Data Publication in Environmental Science

Project funded by DFG

http://umwelt.wikidora.com

Andreas Hense
andreas.hense@brsu.de

Florian Quadt
florian.quadt@brsu.de

Bonn-Rhine-Sieg University of Applied Sciences

MAX-PLANCK-GESELLSCHAFT

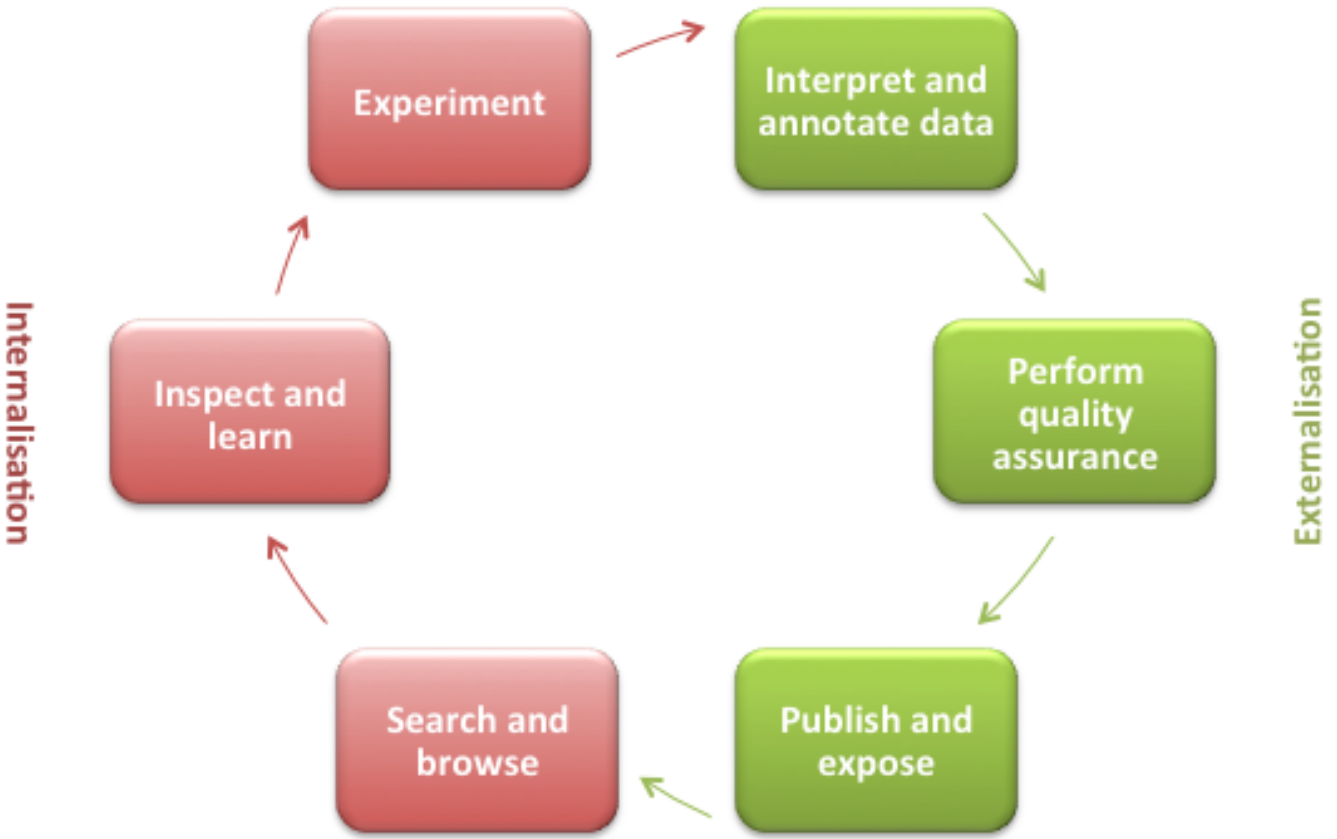
CNI Spring 2012 Member Meeting
April 2-3, Baltimore

Topics of this Talk

- Why is data publication important?
- Which data do we deal with?
- Where are our data stored?
- How do we publish data?

