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:

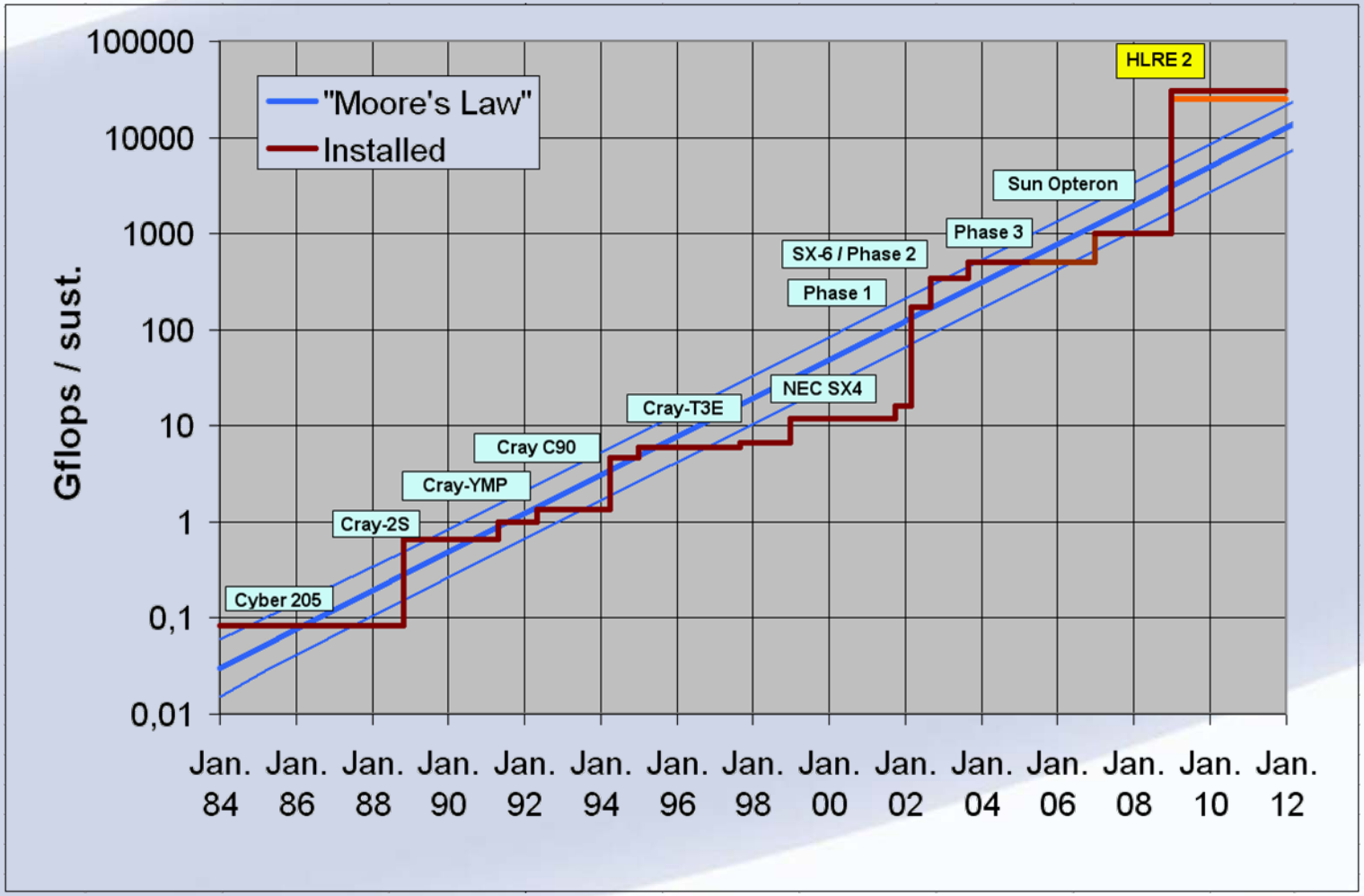
- höchste Rechenleistung,
- ausgereiftes Management größter Datenmengen und
- kompetenter Service garantieren erstklassige Klimaforschung.

