



# 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

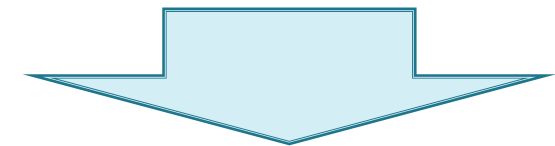
# Overview

- 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

# The World Data Centre for Climate

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

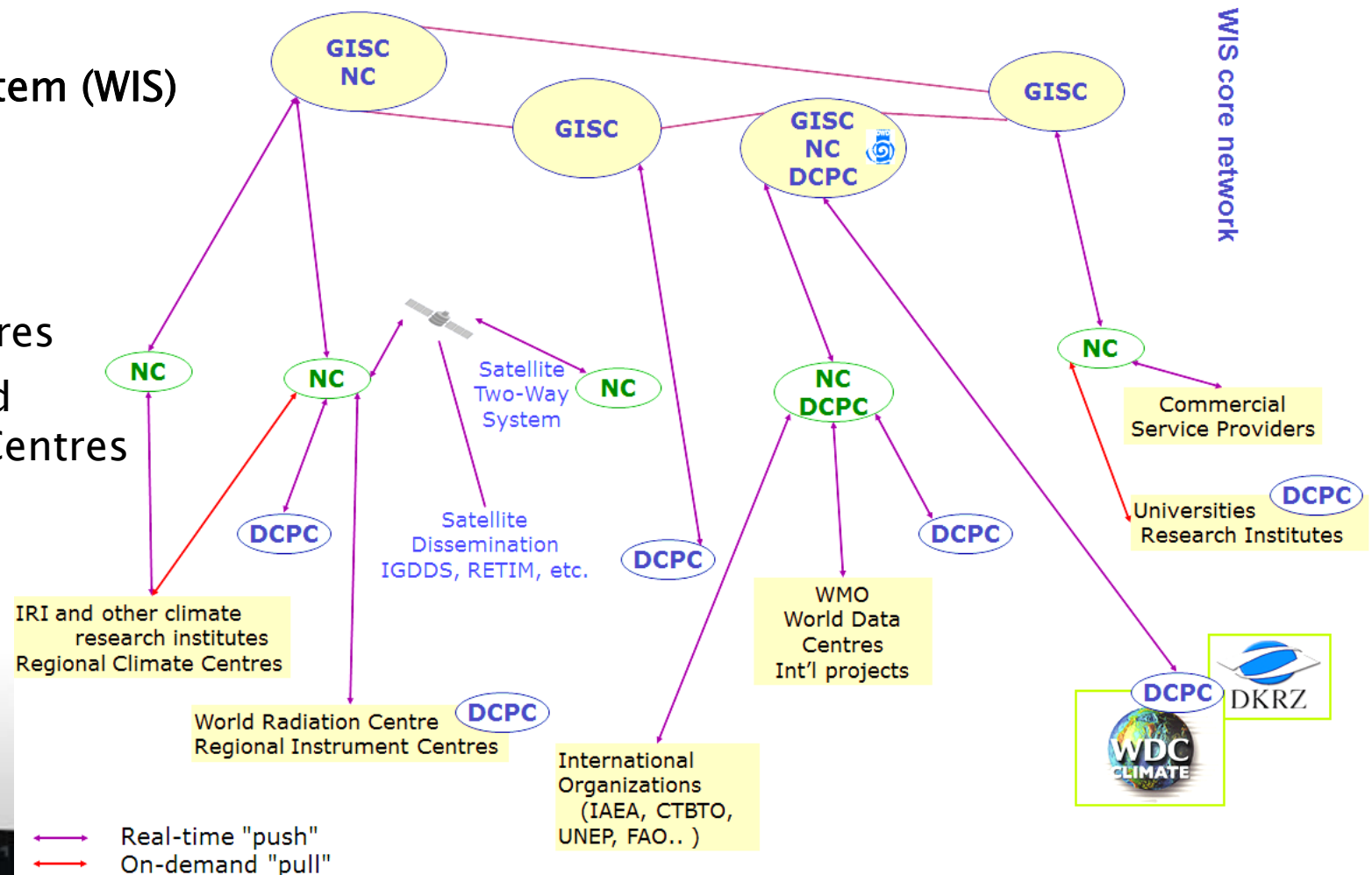
- Max Planck Society, University of Hamburg, and others.
- Mission: Provide HP computing power and storage for the German Earth Science community (mainly)



# The WDCC as WIS Data Collection & Production Centre

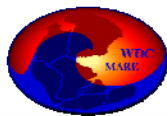
## WMO Information System (WIS)