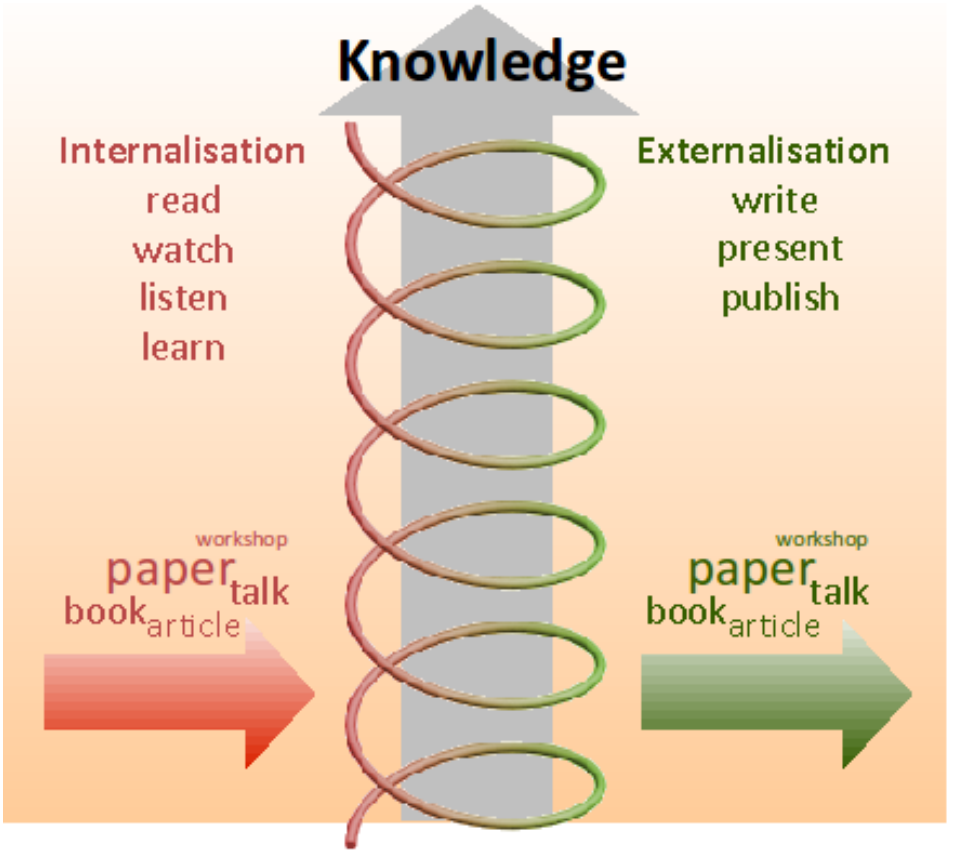
Data Publication

The Data Publication Cycle



[Hense and Quadt, 2011]

The Knowledge Spiral



[Hense and Quadt, 2011]

Differences between Text and Data Publication

	Text	Data
File formats	few widely accepted file formats	many, disciplines, tools
Contents	optimised for human reader	sometimes difficult to inspect
SQA	peer-review long tradition	peer-review data??
FQA and TQA	typesetting, etc.: easy	partial automation possible
Browse&Search	metadata plus full text search	often restricted to metadata
Storage site	single file, stored in a repository	links to huge files, data curation.

[Treloar et al., 2007]

[Klump, 2008]

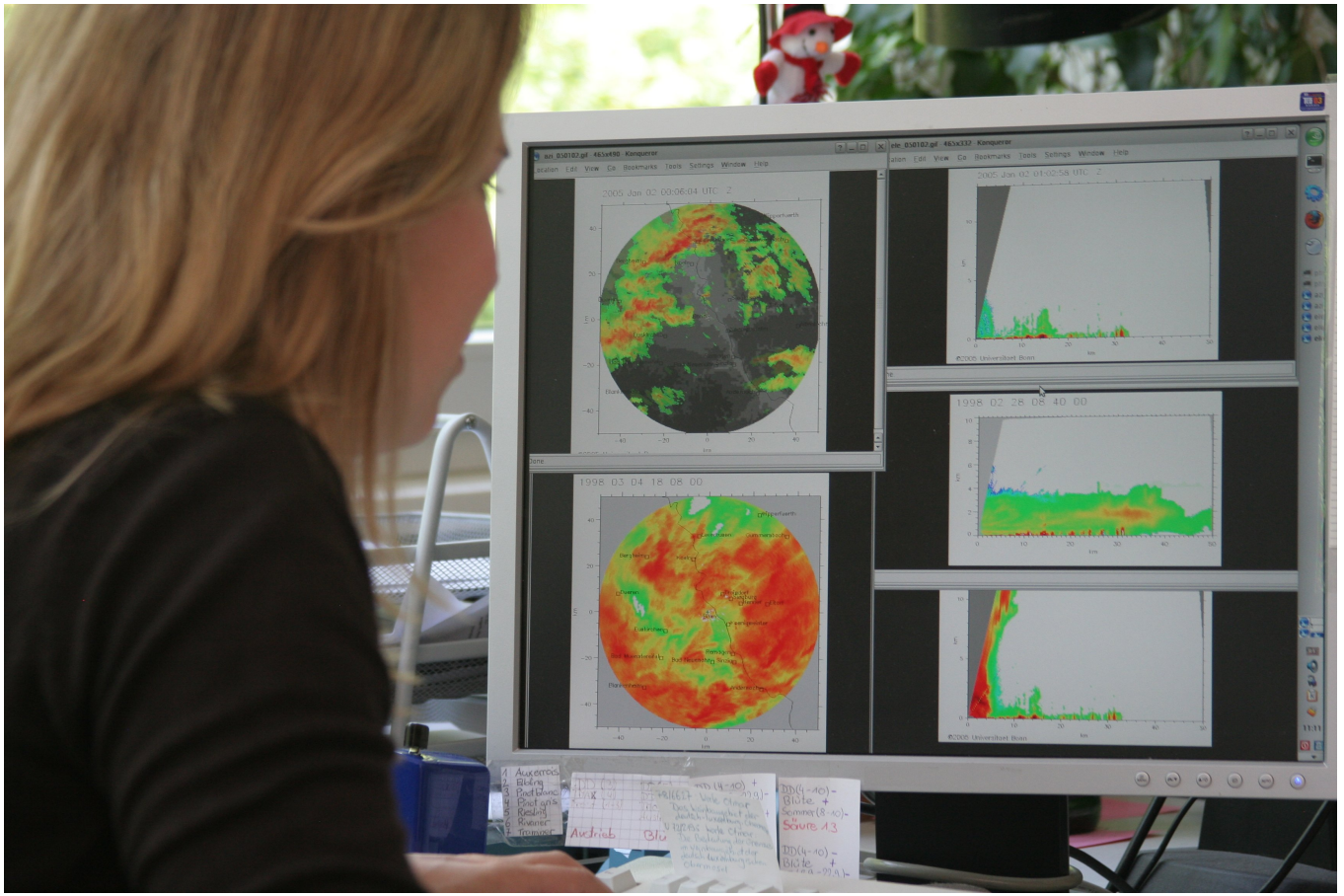
How to Acquire High Quality Data?

