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

GKSS FORSCHUNGSZENTRUM in der HELMHOLTZ-GEMEINSCHAFT

DKRZ

has department

M&D Modelle & Daten

hosts

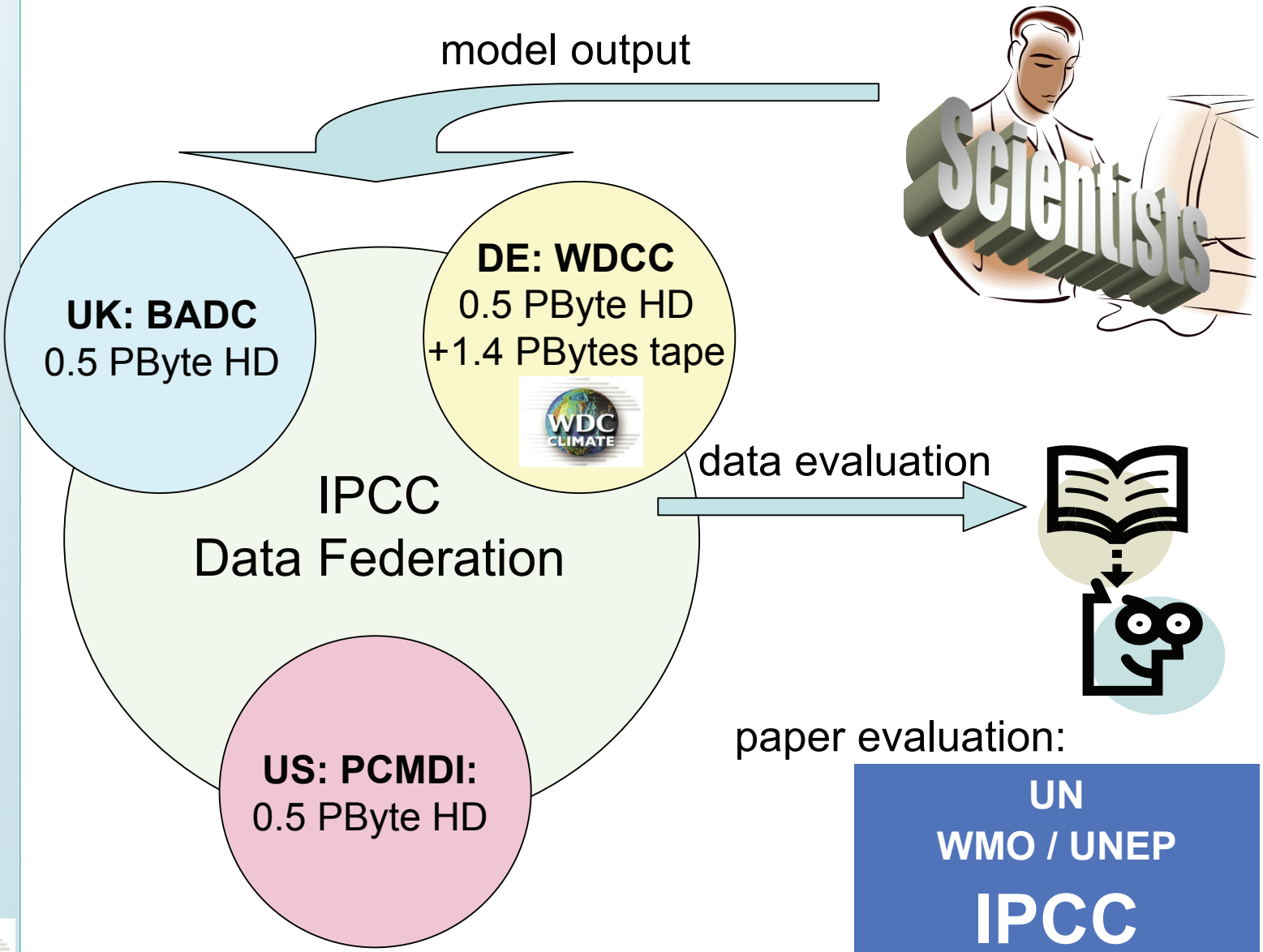
WDC CLIMATE

Computational support for the German climate science community

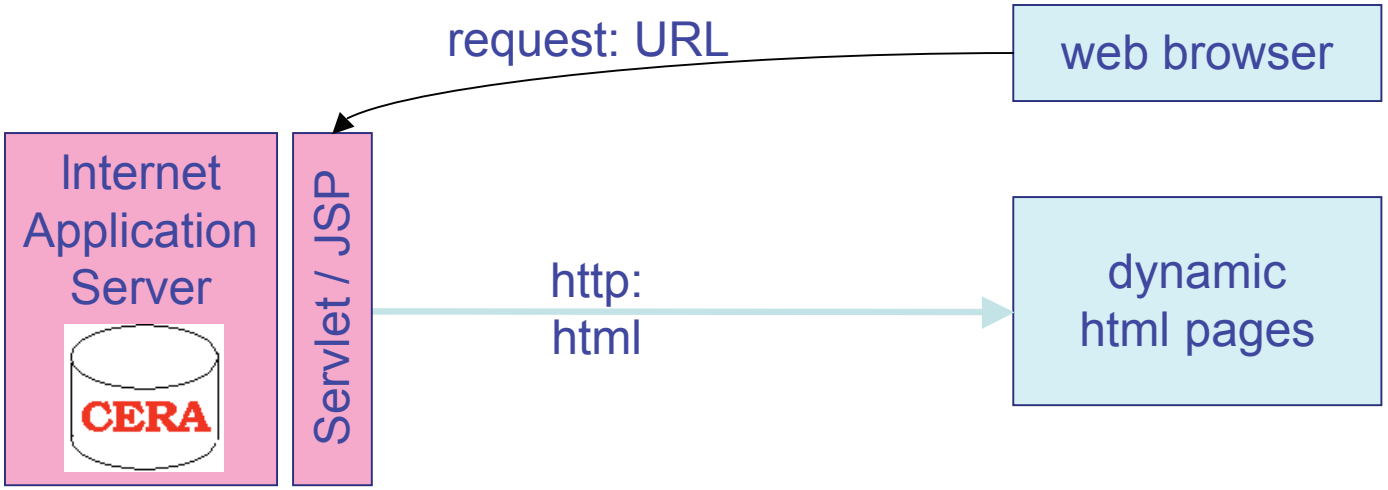
Earth System modelling and data management

