

# Langzeitarchivierung am D K R Z

Workshop

Archivierung sozial- und wirtschaftswissenschaftlicher  
Datenbestände

Deutsche Nationalbibliothek Frankfurt

15./16. Sept. 2011

Hans Luthardt DKRZ/DM

# DKRZ Mission

Deutsches Klimarechenzentrum:

- höchste Rechenleistung,
- ausgereiftes Management größter Datenmengen und
- kompetenter Service garantieren erstklassige Klimaforschung.

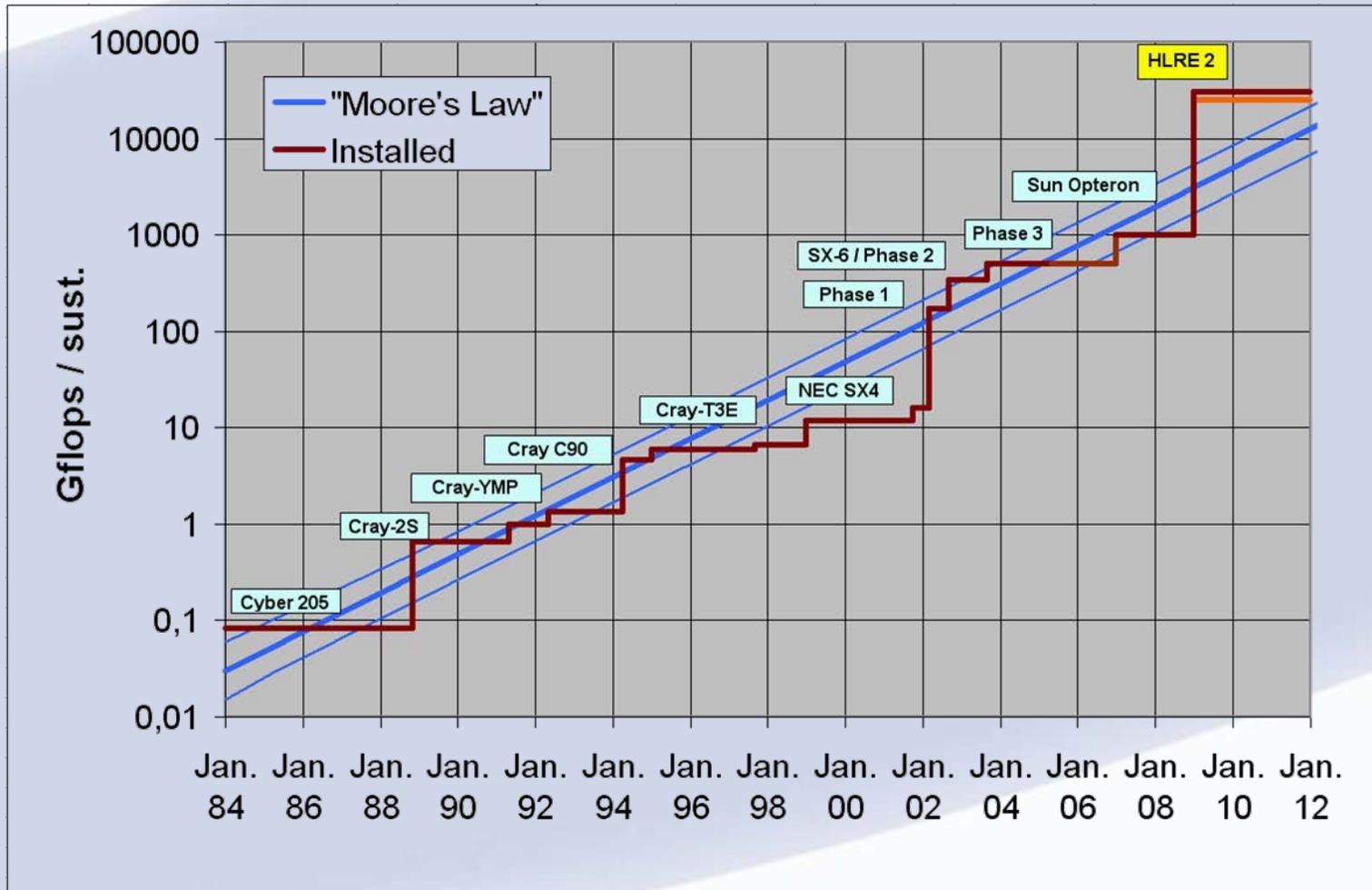
# Rechnerhardware

- 158 TeraFlops ( $158 * 10^{12}$  Gleitkommaoperationen / Sekunde)
- 264 IBM Power6-Rechnerknoten
- 16 Dual-Core-Prozessoren pro Knoten (insgesamt 8.448 Kerne)
- Mehr als 20 TeraByte Hauptspeicher
- 7 PetaByte Festplattenspeicher ( $7 * 10^{15}$  Byte)
- Infiniband-Netzwerk mit 7,6 TeraByte/s aggregierter Übertragungsrate



- grün = Festplatten des Höchstleistungsrechner „Blizzard“
- orange = Höchstleistungsrechner „Blizzard“
- rot = Schaltzentrale (Infiniband) des Höchstleistungsrechners „Blizzard“

