CERA2 Metadata Submission Guide

Description of metadata submissions for DKRZ Long-Term Archive (DKRZ-LTA)

Revision	Author	Scope
July-2016	DKRZ Datamanagement	Public release

Contents

Intended Audience

DKRZ-LTA Users, DKRZ-LTA Data Manager

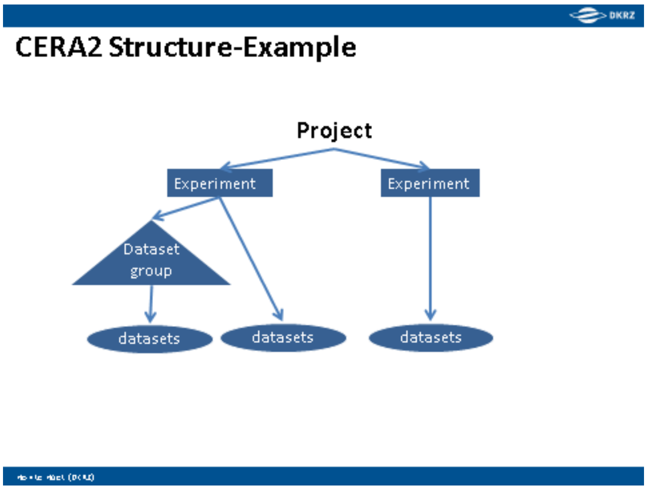
Introduction

The CERA2 metadata standard is used to document data objects archived within DKRZ-LTA. This document contains a description of the parts of the CERA standard relevant for metadata submission, especially the so-called CERA2 hierarchy and the individual elements. Information on the CERA2 metadata is helpful to prepare metadata and data for the archival process[1].

The data, which has been archived under certain rules and standards can be made 'citable' similar to articles in papers. The data will be registered and will get a DOI (Digital Object Identifier).

CERA-2 Hierarchy Structure Description

Data in DKRZ-LTA is organized in multiple layers with a hierarchical structure. These layers are: Projects, Experiments, Dataset groups and Datasets. An option to describe relationships between these different layers is a tree structure – hierarchical order see picture 1. DKRZ-LTA supports single-parent (mono-hierarchical) relationships only.



CERA2 Structure-Example

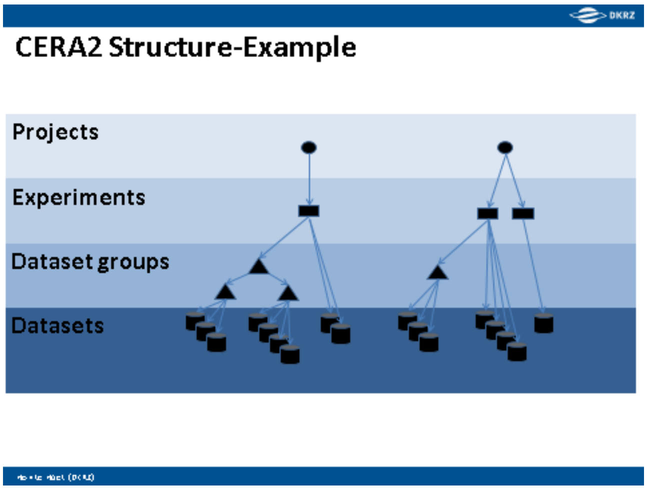
There is only one parent allowed for each element in the hierarchy. A CERA2 project is the top hierarchy element.

Picture 1

The Project, Experiment and Dataset groups contain the overall descriptions of the data. The datasets contain the description of the data and the climate data themselves. Therefore the DKRZ-LTA database CERA contains not only the metadata but also the data (digital object).

In the framework of a hierarchical structure different combination of connections are possible. As an example see picture 2.

[1] http://cera-www.dkrz.de/docs/DataSubmissionPreparationGuide.pdf



CERA2 Structure-Example

An Experiment or dataset groups contains zero or more dataset groups or zero or more datasets. A dataset has only a single parent.

