Die Repositorien-Infrastruktur des Deutschen Klimarechenzentrums

DINI/nestor Workshop "Forschungsdatenrepositorien"

27.11.2017

Hannes Thiemann
Deutsches Klimarechenzentrum (DKRZ)

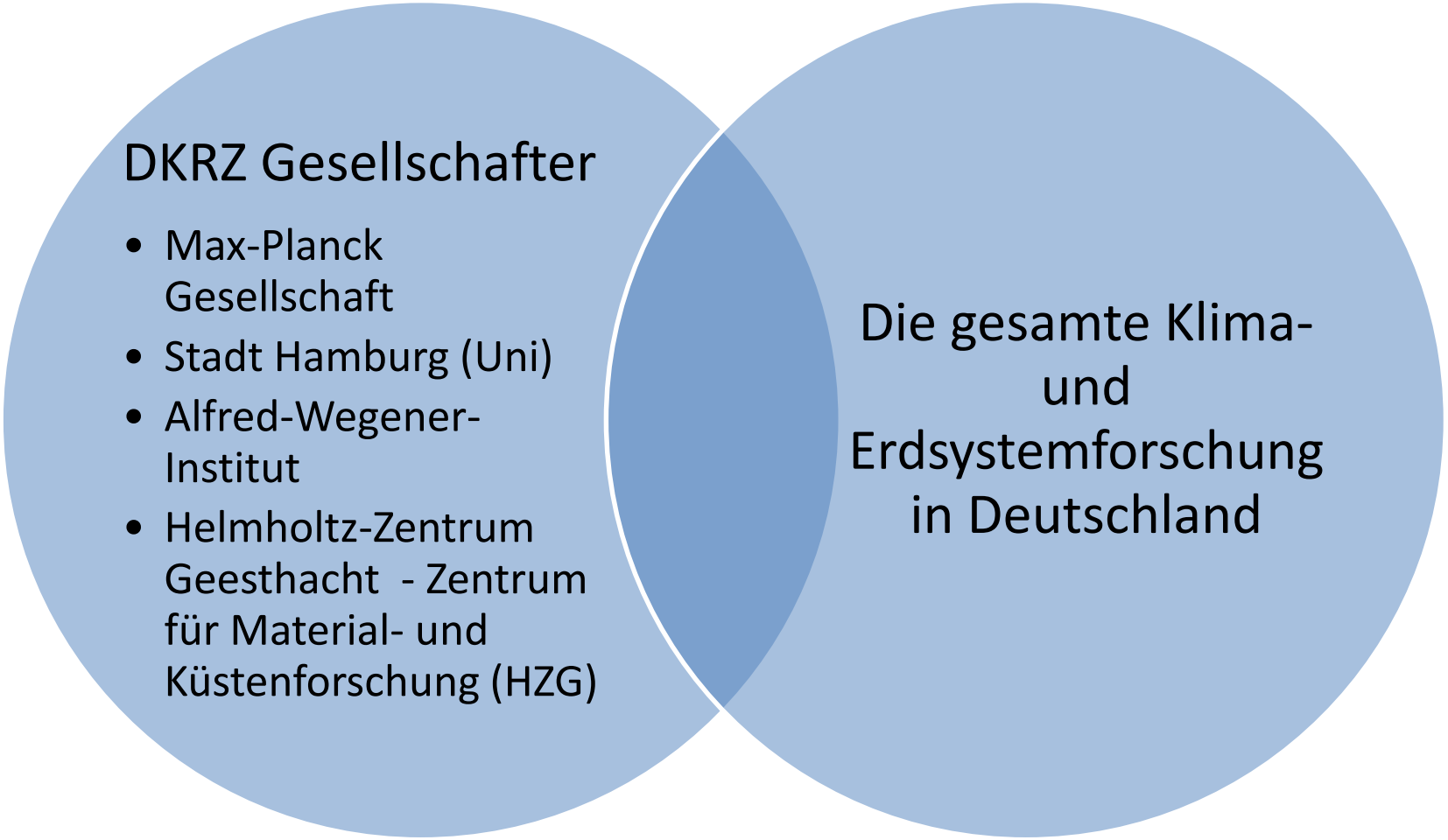
- DKRZ – a brief overview
- Large Repositories at DKRZ
  - ESGF
  - WDCC

