International

Virtual

Observatory

Alliance

Scientific Workflows in the VO

Version 1.00 (Draft for comments)

IVOA Note 2011 October 14

This version:
ThisVersion-20111014
Latest version:
http://www.ivoa.net/Documents/latest/latest-version-name
Previous version(s):

Author(s):
A. Schaaff, J.E. Ruiz et al.

Editor(s):

Abstract

We will soon be facing a new generation of facilities and archives dealing with huge amounts of data (ALMA, LSST, Pan-Starrs, LOFAR, SKA pathfinders...) where scientific workflows will play an important role in the working methodology of astronomers. A detailed analysis about the state of the art of workflows in the frame of the VO involves languages, design tools, execution engines, use cases, etc. A major topic is also the preservation of the workflows and the capability to replay a workflow several years after its design and implementation. Several talks concerning these issues have been presented during the past IVOA Interoperability meetings. In order to undertake this task within our community we have decided as a first step to write this Note. We have collected experiences (including use cases, tools, etc.), references, remarks, etc.

Status of This Document

This is a Note. The first release of this document was 2011 October 14.

This is an IVOA Note expressing suggestions from and opinions of the authors. It is intended to share best practices, possible approaches, or other perspectives on interoperability with the Virtual Observatory. It should not be referenced or otherwise interpreted as a standard specification.

A list of current IVOA Recommendations and other technical documents can be found at http://www.ivoa.net/Documents/.

Acknowledgements

All the participants registered on the workflow@ivoa.net mailing list.

Contents

1 Introduction

One of the current challenges in Astronomy is the efficient exploitation of the huge volume of data currently available. This efficiency is needed in order to ensure the prompt return of the big investments made in terms of facilities to obtain those data, something that clearly the traditional methods of analysis are not currently achieving. This is one of the most important reasons why scientific workflows are becoming a need in Astronomy.

Publishing data and processes as the methodology used in an astronomical digital experiment will need Virtual Observatory standards for the characterization of workflows, in order to be indexed, shared, and retrieved.

In this Note we intend to provide a very quick revision of the state of the art in the domain of scientific workflows, from general technical topics like languages and formalisms, composition tools and engines, through more astronomy specific related initiatives and concerns in the frame of the VO as well as in different VO Working Groups.

2 State of the Art

2.1 Definition

This concept is used to refer in general to modelling and IT management of all tasks and actors in the composition of a business process. The goal is to automate the best working procedures. It should be noted that, commonly, the term workflow is the process that the system used in modelling. For the Workflow Management Coalition, the term procedure is mainly used to discuss the process and workflow for the software to model it. In this note we will use the term workflow also for the process. There are two main types of workflows: business workflows and scientific workflows. We give a quick definition of business workflows even if the goal of this note is to focus on the scientific workflows. Workflow's world is very wide and it is possible de give many examples of workflow languages, engines, integrated tools, etc.

2.1.1 Business workflows

The business workflows (BWFs) appeared in the 70s and the definition that we retain is that given by the WFMC:

The automation of a business process, in whole or parts, where documents, information or tasks are passed from one participant to another to be processed, according to a set of procedural rules.

More practically, they can automate work processes within companies, to which was previously done by hand. The BWFs are software-oriented tasks to perform complex workflows with a major control-flow.

2.1.2 Scientific workflows

The scientific workflows (SWFs) are a variant of BWFs. They are relatively similar but have different features that are not present in BWFs. We retain Bertram Ludäscher's definition:

These are networks of analytical steps that may involve, e.g., database access and querying steps, data analysis and mining steps, and many other steps including computationally intensive jobs on high performance cluster computers.

This type of workflow is designed for scientists and, therefore, is able to meet their specific needs. Therefore, while BWFs are oriented control-flow, the SWFs are in contrast, data-flow oriented. They give the opportunity for users to operate easily in a large number of complex and heterogeneous data, computationally intensive and distributed processing.

Workflows are useful to capture scientific methodology and to provide provenance information for their results. They provide also a formalization of the Scientific analysis (routines to be executed, dataflow, execution details . . .) and they are structures useful to manage computation at a large-scale. A large number of projects have defined their workflow language and the associated tools (engine, composition . . .).

