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

Stephan Kindermann [1], Gregory Foell[1]
Bernadette Fritzsch [2], C3Grid Team

[1] Deutsches Klimarechenzentrum (DKRZ), [2] Alfred Wegner Institute (AWI)

Overview

- The Context:
  - Climate data e-science infrastructures and climate data processing
- Climate data processing workflows:
  - The integration problem
  - The C3Grid experience
- A refined approach:
  - stepwise workflow development and service provisioning

The Context

Climate data infrastructures:

- Consistent data search and access needed
  → metadata, security
- Distributed data management needed
  → versioning, replication, archival..
- „Download and process at home“ approach is a dead end
  → processing at data center, distributed processing workflows
- Support for reproducible science
  → persistent data identification, data provenance, data citation

The Context: e-science infrastructures

Existing infrastructures:

- Earth System Grid Federation (ESGF): Worldwide
- IS-ENES data federation: European (ESGF based)
- C3Grid: German (interoperability to ESGF + processing)

Emerging infrastructures:

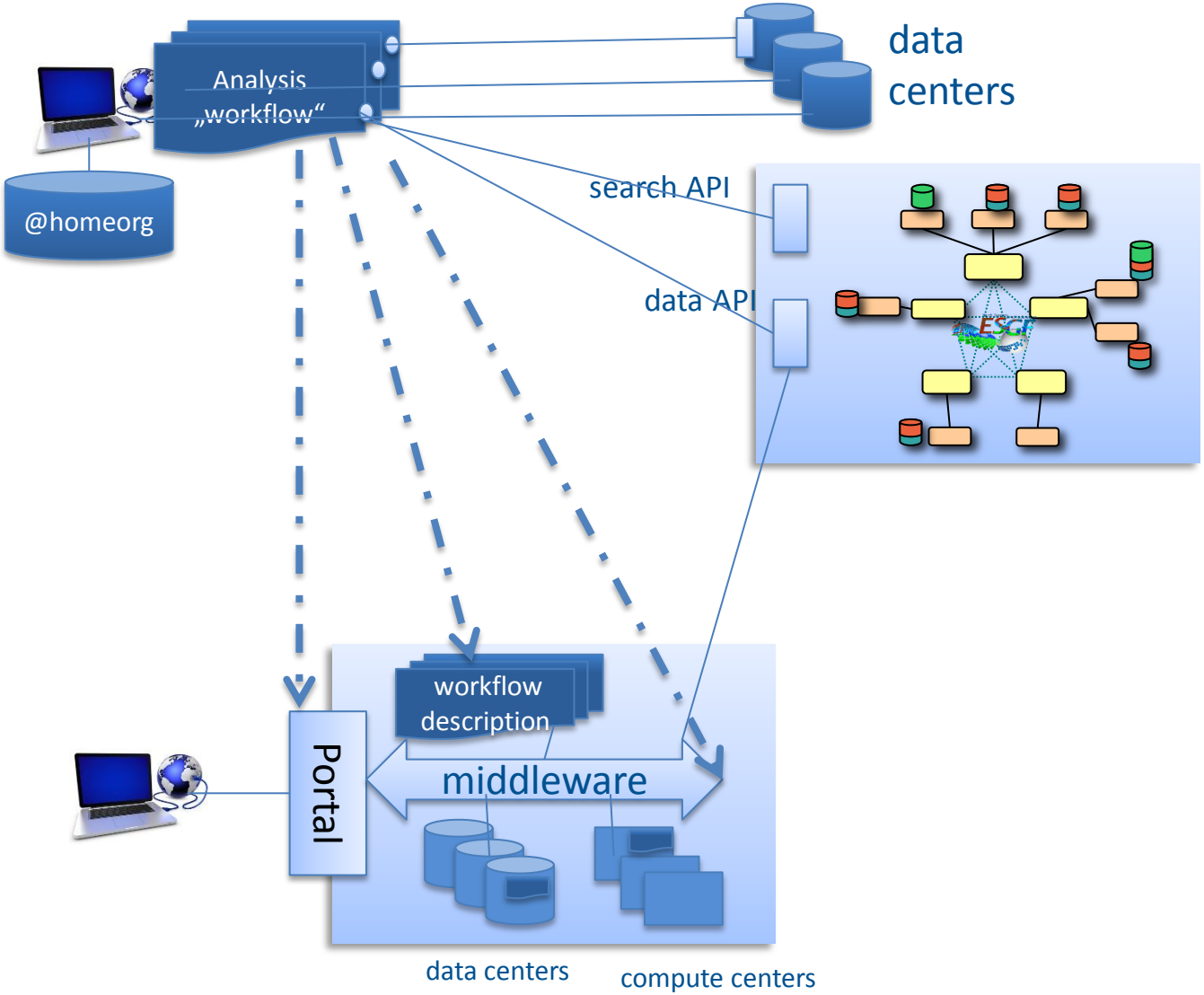
- EUDAT (FP7 project)
- Large Scale Data Management and Analysis (LSDMA - German Helmholtz Association)

