



The Repository of the World Data Centre for Climate

Frank Toussaint, Michael Lautenschlager
Max-Planck-Institut für Meteorologie

Repositories in Research Institutions

MPG eScience Seminar, 25./26.6.2009 Garching





Content



- Structure and mission of WDC-Climate
- Access structure and interoperability
- Central information storage in RDBMS
- Accounting and rights management
- Availability and permanency
- Quality control
- Outlook





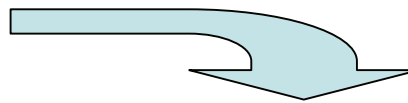
Data Services at DKRZ / WDCC



MAX-PLANCK-GESELLSCHAFT



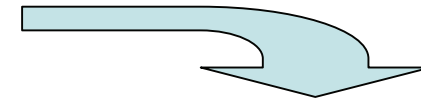
has department



Computational
support for the
German climate
science community

Earth System
modelling and
data
management

hosts

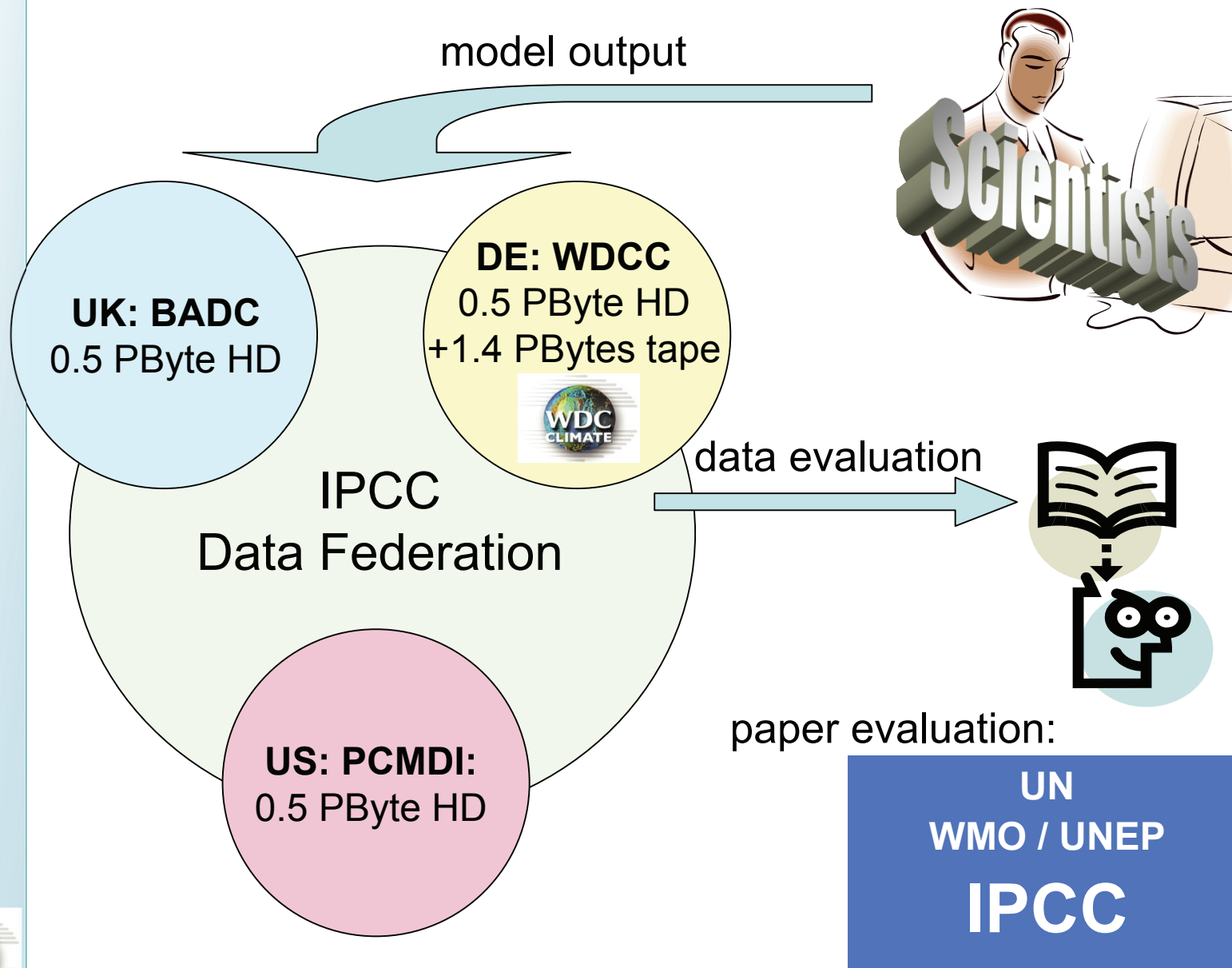


Store & disseminate
climate research data
for the scientific
community