# Entwicklung der Rechenleistung

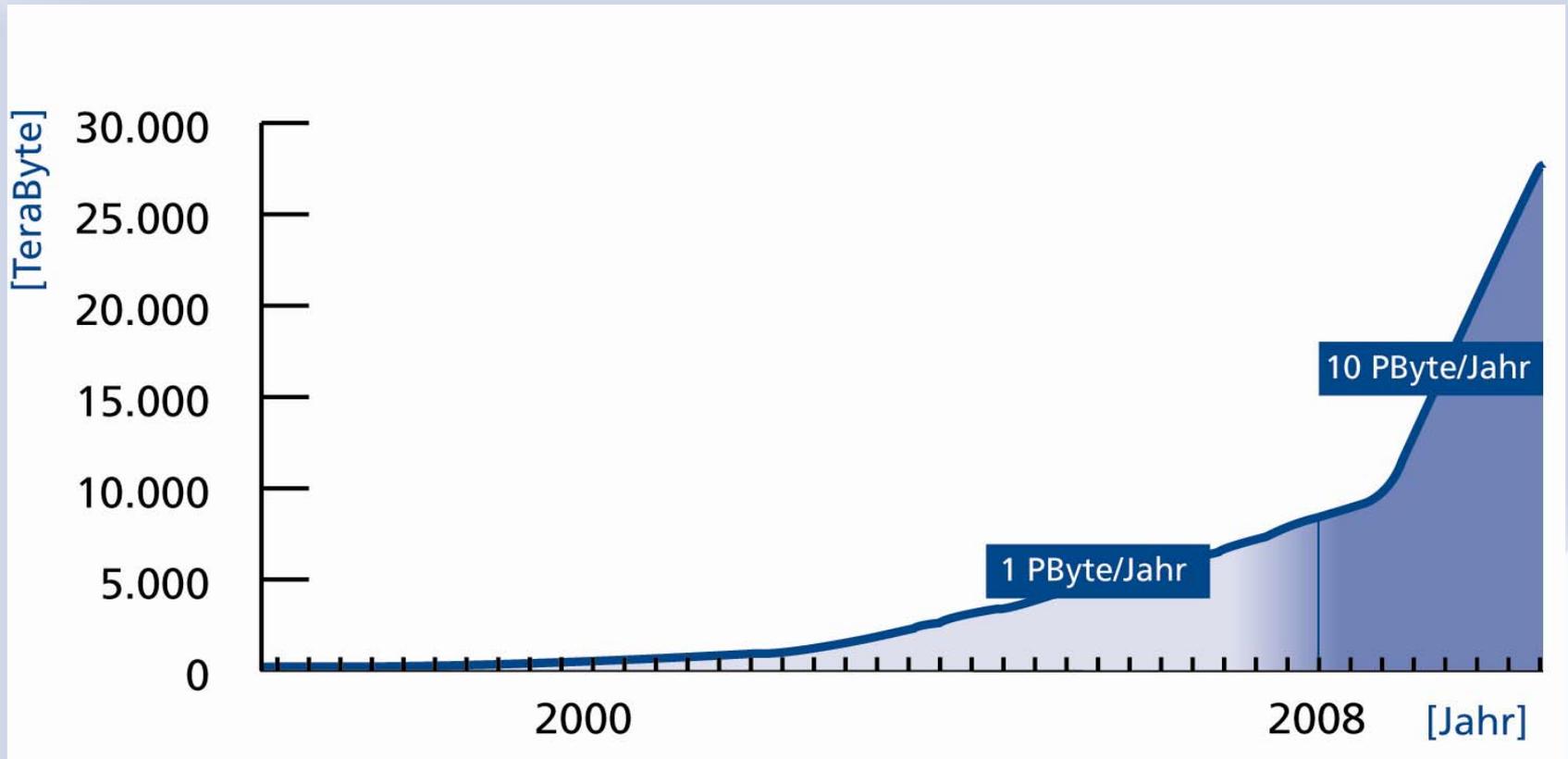


# Datenspeicher

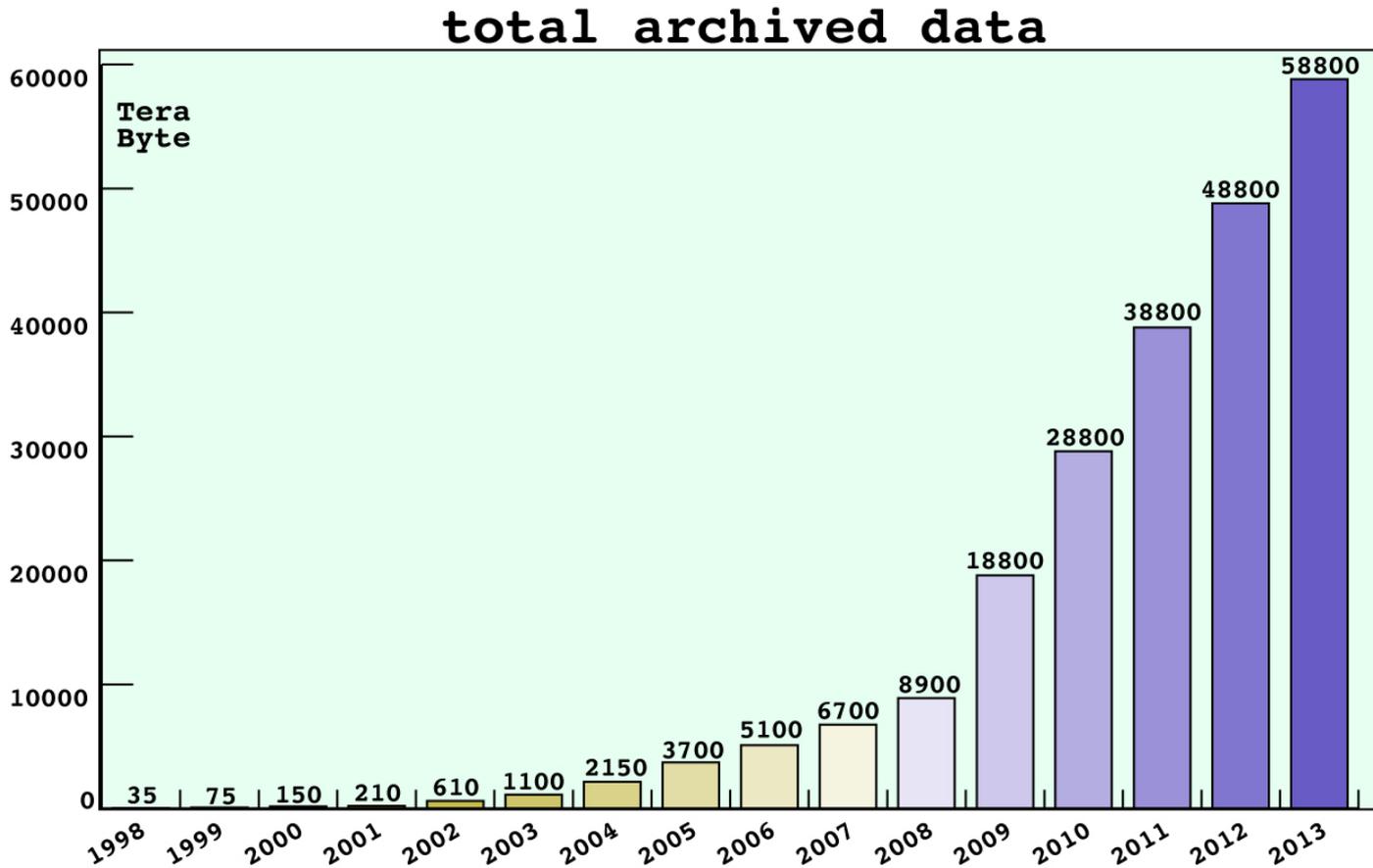
- 7 automatische Sun StorageTek SL8500-Bandbibliotheken
- 8 Roboter je Bibliothek
- mehr als 67.000 Stellplätze für Bänder mit Gesamtkapazität von ca. 100 Petabyte
- 88 Bandlaufwerke
- bidirektionale Bandbreite von 5 GigaByte/s