The workflow / e-science infrastructure problem

scientific prototype under researchers control

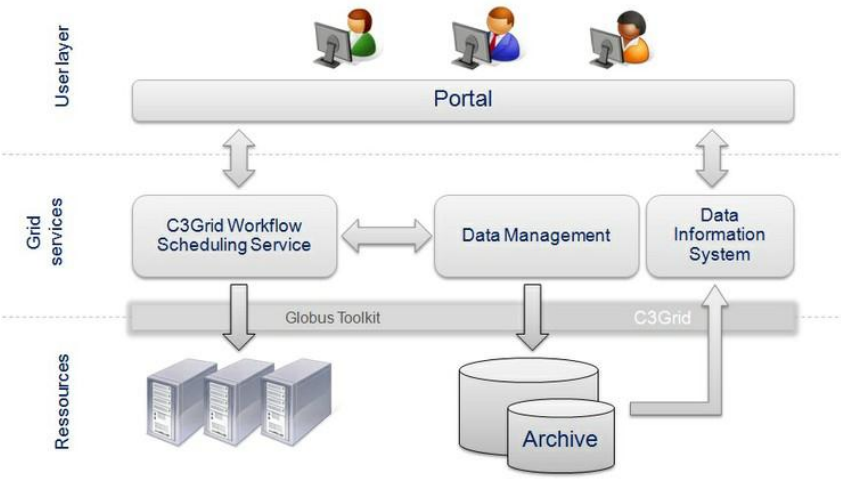
???

stable portal and infrastructure integrated service



C3Grid workflows development

- Decomposition into clearly separated data staging steps with local preprocessing and compute steps
- Deployment at a C3Grid center
- XML wflow language based description



- Upload XML description to C3Grid portal for test
- Interpreted in a co-scheduling middleware (data / compute)
- Debugging ..
- After test a tailored GUI component is integrated in the portal

The C3Grid experience

- Workflow developer training necessary

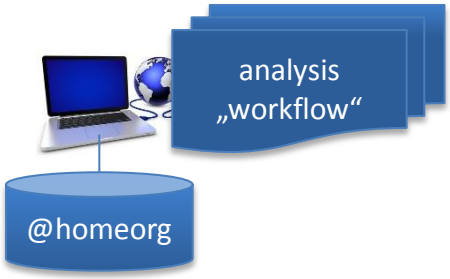
  (workflow description, concept of workflows in a distributed context)
- Time consuming communication between C3Grid developers and workflow developers

  (data constraints, GUI / Portal component, deployment in computing center, debugging)
- Difficult to support „rapid prototyping“
- Different types of „end-users“ requirements:
  - Scientists → „no black boxes“, „we want to know what, where, when is done ..“, „I want to quickly enhance my workflow logic“
  - Non Scientists → „easy to use, transparent front end“, „what does this error message mean ?“

A stepwise integration approach: Overview

Step 0:

- Climate scientist / project develops private prototype

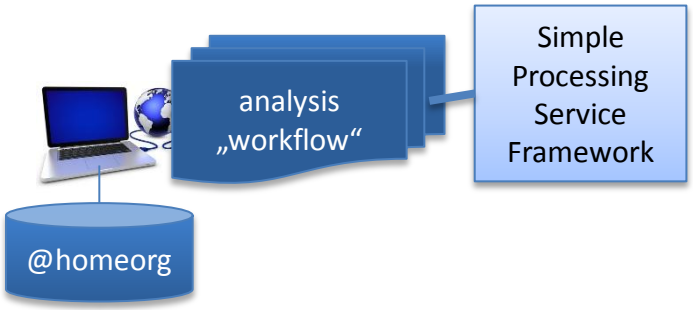


Step 1:

- Climate scientist exposes stable prototype as a web service

→ Required:

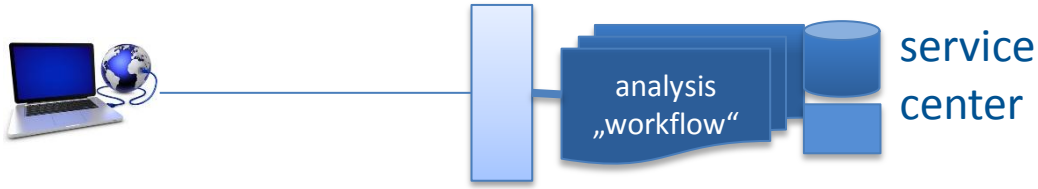
- Easy to install WS-Framework
- Simple workflow integration
- Support of interface standards



