

Review on 5 Years DataCite and 10 Years DOI Registration for Data

DataCite Annual Conference 2014

Nancy, August 25th – 26th

Michael Lautenschlager

(DKRZ – German Climate Computing Centre, Hamburg)

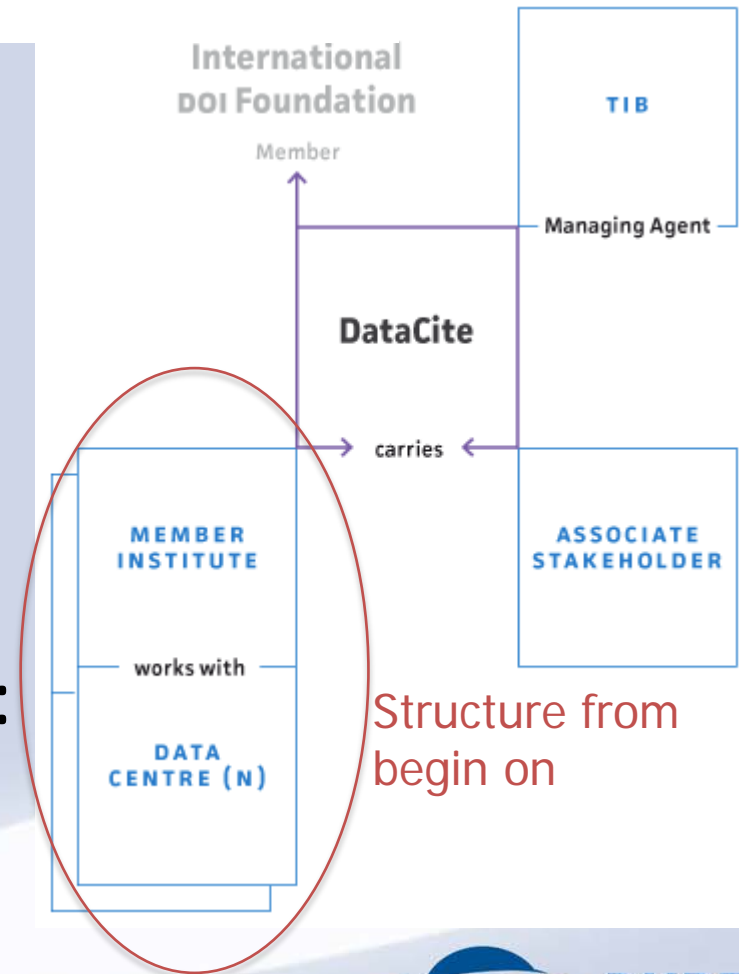


DataCite

Helping you to find,
access, and reuse data

Content

- DataCite:
5 years in a nutshell
- Towards DataCite:
Development before 2009
- DataCite in Climate Research:
Long-term Archiving at
WDCC/DKRZ



Structure from
begin on

DataCite in a Nutshell

- founded in December 2009 by 7 members from 6 countries and started with the **TIB handle server containing about 600,000 scientific data STD-DOIs.**
- enfosters citation of scientific data, data citation can help by
 - enabling easy reuse and verification of data
 - allowing the impact of data to be tracked
 - creating a scholarly structure that recognises and rewards data producers
- Member development
 - 2010: 15 members from 10 countries
 - 2014: 22 full members and 9 associated members from 19 countries
- DOI development
 - 2010: 1 Mio
 - August 2014: 3.5 Mio (expecting more than 4 Mio by end of 2014)
- More details: <https://www.datacite.org>

DataCite Evolution in 5 Steps

- (1998) – 2000: Basic Ideas
 - In the last quarter of 2000 the evolution towards DataCite started.
 - Continuous funding of the development by DFG

DFG funded development projects

- 2001 – 2002: Concept and Implementation Plan
- 2003 – 2005: Pilot Implementation
 - 2004: Start of STD-DOI Handle Server at TIB, Hannover
- 2006 – 2009: STD-DOI Data Publication as a Service

- 2009 – 2014: DataCite
 - 2012: Shut down of STD-DOI Handle Server at TIB and final transition to DataCite Handle Server(s)

DataCite Evolution (i)

- (1998) – 2000: Basic Ideas
 - **Mundt (1998)**: term paper at Uni. Potsdam / GFZ entitled “Der DOI (digital object identifier) ein verlagsorientiertes Indexierungswerkzeug auch anwendbar auf Datensätze?”
 - Oct. 2000: **17th International CODATA Conference** (Committee on Data for Science and Technology): discussion of basic ideas for scientific data publication by Michael Lautenschlager (DKRZ) and Joachim Wächter (GFZ)
 - Nov. 2000: presentation of a concept paper „Publikation und Zertifizierung von wissenschaftlichen Daten“ by and M.L. and J.W. at the **regular meeting of the CODATA national committee** and discussion with DFG (S. Eckelmann) about funding opportunities
 - End of 2000: **decision of DFG to fund a CODATA Working Group** to discuss between librarians and scientists the concept of scientific data publication and come up with an implementation plan

Mundt (1998)

Michaela Mundt

02.10.1998

Weiterbildung zur Wissenschaftlichen Dokumentarin

Feldseminar im Daten- und Rechenzentrum des Geoforschungszentrums Potsdam

20. Juli-2. Oktober 1998

Der DOI (digital object identifier)

ein verlagsorientiertes Indexierungswerkzeug auch anwendbar auf Datensätze?

