

**Acquired  
Analysed  
Archived**

## **Climate Data for Our Future**

Prof. Dr. Andreas Hense  
[andreas.hense@h-brs.de](mailto:andreas.hense@h-brs.de)

DACH Conference Bonn  
September 21, 2010



Hochschule  
Bonn-Rhein-Sieg



# Project Partners



Bonn-Rhine-Sieg University  
oAS, Computer Science,  
Sankt Augustin



**Prof. Dr. Andreas V. Hense**

Professor for Business  
Information Systems

**Project management &  
software development**



Bonn University,  
Meteorological Institute  
Bonn



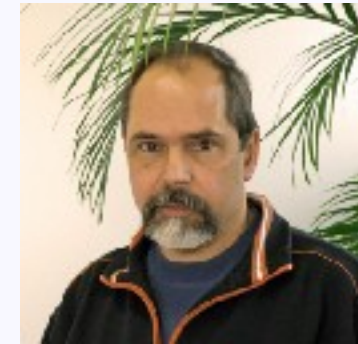
**Prof. Dr. Andreas N. Hense**

Professor for Climate  
dynamics

**Experimental data &  
routines for scientific QA**



Deutsches Klimarechen-  
zentrum GmbH  
Hamburg



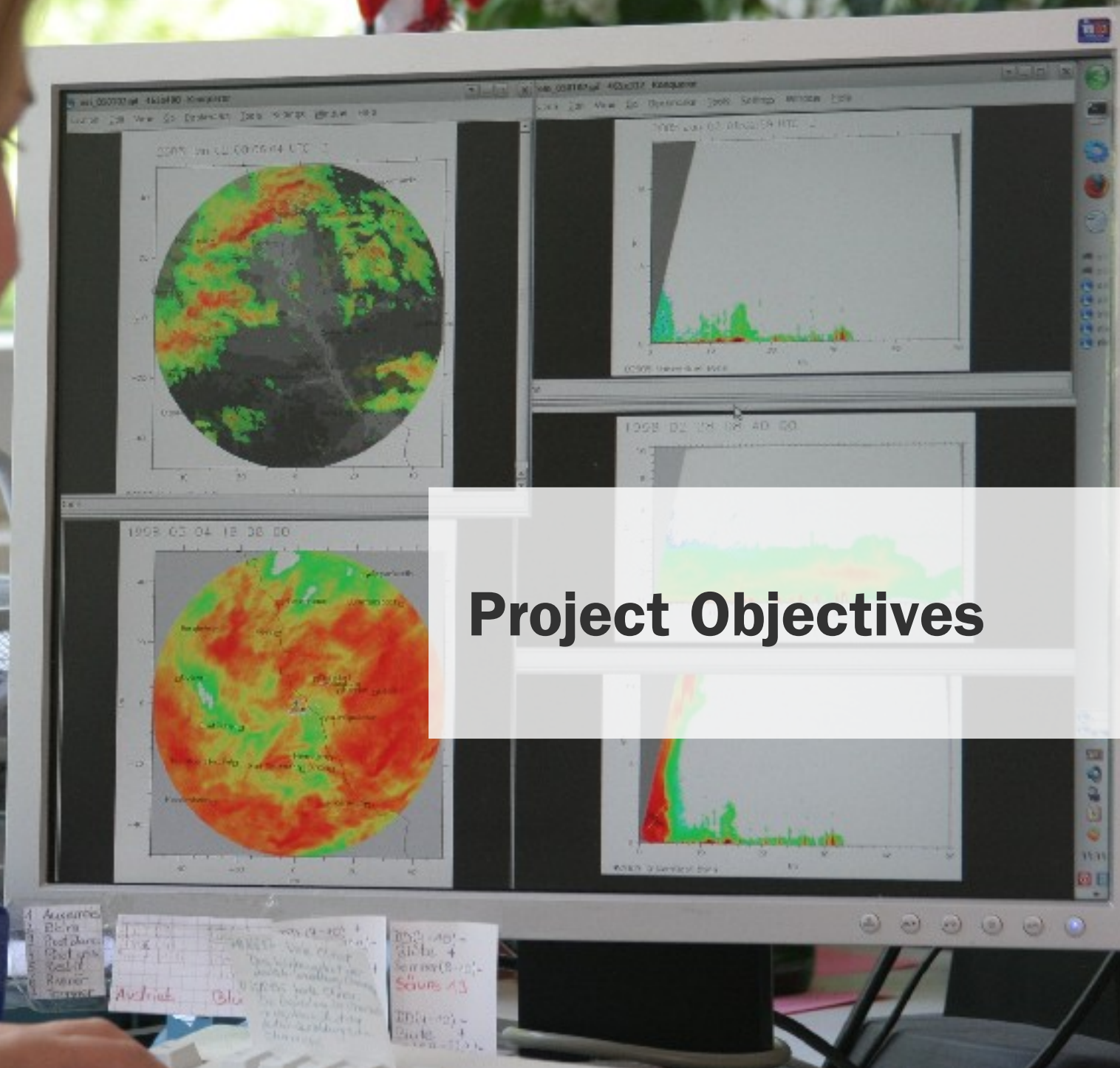
**Dr. Michael Lautenschlager**

Head of Data Mgmt,  
Director WDC for Climate  
(WDCC)

