Workflows in the VO

André Schaaff

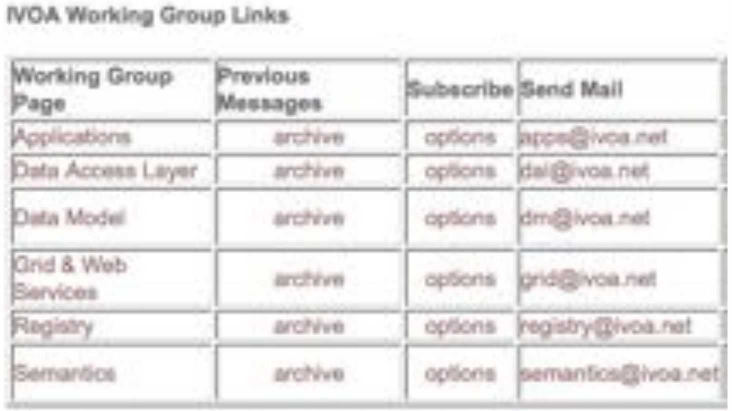
PREDON meeting, Strasbourg 9/12/2015

□ Foreword

- Studies and experiments around workflows in the Virtual Observatory
  - Started in 2005 in the frame of the VO France Workflow Working Group
  - Papers, posters, developments & presentations (IVOA, Euro-VO projects), etc..
  - Study at the IVOA level producing a survey (IVOA Note in 2013) of the tools, methods, …, inside (but also outside) the community

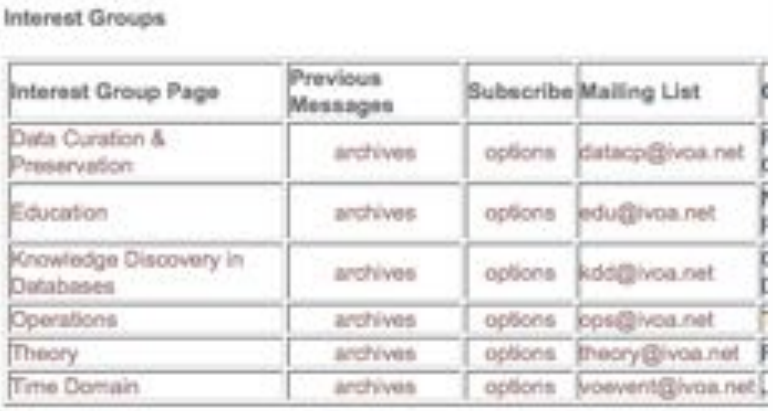
□ IVOA ? WG ? IG ?