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

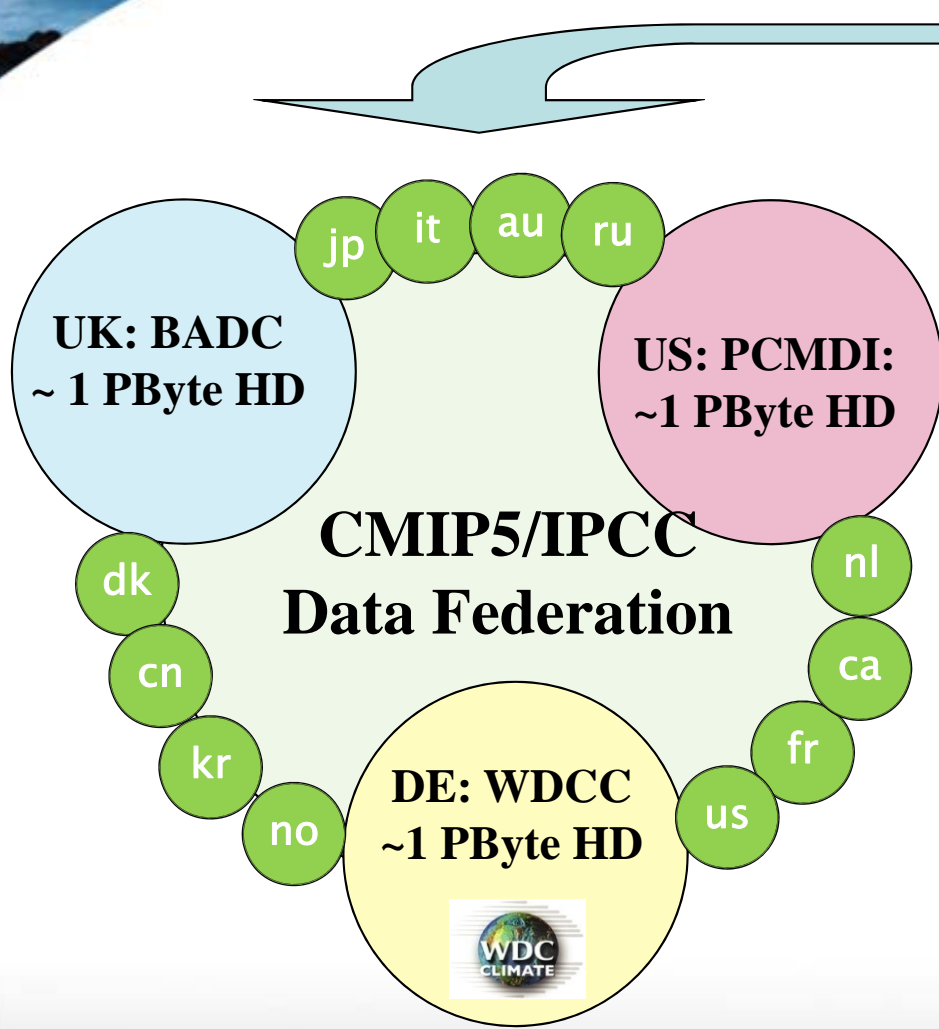


# The WDCC in the ICSU World Data System

- International Council for Science (ICSU)  
World Data System (WDS)  
World Data Centres (WDC)
- WDC Cluster **Earth System Research:**  
WDC–Mare, WDC–RSAT, WDC–Climate