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“

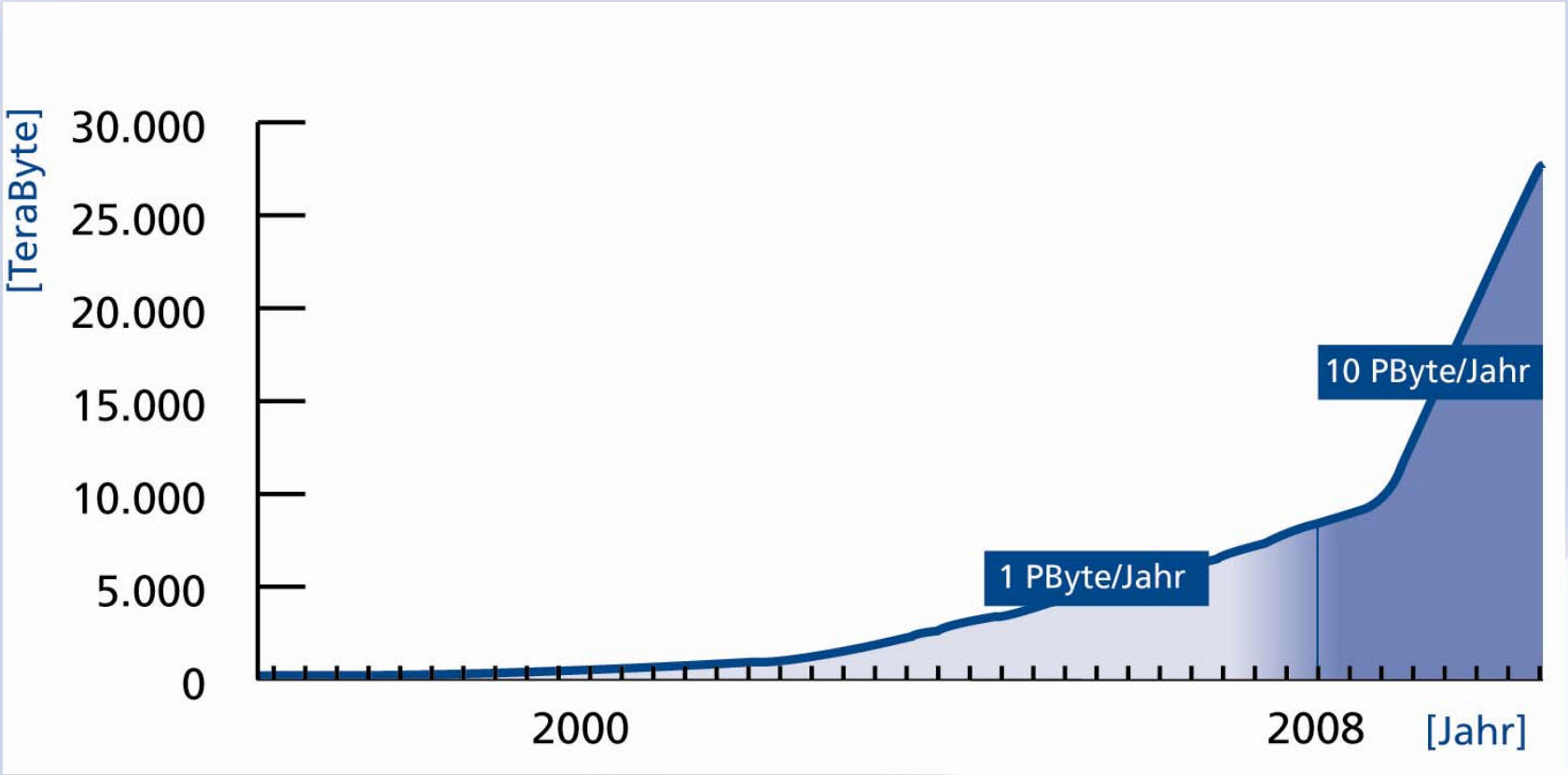
Entwicklung der Rechenleistung



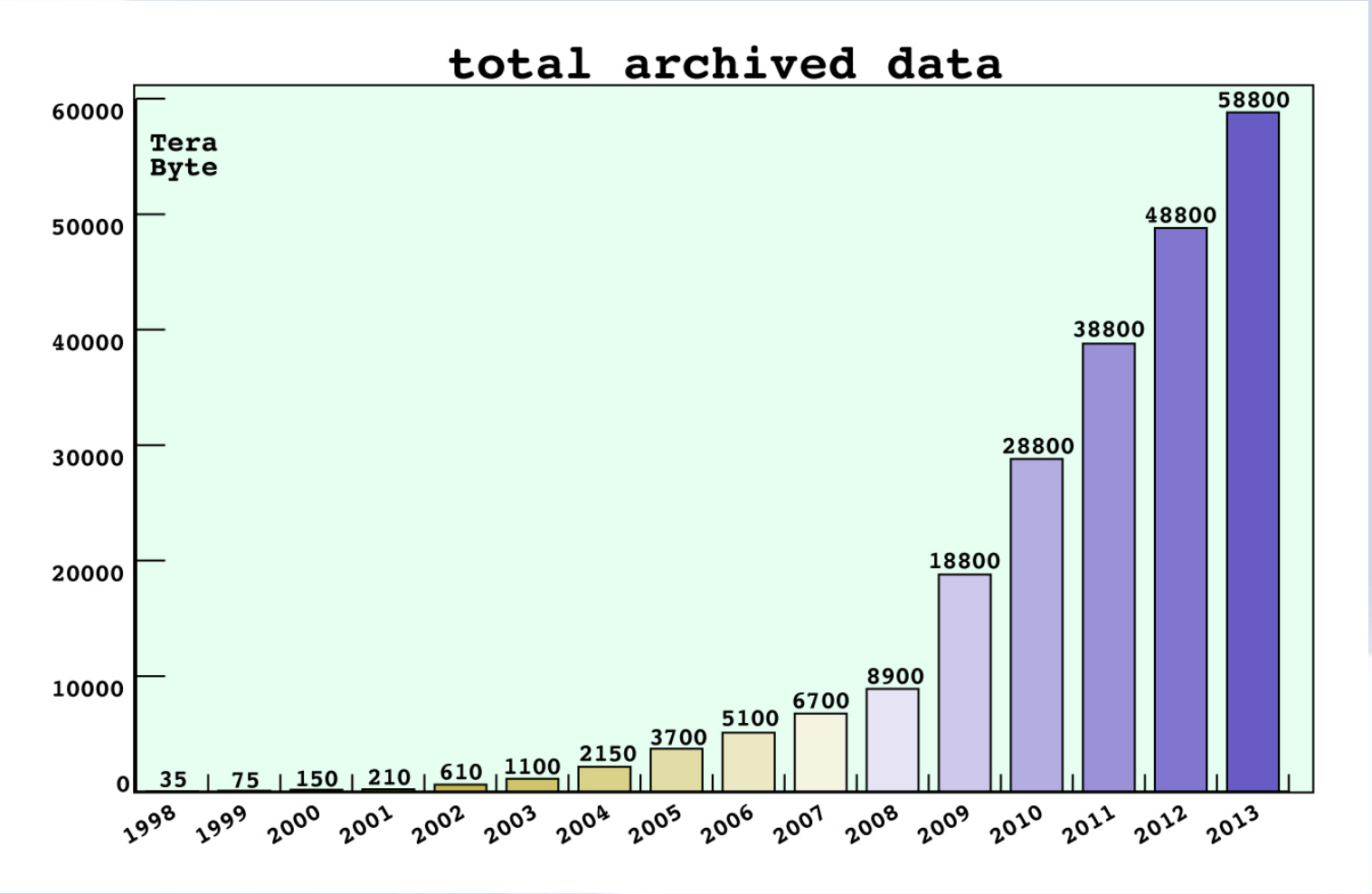
Datenspeicher

- 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

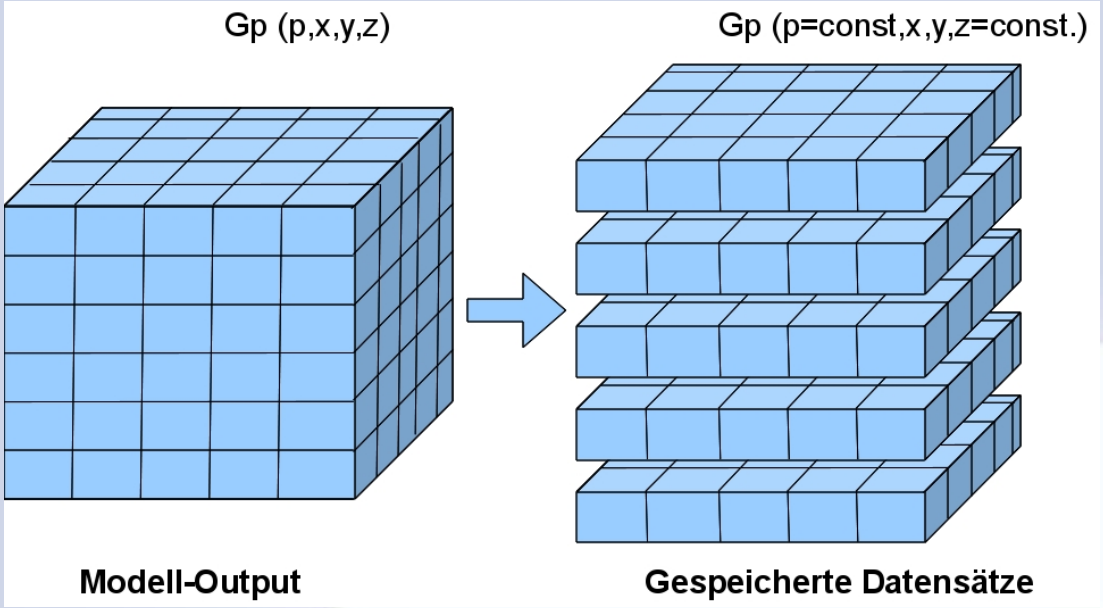