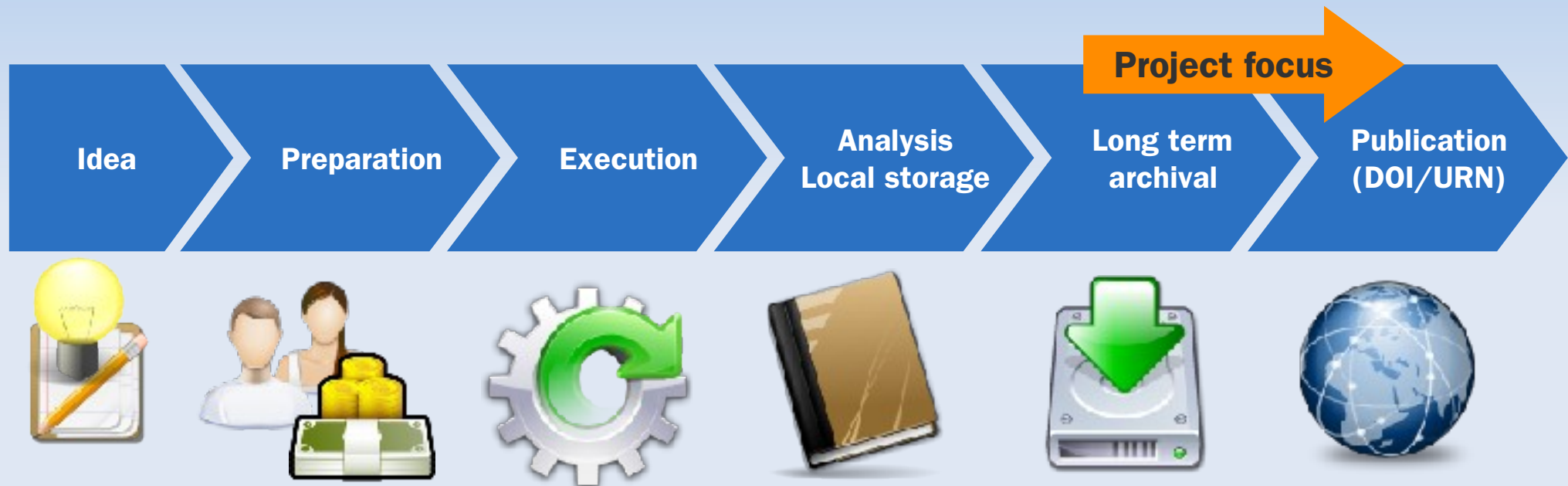
**Process definitions,  
routines for technical QA,  
hosting**

# Agenda

- Project Objectives
- Meteorological Background
- The World Data Center for Climate (WDCC)
- The Publication System Atarrabi



# Location in the Scientific Process



# Data Publication in Research

## Problem:

- Publication and citation have always been common practice for scientific articles.
- Scientific articles are often based on data.
- To check the results of an article or to do further research the data are necessary.

## Solution:

- Publish the article **AND** the data.

# Aspects of Data Publication

- **Storage location** – The volume of data can be huge (e.g. meteorological data). Who can reliably store the data and assure long term availability and fast access?
- **Formats** – There can be various formats to represent data. Which (meta) data format is the most commonly used?
- **Exposition/Registry** – It is not sufficient to save data "somewhere" on the web. Scientists have to notice the existence of data. What is the best way to expose data to search engines? Are there well known (domain specific) catalogues where data can be registered?
- **Quality** – Not all data are qualified for publication. What are the minimum requirements? What are (scientific and technical) quality assurance procedures?
- **Stability** – Can data be changed after publication? How are new versions published?
- **Identifier** – How can data uniformly be referenced? Are there any standards?