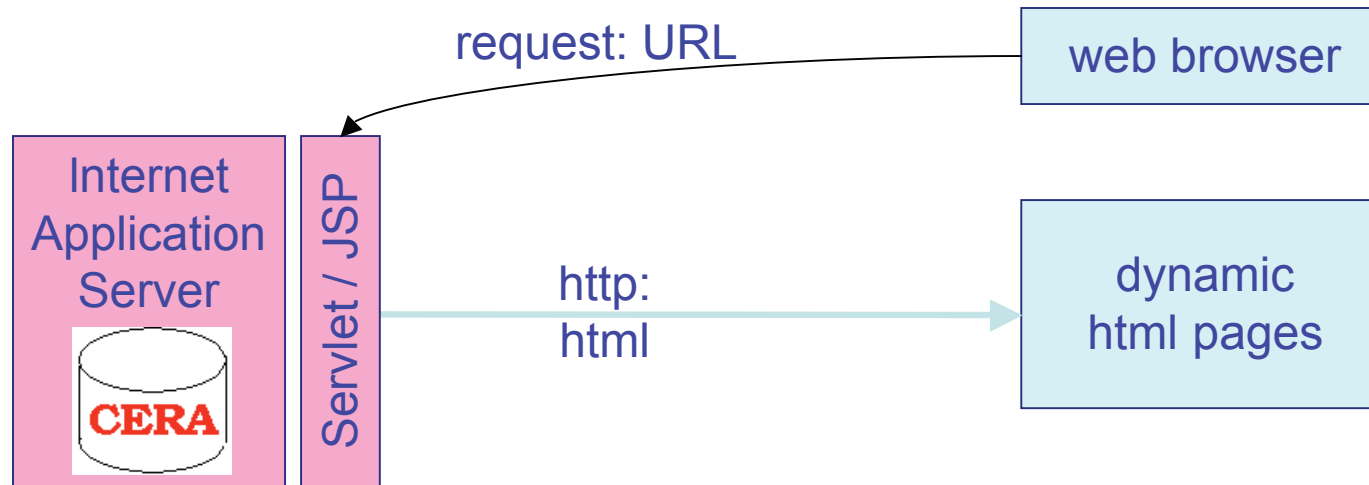


WDCC as IPCC Data Node





Catalogue Access: the GUI



Catalogue access via WWW

- URL parsed by JSP
- integrated DB retrieval by JSP
- response in standard html
- efficient administration of detailed meta information

The screenshot displays the 'World Data Center for Climate, Hamburg' website. The main content area shows 'Parameter information for dataset ERA40_PL00_6H_D1'. The page includes a search bar, navigation links, and detailed metadata for the dataset.

Variable	Value
Topic	oceans
Parameter	surface
Variable	divergence
Variable description	divergence

Code	Value
Number	155
Type	ERA40
Acronym	D
Description	divergence

Unit	Value
Acronym	1/s
Name	1/second
Description	not filled

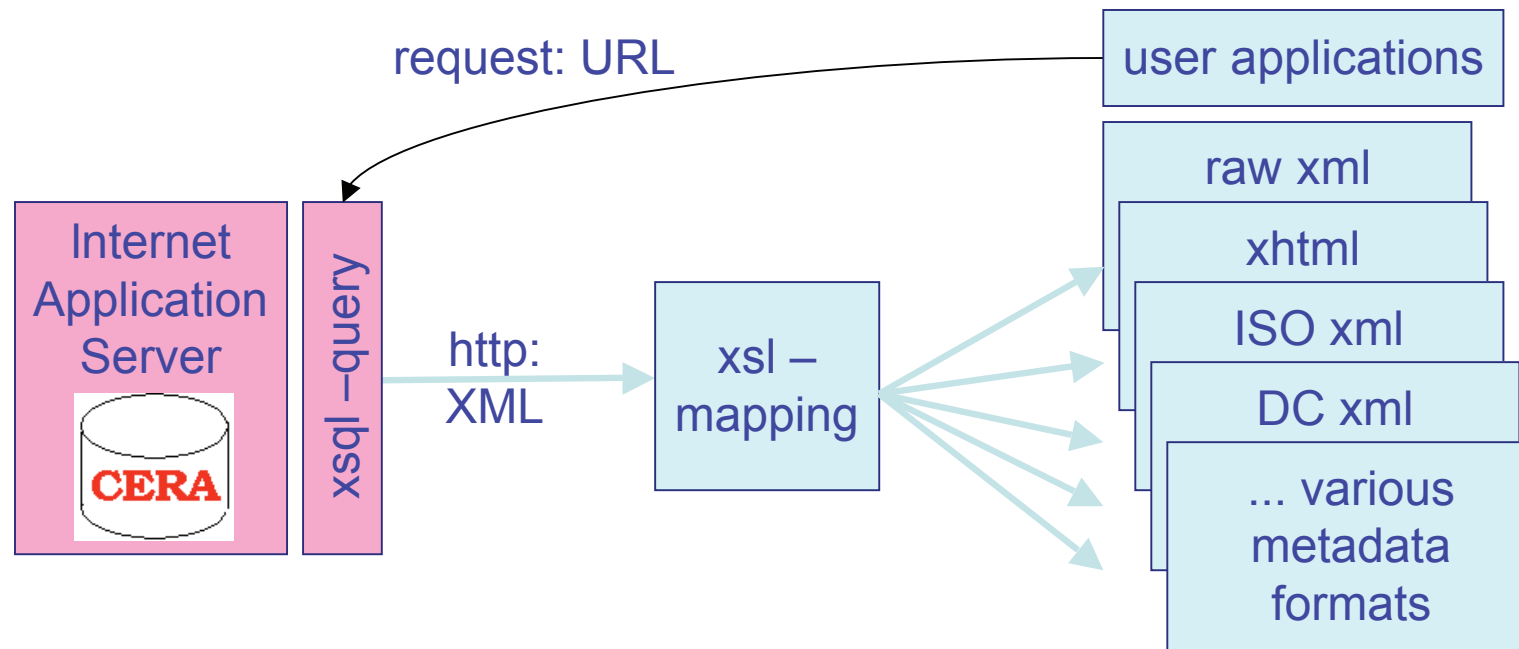
Aggregation	Value
Description	6 hours interval

