Geophysical Research Abstracts
Vol. 15, EGU2013-5880, 2013
EGU General Assembly 2013

A stepwise approach to integrate climate data analysis workflows into e-science infrastructures

Stephan Kindermann (1), Bernadette Fritzsch (2), Wolfgang Hiller (2), Benny Bräuer (2), Nils Hempelmann (3), and Gregory Föll (1)

(1) German Climate Computing Centre (DKRZ), Hamburg, Germany (kindermann@dkrz.de), (2) Alfred Wegener Institute for Polar and Marine Research (AWI), Bremerhaven, Germany (Bernadette.Fritzsch@awi.de), (3) Climate Service Center (CSC), Hamburg, Germany (nils.hempelmann@hzg.de)

Large scale climate data analysis could not be constrained to a single data archive. Rather, linking many different data archives delivers new insights into the fundamental processes in climate science. Thus infrastructures emerged supporting the development of distributed climate data analysis workflows by providing integrated data and workflow management middleware. Two examples of this development are the ESGF infrastructure and the C3Grid.
The ESGF data federation is based on an international effort to establish a distributed archive serving climate model data from large climate modeling intercomparison projects like CMIP5. C3Grid was developed as a national initiative for a climate data processing and data access infrastructure in Germany, integrating national climate data centers with compute providers.
While no support for processing workflows is currently integrated into the ESGF infrastructure (some development efforts are on the way), the C3Grid offers several processing features with different diagnostic workflows. The experiences in this context show the complexity and fragility of efforts to provide operational workflows relying on a distributed e-science infrastructure composed of middleware services interacting with distributed compute and storage resources.

Based on this experience we present a generic workflow integration process. The basic idea is to modularize the necessary integration steps of workflows into the C3Grid e-science infrastructure. The presented workflows also use data sources integrated in the worldwide ESGF data federation.

In a nutshell the process is structured as follows:
1) local development and test of a data analysis components / workflows with example data
2) Exposition of the workflows as a processing web service based on the WPS OGC standard using locally hosted data
3) Generalizing the workflows by explicit inclusion of data staging tasks (thus extension to externally hosted data) alongside with persistent data identification (e.g. for provenance tracking of generated data products)
4) Development of a workflow parametrization GUI
5) Integration of workflows and GUI into the C3Grid infrastructure

The key advantages of this process are:
- early access and testing of newly developed functionality
- ease of access and reuse of analysis components (with a restricted set data sources) independent of the C3Grid middleware e.g. for testing and research
- additional effort for integration into the C3Grid infrastructure is only done for well tested workflows which have reached production status
- clear interaction points between domain scientists ("workflow developers") and e-science experts ("infrastructure developers")

The approach is illustrated by workflows developed by the climate service center (CSC).