- International Virtual Observatory Alliance



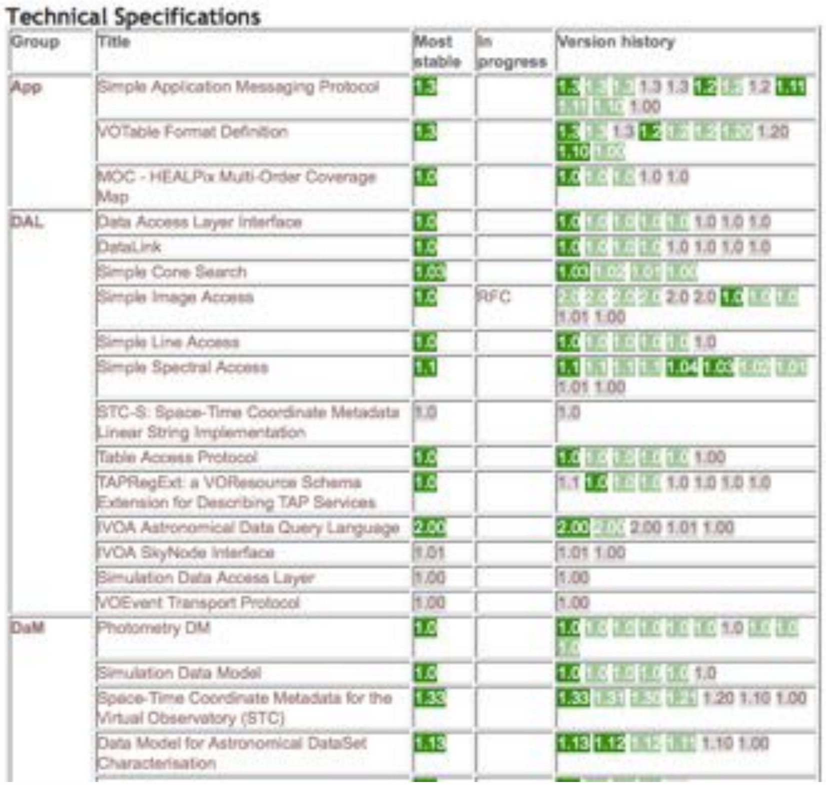
IVOA Working Group Links

Working Group Page	Previous Messages	Subscribe	Send Mail
Applications	archive	options	apps@ivoa.net
Data Access Layer	archive	options	dal@ivoa.net
Data Model	archive	options	dm@ivoa.net
Grid & Web Services	archive	options	grid@ivoa.net
Registry	archive	options	registry@ivoa.net
Semantics	archive	options	semantics@ivoa.net



