CERA – the technical basis for WDCC

Hannes Thiemann

Michael Lautenschlager

Deutsches Klimarechenzentrum GmbH, Germany
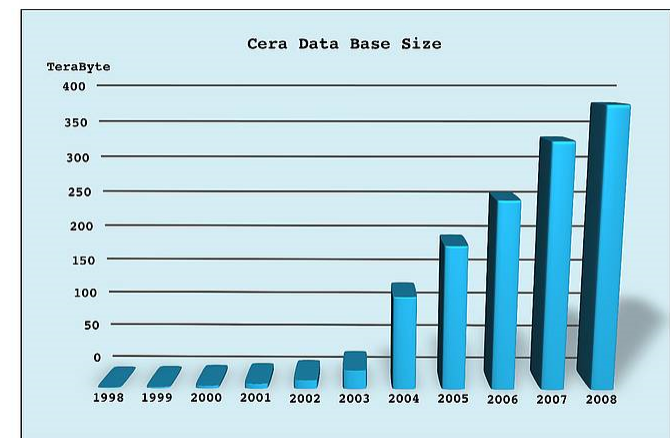
EGU 2010

World Data Centre 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

- Calendar year 2009: 800 active users
- Data from
  - 80 projects
  - 1400 experiments
  - 170000 datasets
  - 8.7 Billion records
  - 1 Million downloads
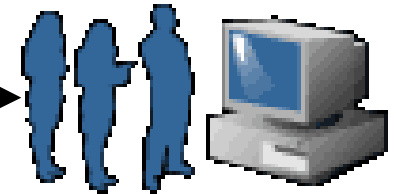- more than 255 TByte in total

Cera Data Base Size
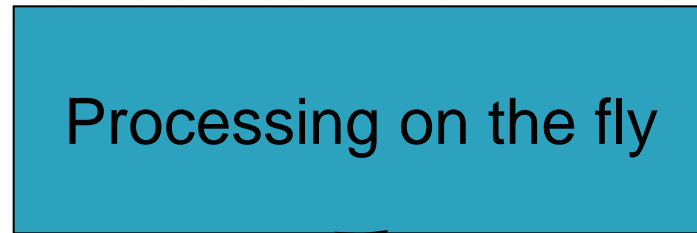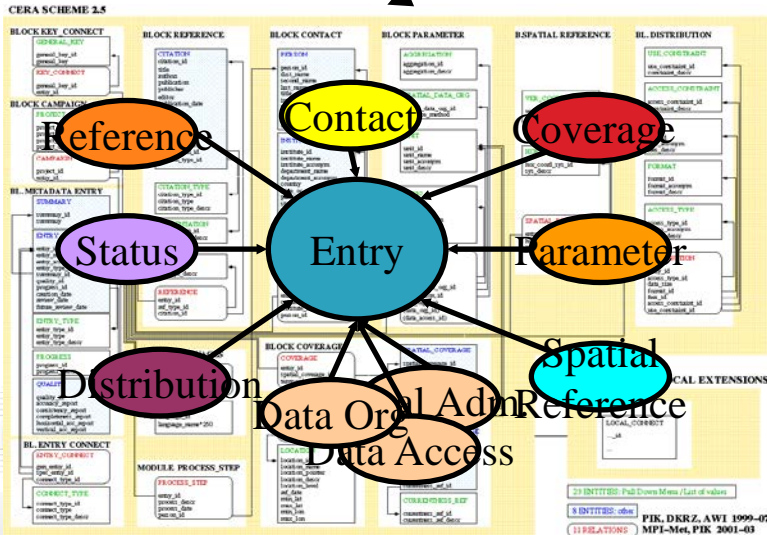
World Data Centre on Climate

- Most active German Projects
  - COPS
  - REMO-UBA / BFG
  - CLM Consortial Runs
  - MILLENNIUM_COSMOS

