Data management services at DKRZ

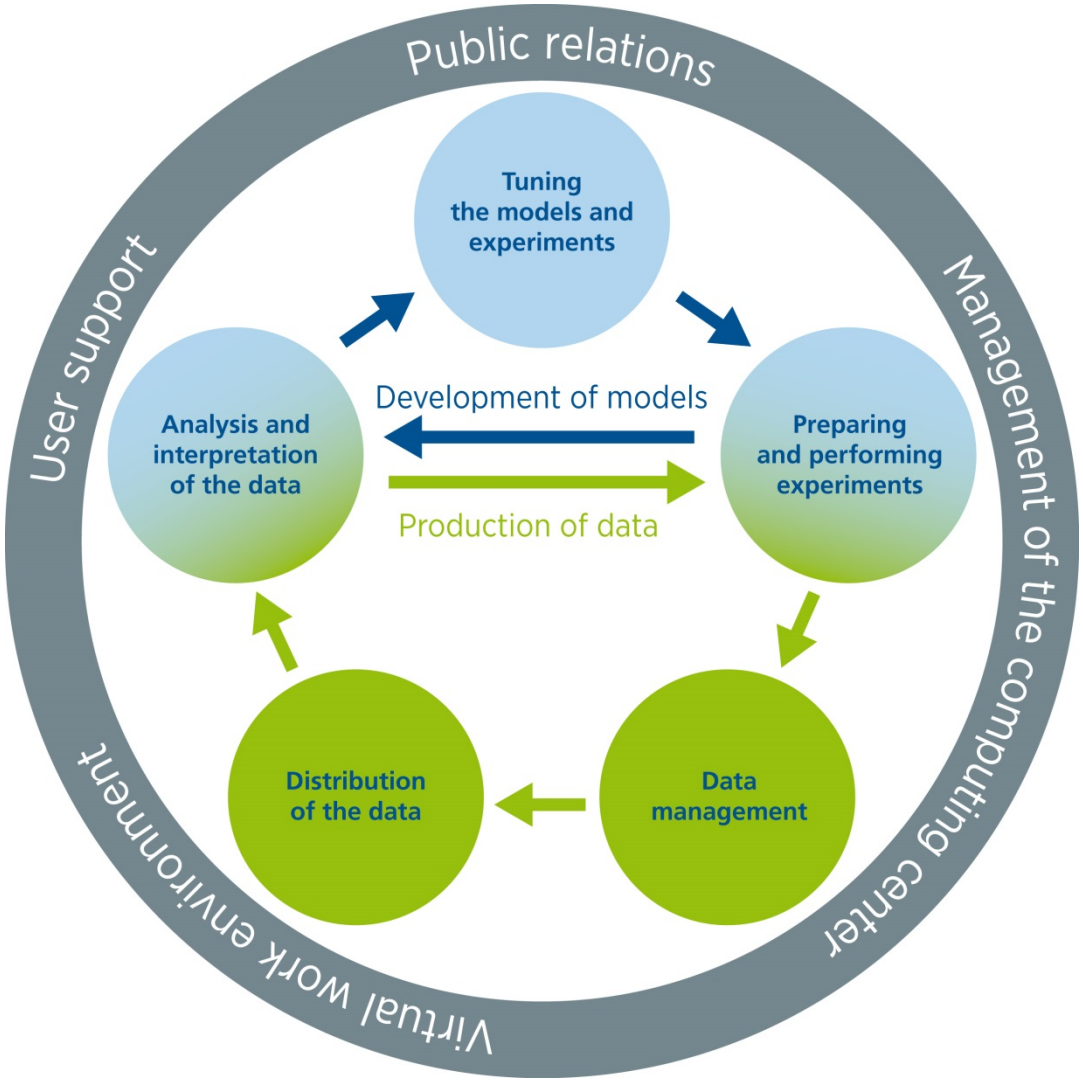
Mistral Training

06-October-2015

Hannes Thiemann
Deutsches Klimarechenzentrum (DKRZ)

Data services at DKRZ

- Disk and tape resources
- Data sharing: DKRZ-cloud and ESGF
- Archiving: DOKU, DKRZ-LTA/WDCC
- Publication / DOI
- Data management plans


