Preservation and Long Term Access of Data at the World Data Centre for Climate

Frank Toussaint

N.P. Drakenberg, H. Höck, S. Kindermann, M. Lautenschlager,
H. Luthardt, H. Ramthun, M. Stockhause, H. Thiemann

World Data Centre for Climate
at the German Climate Computing Centre (DKRZ)
Hamburg, Germany

Overview

- The World Data Centre for Climate
- Data Storage: Technology – Tapes and Disks
- Data Storage: LObStER – the Tape Storage Tool
- Storage Policy
- Long Term Archiving
- DOI – Digital Object Identifier

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

The World Data Centre for Climate

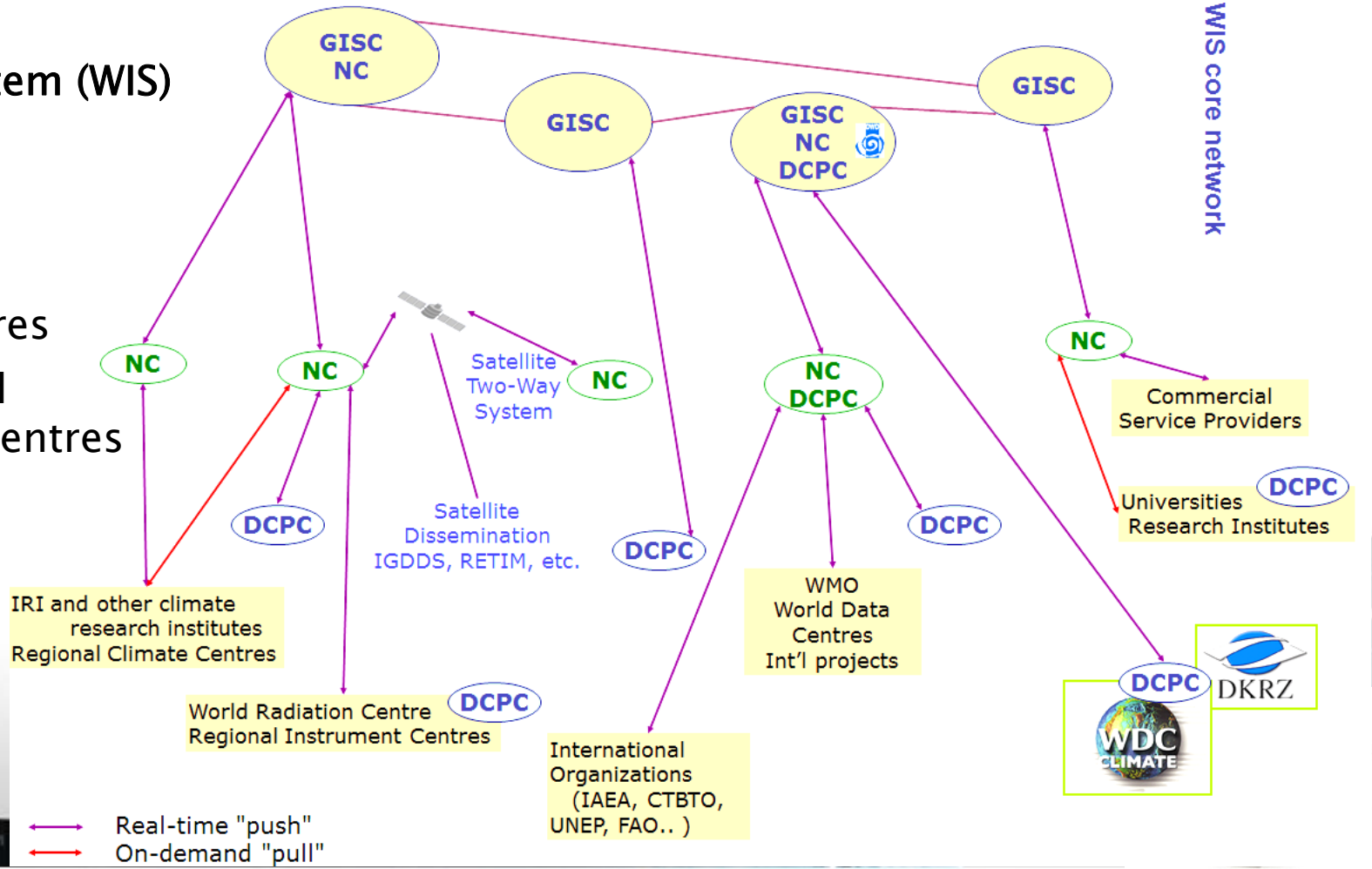
The German Climate Computing Centre (DKRZ) is held by…

