A Collaborative Data Management Infrastructure for Climate Data Analysis

S. Kindermann[1], F. Schintke[2], B. Fritzsch[3]
& C3 Team

1) Deutsches Klimarechenzentrum Hamburg DKRZ, 2) Zuse Institute Berlin ZIB, 3) Alfred Wegener Institute for Polar and Marine Research Bremerhaven AWI

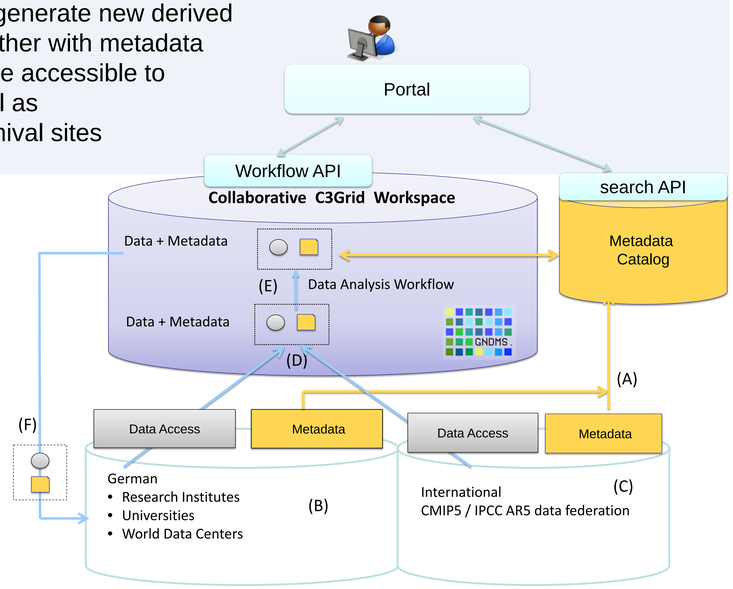
Motivation

- Fast growing databases of climate data require a data management infrastructure that supports climate researchers in their complex data analysis tasks.
- Common sequential tasks for climate researchers up to now:
  - (1) search for appropriate archives and data sets
  - (2) retrieve data from archives using their individual access and query methods
  - (3) process / analyze the data on own resources
- Automation and consolidation is needed for globally distributed data archives and data access on a petabyte scale.
- The C3Grid (Collaborative Climate Community Data and Processing Grid) infrastructure provides automation and offers a collaborative workspace, which provides
  - (1) uniform search interface across all connected data archives
  - (2) uniform data access and
  - (3) distributed storage and analysis management of accessed data

Data Integration

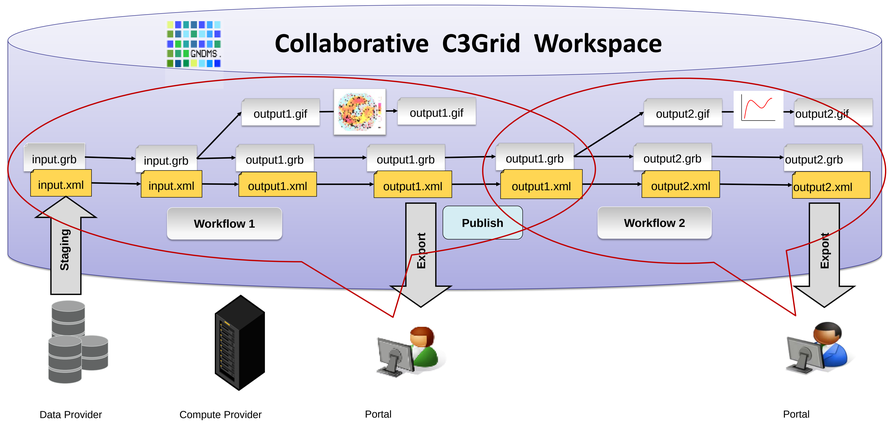
Data archives are connected to C3Grid via standard interfaces (REST, HTTP, GridFTP):

- (A) metadata are harvested to a central catalog for global, uniform search (based on ISO19115)
- (B) national archives implement C3Grid data and metadata interfaces
- (C) C3Grid makes the international ESGF/CMIP5 data federation accessible
- (D) accessed data is staged into the distributed C3Grid workspace – including metadata
- (E) C3Grid workflows generate new derived data products together with metadata
- (F) results can be made accessible to third parties as well as long term data archival sites



Collaborative C3Grid-Workspace (GNDMS)

- Temporary, distributed Grid storage space
  - supports C3Grid security and delegation of certificates
  - data transfer (import and export) via GridFTP und HTTP
  - self-cleaning (data automatically deleted when given life-time expires)
  - separates data sets logically (e.g. slices and slice-IDs)
  - workflows can refer to data using slice-IDs for further analysis of results
  - coupled with the C3Grid workflow scheduler (estimation of data staging duration etc.)



C3Grid Workflows

- always generate valid ISO-metadata describing the results – including provenance information
- hence, results are reproducible and can be re-used as valid input data in C3Grid
- workflows modularize and distribute the whole data analysis process

Next Steps

sharing of results:

- with a user group (several Grid users) or public
- use of handle.net based persistent identifiers

GNDMS will support publishing:

- results are stored on an export site
- data and metadata transfer to a publishing site (another GNDMS instance)
- publishing site is scanned by metadata catalog

→ results are visible and accessible in C3Grid

Further details:

- www.c3grid.de
- esgf.org
- gndms.zib.de

SPONSORED BY THE
Federal Ministry of Education and Research