- Anticipated projects
  - CMIP5
  - IPCC AR5
    - Global and Regional
  - STORM
  - EUCLIPSE
  - And many more

- Most active International Projects
  - CEOP
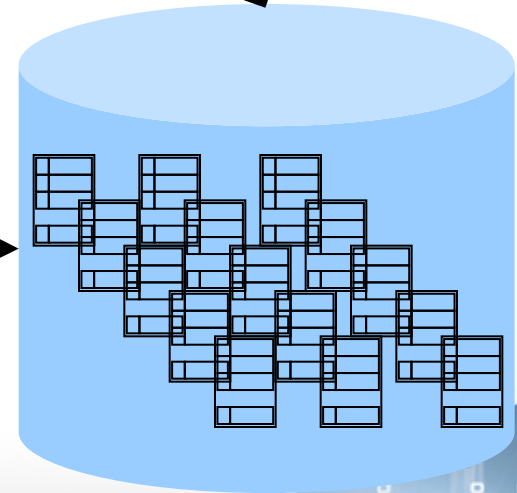 - ENSEMBLES
  - DPHASE
  - Metafor
  - IS-ENES
  - IPCC
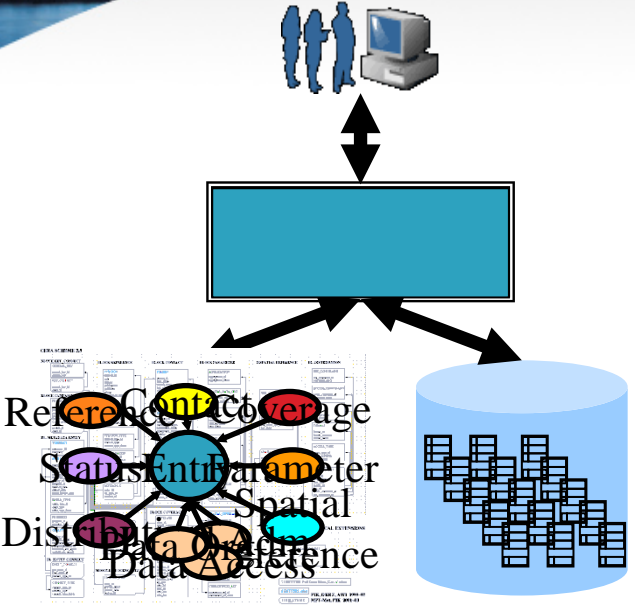
CERA
General Architecture