# 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:

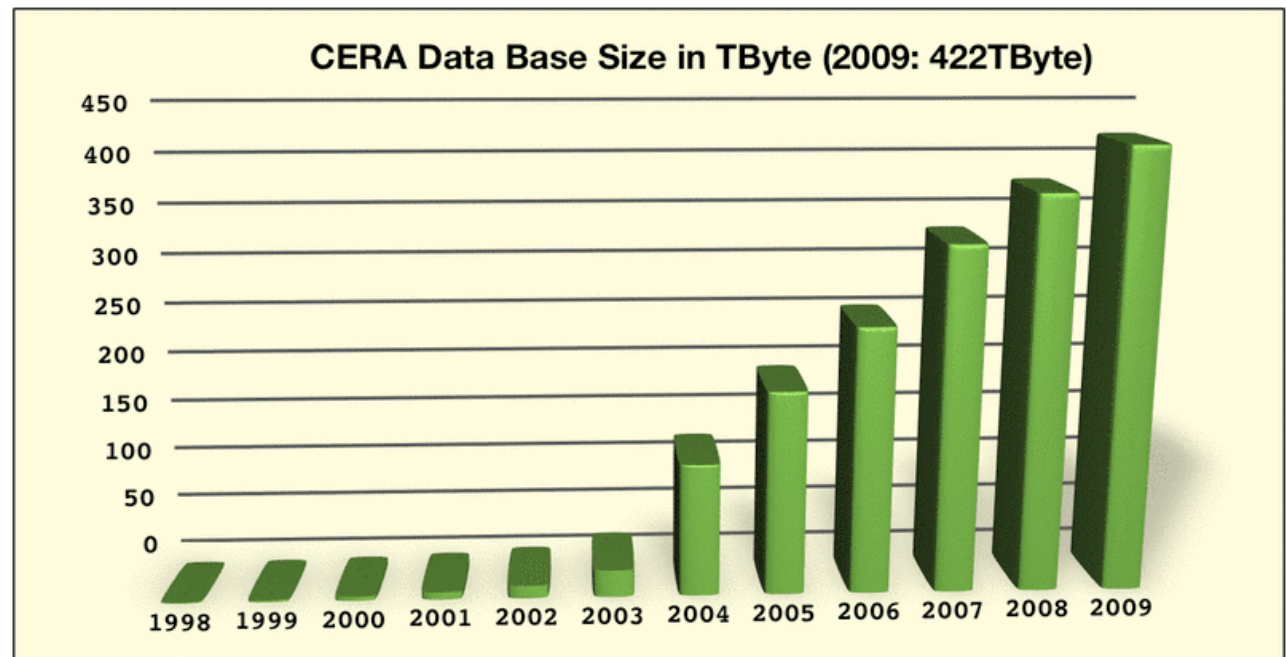
Relative

homogeneous

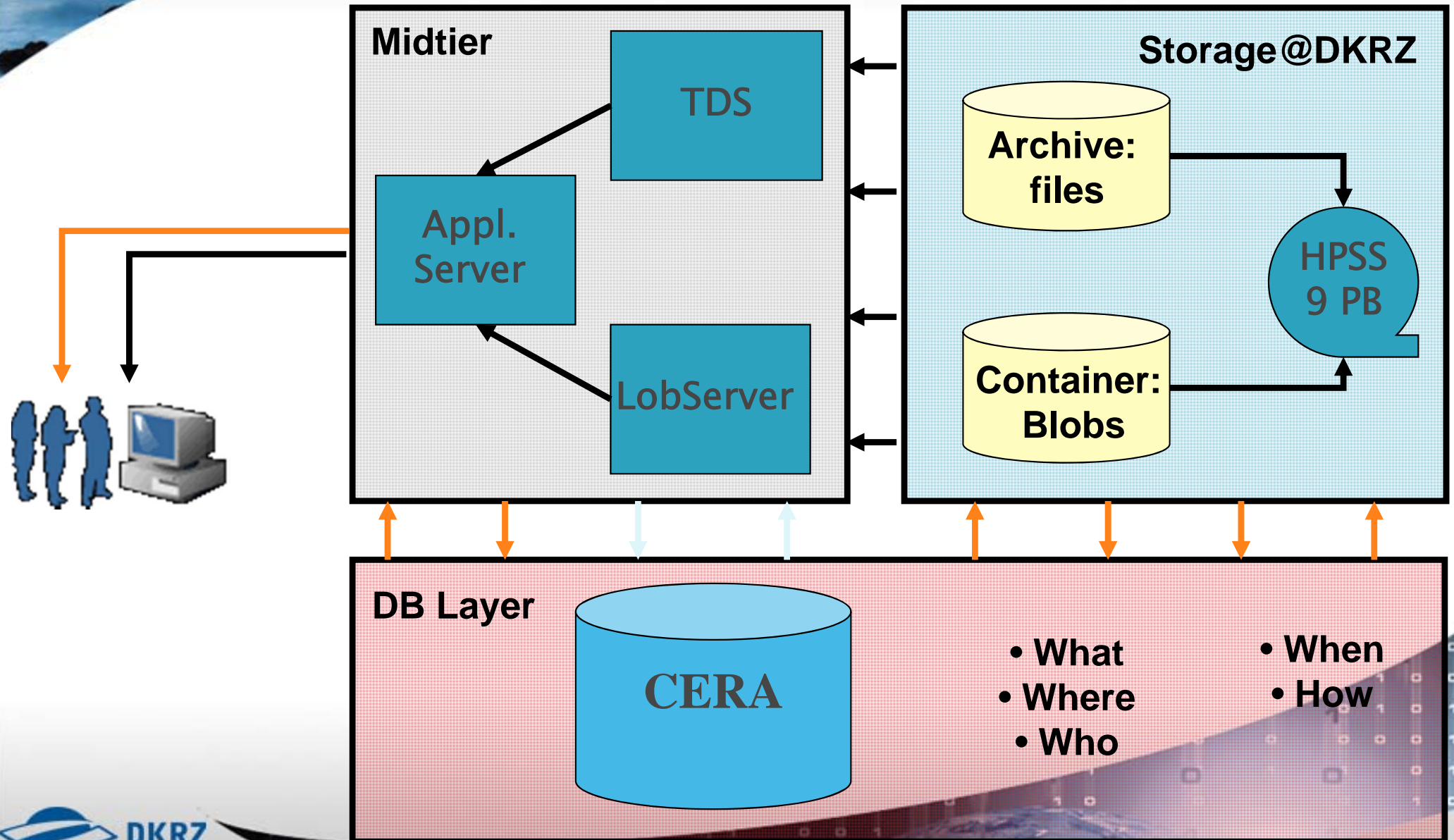
but huge amounts!

