European Collaborative Data Infrastructure
Grant agreement number: RI-283304

Workflows Barcelona Workshop Report

25 – 26 September 2013

EXECUTIVE SUMMARY

The goal of the *EUDAT Workshop on Workflows* 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 by the 20 international experts in the field of "Scientific workflows" present. They shared their insights and experiences towards the need of common services that EUDAT might be able to provide and concluded that 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. The consensus of workshop participants on potential *'common workflow services elements'* are reflected in the following four recommendations and corresponding actions for EUDAT that require more elaborate exploration.

RECOMMENDATIONS

Action	Short Description	Priority	Next Steps
Provide EUDAT Service APIs for use within Workflows	Initiatives like EUDAT should provide service APIs for workflows rather than creating new WF tools. This would enable researchers to seamlessly take advantage of current/new EUDAT services such as data-staging, data-transfer, data replication, or simple store, PID assignments.	High	Some EUDAT services already offer APIs; Create a document that provides an overview how EUDAT service APIs can be used.
Explore solutions for EUDAT Workflow Provenance Service(s)	There is an increasing variety of WF systems and many communities already have chosen their solutions, but might re-use components of others. EUDAT could offer a service that enables 'workflow component sharing' a repository/registry where components of workflows are stored including provenance information; Such information includes but is not limited to assignments of PIDs for workflow components, including concrete software elements, information about concrete execution runs of it, and sample data that enables other researchers to better understand the shared workflow components.	High	PNNL has performed some work on sharing components and describing workflows independently from concrete implementations; Such work needs to be surveyed and could be a baseline for a potential new EUDAT service.
Provide higher-level Analysis & Analytics Workflow Components & Service APIs	The presentations spanning all fields have shown that statistical computing, data mining, and machine learning algorithms (e.g. classification, clustering, or regression techniques) are used in some parts of the workflows; A potential set of 'higher-level data analysis/analytics services' could be hosted by EUDAT on front-end servers of data centers. This includes the provisioning of service APIs for seamless integration in (existing) analysis workflows and their 'application enabling process'.	Medium	Used statistical computing (e.g. R) or machine learning (e.g. Apache Mahout) software already exists; Provide an overview which of these packages could be conveniently hosted by EUDAT and which service APIs could be provided.
Investigate solutions for data workflow recommender services	Data formats are set by user communities and consequently the lack of standardization impacts by preventing easy data sharing across communities ; EUDAT could investigate the possibility of recommender services that provide advice on suitable workflows in context depending on data formats, scalability, portability, etc.; This might include benchmarks of workflows in context and access to (captured) best practices in the community;	Medium	Some data formats are favoured across communities such as HDF5 or NETCDF; EUDAT should survey the use of common data formats in communities;

The field of 'harmonizing security' across workflows was often mentioned during discussions and as a potential recommendation but it was also acknowledged that this is a problem that needs a broader approach.

Overall Next Steps

Discussion must be continued and encouraged in order to not lose momentum and to focus on some concrete work actions derived from the 'broad recommended actions'. Forming a more structured working group was suggested and it could start with identifying overlaps with existing work from the experts.

WORKSHOP PARTICIPANTS

The following experts participated in the discussions:

Name	Organisation	Role
Christian Pagé	CERFACS	Chair
Morris Riedel	JUELICH	Co-Chair
Bertram Ludaescher	University of California	Expert
Erhard Hinrichs	Tübingen University	Expert
Emanuel Dima	Tübingen University	Expert
Ilkay Altintas	San Diego Supercomputer Center	Expert / Minutes taker
Kerstin Kleese Van Dam	Pacific Northwest National Laboratory (PNNL)	Expert
Alan Williams	Manchester University	Expert
Roman Valls Guimera	International Neuroinformatics Coordinating Facility	Expert
Pavel Stranak	Charles University in Prague	Expert
Alex Hardisty	Cardiff University	Expert
Jozef Mišutka	Charles University in Prague	Expert

