFL, WSCL, WSCI, WSFL, XLANG, YAWL, WPDL, PIF, PSL, OWL-S, xWFL.

3.2. Workflow language formalisms

The UML activity diagram, Petri net, BPMN, DAG, IPO, GPSG, Workflow Patterns, Pi Calculus, Finite-State Machine, Gamma-calculus.

3.3. Workflow engines

Taverna is starting to be used in the VO community. Other engines : BioPipe, BizTalk, BPWS4J, DAGMan, GridAnt, Grid Job Handler, GRMS, GWFE, GWES, IT Innovation Enactment Engine, JIGSA, JOpera, Kepler, Karajan, Moteur, OSWorkflow, Pegasus (uses DAGMan), Platform Process Manager, ScyFLOW, SDSC Matrix, SHOP2, Triana, wftk, YAWL Engine, WebAndFlo, WFEE, . . .

3.4. Workflow composition/designing tools

Taverna contains also a composition tools. Other tools for the design : ilog's BPMN Modeller, CAT, GWUI, XBaya GUI for Workflow Composition, Triana, JOpera, Platform Process Manager.

3.5. Workflow mapping from abstract to concrete workflows

CWG, ACWG, Grid Job Handler, GWES.

4. Workflows in the Virtual Observatory

An increasing number of services are developed and deployed in the frame of the Virtual Observatory (registries, data services, Web Services, computing and Grid services,). It become possible to create new (sometime complex) use cases involving these services. The implemention and the coordination of these use cases are possible through workflows.

5. VO France workflow working group

The working group (Schaaff 2006) has started to work in 2005 in the frame of OV France. In a first step it was necessary to give our own definition of a workflow : a sequence of tasks executed within a controlled context by an architecture taking into account VO standards.

The main goals of the working group are :

- the definition of use cases of general interest in different domains (image processing (Slezak 2006), spectroscopy, data mining,. . .).
- the suggestion of solutions for designing and exploiting easily such workflows.
- the identification of the simplest workflow structure allowing portability.
- the definition of elementary bricks (usable in different domains).

The aim of the working group is not to decide which tool is better than another but in a

