DOKU User Guide

(Version 1.0)

Long Term Archive (LTA) group

July 19, 2022

Contents

1 Introduction

DOKU offers long-term archiving for Earth System data and climate-related products. It is hosted by the German Climate Computing Center (DKRZ) in Hamburg and is maintained by its Data Management (DM) department.

DOKU is a service that is only available for DKRZ HPC (high performance computing) projects with a quota for long-term archiving. In comparison to WDCC, DOKU is the simpler solution to preserve your research data while still meeting the requirements of good scientific practice.

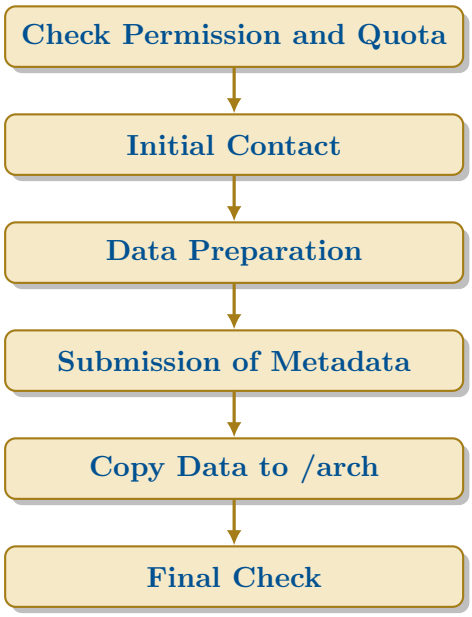
This document supports the DOKU Step-by-Step Guide (https://www.wdc-climate.de/WDCC/ui/cerasearch/info?site=guide_doku) and aims at guiding users through the process of archiving data in DOKU.

1.1 Submission Steps

DKRZ's Data Management (DM) will support you, starting from your initial intent to archive data until they are archived in DOKU.

Before you start the submission process in DOKU, please make sure that you have the authorization to publish the data.

In DOKU the submission process includes the following steps, which are described in detail in the remainder of this document:



2 Check Permission and Quota

Only data that is part of a DKRZ project can be archived in DOKU. Before the archiving process is started, you therefore need to make sure that your data belongs to a project that requested and was granted quota for DOKU.

For that log in to https://luv.dkrz.de with your DKRZ username and password. Under "Projects" all projects that you are a member of are listed. By clicking on the correct project all current allocations are listed. For DOKU the quota for "Archive Long Term" is of importance.

If you find on luv that you have quota for DOKU in your project remaining (and have discussed in your group/with your PI that you may use it), you can start the archiving process.

General information about resource allocation and how to apply for it (including DOKU-quota) at DKRZ can be found here: https://docs.dkrz.de/doc/getting_started/resource-allocation.html#resource-allocation.

3 Initial Contact

Whenever you wish to archive data in the DOKU, please get in contact with DKRZ Data Management (DM) by writing an email to data@dkrz.de. In this email state the project (DKRZ account number) you would like to archive data under and the amount of data (approximate number of files and size).

Your first email to data@dkrz.de will open a ticket in a request tracker. DM will contact you shortly with further instructions. Please do not write a new email, but only answer to this thread. Otherwise a new ticket would be opened.

4 Data Preparation

Please prepare your data files for submission. In the following, you can check what to take into consideration.

Most importantly you need to structure and document your data in a way that reusability is ensured even in years' time.

4.1 Data Format

Only open source data formats are accepted in DOKU. The preferred file format for archiving is NetCDF (Network Common Data Format). More information on NetCDF can be found at https://www.unidata.ucar.edu/software/netcdf/.

Other supported data formats are for example GRIB (GRIdded Binary), CSV, and ASCII. However, in many cases it might be sensible to create tar-bundles from your files.

4.2 File Size

The preferred file size is 10 GB to 100 GB. There is no lower limit, but small files are not optimal for tape storage and each file smaller than 1 GB will be charged 1 GB of quota. Therefore, it is encouraged to pack small files into tar-bundles. Files larger than 100 GB are possible. Even a file of a size of a few TB could be possible, but it is recommended not to exceed 500 GB.

