



World Data Center Cluster "Earth System Research"

- an approach for a common data infrastructure in geosciences

Michael Lautenschlager - WDC Climate
(Max-Planck-Institut für Meteorologie, Hamburg, Germany)

With Contributions of:

Michael Diepenbroek, Hannes Grobe – WDC MARE

Eleni Paliouras – WDC RSAT

Jens Klump – WDC TERRA (candidate)

Irina Sens, Jan Brase – STD-DOI Registration Agency

AGU Fall Meeting, San Francisco, December 2005





Content

- ◆ Scope of German WDC Cluster ESR
- ◆ Data Provision for e-Science: STD-DOI Concept



The WDC cluster „Earth System Research“

Founded in 2004

Members:

- WDC-MARE - World Data Center for Marine Environmental Sciences
(AWI, MARUM, Bremen)
- WDC-C - World Data Center for Climate (MPI-M + DKRZ, Hamburg)
- WDC-RSAT – World Data Center for Remote Sensing
(DLR, Oberpfaffenhofen)
- WDC-TERRA – World Data Center of the Lithosphere
(candidate) – (GeoForschungsZentrum, Potsdam)





WDC-MARE

www.pangaea.de

World Data Center for Marine Environmental Sciences

Hosted by:

- *AWI – Alfred Wegener Institute for Polar and Marine Research, Bremerhaven*
- *MARUM – Institute for Marine Environmental Sciences, Univ. Bremen*

Scope of managed data types:

- *Focus on georeferenced observational data in the fields of environmental oceanography, marine geology, paleoceanography, and marine biology.*

Resources / Staffs:

- *~ 6 positions for data management*
- *4 positions technical and scientific organisation & development*
- *AWI Computer center*

Size of data inventory:

- *260,000 data sets from more than 300,000 geological and hydrological stations*

Specials:

- *Project data management (54 European and international projects)*





WDCC World Data Center for Climate



<http://www.wdc-climate.de>

Hosted by:

- *Model and Data Group at Max-Planck Institute for Meteorology (M&D/MPI-M), Hamburg
German Climate Computing Centre (DKRZ)*

Scope of managed data types:

- *Focus is on data from climate modelling and related observations*

Resources / Staffs:

- *5 positions for technical and scientific organisation and development*

Data inventory:

- *175 Terrabyte data comprising 580 experiments with 68,000 time series and*
- *4.6 Billion BLOBs (individual data tale entries) in a relational DB-System*

Specials:

- *International Panel on Climatic Change (IPCC) Data Distribution Center (DDC),
Coordinated Enhanced Observing Period (CEOP) Model Data Center*



WDC-RSAT World Data Center for Remote Sensing of the Atmosphere

Hosted by:

- DFD – German Remote Sensing Data Center at the German Aerospace Center (DLR), Oberpfaffenhofen

Scope of managed data types:

- Focus on satellite data (trace gases, aerosols, clouds, UV-radiation, surface parameters, dynamics, spectra).

Resources / Staffs:

- *4-5 fulltime positions for technical and scientific organisation and development*

Data inventory:

- 60 Terrabyte ranging from raw data to high level data products produced with sophisticated techniques in near real time



WDC-TERRA

<http://www.gfz.de>

World Data Center of the Lithosphere (candidate)

Hosted by:

- GFZ – Geoforschungszentrum Potsdam

Scope of managed data types:

- Focus on observational data from continental drilling and data from geodetical investigations (satellite – CHAMP), seismology

Resources / Staffs:

- 4 fulltime positions for technical and scientific organisation and development

Data inventory:

- 15 Terrabyte (CHAMP), 5 Terrabyte seismology, ~2000 geological data sets

Specials

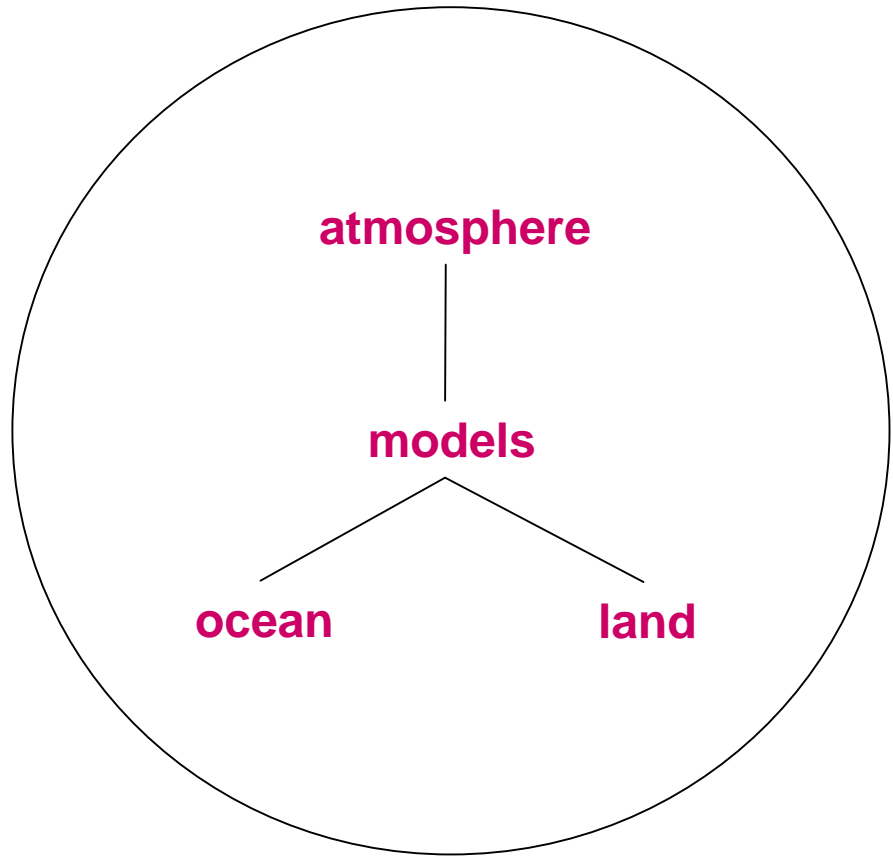
- Geophone as part of the international array of seismologic stations, seismologic task force





WDC cluster „Earth System Research“

Data Type Coverage





Activities & characteristics of the WDC cluster

Longterm_archiving facilities

- *Clear commission as data libraries*
- *Data management infrastructure, expertise, and manpower*
- *Longterm commitment and funding*

Peer review for scientific data (*Publication Agent*)

- *Completeness of data set descriptions (metadata)*
- *Validity of methods used*
- *Data values (precision, sequence, and ranges)*
- *Identifikation of independent data entities which are suitable for publication*
- *Data publication based on citable data entities having persistent identifiers (DOI)*

Userfriendly and reliable systems for data retrieval and distribution

- *General nonrestricted web-access*
- *Offline products (e.g. data collections, DVD)*

Fostering common standards and protocols

Clear commitment to the rules for good scientific practice and open access!



Shortcomings in data provision and interdisciplinary use

Rules of good scientific practise are not taken into account in all cases.

Data sources are widely unknown.

Data are archived without context.

Data cannot be cited as independent entities

Method of solution: publication of primary data as independent entities

Persistent Identifier (DOI/URN) with global resolving mechanism for data archive and context referencing (scientific datamodel at archive level)

Integration into **library catalogues** in order to find data together with articles

STD-DOI application profile: DOI meta data kernel + items for electronic publication (interface between scientific data archives and libraries / e-science)

Motivation for Scientists: Citable data publication gives credit to scientific data management e.g. data documentation, quality assurance,

(1) STD-DOI: Scientific and Technical Data – Digital Object Identifier

STD - DOI Implementation (I)

1. Persistent identifiers (DOI/URN) provide a **global resolving mechanism** of digital objects which is independent from their current URLs.
2. The **connection between data object and URL of the storage place** is archived in the global handle server and is transparent to the user.
3. The **DOI Registration Agency** is responsible for maintaining the integration in the handle server. TIB (German National Library of Science and Technology, Hannover) is the global registration agency for STD-DOIs.
4. DOIs are normally used for classical publications and we use them for publishing scientific data. That means data which have assigned a DOI are **accessible without restrictions** and are **no longer matter of change** like articles in scientific journals.

