Volume Title
ASP Conference Series, Vol. **Volume Number**
Author

Scientific Workflows in Astronomy

André Schaaff[1], L. Verdes-Montenegro[2], J. E. Ruiz[2], J. Santander Vela[3]

[1]*CDS, CNRS, Observatoire Astronomique de Strasbourg, 11 rue de l'Université 67000 Strasbourg, France*

[2]*Instituto Astrofisica Andalucia (CSIC), Glorieta de la Astronomia s/n, 18008 Granada, Spain*

[3]*European Southern Observatory, ALMA Archive Subsystem Data Flow Infrastructure Department, Software Development Division, 85748 Garching bei München, Germany*

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. While the traditional pipelines tend to produce exploitable products, scientific workflows are aimed at producing scientific insight. Virtual Observatory standards provide the tools to design reproducible scientific workflows. A detailed analysis about the state of the art of workflows 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. Discussions on these topics are being held recently in IVOA forums and are part of the work that is being done in the Wf4Ever project. The purpose of the BoF was to present to the community the work in progress at the IVOA, collect ideas and identify needs not yet addressed.

1. Definition of a Scientific Workflow

The workflow concept is used to refer in general to modelling and IT management of all tasks and actors in the composition of a (control-flow oriented) 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. The scientific workflows (data-flow oriented) are a variant of business workflows. They are designed for scientists and, therefore, are able to meet their specific needs and they are useful to provide a formalization of the Scientific analysis (routines to be executed, dataflow, execution details . . .).

2. Related initiatives

2.1. ESO Reflex

ESO Reflex (`http://www.eso.org/sampo/reflex`) is a graphical workflow system for running ESO reduction recipes and related tools in a flexible manner. Initially de-

veloped within the SAMPO project as a proof of concept, it was based on a modified version of the original implementation of Taverna (now based on the Kepler workflow engine) and could be used through Web Services ((Järveläinen et al. 2008)). It allows 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 (Hook et al. 2009) for astronomical data reduction.

2.2. Wf4Ever

The Wf4Ever project (`http://www.wf4ever-project.org/`) : 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.

2.3. AstroGrid

AstroGrid (`http://www2.astrogrid.org/`) developed 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. AstroGrid also developed a version (with VO plug-ins) of Taverna ((Walton et al. 2008)) which relies on the Astro Runtime, a client side library of functions to access the Virtual Observatory ((Benson & Walton 2009)).

2.4. VO France Workflow working group and CDS

The aim was to provide use cases and to implement them with VO (or other) workflow tools ((Schaaff et al. 2008)). A tool, AÏDA (Astronomical Image processing Distribution Architecture), was developed during the MDA (Masses de Donnes en Astronomie) project and the European VOTECH project. It has both a server level to execute a workflow and a user graphical composition tool based on JGraph. It implements the IVOA Characterisation ((Schaaff et al. 2009)) to validate the data (FITS images). See `http://www.france-ov.org/twiki/bin/view/GROUPEStravail/Workflow`.

2.5. Helio-VO

The HELIO project (`http://www.helio-vo.eu/`) 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. These capabilities will be orchestrated with the data and metadata services using the Taverna workflow tool.

2.6. CyberSKA

CyberSKA (`http://www.cyberska.org/`) 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.

3. Workflow 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 softwares 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.

4. Workflows in the Virtual Observatory

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. In distributed data analysis workflows 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 VOSpace, driving simpler services at the individual sites. Data processing pipelines e.g., instrumental or survey data processing pipelines 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. 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.

5. BoF : Scientific Workflows in Astronomy

We had a presention of Wf4Ever (cf. 2.6) followed by a presentation of the IceCore Portal (LifeRay based). This portal is motivated by working both with Taverna and Kepler in the Sampo project. It hides the complexity of installing analysis software and it provides a complete platform independency. It is extended with a knowledge bus which provides interoperabily in messaging between applications both server-side and client-side. Results are kept along with the runs. It was followed by a demo of ESO Kepler Workflows for UVES. We had then a presentation which was a reflexion about the workflows a the different levels of a project (top-down design typically used for developments and many levels down sometimes). A high-level grouping allows for knowing the algorithms, not the code but we will need the detail later on. As a conclusion, workflows as tools for better thinking. The last presentation was about OWL which could be seen as a sum of Condor, Blackboard and Dynamic workflows. Open questions and discussions are preserved at : `http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/ADASSXXIBoF4`

6. Conclusion

The aim of the BoF was to have a discussion with all the interested people to take into account their experiences and needs concerning Scientific Workflows in the future plans of both VO and Wf4Ever projects. The VO provides and works on standards which should be useful for the execution and the long term preservation of workflows.

References

Benson, K. M., & Walton, N. A. 2009, "Memorie della Società Astronomica Italiana", 80, 574

Hook, R., Ullgrén, M., Romaniello, M., Maisala, S., Oittinen, T., Solin, O., Savolainen, V., Järveläinen, P., Tyynelä, J., Péron, M., Ballester, P., Gabasch, A., & Izzo, C. 2009, "Memorie della Società Astronomica Italiana", 80, 578

Järveläinen, P., Savolainen, V., Oittinen, T., Maisala, S., & Ullgrén, M. H., R. 2008, in Astronomical Data Analysis Software and Systems XVII, edited by R. W. Argyle, P. S. Bunclark, & J. R. Lewis, vol. 394 of Astronomical Society of the Pacific Conference Series, 273

Schaaff, A., Bonnarel, F., Louys, M., Slezak, E., Gassmann, B., Pestel, C., Benjelloun, O., & Mantelet, G. 2009, "Memorie della Società Astronomica Italiana", 80, 559

Schaaff, A., Petit, F. L., Prugniel, P., Slezak, E., & Surace, C. 2008, in Astronomical Data Analysis Software and Systems XVII, edited by R. W. Argyle, P. S. Bunclark, & J. R. Lewis, vol. 394 of Astronomical Society of the Pacific Conference Series, 77

Walton, N. A., Witherwick, D. K., Oinn, T., & Benson, K. M. 2008, in Astronomical Data Analysis Software and Systems XVII, edited by R. W. Argyle, P. S. Bunclark, & J. R. Lewis, vol. 394 of Astronomical Society of the Pacific Conference Series, 309