Internetstudie zur möglichen Anwendbarkeit des DOI

für die im ICDP-Clearinghouse angebotenen Daten

E

17th International CODATA Conference (2000)



**Tischvorlage CODATA Landesausschußsitzung
29.11.00 in Bonn:**

Publikation und Zertifizierung von wissenschaftlichen Daten

Michael Lautenschlager (MPIM/M&D)
Joachim Wächter (GFZ)

Der Fortschritt der modernen wissenschaftlichen Forschung geht mit einem enormen Gewinn an Daten einher. Die einzelnen Disziplinen, z.B. Biologie oder die Geowissenschaften sind dabei dieses Problem zu betrachten. Von zentraler Bedeutung sind wissenschaftliche Netzwerke, mit denen die Erhebung, das Verfügbarmachen, sowie die Integration und Nutzung der großen, heterogenen, interdisziplinären Datenmengen unterstützt wird.

Concept Paper
(Nov. 2000)

DataCite Evolution (ii)

- 2001 – 2002: Concept and Implementation Plan
 - Begin of 2001: Foundation of CODATA working group “Possibility of Citing Scientific Primary Data”
 - Composition of WG: 50% librarians and 50% scientists (request of DFG as funding agency)
 - Members: **Carola Kauhs** (MPI-M, Hamburg), **Dr. Michael Lautenschlager** (Speaker, MPI-M/DKRZ, Hamburg), **Dr. Manfred Reinke** (AWI, Bremerhaven), **Prof. Dr. Gerhard Schneider** (Uni. Freiburg), **Dr. Irina Sens** (TIB, Hannover), **Dr. Uwe Ulbrich** (Uni. Cologne), **Dr. Joachim Wächter** (GFZ, Potsdam)
 - Funding period: 2 years funding from DFG
 - Final Report: “**Conception of citing scientific primary data**” (May 29th, 2002)
 - Presentation of results: March 2002 to Library Committee of DFG
 - Conditional funding recommendation with geoscience as pilot community but with URNs as second persistent identifier beside DOIs and with contemporary reporting to DFG

Conception of citing scientific primary data

Final Report

CODATA AG "Possibility of citing scientific primary data"

(May 29th, 2002)

Participants:

Carola Kauhs

(Head of Library, Max Planck Institute for Meteorology, Hamburg)

Dr. Michael Lautenschlager

(Speaker of the Working Group, Max-Planck Institut)

Dr. Manfred Reinke

(Scientific Information System, Marine Research, Bremerha)

Prof. Dr. Gerhard Schneider

(Head of Computer Center,

Dr. Irina Sens

(Director's Representative, Information Library)

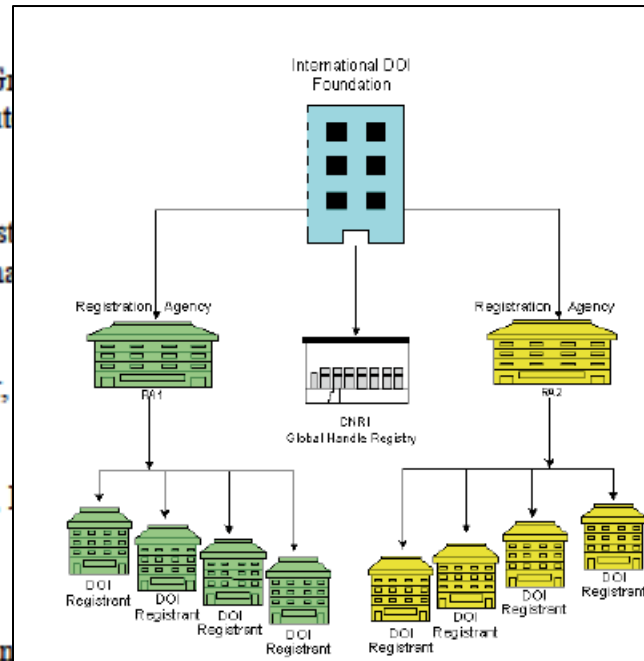
Dr. Uwe Ulbrich

(Institute for Geophysics an

Dr. Joachim Wächter

(Head of Data and Computer Center, Geophysikalisches Institut Potsdam)

Guiding Principle:
Analogy to scientific literature as close as possible, which means scientific data are irrevocable after publication and have a suitable granularity.



Central registration and "Handling System"

subject level with quality assurance and long-term archiving

Figure: Organization of the IDF (Source: DOI Handbook 2002)

- 2003 – 2005: Pilot Implementation
 - Project: 2 years funded by DFG
 - Partners: TIB and 3 data archives, WDCC/DKRZ (coordination), WDC Mare, GFZ Potsdam
 - **Focus:** Scientific reference data which are ready be used in scientific literature (final research data products)
 - **Concept:** quality controlled scientific data are classified as irrevocable and are registered together with DOI (URN) and citation reference in a CNRI Handle Server and in library catalogues in order ...
 - To allow for transparent data access
 - To foster verification of scientific results
 - To allow for data citation in scientific literature
 - To give credit to data authors
 - **STD-DOI metadata** (STD - Scientific and Technical Data): DOI kernel includes Dublin Core and ISO 690 in order to integrate metadata for electronic publication, **cross-disciplinary usability**
 - **Problem:** definition of data granularity across disciplines which is suitable for citation in scientific literature
 - **2004: Start of STD-DOI Handle Server** at TIB, Hannover

2004: Start of STD-DOI Handle Server at TIB

First DOIs from project partners:

18.03.2004: doi:10.1594/WDCC/EH4_OPYC_SRES_A2 (DOI #1)