Picture 2

Definition of Layers

Project: A CERA2-project is a scientific research activity consisting of experiments. It enables data depositors to organize their collections of data from a number of potentially different data sources. Examples: IPCC-AR5_CMIP5, HZG, CliSAP.

Please refer to links provided in the Data Publication Examples.

Experiment: A CERA-2 experiment is a compilation of datasets or groups of datasets. For example the time series of physical quantities like wind or precipitation from different datasets belonging to the same scientific experiment.

Dataset_group: A CERA2 dataset group is a compilation of datasets or dataset groups. This element can be used to make finer distinctions within a higher-ranking compilation.

> **For data citation CERA2-experiments or CERA2 dataset groups can be used. For this a DataCite DOI assignment is possible.**

Datasets: A CERA2 dataset consists of the data itself (the bits) and its description (metadata) needed to make the data understandable to the designated community. A dataset is assigned to an experiment or a dataset group.

Add_Info: A CERA2 additional_info is defined as a compilation of documents or plots enhancing further understanding of the datasets.

Data Publication Examples:

Model Data

IPCC-AR5_CMIP5

CMIP5 simulations of the Max Planck Institute for Meteorology (MPI-M) based on the MPI-ESM-P model: The historical experiment, served by ESGF
http://dx.doi.org/doi:10.1594/WDCC/CMIP5.MXEPhi

Climate Model Simulations at Helmholtz-Zentrum Geesthacht (HZG)

Regional climate model hindcasts for Siberia over the last decades based on the COSMO-CLM model driven by NCEP1 and ERA-40, run by Helmholtz-Zentrum Geesthacht
http://dx.doi.org/doi:10.1594/WDCC/COSMO-CLM_siberia

Observational Data

Integrated Climate System Analysis and Prediction(CliSAP)

Gridded Melt Pond Cover Fraction on Arctic Sea Ice derived from TERRA-MODIS 8-day composite Reflectance Data bias corrected Version 02
http://dx.doi.org/doi:10.1594/WDCC/MODIS__Arctic__MPF_V02

Graphical User Interface for Metadata Submission

To prepare and submit the metadata required for data publication at DKRZ-LTA a web application[2] is provided. Information on how to use this application can be found on the Intro/Help page[1]

The metadata submission GUI provides a key part of the data publication process of data at the World Data Center for Climate (WDCC) at DKRZ-LTA by supporting the metadata ingest of the CERA2 levels Project-Experiment-Datasets. We also support an interface for metadata batch xml submission of CERA2 levels Project-Experiment-Dataset_group-Datasets. Please contact data@dkrz.de if you want to know more about it.

A persistent identifier DOI can be applied for CERA2 Exp eriment or CERA2 Dataset_group after long-term archiving of the data and metadata.

To start the DKRZ-LTA publication process certain requirements exist which are described in the “Data Submission Preparation Guide”.[3]

[2] http://cera-www.dkrz.de/LTA_metadata
[3] http://cera-www.dkrz.de/docs/DataSubmissionPreparationGuide.pdf

Description of relevant CERA2 Elements

This section is divided into parts which correspond to the sections (Tabs) in the submission GUI. Information entered for these CERA2 metadata elements will be displayed in the CERA Search GUI[4] and will also be harvested by external metadata repositories.

The metadata submission GUI does not differentiate between the levels experiment, dataset_group and dataset. For all Experiments and Dataset_groups which receive a citation a separate block Metadata_Entry needs to be filled out.

For all other entries(Datasets and Add infos) this description is optional but recommended.

All characters defined in ISO-8859-1 are allowed unless otherwise stated.

WDCC/CERA2 Submission Form

The core metadata properties are chosen for accurate and consistent data identification in citations and for data retrieval, along with recommended use instructions.