STD-DOI Implementation (II)

5. Scientific data which are suitable for publication undergo a **review process** to ensure the completeness and correctness of their **description** (meta data) and to ensure **data quality** (Peer Review). The **Publication Agents** at the scientific level like the WDCs in the ESR Cluster are responsible for this process.
6. Within the STD-DOI concept we distinguish between Publication DOIs and Identification DOIs. The **Publication DOIs** are connected in addition to the standard DOI meta data with additional metadata which are suitable for electronic publication.
7. These Publication DOIs are included in **standard library catalogues** in order to make scientific primary data searchable together with scientific articles and to make them citable within scientific articles:

Data Authors, Publication year: Title of Data Entity. [Publication-DOI]

8. The definition of the **data granularity for Publication DOIs** is adapted to their scientific use and to their appearance in scientific literature. Data references should be well balanced with references of classical literature. This granularity definition is clearly dependent on scientific disciplines and data types.
9. Data access is normally of finer granularity than the publication entities. This is taken into account by the **Identification DOIs**. They are part of the same DOI handle system, they can be accessed by the same mechanism, and they undergo the same review process, quality rules etc. as for Publication DOIs. The Identification DOIs are not connected with publication meta data and they are not included in classical library catalogues but they are **related to Publication DOIs**.
10. The system has been originally developed for DOIs but we use **URNs** (Uniform Resource Names) in parallel in order to guarantee non-profit and low cost implementation of the STD-DOI concept.

schauer.dkrz.de:/pf/m/m214002/
NEWEXP/EXP300/run365

Experiment

Exp.-Acronym: [EH5_T63L19_AMIP_6H](#)

[Publ.-DOI](#)

Dataset (BLOB-Table)

DS-Acronym: [EH5_T63L19_R365_TEMP2](#)

Variable: 2m temperature

[Ident.-DOI](#)

Dataset (BLOB-Table)

DS-Acronym: [EH5_T63L19_R365_WIND10M](#)

Variable: 10m wind speed

[Ident.-DOI](#)

Number of datasets: 350 time series of 2D global fields
Total amount of GRIB data: 350 * 1.6 GB = 560 GB

350 * [Ident.-DOI](#)

Integration into eScience Information Systems

TIBORDER

(Reference in Library Catalogue)

Integration of DOIs into Handle Server and
of Publication Metadata into Library Catalogue

Data Integration
Layer

TIB



DDB (URN-Knot)

(Registration Agency)

Transfer of Metadata and DOIs
Execution of Data Review Process

Data Publication
Layer

WDC Climate

WDC MARE

WDC RSAT

WDC TERRA

(Publication
Agents)

Project is granted by DFG

STD-DOI: Implementation Status

The [German National Library of Science and Technology \(TIB\)](#) in Hannover has been accepted by IDF as **Registration Agency** for primary scientific and technical data according to the [STD-DOI application profile](#).

Presently the TIB registered **240.000 Identification-DOIs** and **36 Publication-DOIs**. Presently not all Ident.-DOIs are connected with Publ.-DOIs.

The usage of the DOI system is twofold:

- a) **Reference of individual scientific data entities** in e-science environments (Publ.-DOI)
- b) Direct, URL **transparent data access** of finer granularity at archive level (Ident.-DOI)

The Publ.-DOIs and related Ident.-DOIs can be obtained from TIBORDER by searching for "exk primaerdaten" (<http://tws.gbv.de/>).

Katalog

exk primaerdaten

suchen

Kundennummer: [Datenbanken](#)

Suchgeschichte

Kurzliste

Titeldaten

Download
Zwischenablage

■ Ihre Aktion suchen [und] (Kommentar (Exemplar)) primaerdaten

10 von 37

Datenbanken
Bestellung ohne
Recherche
Benutzerinfo
TIB Homepage

Titel: [ECHAM4_OPYC_SRES_B2: 110 YEARS COUPLED B2 RUN 6H VALUES](#) / World Data Center for Climate (WDCC), Hamburg. Monika Esch

Beteiligt: [Monika Esch](#)

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

Erschienen: 2005-02-09

Umfang: Online-Ressource (892250085420 Bytes).