# Storage Sites

Experiment  
Analyze  
Collaborate

Publish

Expose

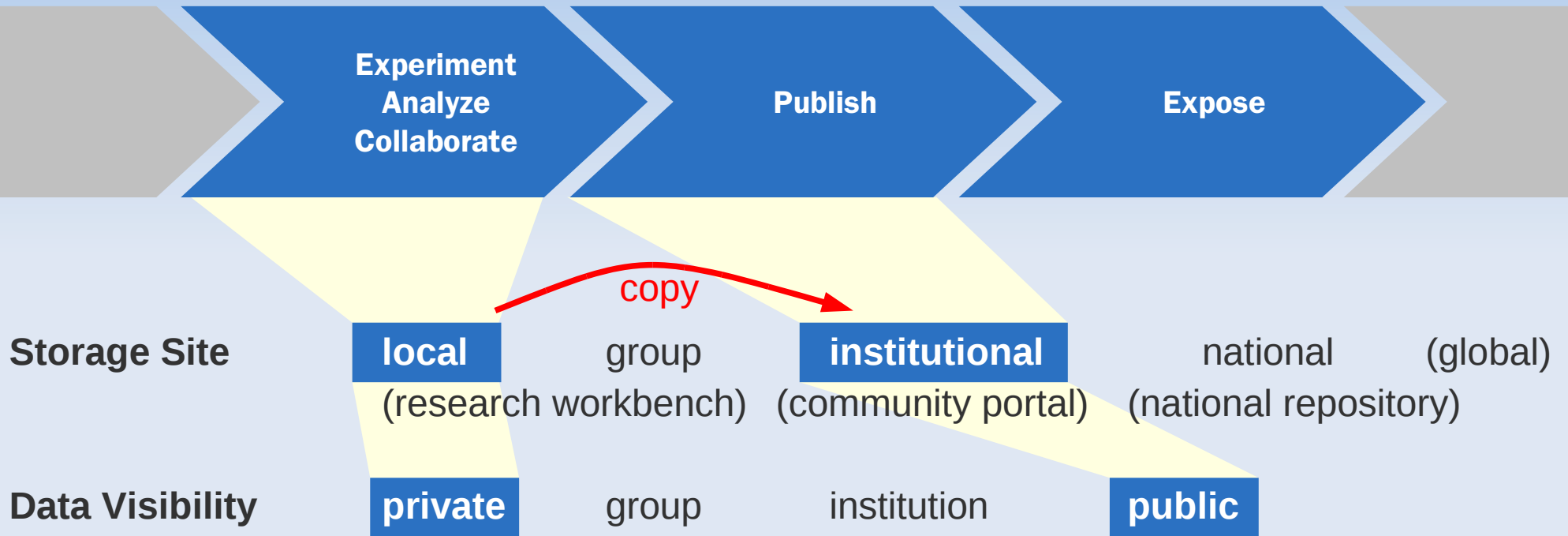
**Storage Site**

local (research workbench)    group (community portal)    institutional (national repository)    national (global)

**Data Visibility**

private    group    institution    public

# A Common Scenario



# The National Storage Solution

Experiment  
Analyze  
Collaborate

Publish

Expose

Storage Site

local

group

institutional

**national**

(global)

(research workbench)

(community portal)

(national repository)

Data Visibility

private

group

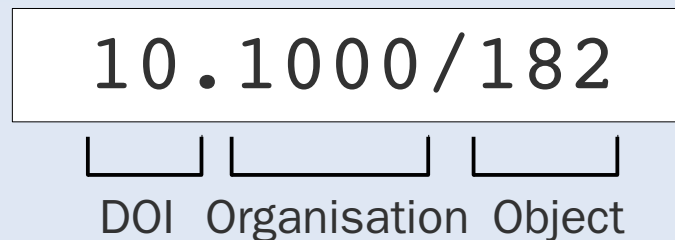
institution

public

# Persistent Identifiers: Digital Object Identifier (DOI)

- Idea
  - Uncouple the identifier and the object location
  - Very similar to DNS and other directories

- DOI structure



- Resolution of a DOI

- visit `http://dx.doi.org/10.1000/182`
- in most cases you will get a landing page showing metadata about the object and a download link

# Project Objectives

- **Definition of a standard procedure** for publication of observational data including documentation of quality assurance actions.
- **Development of a web-based software system** that leads the researcher through metadata entering as well as assists the publication agent to finalize the process.
- **Integration of the software system** into the existing central data repository for meteorology (World Data Center for Climate (WDCC)).
- **Generalisation** of the defined process for other environmental sciences.