A stepwise integration approach: Overview

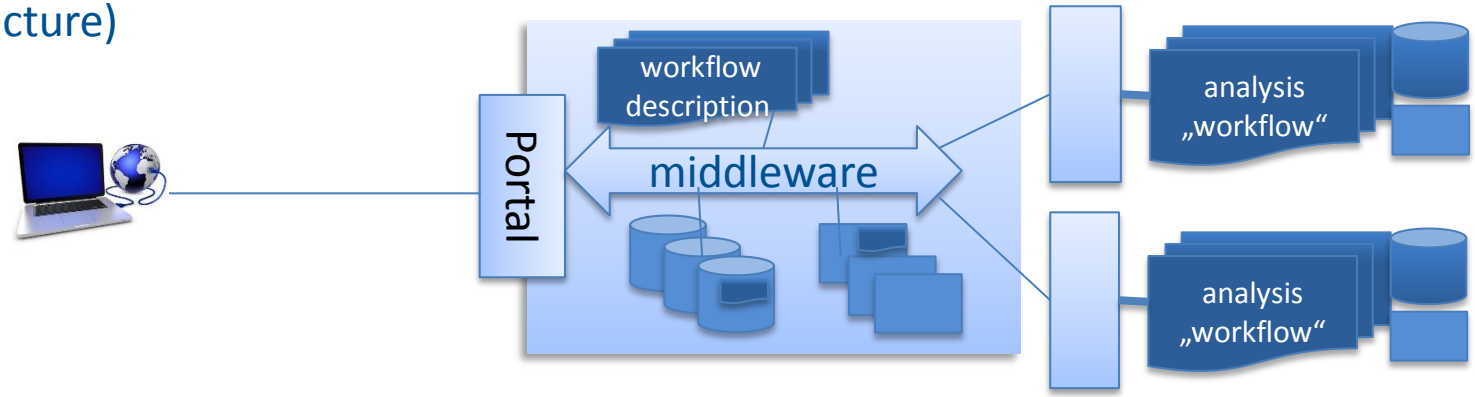
Step 3:

- Climate scientist / sys-admin deploys stable prototype at a service center



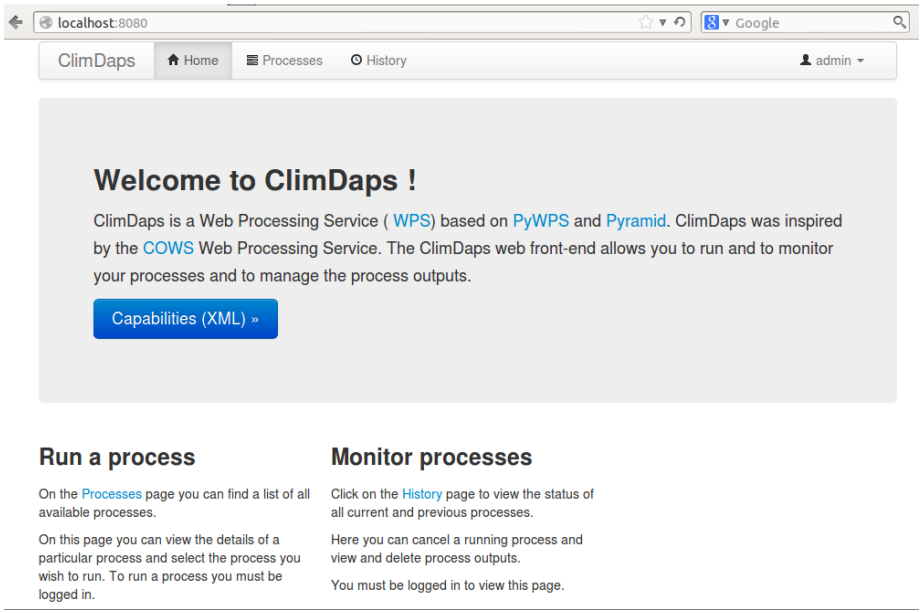
Step 4:

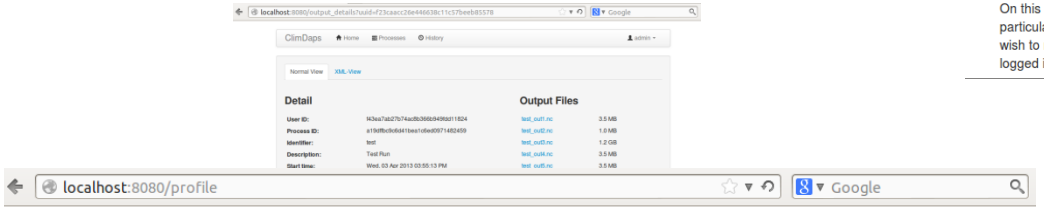
- Workflow integration in portal (and associated e-science infrastructure)