Store & disseminate climate research data for the scientific community

WDCC as IPCC Data Node

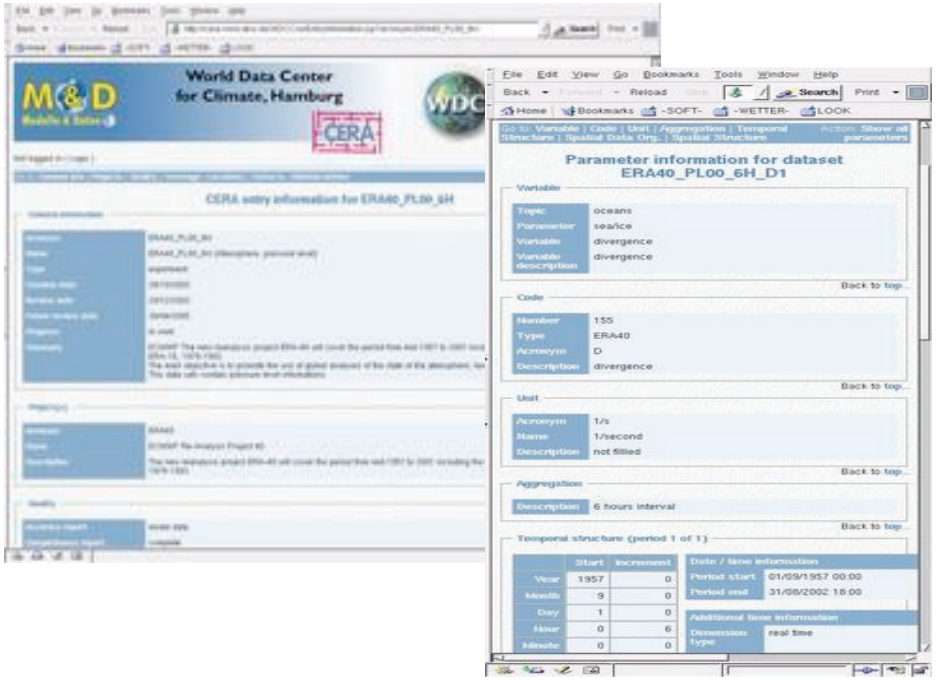


Catalogue Access: the GUI

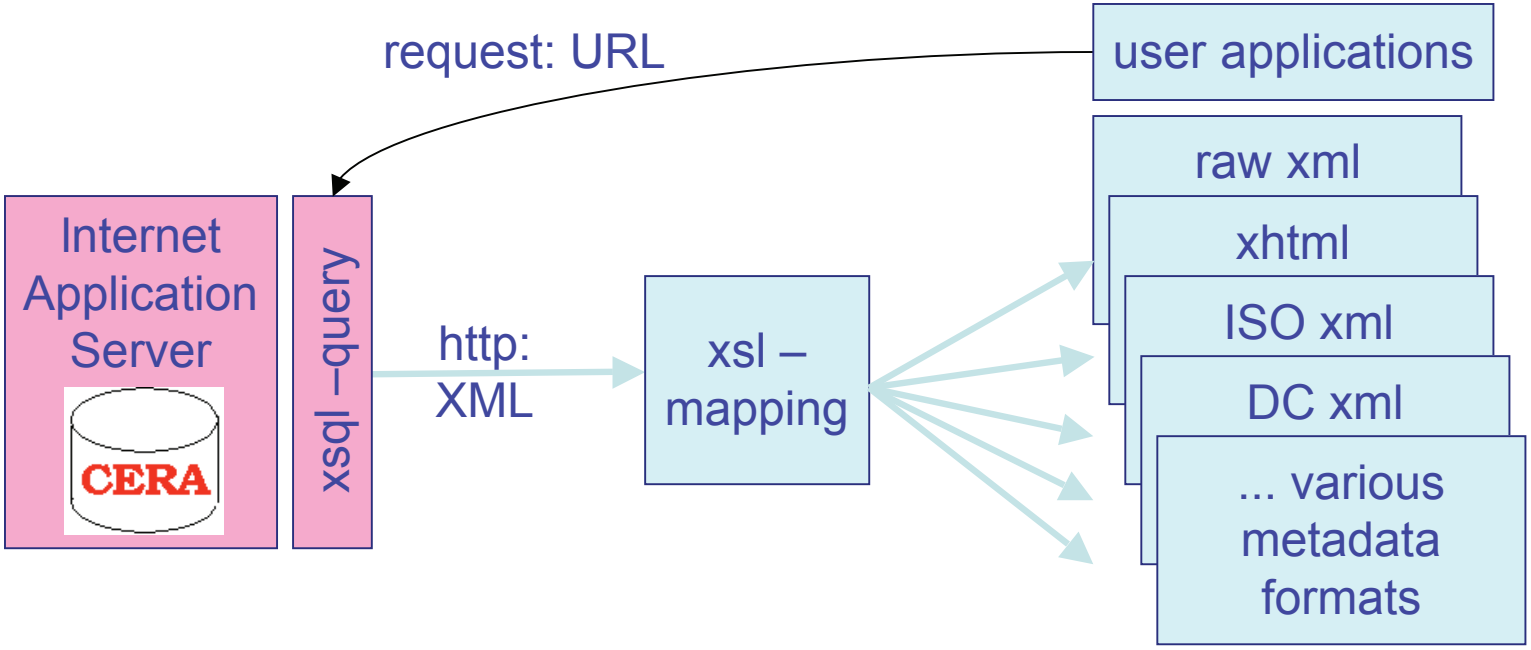