- Model like Observations

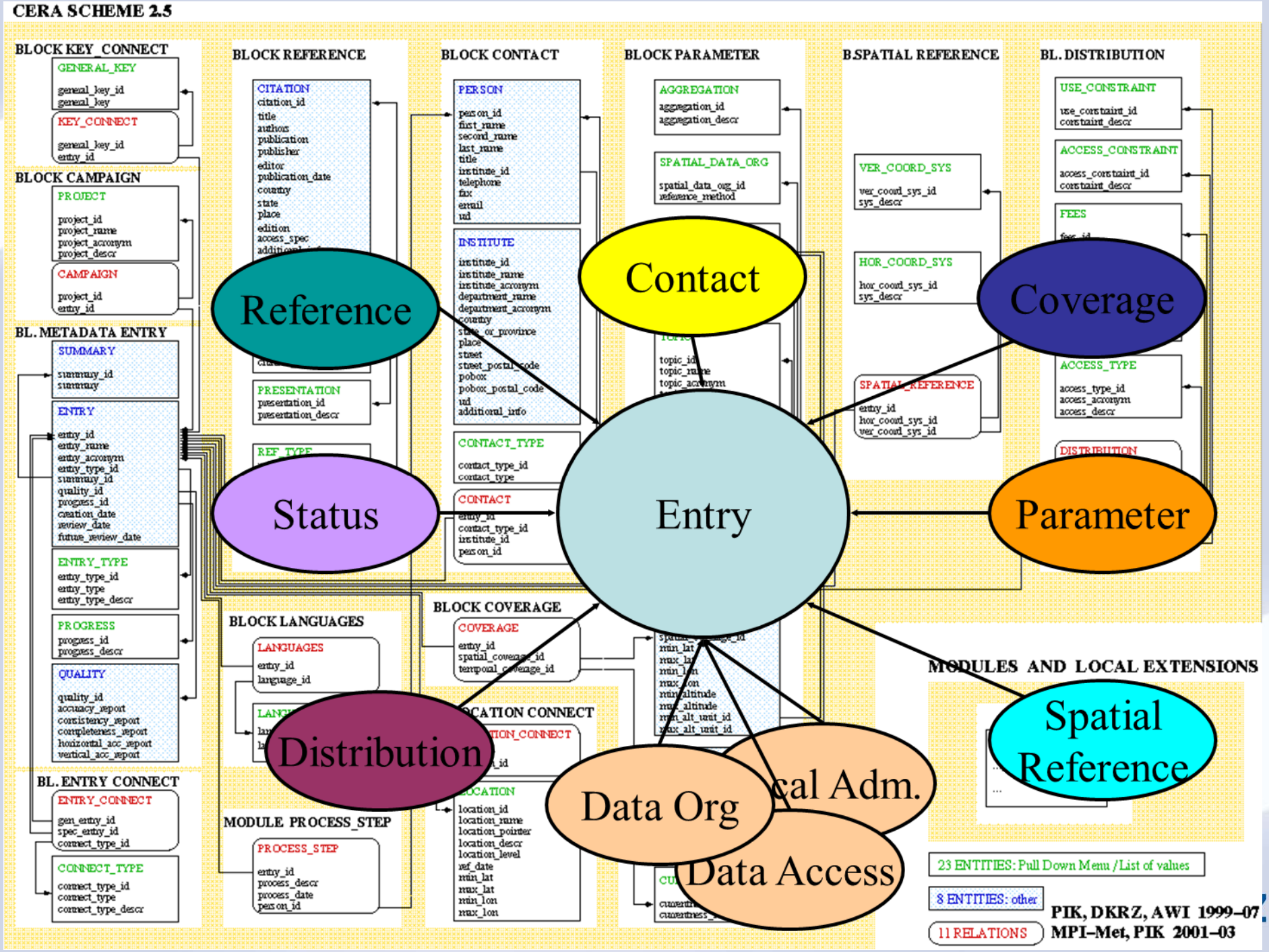
  Reanalyses data

Aufbereitung der Datensätze

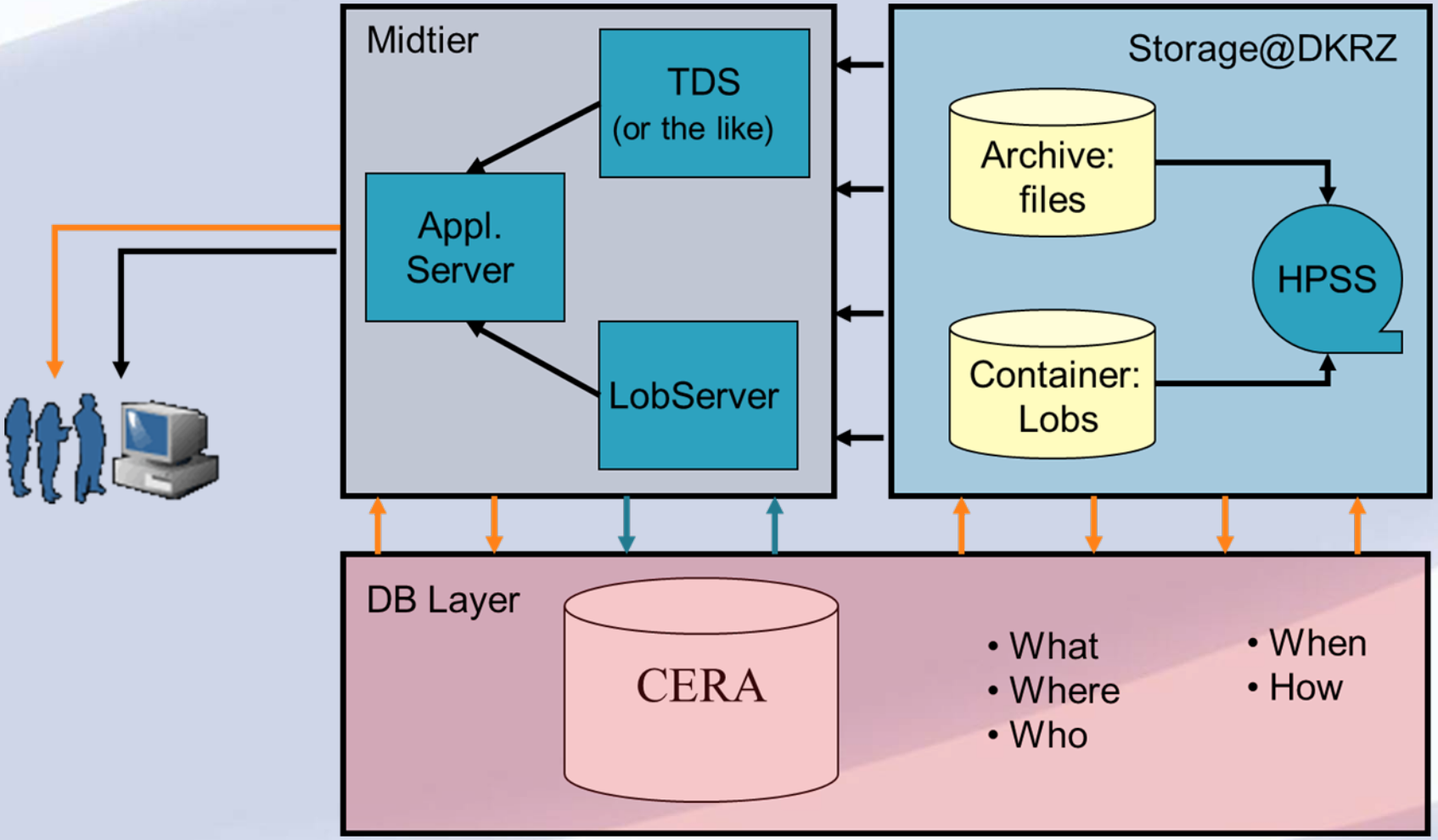
Aufbereitung der Datensätze

- Benutzerbedürfniss
- Reduzierung der herunterzuladenden Datensatzvolumen
- Verkleinerung der Transfervolumens