4.3 Labeling of files and directories

Files should be labeled in a consistent manner. Special characters and spaces in directory and file names should be avoided. The file names have to be unique (even if spread over multiple directories). Please also take care that the file names describe the content, so that others are able to "read" the names.

5 Submission of Metadata – MetaXA Guide

The metadata belonging to your data can be entered via the graphical user interface, MetaXA (https://www.wdc-climate.de/metaXA). Before you can use this application, you need an WDCC account. If you do not have one already, please apply for an account here: https://www.wdc-climate.de/WDCC/ui/cerasearch/register.

In MetaXA you can enter metadata for DOKU-Datasets. Furthermore, you can add (or edit) a Person, add an Additional Info or a Citation/Reference. These elements will explained in the following paragraphs in the order in which they appear in MetaXA.

Only marked fields are required, but try to fill out as many details as possible. Help texts are available directly in MetaXA (via the little `?`) and more details for each field are given in the Appendix A in the respective tables.

5.1 Person

Please start with the Person entry. Here you should enter all persons relevant to your data (e.g. all authors who should be listed in the citation). Via the search form you can check if the person in question is already listed. Please do so to avoid duplicate entries. If you find the person in the database, check if all information is still correct. Make corrections, if necessary.

If you cannot find the person, you need to click the button "Add new person". Please enter the details belonging to the person here.

External Identifier: We would like to get the ORCID from all users (even though it is not a required field). If you do not have an ORCID, you can get it here: https://orcid.org. Alternatively (or additionally) you can also add the ID for ResearchGate or ResearcherID.

Please enter it in the "External Identifier" field. For that you need to choose the correct identifier type in the first row and enter the corresponding ID in the second row.

Institute: When entering information for a person, please choose an institute from the list. If you do not find the correct institute, click on "Other (new) Institute" and enter the details for the institute here.

5.2 DOKU-Dataset

There are different methods to enter metadata for Datasets. If you have only a few DOKU-Datasets you will use MetaXA. Otherwise (only for a large number of DOKU-Datasets), you can discuss with DM whether submission via CSV-lists might better.

5.2.1 MetaXA

The following elements are important for DOKU-Datasets. The tables in Appendix A will help you to enter metadata into MetaXA.

Once you have finished editing your metadata and want to inform DM that you have completed all entries for the DOKU-Dataset, click on "Notify Completion". If, later on, changes need to be made, this is still possible.

Metadata Entry: Please refer to table A1 for a detailed description.

Data Citation: Please refer to table A2 for a detailed description.

Temporal Coverage: This block is only available if the box "Data has no temporal information" has NOT been checked. Please refer to table A3 for a detailed description.

Spatial Coverage: This block is only available if the box "Data has no spatial information" has NOT been checked. Please refer to table A4 for a detailed description.

5.2.2 CSV-Lists

In case you have many Datasets that do not differ strongly from each other, it might be more convenient to create a CSV-list containing all relevant metadata. DM will assist you with this and will give you further instructions in case this applies to your data.

5.3 Additional Info

Fewer fields are needed for the Additional Info (please refer to Appendix C for an explanation of the different hierarchical elements). Please do not forget to go to "Notify Completion" once you are finished.

Metadata Entry: Please refer to table A5 for a detailed description.

Citation Information: Please refer to table A6 for a detailed description.

File upload: The Additional Info can consist of one or more files. You can upload the file(s) directly in MetaXA. The preferred file format for Additional Infos is PDF. If you cannot upload your file, this is most probably caused by an unsupported file format. Please convert it to PDF or contact DM to discuss if it possible to upload your preferred file format.

5.4 Citation / Reference

You can add references to your entries. Use the upper left field to search for existing citations and update them if necessary.

If you want to add new citations, use the button in the upper right corner "Create Reference Template". This opens the form to enter metadata for a new Citation/Reference. If a DOI is available, please enter it in the respective field in the format `doi:10.1045/january2015-brase` and click on "Insert from CrossRef". Otherwise, fill in the fields below (required fields are marked with a red asterisk).

If you edit or add a reference via MetaXA, you can also connect it with your entry. If you have already entered the DOKU-Dataset in MetaXA you will find it in a list and can choose one or more of the entries. Otherwise contact DM and they will assist you in connecting the reference to your entries.