Catalogue access via WWW

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



http Metadata Output



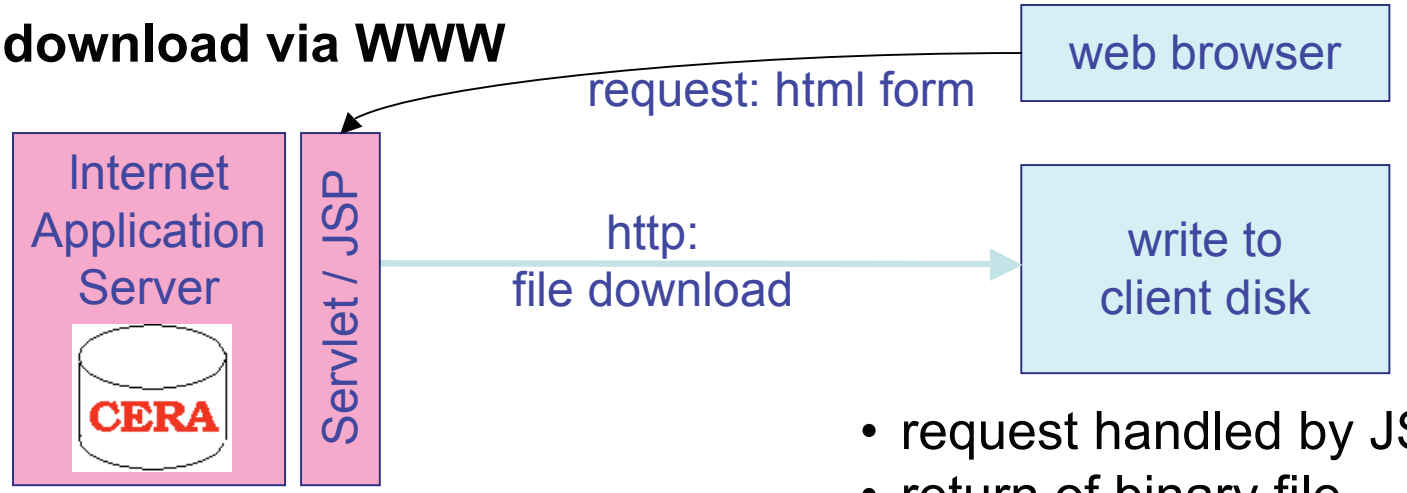
Metadata access via WWW:

- 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/wdcc/
          </CharacterString>
        </organisationName>
      - <role>
          <CI_RoleCode codeList="http://mad.dkrz.de/Daten/Metadata_Fill/contact_type.xsql" codeListValu
        </role>
      </CI_ResponsibleParty>
    </contact>
```

Data Download

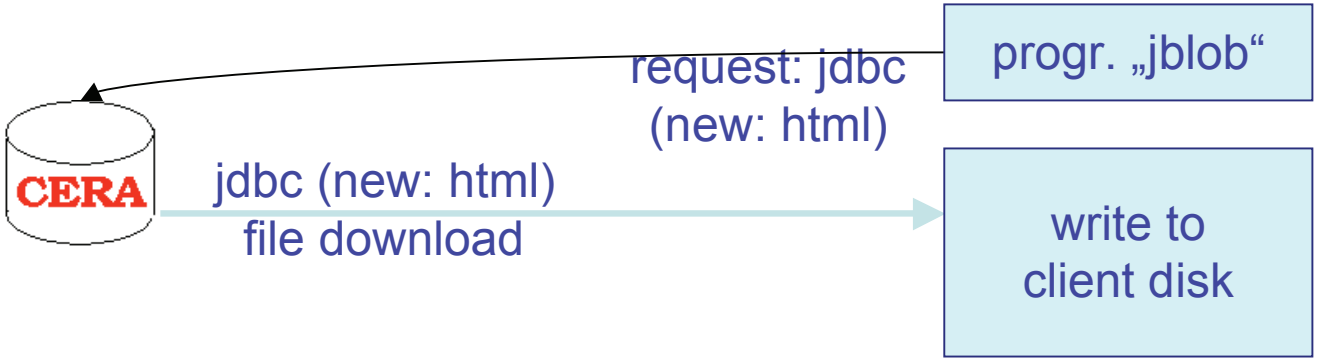
Data download via WWW



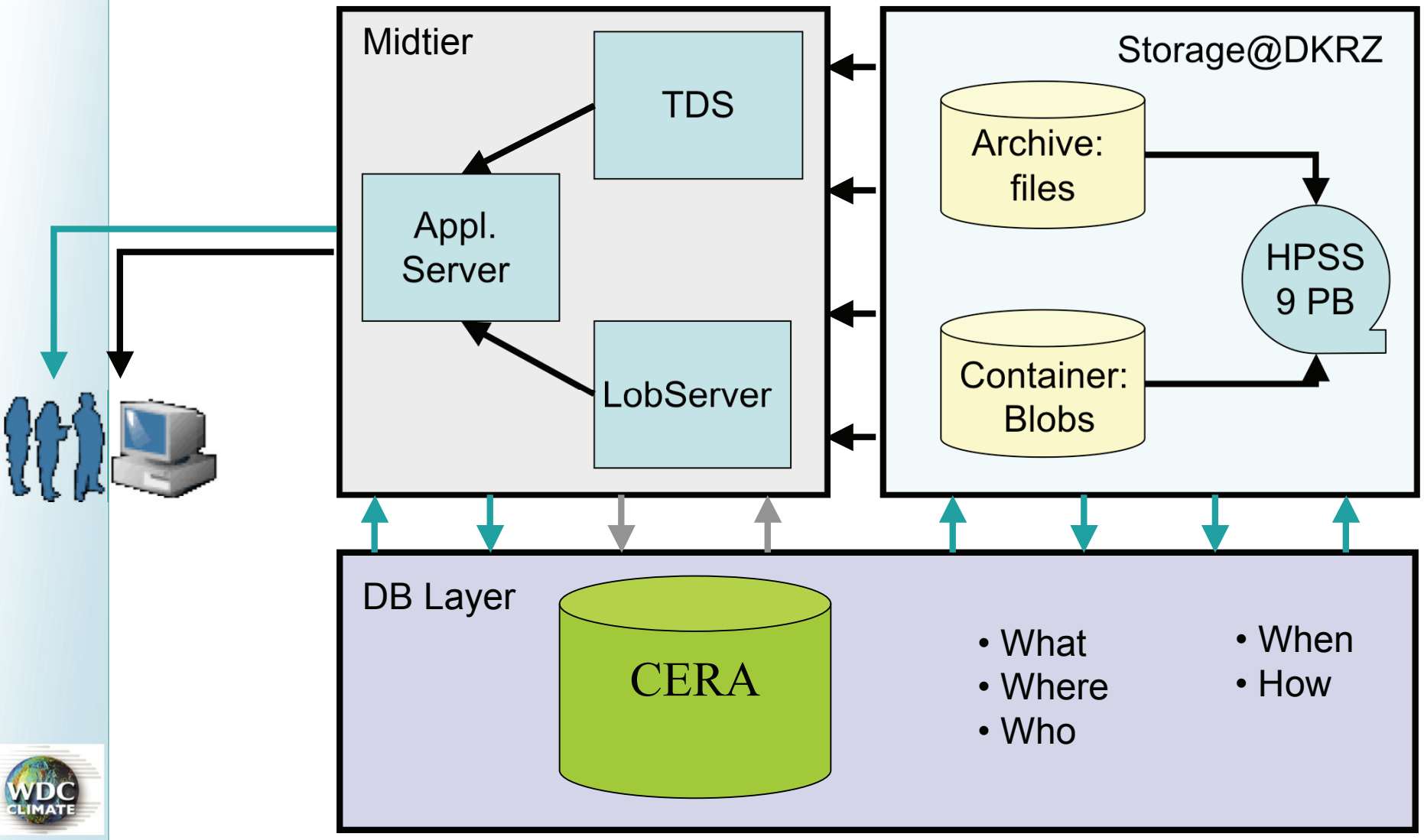
- request handled by JSP