Temporal structure (period 1 of 1)	Value
Start	1957
Increment	0
Date / time information	
Period start	01/09/1957 00:00
Period end	31/08/2002 18:00
Additional time information	
Dimensions	real time
Type	





http Metadata Output



Metadata access via WWW:

- xsql query to DB
- xml output from DB
- xsl mapping to any metadata format

```
- <MD_Metadata>
- <fileIdentifier>
  <CharacterString>EH4_OPYC_SRES_B2_TEMP2EH4_OPYC_SRES_B2_TEMP2:</CharacterString>
</fileIdentifier>
- <language>
  <CharacterString>en</CharacterString>
</language>
- <parentIdentifier>
  <CharacterString>ECHAM4_OPYC_SRES_B2: 110 YEARS COUPLED B2 RUN 6H VALUES
</CharacterString>
</parentIdentifier>
- <hierarchyLevel>
  <MD_ScopeCode codeList="http://mad.dkrz.de/Daten/Metadata_Fill/scope.html" codeListValue="data"
</hierarchyLevel>
- <contact>
- <CI_ResponsibleParty>
  - <organisationName>
    <CharacterString>World Data Center for Climate http://www.mad.zmaw.de/wdccc/
    </CharacterString>
  - <organisationName>
    </organisationName>
  - <role>
    <CI_RoleCode codeList="http://mad.dkrz.de/Daten/Metadata_Fill/contact_type.xsql" codeListVal
    </role>
  </CI_ResponsibleParty>
</contact>
```