5.5 Notify Completion

Once you have completed entering all relevant metadata for an entry and want to notify DM, please do not forget to go to "Notify Completion". This needs to be done for DOKU-Datasets, and Additional Infos. Persons and Citations can be added without "Notify Completion".

Check all appropriate entries and click on "Finalize Selected Entries". After you have finished your work with "Notify Completion", DM will further process your entries.

After this changes can still be made. Please contact DM, if you find that amendments are needed.

6 Copy Data to /arch

For your data to be archived in /doku, you need to transfer it to the tape archive first. It needs to be located in /arch under the correct DKRZ project directory. There, please create for each DOKU-DS a separate subdirectory that contains all files to be archived, but nothing else. The paths will look like this: `/arch/bm0000/path/to/your/data/<name_of_first_dataset>/`, `/arch/bm0000/path/to/your/data/<name_of_second_dataset>/`, ...

Further documentation on how to use the tape archive is provided here: https://docs.dkrz.de/doc/datastorage/hsm/man_pages.html

Once your data is located in /arch in the correct subdirectories, please write to DM to tell them:

- the data path for each DOKU-Dataset
- the number of files for each DOKU-Dataset
- the approximate size for each DOKU-Dataset

With this information, DM can start the process of data filling.

7 Final Check

Once your data has been filled and is set to completely archived, DM will inform you and ask you to make final checks. Please check the metadata one last time. Also see if all your files are ready for download. It might be also advisable to try to download at least one sample file to see if everything works well.

8 How to cite your data

All DOKU-Datasets are assigned a persistent URL once the metadata is online. So even before the data filling processes has been completed, the URL is final and can, for example, be used for publications. For journals it is often sufficient to provide a persistent URL with the metadata when submitting a paper – of course, the data needs to follow.

8.1 PIDs for DOKU-Datasets

Once the data is available for download the DOKU-Datasets are assigned a PID (persistent identifier). This is a unique code of digits and letters that is assigned to your DOKU-Dataset

to make it better citable. A PID is even more suitable for referencing your data in a journal publication than a persistent URL.

Appendix A: Description of MetaXA fields

List of Tables

Table A1: **DOKU-Datasets – Metadata Entry**

Field	Required	Comment
DKRZ Project Number	✓	The internal DKRZ Project Number, eg. "bm0999"
Name	✓	The entry name uniquely identifies the metadata entry (DOKU-Dataset). Maximum length: 160 characters. The name should meet the following criteria: • The first letter should be in upper-case • Do not use underscores if not necessary • Whitespaces are the preferred word delimiters • It should provide proper information about the Entry • Give non-scientists an idea about the content • Avoid constant change of letters and numbers as well as constant change of upper and lower case letters. This will worsen the searchability. • If a version number needs to be provided, use the following format: Title of the entry (Version 2.1)
Summary	✓	It should be described here what is covered by the DOKU-Dataset. Give information on the entry that has not been entered into any other fields. The summary may contain links to further information. Maximum length: 4000 characters.
Responsible Person		The person selected here will be responsible for the metadata published for the DOKU entry.
Use Constraint (License)		"Use Constraint" reflects constraints regarding the use of the data. Refer to Appendix B for further details.
Format		Select format of data from the list. For details see section about Data Format.
Data has no temporal/spatial information		Check these boxes (or one of these) if the data has no temporal and/or no spatial information.

Table A2: **DOKU-Datasets – Data Citation**

Field	Required	Comment
Citation Title		The title of the citation. Please follow the same rules as for the entry names. The entry name can also be used here. If left empty, the entry will have no citation.
Authors List		The authors of this data for the citation. Format: `lastname,firstname[; lastname, firstname; ...]` Cite by institute is also possible.

Table A3: **DOKU-Datasets – Temporal Coverage**

Field	Required	Comment
Start Year	✓	The first year of the considered time period.
Month	✓	The first month of the considered time period.
Day	✓	The first day of the considered time period.
Stop Year	✓	The last day of the considered time period.
Month	✓	The last day of the considered time period.
Day	✓	The last day of the considered time period.
Currentness Reference		Currentness Reference reflects the construction of the time axis. Choose the correct calendar from the list. Use "not filled" if no entry matches.