- return of binary file

Data download via script/batch

- standard client side jdbc retrieval
- return of binary file



WDCC Data Access: Autumn 2009



Interoperability

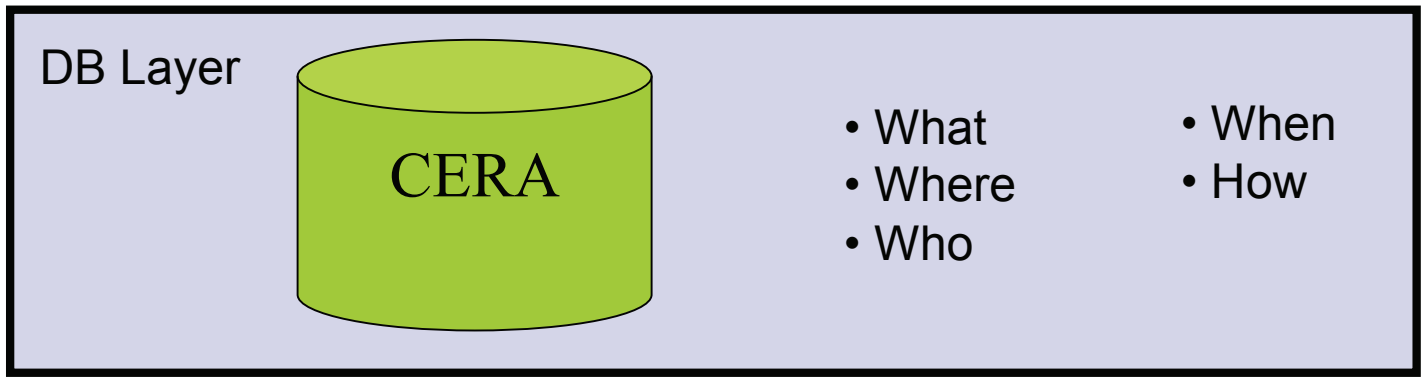
Metadata interoperability

- 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

DB Data Holding: A & A

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.

DB Data Holding: Metadata

Efficient handling of detailed metadata