22.07.2004: doi:10.1594/GFZ/ICDP/KTB/KTB-GEOCH-GASCHR-P

14.12.2004: doi:10.1594/PANGAEA.119754

End of 2004: about 30 DOIs registered and in library catalogue TIBORDER

DKRZ DEUTSCHES KLIMARECHENZENTRUM World Data Center for Climate Hamburg

DOI #1

DOI for Scientific and Technical Data
10.1594/WDCC/EH4_OPYC_SRES_A2

Citation elements

Creator
(person(s) or institute(s) responsible for this assemblage of data: e.g. author, data collector, editor...)
Stendel, Martin; Schmith, Torben; Roeckner, Erich; Cubasch, Ulrich

Publication Year
2004

Title
ECHAM4_OPYC_SRES_A2: 110 YEARS COUPLED A2 RUN 6H VALUES

DOI Publisher
WDCC at DKRZ

Identifier
DOI:10.1594/WDCC/EH4_OPYC_SRES_A2

The DataCite consortium proposes a citation format (Creator (PublicationYear): Title. DOI Publisher. Identifier) citation export in different formats.

Detailed Metadata
http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=EH4_OPYC_SRES_A2

Data Access
http://cera-www.dkrz.de/WDCC/ui/EntryList.jsp?acronym=EH4_OPYC_SRES_A2

Summary
The SRES data sets were published by the IPCC in 2000 and classified into four different scenario families (A1, A2, B1, B2). SRES-A2 storyline describes a very heterogeneous world with the underlying theme of self-reliance and preservation of local identities. It results in this scenario a continuous increasing population together with a slower economic growth and technological change. The model consists of the atmospheric component which based on the weather forecast model of ECMWF. The atmospheric component is the standard model version of a 19-level hybrid sigma-pressure coordinate system.

TIBORDER - Dokumentlieferdienst der TIB Hannover - results/titledata - Microsoft Internet Explorer - zur Verfügung gestellt von

Adresse <http://tws.gbv.de/DB=2.63/SET=11/TTL=1/SHW?FRST=1>

Suchen | Suchergebnis | Erweiterte Suche | Zwischenablage | Hilfe

suchen [und] Person, Autor sortiert nach Erscheinungsjahr

stendel, martin suchen

Kundennummer: | Datenbanken

Suchgeschichte | Kurzliste | Titeldaten

Ihre Aktion suchen [und] (Person, Autor) stendel, martin 1 von 8

Titel: ECHAM4_OPYC_SRES_A2: 110 years coupled A2 run 6H values / World Data Center for Climate (WDCC), Hamburg. **Martin Stendel**; Torben Schmith; Erich Roeckner ...

Beteiligt: [Martin Stendel](#); [Torben Schmith](#); [Erich Roeckner](#)

Körperschaft: [World Data Center for Climate \(WDCC\)](#)

Erschienen: 2003-01-26

Umfang: 961261722720 Bytes.

Anmerkung: mode: abstract
resourceType: PrimaryData
issueNumber: 1

Inhalt: This is a transient scenario experiment where changes in greenhouse gas concentrations and sulfur emissions from 1990 until 2100 are prescribed according the SRES A2 scenario (IPCC, Third Assessment Report). The model experiment is initialized in 1990 using data from the 'GSDIO' experiment described in Roeckner et al. (1999) J. Climate 12, 3004-3032. The model consists of the atmospheric component ECHAM4 (MPI report No. 218) and the isopycnal ocean model OPYC (DKRZ technical report No. 7). The coupling procedure and results are documented in Roeckner et al. (1996), Climate Dynamics 12, 737-754 and in Bacher et al. (1998) Climate Dynamics 14, 431-450.

Technische Angaben:format: GRIB

Bestandsinfo: <89> Hannover, Universitätsbibliothek Technische Informationsbibliothek
Volltext/Image: [Anzeigen](#) [lizenzpflichtig!](#)

1 von 8

enschlager **DKRZ**

- 2006 – 2009: STD-DOI Data Publication as a Service
 - Project: 2 1/2 years funded by DFG
 - Partners: TIB (coordination) and 4 data archives WDCC/DKRZ, WDC Mare, GFZ Postdam, and WDC RSAT
 - Focus of Funding period:
 - STD-DOI data publication as operational service
 - Expansion from geosciences to more disciplines
 - Integration of international partners
 - Sustainability and business model
 - TIB Advisory Board 2005:
 - integration of operation of a non-commercial registration agency for scientific-technical primary data into TIB operational services
 - 17 new partners partners from different disciplines:
 - International Ocean Drilling Project, Freiburg Materials Research Center, FLOSS project, Verlag Thieme Chemistry, Office of Scientific & Technical Information, Lamont-Doherty Earth Observatory, and libraries
 - Increasing international interest led to a discussion about the organisational structure and finally led to the foundation of DataCite