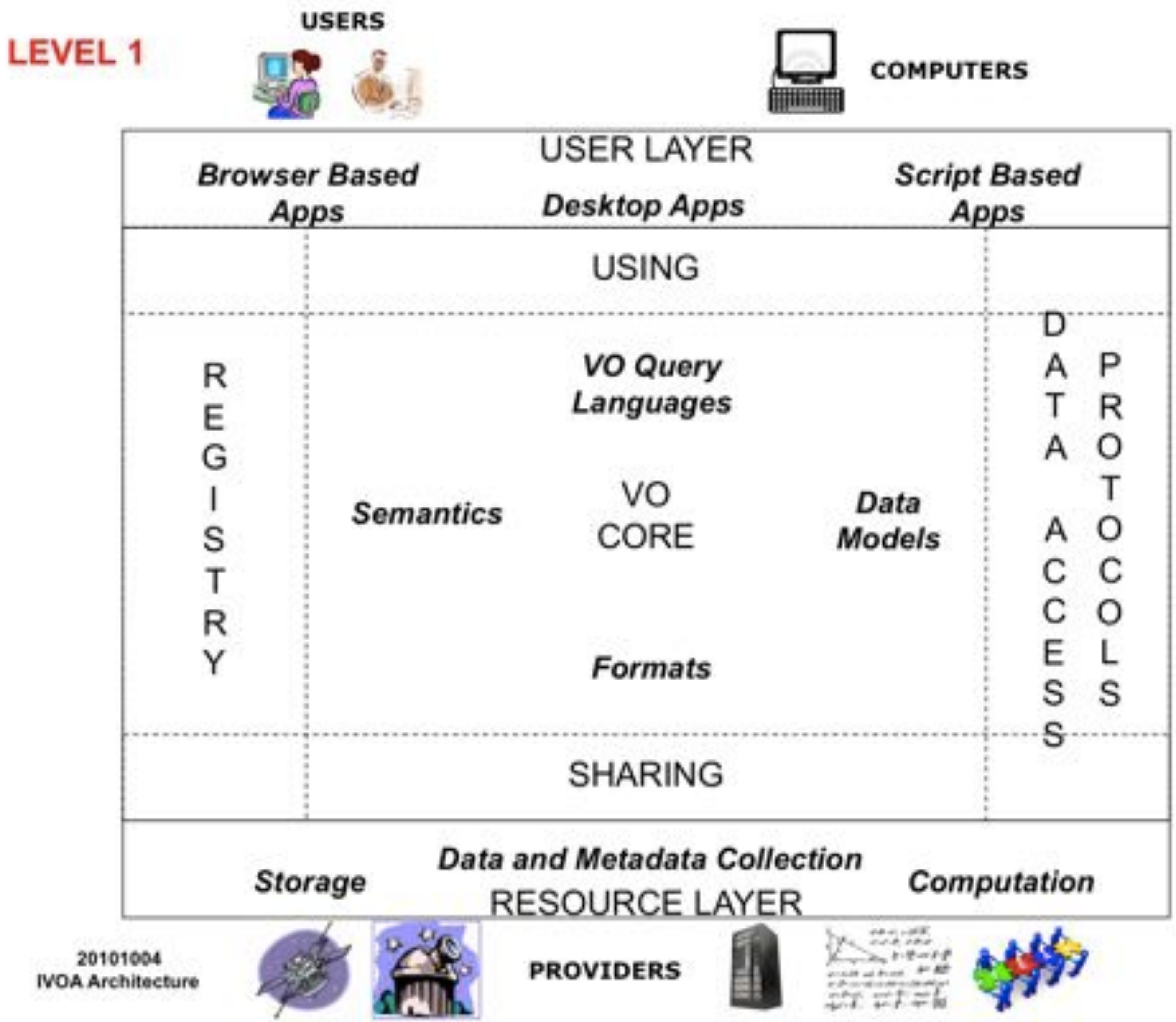
Interest Groups

Interest Group Page	Previous Messages	Subscribe	Mailing List
Data Curation & Preservation	archives	options	datacp@ivoa.net
Education	archives	options	edu@ivoa.net
Knowledge Discovery in Databases	archives	options	kdd@ivoa.net
Operations	archives	options	ops@ivoa.net
Theory	archives	options	theory@ivoa.net
Time Domain	archives	options	voevent@ivoa.net



Technical Specifications

Group	Title	Most stable	In progress	Version history
App	Simple Application Messaging Protocol	1.3		1.3 1.3 1.3 1.3 1.3 1.2 1.2 1.2 1.11 1.11 1.10 1.00
	VOTable Format Definition	1.3		1.3 1.3 1.3 1.2 1.2 1.2 1.20 1.20 1.10 1.00
	MOC - HEALPix Multi-Order Coverage Map	1.0		1.0 1.0 1.0 1.0 1.0
DAL	Data Access Layer Interface	1.0		1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	DataLink	1.0		1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	Simple Cone Search	1.03		1.03 1.03 1.03 1.00
	Simple Image Access	1.0	RFC	2.0 2.0 2.0 2.0 2.0 2.0 1.0 1.0 1.0 1.01 1.00
	Simple Line Access	1.0		1.0 1.0 1.0 1.0 1.0 1.0
	Simple Spectral Access	1.1		1.1 1.1 1.1 1.1 1.04 1.03 1.02 1.01 1.01 1.00
	STC-S: Space-Time Coordinate Metadata Linear String Implementation	1.0		1.0
	Table Access Protocol	1.0		1.0 1.0 1.0 1.0 1.0 1.00
	TAPRegExt: a VOResource Schema Extension for Describing TAP Services	1.0		1.1 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	IVOA Astronomical Data Query Language	2.00		2.00 2.00 2.00 1.01 1.00
	IVOA SkyNode Interface	1.01		1.01 1.00
	Simulation Data Access Layer	1.00		1.00
	VOEvent Transport Protocol	1.00		1.00
DaM	Photometry DM	1.0		1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	Simulation Data Model	1.0		1.0 1.0 1.0 1.0 1.0 1.0
	Space-Time Coordinate Metadata for the Virtual Observatory (STC)	1.33		1.33 1.31 1.30 1.21 1.20 1.10 1.00
	Data Model for Astronomical DataSet Characterisation	1.13		1.13 1.12 1.12 1.11 1.10 1.00