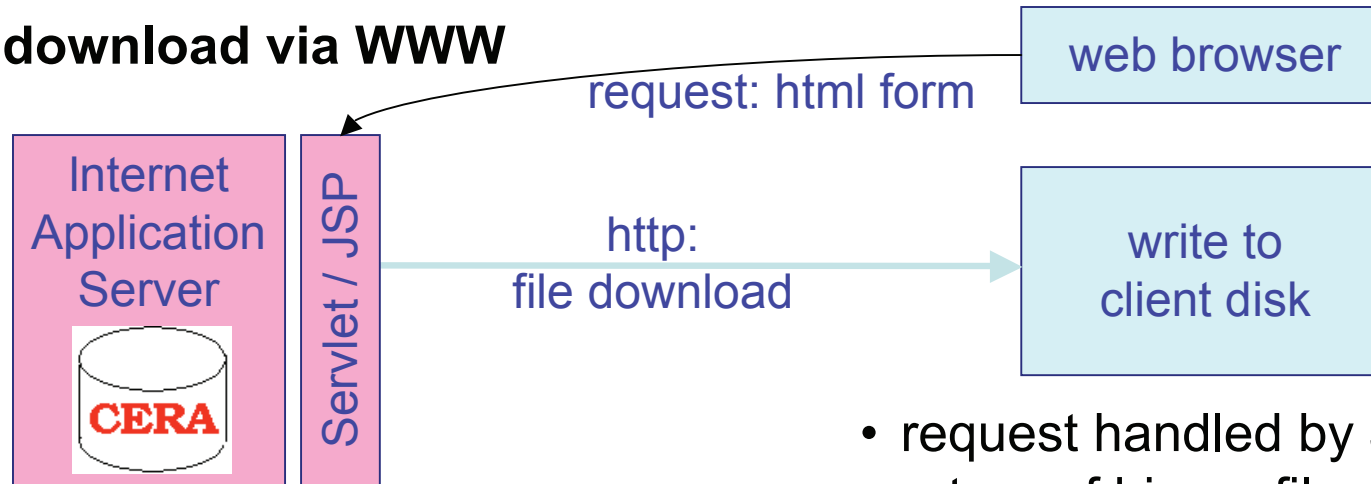




Data Download

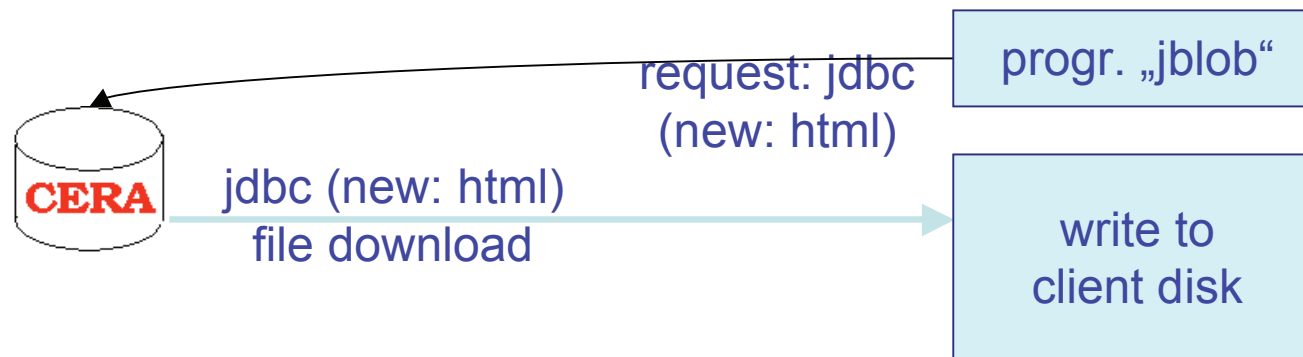


Data download via WWW



- request handled by JSP
- return of binary file

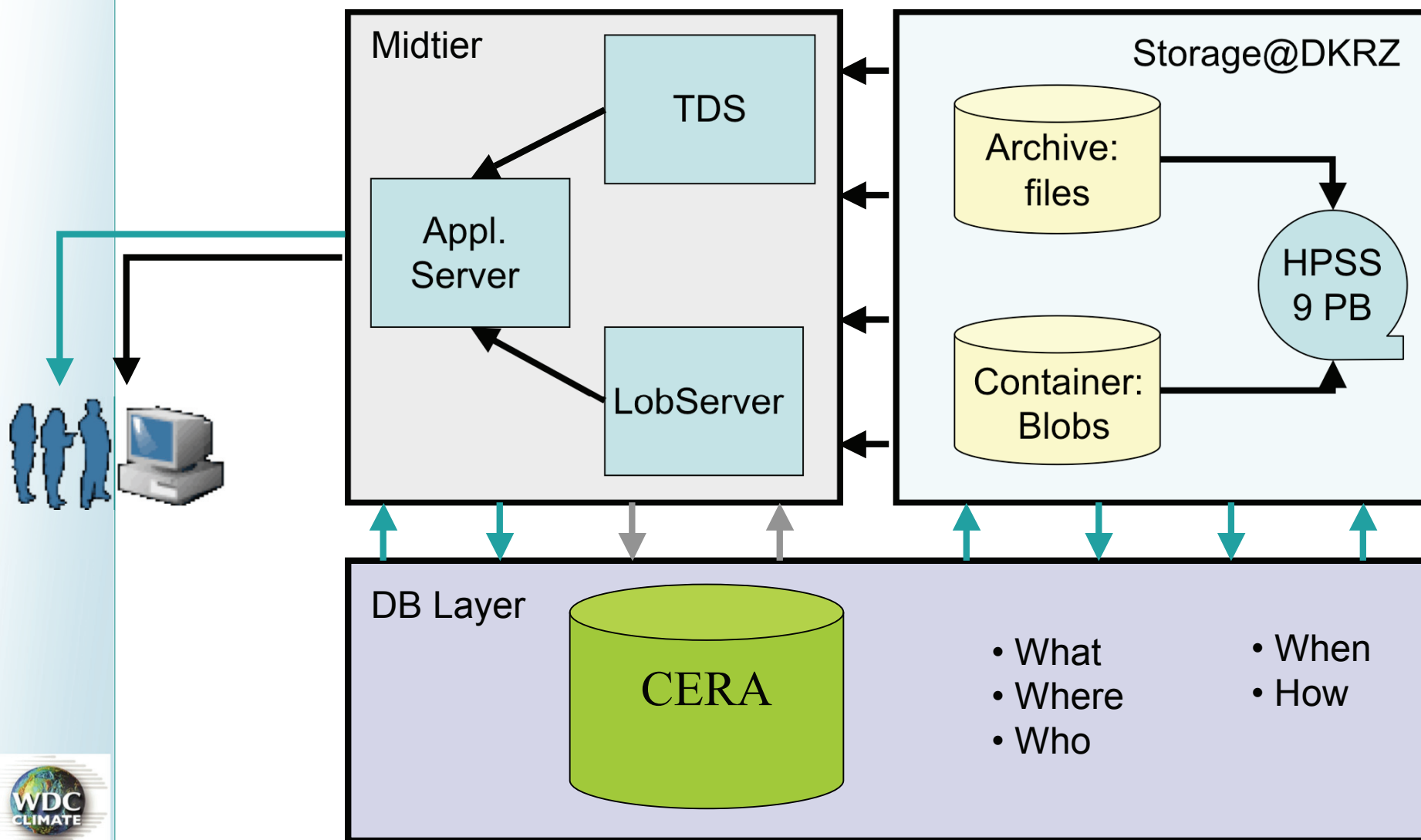
Data download via script/batch



- standard client side jdbc retrieval
- return of binary file



WDCC Data Access: Autumn 2009





Metadata interoperability

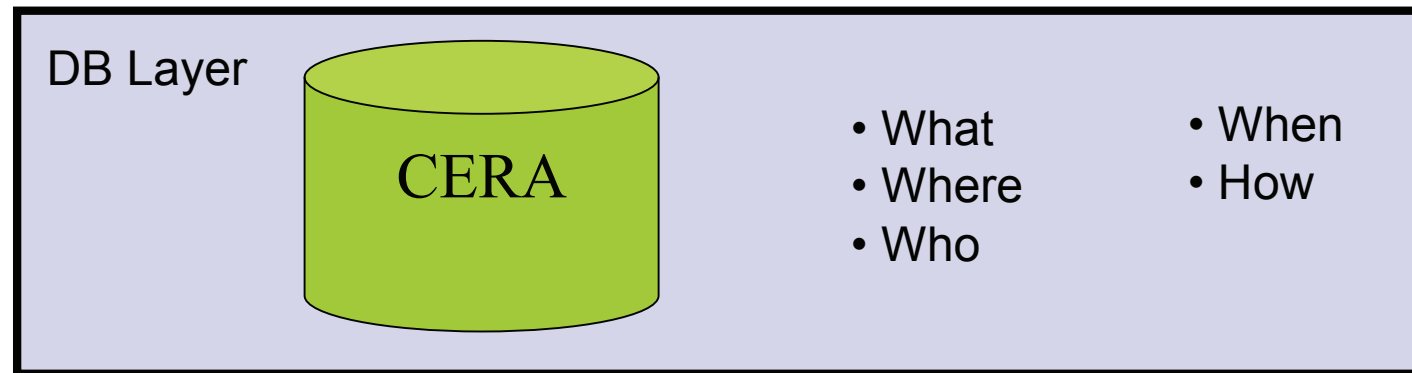
- xml – TDS: defined protocol for http download including:
 - OAI / OpenDAP
 - Web services WMS / WCS
 - harvesting
 - spatiotemporal selections

Data interoperability

- limited by data formats: WMO-Grib, netCDF
- not yet GIS formats (arcInfo)



The Relational Data Base



A priori

Efficient handling of

- authentication
 - authorization
 - detailed metadata
 - fine granularity data
- independent of
DKRZ accounts

A posteriori

- download statistics