Public relations
Management of the computing center
Virtual work environment
User support
Tuning the models and experiments
Preparing and performing experiments
Data management
Distribution of the data
Analysis and interpretation of the data
Development of models
Production of data


Analysis and interpre-
tation of the data
DKRZ services:
• Selection and processing of the data
• Visualization
Preparing and perfor-
ming experiments
DKRZ services:
• Operation of the computers and storage systems
• Performance monitoring
Data management
DKRZ services:
• Standardization
• Documentation
• Quality control
• Archiving
Distribution of the data
DKRZ services:
• Data made available via WDCC/CERA and ESGF
• DOI assignment for referencing

Data resources (Disk)

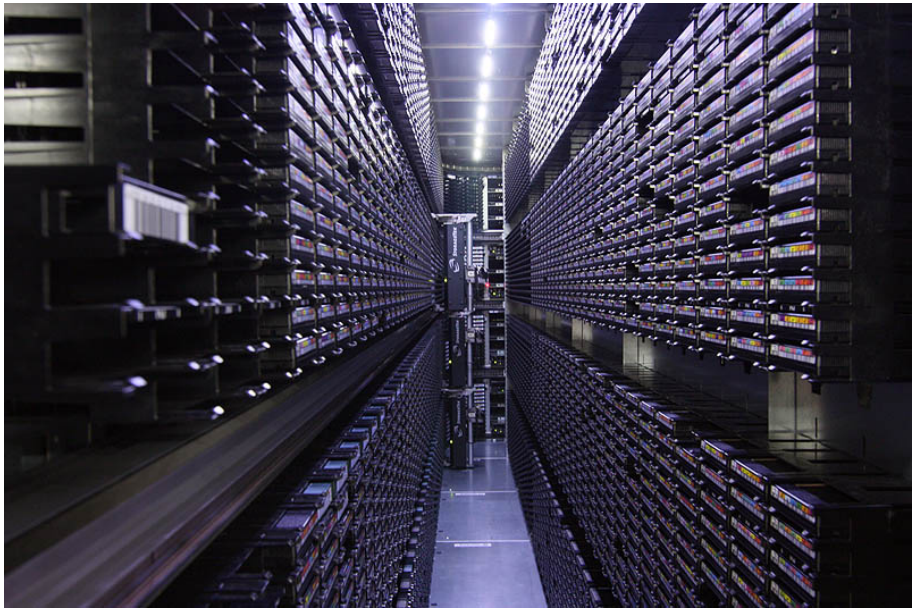
Diskspace (Lustre)

- HOME
  - /pf
  - < 24 GB (per User)
  - for source codes, …
  - Backed up regularly
- SCRATCH
  - /scratch
  - <15 TB (per User)
  - no backup
  - old data automatically removed, granted period is 14 days
- WORK
  - /work
  - project based, quota applies
  - no backup
  - no automatic data deletion, but to be removed 1 month after project expires

Data resources (Tape / HSM system)

ARCH	/hpss/arch	Single copy on tape	Size 20-200 GB recommended	Data removed one year after project has expired
DOUBLE	/hpss/double	2 copies on tapes in different fire compartments		

One common project based quota for ARCH and DOUBLE



DKRZ data services for data sharing

- DKRZ-Cloud
- Earth System Grid Federation (ESGF)

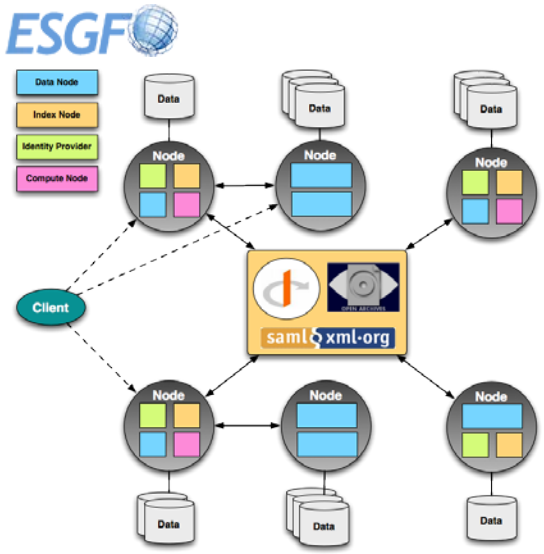
DKRZ Cloud Storage System

The cloud storage system can be used to share data in various ways:

- Access data from different devices, e.g. notebook, smartphone and supercomputer
- Share data with colleagues with or without account at DKRZ
- Replicate data between different sites automatically

International data distribution via ESGF

The Earth System Grid Federation (ESGF) supports the distribution of climate data in the context of international model intercomparison experiments
(MIPs like CMIP5/6 or CORDEX)

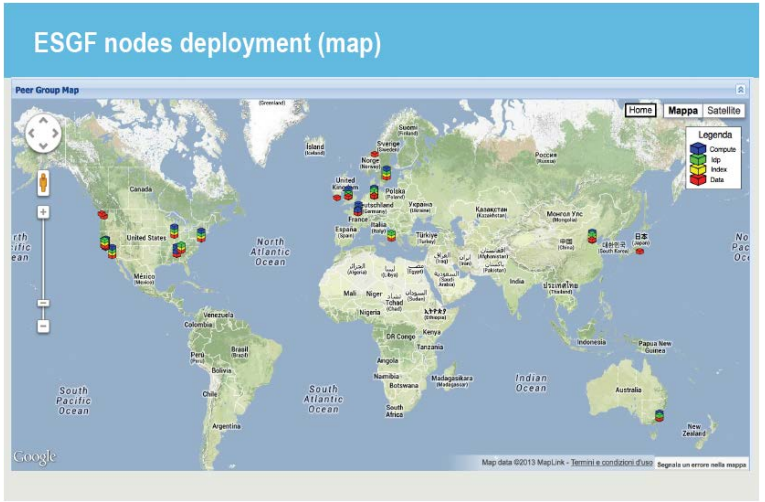


The DKRZ offers:

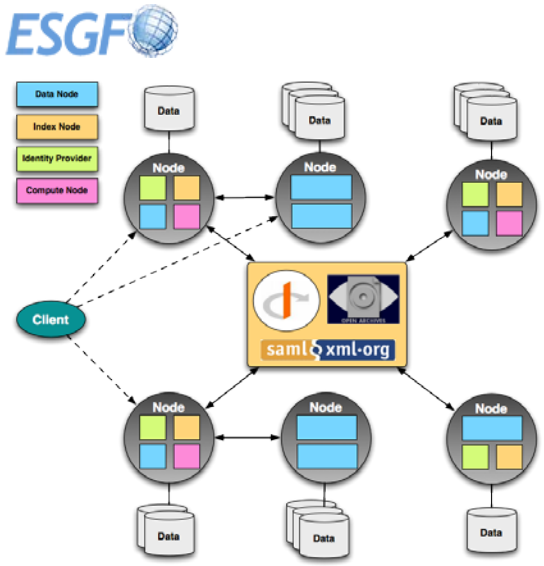
- Support for data-preparation and data quality assurance (CF standard, MIP conventions)
- Data distribution via the (DKRZ) ESGF data node(s)
- Data publication via the (DKRZ) ESGF portal
- Support for long term archival of ESGF data, data distribution via the WDCC as well as DOI assignment

Advantages:

- International visibility of data products
- Uniform interfaces for data search and access and subsetting (opendap)
- Accessibility to community portals (like e.g. climate4impact for the impact community))



Data distribution via ESGF



Preconditions:

- The data are conformant to the standards and conventions of the respective model intercomparison context (e.g. cf-conventions, CMIP6 conventions , Obs4Mips conventions ..)
- The data products are subject to common data access policies (e.g. open accessible or open for non commercial use)

Step by step workflow:

- Submit data publication request to data@dkrz.de
- Fill the respective data publication form with the details of the data to be published (projekt, where to get the data, which variables ..)
- Quality assurance with feedback by DKRZ
- Data publication in the ESGF infrastrukture (on DKRZ ESGF data node and portal)
- Same workflow for new versions of data. Additionally provide information about the reason(s) of the data changes

Data preservation

- DKRZ Doku
- WDCC

https://www.dkrz.de/daten/data-services/langzeitarchivierung

Archiving / DOKU

Service for WLA or shareholder projects to support traceability of the scientific project after project expiration.