LEVEL 1

USERS

COMPUTERS

USER LAYER

Browser Based Apps

Desktop Apps

Script Based Apps

USING

REGISTRY

VO Query Languages

Semantics

VO CORE

Data Models

Formats

DATA ACCESS PROTOCOLS

SHARING

Storage

Data and Metadata Collection

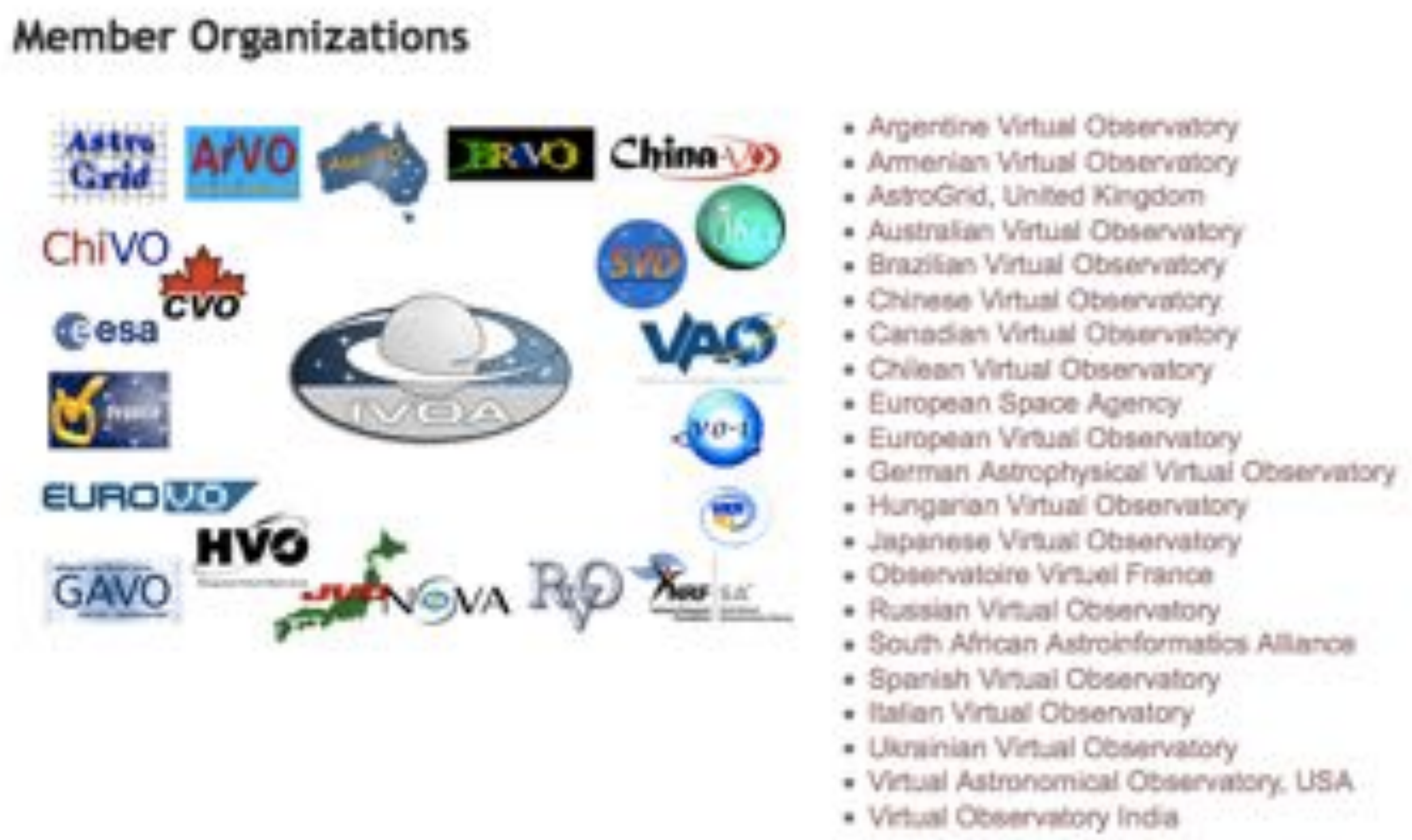
Computation

RESOURCE LAYER

20101004
IVOA Architecture

PROVIDERS

□ IVOA, who is involved ?