Needed: Tape

access (nearline)



# The Data Flows

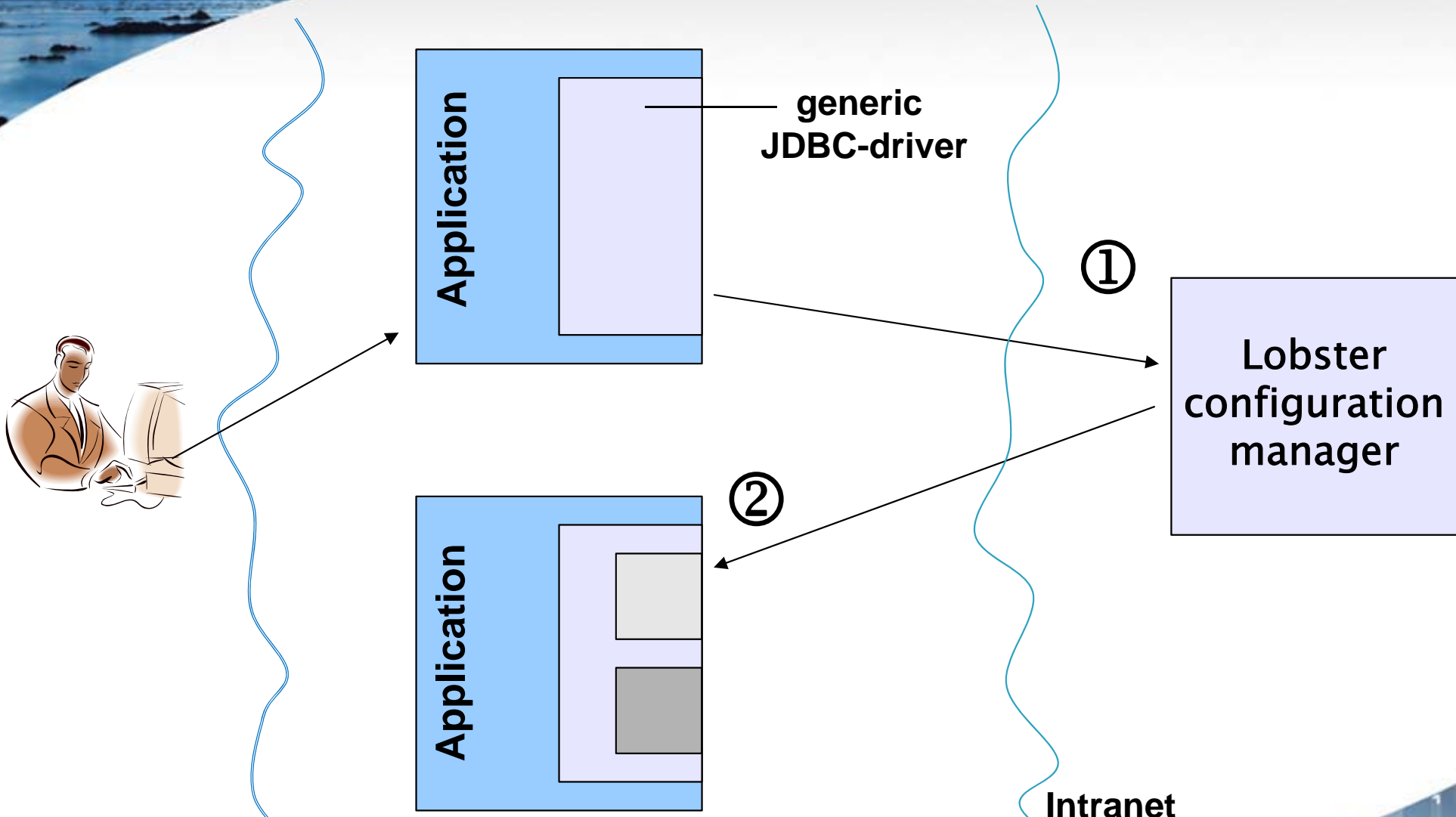




# LObStER: Large Object Storage and Efficient Retrieval

- ❑ 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

# LObStER

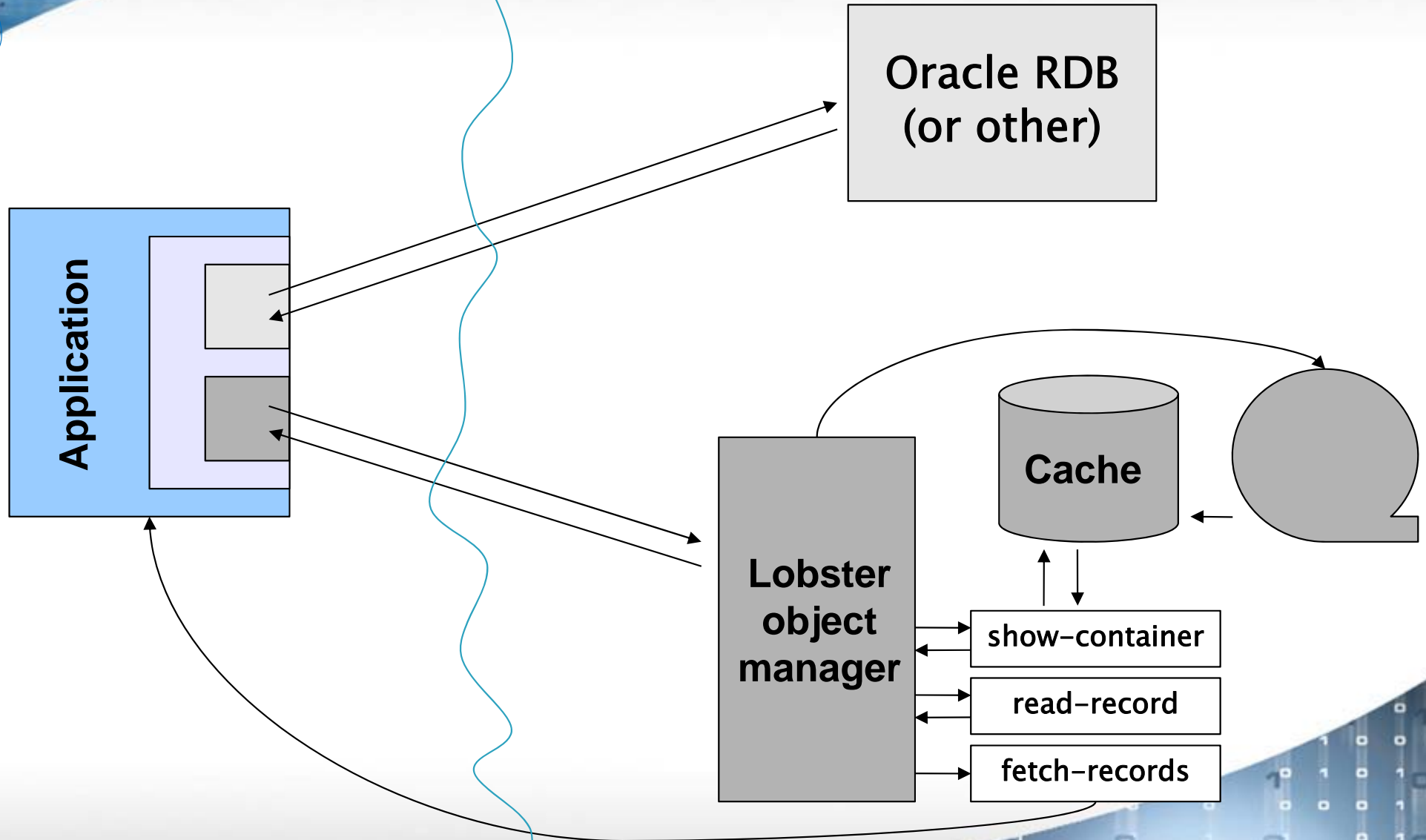


