

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

Stephan Kindermann <sup>1)</sup> - Gregory Foell<sup>1)</sup> Bernadette Fritzsch <sup>2)</sup> -C3Grid Team

Deutsches Klimarechenzentrum (DKRZ), <sup>2)</sup> Alfred Wegner

Thetitute (AUT

Stephan Kindermann



### 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
  - $\rightarrow$  versioning, replication, archival..
- "Download and process at home" approach is a dead end

   processing at data center, distributed processing workflows
- Support for reproducible science
  - $\rightarrow$  persistent data identification, data provenance, data citation



### The Context: e-science infrastructures

### Existing infrastructures:

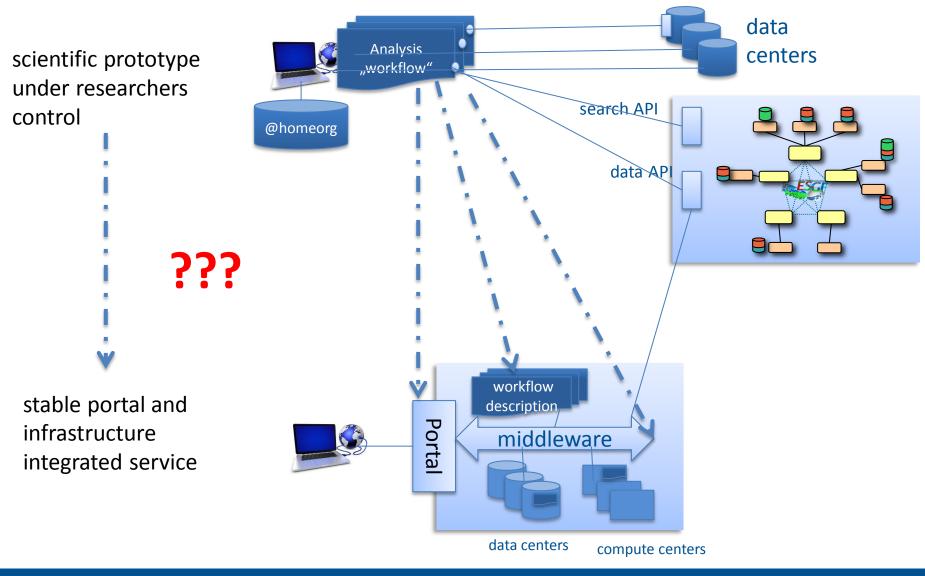
- Earth System Grid Federation (ESGF): Worldwide
- IS-ENES data federation: European (ESGF based)
- C3Grid: German (interoperability to ESGF + processing)

# Emerging infrastructures:

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



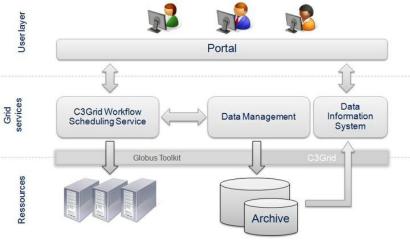
## The workflow / e-science infrastructure problem





# 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

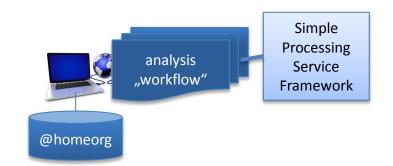


### Step 1:

 Climate scientist exposes stable prototype as a web service

### $\rightarrow$ Required:

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





# A stepwise integration approach: Overview

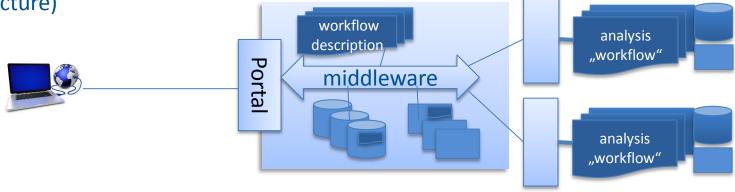
### Step 3:

 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)

Iocalhost:8080/output\_details?uuid=f23caacc26e446638c11c57beeb85578

ClimDaps + Home Processes © History

Processes

Add new group

+ Add

Wed, 03 Apr 2013 03:55:13 Pt

Users Requests 1

O History

-

Normal View

Detail

Description

Groups

A Home

Password

Select a group

Group name

🕘 localhost:8080/profile

ClimDaps

Profile

 fully automatic installation process (on any linux box)

🛛 🛪 🔿 🚺 🖲 Google

1.0 MB

3.5 MB

3.5 MB

Max output storage: 10000

30

Delete outputs after:

Extend outputs delete:

マ わ 🛛 😽 🔻 Google

GB

Days

Days

1 admin -

**Output Files** 

 developed within ExArch G8 project at DKRZ

÷	localhost:8080				ි 🔻 නි 🔀 🖲 Google		Q,
	ClimDaps	A Home	Processes	O History		よ admin 👻	

### Welcome to ClimDaps !

ClimDaps is a Web Processing Service (WPS) based on PyWPS and Pyramid. ClimDaps was inspired by the COWS Web Processing Service. The ClimDaps web front-end allows you to run and to monitor your processes and to manage the process outputs.

Capabilities (XML) »

#### Run a process

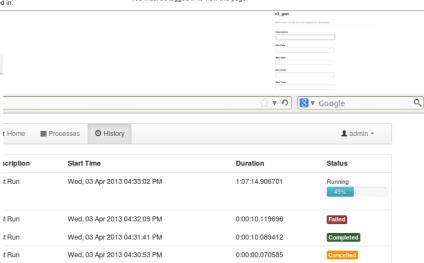
logged in.

Q

#### Monitor processes

On the Processes page you can find a list of all available processes. On this page you can view the details of a Click on the History page to view the status of all current and previous processes. Here you can cancel a running process and view and delete process outputs.

particular process and select the process you view and delete process outputs. wish to run. To run a process you must be you must be logged in to view this page.



### Stephan Kindermann

10



### "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) <u>http://climate4impact.eu</u>



### 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, ..)