Processing on the fly

CERA2 Data Model

CERA2 Data Storage

CERA as part of DKRZ infrastructure

Reference Coverage

Status Entry Parameter

Spatial

Distribution Data Reference

Data Access

StorageTek Silos
Total Capacity:
60000 Tapes
Approx. 60 PB
(LTO and Titan)
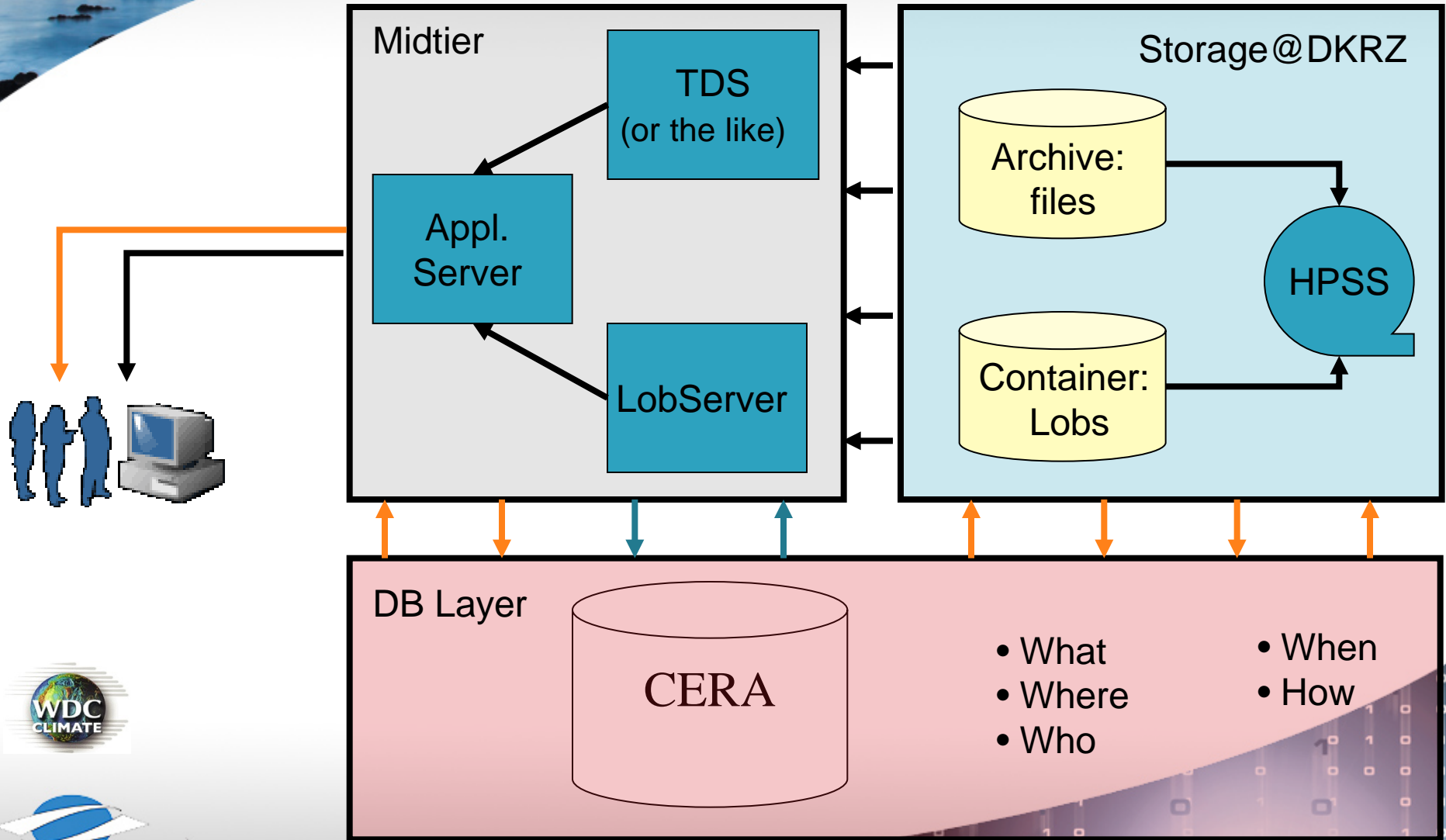
HPSS
(10 Pbyte /a)

WDC CLIMATE

DKRZ

WDCC / DOI

- Additionally WDCC offers the primary data publication service for final data entities which are of general scientific interest

  - Following the STD-DOI concept (Scientific and Technical Data – Digital Object Identifier, URL: www.std-doi.de)

  - Important aspects of the publication process are

    - The identification of independent data entities which are suitable for publication at the level of scientific literature,

    - The execution of an elaborated review process for metadata and climate data,

    - The assigment of additional metadata for electronic publication (ISO 690-2) and of persistent identifiers (DOI / URN) and

    - The integration of publication metadata and persistent identifiers into the TIB library catalogue (Technical Information Library, Hannover) so that primary data entities are searchable and citable together with scientific literature.

    - Quality characteristic is presently "approved by author", future development should be "peer reviewed".