# Entwicklung der Datenspeicherung



# Datenspeicherung : Filesbasiert



# WDCC – Word Data Center on Climate

**Start:** Approved in January 2003

**Maintenance:** Model and Data (M&D/MPIMET) and German Climate Computing Centre (DKRZ)

**Mission:** Data for climate research are collected, stored and disseminated

**ICSU Policy:** long-term archiving and unrestricted data access for scientists

**Restriction:** Only climate data products in CERA DB, no raw data storage.

**Content:** Emphasis is spent on climate modelling and related data products.

**Co-operation:** with thematically corresponding data centres like WDC-MARE (Bremen) and WDC-RSAT (Oberpfaffenhofen)

URL: <http://www.dkrz.de/daten-en/wdcc>

# WDCC – Word Data Center on Climate



- Approved in 2003
- Hosts several projects and Data Centres
- WDCC operates as a long-term data archive (10years +)
- WDCC is implemented within the CERA data and information system.
- Data are stored in conjunction with metadata.
- WDCC offers the publication service for primary data. (DOI)
- Approximately 5 person staff and 500 TB of data.
- Increase of a 1 PB/year starting in year 2011

**CERA: General Statistics at 01-09-2011 00:00:18**

**Internal data**

**Database Size (TByte): 434**

**Number of container: 183038**

**Number of blobs: 8586769505**

# Klimadaten

- Climate model results from global and regional climate models from different climate modelling centres

CCCma, CCSR/NIES, CSIRO, GFDL, HADLEY, MPIfM , NCAR  
based on IPCC-emission scenarios

- Data from scientific projects

HOAPS (satellite data), CARIBIC (civil aircraft data), GOP, COPS, CEOP

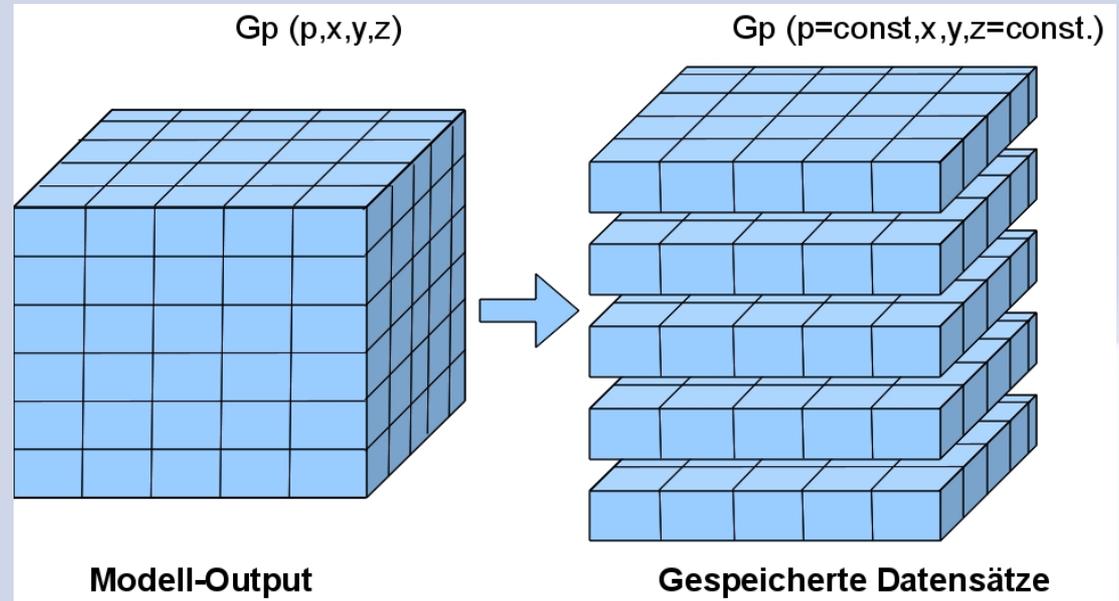
- Model like Observations

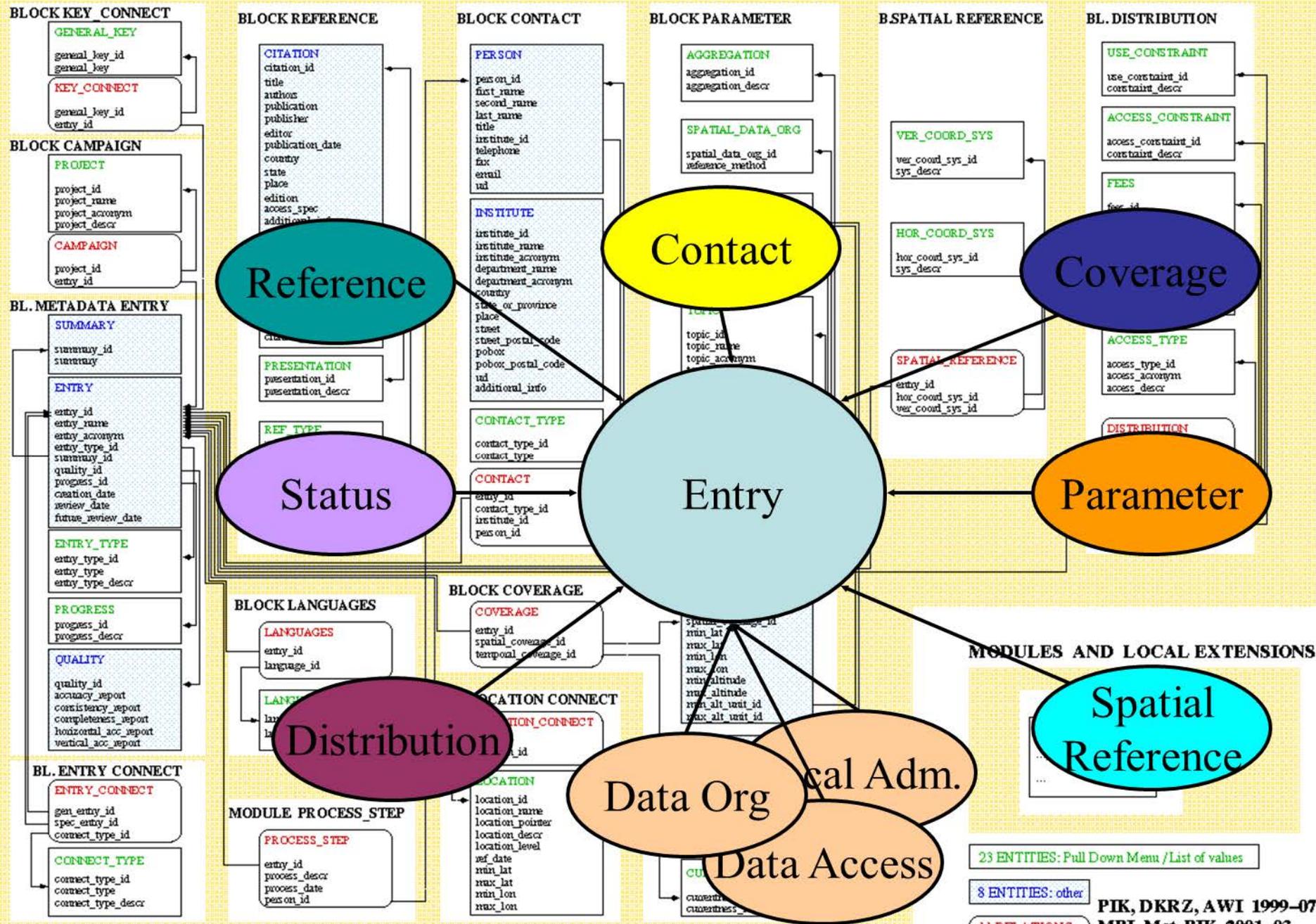
Reanalyses data

# Aufbereitung der Datensätze

## Aufbereitung der Datensätze

- Benutzerbedürfniss
- Reduzierung der herunterzuladenden Datensatzvolumen
- Verkleinerung der Transfervolumens