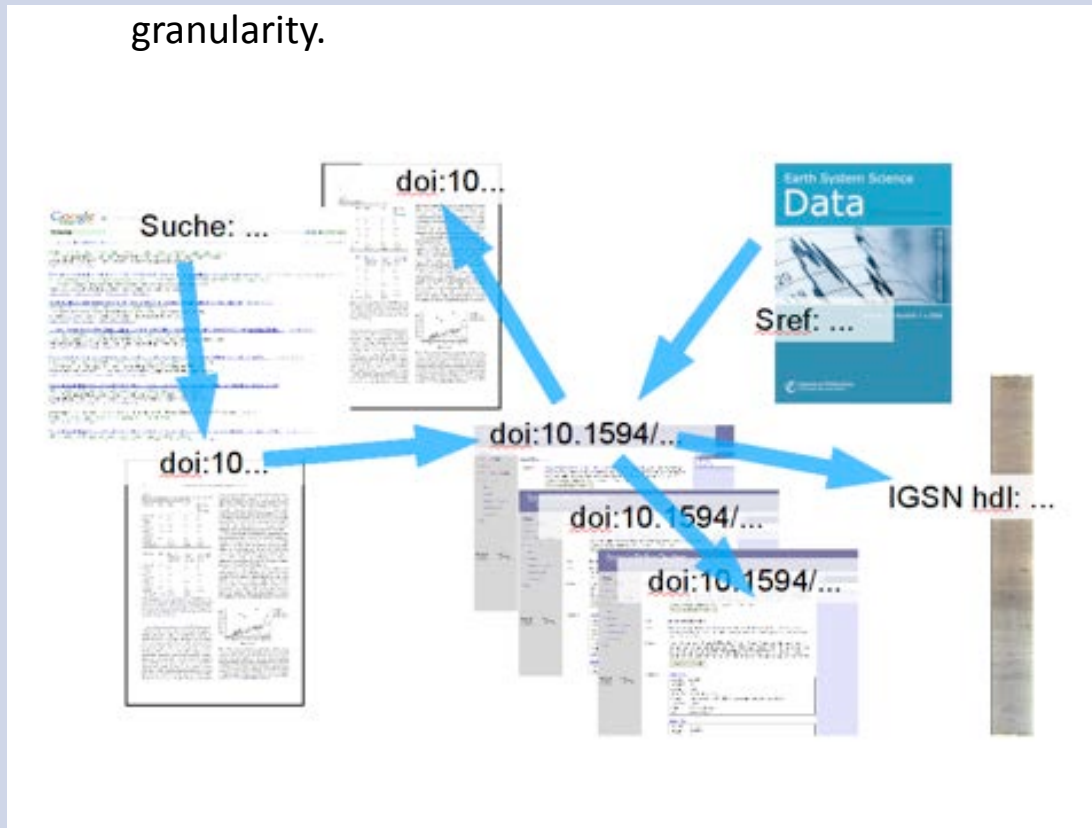
- Points of discussion:

- Use of **scientific data in literature**:

- Supplement: the data are part of an article, e.g. geophysical measurements
 - Independent data publication: an article references to the data, e.g. climate model data, satellite data

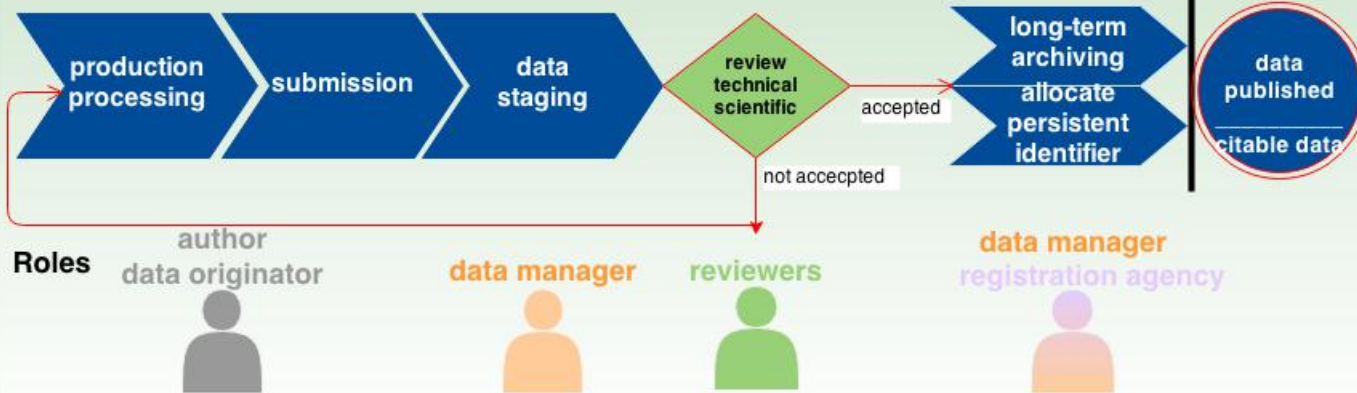
- **Granularity**:

- Definition of data entities for STD-DOI data publication which are suitable to appear in reference lists is discipline dependent and normally different from data access granularity.

