PIT adoption for Climate Data Management

The German Climate Computing Center (DKRZ)

5th RDA Plenary

Stephan Kindermann

Overview

- DKRZ: A climate science service provider
- Data services for the international climate science community
- RDA PIT adoption plans
  - Generic underpinning for future data services:

    Replication, Load balancing, collection management

DKRZ

A service provider for the german climate (modeling) community

- Non profit company (GmbH) established 1987
- Located in Hamburg, Germany

Balanced HPC / storage system

- 3 PFlop Bull system
- 45 PByte Lustre parallel file system
- 335 PByte HPSS tape backend

Data Services:

- Long term data archival
- World Data Center for Climate
- Core node in international climate data federation (ESGF, IS-ENES)

Climate data services

Cross Community Context

National climate modeling community

International climate community

(modeling + impact + ..)

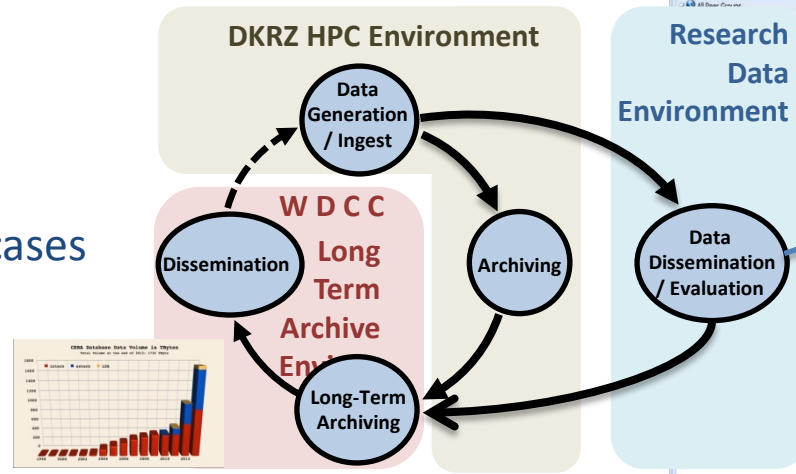
World Data Center Climate

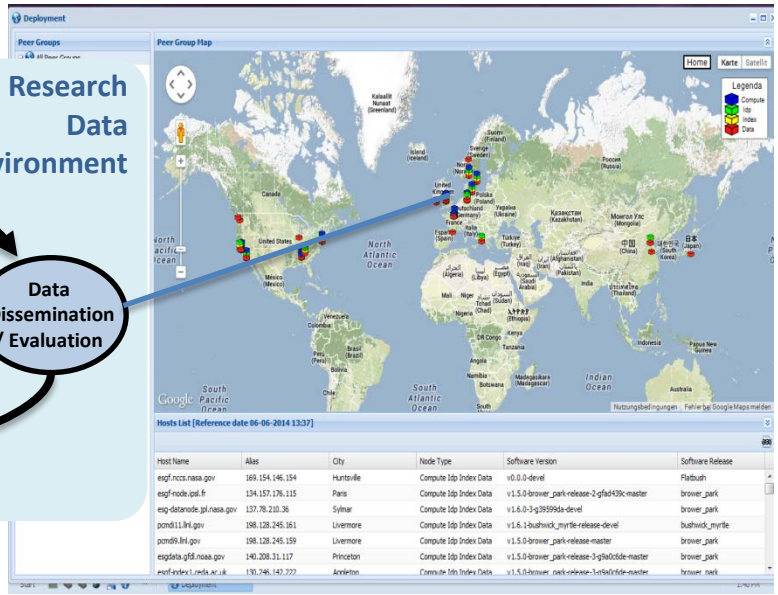
DKRZ

ESGF data federation

Interdisciplinary use cases

ICSU WORLD DATA SYSTEM





Climate data services

Cross Community Context

National climate modeling community

International climate community
(modeling + impact + ..)