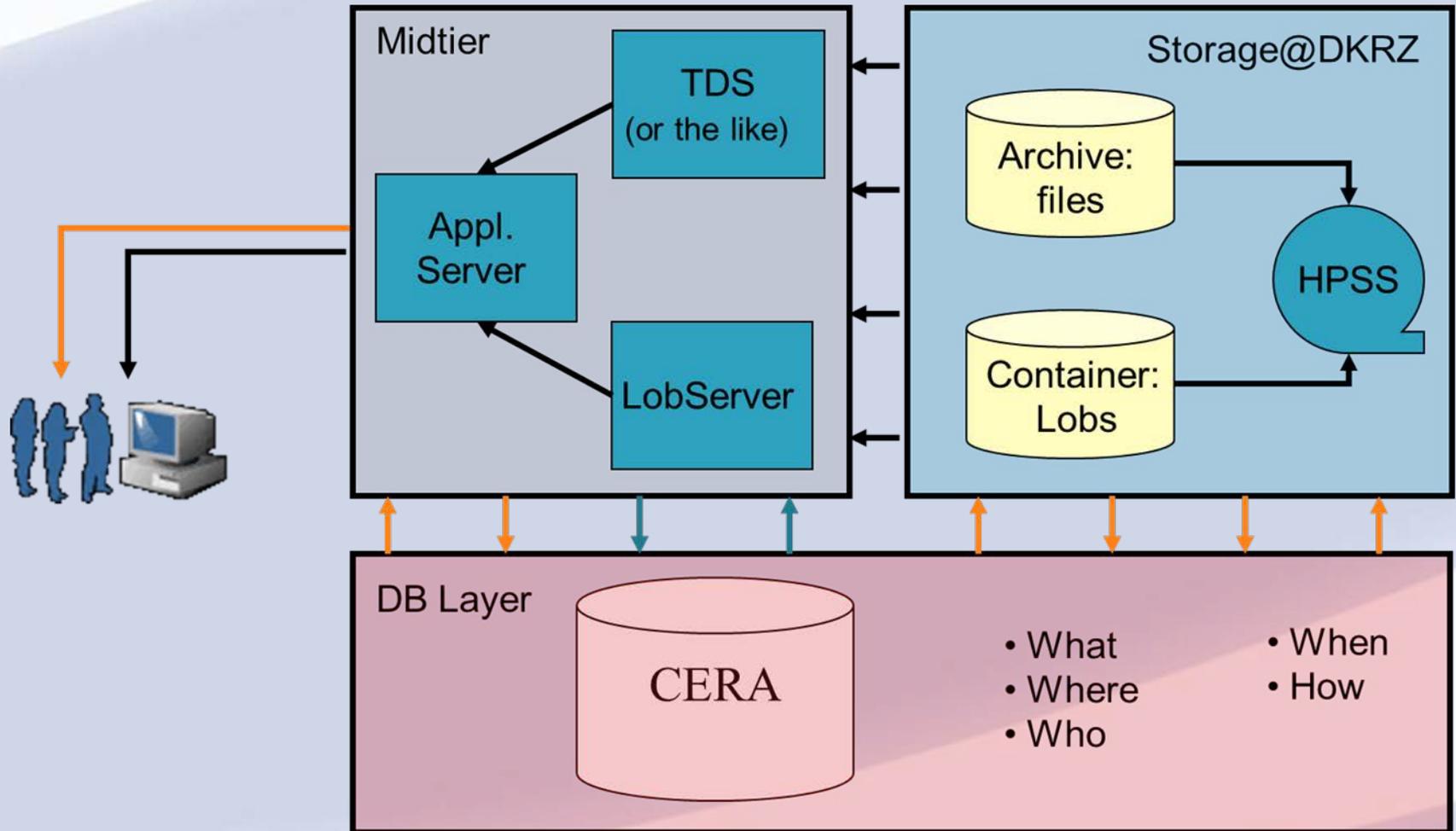


23 ENTITIES: Pull Down Menu / List of values

8 ENTITIES: other

11 RELATIONS

# WDCC Datenzugriff

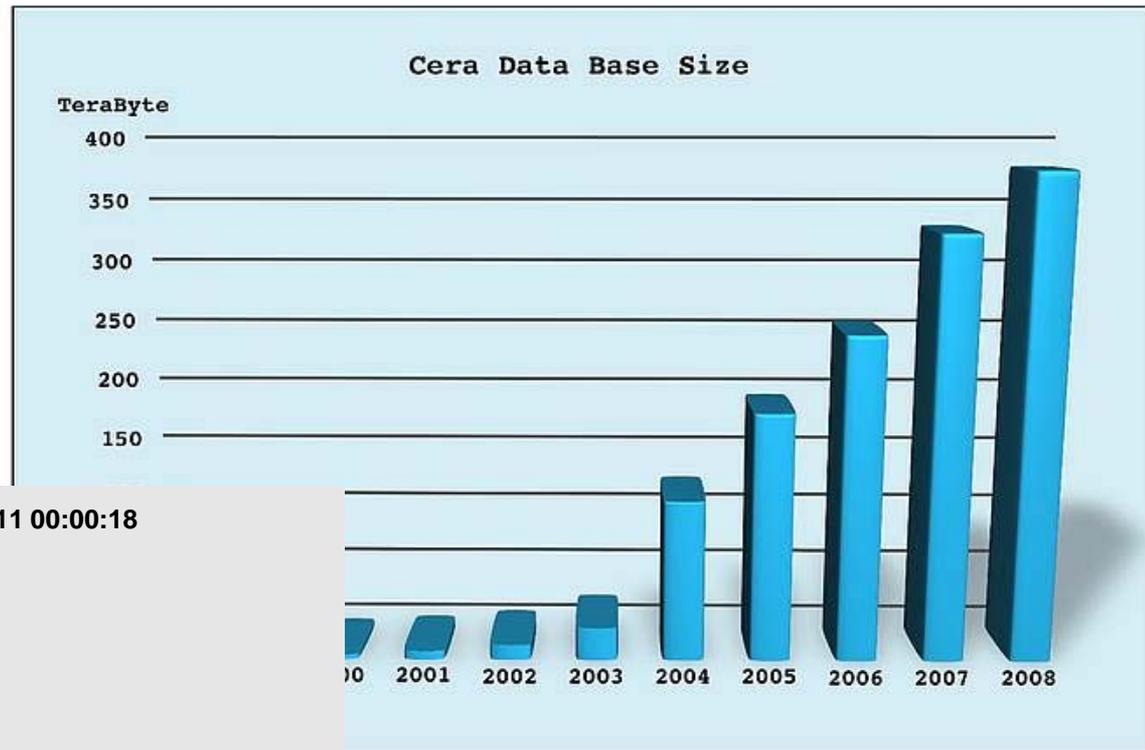


# Web Portal der CERA- Datenbank : Suchen und Download

The screenshot shows the 'Welcome to the CERA WWW-Gateway V1.4.0' interface. At the top, there are navigation buttons: 'Back', 'Home', 'Process List', 'Search by Name', 'Search by Topic', and 'Manual'. The main area is divided into several sections:

- Keyword:** A list of project codes including 1PCTT02X, 1PCTT04X, 20C3M, AMIP, AMIP2, BCC, BCC-CM1, BCCR, BCCR-BCM2.0, BMRC, CARIBIC, CCCma, CCCma CGCM2, and CCCma CGCM2. Below the list are 'Select' and 'Clear Selection' buttons, and a 'Selected Keyword' input field.
- Project:** A list of project names including IMPETUS: An integrated approach to the efficient management of climate data, IPCC Data Distribution Centre, IPCC simulations at the University of Bonn/METRI, IPCC-Hamburg Climate Model Simulation (highlighted), NCEP Re-Analysis Project, Observational Data, Ocean Gateways, Paleo-Climate Model Simulations at University of Bremen, Paleo-Climate Model Simulations at University of Utrecht, Publication and Citation of Scientific Primary Data, SFB Subproject F2, Time Slice Experiments at CSCS, and Time Slice Experiments at MIPM. Below the list are 'Select' and 'Clear Selection' buttons, and a 'Selected Project' input field.
- Experiments:** A list of experiment identifiers including 19990906\_MLE\_DUS, 19991211\_MUC\_CMB, 19991212\_CMB\_MUC, 20000118\_MUC\_MLE, 20000119\_MLE\_DUS, 20000208\_MUC\_MLE, 20000209\_MLE\_DUS, 20000325\_DUS\_MLE, 20000326\_MLE\_MUC, and 20000414\_DUS\_CMB. To the right are buttons for 'Information', 'Contacts/Refs', and 'Datasets'.
- Login Information:** A text box showing 'You are logged in as BOSCH', a 'Login to CERA database' button, and an 'About' section with an 'About CERA WWW-Gateway' button.