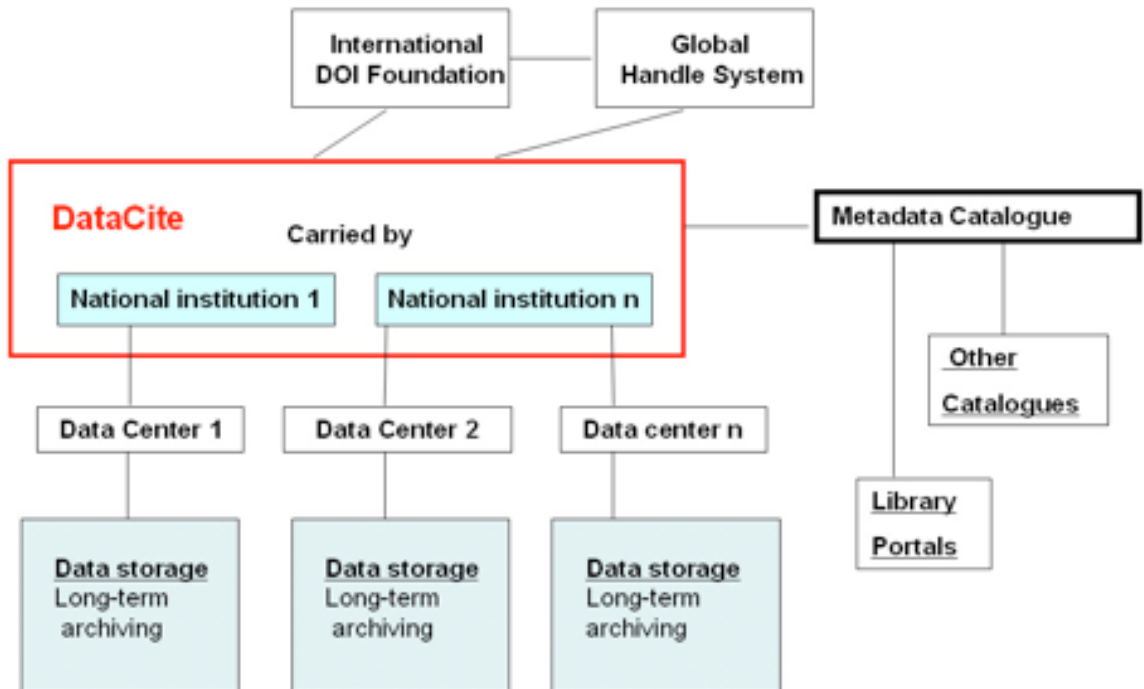


STD-DOI network of articles, data and samples

Data Publication



Organisational Structure of publications agency, publication agent and finally DataCite



DataCite Evolution (v)

- 2009 – 2014: DataCite
 - **December 2009: Formation of DataCite** as non-profit organisation in London with 7 members from 6 countries, starting with about 600,000 scientific data STD-DOIs from the TIB Handle Server
 - TIB operated as Managing Agent with office in Hannover and with Jan Brase in person
 - 2010: 15 members from 10 countries and 1 Mio DIOs
 - 2012: Shut down of STD-DOI Handle Server at TIB and final transition to DataCite Handle Server(s)
 - 2014: 22 full members and 9 associated members from 19 countries and 3 Mio DIOs
 - DataCite Metadata Schema
 - list of core metadata properties chosen for the accurate and consistent identification of data for citation and retrieval purposes, along with recommended use instructions
 - Integration into repositories of research data
 - DataBib initiative and their list of repositories
 - Google Docs
 - re3data.org

Global Distribution of DataCite Members



DataCite in Climate Research

Examples are

- DKRZ Data Services and Atarrabi as publication tool
- CMIP5 and DataCite data publication for reference data

DKRZ Projects

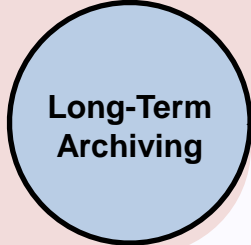
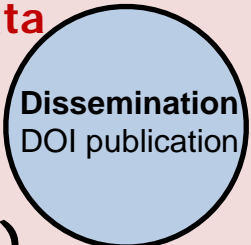
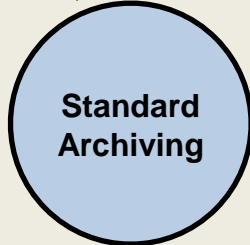
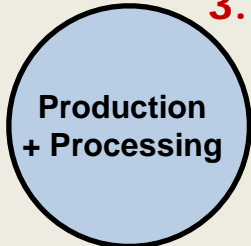
ESGF