Reputation If a repository has the reputation of accepting only high quality research data and the visibility of its publications is high, a publication in such a repository is of great value for the researcher.

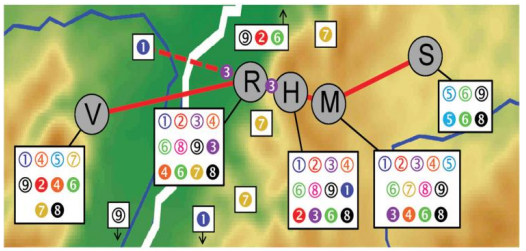
Reliability If the organisation which is running the repository is known to be well-financed and has a long tradition, researchers will believe that their data are stored safely and for the long term. (Scientists love DOIs)

Process If the submission and quality assurance process is well-documented and user-friendly, it will engender trust and lower the workload for potential submitters. [Hense and Quadt, 2011]

Environmental Data



Observational Data



Airborne lidar platforms: DLR Falcon (①③) and SAFIRE Falcon (①)
2 mobile Doppler-On-Wheels (❷❷)

Legend:

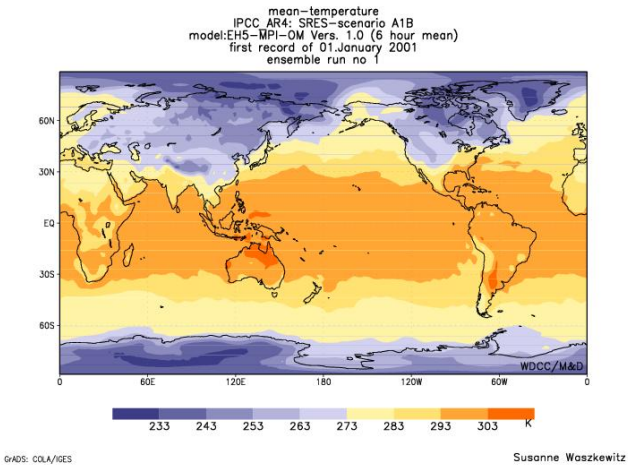
① Water Vapor Lidar	❶ C-Band Polarization Radar
② Temperature Lidar	❷ Precipitation Radar (other)
③ Wind Lidar	❸ Micro-Rain-Radar
④ Aerosol Raman Lidar	❹ Wind Profiler
⑤ Ceilometer	❺ Wind-Temperature-Radar
⑥ Microwave Radiometer	❻ Energy balance station
⑦ FTIR Radiometer	❼ Sodar
⑧ Cloud Radar	❽ GPS receiver
⑨ Radiosonde station	

Characteristics Small amount, heterogeneous.

Exposition/Registry WDCC, GetInfo TIB (DOI).

Provenance Convective and Orographically-induced Precipitation Study [COPS, 2007], General Observation Period [GOP, 2007].