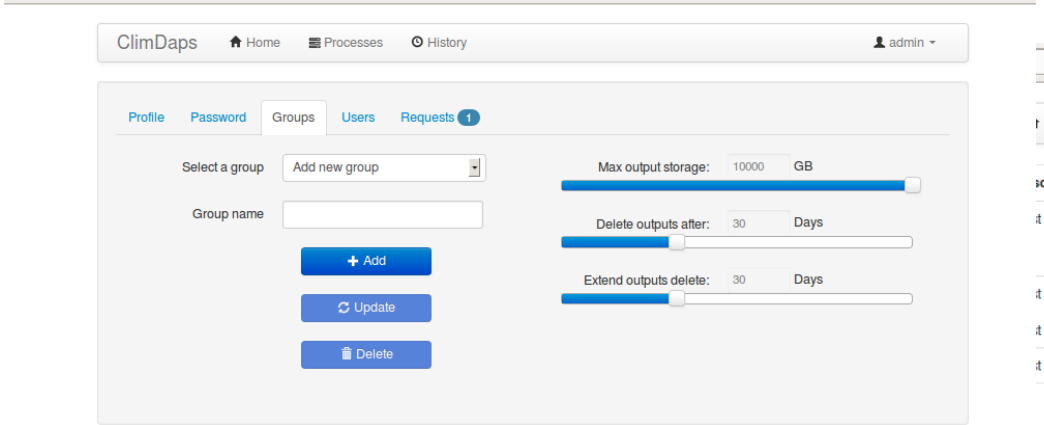


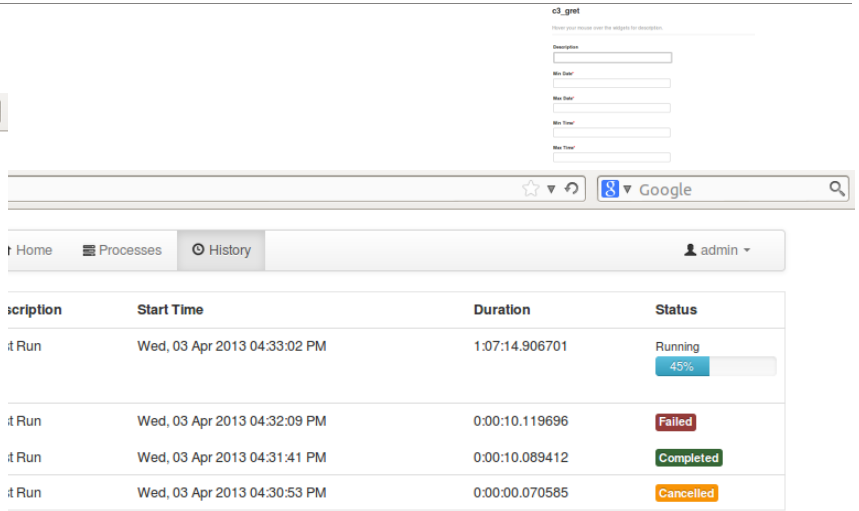
„Workflow as a service“

- The Climate Data Processing Service (ClimDaPS)
  - light weight OGC WPS based framework (based on pyWPS)
  - fully automatic installation process (on any linux box)
  - developed within ExArch G8 project at DKRZ









„Workflow as a service“

- ClimDaPS status:
  - stable first release of framework
    - https://redmine.dkrz.de/collaboration/projects/climdaps/wiki
  - first Climate Service Center workflow integrations:
    - e.g. grass reference evapotranspiration (input CORDEX data)

- ClimDaPS next steps:
  - C3Grid workflow provisioning at DKRZ
  - Integration with ISO metadata generation framework at DKRZ
  - Integration with EPIC PID service at DKRZ (see poster EGU2013-8371 ESSI 2.4)
    - data+metadata+code PIDs → data provenance !
  - European IS-ENES infrastructure integration
    - ESGF CMIP5 and CORDEX data processing
  - Project workflow prototypes:
    - Miklip workflow provisioning
    - LSDMA tests
    - EUDAT tests

„Workflow as a service“

- Related developments in Europe:
  - COWS WPS (BADC, UK): based on own OGC WPS implementation, targets more resource centers (with e.g. job scheduling etc.)

    http://ceda-wps2.badc.rl.ac.uk

  - KNMI impact portal (IS-ENES): exposes OGC WPS based functionality (pyWPS based implementation)

    http://climate4impact.eu

Summary

- C3Grid experience showed the need for an „agile“ approach to workflow service provisioning
- Web processing framework for rapid prototyping and „added value services integration (e.g. pid, metadata generation)
- Parallel activities to develop, test and deploy OGC-WPS services and integrate into e-science infrastructures (C3Grid, IS-ENES, ..)