Block: Metadata Entry		
Experiment or Dataset_group description: all fields are required. **For data citation CERA2-experiments or CERA2 dataset groups can be used**.		
Field	**Description**	**Extra**
Entry Name	The Entry Name uniquely identifies the data and metadata. This is the name element of the CERA-2 entity.	160 bytes
Summary	Information on the entry that does not fit into any other field. This could be the motivation for conducting the experiment, models. … This information may contain links to further information unless this information is not entered as an additional info. URL links should be followed by a blank or “)”.	4000 bytes.
Authors	The main researchers involved in producing the data, or the authors of the data publication. The complete citation will be created with the elements: [author(s)][PublicationYear];[Entry Name][Publisher].[Identifier] on the WDC-Climate internet page for CERA2-experiments or CERA2 dataset groups. For this a DataCite DOI assignment is possible see	

[4] http://cera-www.dkrz.de/WDCC/ui

	Section: DOI Submission.
Investigator	A person who created the resource or is involved in analyzing data or the results of an experiment or formal study. The person is responsible for the data and he/she is the contact person on the WDC-Climate internet page in case no other contact is given.
Metadata	Party who is the author of metadata

Block: Temporal coverage		
This information is required to classify your data and to help potential users decide whether your entry is of interest to them. The time indicated by the temporal coverage is the time the data refers to.		
Field	**Description**	**Extra**
Start Year	First year of the time period	Integer
Start Month	First month of the time period	Positive integer
Start Day	First day of the time period	Positive integer
Stop Year	Last year of the time period	Integer
Stop Month	Last month of the time period	Positive integer
Stop Day	Last day of the time period	Positive integer
Currentness Ref	Basis, for determining the time period covered by the data, e.g. calendrical, model time, 365 days per year.	2000 bytes.

Block: Spatial Coverage		
The values for spatial coverage contain information on the volume of space that the data relates to. This information is required to classify the data and to help potential users decide whether the entries are of interest to them. The spatial coverage of your data is restricted to a rectangular area that is defined by the northern, eastern, southern and western geographical coordinates. Values shall be expressed in decimal degrees. It is required that MinLat <= MaxLat and MinLon <= MaxLon. The spatial coverage also has to be specified with respect to its vertical dimension. If your data entities use rotated grids the derotated coordinates shall be entered here.		
Field	**Description**	**Extra**
Min Lat	northernmost latitude (south of Equator negative)	-90 <= MinLat <= 90
Max Lat	southernmost latitude(south of Equator negative)	-90 <= MaxLat <= 90
Min Lon	westernmost longitude (west of Greenwich negative)	-180 <= MinLon <= 180
Max Lon	the easternmost longitude (west of Greenwich negative)	-180 <= MaxLon <= 180

Min Altitude	The lowest depth/altitude of the vertical coverage as lower value	Float
Min Alt Unit	Unit of ‘Min Altitude’	LOV
Max Altitude	The highest depth/altitude of the vertical coverage	Float
Max Alt Unit	Unit of ‘Max Altitude’	LOV

Block: Data Description		
Format of the data.		
Field	**Description**	**Extra**
Format	Format of the data. For a description of supported formats please have a look at http://cera-www.dkrz.de/docs/DKRZ-LTA-Formats.pdf.	LOV[5] If your format is not in the selection list please contact data@dkrz.de 2000 bytes

Block: Quality	
The sole responsibility for scientific quality lies with the data producer= author and the level of quality must be approved by them. Use this block to provide summarized information on quality assurance measures that have been applied to the publication entity.	
Field	**Description**
Accuracy Report	The accuracy report should contain information on the procedure of data quality checking and its findings. Information on quality that does not fit into the space provided for the summarized description of quality checks should be submitted as an attached file, e.g. details of the procedure, quality check protocols, images of the quality check findings, etc. The results of the quality checks have to be attached as additional information. See Additional Documentation Block Upload Area. • For model data the accuracy report should contain a link to the homepage of the model documentation, short descriptions of the resolution, parameterization, boundary conditions, input data, and constants of the experiment. • For observational data the accuracy report should contain the data level e.g. http://www.godae.org/Data-definition.html. It should include description of campaigns, supersites, resolution in time and space, instruments and platforms. The accuracy report should be completed with the confirmation of the contact person approval: “SQA – Scientific Quality Assurance ‘approved by contact’ “

[5] List of Values (predefined)

	<date>.
Completeness Report	The completeness report should contain a summary of the temporal, spatial and parameter completeness with description about how missing values are indicated (if applicable).