Standardized
Research Data
Environment

1. Data Management Plan

2. DKRZ Storage

3. ESGF Standardization



7. DataCite Data
Publication

3.7 PB (07.14)

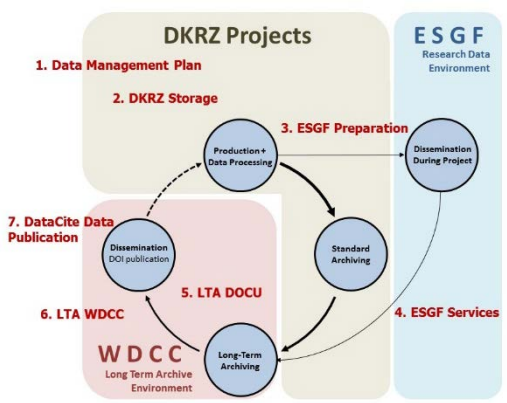
5. LTA DOKU

6. LTA WDCC

WDCC

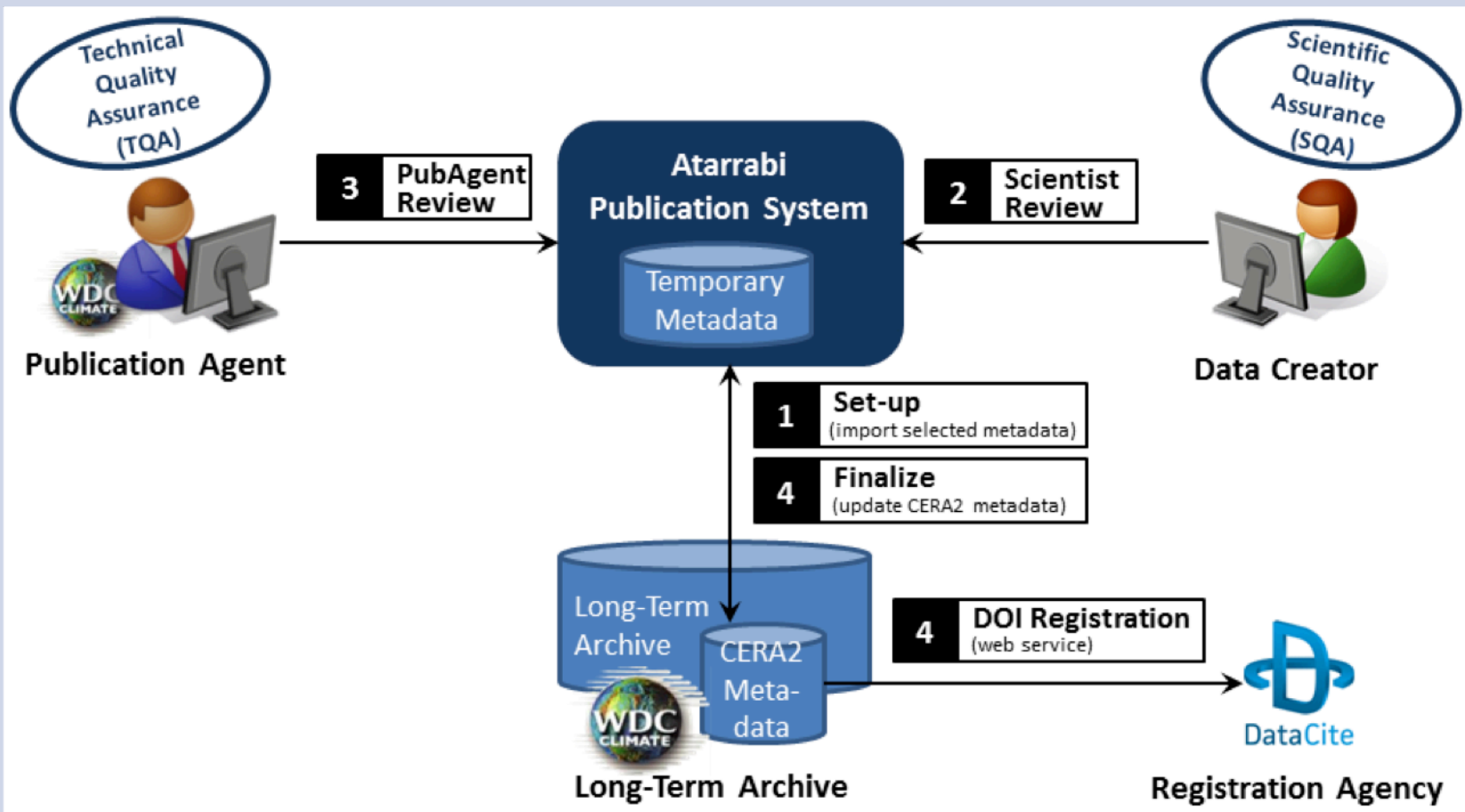
Long Term Archive
Environment

4. ESGF Services

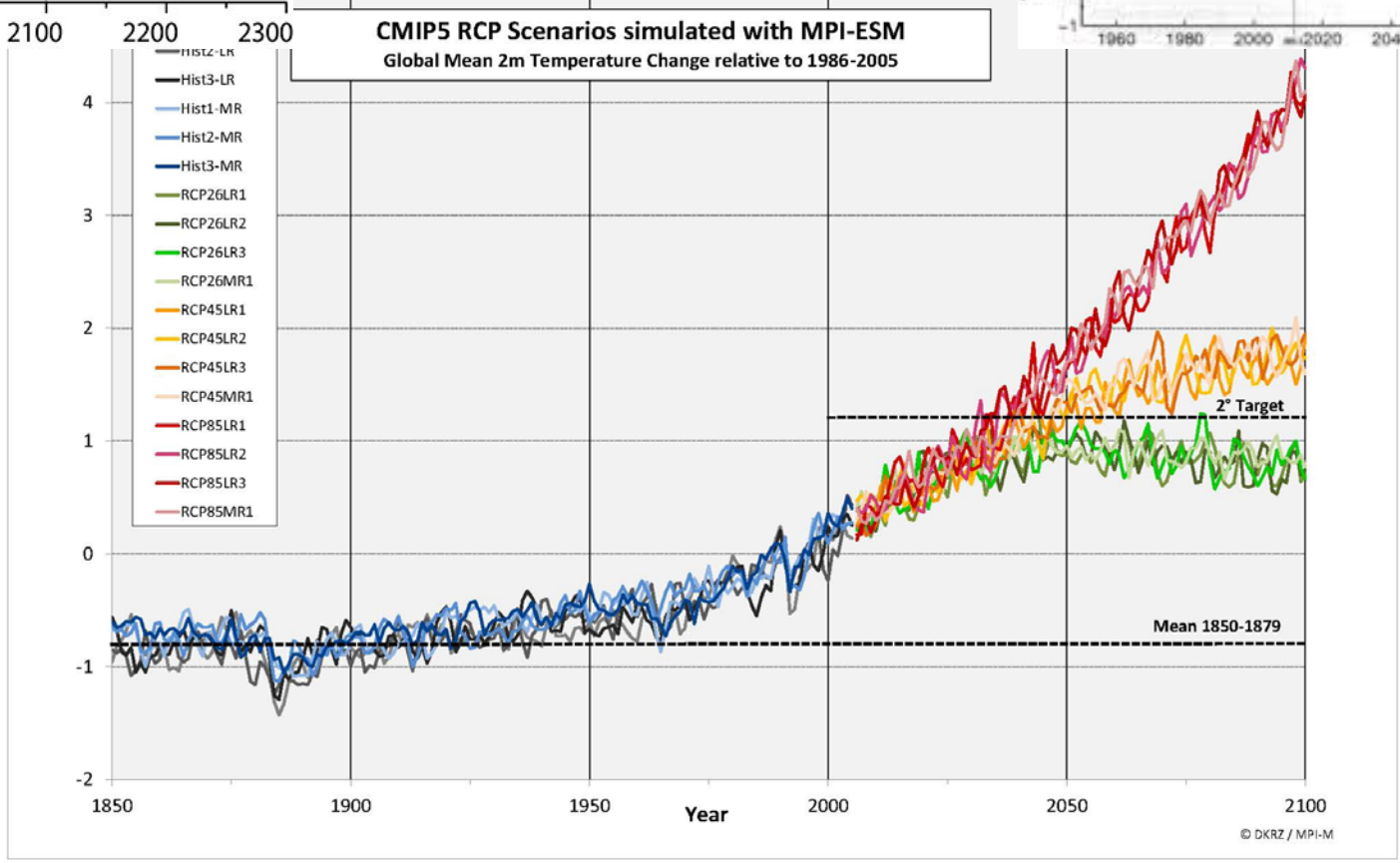