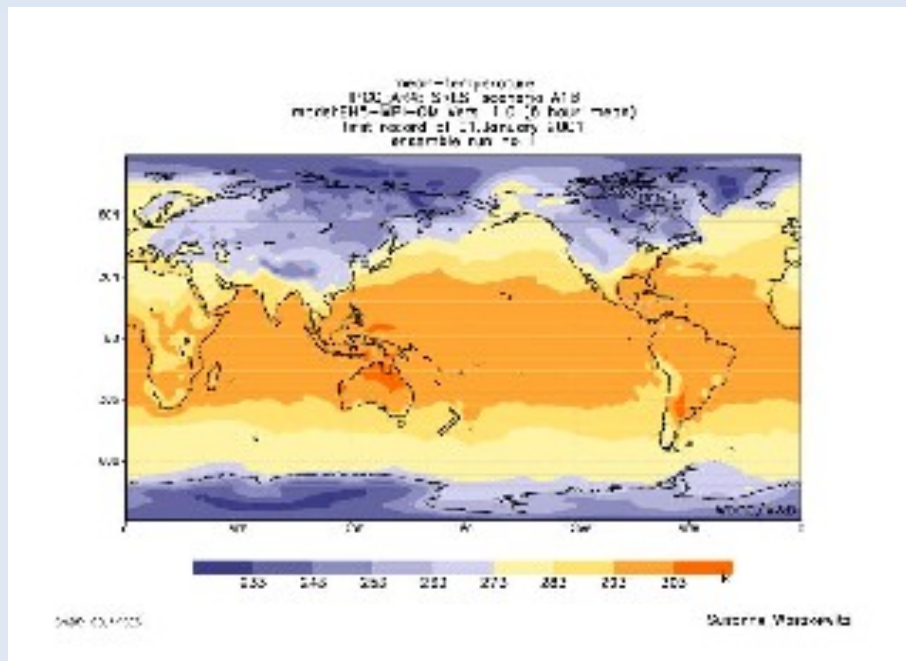


# **Meteorological Data**

# Meteorological Data Sources

## Climate Simulations

Data from Models: grid-oriented, 2-3 spatial- & 1 time-dimensions & 1 variable dimension & 1 sampling/probability dimension



**Large amount, but simple structure**

## Experimental data

Empirical Data: various structures in time and space

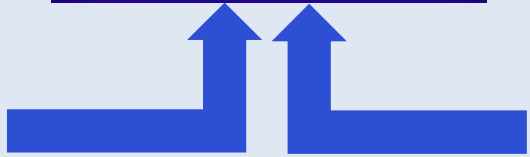
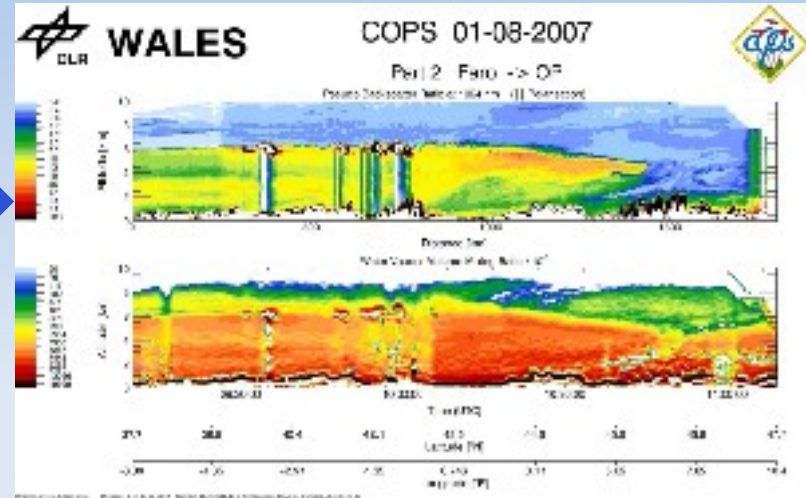
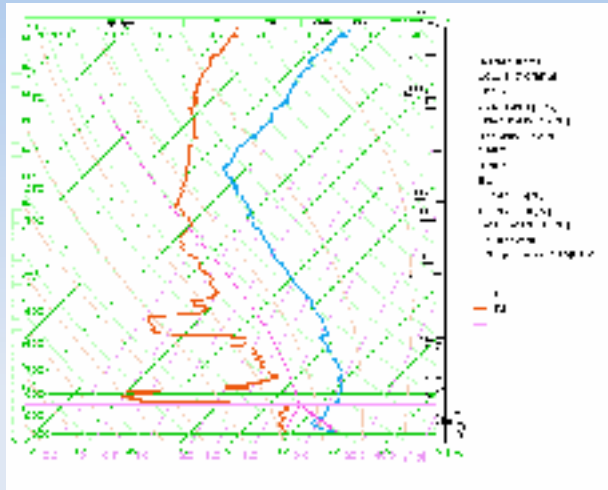


Airborne lidar platforms: DLR Falcon (V) and SAFIRE Falcon (H)  
2 mobile Doppler On Wheels (S)