Simulation Data

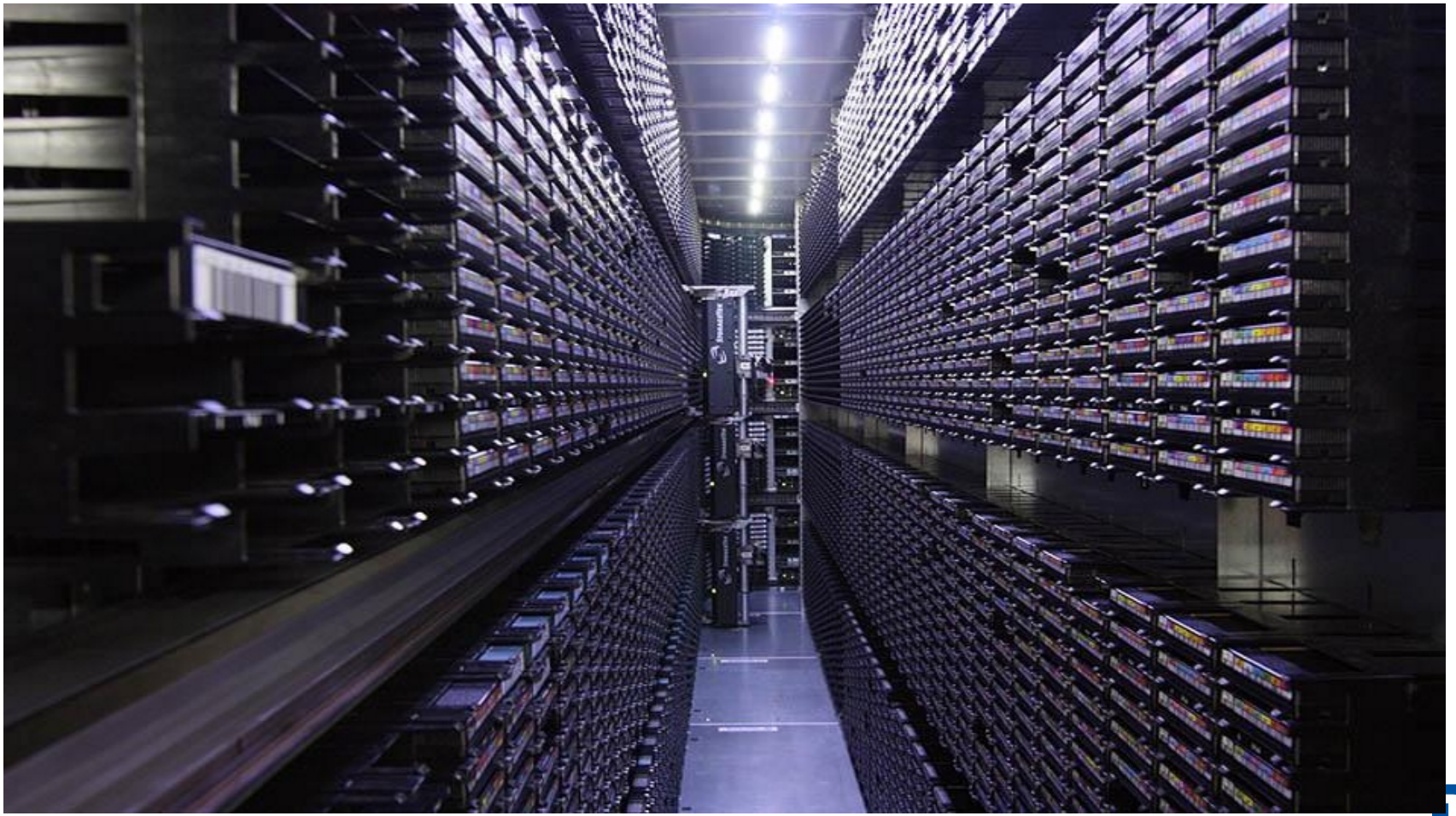


Characteristics Huge amount, homogeneous.

Exposition/Registry WDCC, GetInfo TIB (DOI).

Provenance Coupled Model Intercomparison Project Phase 5 (CMIP5) [LLNL, 2010]

Data Repositories



. . . and examples

	domain-specific	general
institutional	unusual	[ARROW, 2008]
national	[AuScope, 2009] [IMOS, 2009]	[TIB, 2010] [ANDS, 2009]
global	[DKRZ, 2010]	[DataCite, 2010]