# Datenspeicherung : Semantisch-Datenhaltung



CERA: General Statistics at 01-09-2011 00:00:18

Metadata

Number of projects: 80

Number of experiments: 1439

Number of ds groups: 280

Number of datasets: 167802

Number of add\_info: 237

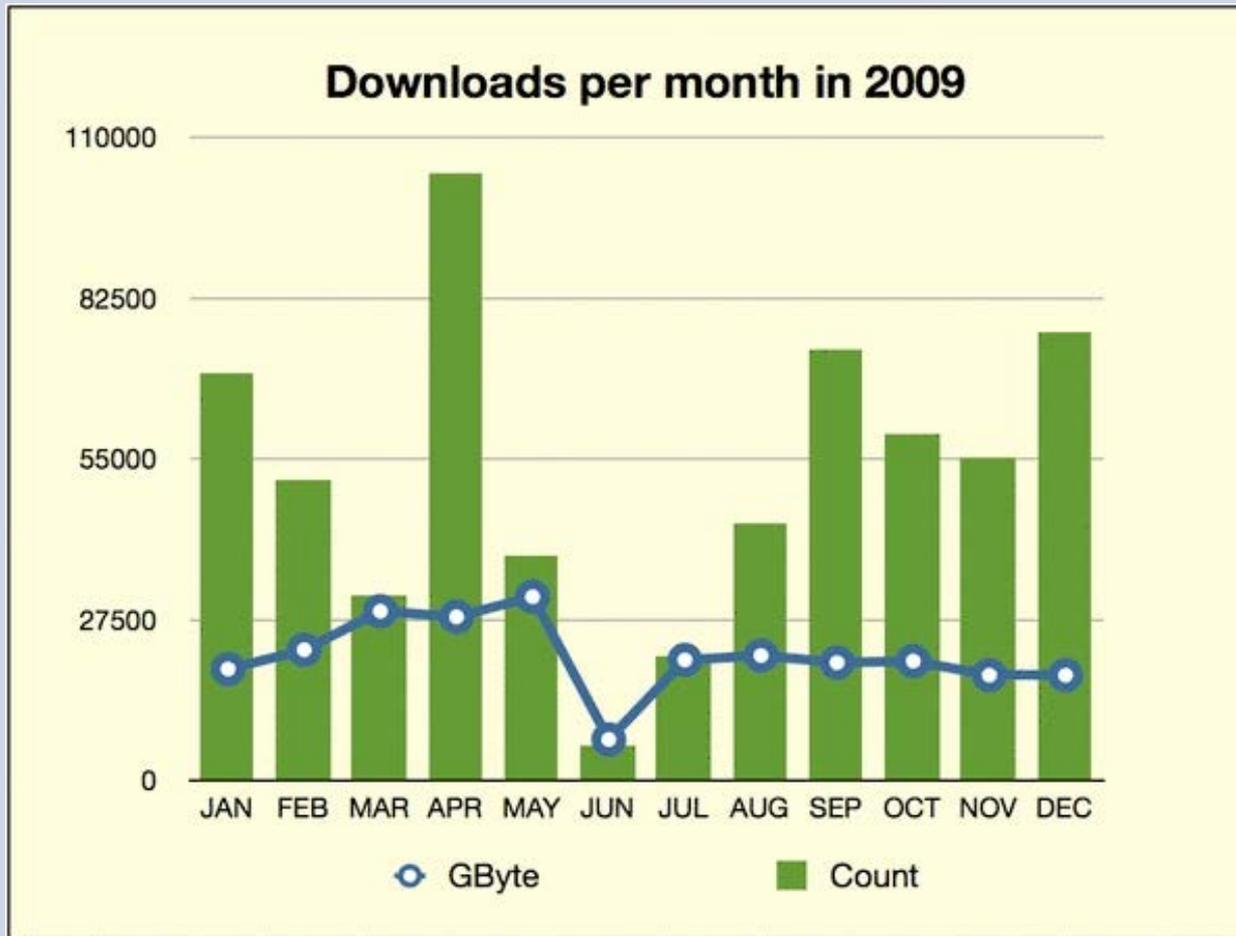
Internal data

Database Size (TByte): 434

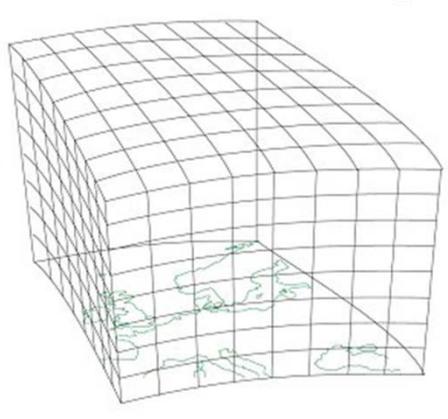
Number of container: 183038

Number of blobs: 8586769505

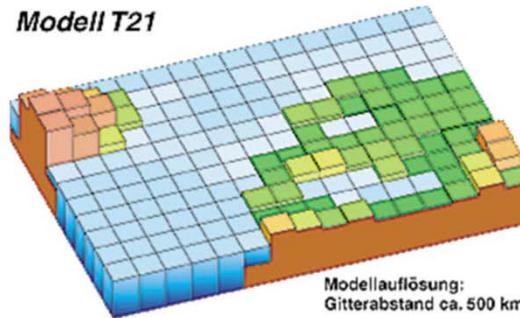
# Downloads aus dem WDCC



# Datenvolumen

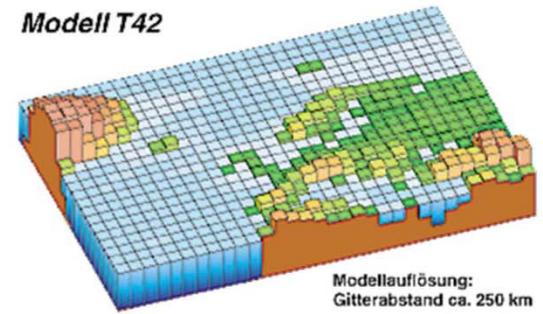


**Modell T21**



Modellauflösung:  
Gitterabstand ca. 500 km

**Modell T42**



Modellauflösung:  
Gitterabstand ca. 250 km

10-fache Rechenzeit!  
>6-fache Datenmenge!

# Langzeitarchivierung am DKRZ

Service wird angeboten für:

Nutzer des DKRZ :

- aus den Einrichtungen der Gesellschafter entsprechend ihren Anteil
- aus dem BMBF-Anteil geförderten Nutzen

Externe Nutzen :

gegen Kostenerstattung

Geplante Speicherkapazität pro Jahr : 2 PetaByte

# DFG-Request zur Archivierung von Forschungsdaten

## Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten

Stand 26.6. 2008

Forschungsprimärdaten bilden einen wertvollen Fundus an Informationen, die mit hohem finanziellem Aufwand erhoben werden. Je nach Fachgebiet und Methode sind sie replizierbar oder basieren auf nicht wiederholbaren Beobachtungen oder Messungen. **In jedem Fall** sollten die erhobenen Daten **nach Abschluss der Forschungen öffentlich zugänglich und frei verfügbar** sein. Dieses ist die wesentliche Voraussetzung dafür, dass Daten im Rahmen neuer Fragestellungen wieder genutzt werden können sowie dafür, dass im Falle von Zweifeln an der Publikation die Daten für die Überprüfung der publizierten Ergebnisse herangezogen werden können.

1997 veröffentlichte die DFG „Vorschläge zur Sicherung guter wissenschaftlicher Praxis“ mit 16 Empfehlungen. Die Empfehlung 7 lautet **„Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden.“**

...

# Datenvolumen

- **Horizontalaufösung des Klimamodells**

- T42:  $128 * 64 = 8192$  Punkte pro Globalfeld
- T106:  $160 * 320 = 51200$  Punkte pro Globalfeld

- **Erforderliche Speichereinheiten (GRIB Format)**

- **Horizontalfeld (Zugriffseinheit):**

- 17.1 kB (T42)
- 100.1 kB (T106)

- **Unix Filegröße für monatsweise akkumulierte Ergebnisse mit 6 Std. Speicherintervall und 300 2d Variablen (Physikalische Einheit):**

- 616 MB (T42)
- 3500 MB (T106)

- **240 Jahre Modellintegration (Logische Einheit):**

- 1.7 TB (T42)
- 10 TB (T106)