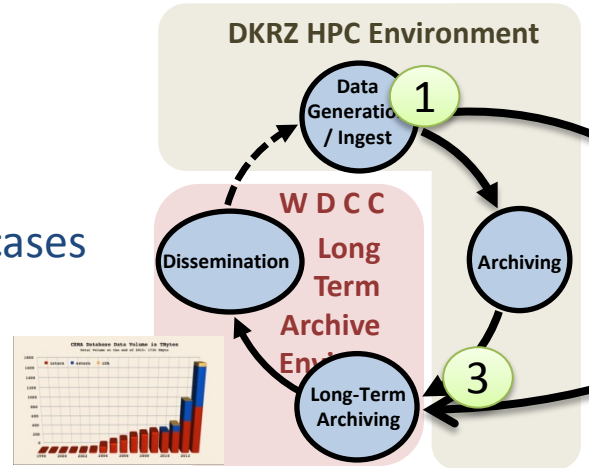
World Data Center Climate

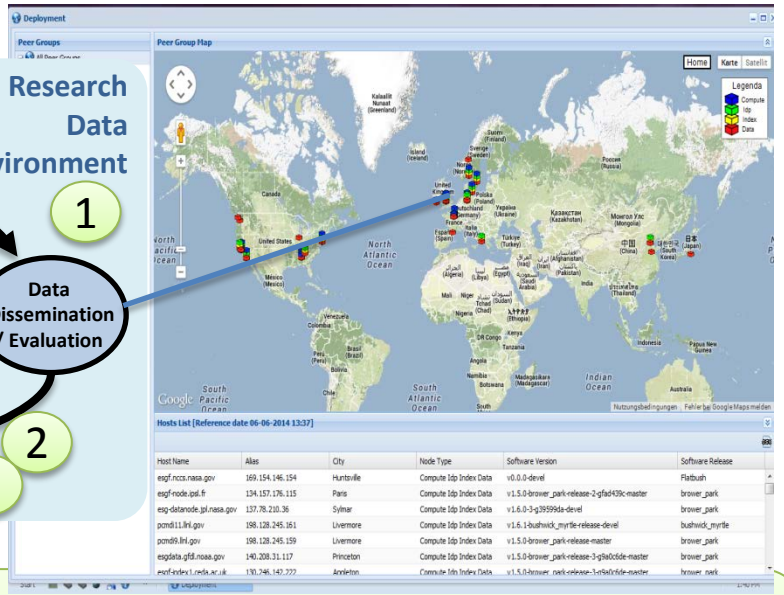
DKRZ

ESGF data federation

Interdisciplinary use cases

ICSU WORLD DATA SYSTEM





1 PID assignment / management

2 Replication and PIT

3 Balanced data access and PIT

PID Information Types (PIT) adoption

DKRZ RDA adoption plans

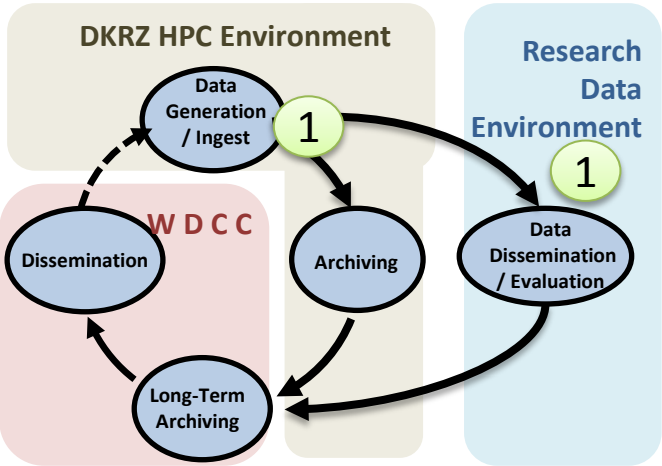
Early PID assignment in data life cycle: 1

- PID registration as part of data center ingest process

First steps:

- DKRZ is partner in European Persistent Identifier Consortium (EPIC)
- Operates PID (Handle) server
- PID registration for (existing) LTA data products

- ESGF PID integration → see large scale data projects meet RDA session 1



PIT adoption: core services

Collection / hierarchy discovery:

- climate data sets are organized in collections (e.g. time series, data + metadata) and hierarchies (e.g. according to experiment organisation)
- Collections are built through PID Information types (agreement on specific PID metadata elements needed for collection/hierarchy management)

→ Generic collections / hierarchy discovery service to be used in many use cases (e.g. replication, balancing,..)