Deutsches Klimarechenzentrum GmbH (DKRZ)

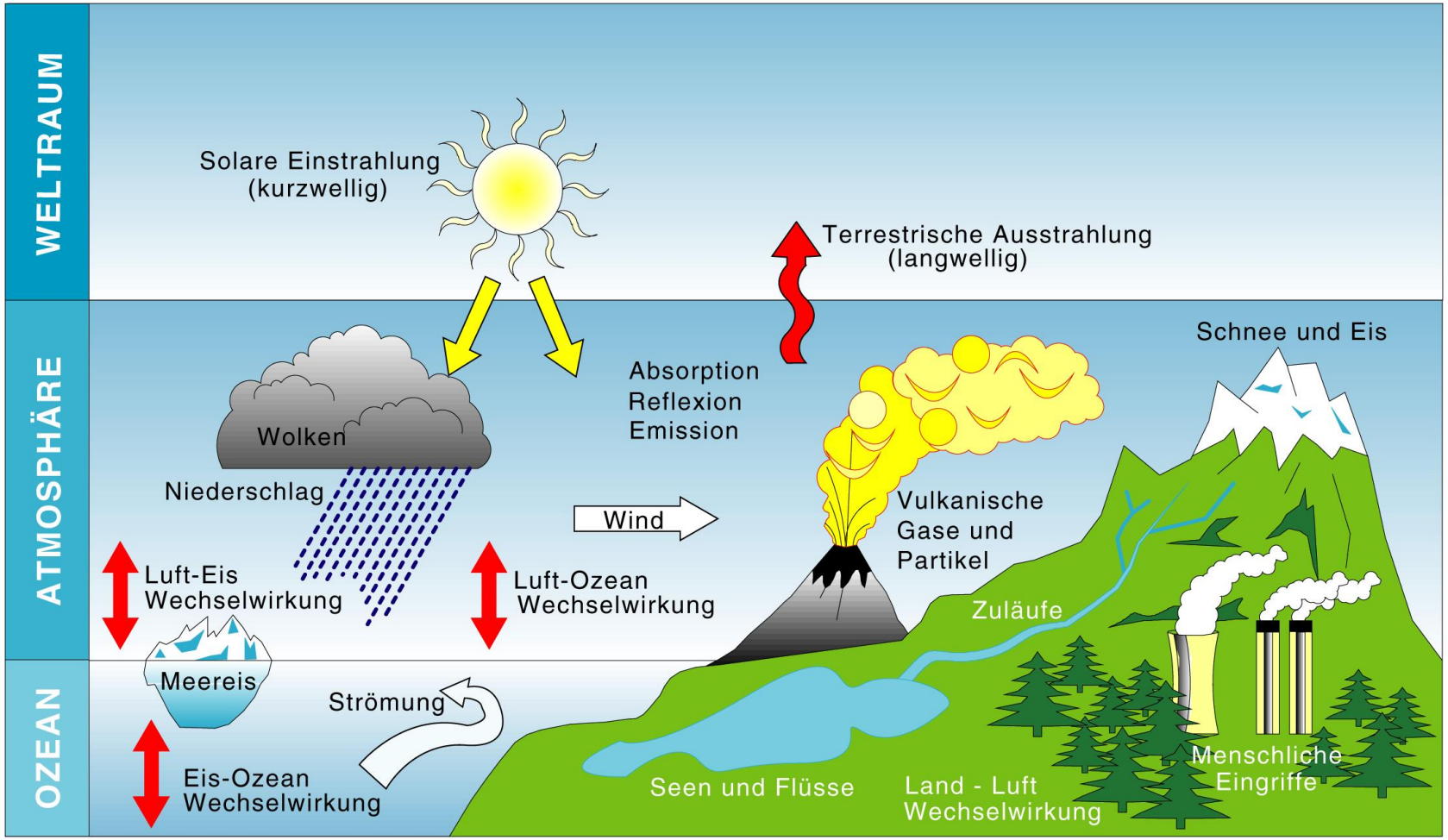
DKRZ - to provide high performance computing platforms, sophisticated and high capacity data management, and superior service for premium climate science.