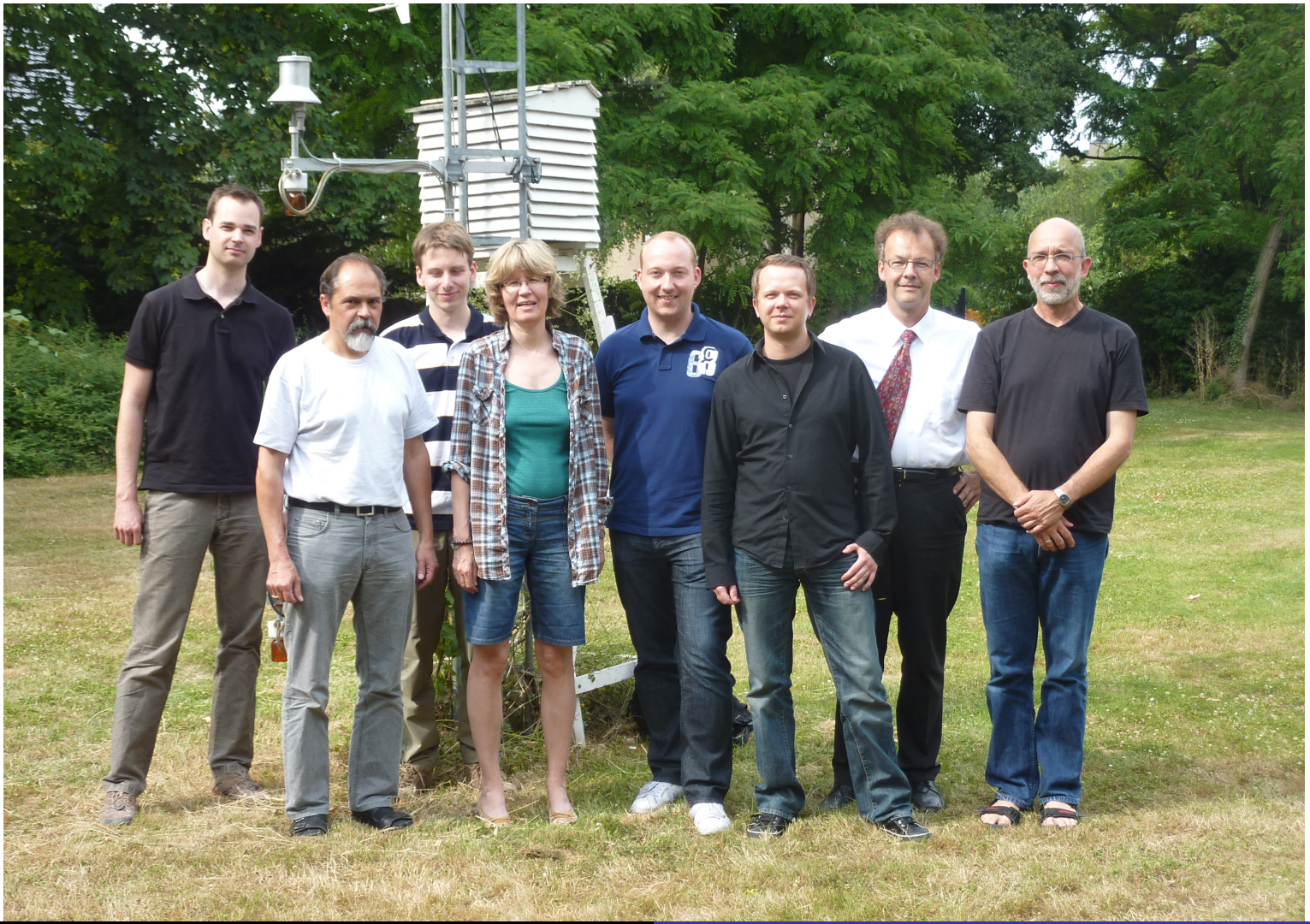
The World Data Center for Climate (WDCC)

Operator Deutsches Klimarechenzentrum GmbH: The German High Performance Computing Centre for Climate- and Earth System Research [DKRZ, 2010]

Long-Term Archival Storage 60 PB: 6 Sun StorageTek SL8500 tape libraries, 10 000 media slots per library, 8 robots per library.

Annual Data Production Total: 50 PB, Limit for mass storage archive (expiration date): 10 PB, Limit for long-term data archive: 1 PB