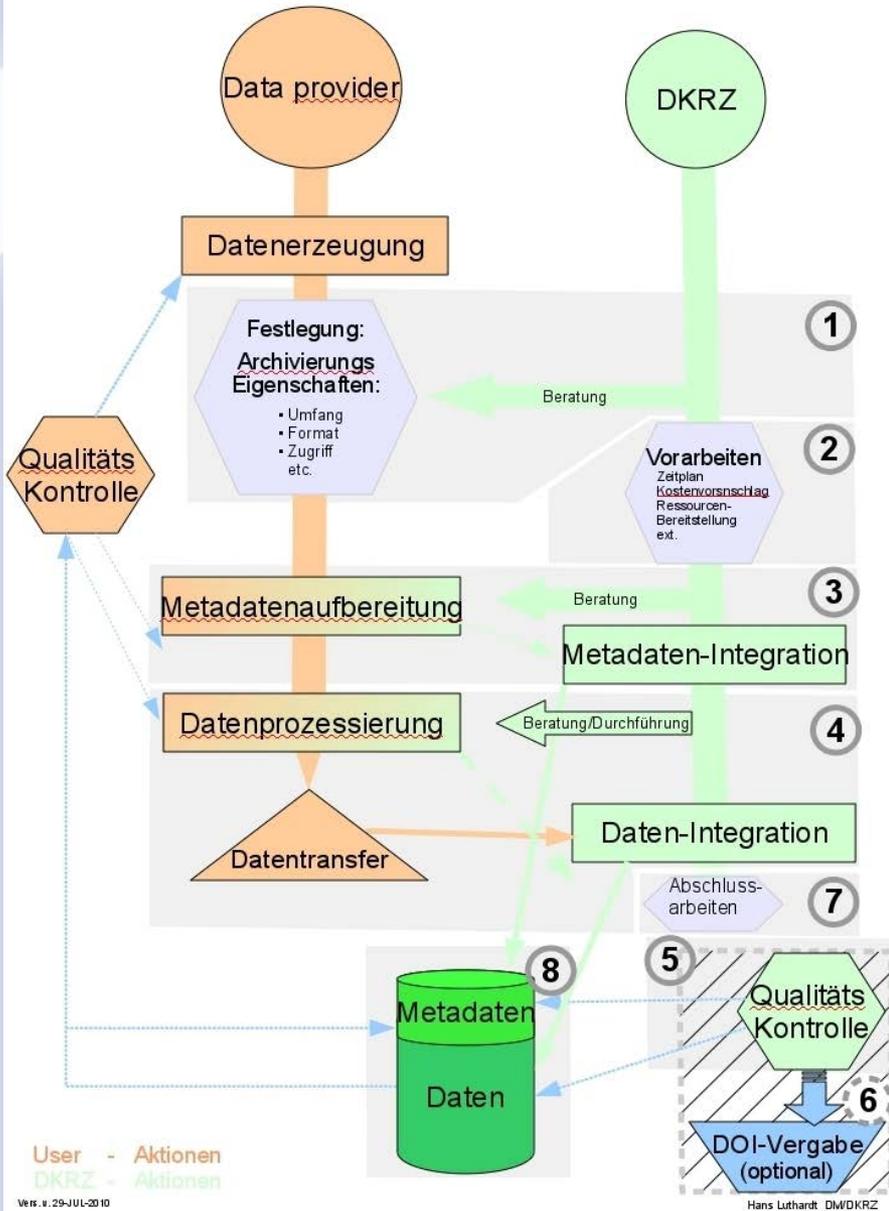
# Kostenfaktoren

## Kostenfaktoren bei der Langzeitarchivierung:

- Arbeitsaufwand bei Beratung, Einrichtung, Erstellung und Prüfung der Metadaten, ggf. Preprocessing, Einfüllen
- Rechenzeit
- Datenträger (für 10 Jahr)
- Betriebskosten: Datenarchiv, Internetzugang , ...

# Workflow

## Langzeit-Datenspeicherung am DKRZ



# Erzeugungstools für Metadaten

Form On Dkrz Lta - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://cera-www.dkrz.de/pls/apex/f?p=112:2:4484077964503547::NO

Most Visited LEO Getting Started Latest Headlines

Form On Dkrz Lta



DEUTSCHES KLIMARECHENZENTRUM

K204026 Logout

Home WDC Metadata Form New Person New Institute New Project New Citation New Code List Upload Area Help