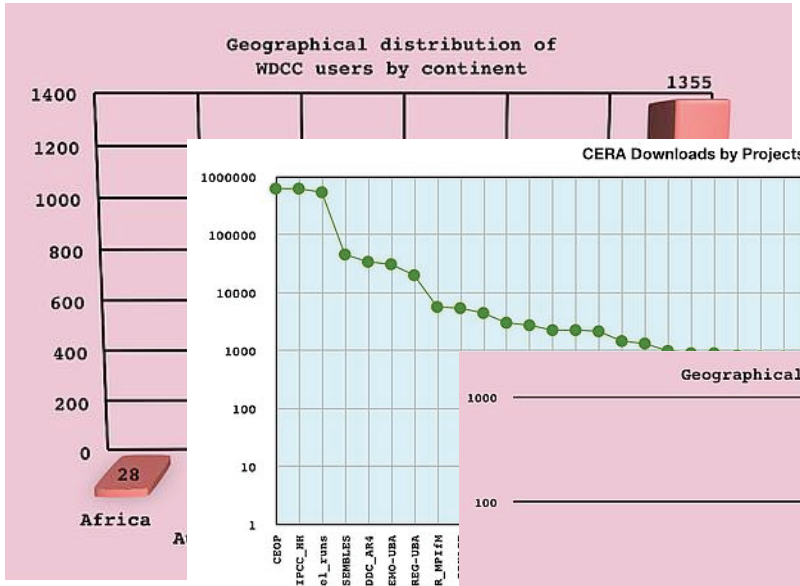
- easy and structured administration of > 60 tables
- detailed catalog information for users
- access support:
  Java Server Pages (JSP), Servlets, jdbc, xsql
  including standard DB features (views, ...)

DB Data Holding: Data

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)

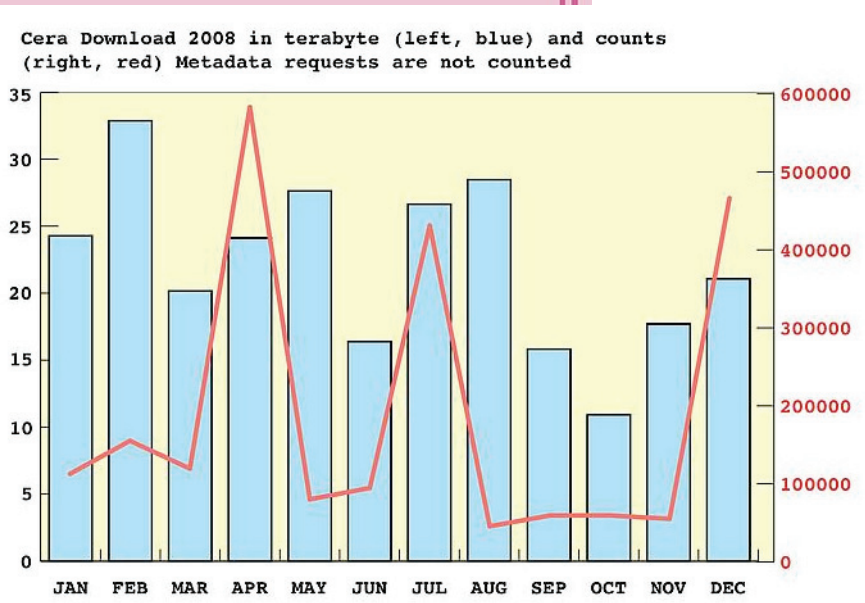
User Statistics



... by continent

... by project

... by country



... by time