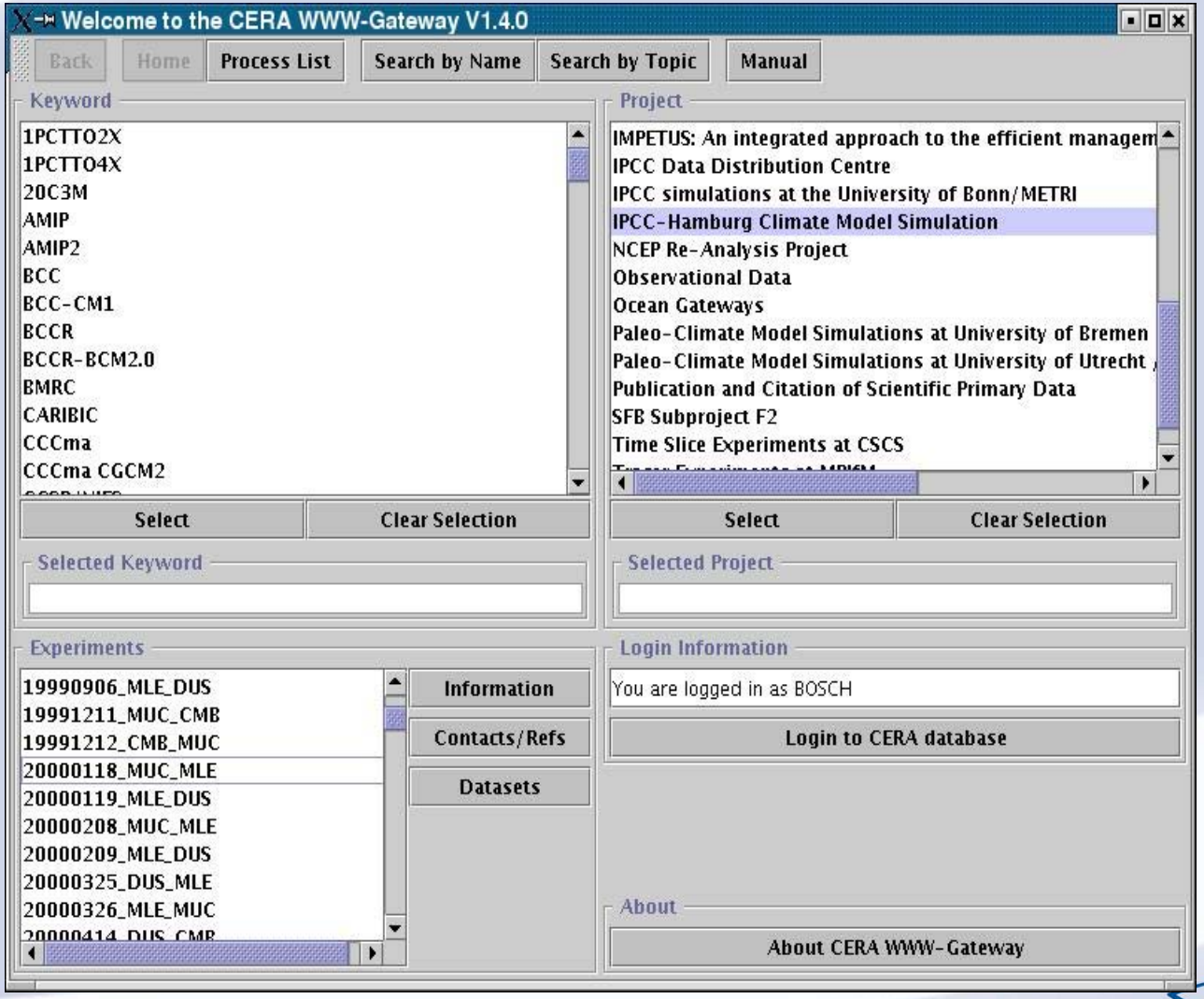



CERA SCHEME 2.5
BLOCK KEY_CONNECT
GENERAL_KEY
general_key_id
general_key
KEY_CONNECT
general_key_id
entry_id
BLOCK CAMPAIGN
PROJECT
project_id
project_name
project_acronym
project_descr
CAMPAIGN
project_id
entry_id
BL. METADATA ENTRY
SUMMARY
summary_id
summary
ENTRY
entry_id
entry_name
entry_acronym
entry_type_id
summary_id
quality_id
progress_id
creation_date
review_date
future_review_date
ENTRY_TYPE
entry_type_id
entry_type
entry_type_descr
PROGRESS
progress_id
progress_descr
QUALITY
quality_id
accuracy_report
consistency_report
completeness_report
horizontal_acc_report
vertical_acc_report
BL. ENTRY CONNECT
ENTRY_CONNECT
gen_entry_id
spec_entry_id
connect_type_id
CONNECT_TYPE
connect_type_id
connect_type
connect_type_descr
BLOCK REFERENCE
CITATION
citation_id
title
authors
publication
publisher
editor
publication_date
country
state
place
edition
access_spec
PRESENTATION
presentation_id
presentation_descr
REF_TYPE
BLOCK CONTACT
PERSON
person_id
first_name
second_name
last_name
title
institute_id
telephone
fax
email
uid
INSTITUTE
institute_id
institute_name
institute_acronym
department_name
department_acronym
country
state_or_province
place
street
street_postal_code
pobox
pobox_postal_code
uid
additional_info
CONTACT_TYPE
contact_type_id
contact_type
CONTACT
entry_id
contact_type_id
institute_id
person_id
BLOCK PARAMETER
AGGREGATION
aggregation_id
aggregation_descr
SPATIAL_DATA_ORG
spatial_data_org_id
reference_method
topic_id
topic_name
topic_acronym
B.SPATIAL REFERENCE
VER_COORD_SYS
ver_coord_sys_id
sys_descr
HOR_COORD_SYS
hor_coord_sys_id
sys_descr
SPATIAL_REFERENCE
entry_id
hor_coord_sys_id
ver_coord_sys_id
BL. DISTRIBUTION
USE_CONSTRAINT
use_constraint_id
constraint_descr
ACCESS_CONSTRAINT
access_constraint_id
constraint_descr
FEES
ACCESS_TYPE
access_type_id
access_acronym
access_descr
DISTRIBUTION
Reference
Contact
Coverage
Status
Entry
Parameter
BLOCK LANGUAGES
LANGUAGES
entry_id
language_id
BLOCK COVERAGE
COVERAGE
entry_id
spatial_coverage_id
temporal_coverage_id
min_lat
max_lat
min_lon
max_lon
min_altitude
max_altitude
min_alt_unit_id
max_alt_unit_id
MODULES AND LOCAL EXTENSIONS
Spatial Reference
Distribution
Data Org
cal Adm.
Data Access
MODULE PROCESS_STEP
PROCESS_STEP
entry_id
process_descr
process_date
person_id
LOCATION
location_id
location_name
location_pointer
location_descr
location_level
ref_date
min_lat
max_lat
min_lon
max_lon
23 ENTITIES: Pull Down Menu /List of values
8 ENTITIES: other
11 RELATIONS
PIK, DKRZ, AWI 1999–07
MPI–Met, PIK 2001–03

