

Data Submission Preparation Guide

Recommendations on how to prepare data submissions for DKRZ Long-Term Archive (DKRZ-LTA)

Revision	Author	Scope
July-2016	DKRZ Data management	Public release

Intended Audience

DKRZ-LTA Users, DKRZ-LTA Data Manager

Introduction

How do you know if your data is ready to be long-term archived in DKRZ Long-Term Archive (DKRZ-LTA)? What do you need to think about in advance of depositing?

Before you deposit in DKRZ-LTA you need to organize and prepare your material, and decide on a number of things.

This document explains the steps which need to be accomplished before you submit your material. These steps are

- Verify data publication requirements.
- Structure the material.
- Prepare files.
- Prepare metadata and documentation.
- Prepare a task list.

Data Publication Requirements

These are the preconditions to start the data publication:

- Make sure you have the authorization to publish the data you would like to archive in DKRZ-LTA.
- Please contact DKRZ-LTA at data@dkrz.de to coordinate the publication process.

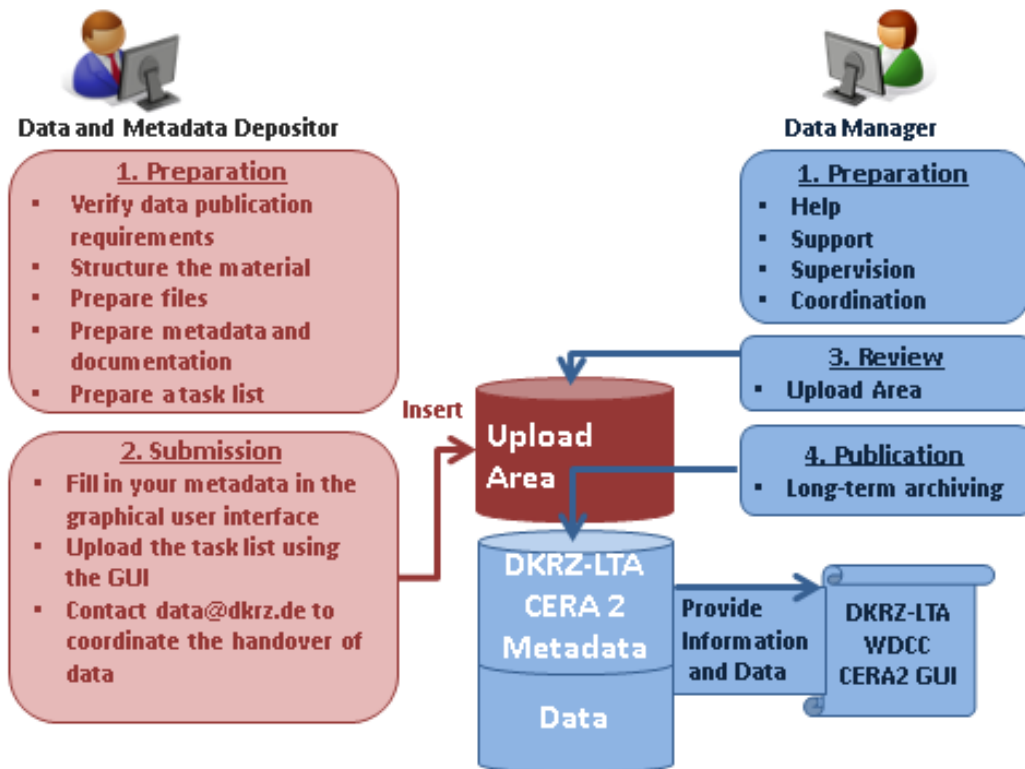


Figure 1: Steps performed for preparation and submission of data and metadata for DKRZ-LTA

Structure the material

Main questions are:

- Are you depositing an individual dataset or item?
- Are you depositing multiple datasets or items?

Think about the granularity of your data and the dependent ancillary files which compose your research data.

Decide on how to structure your research data for making reuse and proper citation possible. In some cases your data will be composed of different files or items that belong to one experiment, but in other cases your data might need to be segregated across a number of dataset groups or experiments in consultation with DKRZ-LTA. These layers are described in the CERA2 Metadata Submission Guide¹.

Decide about:

- ✓ Labelling of files and directories - does a controlled vocabulary in your project exist? If possible use CF-Conventions/CF-Standard Names.²

¹ <http://cera-www.dkrz.de/docs/CERA2MetadataSubmissionGuide.pdf>

² <http://cfconventions.org>

- ✓ Data versioning
- ✓ Data format in consultation with DKRZ-LTA.