User Statistics – what for?

- When are times of low traffic / are good times for maintenace?
- 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

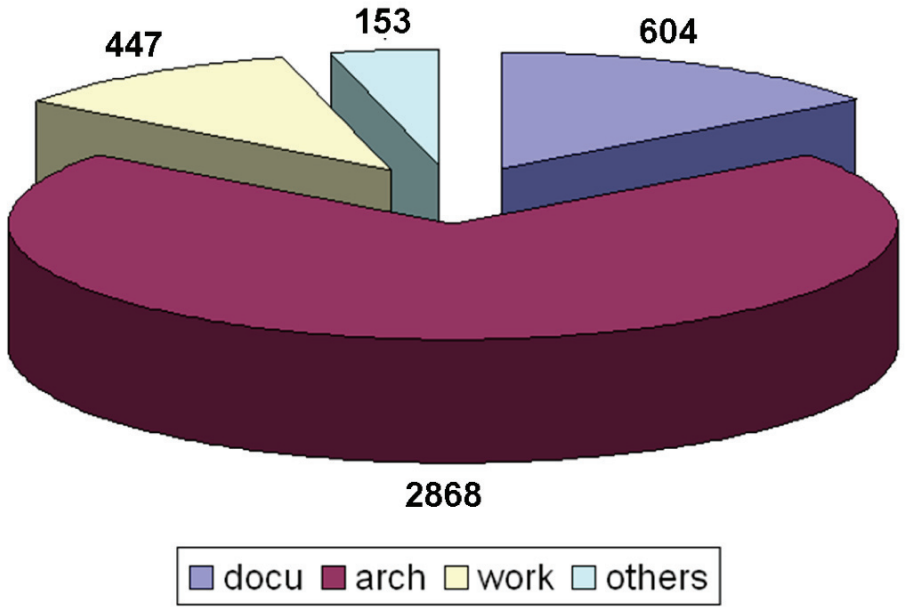
Necessity to Handle Different Rights

- 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

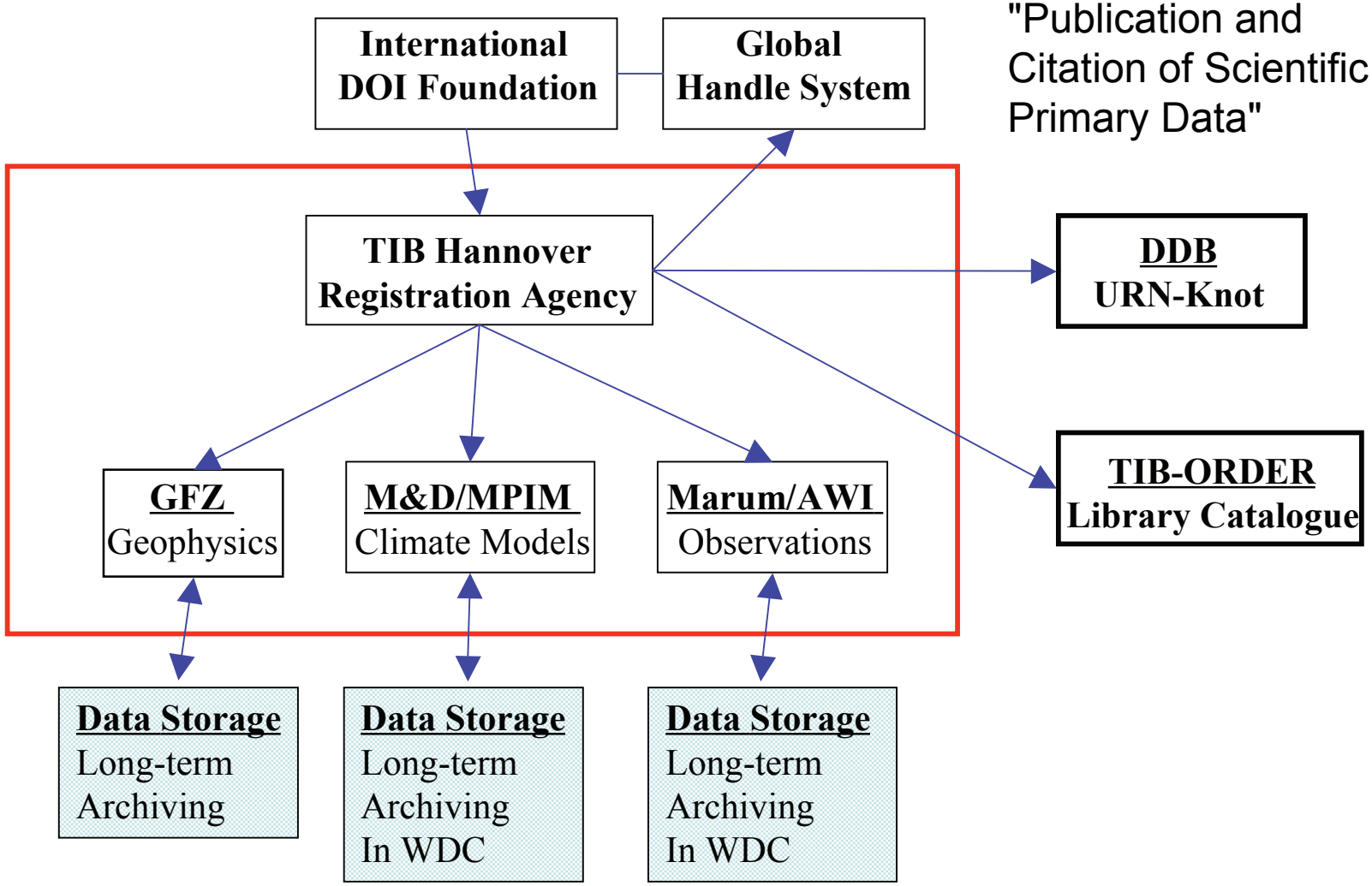
1. Test data — life cycle: weeks to months
2. Project data — life cycle: 3 – 5 years
3. Final results — life cycle: 10 years and longer