Create Metadata for WDC Cancel

Create New Metadata Entry

Entry Name

Summary

Authors

Project

Investigator

Metadata

Temporal Coverage

Start Year

Month

Day

Stop Year

Month

Day

Currentness Ref

Spatial Coverage

Min Lat

Max Lat

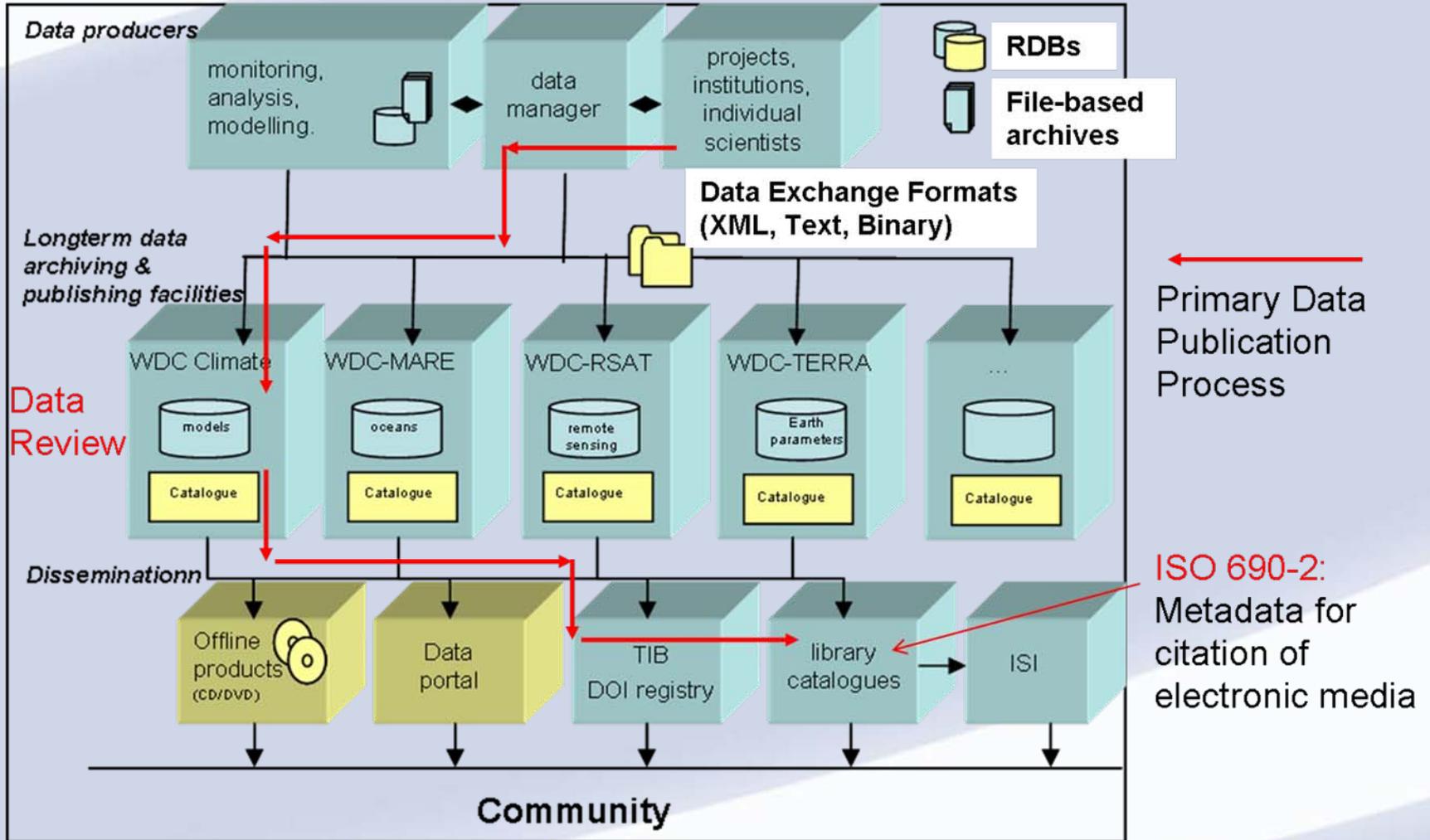
Min Lon

Max Lon

# Qualitätskontrolle

- Durch den Datenerzeuger :
  - wissenschaftliche Korrektheit
  - Richtigkeit und Vollständigkeit der Metadaten
  - Konsistenz von Daten/ und Metadaten im Langzeitarchiv
  
- Durch Datenmanagement :
  - Korrektheit des Postprozessrings/Dateneinfüllens
  - Konsistenz und Vollständigkeit der Datensätze
  - Überprüfung der Zugriffs/Download-Mechanismen