Legend:			
⊙	Water Vapor Lidar	■	C-Band Polarization Radar
⊕	Temperature Lidar	⊗	Precipitation Radar (other)
⊗	Wind Lidar	⊙	Micro-Rain-Radar
⊗	Aerosol/Raman Lidar	⊗	Wind Profiler
⊙	Cellometer	⊗	Wind-Temperature-Radar
⊙	Microwave Radiometer	⊗	Energy balance station
⊙	FTIR Radiometer	⊙	Sodar
⊗	Cloud Radar	⊙	GPS receiver
⊗	Radiosonde station		

**Not so big amount but much more complex**

# Weather Experiments



# Meteorological Data

- We can distinguish
  - experimental (observational) data (small amount, heterogenous) and
  - climate simulation data (huge amount, simple structure)

	Experimental data	Climate simulation data
Storage location	WDCC (work in progress)	WDCC
Formats	NetCDF (with restrictions, work in progress)	NetCDF
Exposition/Registry	WDCC (work in progress), TIBORDER (work in progress)	CERA catalogue, TIBORDER
Quality	<b>Project focus</b>	QA more technical than scientific
Stability	No changes to primary data allowed, changes to metadata are restricted	
Identifier	Digital Object Identifier (DOI), Uniform Resource Name (URN)	

# Relevant Meteorological Projects

## Experimental data:

- As part of the "Quantitative Precipitation Forecast" (DFG SPP1167):
  - Convective and Orographically-induced Precipitation Study (**COPS**), measurements in the Black Forest in 2007, <http://www.cops2007.de>.
  - General Observation Period (**GOP**), extended measurements in Central Europe in 2007, <http://gop.meteo.uni-koeln.de/gop/doku.php>.
  - All participants have agreed to publish the data to support further research.

# Relevant Meteorological Projects

## Climate simulation data:

- Coupled Model Intercomparison Project Phase 5 (**CMIP5**):
  - Standard experimental protocol for studying the output of coupled ocean-atmosphere general circulation models (GCMs)
  - Provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access.
  - Addresses outstanding scientific questions that arose as part of the IPCC AR4 (the Intergovernmental Panel on Climate Change 4th Assessment Report) process.
  - Provides estimates of future climate change that will be useful to those considering its possible consequences.



**The World Data Center  
for Climate (WDCC)**

# Long term archival

- The WDCC in Hamburg, Germany operates large databases (60 PB) for the long-term archival of data from climate simulation and weather experiments.
- WDCC is controlled by "Deutsches Klimarechenzentrum" (German climate data processing center)
- Data production: 50 PB/year
- Limit for mass storage archive: 10 PB/year
  - Data with expiration date
- Limit for long-term data archive: 1 PB/year
  - Data without expiration date
- Currently only a very small amount of data is published (approx. 1,5 TB), this is expected to grow significantly.

# WDCC equipment



- HPC Cluster ("blizzard")
  - IBM p575 "Power6" cluster
  - water cooled, 16 dual core CPUs per node, total: 264 nodes, 8448 cores
  - Total system peak performance: 158 TeraFlops/s
  - Top500: Rank 27 in 06/09
  - 20 TeraByte memory
  - 3 PetaByte GPFS file system (additional 3 PetaByte in 2011)

