Working Groups
Workflows

Workflows at the Second EUDAT Conference

Overview

Workflows are a joint research activity in EUDAT in which communities (e.g. ENES and CLARIN) are assessing solutions in which community workflows can make use of the EUDAT services. During the 2nd EUDAT Conference, 28-30 October 2013 – Rome, Italy[1], the New Services track included a specific session on Workflows where the results from the working group workshop (Barcelona, Sept 2013) were presented followed by the ENES and CLARIN use cases on workflows and a proposal for a generic execution framework (GEF).

Workflow working group workshop

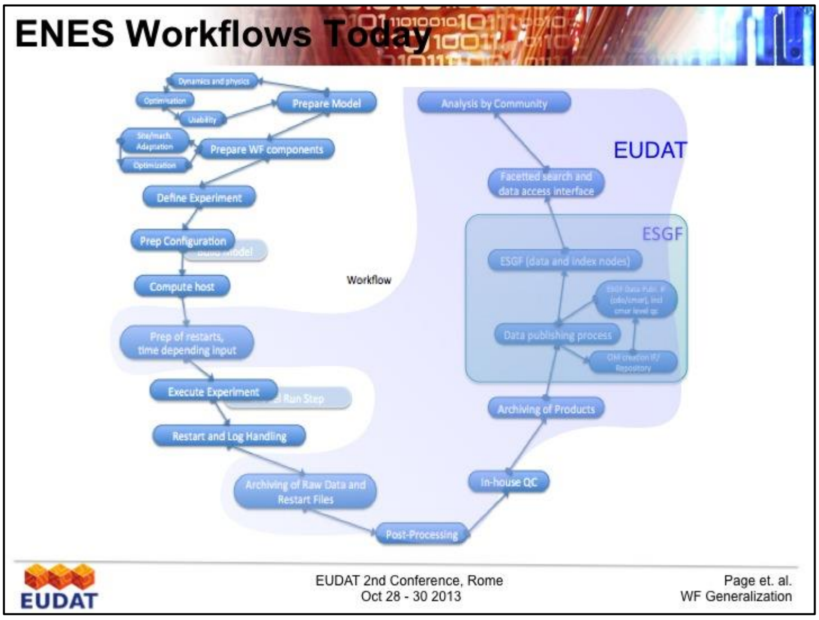
The goal of the working group workshop in Barcelona was to understand the needs of the community experts on common services, how to orchestrate data processing and how scientific workflows can make use of EUDAT services. Support for workflow provenance and services to register and describe workflow components and make them discoverable, referable (e.g. assigning PIDs to components) and to capture best practices were intensively discussed. It is very important to describe the functionality of a workflow component, input and output data formats and test data to certify the functionality of a component. Additionally, it is recommended that EUDAT does not develop a new workflow system but rather clearly define an API to be used within workflows. This is in line with the EUDAT GEF developments. The next steps are: not to lose momentum, to focus on concrete work and formalize and continue the work group.

ENES Workflow use case

The European Network for Earth System (ENES) community represents the European community for climate modelling providing predictions for the IPCC[2] report and for EU mitigation and adaptation of policies. Climate modelling studies are very compute intensive and there is a strong need for climate numerical models tailored for HPC computing. Therefore ENES is collaborating with PRACE to get access to world-class computing resources. Climate is a global event influencing all aspects of the environment. It impacts researchers and modellers from agriculture, water management, shipping, dikes, etc. EUDAT provides a platform and building blocks to enable this kind of inter-disciplinary research.

The ENES workflows must be integrated with the Earth System Grid Federation (ESGF), which is the workhorse of the ENES community. The workflow of tomorrow must provide better and more automated metadata management, provenance data, integrated into the daily work of the scientists and better interoperability between workflows in different communities. These were the main reasons to start working on workflows activity in EUDAT. The figure below gives a global overview of the relationship between the ENES workflows and the ESGF and EUDAT service domain.