Anmerkung: Mode: Abstract

StructuralType: Digital

CreationDate: 2002-05-06

Inhalt: The SRES data sets were published by the IPCC in 2000 and classified into four different scenario families (A1, A2, B1, B2). SRES_B2 storyline describes a world in which the emphasis is on local solutions to economic, social and environmental sustainability. The global population is increasing at a lower rate than A2. It has an intermediate level of economic development and a less rapid and more diverse technological change than in A1 and B1. The model consists of the atmospheric component which is 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. The ocean component is a model which computes with isopycnal coordinates. This data set is an enlargement of the IPCC data set and provides additional meteorological parameters. The run produces 6h values of the variables. ECHAM4/OPYC3 (http://cera-www.dkrz.de/IPCC_DDC/SRES/ECHAM4/echam4opyc3.html) Changes of anthropogenic emissions of CO₂, CH₄, N₂O and sulphur dioxide are prescribed according to the above mentioned scenario. The model run starts in 1990 from the results of the scenario run GSDIO (Experiment "EH4OPYC_22723GSDIO") which has been run with observed conditions for the time period 1860-1990.

Technische Angaben: Format: GRIB

Links:

doi: [10.1594/WDCC/EH4_OPYC_SRES_B2](https://doi.org/10.1594/WDCC/EH4_OPYC_SRES_B2)

URN: [urn:nbn:de:tib-10.1594/WDCC/EH4_OPYC_SRES_B20](https://nbn-resolving.org/urn:nbn:de:tib-10.1594/WDCC/EH4_OPYC_SRES_B20)

Bestandsinfo: [Anzeigen](#) lizenzfrei

Anmerkung: **Primaerdaten**

Fertig





[DOI Home](#) | [CERA Home](#) | [WDCC Home](#)
 Always quote citation when using data!

DOI for Scientific and Technical Data Publ.-DOI
 10.1594/WDCC/EH4_OPYC_SRES_B2

Citation
 Esch, Monika (2005): ECHAM4_OPYC_SRES_B2: 110 YEARS COUPLED B2 RUN 6H VALUES. [doi: 10.1594/WDCC/EH4_OPYC_SRES_B2]

Publication Date
 2005-02-09

Author(s)
 Esch, Monika

Title
 ECHAM4_OPYC_SRES_B2: 110 YEARS COUPLED B2 RUN 6H VALUES

Summary
 The SRES data sets were published by the IPCC in 2000 and classified into four different scenario families (A1, A2, B1, B2). SRES_B2 storyline describes a world in which the emphasis is on local solutions to economic, social and environmental sustainability. The global population is increasing at a lower rate than A2. It has an intermediate level of economic development and a less rapid and more diverse technological change than in A1 and B1. The model consists of the atmospheric component which based on the weather forecast model of ECHAM4. The atmospheric component is the standard model...

Fertig



GRIB

Data Size
889597831680 Bytes **830 GB**

Contact
Monika Esch
Max-Planck-Institut fuer Meteorologie
Atmosphere in the Earth System
Bundesstrasse 53
D-20146 Hamburg, F.R. Germany
<http://www.mpimet.mpg.de/>

Project
IPCC-Hamburg Climate Model Simulation (IPCC_HH)
The Intergovernmental Panel on Climate Change (IPCC) has been established by WMO and UNEP to assess scientific, technical and socio-economic information, relevant for the understanding of climate change, its potential impacts and options for adaption and migration. Continued description about the work of the IPCC will be found at the homepage (<http://www.ipcc.ch>) and (www.grida.no/climate/ipcc). As a further development the Special Report on Emission Scenarios (SRES) have been constructed, to explore future developments in the global enviroment with special reference to the production of greenhouse gases and aerosol precursor emissions. A set of four scenarios families (A1, A2, B1, B2) have been developed (see also <http://www.grida.no/climate/ipcc/emission/index.htm>) These data are available at the World Data Center for Climate, Hamburg. (wdc-climate.de).