Table A4: **DOKU-Datasets – Spatial Coverage**

Field	Required	Comment
Min Lon	✓	The minimum longitude of the considered domain [0.0, 360.0].
Max Lon	✓	The maximum longitude of the considered domain [0.0, 360.0].
Min Lat	✓	The minimum latitude of the considered domain [-90.0, 90.0].
Max Lat	✓	The maximum latitude of the considered domain [-90.0, 90.0].
Min Altitude		Minimum altitude. Example for ocean/atmosphere data: -6200m (ocean) to 10hPa (atmosphere)
Min Alt Unit		Unit for minimum altitude.
Max Altitude		Maximum altitude. Example for ocean/atmosphere data: -6200m (ocean) to 10hPa (atmosphere)
Max Alt Unit		Unit for maximum altitude.

Table A5: **Additional Infos – Metadata Entry**

Field	Required	Comment
Name	✓	The entry name uniquely identifies the metadata entry (Additional Info). Maximum length: 160 characters. The name should meet the following criteria: • The first letter should be in upper-case • Do not use underscores if not necessary • Whitespaces are the preferred word delimiters • It should provide proper information about the Entry • Give non-scientists an idea about the content • Avoid constant change of letters and numbers as well as constant change of upper and lower case letters. This will worsen the searchability. • If a version number needs to be provided, use the following format: Title of the entry (Version 2.1)
Summary	✓	It should be described here what is covered by the Additional Info. Give information on the entry that has not been entered into any other fields. The summary may contain links to further information. Maximum length: 4000 characters.
Parent	✓	Select an Experiment (or Dataset group) to be the parent of this entry. If you have already entered the element it should appear in this list.
Publication Type		If your publication is well described by an "additional info", but something from this list fits better, you can specify that here.
Use Constraint (License)		"Use Constraint" reflects constraints regarding the use of the data. Refer to Appendix B for further details.

Table A6: **Additional Infos – Citation Information**

Field	Required	Comment
Citation Title		The title of the citation. Please follow the same rules as for the entry names. The entry name can also be used here. If left empty, the entry will have no citation.
Authors		The authors of this additional info for the citation. Please select from the list.

Appendix B: License

You need to decide under which license to publish your data. We suggest CC-BY 4.0 for data submissions. A short overview is given here.

B.1 Creative Commons Attribution 4.0 International

→ https://creativecommons.org/licenses/by/4.0/

→ https://creativecommons.org/licenses/by-nd/4.0/

→ https://creativecommons.org/licenses/by-nc/4.0/

→ https://creativecommons.org/licenses/by-nc-nd/4.0/

→ https://creativecommons.org/licenses/by-nc-sa/4.0/

→ https://creativecommons.org/licenses/by-sa/4.0/

Summary (human readable): To see which terms apply to which Creative Common license in detail, please refer to table B1.

You are free to (green in table B1):

- *Share* – copy and redistribute the material in any medium or format
- *Adapt* – remix, transform, and build upon the material.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms (blue in table B1):

- *Attribution* – You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- *NoDerivatives* – If you remix, transform, or build upon the material, you may not distribute the modified material.
- *NonCommercial* – You may not use the material for commercial purposes.
- *ShareAlike* – If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions: You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices: You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