DKRZ – to provide a unique combination of world-class computer power and expert personnel to enable superior climate modelling.

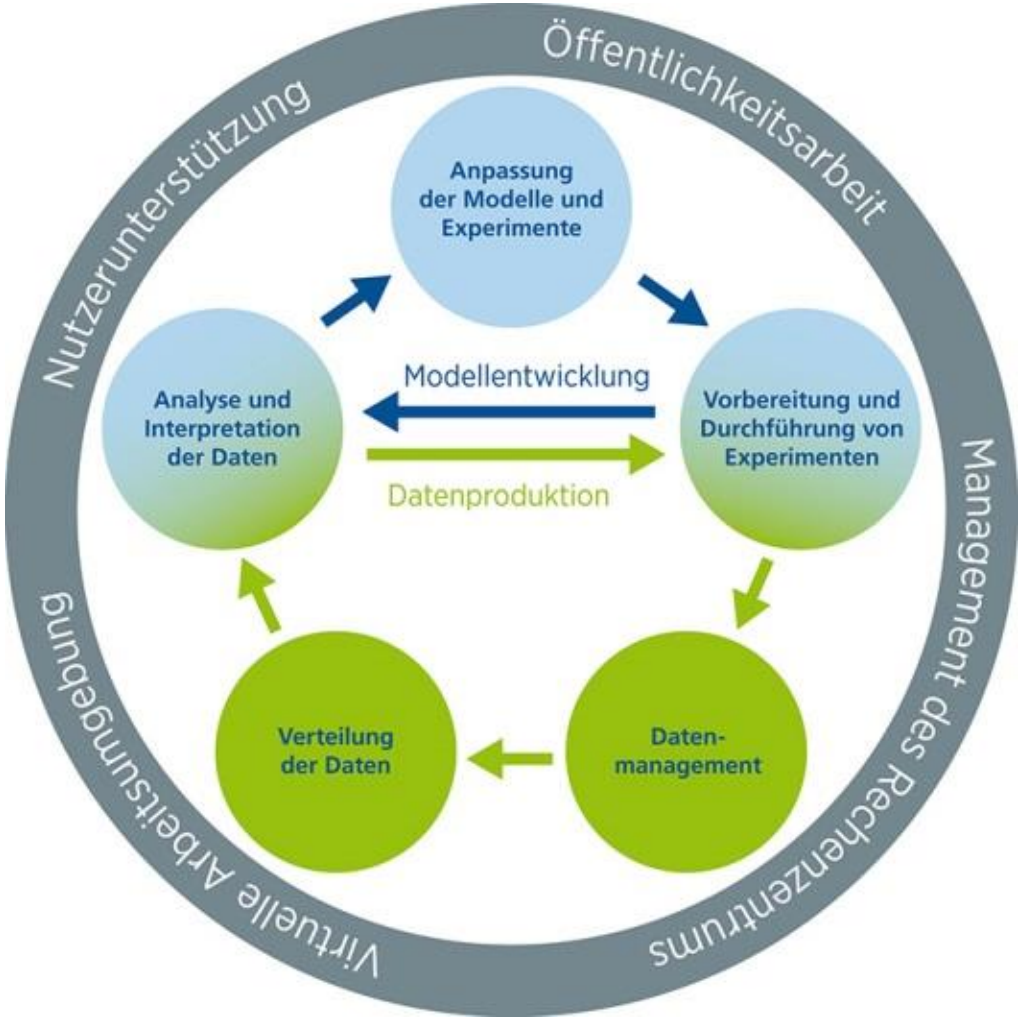
Für wen?


DKRZ Gesellschafter
• Max-Planck Gesellschaft
• Stadt Hamburg (Uni)
• Alfred-Wegener-Institut
• Helmholtz-Zentrum Geesthacht - Zentrum für Material- und Küstenforschung (HZG)
Die gesamte Klima- und Erdsystemforschung in Deutschland

Klimasystem



Das DKRZ bietet viele Dienste um die wissenschaftlichen Arbeitsflüsse der Klimamodellierer zu unterstützen