- accounting



Handling of authentication & authorization

Authentication

- DB login by http / servlet
- DB login by direct connection to the DB (old script tool *jblob*)
- DB login by http connection to the DB (new script tool *jblob*)

Authorization

- Control of access rights (authorization) at level of finest granularity.



Efficient handling of detailed metadata

- easy and structured administration of > 60 tables
- detailed catalog information for users
- access support:
Java Server Pages (JSP), Servlets, jdbc, xsql
including standard DB features (views, ...)



Efficient handling of fine granularity data

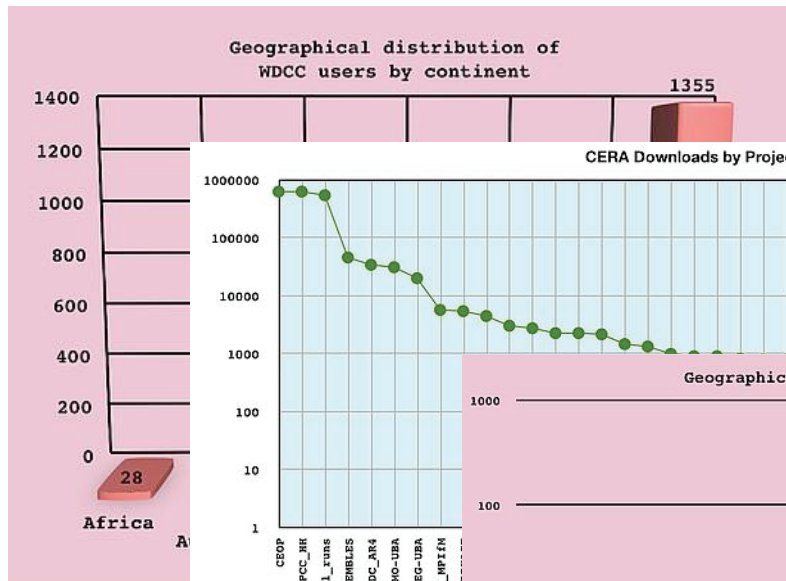
- field based data access to arbitrary time steps of single parameters
- access support:
Servlets, jdbc, web download
including standard DB features
- transparent migration of bulk data between tape & disk (nearline access)