Internet

specific JDBC-drivers loaded

Intranet

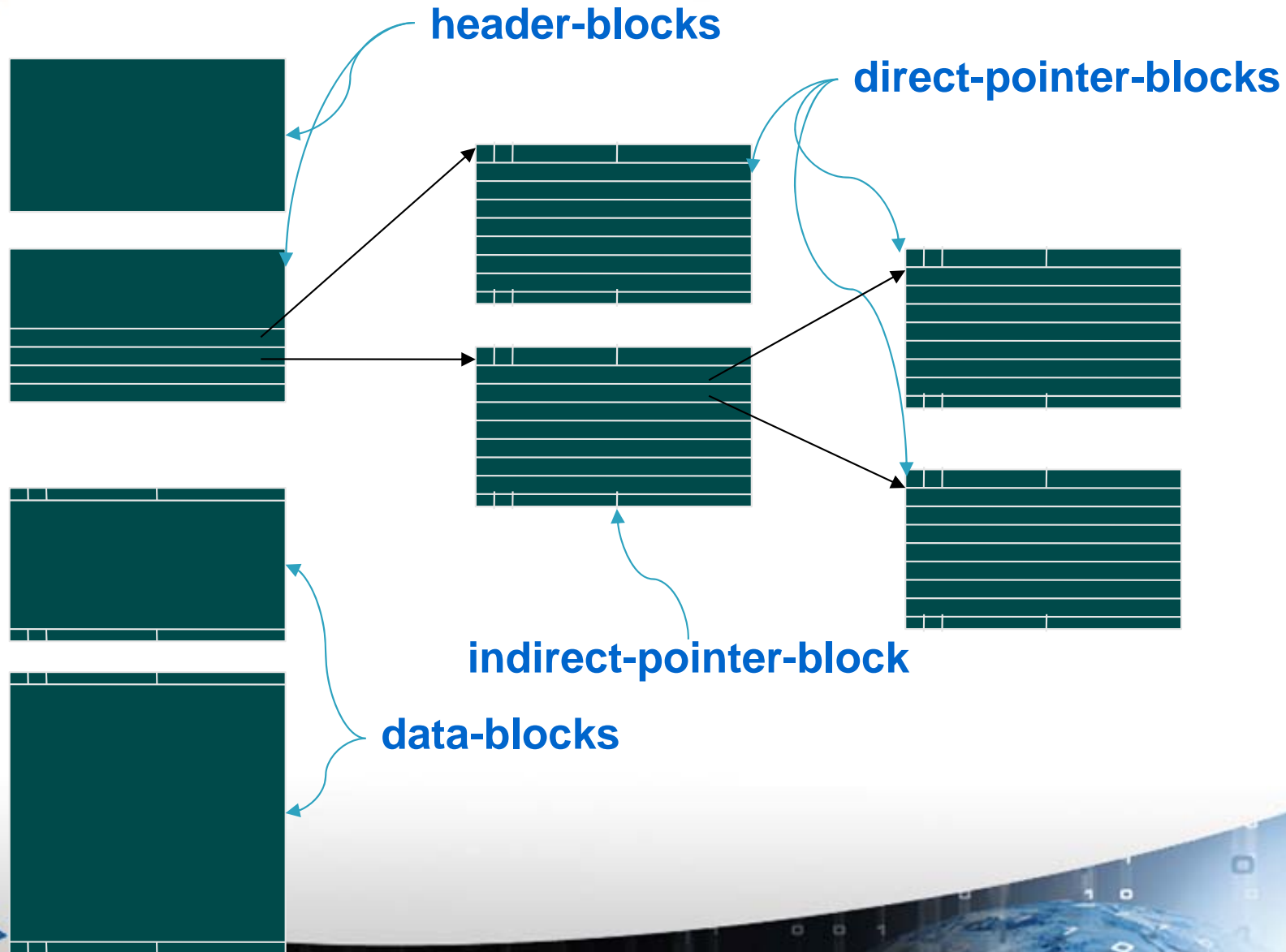
# LObStER



# 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  $\geq 1$  blocks
- ❑ Block sizes  $2^k$ ,  $k \in \{ 8, 9, 10, \dots, 62 \}$