- Max Planck Society, University of Hamburg, and others.
- Mission:
  Provide HP computing power and storage for the German Earth Science community (mainly)

The WDCC as WIS Data Collection & Production Centre

WMO Information System (WIS)

- National Centres
- Global Information System Centres
- Data Collection and Production Centres



DKRZ DEUTSCHES KLIMARECHENZENTRUM

The WDCC in the ICSU World Data System

- International Council for Science (ICSU)

  World Data System (WDS)

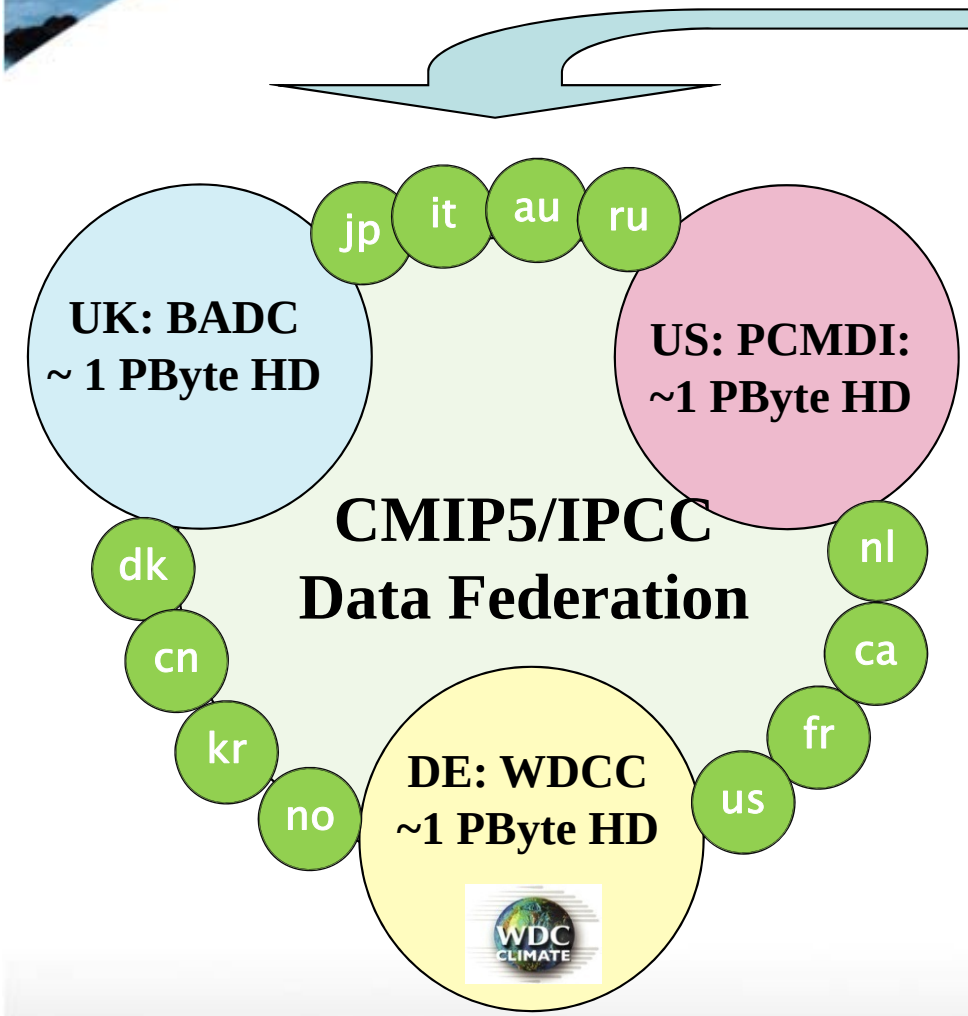
  World Data Centres (WDC)

- WDC Cluster Earth System Research:

  WDC-Mare, WDC-RSAT, WDC-Climate

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

The ESG Data Nodes



Earth System Grid Federation (ESGF):

>10 data nodes

Primary responsibility for CMIP5/IPCC-AR5: PCMDI, BADC, & WDCC

About 1 PB Data are replicated

The Evolution of Data Quantities

Climate Model Data:

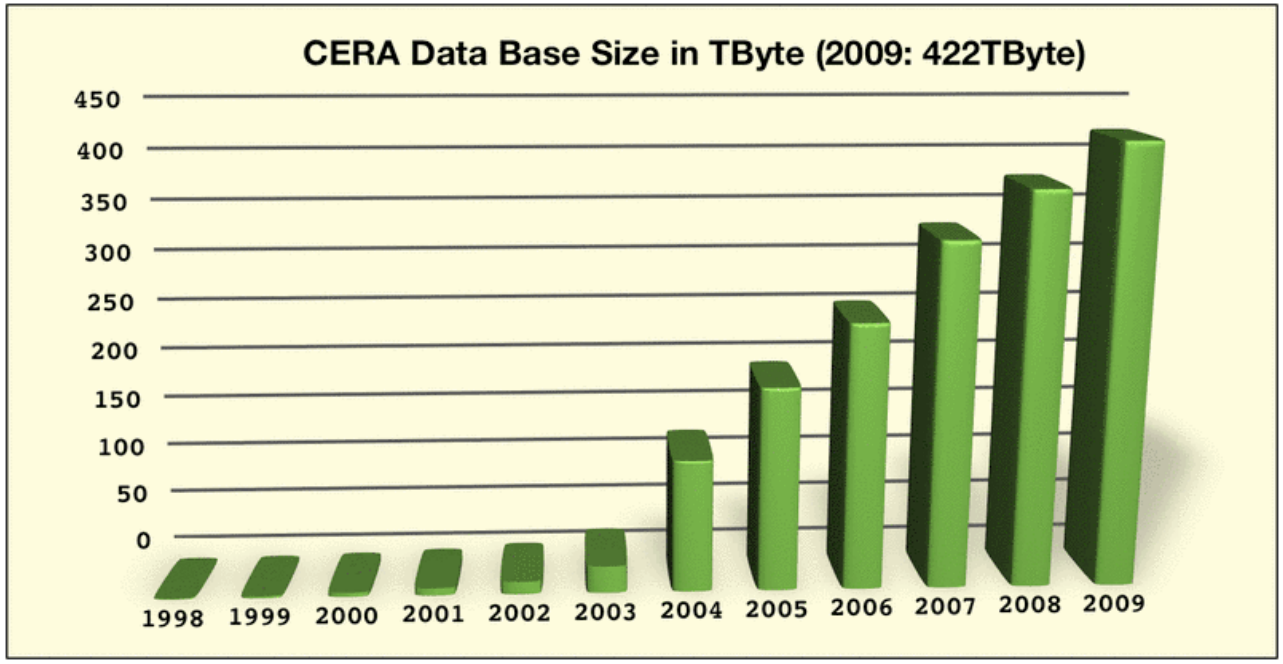
Relative

homogeneous

but huge amounts!

Needed: Tape

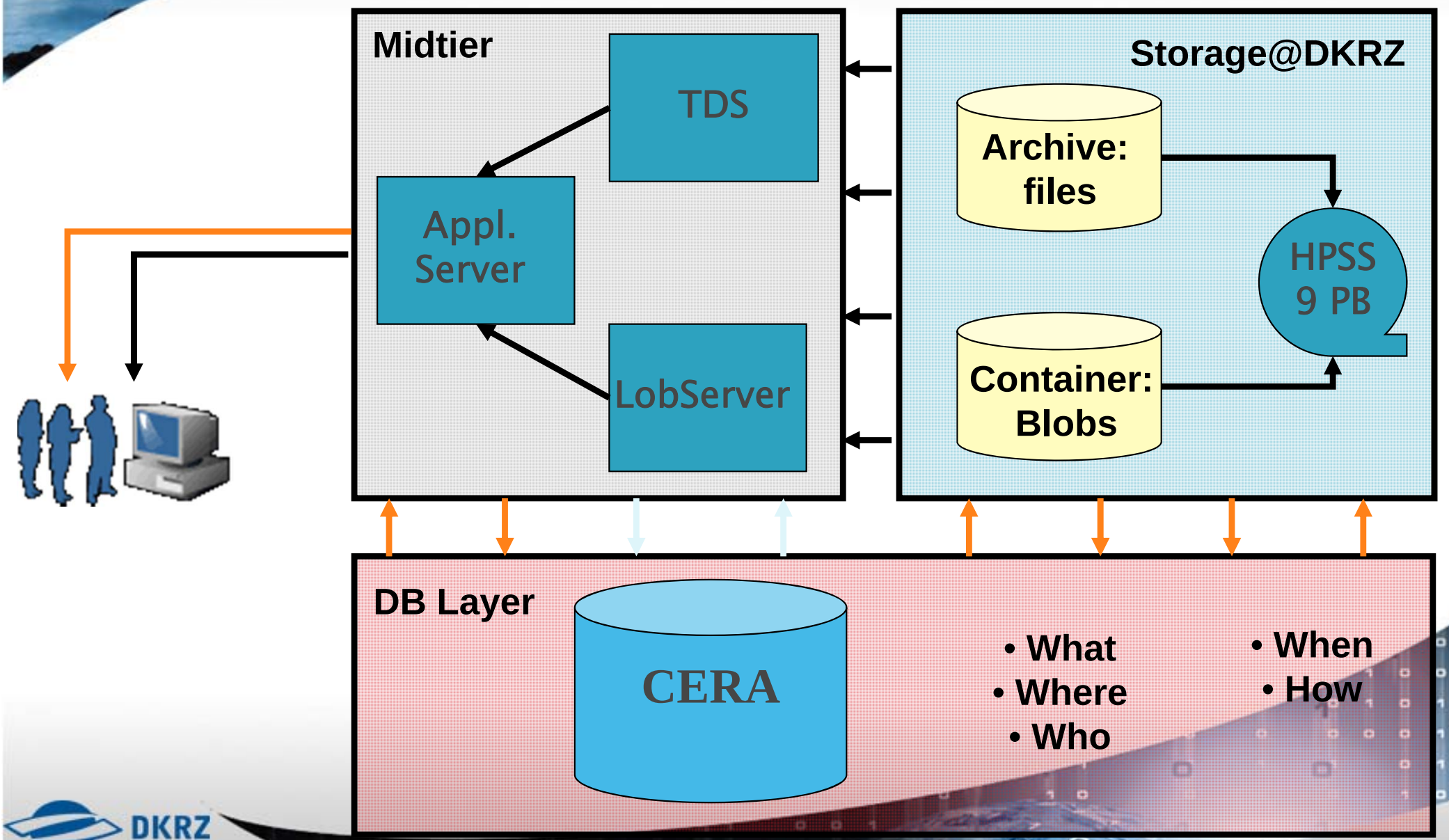
access (nearline)



DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

The Data Flows



DKRZ
DEUTSCHES KLIMARECHENZENTRUM

LObStER: Large Object Storage and Efficient Retrieval

- Huge amounts of data in each container file
- Very different sizes of records: 64b .. 2 Gb
- Efficient administration of all records
- Irregular access patterns
  (access latency independent of the record position)
- Transactional behaviour for read/write
- Fault tolerance for HD, controller, tapes, etc

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

LObStER

Application

generic
JDBC-driver

①

Lobster
configuration
manager

②

Application

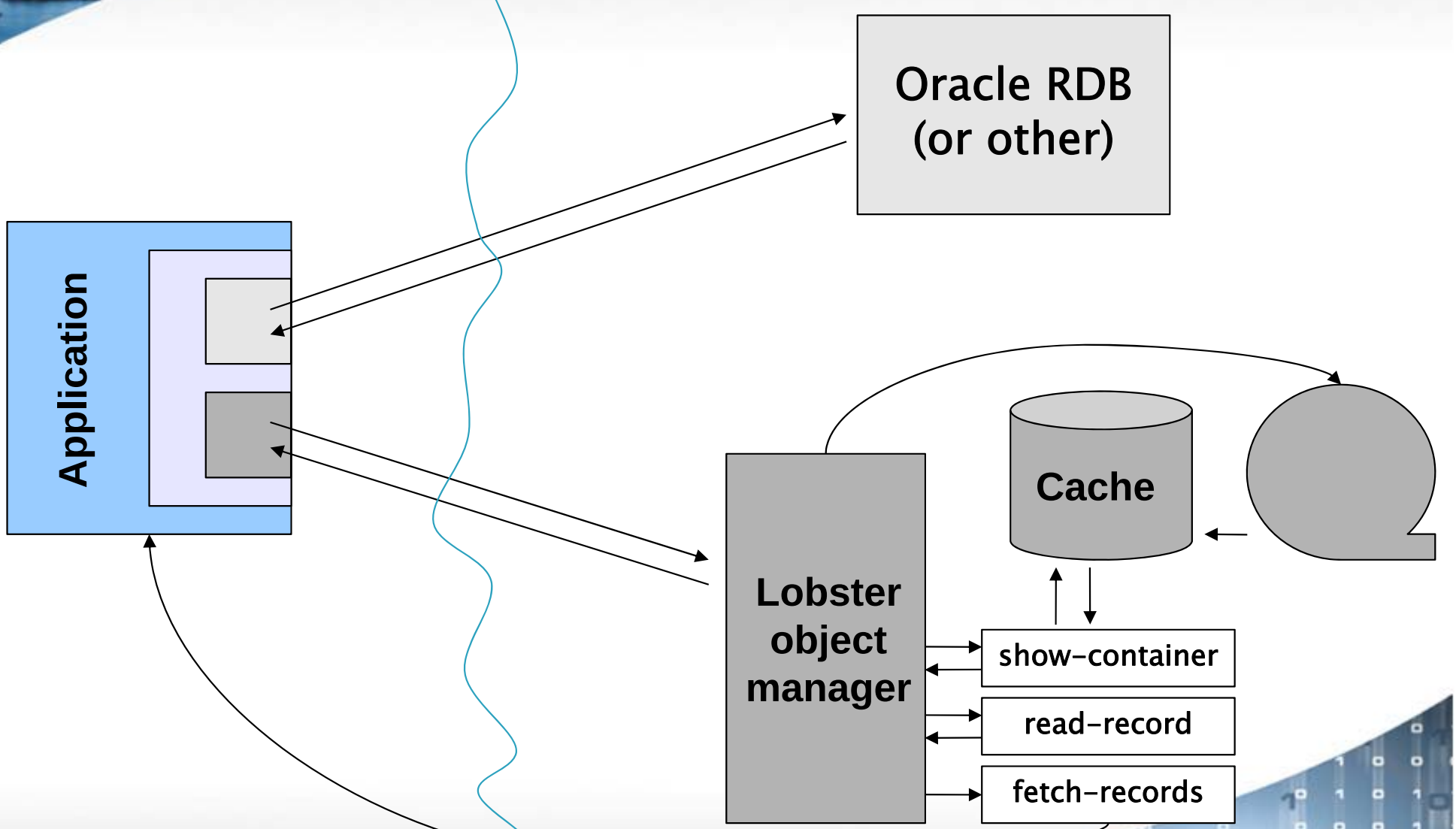
Internet

Intranet

specific JDBC-
drivers loaded

DKRZ
DEUTSCHES
KLIMARECHENZENTRUM

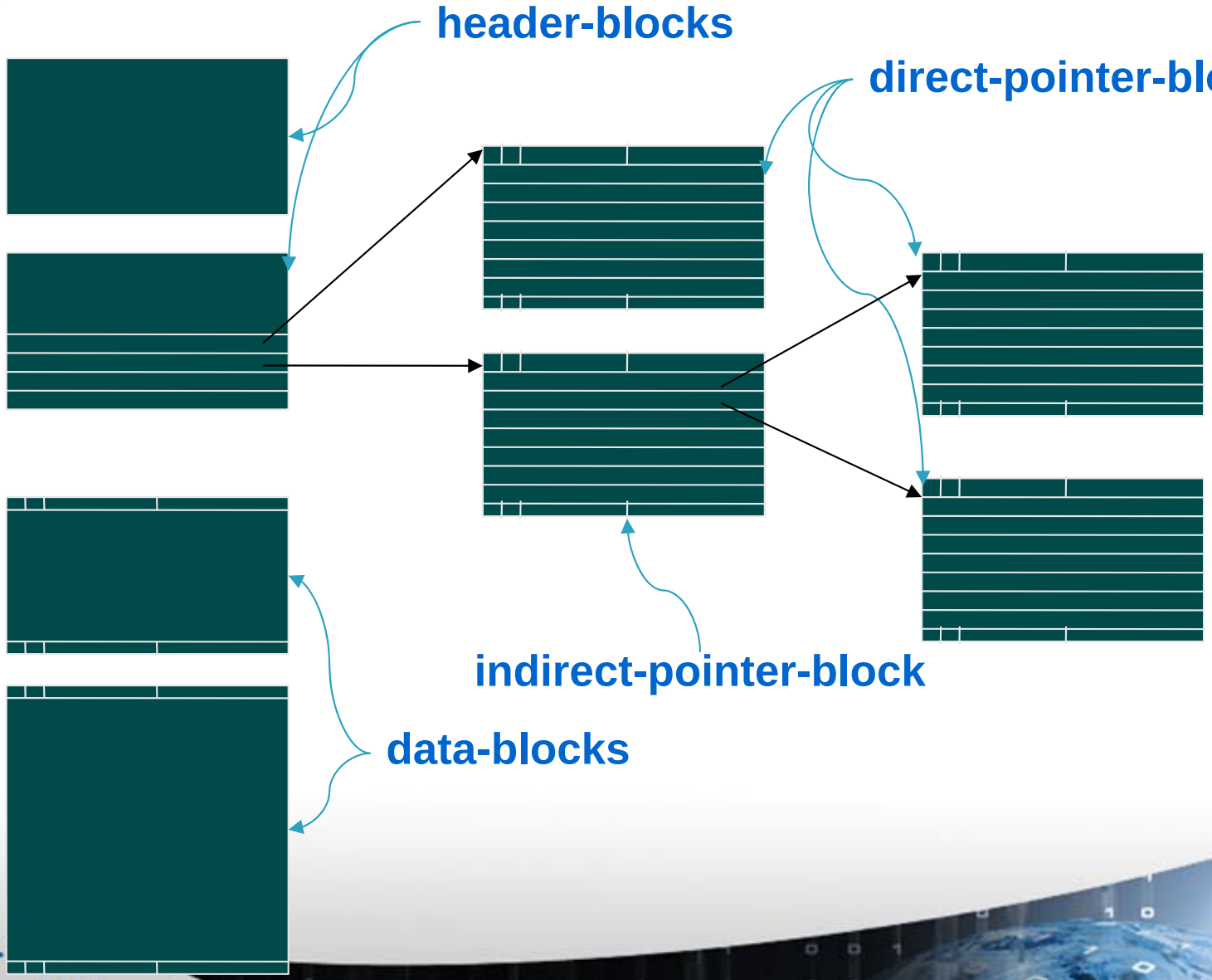
LObStER


Oracle RDB
(or other)
Application