A) Check service applicability: PIT profile conformance test
B) Get components of collection / hierarchy, get values of properties with specified types

PIT adoption: replication use case

Data collection replication:

- DKRZ acts as a core replication center for data e.g. from large international climate intercomparison projects (CMIPs)
- We are on the way to define generic PID types for collections (coll_t) as well as replicas (rep_t), will feed experiences back to RDA

→ Based on this the next step is to develop a generic replication service

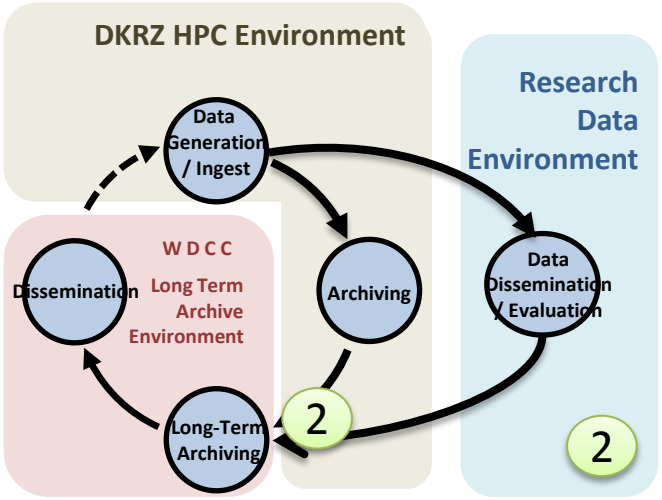
Replication service:

A) check_service_preconditions (PIT profile conformance check)

B.1) Get parts of collection: PID1 → (PIDa,..PIDx)
(See collection management service)

B.2) Get access urls and chcksums
(PIDa,..,PIDn) → (URLa,..,URLn), (csum1,..,csumn)

C) Replicate/check and create new (replica) collection



PIT adoption: load balancing use case

High volume data access:

- High bandwidth requirements for data access
- Exploit replicas for data download

→ Generic data access service exploiting replicas

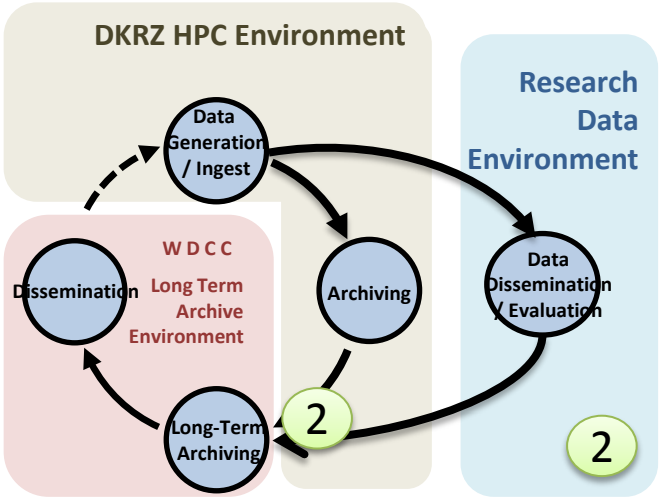
Data access service:

A) user/tool provides either a collection PID or set of PIDs

B) Check applicability: PIT profile conformance check

C) Walk over PIDs (in collection) and get replica locations + chksums

C) Select “nearest” and generate download script



DKRZ RDA adoption plans

- End User services based on PITs and the Type Registry
  - rich landing pages for collections aggregating PIT metadata info with external sources (QA records, annotations,..)
  - early data referencing service for data collections

- PIDs to support international climate model intercomparison project (CMIP) data management
  - collaboration with ESGF – see „Big data projects meet RDA session“
  - DKRZ acts as long term archival and DOI assignment center in CMIP context, transition PIDs – long term DOIs

- Interested in „data fabric“ concepts etc.
  → see ENES data fabric use case

Thank You

Stephan Kindermann: kindermann@dkrz.de
Tobias Weigel: weigel@dkrz.de