The DKRZ offers

- Persistent data and metadata storage 10 years beyond project expiration
- Public access to data using either DKRZ or CERA account

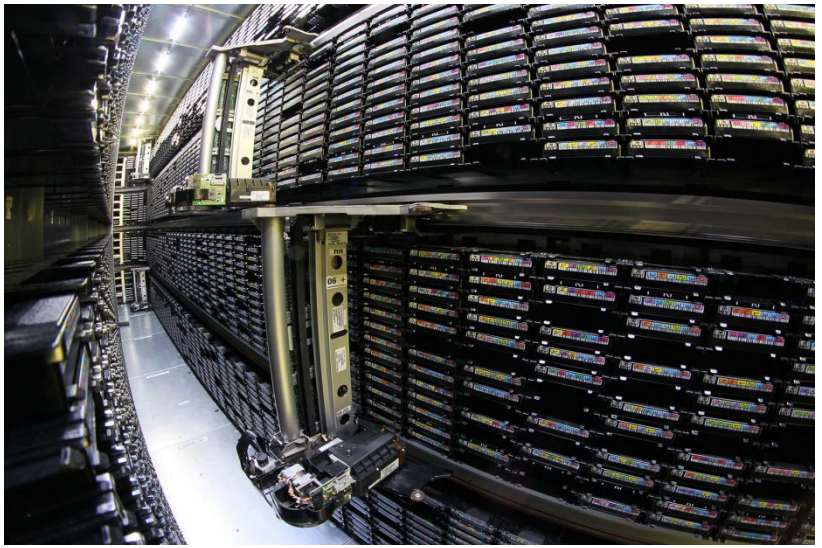
Advantages

- Two copies of data on tape (2nd in Garching (soon))
- Preserve existing directory structure
- Preserve data beyond project lifetime
- Keep descriptive metadata including public project reports

Archiving / DOKU

Preconditions

- Sufficient quota in DOKU
- Basic metadata available
- List of files to be archived



Workflow

- Within user portal http://luv.dkrz.de: Check project quota in DOKU, ask project administrator for consent
- Fill in basic metadata at http://cera-www.dkrz.de/LTA_metadata
- Create filelist and submit it at http://cera-www.dkrz.de/LTA_metadata
- DKRZ staff handles request, moves (or copies) data from /hpss/arch to /hpss/doku

Archiving / WDCC

Service for scientific projects to support reuse of data after project expiration

DKRZ offers

- Persistent data and metadata storage in certified repository beyond project lifetime >10 years
- Access using CERA account
- Implementation of access policies (embargo)
- Distribution of metadata to external research networks
- Archiving of ESGF data
- Access through ESGF

Advantages

- Preserve data beyond project lifetime
- Rich set of metadata facilitates subsequent reuse of data
- Two copies of data on tape (2nd in Garching (soon))
- Building trust through certification: data producer, data consumers, funders
- Prerequisite for DataCite DOI publication
- Easy to find in external search portals

Archiving / WDCC

Preconditions

- Sufficient quota in DOKU or separate agreement
- Full set of metadata available
- List of files to be archived

Workflow

- Within user portal http://luv.dkrz.de: Check project quota in DOKU, ask project administrator for consent
- Fill in metadata at http://cera-www.dkrz.de/LTA_metadata
- DKRZ staff handles request in close cooperation with data producer

DataCite DOI Data Publication at WDCC

Formal citation of data using DataCite data DOIs allows to give and to get credit for the preparation of high-quality research data.

WDCC/DKRZ services for data with a registered DataCite DOI:

- Data are long-term available and usable by an interdisciplinary user community (data curation).
- Data and formal data citation information are persistent.
- Data and metadata are continuously accessible via the unique persistent identifier DOI.

Requirements for the DataCite DOI data publication at WDCC/DKRZ:

- Detailed information on the data is provided (complete set of CERA2 metadata).
- Data are long-term archived at WDCC/DKRZ.
- Applied scientific data quality procedures are documented.

More Information: http://www.dkrz.de/daten-en/Datapublication

Data Management Plans (DMP)

Contact us. We can help!

Why DMP plans

- Mostly obligate for funded projects (EU, DFG...)
- It helps !
  - Clear data structures
  - data production workflow
  - access rights

Summary

Services:

- Data storage – disk and tape
- Data distribution – DKRZ-cloud and ESGF
- Data preservation – DOKU and WDCC
- Data publishing
- Data management plans

- Contact us at data@dkrz.de
- Don‘t forget to apply for (data) resources until 31-October-2015