Available datasets (338)

- EH4_OPYC_SRES_B2_TEMP2
- EH4_OPYC_SRES_B2_TDCL
- EH4_OPYC_SRES_B2_TDMSE
- EH4_OPYC_SRES_B2_TEFF
- EH4_OPYC_SRES_B2_TEMP2**
- EH4_OPYC_SRES_B2_TOPMAX
- EH4_OPYC_SRES_B2_TPREC
- EH4_OPYC_SRES_B2_TRAD0
- EH4_OPYC_SRES_B2_TRADS
- EH4_OPYC_SRES_B2_TRADSU
- EH4_OPYC_SRES_B2_TRAF0

Dataset Information >>
with acronym EH4_OPYC_SRES_B2 can be obtained from the World Data Center Climate (WDCC). Please take a [RA/index.html](#) for information on downloading data.



Always quote experiment citation when using data!

DOI for Scientific and Technical Data

10.1594/WDCC/EH4_OPYC_SRES_B2_TEMP2

Ident.-DOI

Dataset Name

EH4_OPYC_SRES_B2_TEMP2

Dataset Acronym

EH4_OPYC_SRES_B2_TEMP2

Summary

See summary of corresponding experiment. This dataset contains 6H values.

Spatial Coverage

Latitude: -87.863 to 87.863; Longitude: 0.0 to 357.124; Altitude: 2.0 metre to 2.0 metre

Temporal Coverage

1/1/130 to 30/12/240 (climate model time)

Topic

near surface air temperature

Unit

Kelvin

Data Format

GRIB

Data Size

Fertig

Data retrieval procedure is given at the end





STD-DOI: Prototype Bussiness Model



- The cost model on non-profit basis is divided in two parts, one is the direct payment to the TIB for the **registration agency service**. The range of registration costs is **¼ - ½ EUR / DOI** depending on complexity and quantity.
- The second part consists of time and effort which have to be covered by the publication agent for the **publication process itself** including data review, quality assurance and long-term archiving. The peer review process as well as long-term archiving has to be established by the scientific community. They are not contained in the registration fee of TIB.
- The **WDCC publication example** (1 model experiment connected to 350 individual time series) results in the assignment of **1 Publ.-DOI and 350 Ident.-DOIs** and about **90 EUR for the registration**. The amount of data behind the publication is about **3 Tera Byte**. The time we spent at WDCC level for the **publication process** is small (1 week) compared to the production time (3-6 months).



Critical points are securing of **data quality and integrity** and the **stable connection** between identifier and data entity. This includes:

- Definition of **independent data entities** which are suitable for publication, **metadata** implementation (expert data description) and **long-term archiving** (**Publication Agent**)
- Scientific **quality assurance** is expected by peer review as part of the publication process (**Publication Agent**).
- **Stable connection** between persistent identifier and data entity must be ensured (**Registration Agency + Publication Agent**)
- Published primary **data are no longer matter of change**. They can be accessed without restrictions for science and they are fixed like published articles in scientific literature. (**Publication Agent**).

These responsibilities are included in a formal contract between Publication Agent and Registration Agency. For details refer: www.std-doi.de

Summary:

The German **WDC Cluster ESR** offers:

- **Long-term data storage** and data dissemination within the framework of each WDC
- **Project data management** within the thematic scope of each WDC
- **Primary data publication** within the STD-DOI concept. WDCs act as **Publications Agents**.



Further information



WDC Cluster ESR

WDC Climate:

<http://www.wdc-climate.de/>

WDC MARE

<http://www.wdc-mare.org/>

WDC RSAT

<http://wdc.dlr.de/>

WDC TERRA (candidate)

<http://www.gfz-potsdam.de/>

STD-DOI Concept:

<http://www.std-doi.de/>

TIBORDER Catalogue:

<http://twsgbv.de/>

(Primary Data Search: exk +primaerdaten)





End of Presentation

Questions?