Cache
Lobster
object
manager
show-container
read-record
fetch-records

LObStER: The Data Containers

- Container files with blocked format
- 64-bit files and 64-bit internal position referencing
- Max file size: 16384 PBytes
- Entries stored in ≥1 blocks
- Block sizes 2^k, $k \in \{ 8, 9, 10, \ldots, 62 \}$

LObStER: The Data Containers



Long Term Archiving

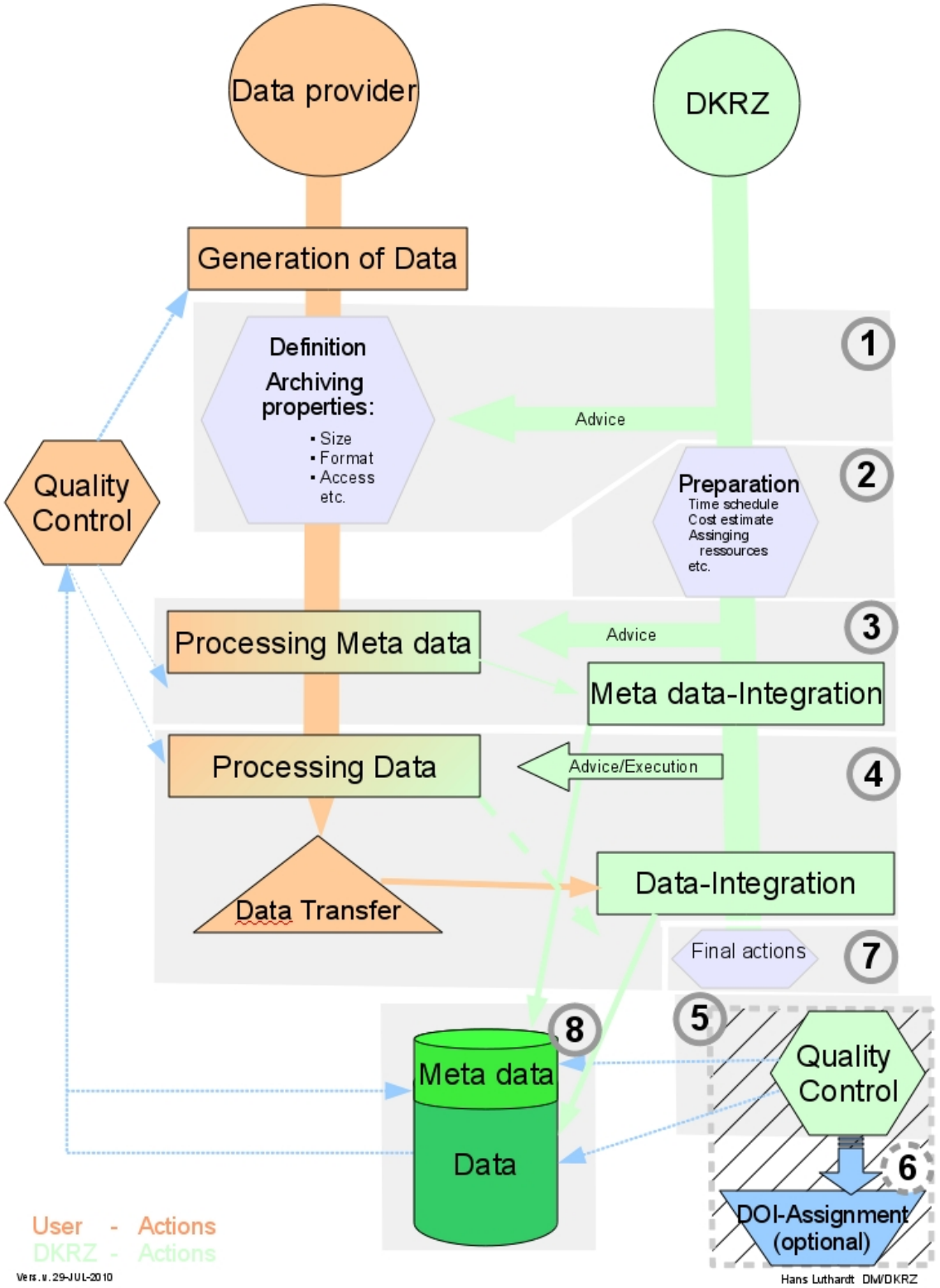
Several steps:

- specification & concept
- filling of metadata & data
- quality checks & DOI

- LTA for, e.g., EUCLIPSE, MedCLIVAR, combine

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE



LTA

Costs depend on complexity and efforts at our site:

- metadata
- reformatting
- etc

Long Term Archiving

- Quality Checks on three levels for LTA

QC L1: conformity to general standards

(format, ...)

QC L2: coarse automated content checks

QC L3: detailed spot checks:

TQA – Technical Quality Assurance

SQA – Scientific Quality Assurance

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

The WDC-Climate as Publishing Agency of the IDF

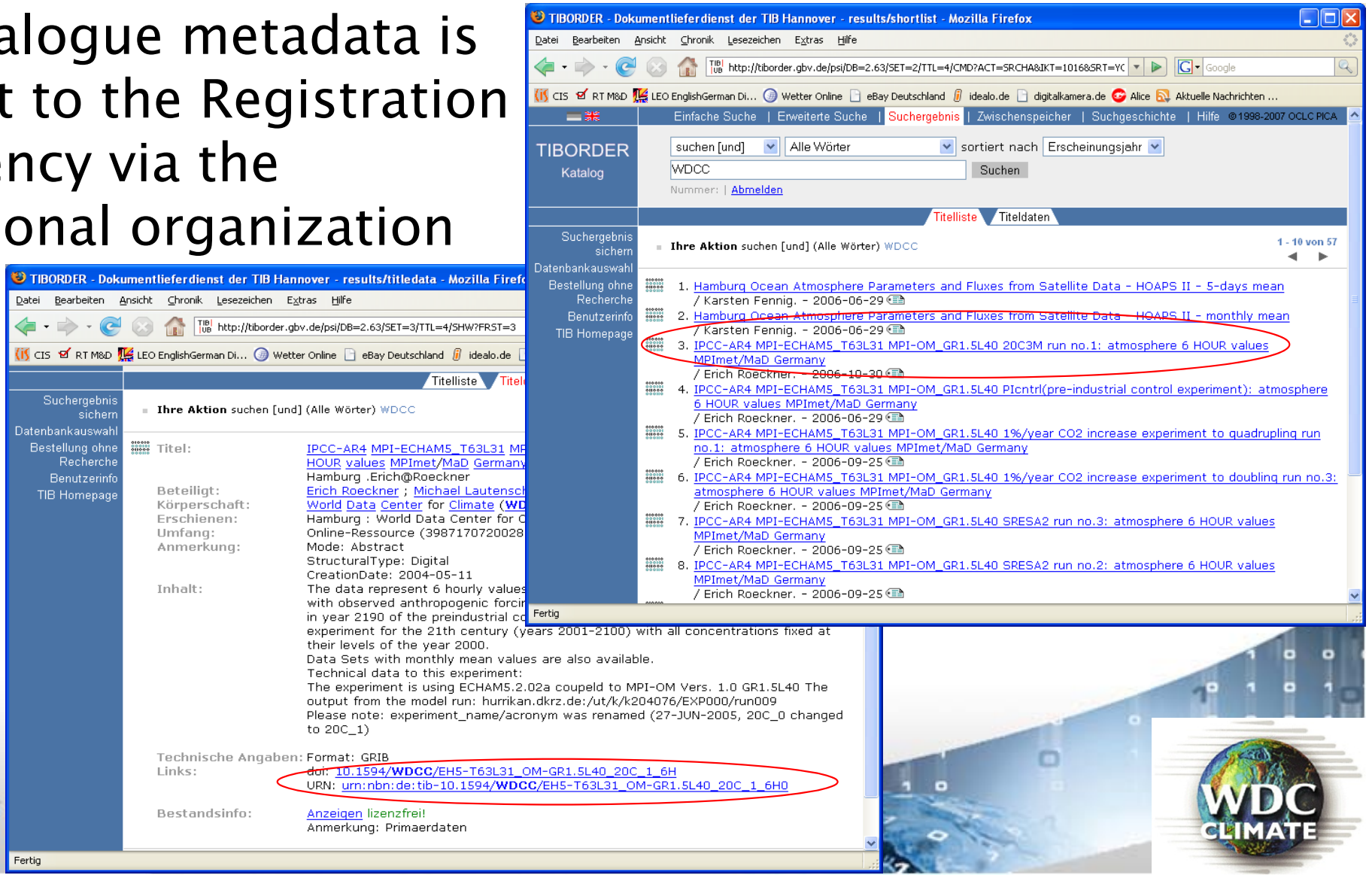
International DOI Foundation	doi.org
Registration Agencies	DataCite.org
National Organizations	tib-hannover.de
Publisher	wdc-climate.de