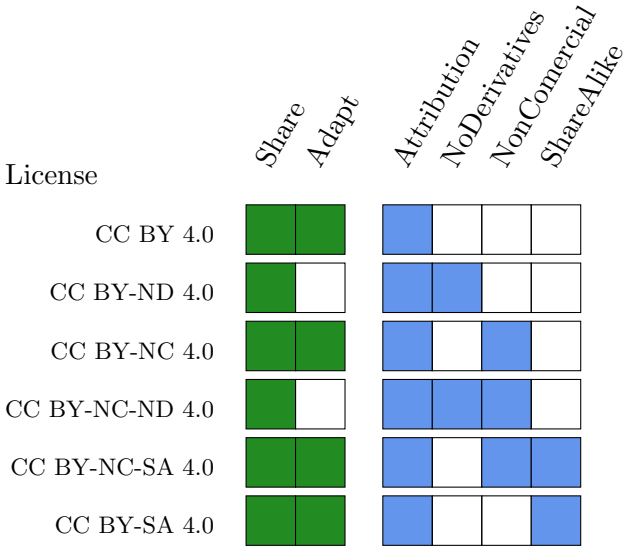


Table B1: **Summary of Creative Commons Attribution 4.0 International**

Appendix C: Hierarchy Structure Description

Data in DOKU is organized in multiple layers with a hierarchical structure. These layers are named: *Projects*, *Experiments*, *Dataset groups* and *Datasets*. (Even though Dataset groups are currently not used in DOKU.) The relationships between them can be described with a tree structure as in figure 1. For the top-level elements only single-parent (mono-hierarchical) relationships are supported. That means that only one parent is allowed for each Experiment and Dataset group, but a parent can have multiple children-elements. Datasets and *Additional Infos* (which are a bit of an exception and will be described later in this section) need to have at least one parent, but it is possible that they have more than one.

The Project, Experiment and Dataset groups contain the overall description of the data, i.e. the metadata. Metadata is the description of the data itself. Detailed metadata is essential for later reuse of the data and therefore for long-term archiving. These elements (Project, Experiment and Dataset groups) can therefore be used to structure the data. The data itself is located in the Datasets (which also contain description of the data).

Each DOKU-Dataset is part of a hierarchical tree with an Experiment and a Project. The most minimalistic data to be archived needs to contain at least one Project, one Experiment and one Dataset in the hierarchical structure.

Project: A Project is the top hierarchical element. All data that is archived under DOKU belongs to the same Project: "Long-term Archiving of Climate Model Data at WDC Climate and DKRZ (DOKU)" (https://cera-www.dkrz.de/WDCC/ui/cerasearch/project?acronym=DKRZ_lta).

Experiment: An Experiment is the hierarchical child element of a Project and a compilation of Datasets. In DOKU each DKRZ project (e.g. bm0000) becomes an Experiment and will compile all DOKU-Datasets that are archived under the project account bm0000. (It has to be noted that "Experiment" is just a name for a hierarchical element and should not be confused with a model run for example.)

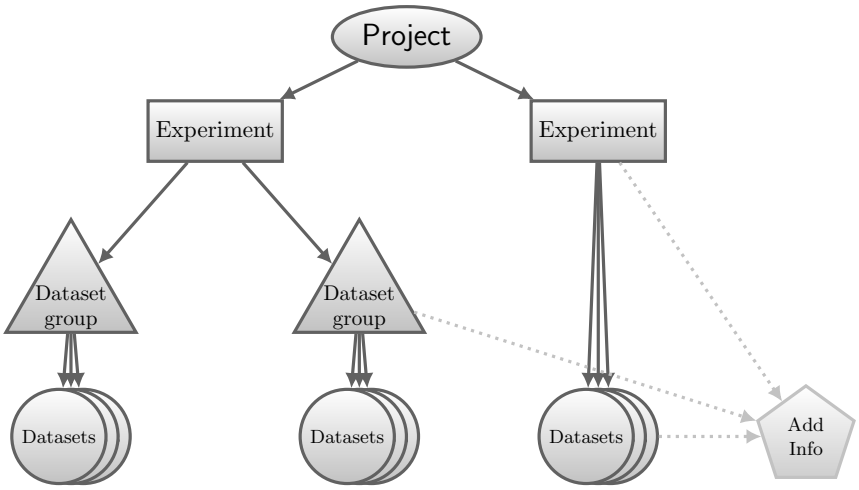


Figure 1: **Hierarchical structure elements**

Dataset: A Dataset consists of the data itself and its description (the metadata) needed to make the data understandable. A DOKU-Dataset is a child of an Experiment.

Additional Info: There is another element a bit outside of the hierarchical tree in figure 1. An Additional Info is an optional element that is located in the same layer as a Dataset, i.e. is the child of an Experiment (but can also be the child of a Dataset) and can have more than one parent. It is defined as a compilation of documents or plots enhancing further understanding of the data.

C.1 Nomenclature

DM will help you to organize your data into the hierarchical structure described above. In this context the nomenclature of the elements will also be discussed. You have to give each element a name which is the title for this entry and will also be used in the WDCC search interface. Hints for good entry names will be given in the section “Submission of Metadata”.

You can decide together with DM which structure and nomenclature makes sense for your data.