Publication of Environmental Data



About the Project

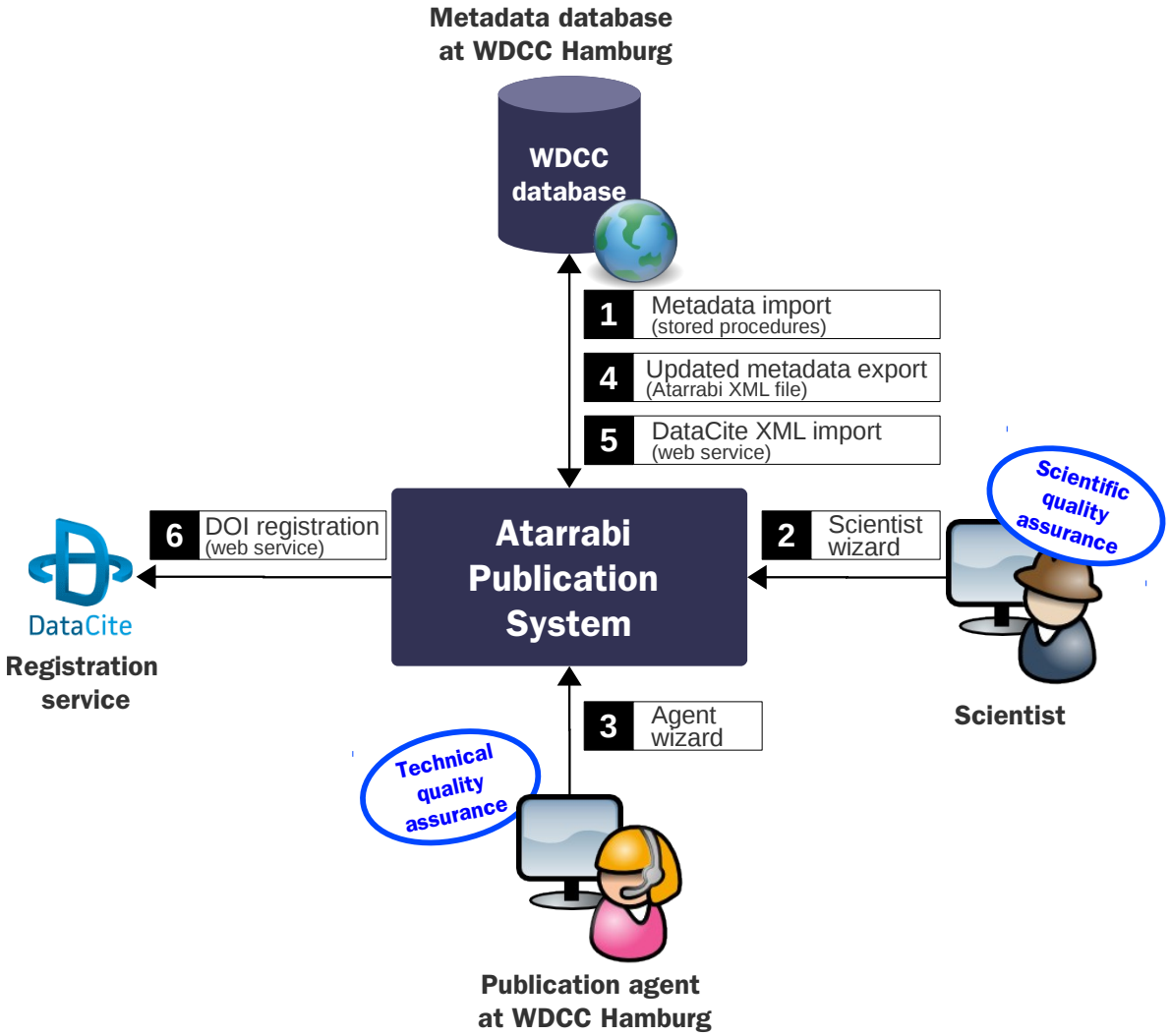
Funding DFG, April 2009 – March 2012

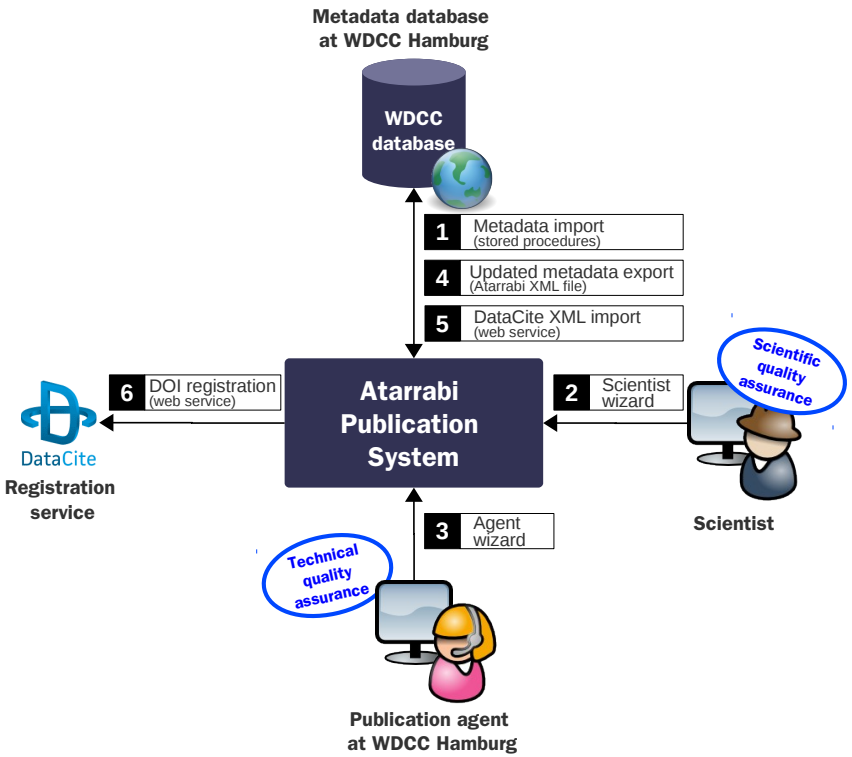
Partners

- Bonn-Rhine-Sieg University oAS
- Bonn University, Meteorological Institute
- German Climate Computing Centre (DKRZ)