Block: New Project		
A CERA2-project is a scientific research activity consisting of experiments. It enables data depositors to organize their collections of data from a number of potentially different data sources.		
Field	**Description**	**Extra**
Project Acronym	Abbreviated version of the project's name	31 bytes, only alphanumeric and “/” “_” “-“ “ “are allowed.
Project Name	Name of the project during which the data were generated.	250 bytes
Project Description	Summary of the project. If possible, provide a link to the homepage of the project, describe its scientific ideas and name its funders.	2000 bytes.

Block: New Person	
It will be used for contact purposes and will also be provided to any other user of the data.	
Field	**Extra**
First Name	Required, 80 bytes
Second Name	80 bytes
Last Name	required, 80 bytes
Title	31 bytes
Institute	required, from LOV
Telephone	80 bytes
Fax	80 bytes
Url	80 bytes
Email	Required, 80 bytes

Block: New Institute	
It will be used for contact purposes and will also be provided to any other user of the data.	
Field	**Extra**
Institute Name	required, 80 bytess
Institute Acronym	required, 31 bytes
Department Name	80 bytes
Department Acronym	31 bytes
Country	required, 80 bytes
State Or Province	80 bytes
Place	80 bytes

Street		80 bytes
Street Postal Code		10 bytes
Pobox		80 bytes
Pobox Postal Code		10 bytes
Url		80 bytes
Additional Info		250 bytes

Block: New Citation		
Publications in conjunction with the data. Relations should be used to build up a system supporting data discovery that is based on the specification of different kinds of cross-references to other data or print publications.		
Field	**Description**	**Extra**
Title		required, 2000 bytes
Authors		required, 2000 bytes
Publication		80 bytes
Publisher		31 bytes
Editor		200 bytes
Publication Date		required, Date
Country		80 bytes
State		80 bytes
Place		required, 80 bytes
Edition		80 bytes
Access Spec	E.g. doi:…	2000 bytes
Additional Info		250 bytes

Additional Documentation

This documentation should give adequate information about what data is included, the data quality and how it is structured thus enhancing further understanding of the datasets, e.g. a readme or a methodology report.
For this detailed documentation it is possible to upload files in the Upload Area. After publication the files are stored in the CERA database as Add_infos.
In case of published documentation available as print publications the 'Block New Citation' should be used to build up a system of references to the data e.g. references to evaluation results (data) and methods.

Block: Upload Area – Additional Infos		
The upload area is used to upload additional files (not the data itself!) containing the additional documentation or plots. It is not required that each of the fields is addressed by a separate file.		
Field	**Description**	**Extra**
Add_info files: quality checking	The Add_info files should contain information on the procedure of data quality checking and its finding submitted as attached files, e.g. details of the procedure, quality check protocols, images of the quality check findings, etc. It should include documentation about:	Attached files preferred format: pdf

	• How are missing values indicated • Documented procedure of statistical quality control (Examples see Appendix 1) • Estimation of technical (Examples see Appendix 2) and methodological errors and deviations • Documented procedure with validation against independent data	
Add_info files: model data methodology report	It should include description of the models, components and their equations (link to model homepage). Detailed description of simulations with: • Resolution in time and space, dependencies of time and space resolutions. • Structure – grid description, extraction possibilities • Boundary conditions – forcing • Input data • Constants – for initialization and run e.g. orography, solar constant, drag coefficient, area leaf index • Information about benchmark tests and the reproducibility of simulation runs Description of family trees of models like: http://www.gfdl.noaa.gov/jrl_gcm or http://www.aip.org/history/climate/xAGCMtree.htm	Attached files preferred format: pdf
Add_info files: observational methodology report and data level classification	It should include description of campaigns, supersites, resolution in time and space, instruments and platforms. For example see: https://icdc.zmaw.de/fileadmin/user_upload/HDCP2_Docs/hdcp2_obs_data_product_standard_v2.2.pdf Detailed description of classification into a level system e.g. http://www.godae.org/Data-definition.html.	Attached files preferred format: pdf
Add_info files: readme	How to use the data for example see: http://cera-www.dkrz.de/WDCC/ui/Entry.jsp?acronym=CLM_README_2010	Attached file preferred format: pdf