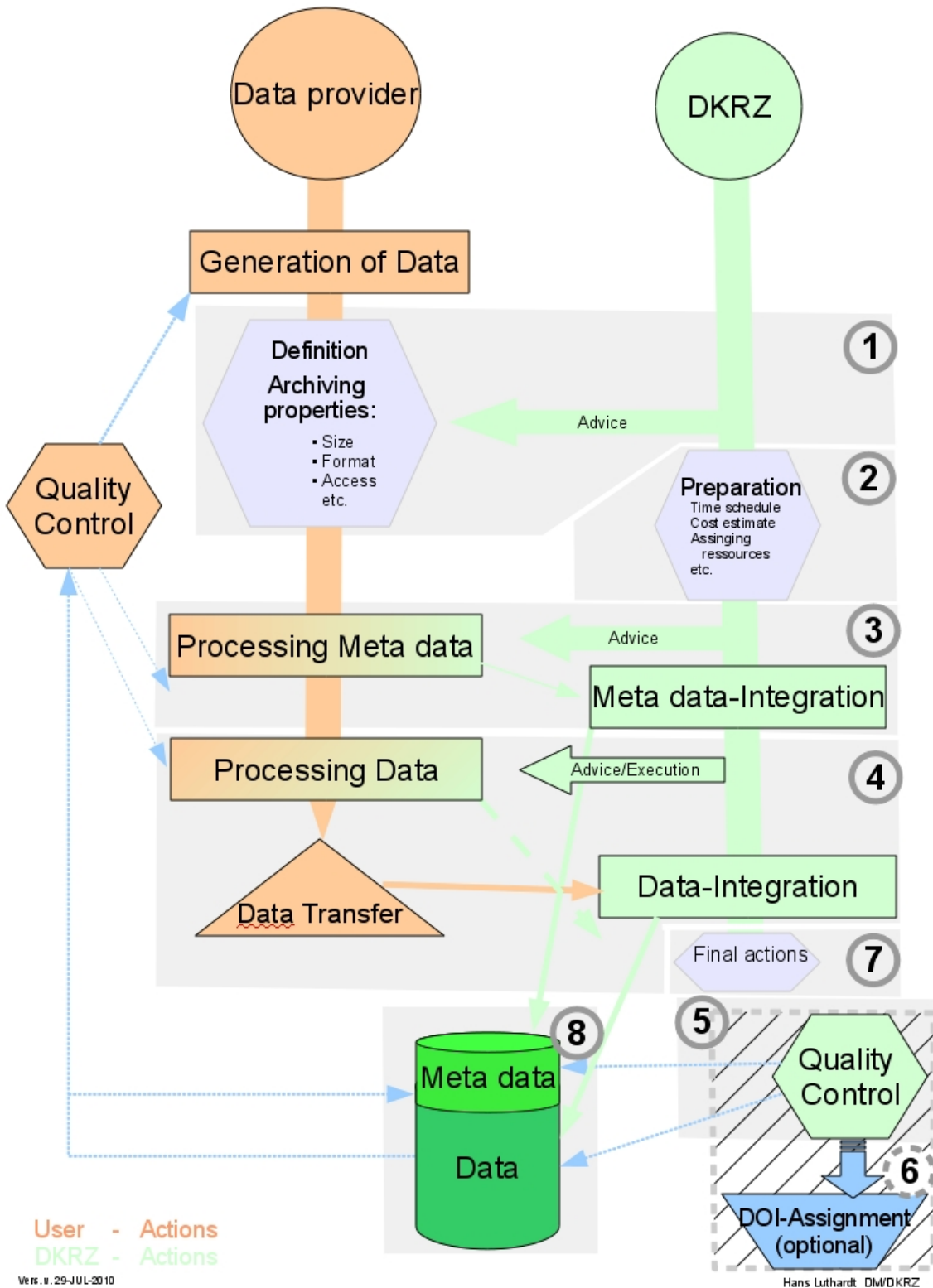
# LObStER: The Data Containers



# Long Term Archiving

Several steps:

- specification & concept
  - filling of metadata & data
  - quality checks & DOI
- 
- LTA for, e.g., EUCLIPSE, MedCLIVAR, combine



# 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



# The WDC-Climate as Publishing Agency of the IDF

International  
DOI Foundation

International DOI  
Foundation

[doi.org](http://doi.org)

Registration  
Agencies

DataCite

[DataCite.org](http://DataCite.org)

National  
Organizations

TIB, BL, ...

[tib-hannover.de](http://tib-hannover.de)

Publisher

WDC, ...

[wdc-climate.de](http://wdc-climate.de)