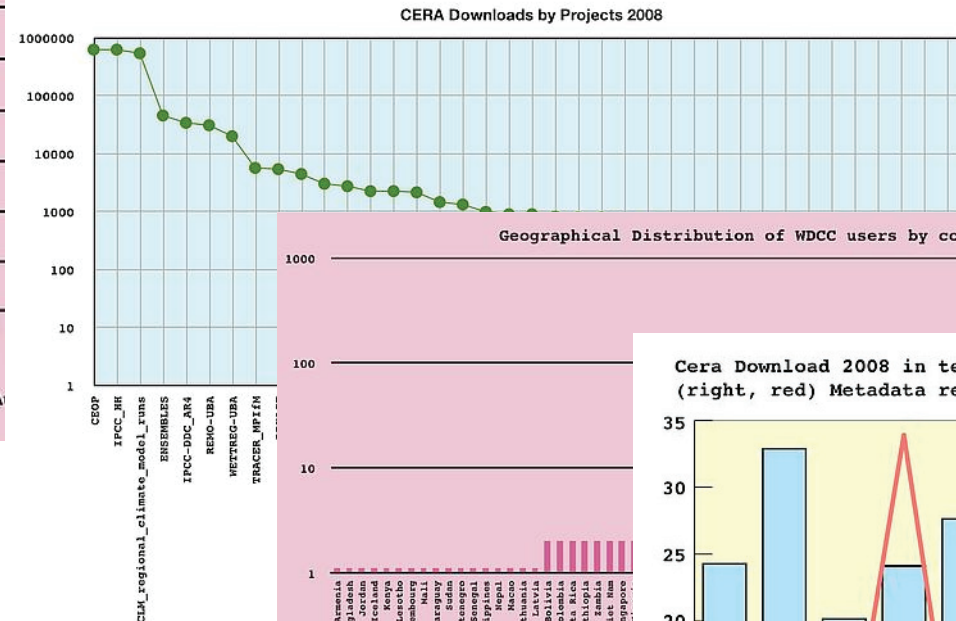
User Statistics



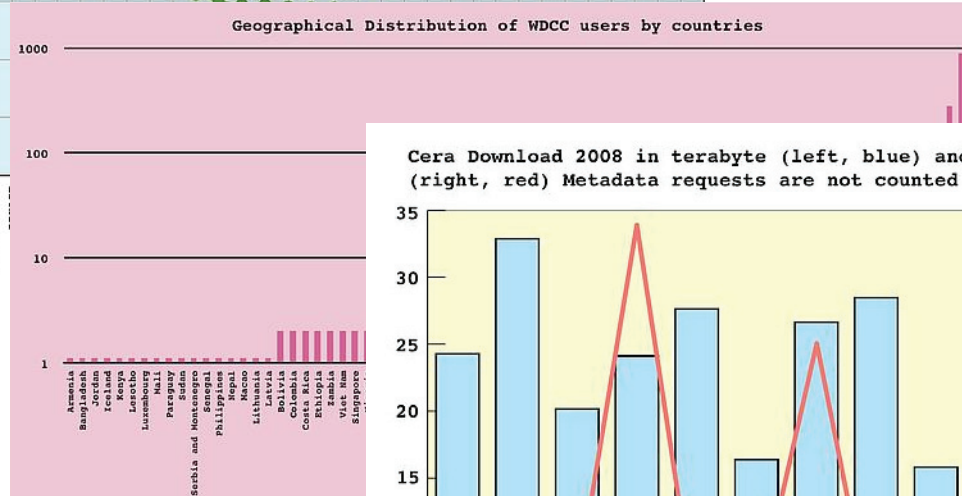
... by continent



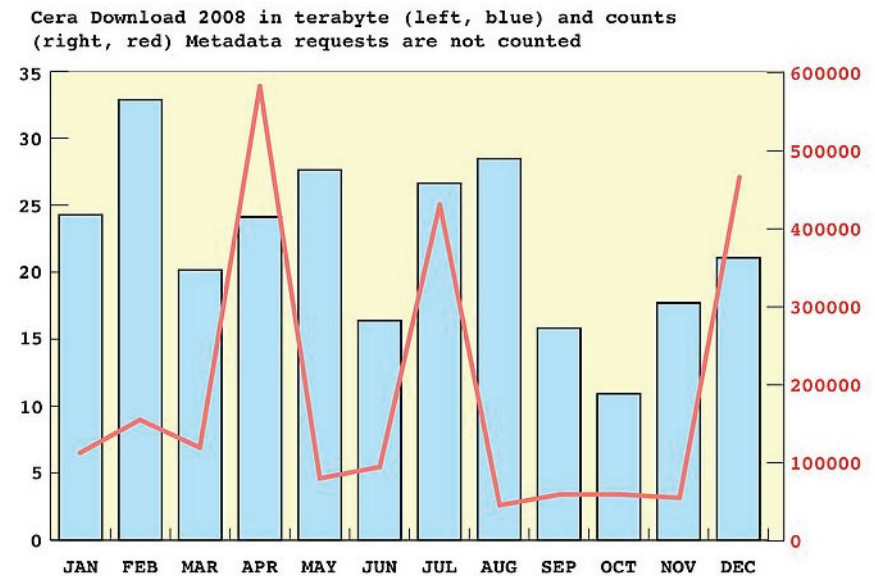
... by project



... by country



... by time





User Statistics – what for?



- When are times of low traffic / are good times for maintenance?
- Who downloaded special classified data (accounting for data providers)?
- Who downloaded which data / is to notify in case of data withdrawal?
- What data quantities have been downloaded by how many logins?