2.1.3 Towards a new type of workflow

In recent years, given the popularity of workflows and to meet new expectations, a new type of workflow has emerged, which we can call "adaptive workflows". In the literature, we can see it under different names, i.e. "WDOs" (Workflow-Driven Ontologies) or "flexible workflow." The main characteristic of this type of workflow is to offer the ability to change, more or less automatically, the structure of a workflow during its execution. It takes into account the execution environment of the workflows. This new kind of workflows is based on ontologies.

2.2 Languages and formalisms

A workflow language gives a way to describe a workflow and to make its execution possible through a workflow engine. It is like a programming language. A workflow could be defined at least with a simple script language. Sculf is an XML-based language associated to Taverna.

Other examples: AGWL, BPEL4WS, BPML, DGL, DPML, GJobDL, GSFL, GFDL,GWorkflowDL, MoML, SWFL, WSCL, WSCI, WSFL, XLANG, YAWL, WPDL, PIF, PSL, OWL-S, xWFL…

Workflow formalism is at the modelling level. UML activity diagram is a well-known example.

Other examples: Petri net, BPMN, DAG, IPO, GPSG, Workflow Patterns, Pi Calculus, Finite-State Machine, Gamma calculus . . .

The need for a standard is justified by the fact that all the workflow tools are based on a language of their own as well as a model of relations between objects and a set of commands for the transfer of information between participants.

2.3 Workflow composition and enactment

2.3.1 Design tools

The process definition tools are tools to model the workflow to be performed. Thus, in most cases, these tools have graphical features for easy drag and drop tasks and actors in the composition of their processes. Existing communications between the entities are then defined by linking them just as easily.

Examples: ilog's BPMN Modeller, CAT, GWUI, XBaya GUI for Workflow Composition, Triana, JOpera, Platform Process Manager. . .

2.3.2 Workflow engines

The workflow engine is a software service that provides and controls all or only a part of the runtime of a workflow instance.

Examples: BioPipe, BizTalk, BPWS4J, DAGMan, GridAnt, Grid Job Handler, GRMS, GWFE, GWES, IT Innovation Enactment Engine, JIGSA, JOpera, Kepler, Karajan, OSWorkflow, Pegasus (uses DAGMan), Platform Process Manager, ScyFLOW, SDSC Matrix, SHOP2, Taverna, Triana, wftk, YAWL Engine, WebAndFlo, WFEE. . .

2.3.3 Workflow Enactment System

At the heart of the workflow is the Workflow Enactment System. It is service to create, manage, run instances of procedures and manage their interactions with the outside. It is composed of one or more workflow engines that allow it to maintain an internal control data centrally or distributed.

2.4 User tools

Our goal here is not to give an exhaustive list of all existing workflows tools for final users. We look more toward scientific workflows and we include a few as examples, but there are many others.

Taverna is a strongly typed bioinformatics workflow management system developed by the European Bioinformatics Institute and the University of Manchester. It aims to provide a language and software tools to facilitate the use of workflow and distributed computing in the scientific community. The Taverna suite includes the Taverna Engine that powers both the Taverna Workbench and Server which allows remote execution of workflows.

Kepler is a generic science oriented workflow system (ecology, bioinformatics, geology...), which would tend to be universal.
It is based on Ptolemy II system developed by the EECS (Electrical Engineering and Computer Science). A set of actors is defined and their performances are under the supervision of one or more directors who determine the semantics to apply to the links between the actors.

Triana is a workflow system originally built to provide a tool for rapid analysis of data from gravitational waves. At the beginning, the procedures were modelled and executed locally or remotely using RMI. Recently, Triana has been extended to incorporate components that are distributed, grid computing-oriented or Web Services oriented.

MyExperiment is a social networking site for workflow exchange and sharing, with 3000 members and 1000 workflows representing 10 workflows management systems. As in the case of Taverna, this Virtual Research Environment is mainly used by bioinformatics, enabling users to upload and find publicly shared workflows, promoting building of communities, forming of relationships and collaboration.