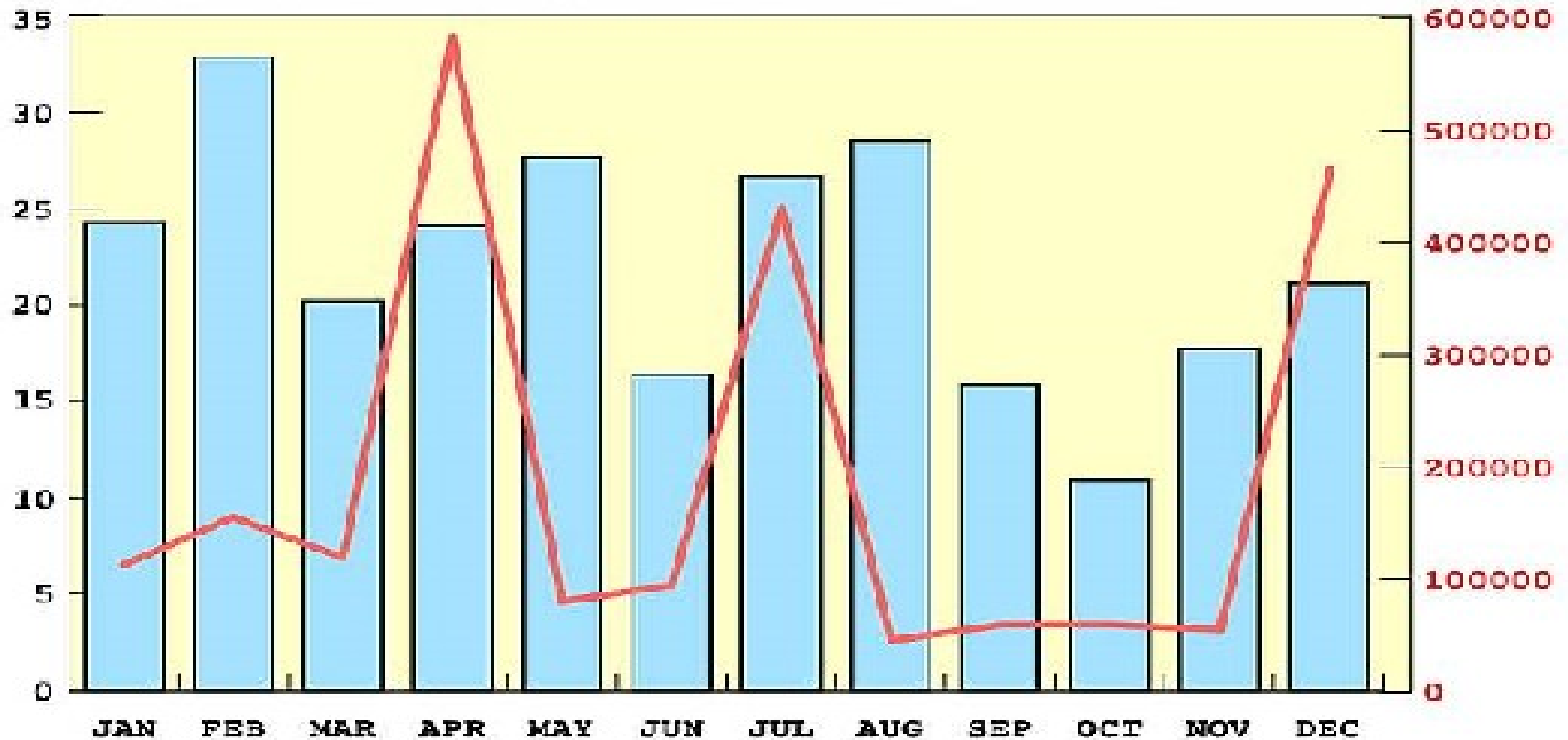
# HLRE2 Data Archive: HPSS

- 6 Sun StorageTek SL8500 tape libraries
  - 10 000 media slots per library, 8 robots per library, 73 tape drives
  - total capacity: 60 PetaByte.
- projected fill rate: 10 PetaByte/year



# Strong demand for scientific data

Cera Download 2008 in terabyte (left, blue) and counts (right, red) Metadata requests are not counted



