

PID usage at DKRZ, the role of RDA and ePIC policies

DataCite-ePIC event,
September 21, 2015, Paris

Tobias Weigel, Stephan Kindermann, Michael Lautenschlager
Deutsches Klimarechenzentrum (DKRZ)

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

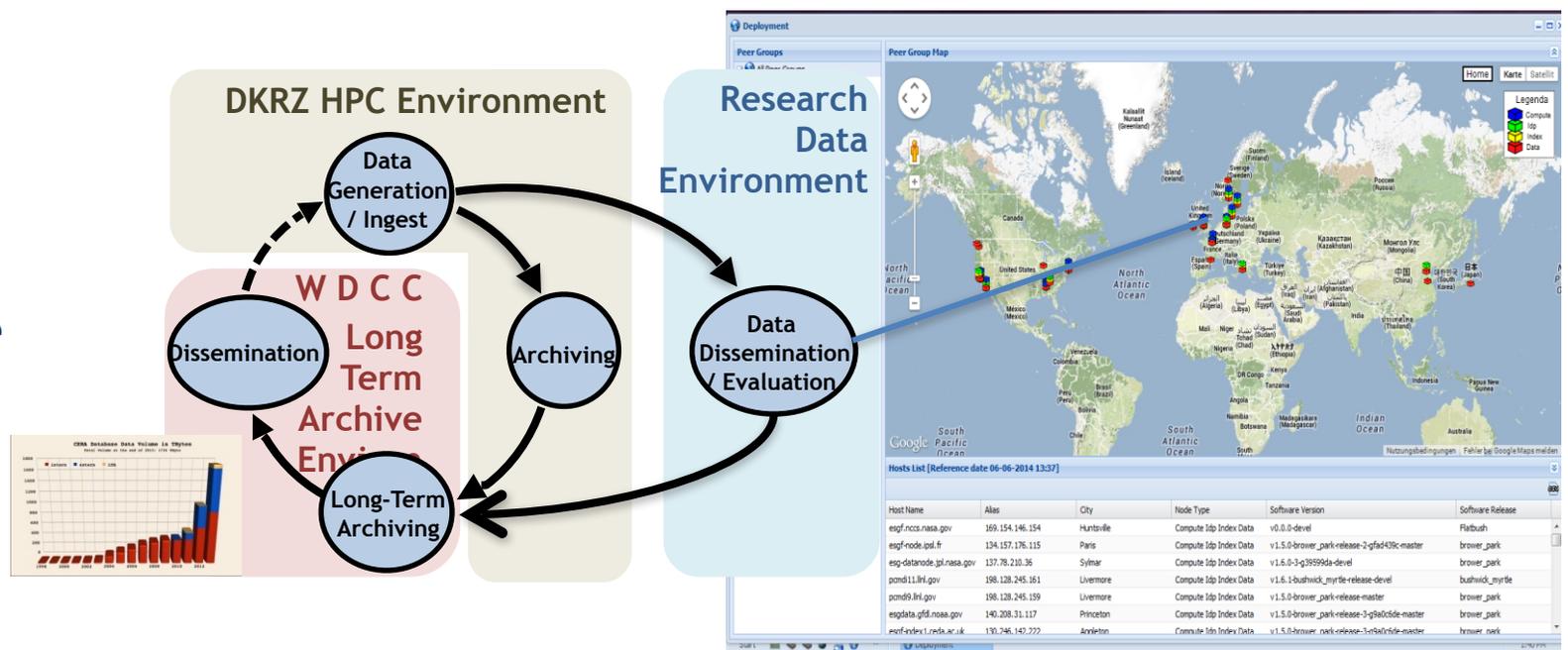
International climate community (modeling + impact + ..)

World Data Center Climate

DKRZ

ESGF data federation

Interdisciplinary use cases



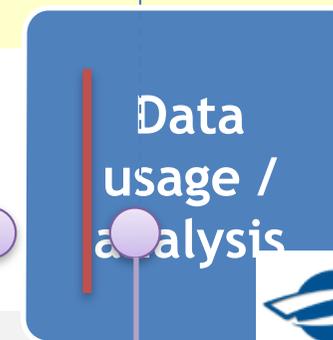
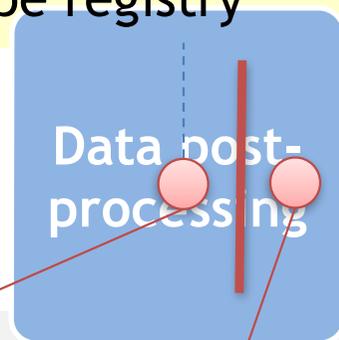
Actionable PIDs for CMIP6



- **PID infrastructure: DONA-ePIC - prefix 21.14100**

- PID information types and type registry

- PID Collections



- Definition of PID names

- Infrastructure and end user services

- Management of PIDs: Integration into ESGF data publication (and versioning/replication) process



- **ESGF PID backend infrastructure: Handle system, message queue, operational agreements ...**

- PID ↔ DOI transition

No Delete policy driven by this scenario

- PIDs are kept beyond object lifetime:
No control over their use once they are visible
 - How can a *temporarily persistent ID* be trustworthy?
- Added value from keeping records:
 - Transparency: Enable users to verify infrastructure actions such as retraction and versioning
 - Improve scientific practice: Preclude use of deprecated data, warn and point to new versions



Parallel use of ePIC and DataCite

- Low-level ePIC Handles for
 - mass data with unclear preservation status
 - more automated data management
 - constructing flexible hierarchies via PID collections
- More rigid DataCite DOIs for
 - fully citable data
 - strong preservation policies
 - towards the end of the data life cycle
- We should clearly communicate these differences
 - Establish and advertise quality levels for PIDs



We need an operational transition process.



- Transition from ePIC Handles to DataCite DOIs
 - Harmonization of policies: well-defined policy sets
 - Differences in policies visible for individual PIDs
 - Clear workflow and additional support services
 - Metadata transition and requirements
- Avoid maintenance of separate services
 - Handles and DOIs in the same infrastructure

It may also work the other way.



- There is also a potential connection back from DataCite DOIs to ePIC Handles
 - Diving into collections, given a DOI
 - Again: preferably use the same tools, unified APIs

RDA efforts and ePIC (and DataCite?)

- PIT: What to put in PID records and how to structure them
 - Conceptual backing, common API, also across PID providers
- TR: Can we register ePIC data types?
- Collections (BoF): Who uses them? What models are there?
- We need to come in ePIC to a better view how to use these things
 - Then we can develop shared services

Thank you for your attention.