Member Organizations

- Argentine Virtual Observatory
- Armenian Virtual Observatory
- AstroGrid, United Kingdom
- Australian Virtual Observatory
- Brazilian Virtual Observatory
- Chinese Virtual Observatory
- Canadian Virtual Observatory
- Chilean Virtual Observatory
- European Space Agency
- European Virtual Observatory
- German Astrophysical Virtual Observatory
- Hungarian Virtual Observatory
- Japanese Virtual Observatory
- Observatoire Virtuel France
- Russian Virtual Observatory
- South African Astroinformatics Alliance
- Spanish Virtual Observatory
- Italian Virtual Observatory
- Ukrainian Virtual Observatory
- Virtual Astronomical Observatory, USA
- Virtual Observatory India

□ Workflow systems

- Different tools executed by hand under shell
  - Not very efficient but could be a solution if the tools are interactive
- Script files
  - Same as previous but a first step to the formalization
  - Involved tools are mainly black boxes (or converted to) with just I/O
- Description language
  - The workflow is described in a specific language (often in XML format) interpreted by a workflow engine

□ Workflow systems (2)

- « Sophisticated » workflow system
  - Graphical design tool
  - Workflow description (XML, ...) is sent to an engine which executes the workflow (tasks dispatching)
  - Execution (often) visible step by step
  - Possible storage of intermediate data to change some parameters without re-executing the whole workflow
  - Result(s) exploited through tools related to the kind of output data (FITS, ...)

□ Workflow jungle

- Languages: AGWL, BPEL4WS, BPML, DGL, DPML, GJobDL, GSFL, GFDL, GWorkflowDL, MoML, SWFL, WSCL, WSCI, WSFL, XLANG, YAWL, SCUFL/XScufl, WPDL, PIF, PSL, OWL-S, xWFL, …
- Formalisms: Petri net, UML activity diagram, BPMN, DAG, IPO, GPSG, Workflow Patterns, Pi Calculus, Finite-State Machine, Gamma-calculus, …

□ Workflow jungle (2)

- … and also engines: 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, …
- … and composition tools: ilog's BPMN Modeller, CAT, GWUI, XBaya GUI for Workflow Composition, Triana, JOpera, Platform Process Manager, ...

□ Initial motivation

- Many services are developed/deployed in the frame of the Virtual Observatory (registry, data services, Web Services, computing and Cloud services, …).
- Complex implementation and coordination of the services are possible through workflows
  - Evolution from an execution of one service to a combination of services (exchanging data, …)

Initial motivation (2)