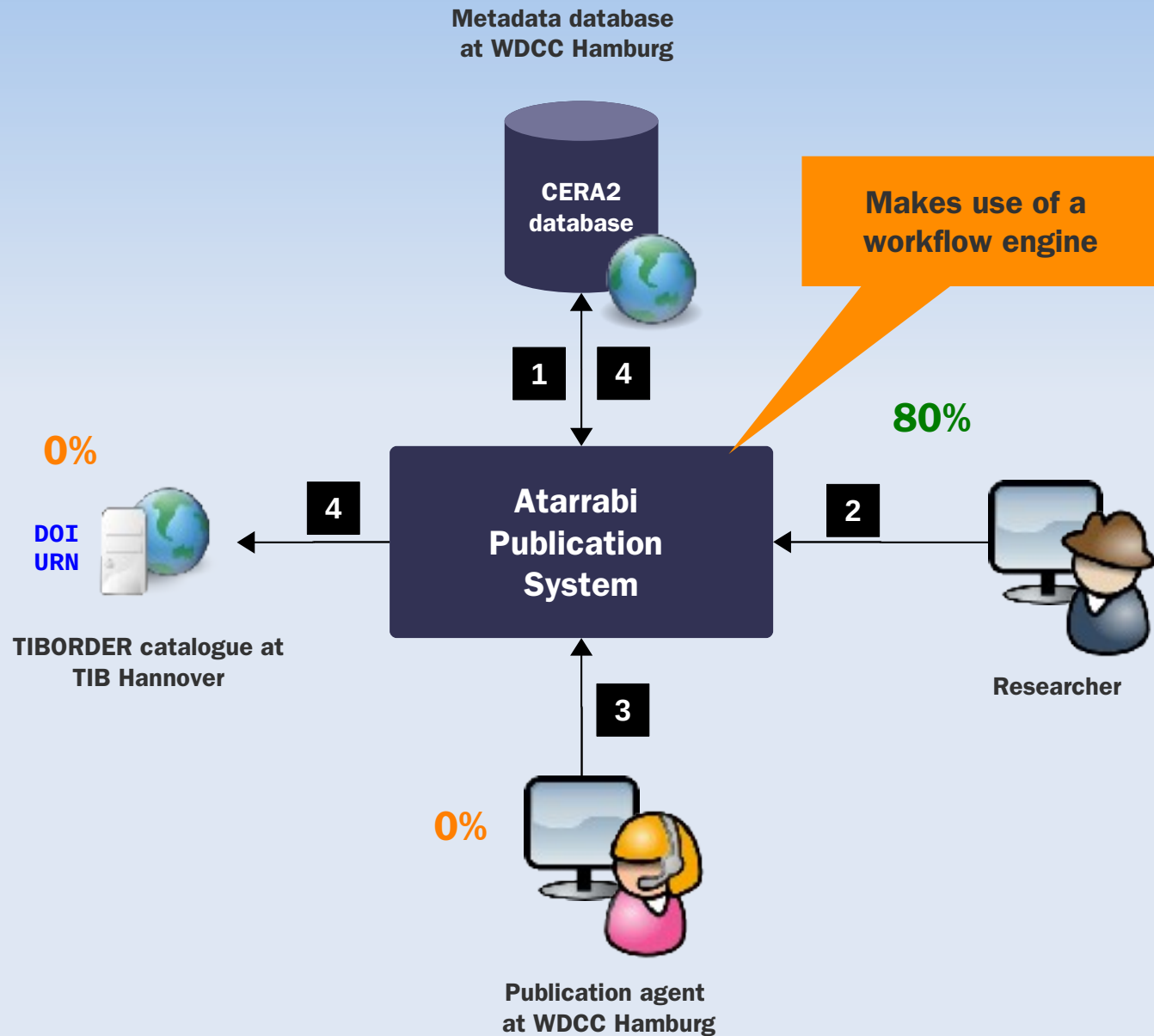


**The Publication  
System *Atarrabi***

# Publication via DOI and URN

- Experiments of particular importance can be published with a DOI and a URN.
- The decision making will take place at WDCC.
- DOI and URN registration by "TIB Hannover".
- Data is double-checked before publication (scientific and technical quality assurance).
- Most important is a complete and correct metadata record.

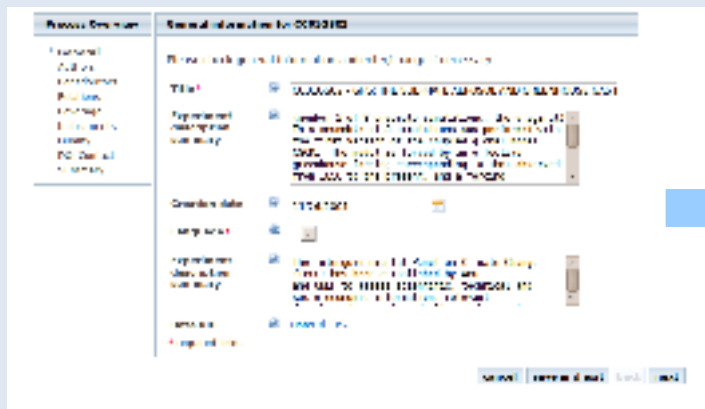
# System Context



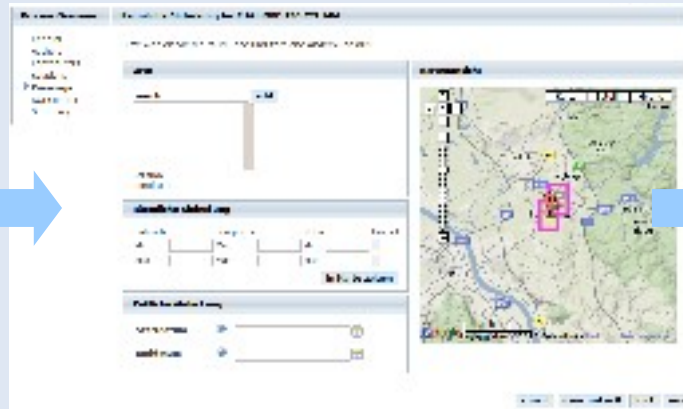
# Wizard-based Metadata Entering

- Divide metadata fields into several logical units.
- The user can leave the wizard at any time and return later to continue.