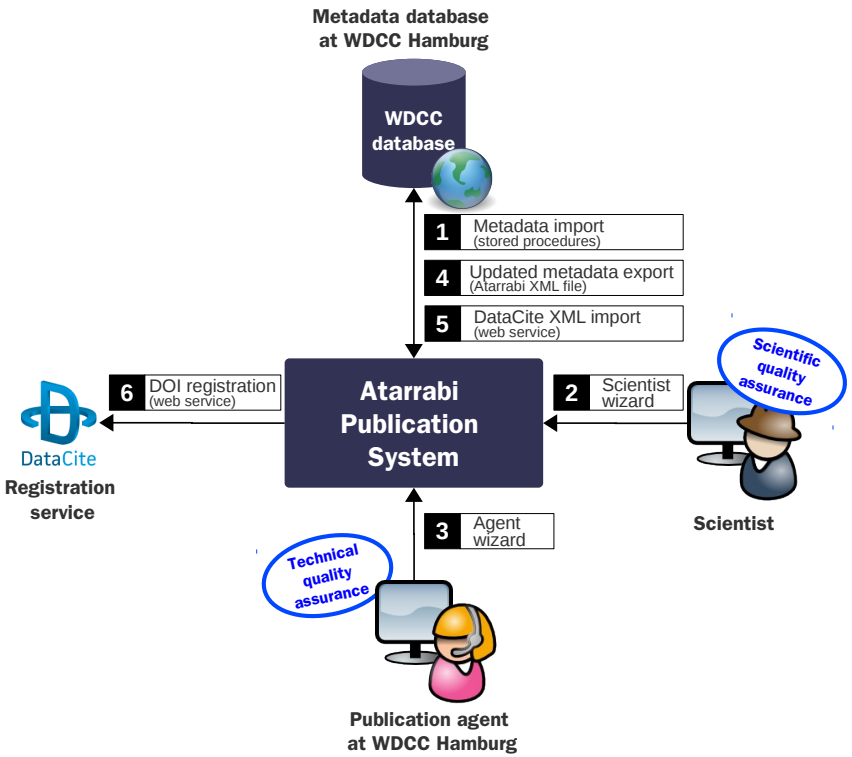
Objectives

- Definition of a *standard procedure* for the publication of data including documentation of quality assurance actions.
- Development of the *workflow system* **Atarrabi** (Basque: good weather spirit) for the researcher and the publication agent.
- *Integration* into the Word Data Center for Climate (WDCC).
- *Generalisation* of the defined processes for other environmental sciences.

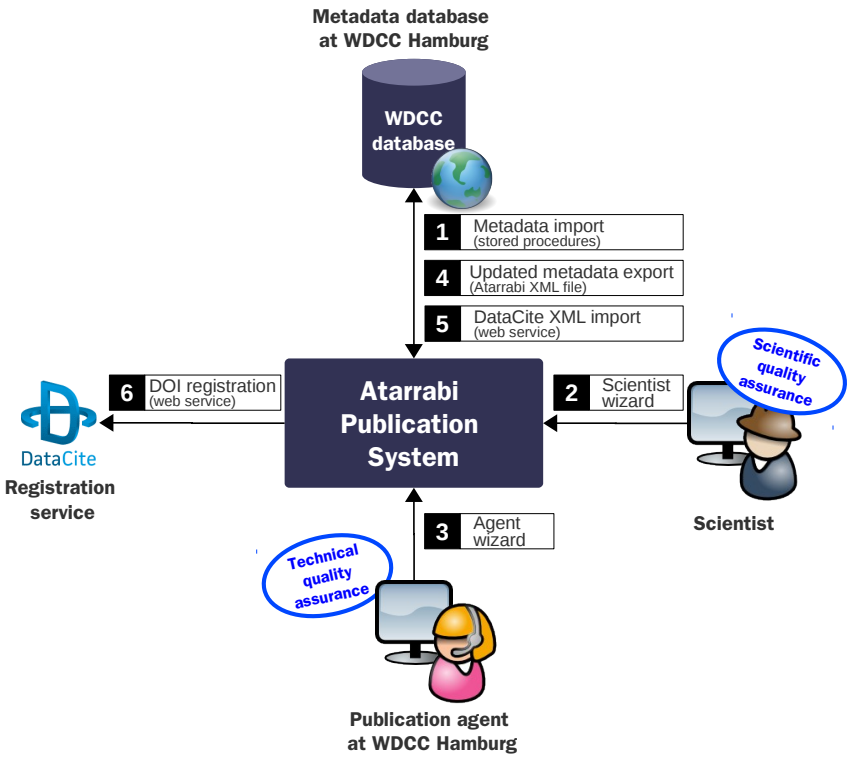

Metadata database
at WDCC Hamburg
WDCC
database
1 Metadata import
(stored procedures)
4 Updated metadata export
(Atarrabi XML file)
5 DataCite XML import
(web service)
6 DOI registration
(web service)
Atarrabi
Publication
System
2 Scientist
wizard
Scientific
quality
assurance
DataCite
Registration
service
Scientist
3 Agent
wizard
Technical
quality
assurance
Publication agent
at WDCC Hamburg



1. Atarrabi imports all relevant metadata of the meteorological experiment to publish from the CERA2 database at WDCC Hamburg.
2. In the course of the 'scientist wizard' the researcher checks, changes and augments the metadata and performs and documents the scientific quality assurance (SQA).



3. A publication agent at WDCC Hamburg double-checks the results of the scientist wizard and performs the technical quality assurance (TQA).
4. If all checks are passed successfully, the updated metadata are exported back to the CERA2 database.