Accounting



- free personal accounts for named user
access to most of the data
- metadata visible for all accounts
- presently not basis for data charges
- for classified data: WDC-Climate
establishes the contact to the data
providers for permission





Necessity to Handle Different Rights



- all rights are project/provider dependent
- ICSU-WDC requirement: open data access
- no common EU policy (e.g., like Crown Copyright in UK and others)

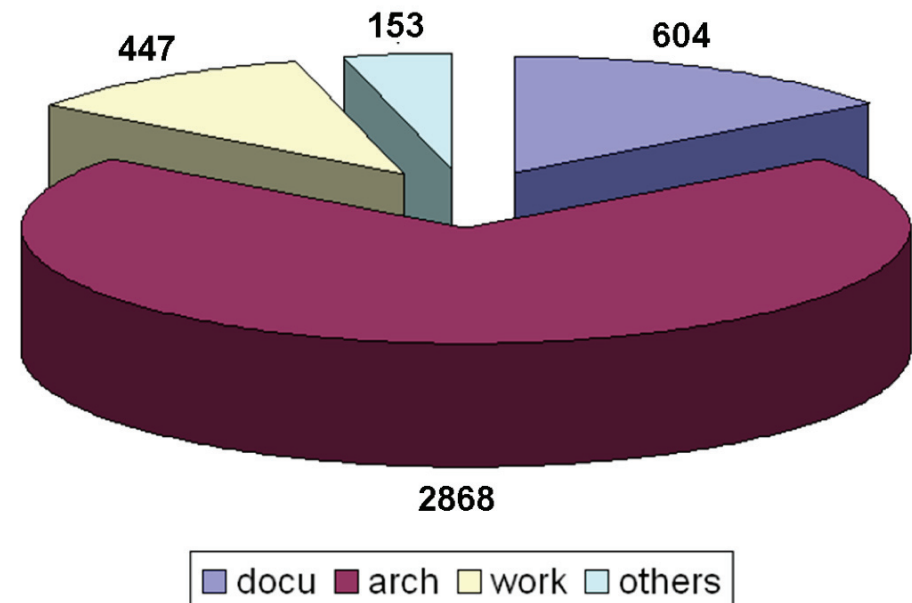


The Increasing Data Amount



Tape space distribution to archive classes at DKRZ

- part of the “work” space on tape because GFS too small
- “docu” domain consists of WDCC
- no expiration dates in “arch” domain
- parts of “arch” domain belongs to “docu” but not yet documented





Ways to Drain the Data Tsunami



A) Introduction of three data classes

1. Test data life cycle: weeks to months
2. Project data life cycle: 3 – 5 years
3. Final results life cycle: 10 years and longer

B) Introduction of four archive classes

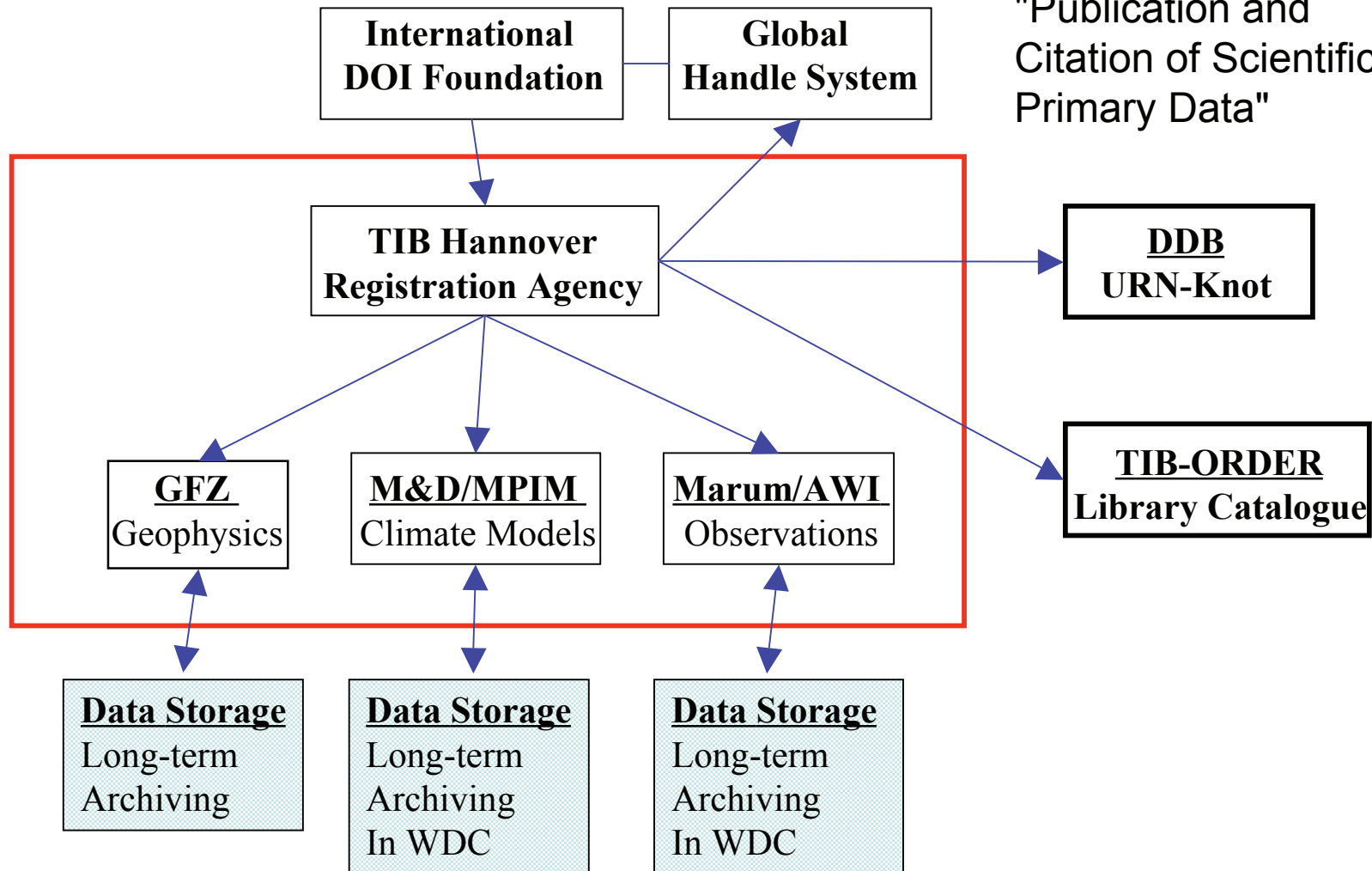
1. Temp(orary) scratch discs at compute server
2. Work fixed disc space at project level
3. Arch(ive) tape, single copy, expiring
4. Docu(mentation) tape, security copy, long-term
documented fixed data