# DOI - Vergabe



# WDC-Climate hosted data in the catalogue at TIB Hannover

**TIBORDER - Dokumentlieferdienst der TIB Hannover - results/titledata - Mozilla Firefox**

Titelliste Titeldaten

Suchergebnis sichern  
Datenbankauswahl  
Bestellung ohne Recherche  
Benutzerinfo  
TIB Homepage

Titel: [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_6 HOUR values MPImet/MaD Germany](#) / World Data Center for Climate (WDCC)  
Beteiligt: Erich Roeckner ; Michael Lautenschlager ;  
Körperschaft: World Data Center for Climate (WDCC)  
Erschienen: Hamburg : World Data Center for Climate  
Umfang: Online-Ressource (3987170720028 Bytes)  
Anmerkung: Mode: Abstract  
StructuralType: Digital  
CreationDate: 2004-05-11

Inhalt: The data represent 6 hourly values of a 2000 year experiment with observed anthropogenic forcings(CO2) in year 2190 of the preindustrial control run. Technical data to this experiment: Data Sets with monthly mean values are available. The experiment is using ECHAM5.2.02a control output from the model run: hurrikan.dkrz.de. Please note: experiment\_name/acronym was changed to 20C\_1)

Technische Angaben: Format: GRIB  
Links: doi: [10.1594/WDCC/EHS-T63L31\\_OM-GR1](#)  
URN: [urn:nbn:de:tib-10.1594/WDCC/EHS-T63L31\\_OM-GR1](#)

Bestandsinfo: [Anzeigen](#) [lizenzfrei!](#)  
Anmerkung: Primaerdaten

Fertig

**TIBORDER - Dokumentlieferdienst der TIB Hannover - results/shortlist - Mozilla Firefox**

Einfache Suche | Erweiterte Suche | **Suchergebnis** | Zwischenspeicher | Suchgeschichte | Hilfe © 1998-2007 OCLC PICCA

suchen [und] Alle Wörter sortiert nach Erscheinungsjahr

WDCC Suchen

Nummer: | [Abmelden](#)

Titelliste Titeldaten

1 - 10 von 57

1. [Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data - HOAPS II - 5-days mean](#) / Karsten Fennig. - 2006-06-29

2. [Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data - HOAPS II - monthly mean](#) / Karsten Fennig. - 2006-06-29

3. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-10-30

4. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 Picntrl\(pre-industrial control experiment\): atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-06-29

5. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 1%/year CO2 increase experiment to quadrupling run no.1: atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-09-25

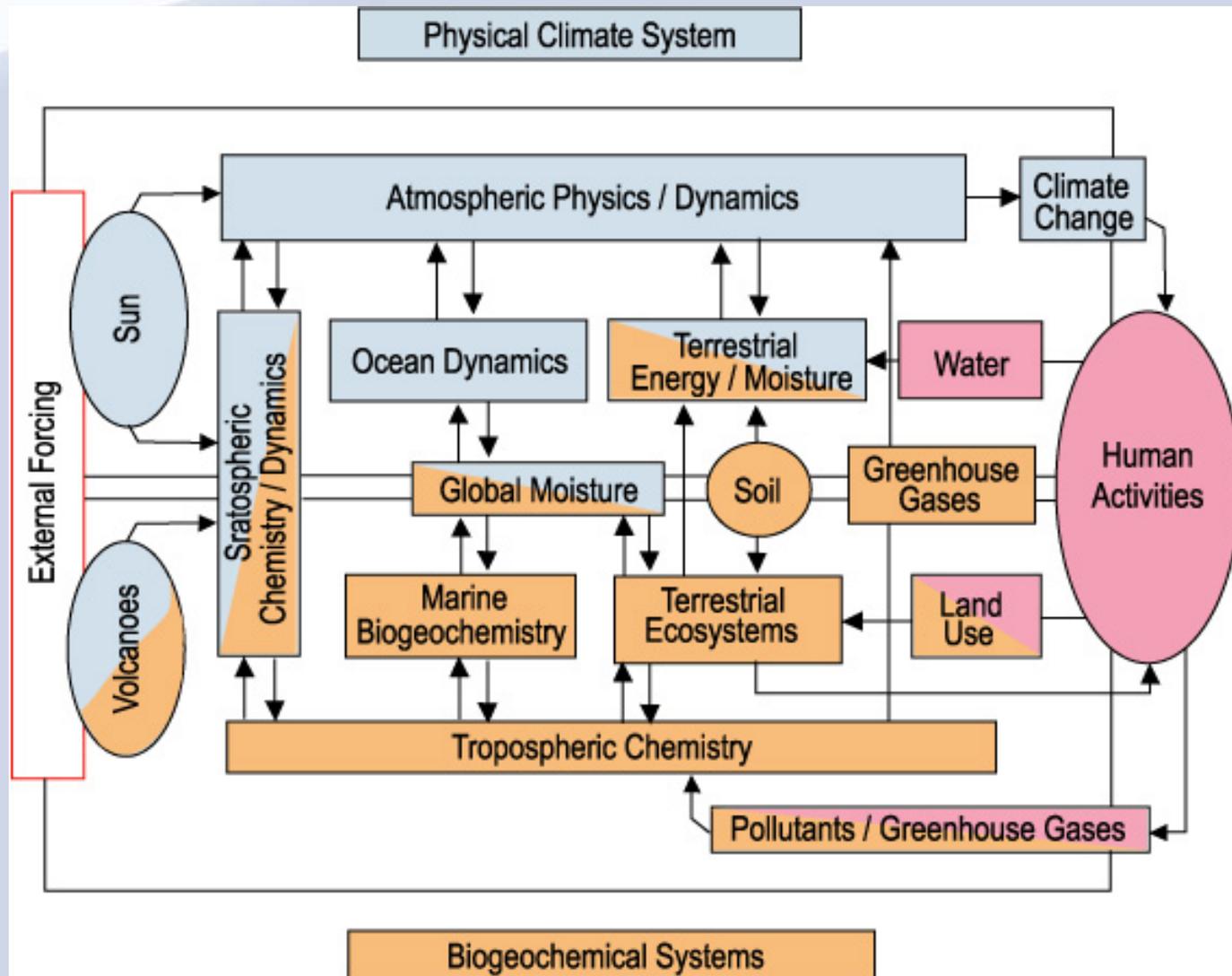
6. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 1%/year CO2 increase experiment to doubling run no.3: atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-09-25

7. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 SRESA2 run no.3: atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-09-25

8. [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 SRESA2 run no.2: atmosphere 6 HOUR values MPImet/MaD Germany](#) / Erich Roeckner. - 2006-09-25

Fertig

# Climate Model → Earth System Model



**Viele Dank  
für die Aufmerksamkeit**