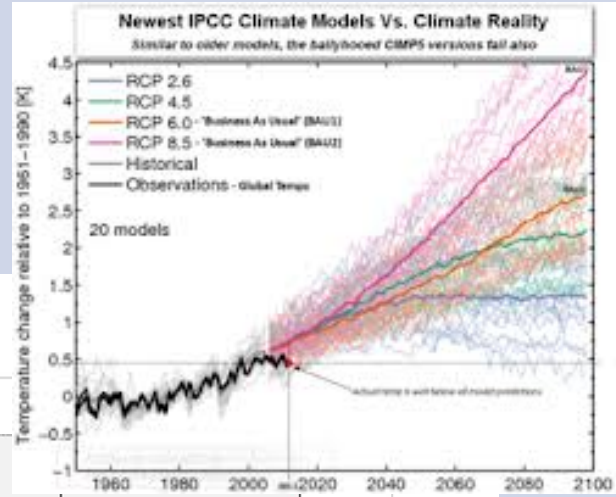
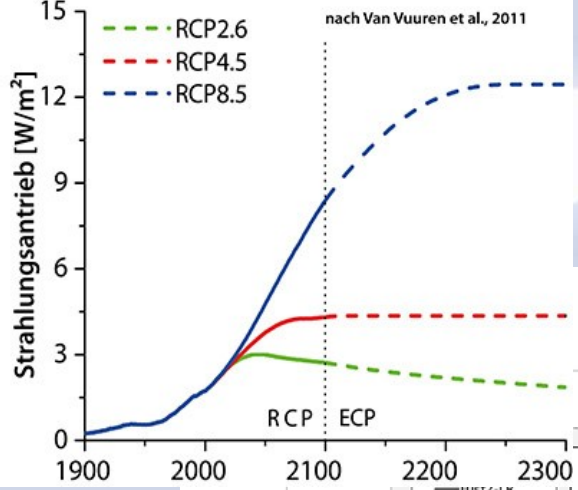


7. DataCite Data Publication at WDCC/DKRZ

Formal citation of data using DataCite data DOIs allows to give and to get credit for the preparation of high-quality research data and allows for transparent data access via DOI and Handle Server.