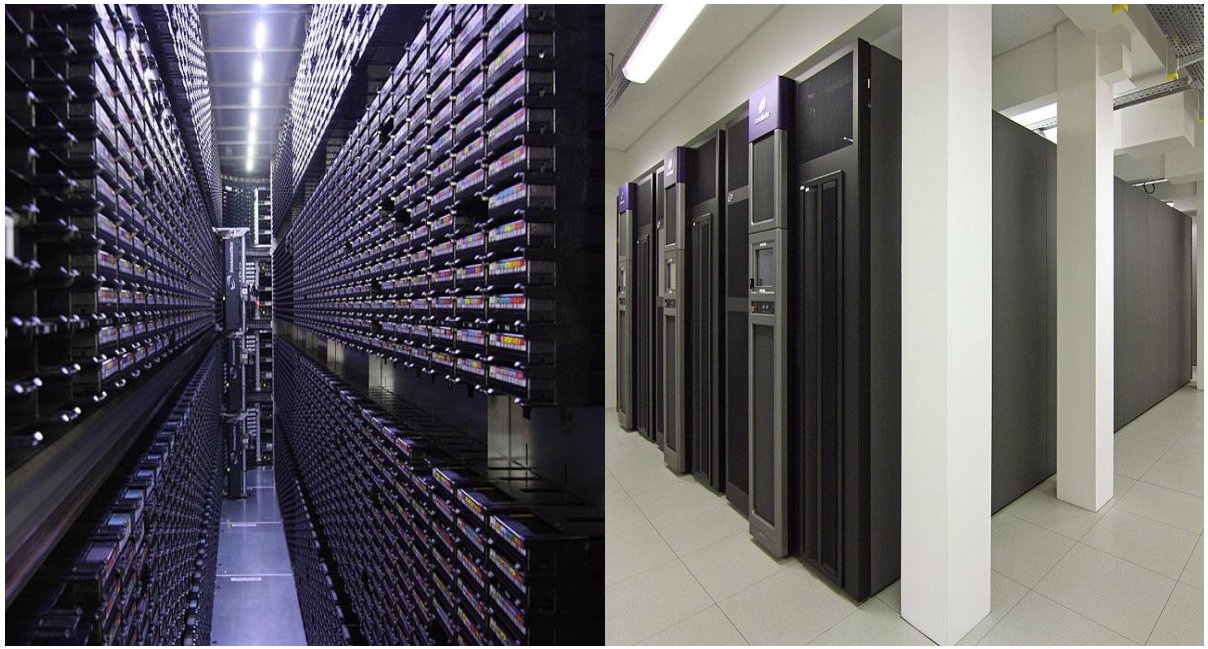


HLRE-3 – Mistral (2015-2020)

bullx DLC 720, 3,300+ nodes, 100,000+ cores, Haswell/Broadwell, 3.6 PFLOPS
240 TB main memory, **54 PB disk storage**, 450 GB/s mem-disk rate, FDR network
21 nodes for visualization
hot liquid cooling with high efficiency

High Volume Data Archive

- 65,000 slots for tapes in Hamburg (10,000 remote)
- 60+ PB of climate simulation data
- increase 8 PB/y before 2015, now up to 9x more
- **500 PB capacity until 2020**



CMIP(5) – Coupled Model Intercomparison Project

- Provides key input for the IPCC report
  - 5th AR, 2013
- ~20 modeling centers around the world
  - DKRZ being one of the biggest
- Produces 10s of PBytes of output data from ~60 experiments ("digital born data")
- Stored in ESGF (Earth System Grid Federation)