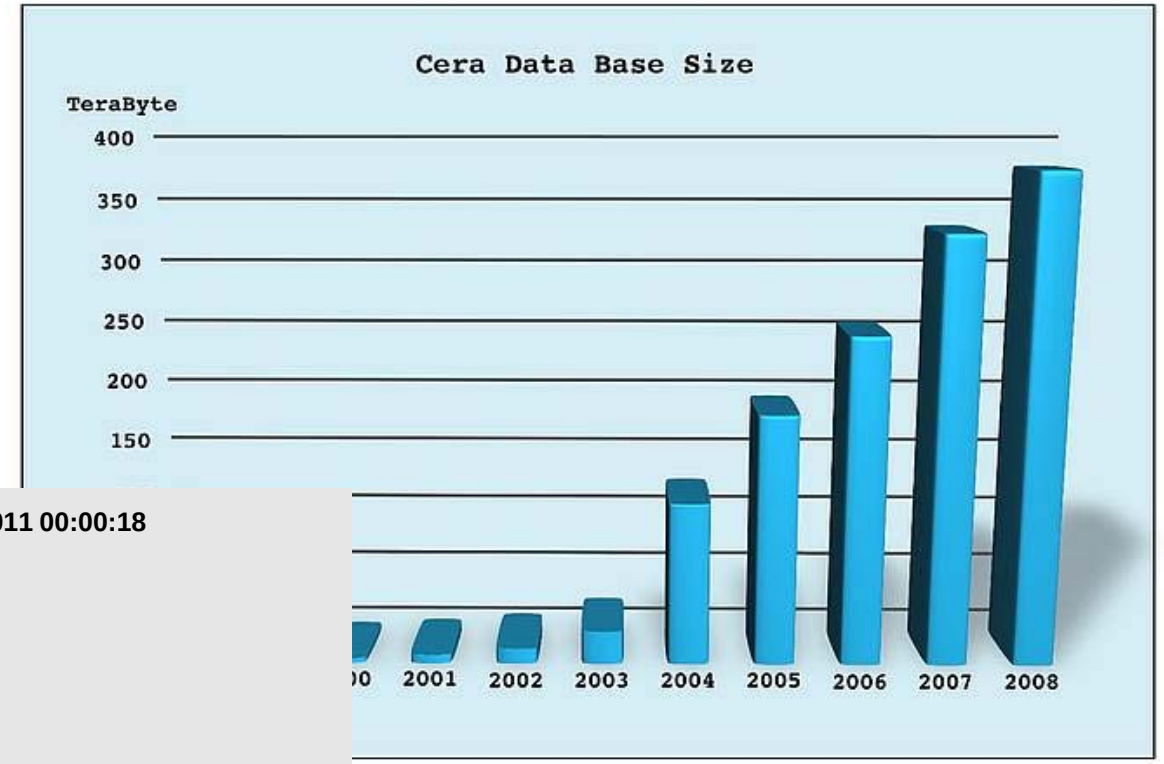
WDCC Datenzugriff



Web Portal der CERA- Datenbank : Suchen und Download

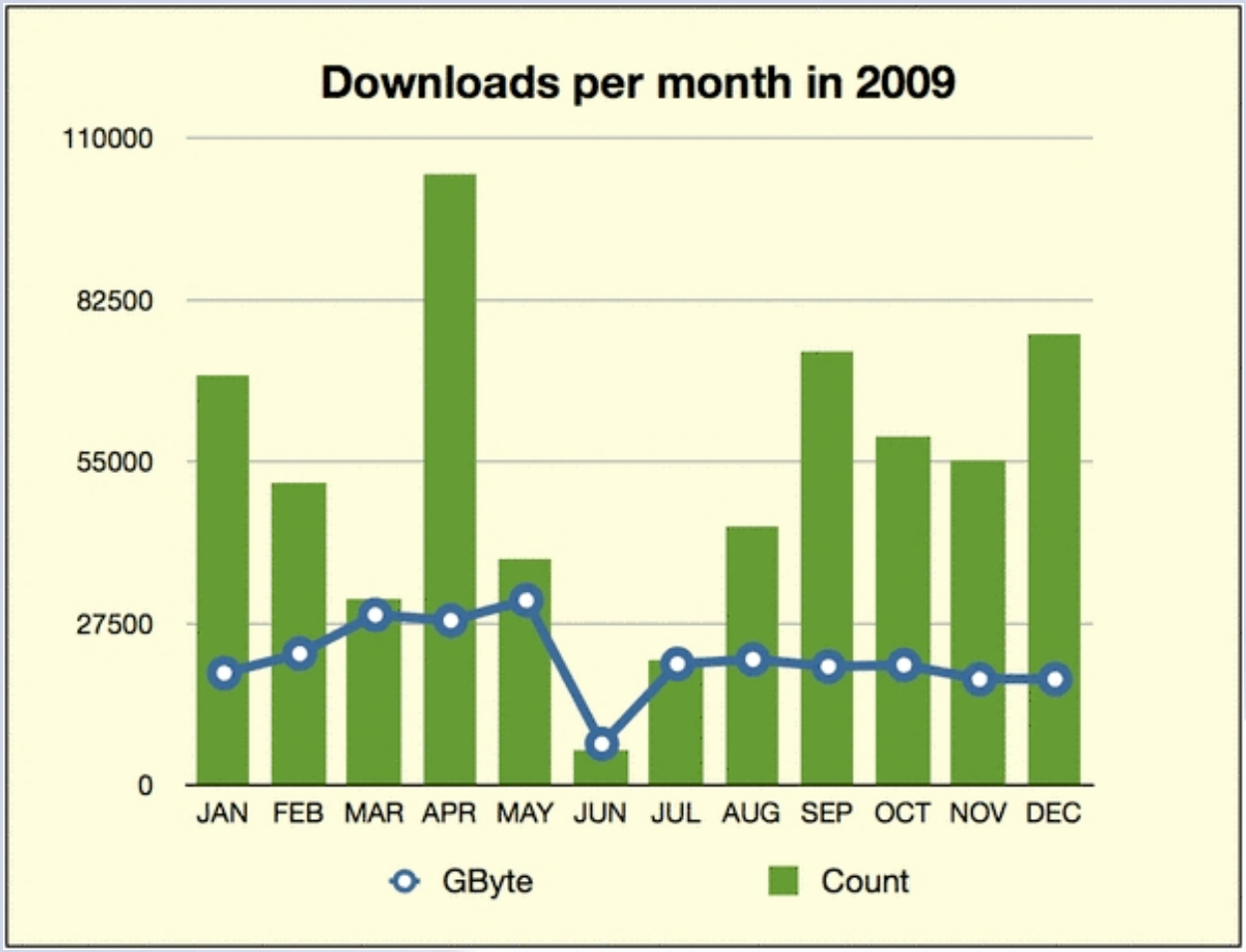


Datenspeicherung :
Semantisch-Datenhaltung

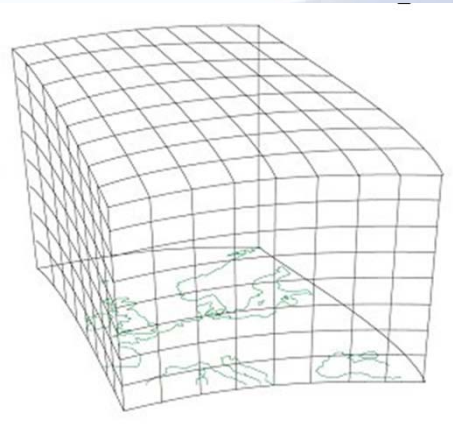


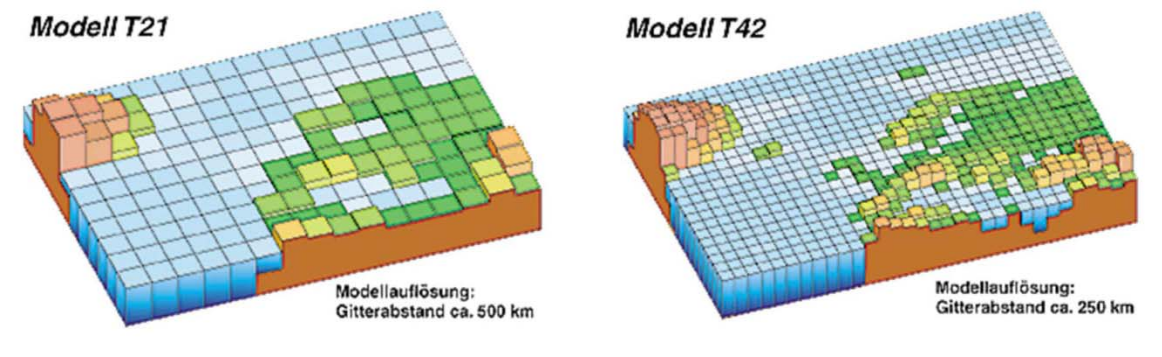
CERA: General Statistics at 01-09-2011 00:00:18
Metadata
Number of projects: 80
Number of experiments: 1439
Number of ds groups: 280
Number of datasets: 167802
Number of add_info: 237
Internal data
Database Size (TByte): 434
Number of container: 183038
Number of blobs: 8586769505