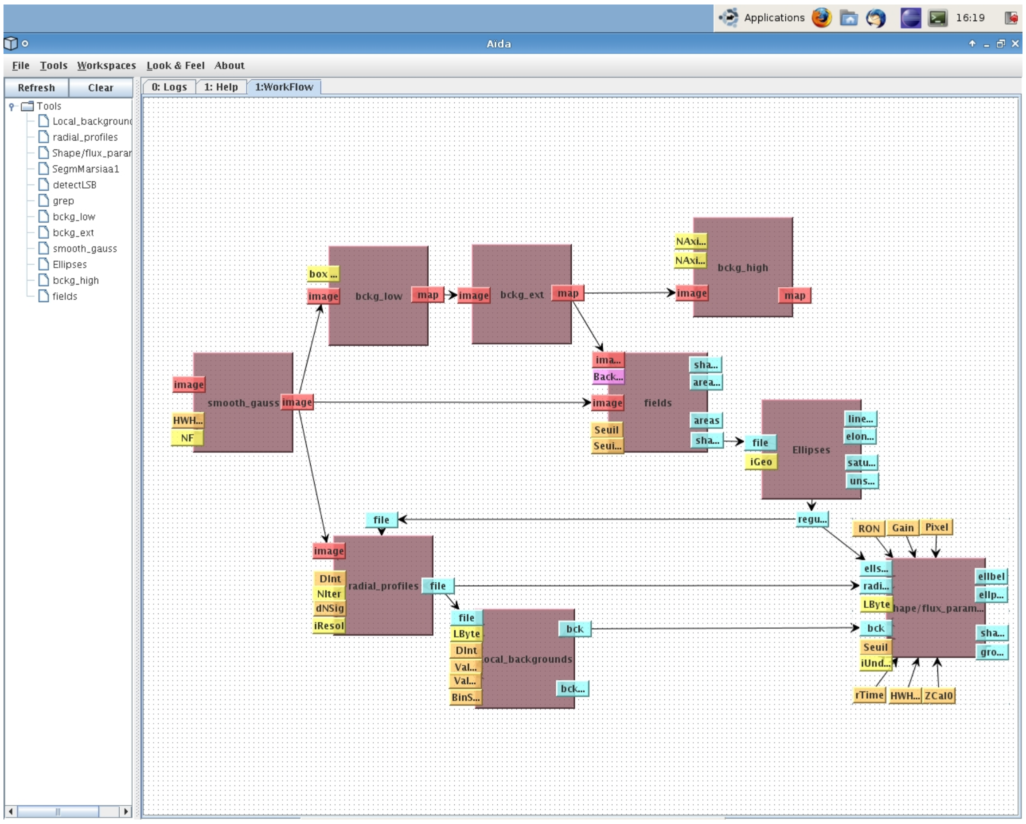

Fig. 1. Example of workflow for image processing

first step we decided to use AÏDA (Schaaff et al. 2005; Louys et al. 2006; Slezak & Schaaff 2006) developed in the French Ministry funded project MDA and continued in the frame of the VOTECH project. It is a pragmatic choice because it is easy to modify the sources to add for example data types (important in workflows) and the VO standards are partially taken into account.

6. Workflow use cases in astronomy

In the frame of the VO France workflow working group we have discussed about use cases in different domains.

6.1. Image processing (E. Slezak)

Astronomers, Phd students and trainees develop image processing tools in different languages and on different plateforms. It is not easy to maintain this work, to create new algorithms including previous developements or to reuse them in other project. The AÏDA architecture has been experimented as a solution to these different problems in the image processing domain. See fig 1.

6.2. Simulation (F. Le Petit et al.)

The use of workflows is useful to formalize the different steps of simulations. Some of these steps can require a long time to end. In this case a workflow managing tool can provide interesting features like the control of the errors,