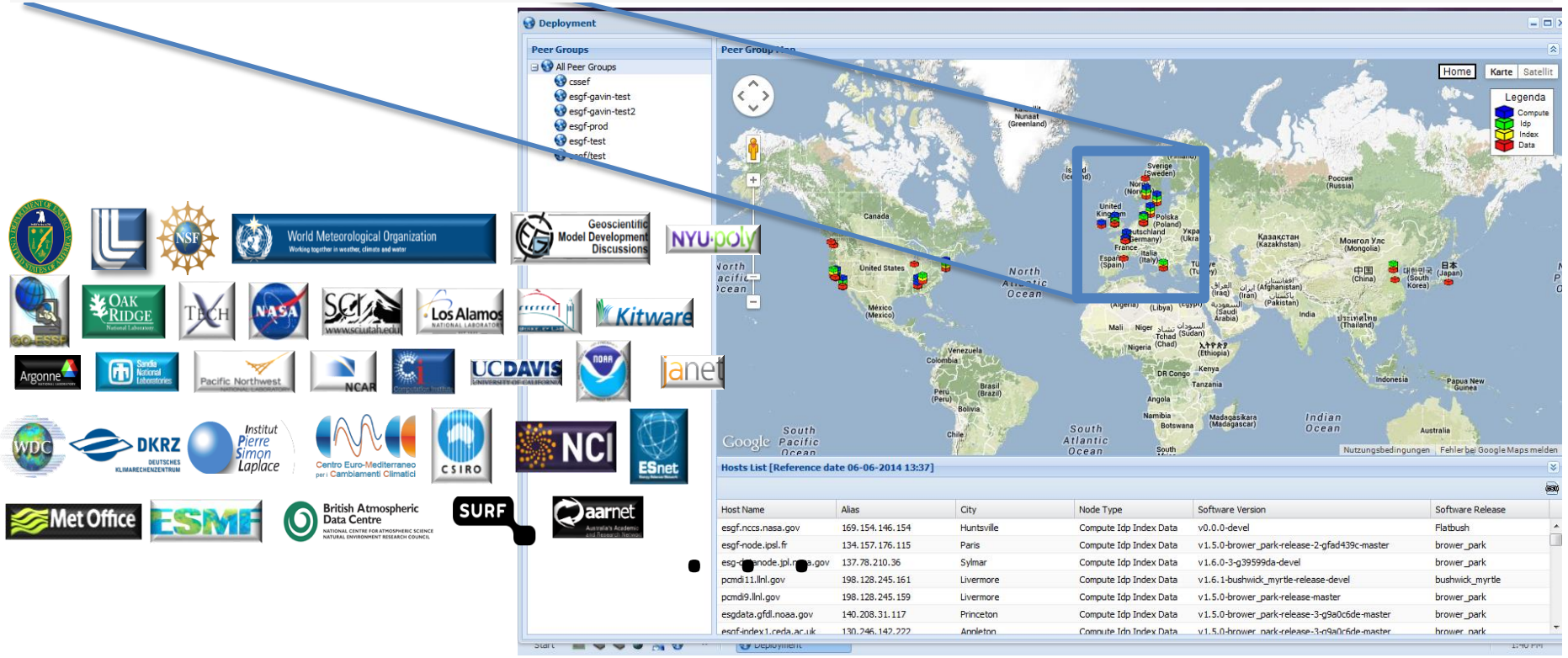
Data are produced without knowing all applications beforehand and these data are stored and archived for interdisciplinary utilization by yet unknown researchers

Die Earth System Grid Federation (ESGF)

Was ist ESGF ?

- Eine weltweite Kollaboration von Universitäten und Forschungseinrichtungen, sowie Daten- und Rechenzentren gefördert aus unterschiedlichen nationalen und institutionellen Quellen
- Eine open source software initiative für eine verteilte Daten- und Rechen-Plattform für Klimadaten
- Eine operationelle Infrastruktur (u.a.) zur Unterstützung von CMIP Projekten

IS-ENES: europäische ESGF Föderation + Klimafolgenforschungsportal



ESGF Datenpublikation

Voraussetzungen:

- …
- Die Daten berücksichtigen die Standards und Konventionen, die im Rahmen des jeweiligen Datenvergleichskontextes festgelegt wurden (z.B. CMIP, CORDEX, Obs4Mips etc.)
- Die Daten unterliegen einer einheitlichen Regelung zur Datennutzung (z.B. offen nutzbar bzw. für nichtkommerzielle Zwecke offen nutzbar)

Vorgehen

- …
- Qualitätsprüfung durch das DKRZ
- Publikation auf DKRZ ESGF Infrastruktur
- Neue Versionen etc. mit Bekanntgabe von Errata Informationen zu den Daten

Long-term archiving at DKRZ

- Two flavours available
  - DOKU
    - support traceability of the scientific project after project expiration
  - World Data Centre for Climate (WDCC)
    - to support reuse of data after project expiration

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 (Hamburg & Garching)
- Preserve data beyond project lifetime
- Keep descriptive metadata including public project reports
- Preserve existing directory structure

Archiving / WDCC

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

WDC CLIMATE

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)
- Building trust through certification: data producer, data consumers, funders
- Prerequisite for DataCite DOI publication
- Easy to find in external search portals

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.