3 Related initiatives

3.1 ESO Reflex

ESO Reflex is a graphical workflow system for running ESO reduction recipes and related tools in a flexible manner. Initially developed within the SAMPO project as a proof of concept, it was based on a modified version of the original implementation of Taverna. It allowed the user to define and execute a sequence of recipes using an easy and flexible GUI. Instead of running the recipes one at a time, a sequence of recipes can be run as a workflow where the output of a recipe is used as an input to another recipe. It was focused on ESO pipelines for astronomical data reduction. The power of the workflow as an entity encompassing the tasks typically assigned to scripts, combined with the

additional semantics which actual encode the data reduction recipes, have made ESO continue the incarnation of ESO Reflex, this time based on the Kepler workflow engine.

3.2 AstroGrid

AstroGrid, the UK's Virtual Observatory System, developed the AstroGrid Workflow System, a multi-user batch system for the execution of potentially long-running astronomical workflows. The input is a workflow document describing which remote applications — data collections and processing packages — are to be used. These applications may be distributed throughout the VO, some of them may be implemented in CEA servers. The CEA (Common Execution Architecture) defines the Web service interfaces, message protocols, and formats that an executable application must support in order to be fully compliant with VO standards. The results and intermediate products of the workflow are stored in MySpace.

AstroGrid also developed a version of the Taverna v1 Workbench (AstroTaverna) with VO plug-ins, which added a number of significant capabilities. The AstroGrid implementation of Taverna relies on the Astro Runtime, a client side library of functions to access the Virtual Observatory.

3.3 VO France & CDS

A Workflow working group has started to work in 2005 in the frame of OV France. The aim was to provide use cases and to implement them as workflows with a (VO or other) workflow tool. The CDS has developed AÏDA (Astronomical Image processing Distribution Architecture) during the MDA (Masses de Données en Astronomie) French Ministry funded project and the European VOTECH project.

AÏDA has 2 sides, one at the server level to execute a workflow and one at the user level as it provides a graphical composition tool based on JGraph. This tool is able to validate the data (FITS images) before the execution at each step of the workflow through the IVOA Characterisation implementation: each FITS image.

3.4 Helio-VO

The HELIO project is a domain-specific virtual observatory for solar physics that is being built, not only with data access and sharing in mind, but with the actual description of the knowledge in the field (via ontologies), and their processes (via workflows). One of its main achievements is having enabled Taverna to run on Grid or Cloud based resources, thus greatly expanding its potential in Astronomy. Processing and storage services will allow the user to explore the data and create the products. These capabilities will be orchestrated with the data and metadata services using the Taverna workflow tool.

3.5 CyberSKA

CyberSKA is a project aimed at exploring and implementing the cyber-infrastructure that will be required to address the evolving data intensive science needs of future radio telescopes such as the Square Kilometre Array. They are developing a web based workflow builder that supports image segmentation, image mosaicking, spatial reprojection, and plane extraction from data cubes. These actions and processes contained in the workflow are provided as web services, which automatically determine the most efficient course of action regarding where data is to be retrieved from and processed.

3.6 Wf4Ever

The EU FP7 funded project - Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science – started in December 2010 with the main intend to contribute to the development of standards and models for the preservation of scientific workflows. Wf4Ever considers complex digital objects (Research Objects) that include workflow models, the provenance of their executions, and interconnections between workflows and related resources. This project will investigate and develop technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows in a range of disciplines, including Astronomy.

4 Workflows preservation

The preservation of workflows as complex digital experiments is an important issue where methodology, processes and data need a common preservation strategy in order to achieve reproducible procedures and repeatable results through large periods of time.

Workflows and their components, as digital entities, need specific applications to be interpreted and re-executed. These, in turn, need specific libraries installed on a specific operating environment**,** which runs on very specific hardware configurations for which drivers are provided. All of these factors combine to ensure that workflows are severely vulnerable to obsolescence: if any of the layers in the dependency tree is lost, the entire object ceases to be accessible and usable. On top of that, we find vulnerabilities regarding the interpretation of workflows and data, documenting their provenance and limitations, and ensuring that they are authentic and trustworthy.