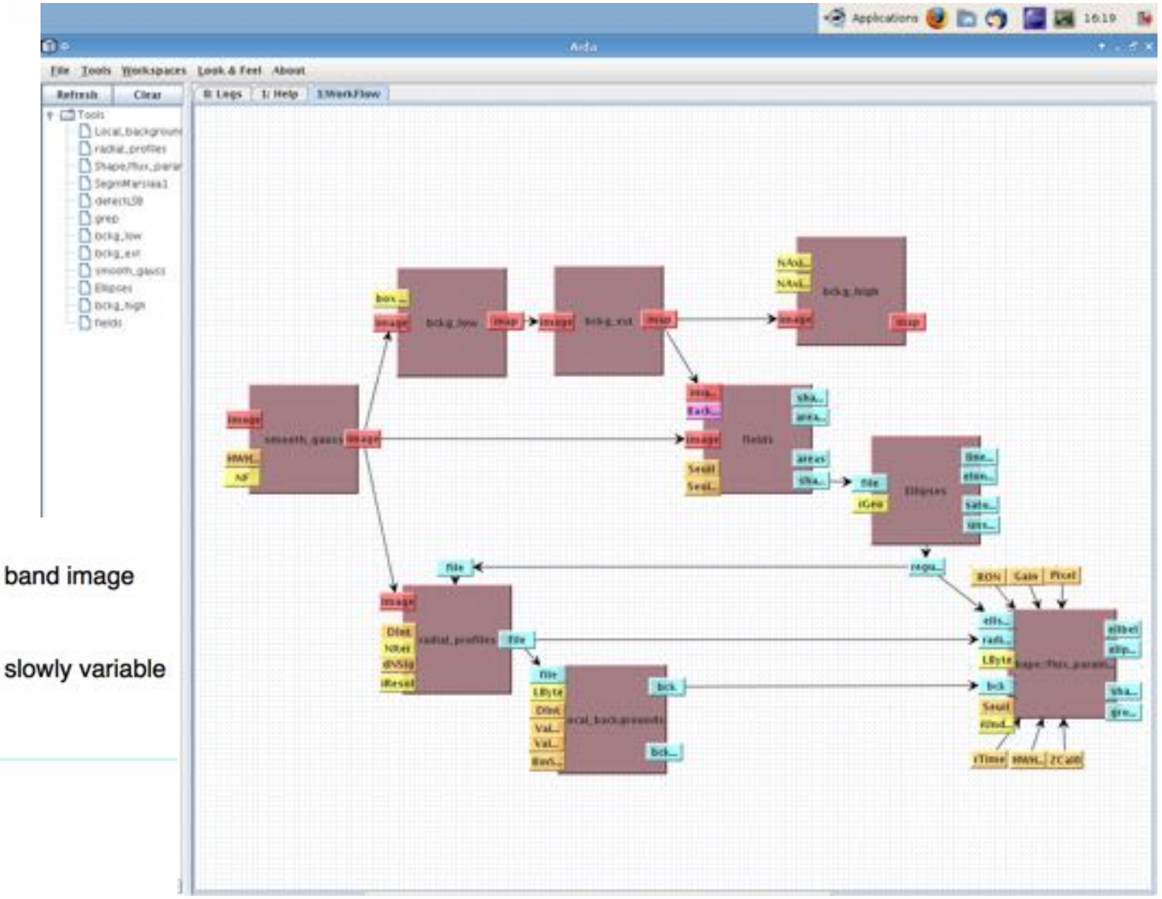
- Workflows are useful to capture scientific methodology and to provide provenance information for their results
- Workflows provide a formalization of the scientific analysis
  - routines to be executed, data flow, execution details, …
- Workflows are structures useful to manage computation at a large-scale
- Collaboratively designed, assembled, validated, analysed

□ Initial motivation (3)

- Definition of workflow use cases
- Easy to reuse in new workflows old applications developed in the past by trainees, Ph D. students, engineers and astronomers in different languages
- Experience sharing with people in different domains (image, spectroscopy, data mining, simulation, …)

Illustration: preservation of the process

Image processing use case, Eric Slezak, Observatoire de la Côte d’Azur



- **operation**
  - detection and evaluation of related objects in 1 band image
- **subjacent model**
  - diffuse disjoined tasks in emission on a bottom slowly variable without defects
- **method**
  - cartography of the background
  - thresholding by segmentation
  - adjustment of an ellipse of form
  - evaluation of the azimuth profile of brightness
  - calculation of measurements of form and flow

IVOA Note about scientific workflows (2013)

- A survey of existing tools, methods, …
- Around 35 participants from the astronomical community

□ In the VO context

- A lot of questions…
  - Existing tools and projects (how to take them into account in workflows) and constraints on the future developments
  - Localisation and checking of the services (execution time, tests, results, …)
  - How to integrate the VO Standards ?
  - How to preserve these workflows ?
  - Etc.