Downloads aus dem WDCC



Datenvolumen





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

- **Horizontalauflösung des Klimamodells**
  - T42: 128 * 64 = 8192 Punkte pro Globalfeld
  - T106: 160 * 320 = 51200 Punkte pro Globalfeld

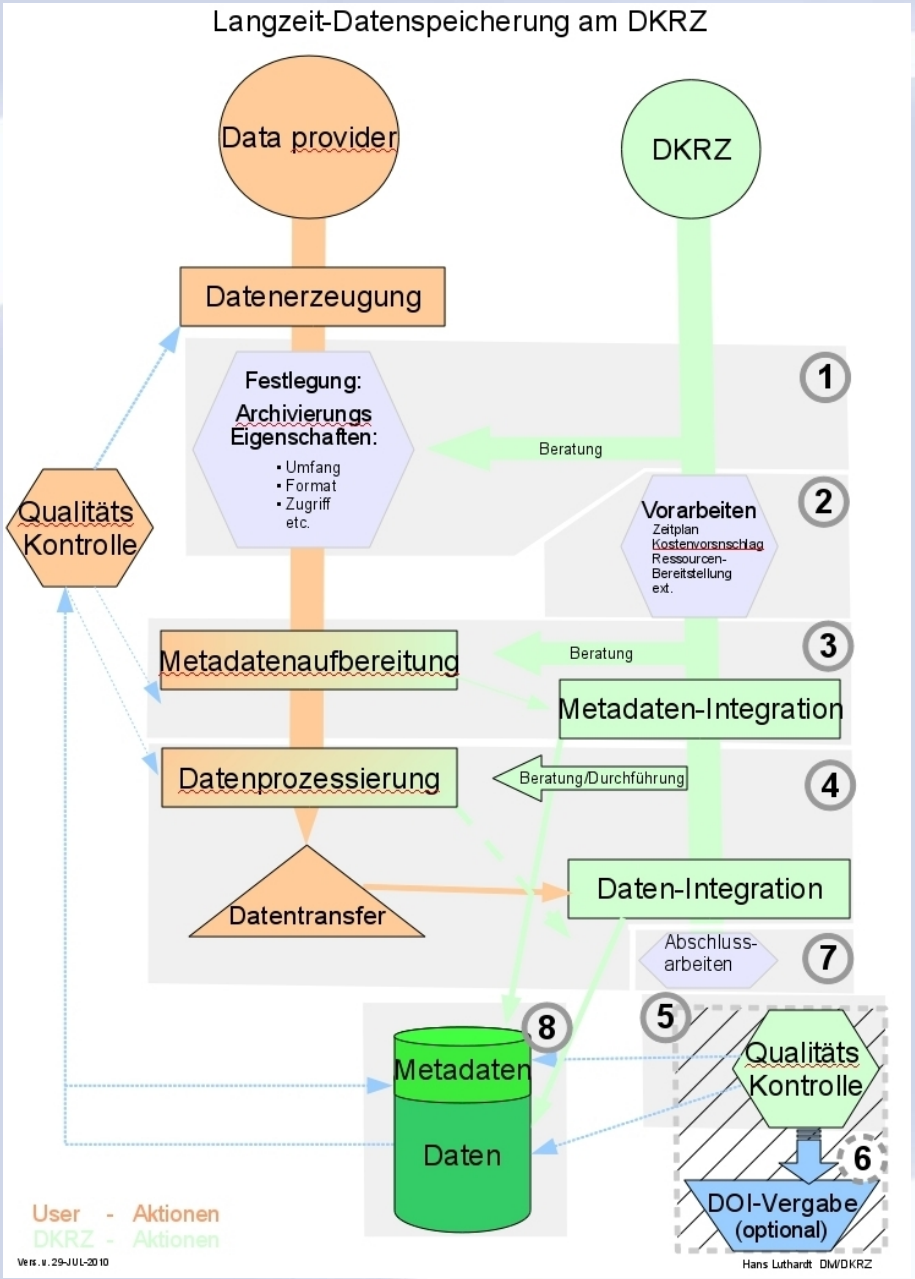
- **Erforderliche Speichereinheiten (GRIB Format)**
  - **Horizontalfeld (Zugriffseinheit):**
    - 17.1 kB (T42)
    - 100.1 kB (T106)
  - **Unix Filegröße für monatsweise akkumulierte Ergebnisse mit 6 Std. Speicherintervall und 300 2d Variablen (Physikalische Einheit):**
    - 616 MB (T42)
    - 3500 MB (T106)
  - **240 Jahre Modellintegration (Logische Einheit):**
    - 1.7 TB (T42)
    - 10 TB (T106)