CMIP5 Results

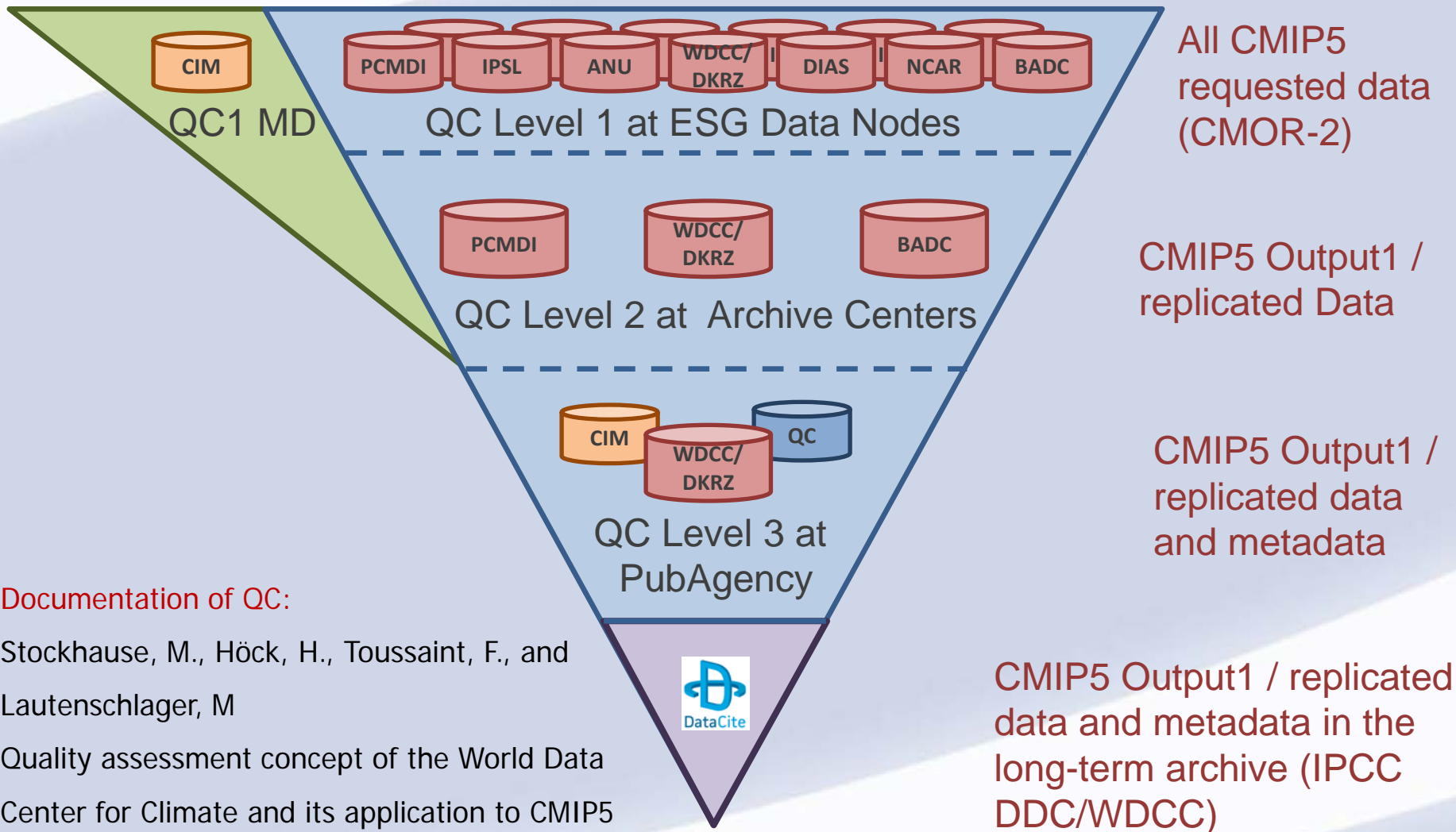


Data Amounts CMIP3/CMIP5

DataCite data publication
unit: Model Experiment

- CMIP3 / IPCC-AR4 (Report 2007)
 - Participation: 17 modelling centres with 25 models
 - In total 36 TB model data central at PCMDI and ca. ½ TB in IPCC DDC at WDCC/DKRZ as reference data
- CMIP5 / IPCC-AR5 (Report 2013/2014)
 - Participation: 29 modelling groups with 61 models
 - Produced data volume: ca. 10 PB with **640 TB from MPI-ESM**
 - CMIP5 requested data volume: ca. 2 PB (in CMIP5 data federation)
 - Data volume for IPCC DDC: 1.6 PB (complete quality assurance process) with **60 TB from MPI-ESM**
- Status CMIP5 data archive (August 2014):
 - 2.3 PB for 69000 data sets stored in 4.3 Mio Files in 23 data nodes
 - CMIP5 data is **more than 50 times** CMIP3

3-Layer Quality Assurance Concept



Documentation of QC:

Stockhause, M., Höck, H., Toussaint, F., and Lautenschlager, M

Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. Geosci. Model Dev., 5, 1023-1032, 2012

DOI :[10.5194/gmd-5-1023-2012](https://doi.org/10.5194/gmd-5-1023-2012)

DataCite Conference 2014 / Lautenschlager (WDCC/DKRZ)

Finalisation of Quality Assessment

- After final control of data and metadata CMIP5 data are transferred from the ESGF archive (*most recent version*) into the reference data archive (*snapshot around March 2013*)
 - Quality status: „**approved by author**“
 - Data are marked as **irrevocable**
 - Long-term archiving in **WDC Climate** of DKRZ
- Final step is the DataCite data publication and integration of associated citation reference into library catalogues
 - Data entity (here one climate model experiment) receives a **citation reference** for direct usage in scientific publications and a **DOI** (Digital Object Identifier) for the transparent data access
 - Citation reference contains data author and title as well as WDC Climate as DataCite DOI publisher and the DOI
 - Resolution of the DOI leads to a „**Landing Page**“, which address is stored in the central data base of the DOI Handle Server at DataCite

Status QC for CMIP5

- QC Status CMIP5 (August 2014)
 - Quality Control 1: 1142 Experiments
 - Quality Control 2: 910 Experiments (finalised 680)
 - Quality Control 3: 994 Experiments
 - DataCite DOI: 240 Experiments (WDCC / IPCC-DDC)
 - RCPs, AMIP, Historical

IPCC Data Distribution Centre:

http://www.ipcc-data.org/sim/gcm_monthly/AR5/index.html

DataCite DOI Interface at WDCC (CMIP5 Reference Data Archive)

- Show me all data entities in the WDCC, which have an DataCite DOI
 - URL: <http://cera-www.dkrz.de/WDCC/ui/FindDoiPublications.jsp>
- Show me for a given CMIP5 NetCDF/CF dataset, whether it belongs to a DataCite published data entity
 - URL: <http://cera-www.dkrz.de/WDCC/CMIP5/Citation.jsp>

My DataCite Wish List

Beside fostering citation of scientific data I would like to see more emphasis from DataCite on

- Keeping the focus on irrevocable and well-documented scientific data products
- Use DOIs for data entities with a granularity that is suitable for reference lists in scientific work
- Integrating trusted, certified data centres into DataCite in order to ensure quality and accessibility of data and services

DataCite should strengthen its branding.