As a first approach to preservation of workflows we can consider the basic steps for software preservation: preserve, retrieve, reconstruct and replay. For retrieval, in addition to knowledge of general software architecture, there is a need for explicit information on the software's functionality. With reconstruction there is a need for understanding the dependencies and components, details on program

language and the libraries required to ensure the correct output. Replay will also need sufficient documentation and might be used as a benchmark to assess the success of the preservation method.

We should consider the preservation of all digital entities involved in a workflow, taking into account the provenance of the final results, which is especially complex in a cloud of services. Given a predicted rise in the number of openly available web services and workflows, it would seem necessary, to curate processes as effectively as we curate the data they consume and the publications they generate. We should be able to find a workflow or process based on what it does, what it consumes as inputs and produces as outputs, and find copies or similar services usable as alternates.

Other issues to be considered are permissions and licenses concerning infrastructure requirements or proprietary data, versioning of workflows and of its components, classification and indexation in semantic repositories for them to be retrievable, referenced and acknowledged.

5 Workflows in the VO

Unlike traditional pipelines, which tend to produce scientifically exploitable results, most of the scientific workflows in the Virtual Observatory should be aimed at producing scientific insight. They should be easily accessible to a wide range of non-highly specialised technical users, allowing an effortless design, composition and execution. The complete digital characterization of workflows should describe the scientific methodology used in an experiment in its entirety.

VO services could be used as components for internet-based workflows. Since their execution is independent of the investigator's platform, they ensure the reproducibility of the results and their dissemination given their modularity, and their universal availability.

5.1 Distributed data analysis workflows

In this case a user or a client defines and executes a distributed workflow, which invokes services on multiple remote sites via the VO infrastructure. The workflow would be entirely in VO-space, driving simpler services at the individual sites.

The AstroTaverna developments provided a graphical tool for the composition and design of workflows based on VO services and data from different archives and facilities.

5.2 Data processing pipelines

Traditional data processing pipelines, e.g., instrumental or survey data processing pipelines, which produce higher, level data products. At present

there are many variants of these and they have little or no direct connection to VO, aside from possibly producing VO-compliant data or being optionally driven from VO.

It is not clear how much VO mechanisms are needed at this level (VO compliant data and metadata, modelling provenance, etc.)

5.3 Driving data processing pipelines from VO

In this case we have a traditional data processing pipeline and the remote user or client software invokes a job to do some pipeline reprocessing, e.g., to custom reprocess an instrumental dataset to produce a new image, cube, etc. The "workflow" in this case runs at a single site, and VO is used to drive the job remotely (SSO, UWS) and manage the results (VOSpace, VO data services).

We could think on integrating the traditional data processing pipelines we already have with VO, to allow VO users to do on-the-fly reprocessing to generate data products which can be analyzed with VO (custom reprocessing of observatory data for example)

Some attempts to integrate general processing applications have been made with CEA and UWS.

6 Workflows and IVOA standards

Remark: this part will be extended and documented for the published version of the Note

6.1 Data Model

Characterisation, Provenance

6.2 Semantics

UCD, Ontologies, Vocabularies

6.3 Data Access Layer

Self-descriptive Web Services

6.4 Grid and Web Services

UWS, VOSpace, SSO

6.5 Data Curation and Preservation

Permanent identifiers

6.6 Application

SAMP

6.7 Knowledge Discovery in Databases

To be defined.

7 Proposal

The quantitative leap in volume and complexity of the next generation of archives will need analysis and data mining tasks to live closer to the data, in computing and distributed storage environments, but they should also be modular enough to allow customization from scientists and be easily accessible to foster their dissemination among the community.

Astronomy is a collaborative science, but it has also become highly specialized, as many other disciplines. Sharing, preservation, discovery and a much simplified access to resources in the composition of scientific workflows will enable astronomers to greatly benefit from each other's highly specialized know-how, they constitute a way to push Astronomy to share and publish not only results and data, but also processes and methodologies.