[1] http://www.eudat.eu/parallel-track-iv-new-services-overview
[2] Intergovernmental Panel on Climate Change, homepage http://www.ipcc.ch


ENES Workflows Today

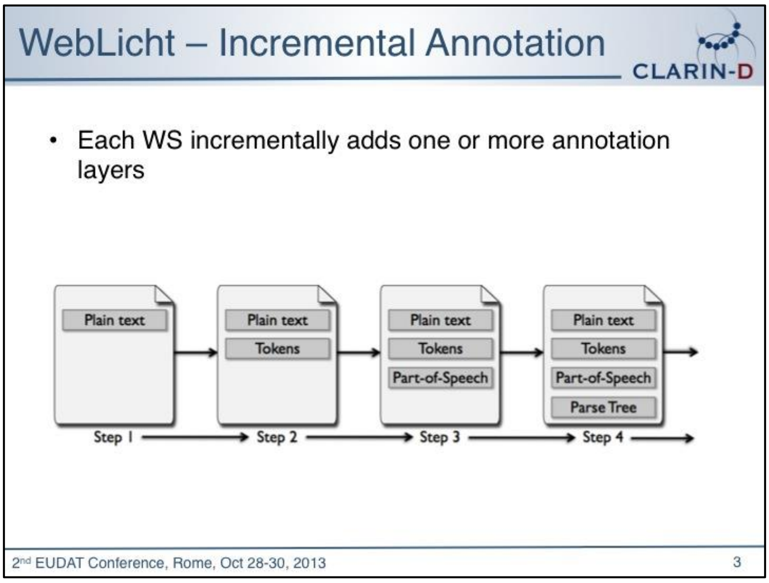
EUDAT 2nd Conference, Rome
Oct 28 - 30 2013

Page et. al.
WF Generalization

Figure 1 - Relationship between ENES workflows and ESGF and EUDAT service domain

CLARIN WebLicht workflow use case

WebLicht (Web-based Linguistic Chaining Tool) is a broadly used workflow engine for linguistic annotations, which is based on Service Oriented Architecture (SOA) principles. Each tool is made available as a web service. The user does not have to install any annotation tool on his/her local machine and is able to visualize the workflow within the web interface. In WebLicht each workflow step incrementally adds one or more annotation layers as shown in the figure below. There are many challenges due to (1) increasing data sizes and (2) an increasing amount of users who want to execute chains. The increasing size and partly also legal issues of data make it hardly possible to move data to the locations where data analysis tools are being executed.. The increasing data volumes and the fact that an increasing amount of users want to execute workflow chains with their data require a change of approach so that data will be stored close to HPC or large cluster servers dependent on the type of algorithms being executed. The question is how WebLicht workflows can take advantage of EUDAT services which take care of storing data and bringing data close to computational facilities? A possible solution is to enrich the EUDAT B2STAGE to interface with workflows, for example with the generic execution framework.


WebLicht – Incremental Annotation

CLARIN-D

- Each WS incrementally adds one or more annotation layers

2nd EUDAT Conference, Rome, Oct 28-30, 2013 3

Figure 2 - Web-Licht incremental annotation

EUDAT Generic Execution Framework

The idea of the Generic Execution Framework (GEF) is to enable processing of datasets close to where the data is stored, allowing faster access, lower network bandwidth usage and easier filtering and sub-setting by only transferring the end results back to the user. The GEF will provide an API layer, which consists of a collection of HTTP web services. The API layer allows easy integration of the GEF into existing workflow engines (e.g. Taverna, Kepler) and community specific data federation interfaces. GEF is built on top of iRODS, which is the current core technology of the EUDAT B2SAFE (Safe Replication) service, but other back-ends are possible. It allows the input and output of data sets to be specified via URIs or handles/PIDs. The GEF API is generic, whereas functions are to be created and maintained by communities, functions can be combined into pipes. A pilot implementation of the GEF framework has been developed and has been tested within the ENES and CLARIN workflows. Tests are on the way to test a full integration within the ENES and CLARIN federations.

Workflow Discussion & Conclusions

The first set of questions were technical about the service implementation and the API functions supported by the GEF framework and how much training is needed to make use of the infrastructure. The service has been implemented in JAVA and provides an HTTP/Rest API interface. The basic functions are: send/get data and send/get workflows, which can only be executed by a community manager. On the subject of training an example was given about the uptake of the WebLicht workflow engine within the CLARIN community. This has been good and unexpected; currently the WebLicht workflow engine is used for teaching purposes at a number of universities to explain linguistics.

During the discussions comments on the results from the Workflow working group were given:

- EUDAT should try to minimise the number of workflows, but adopt a bottom up approach;

- EUDAT should look at cross community aspects and information about workflows should be discoverable. For this EUDAT could provision a workflow repository and registry service in which communities can provide content about workflow execution engines.

Further Information

For more information see the Workflows web Section http://www.eudat.eu/workflows or contact:

- Christian Pagé, CERFACS [christian.page[at]cerfacs.fr] or
- Morris Riedel, FZ Juelich [m.riedel[at]fz-juelich.de]