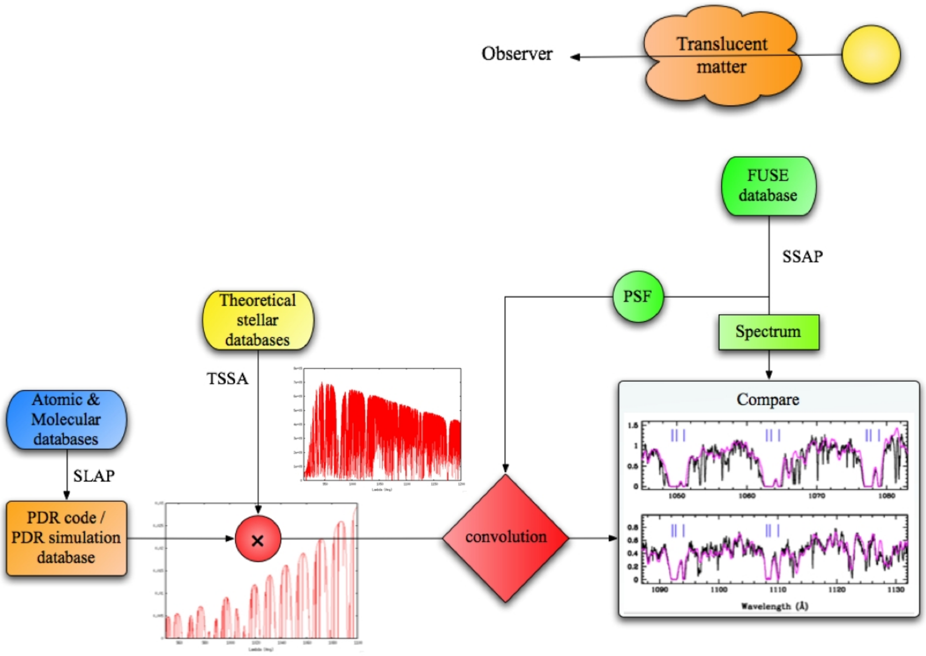

Fig. 2. Example of workflow in the simulation domain

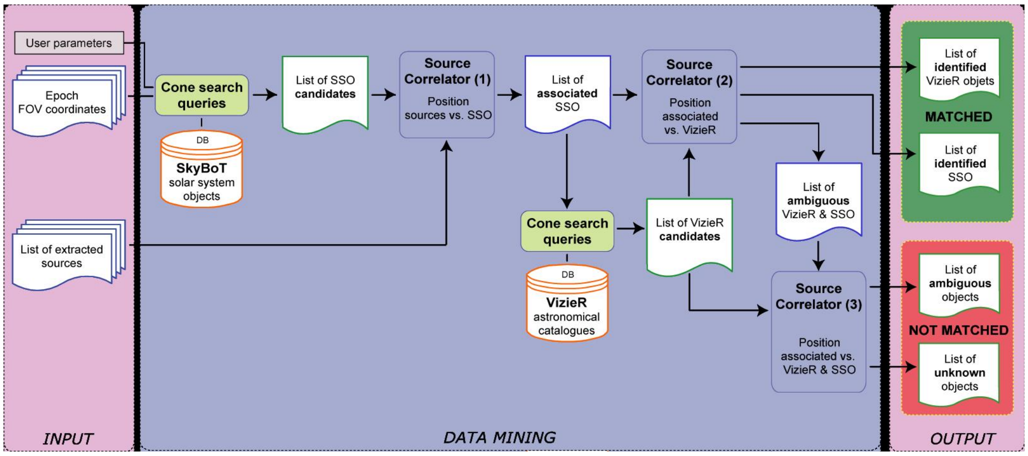

Fig. 3. Example of workflow in the data mining domain

the possibility to replay the scenario from an advanced step, etc. See fig 2.

6.3. Data mining (J. Berthier et al.)

Other example of domain in which the use of workflow is experimented. See fig 3.

7. Workflows and IVOA standards

Workflows are concerned by many IVOA standards but we focus on one of them, the Characterization. (See http://www.ivoa.net).

7.1. Aim of the Characterization

The Characterisation defines a highest metadata level necessary to describe the physical parameter space of observed or simulated astronomical data sets. We have applied the Characterization in the context of FITS images. FITS is a forty years old standard for astronomical images. A FITS file is composed by i) a header describing the metadta and ii) the data itself. The FITS header is based on FITS keywords which are not all standardised. The keywords are not always suffisant (for the use in tools for example) and the Characterization of the data is useful to go further.

7.2. Characterisation in a workflow

For each task of a workflow a constraints file is created. This file describes what should be the entries of the tasks. It is very far from a simple typing.

7.3. How it works

Before the execution of the workflow we do a validation. It starts a review of all the tasks having prefixed data for some of their entries. The entries receiving data during the execution could not be checked in a first step.

During the execution, the controles should be done for the produced data. But it means that a task is able to generate a Characterization file associated to the data. In many cases we will not have this file and in this case it is necessary to produce it automatically from the data.

7.4. The work done

We have created a Characterization generator. It extracts all the keywords of a FITS file and provides a mapping used to produce the Characterization file. As the keywords are often not suffisant an interaction with the user is necessary to create the mapping for a first set of data. The user manipulations are saved in a Knowledge base. As in astronomy we work with collections (common caracteristics) of images, doing this work for a sample of the data is suffisant to be able to automatize it for the rest of the collection.

7.5. Interaction with other IVOA standards

Other standards like the Registry are also interesting for the creation of flexible and efficient workflows. The Registry can be requested to find the best (pertinency, availability, etc.) tools to use in a workflow and these choices can be modified on the fly during the execution if necessary.

8. Conclusion and perpectives

We are now testing the characterization file creation for different workflow use cases in different domains. We are able to validate partially a workflow before its execution and it is possible to improve it by a first interaction with the astronomer (creation of the mapping file). The next step is to apply this concept to other kind of data and to increase the field of the automatization. But we think that in many cases a minimal interaction with the astronmer will be necessary to add metadata.

Acknowledgements. Massive Data in Astronomy project (ACI Masses de Données of French ministry), VOTECH European project and Action Spécifique OV France

References

A. Schaaff, F. Le Petit, P. Prugniel, E. Slezak & C. Surace 2008 in ASP Conf. Ser. XXX, ADASS XVII, ed. J. Lewis, R. Argyle [O4a.4]

M. Louys, F. Bonnarel, A. Schaaff, J.-J. Claudon & C. Pestel 2008 in The Virtual Observatory in Action: New Science, New Technology, and Next Generation Facilities, 26th meeting of the IAU

E. Slezak & A. Schaaff 2006 in The Virtual Observatory in Action: New Science, New Technology, and Next Generation Facilities, 26th meeting of the IAU

A. Schaaff, F. Le Petit, P. Prugniel, E. Slezak & C. Surace 2007 in ASP Conf. Ser. XXX, Journes SF2A, ed. J. Lewis, R. Argyle