This disruptive transformation in the way digital experiments are designed, performed, shared and preserved in Astronomy cannot be done outside the Virtual Observatory, where workflows, processes and services should benefit of the same privileges acquired by data.

References

Remark: references will be ordered for published version of this Note

[1] R. Hanisch, *Resource Metadata for the Virtual Observatory*, http://www.ivoa.net/Documents/latest/RM.html
[2] R. Hanisch, M. Dolensky, M. Leoni, *Document Standards Management: Guidelines and Procedure*, http://www.ivoa.net/Documents/latest/DocStdProc.html
[3] A. Schaaff et al., ADASS London 2007, http://adsabs.harvard.edu/abs/2008ASPC..394...77S
[4] P. Järveläinen et al., ADASS London 2007, http://adsabs.harvard.edu/abs/2008ASPC..394..273J
[5] A. Schaaff, IVOA Trieste GWS session 2008, http://www.ivoa.net/internal/IVOA/InterOpMay2008GridAndWebServices/GWS-Charac-Workflow-20May08.pdf
[6] A. Schaaff and F. Bonnarel, IVOA Baltimore Application session 2008, http://www.ivoa.net/internal/IVOA/InterOpOct2008Applications/GWS-DM-REG-301008.pdf
[7] A. Schaaff et al., Euro-VO DCA Theory and Grid Workshops 2008, http://sait.oat.ts.astro.it/MSAIt800209/PDF/559.pdf
[8] K. Benson and N.A. Walton, Euro-VO DCA Theory and Grid Workshops 2008, http://sait.oat.ts.astro.it/MSAIt800209/PDF/574.pdf
[9] R. Hook et al., Euro-VO DCA Theory and Grid Workshops 2008, http://sait.oat.ts.astro.it/MSAIt800209/PDF/578.pdf
[10] M.J. Graham, IVOA Garching GWS session 2009, http://www.ivoa.net/internal/IVOA/InterOpNov2009GWS/Garching-GWSWorkflow.pdf
[11] J.E. Ruiz, IVOA Nara DCP session 2010, http://www.ivoa.net/internal/IVOA/InterOpDec2010DCP/Wf4Ever.pdf
[12] J.E. Ruiz, IVOA Naples DCP session 2011, http://www.ivoa.net/internal/IVOA/InteropMay2011DCP/Wf4EverNaples.pdf
[13] J.E. Ruiz, IVOA Pune DCP session 2011, http://www.ivoa.net/internal/IVOA/InterOpOct2011DCP/Wf4EverPune.pdf
[14] VO France Workflow WG, http://www.france-ov.org/twiki/bin/view/GROUPEStravail/Workflow
[15] Wf4Ever project, http://www.wf4ever-project.org/web/guest/home
[16] Scientific Workflows in Astronomy BoF, ADASS 2011, http://www.eso.org/sci/php/meetings/adass2011/html/display.php?topic=BoF_Schaaff_1314702958.html
[17] N. A. Walton et al, ADASS London 2007 http://adsabs.harvard.edu/abs/2008ASPC..394..309W
[18] Workflow Coalition Management, http://www.wfmc.org/
[19] Find, use, share scientific workflows and other Research Objects, and to build communities, http://www.myexperiment.org/
[20] Sharing Interoperable Workflows for large-scale scientific simulations on Available DCIs, http://www.shiwa-workflow.eu/
[21] Workflow-Driven Ontologies, http://trust.utep.edu/wdo/
[22] ESO Reflex, http://www.eso.org/sampo/reflex
[23] AstroGrid, http://www2.astrogrid.org/

http://www.ivoa.net/Documents/latest/AstrogridWorkflow.html
http://www.ivoa.net/Documents/Notes/CEA/CEADesignIVOANote-20050513.html
http://www.ivoa.net/Documents/latest/IntroductionCEA_UWS.html
[24] Helio-VO, http://www.helio-vo.eu/
[25] CyberSka, http://www.cyberska.org/
[26] Wf4Ever, http://www.wf4ever-project.org/
[27] SWFs definition, http://www.sdsc.edu/~ludaesch/Paper/kepler-swf.pdf