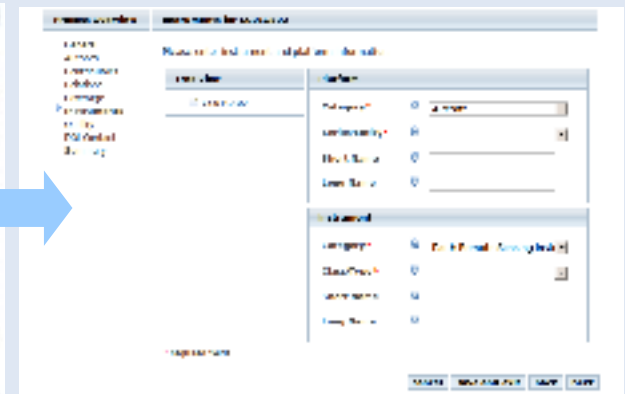
General



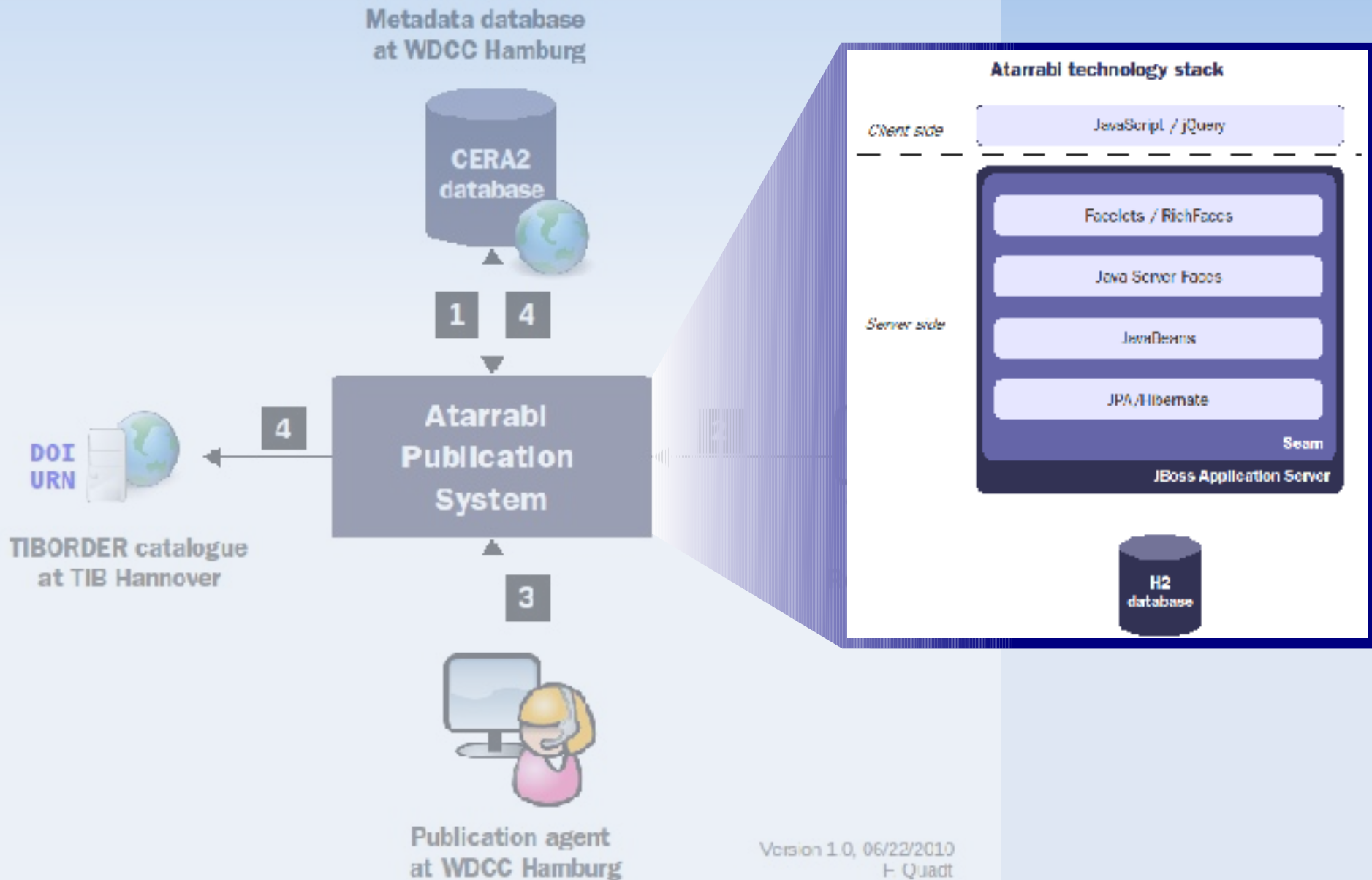
Spatial and temporal coverage



Instruments



# Technology Stack





**Acquired  
Analysed  
Archived**

## **Climate Data for Our Future**

Prof. Dr. Andreas Hense  
[andreas.hense@h-brs.de](mailto:andreas.hense@h-brs.de)

visit us:  
[umwelt.wikidora.com](http://umwelt.wikidora.com)



Hochschule  
Bonn-Rhein-Sieg