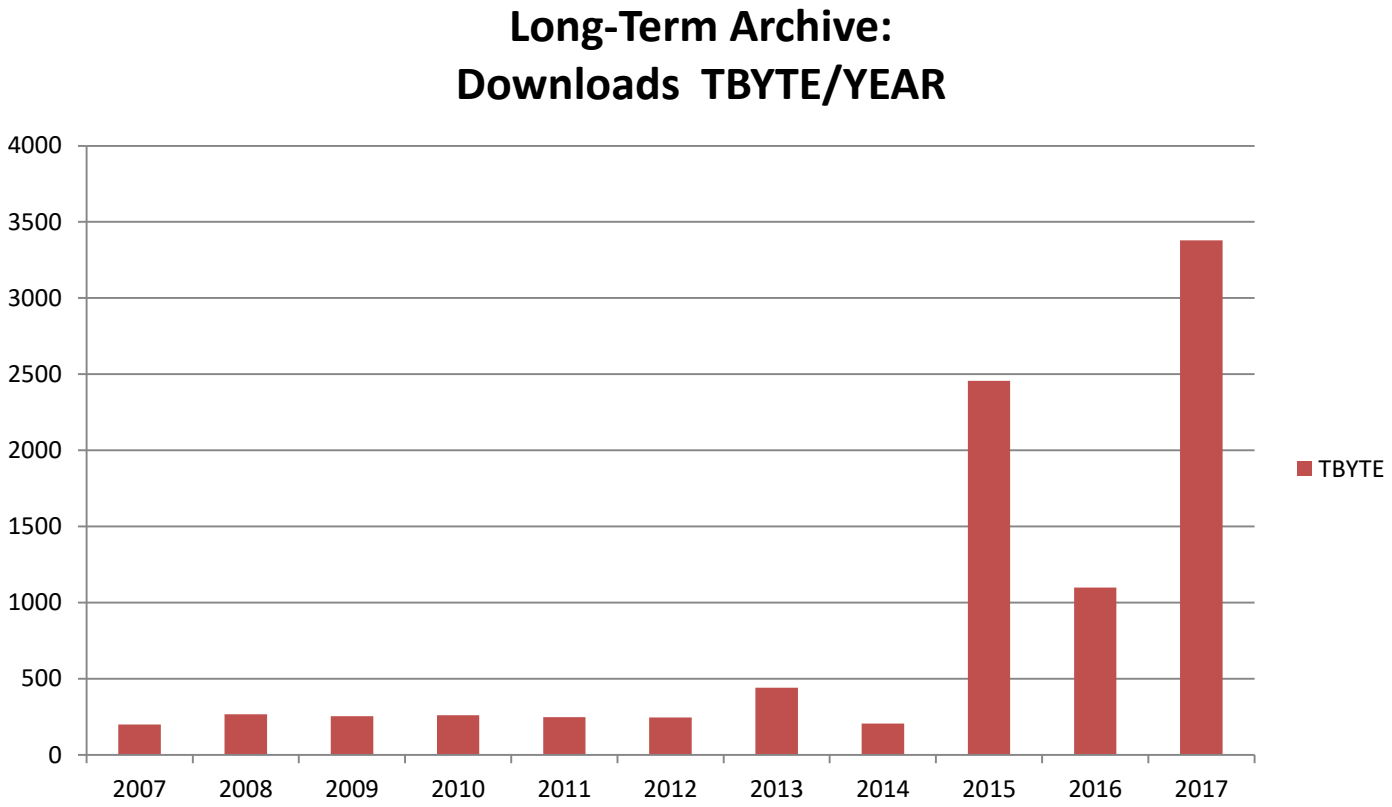
WDCC/DKRZ services for data with a registered DataCite DOI:

- 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.

Usage Statistics



Key points

- Data distribution and preservation in the range of petabytes.
- Quality assessment of all data managed within ESGF and WDCC.
- Free of charge for data producers and data consumers.