5. After the metadata in the CERA2 database have been updated, an XML file will be generated which is compliant to the DataCite XML schema for DOI registration. This file is imported by Atarrabi for the next step.
6. The XML file which has been imported in the step before is used to register DOI, metadata and landing page at DataCite.

World Data Center for Climate, Hamburg

CERA

WDC

Atarrabi
Publication System
Version 2.0

You are not logged in. | About | Impressum

Menu

- Home

Login

Welcome to the Atarrabi Publication System!

Please login using your existing CERA account credentials to publish data or to view work in progress.

Username
Password
Login

Create a new CERA account:
Register new CERA user

News

Regularly Scheduled Maintenance

There will be a regularly scheduled maintenance window every day from 5:00am until 5:05am CET. During this time, the Atarrabi system may be briefly unavailable.

Creation date: 11/08/11 09:38
Reporter: k204060

About Atarrabi

The Atarrabi web application is used to prepare meteorological data for publication via DataCite.

CMIP5 | Environmental Data

The Atarrabi interface provides a key part of the DataCite publication process of CMIP5 core data by supporting the authors final quality control of data and metadata. After publication the data remains unchanged. A copy of the data is stored on tapes in the long-term archive of the World Data Center for

Documentation

Download the Atarrabi 2.0 tutorial (PDF file)

About this software

DFG

This software is funded by the German Research Foundation

Learn more about the research project Publication of Environmental Data

DFG

Selected Concepts of the Atarrabi Publication System I

Web-based No client-side installation necessary. Atarrabi is hosted as a web-based application.

Workspaces After login the user is directed to a role-dependent workspace, which focuses on the relevant information and functions.

Wizards The metadata review process is structured like a software installation wizard. The individual wizard pages deal with certain aspects of the metadata (e. g. authors, spatial coverage, instruments, spatial coverage, quality). The scientist can leave the wizard anytime and continue at a later time.

Selected Concepts of the Atarrabi Publication System II

List-of-Values Based on data already stored in the CERA2 metadata database at DKRZ Atarrabi provides search-as-you type lists or dropdown menus for persons, institutes, location, units, identifier and relation types.

Workflow Using a workflow engine behind the scenes ensures a well-defined and easily changeable process flow.

Adaptability Atarrabi is designed to be adaptable to other ...

- data domains,
- data sources, and
- process flows

Demo

DEMO

How to get the source:

Atarrabi has been published under a GPL license at Sourceforge:
`http://sourceforge.net/projects/atarrabi/`

Uptake

- Atarrabi releases so far:
  - Atarrabi 1.0 (January 2011): start of production with scientist wizard supporting CMIP5 data (climate simulation).
  - Atarrabi 1.5 (April 2011): support for observational data.
  - Atarrabi 2.0 (November 2011): wizard for agent review.
  - Atarrabi 2.1 (March 2012): automated identifier registration using the TIB web services.
- Up until now about 50 DOIs have been successfully registered with the support of Atarrabi.
- Very positive feedback from Atarrabi users.

Next Steps

- Maintenance of Atarrabi is taken over by DKRZ.
- Adaptation to other data repositories and other data domains requires new effort.
- Our vision: extend Atarrabi to support peer-review of data!

Please do not forget to take home a flyer.

Time for Questions

[Hense et al., 2009]

References I

ANDS (2009).
About ANDS.
http://ands.org.au/.

ARROW (2008).
Monash university ARROW repository | home.
http://arrow.monash.edu.au/vital/access/manager/Index.

AuScope (2009).
AuScope discovery portal.
http://portal.auscope.org/portal/gmap.html.

References II

COPS (2007).
COPS - convective and orographically-induced precipitation study.
http://www.cops2007.de/.

DataCite (2010).
Welcome to DataCite.
http://datacite.org/.

DKRZ (2010).
Deutsches klimarechenzentrum GmbH.
http://www.dkrz.de/.

References III

GOP (2007).
GOP - general observation period.
http://gop.meteo.uni-koeln.de/gop/doku.php.

Hense, A. and Quadt, F. (2011).
Acquiring high quality research data.
D-Lib Magazine, 17(1/2).

Hense, A. V., Hense, A. N., and Lautenschlager, M. (2009).
Publikation Umweltdaten.
http://umwelt.wikidora.com/wikidora/.

References IV

IMOS (2009).
Introducing IMOS.
http://imos.org.au/.

Klump, J. (2008).
Anforderungen von e-Science und Grid-Technologie an die Archivierung wissenschaftlicher Daten.
nestor-materialien 9.

LLNL (2010).
CMIP5 coupled model intercomparison project.
http://cmip-pcmdi.llnl.gov/cmip5/.

References V

TIB (2010).
GetInfo - get ahead and create knowledge.
https://getinfo.de/app.

Treloar, A., Groenewegen, D., and Harboe-Ree, C. (2007).
The data curation continuum.
D-Lib Magazine, 13(9/10).