Kostenfaktoren

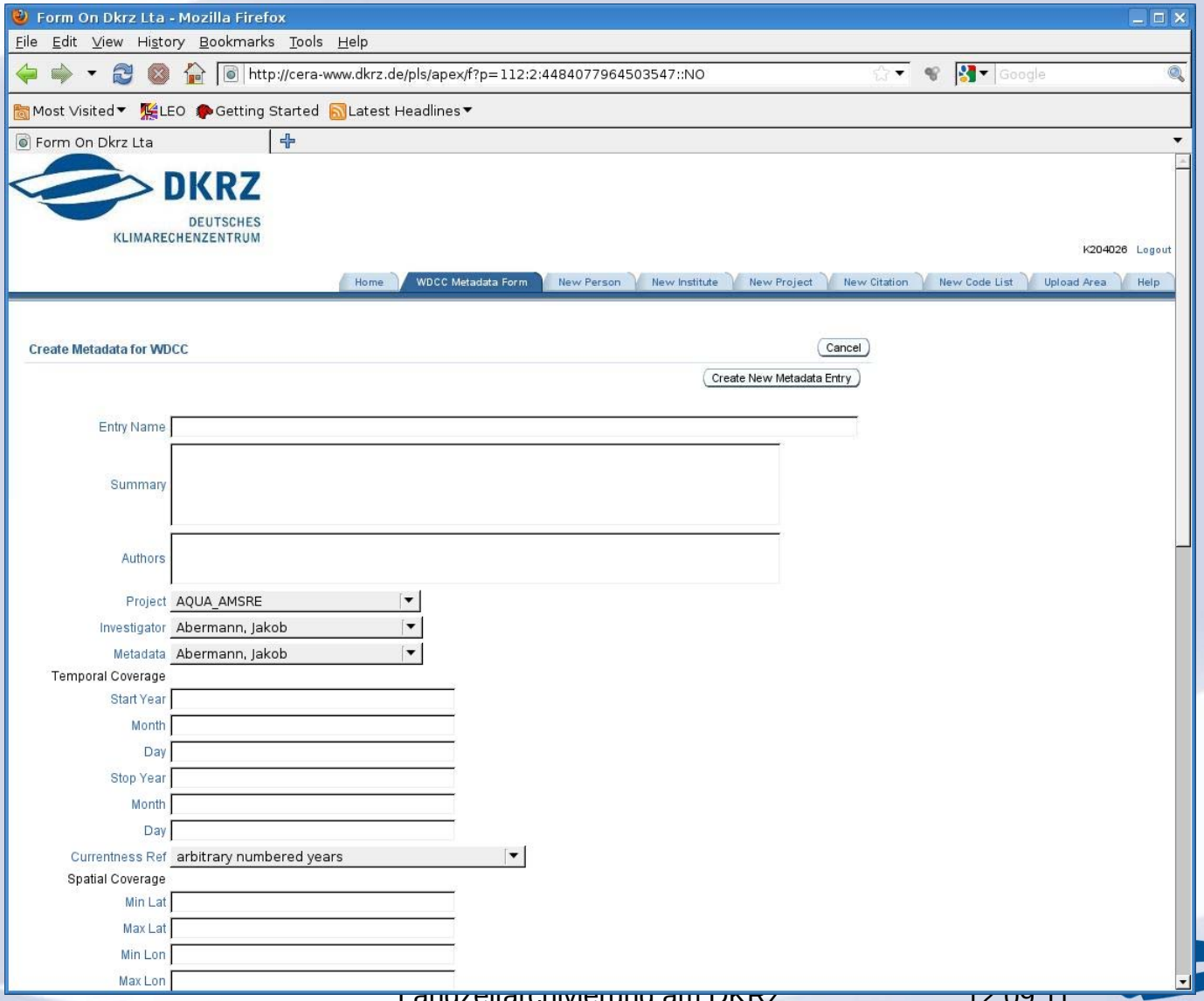
Kostenfaktoren bei der Langzeitarchivierung:

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

Workflow


Langzeit-Datenspeicherung am DKRZ
Data provider
DKRZ
Datenerzeugung
Festlegung:
Archivierungs
Eigenschaften:
• Umfang
• Format
• Zugriff
etc.
Beratung
1
Qualitäts
Kontrolle
Vorarbeiten
Zeitplan
Kostenvorsnschlag
Ressourcen-
Bereitstellung
ext.
2
Metadatenaufbereitung
Beratung
3
Metadaten-Integration
Datenprozessierung
Beratung/Durchführung
4
Datentransfer
Daten-Integration
Abschluss-
arbeiten
7
8
5
Metadaten
Daten
Qualitäts
Kontrolle
6
DOI-Vergabe
(optional)
User - Aktionen
DKRZ - Aktionen
Vers. v. 29-JUL-2010
Hans Luthardt DM/DKRZ