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

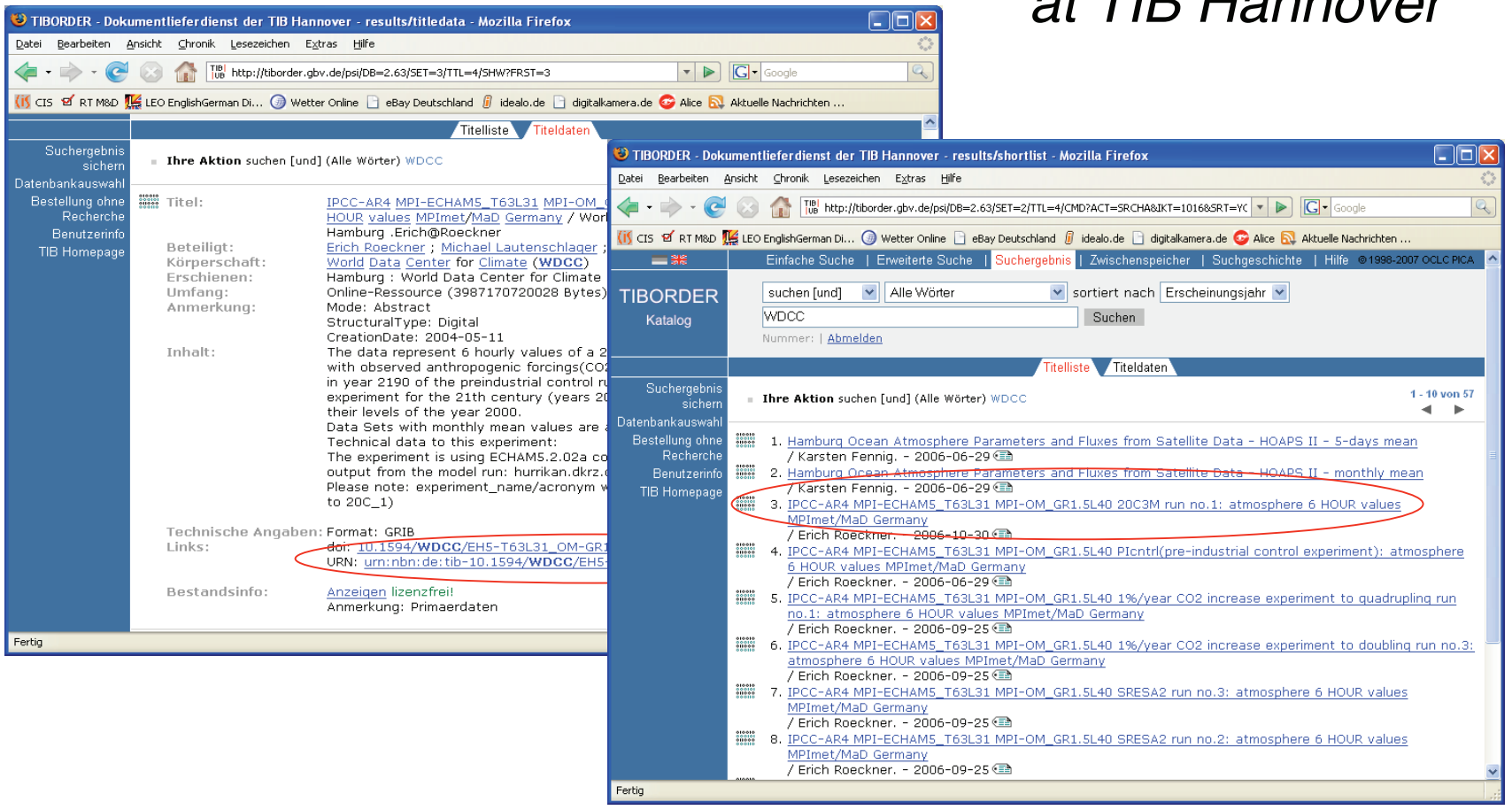
Digital Object Identifiers (DOI)

DFG Project "Publication and Citation of Scientific Primary Data"



Connections to Central Catalogues

WDC-Climate hosted data in the catalogue at TIB Hannover



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

<u>The Earth System Sciences viewpoint</u>

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?