DOI Submission

A persistent identifier DOI can be applied for CERA2 Experiment or CERA2 Dataset_group after long-term archiving of the data and metadata.

For the data publication process with DOI assignment the data and metadata will be reviewed again by the DOI publication agent[6] at WDCC.

In the event of questions occurring during this additional review process and coordination of DOI author list, DOI-contact and title, the contact person will be informed by e-mail. If the agent accepts the data, the entry will be assigned a DOI via the DataCite[7] registration service.

[6] https://www.dkrz.de/daten/data-services/datenpublikation/koncept-datacite-tib-metadatakernel

Block: DOI		
The data, which has been archived under certain rules and standards can be made 'citable' similar to articles in papers. The data will be registered and will get a DOI (Digital Object Identifier).		
Field	**Description**	**Extra**
Authors	The main researchers involved in producing the data, or the authors of the publication, in priority order. The complete citation will be created with the elements: [author(s)][(PublicationYear)];[Title][Publisher].[DOI] on the WDC-Climate internet page for CERA2-experiments or CERA2 dataset groups.	
Title	The title uniquely identifies the data and metadata. This is the title element of the DataCite publication citation. The complete citation will be created with the elements: [author(s)][(PublicationYear)];[Title][Publisher].[DOI] on the WDC-Climate internet page for CERA2-experiments or CERA2 dataset groups.	
DOI-Contact	Party who can be contacted for acquiring scientific knowledge about the resource. The person is responsible for the data and he/she is the contact person on the WDC-Climate internet page. At least one person is required as contact person for a DOI.	

[7] http://www.datacite.org

Appendices

Appendix 1: Examples of statistical quality control tests

a) **Rough Errors Tests**

- LIM-test by Meek and Hatfield[8] are applied. (The test checks every data point on whether it exceeds a predefined range of values.)
- NOC-test by Meek and Hatfield[8] are applied. (The test checks on whether data does not change for more than a predefined number of values. It can be used to detect errors of instrument.)
- ROC-test by Meek and Hatfield[8] are applied. (The test checks the rate of change. The difference between two consecutive elements is checked concerning limits.)

b) **Tests for systematic deviations in time and space (e.g. changes in mean, variance and trends) and random errors**
e.g.: Düsterhus, A. and Hense, A.: Advanced information criterion for environmental data quality assurance, Adv. Sci. Res., 8, 99-104, doi:10.5194/asr-8-99-2012 , 2012.

Appendix 2: Examples of technical sources of errors and deviations

Documentation of benchmark tests:

- Computer specification
- Computing center name
- Question: Under which conditions is it possible to repeat the simulation runs?

Condition examples:

- source code
  Robustness in extreme values
  Exception for overflow and zero
  Used units e.g. mm/d or m/sec
  Errors like latitude, longitude, rotated pole grid conversion confusion
  Statistic problems – samples over time versus instantaneous values
  Unit conversion and confusion
- Compiler
- Software library
- Number of processors
- Computer accuracy – numerical stability for very large and small values

Contacts

In case any questions regarding metadata submission arise, please contact DKRZ-LTA user support at data@dkrz.de .

[8] Meek, D. Hatfield, J. (1994) Data quality checking for single station meteorological databases. Agricultural and Forest Meteorology - AGR FOREST METEOROL , vol. 69, no. 1-2, pp. 85-109, DOI: 10.1016/0168-1923(94)90083-3