Erzeugungstools für Metadaten

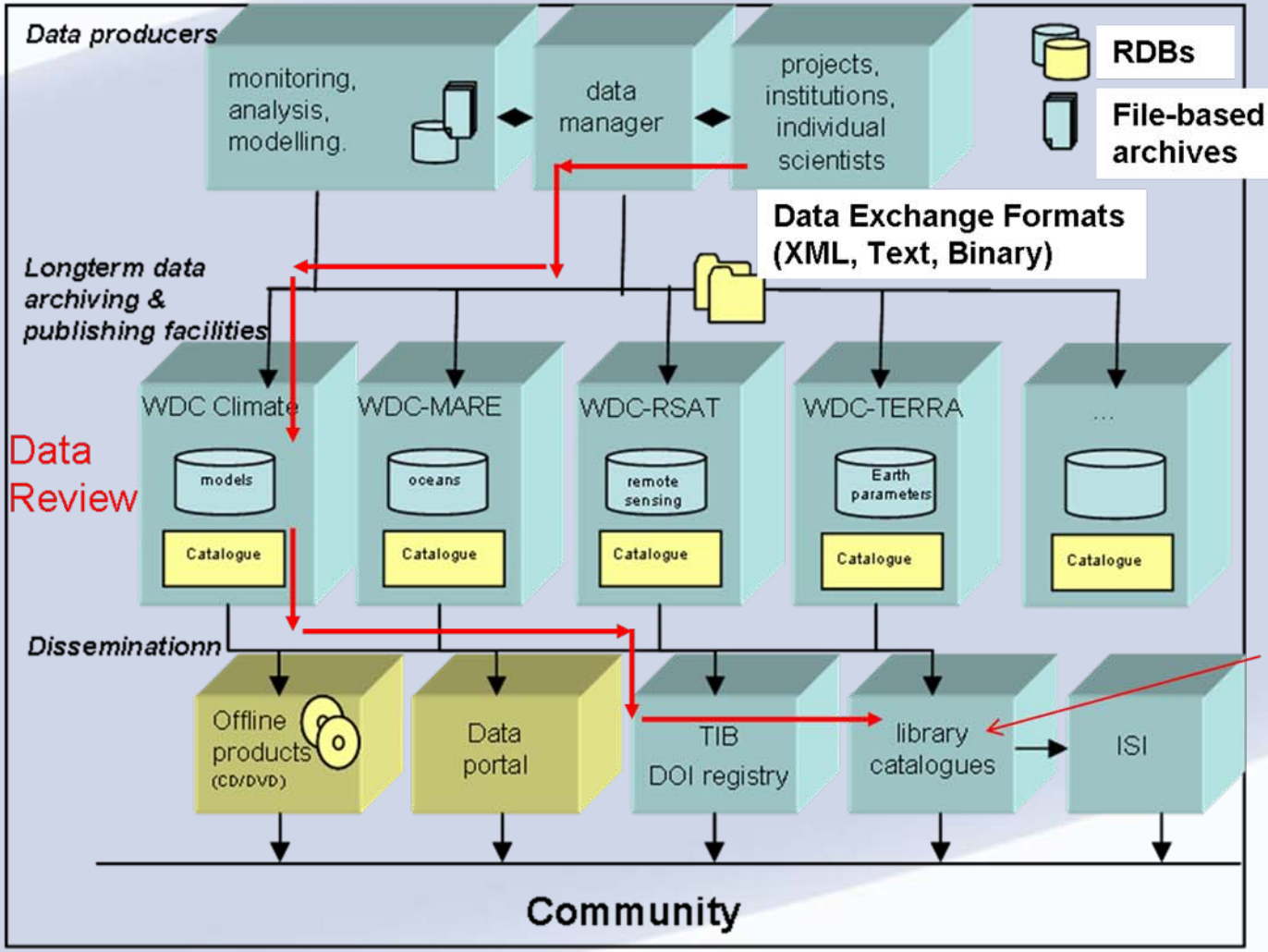


Qualitätskontrolle

- 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-Machanismen

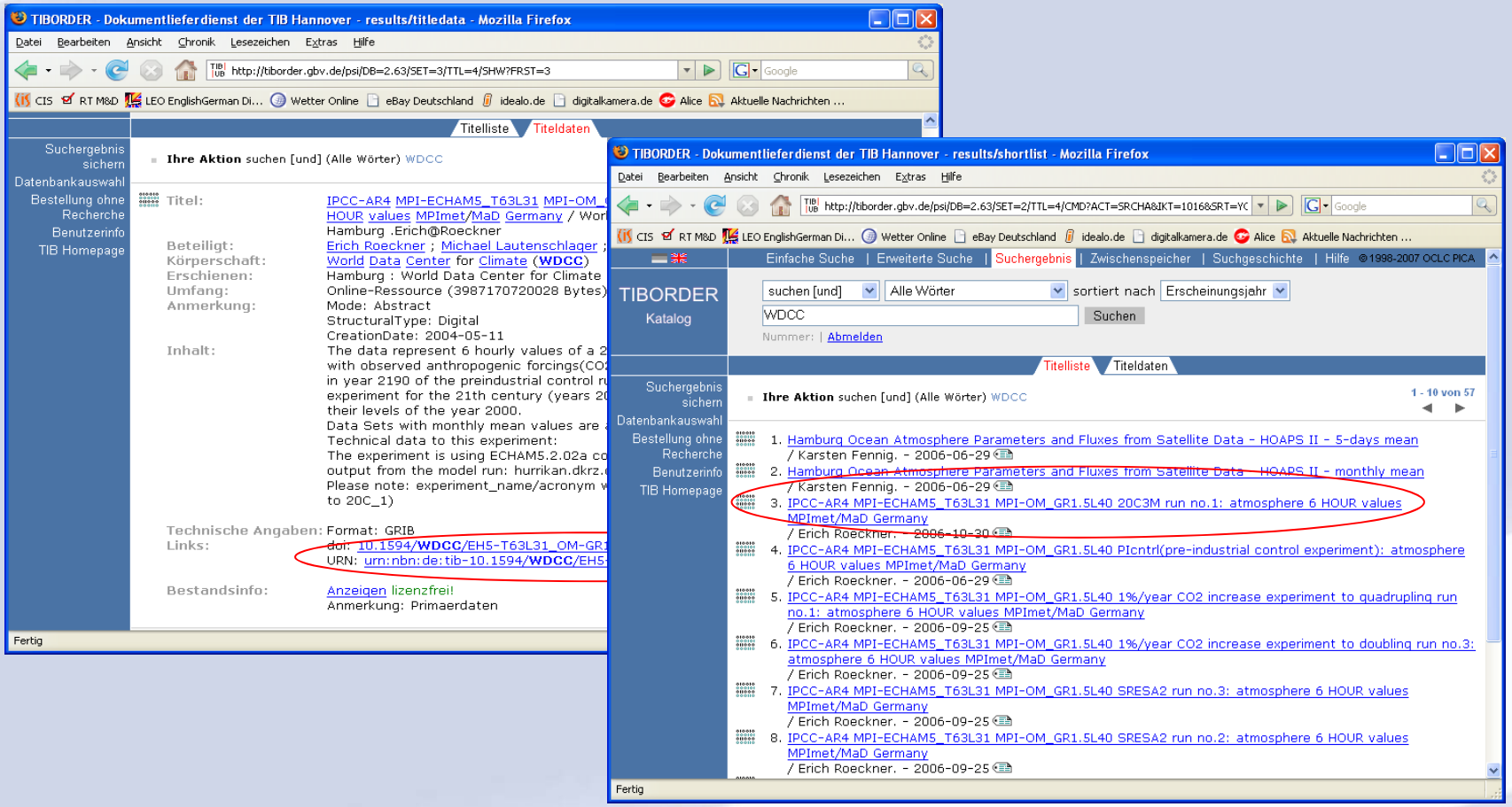
DOI - Vergabe



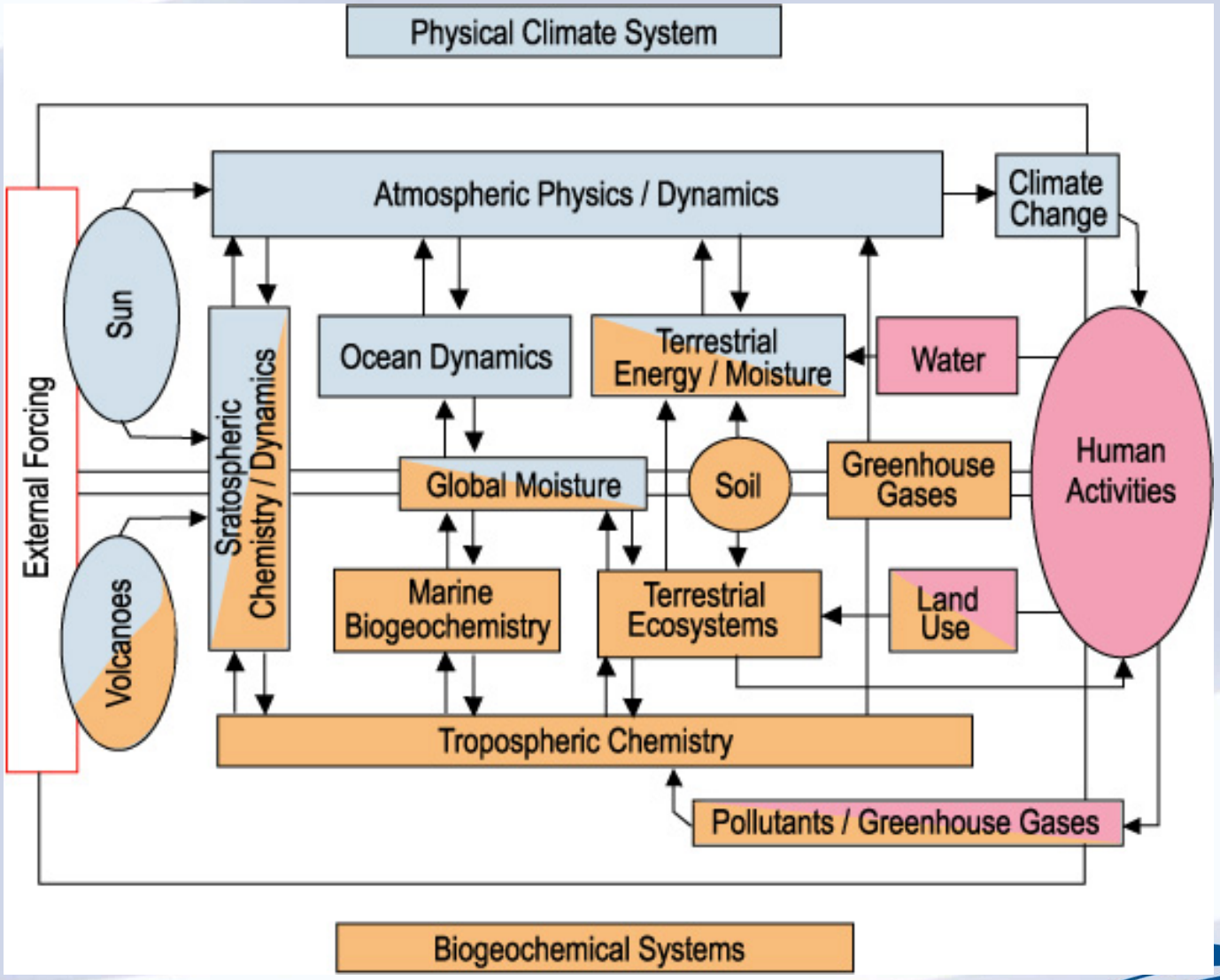
Primary Data Publication Process

ISO 690-2: Metadata for citation of electronic media

WDC-Climate hosted data in the catalogue at TIB Hannover



Climate Model → Earth System Model



Viele Dank
für die Aufmerksamkeit