ACLs and Statistics

▸ It is often required to manage ACLs

  ◦ Data owners want to publish papers before others start using the data

  ◦ Commercial use shall be prohibited

▸ Statistics on data usage are necessary

  ◦ Data owners want to know how often or who uses their data

  ◦ In case of problems or new versions users can be informed

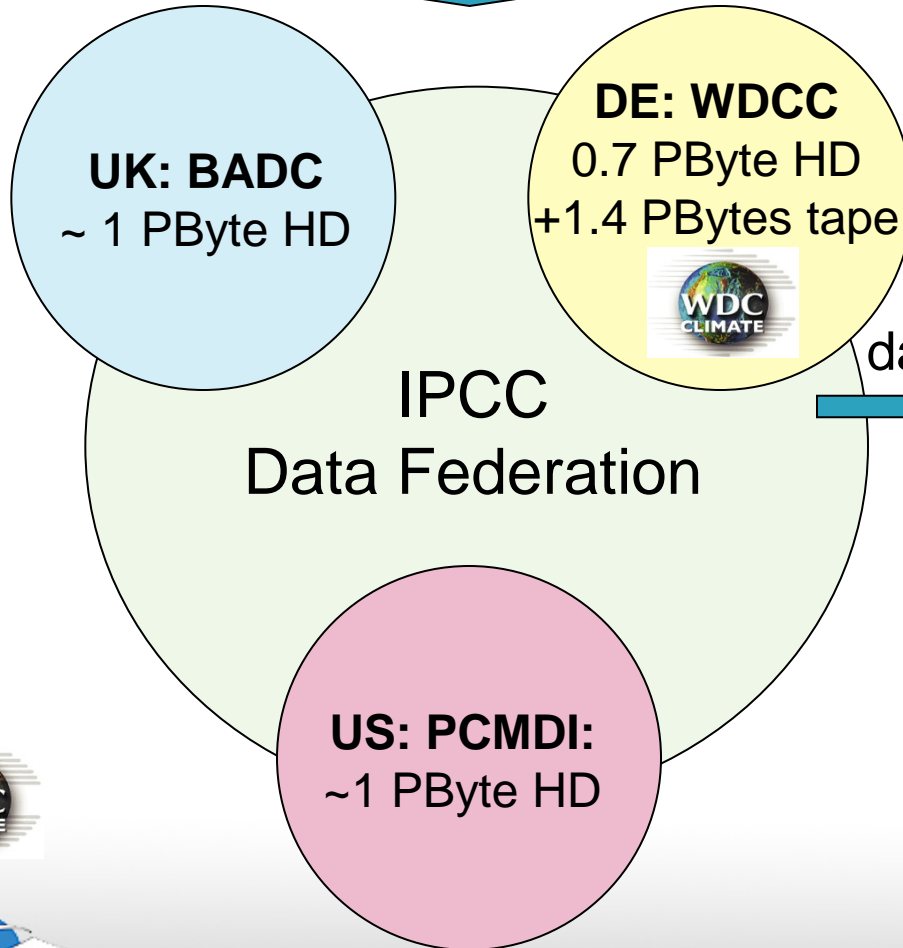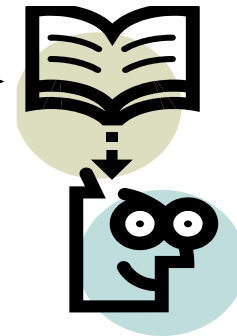 ◦ Gives important information how data shall be stored in future projects

WDCC data access

WDCC as IPCC / CMIP5 Data Node

model output

DE: WDCC
0.7 PByte HD
+1.4 PBytes tape

UK: BADC
~ 1 PByte HD

IPCC
Data Federation

US: PCMDI:
~1 PByte HD

data evaluation

paper evaluation:

UN

WMO / UNEP

IPCC

DKRZ

Summary

▸ CERA as a basis for WDCC

◦ CERA Metadata, DKRZ storage (disk, tape)

▸ Challenge: Integrate project data

management into long term archival

◦ More frequent changes in metadata and data

▸ Transition phase

◦ Metadata and data components

Thank you!

Contact

hannes.thiemann(at)zmaw.de