Digital Object Identifiers (DOI)



DFG Project
"Publication and
Citation of Scientific
Primary Data"





Connections to Central Catalogues



WDC-Climate hosted data in the catalogue at TIB Hannover

The screenshot displays two overlapping browser windows from the TIBORDER web interface. The left window shows a detailed record for a specific data set, while the right window shows a list of search results for the same query.

Left Window: TIBORDER - Dokumentlieferdienst der TIB Hannover - results/titledata - Mozilla Firefox

URL: <http://tiborder.gbv.de/psi/DB=2.63/SET=3/TTL=4/SHW?FRST=3>

Suchergebnis sichern
Datenbankauswahl
Bestellung ohne Recherche
Benutzerinfo
TIB Homepage

Titelliste | **Titeldaten**

■ Ihre Aktion suchen [und] (Alle Wörter) WDCC

Titel: [IPCC-AR4 MPI-ECHAM5_T63L31 MPI-OM_6HOUR values MPImet/MaD Germany](#) / World Data Center for Climate (WDCC) / Hamburg .Erich Roeckner ; Michael Lautenschlager ;

Beteiligt: [World Data Center for Climate \(WDCC\)](#)

Erschienen: Hamburg : World Data Center for Climate

Umfang: Online-Ressource (3987170720028 Bytes)

Anmerkung: Mode: Abstract

Inhalt: StructuralType: Digital
CreationDate: 2004-05-11
The data represent 6 hourly values of a 2190 of the preindustrial control experiment for the 21th century (years 2000-2100). Data Sets with monthly mean values are available. Technical data to this experiment: The experiment is using ECHAM5.2.02a code output from the model run: hurikan.dkrz.de. Please note: experiment_name/acronym w to 20C_1)

Technische Angaben: Format: GRIB

Links: [doi: 10.1594/WDCC/EH5-T63L31_OM-GR1](#)
[URN: urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1](#)

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

Fertig

Right Window: TIBORDER - Dokumentlieferdienst der TIB Hannover - results/shortlist - Mozilla Firefox

URL: <http://tiborder.gbv.de/psi/DB=2.63/SET=2/TTL=4/CMD?ACT=SRCHA&IKT=1016&SRT=YC>

Einfache Suche | Erweiterte Suche | **Suchergebnis** | Zwischenspeicher | Suchgeschichte | Hilfe

suchen [und] | Alle Wörter | sortiert nach Erscheinungsjahr

WDCC | Suchen

Nummer: | Abmelden

Titelliste | **Titeldaten**

■ Ihre Aktion suchen [und] (Alle Wörter) WDCC 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 P1cntrl\(pre-industrial control experiment\): 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



Data Quality Control at WDC-Climate



Quality control of metadata

- some general topics by WDC-Climate
- more detailed by client users via GUI (*GeoNetwork*)
- final QC in the STD-DOI data publication process

Quality control of data in cooperation of data providers and WDC-Climate

- for internally generated data (climate models) by WDC-Climate (automated process)
- final QC in the STD-DOI data publication process
- data from scientific projects are in the responsibility of the data providers





Main Challenges - Metadata



Increasing interdisciplinarity will lead to
increasing semantic problems in the metadata
and in their presentation
→ ontologies needed

Interdisciplinary catalogue access
→ more central catalogues
& data federations needed

more standardisation projects
→ Inhomogeneity within one discipline
seems to be decreasing





Main Challenges - Data

The Earth System Sciences viewpoint

Data quantities will keep growing –
→ huge storage media with application
adapted data structures (selected data access)

Grow of data transport means seems to
follow games and video –
→ better access performance

Data inhomogeneities will –
grow with growing interdisciplinarity
shrink with growing number of standards
→ flexible access tools





Thank you !

Questions?