# The Visibility of LTA Data in Public Catalogues

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

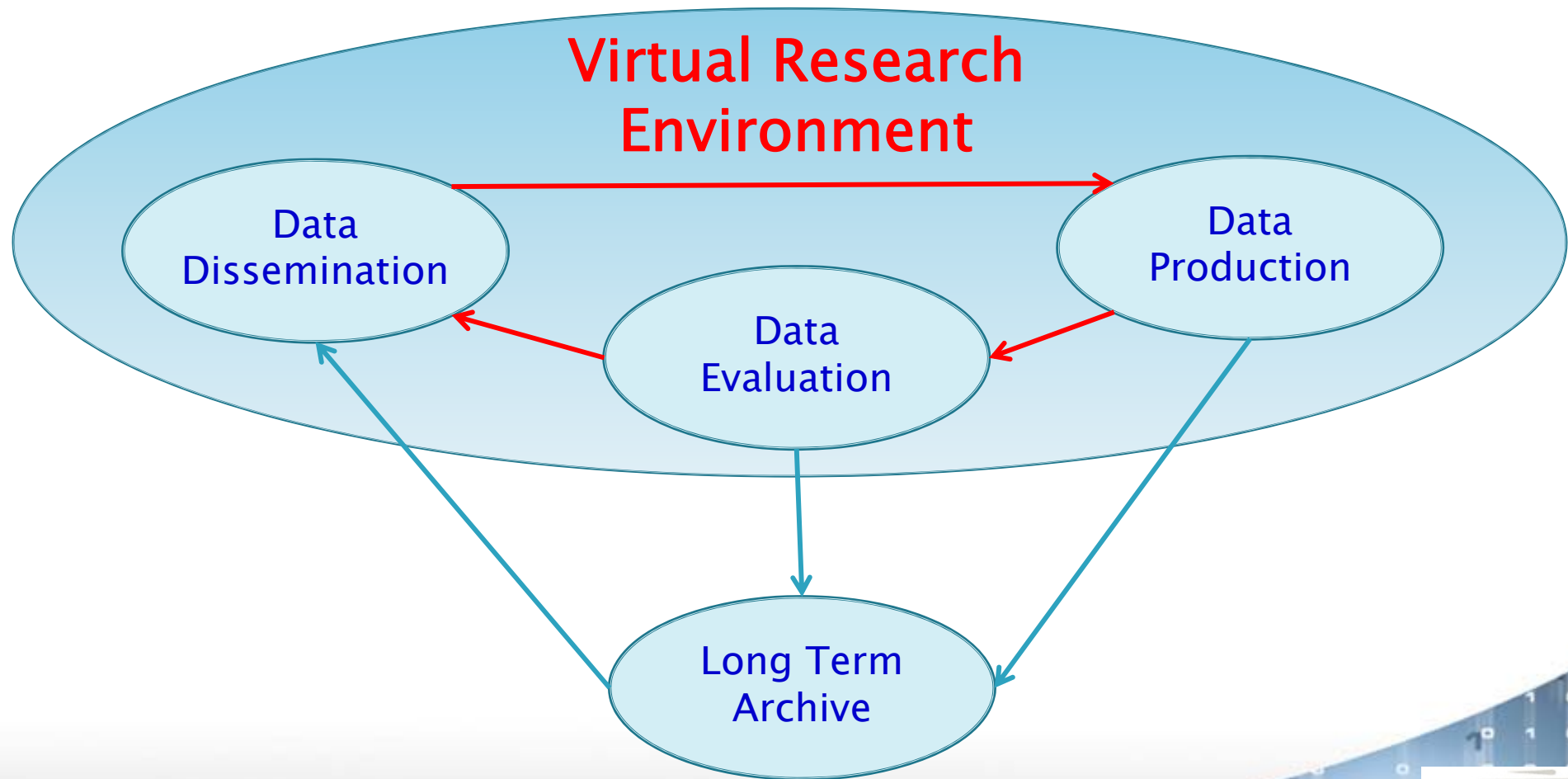
The screenshot shows the search results for 'WDCC' in the TIBORDER catalogue. The search results list several entries, with the third entry circled in red: [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany](#). The details for this entry are shown below, with the DOI and URN circled in red: [10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H) and [urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H0](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H0).