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

The Visibility of LTA Data in Public Catalogues

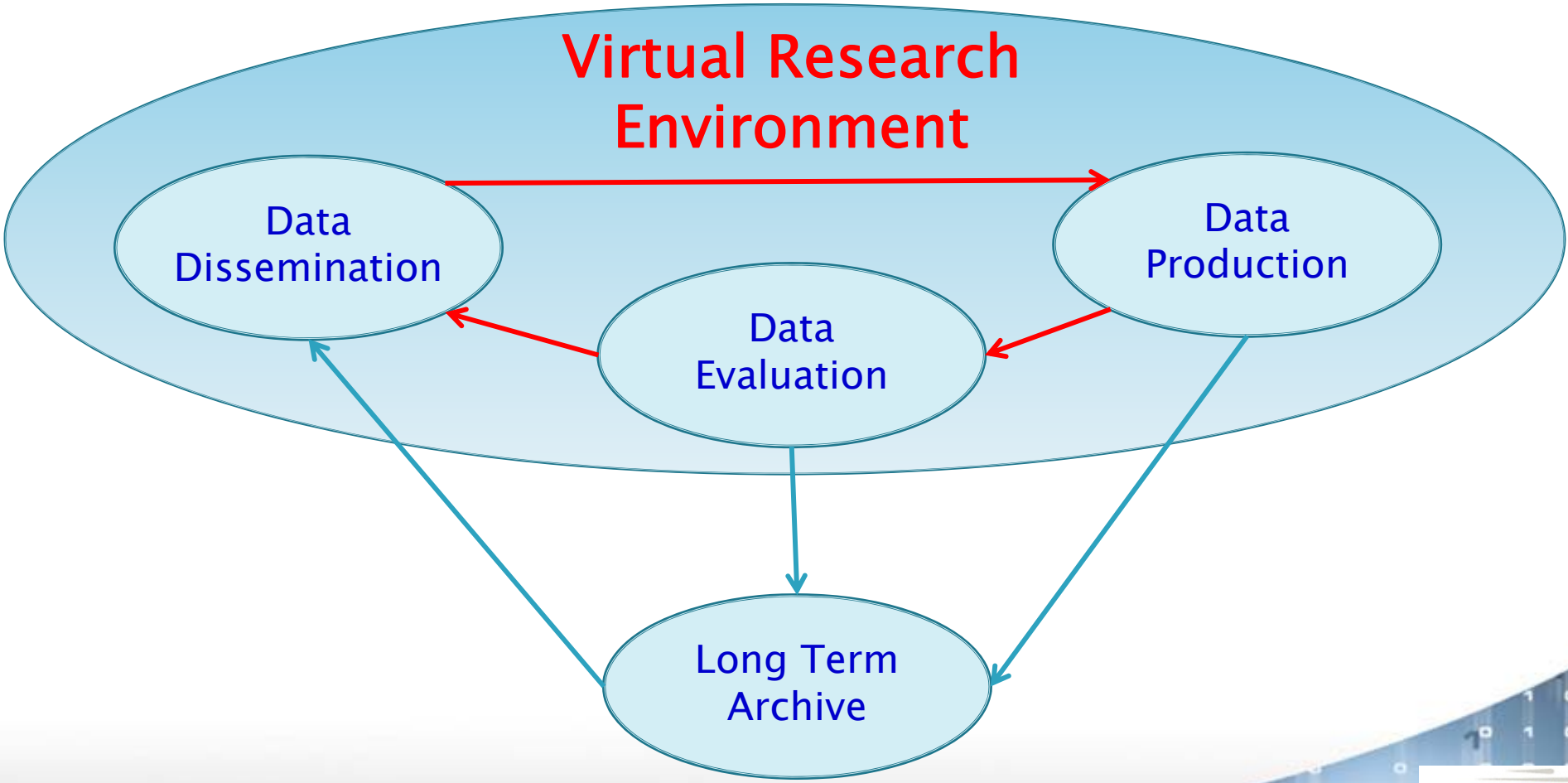
- DOI is given
- Catalogue metadata is sent to the Registration Agency via the national organization



DKRZ DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

The Data Life Cycle Management



DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE

THANK YOU.

QUESTIONS?

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

WDC CLIMATE