□ « Social » preservation

- myExperiment, a social networking site for workflow exchange and sharing, with 10000 members and 3700 workflows representing several 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.

□ Workflow for Ever

- The EU FP7 funded project “Wf4Ever: Advanced Workflow Preservation Technologies for Enhanced Science” (2010)
- Main intend was 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.

□ Workflow for Ever (2)

- This project investigated and developed technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows in a range of disciplines, including Astronomy.
- AstroTaverna (Taverna plugin / IVOA standards)
- But this kind of project has an End !

□ Workflow for Ever (3)

□ Workflows and VO standards

- Registry WG
  - Adaptive workflows with a choose of tools depending on parameters like the availability (IVOA VOSI), ...
- Grid and Web Services WG
  - VOSpace: storage of intermediate (deleted after each execution or temporary conserved to replay partially the workflow, ...) or final data produced during the workflow execution, ...
  - UWS: use of asynchronous VO services in a workflow, ...

□ Workflows and VO standards (2)

- Data Modelling WG
  - Characterisation, Provenance, cf. M. Louys's talk
- Data Curation and Preservation IG
  - Permanent identifiers
- Application WG
  - Workflow management apps, SAMP
- Theory IG
  - Self-descriptive Web Services (cf. PDL)

IVOA Parameter Description Language (2014)

- PDL is particularly addressed to scientists or engineers
  - wishing to expose their research codes (without limits or compromise on the complexity of the exposed code) online as public services
  - wishing to interconnect their codes into workflows

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.

□ Workflow preservation (2)

- 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.

□ Workflow preservation (3)

- 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.
- In this context, Virtual Machines have been considered as a method for capturing a workflow in an executable, mostly portable form.
- But there could also be vulnerabilities regarding the interpretation of workflows and data, documenting their provenance and limitations, and ensuring that they are trustworthy.

Our recommendations concerning workflow preservation

- As a first approach to preservation of workflows we can consider the basic steps for software preservation:
  - retrieve
  - reconstruct
  - 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.

Our recommendations concerning workflow preservation (2)

- 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.

Conclusion

- 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.

□ Conclusion

- 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.

□ Conclusion (2)

- 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

- **Parameter Description Language**, Zwolf C.M., Harrison P., Garrido J., Ruiz J.E., Le Petit F., **IVOA** Recommendation, 2014
- **Scientific Workflows in the VO**, Schaaff A., Ruiz J.E., **IVOA** Note, 2013
- **BoF about Scientific Workflows in the VO**, Schaaff A., Verdes-Montenegro L., Ruiz J.E., Santander-Vela, ADASS XXI, 2011
- **Workflow systems and VO Standards**, Schaaff A., Bonnarel F., Louys M., Slezak E., Gassmann B., Pestel C., Benjelloun O., Mantelet G., Memorie della Società Astronomica Italiana, 2009
- **Workflow in Astronomy : the VO France Workflow Working Group experience**, Schaaff A., Petit F., Prugniel, P., Slezak E., Surace C., ADASS XVII, 2007
- **Implementing astronomical image analysis pipelines using VO standards**, Louys M., Bonnarel F., Schaaff A., Claudon J-J., Pestel C., Highlights of Astronomy, Volume 14, 2006
- **Image processing and scientific workflows in the Virtual Observatory context**, Slezak E., Schaaff A., Claudon J-J., Highlights of Astronomy, Volume 14, 2006
- **Work Around Distributed Image Processing and Workflow Management**, Schaaff A., Bonnarel, F., Claudon J.-J., Louys M., Pestel C., David R., Genaud S., Wolf C., ADASS XV, 2005
- **VO France Workflows WG**, Schaaff A., Surace C., Slezak E., Le Petit F., Prugniel P., http://www.france-ov.org/twiki/bin/view/GROUPEStravail/Workflow, started in 2005