**Titel:** IPCC-AR4 MPI-ECHAM5\_T63L31 MPI-OM\_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany  
**Beteiligt:** Erich Roeckner  
**Körperschaft:** Hamburg  
**Erschienen:** World Data Center for Climate (WDCC)  
**Umfang:** Online-Ressource (3987170720028)  
**Anmerkung:** Mode: Abstract  
StructuralType: Digital  
CreationDate: 2004-05-11  
**Inhalt:** The data represent 6 hourly values with observed anthropogenic forcing in year 2190 of the preindustrial control experiment for the 21th century (years 2001-2100) with all concentrations fixed at their levels of the year 2000. Data Sets with monthly mean values are also available. Technical data to this experiment: The experiment is using ECHAM5.2.02a coupled to MPI-OM Vers. 1.0 GR1.5L40 The output from the model run: hurrikan.dkrz.de:/ut/k/k204076/EXP000/run009 Please note: experiment\_name/acronym was renamed (27-JUN-2005, 20C\_0 changed to 20C\_1)  
**Technische Angaben:** Format: GRIB  
**Links:** doi: [10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H)  
URN: [urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H0](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H0)  
**Bestandsinfo:** [Anzeigen lizenzfrei!](#)  
Anmerkung: Primaerdaten

The screenshot shows the search results for 'WDCC' in the TIBORDER catalogue. The search results list several entries, with the third entry circled in red: [IPCC-AR4 MPI-ECHAM5\\_T63L31 MPI-OM\\_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany](#). The details for this entry are shown below, with the DOI and URN circled in red: [10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H) and [urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H0](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H0).

**Titel:** IPCC-AR4 MPI-ECHAM5\_T63L31 MPI-OM\_GR1.5L40 20C3M run no.1: atmosphere 6 HOUR values MPImet/MaD Germany  
**Beteiligt:** Erich Roeckner  
**Körperschaft:** Hamburg  
**Erschienen:** World Data Center for Climate (WDCC)  
**Umfang:** Online-Ressource (3987170720028)  
**Anmerkung:** Mode: Abstract  
StructuralType: Digital  
CreationDate: 2004-05-11  
**Inhalt:** The data represent 6 hourly values with observed anthropogenic forcing in year 2190 of the preindustrial control experiment for the 21th century (years 2001-2100) with all concentrations fixed at their levels of the year 2000. Data Sets with monthly mean values are also available. Technical data to this experiment: The experiment is using ECHAM5.2.02a coupled to MPI-OM Vers. 1.0 GR1.5L40 The output from the model run: hurrikan.dkrz.de:/ut/k/k204076/EXP000/run009 Please note: experiment\_name/acronym was renamed (27-JUN-2005, 20C\_0 changed to 20C\_1)  
**Technische Angaben:** Format: GRIB  
**Links:** doi: [10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H)  
URN: [urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31\\_OM-GR1.5L40\\_20C\\_1\\_6H0](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH5-T63L31_OM-GR1.5L40_20C_1_6H0)  
**Bestandsinfo:** [Anzeigen lizenzfrei!](#)  
Anmerkung: Primaerdaten

# The Data Life Cycle Management





THANK YOU,

QUESTIONS?