Detailed guidance can be found in our documentation on recommended file formats³.

- ✓ Data compression

If you use data compression, make sure that the compression algorithm and/or software is under an open license and used consistently for all files. Do not apply any compression beforehand without contacting DKRZ data management in advance.

- ✓ Citation of your data

Think about the parties who should be considered in the citation. Should all items in this submission have the same citation?

Prepare files

Structure and Labelling of files and directory

Have you structured and labelled your data in a consistent manner?

- ✓ Is the structure correct in accordance with the data granularity?
- ✓ Are files and directories structured and named in a consistent and documented manner.
- ✓ Is your data dynamic in time? Are there any pending changes?
- ✓ Special characters and spaces in directory and file names should be avoided. If it is required to follow a specific sequence in archiving this has to be indicated as well.
- ✓ Ensure you labelled your data consistent with the controlled vocabulary.
- ✓ Ensure you labelled your data consistent with the versioning.

Create checksum

You may want to prepare checksums of all files before submitting your request for archival. This will help us to guarantee data integrity throughout our archival workflow. If you create checksums, we recommend using md5.

- ✓ Are checksums prepared for all files.

Format correct

- ✓ Are the formats and file extensions correct (checked with format checker)? We recommend applying a cf-netcdf checker⁴ if applicable.

³ <http://cera-www.dkrz.de/docs/DKRZ-LTA-Formats.pdf>

⁴ <http://cfconventions.org/compliance-checker.html>

Compression

- ✓ Is the compression correct?

File size

- ✓ Check the overall data size and make sure it fits to the data amount agreed with DKRZ-LTA.
- ✓ Are the file sizes feasible for access and transmission? If your files are larger than 2 GByte please contact us.
- ✓ Are the data sizes checked and correct? There should be no items with size 0.

Data Accuracy

- ✓ Are missing values indicated, e.g. with fill values or in the metadata?

Completeness

- ✓ Ensure your data is complete e.g. check start and end date, number of items.

Sequence of files

- ✓ Ensure that, if the data items need to be archived in a specific sequence, this is clearly documented.

Prepare metadata and documentation

Metadata in long-term archiving is essential for later data usage. DKRZ-LTA uses a metadata standard (CERA2) to accomplish the requirements for long-term archiving. This metadata standard is described in more detail in the CERA2 Metadata Submission Guide¹.

To prepare the metadata you should use the graphical user interface⁵. We also support an interface for batch submission of metadata. Please contact data@dkrz.de if you want to know more about it.

Before you start submitting your metadata you should make yourself aware of the CERA2 hierarchy which consists of 3 to 5 levels and is described in the CERA2 Metadata Submission Guide¹. You will be asked for further metadata information once you start submission of metadata using the GUI. The amount of information required depends on various factors, e.g. project specifications.

⁵ http://cera-www.dkrz.de/LTA_metadata

You should submit additional documentation, e.g. a readme or a methodology reports or any other relevant material later on in the submission process. This documentation should give adequate information about what data is included, its data quality, and how it is structured.

Decide about:

- ✓ Do you know how to map your data onto the CERA2 hierarchy?
- ✓ Have you checked the metadata policies of your project?
- ✓ Have you prepared information about the access rights correspond to project requirements?
- ✓ Do you have data citations ready e.g. titles and author lists of your structured data?
- ✓ Have you prepared references or links to research methodology reports, evaluation results and any other relevant information?

Prepare task list

To perform the long-term archiving of data files, it is necessary that the elements to be archived and their assignment to metadata in the long-term archive are described clearly and completely. To facilitate this process a format for archiving task lists is defined by DKRZ-LTA⁶.

- ✓ Prepare task list in defined format.

Submission of material

As described in this document the submission of your material comprises the following steps:

1. Check preconditions for publication.
2. Structure your materials and prepare files.
3. Fill in your metadata in the graphical user interface⁷.
4. Upload the task list using the GUI.
5. Contact data@dkrz.de to coordinate the handover of data.

Supplementary rules

Terms and conditions for the deposition of material in the DKRZ Long Term Archive are laid down in the depositor agreement⁸.

- ✓ Check if your submission fits the terms and conditions written down in the depositor agreement.

⁶ http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf

⁷ https://www.dkrz.de/daten-en/data-services/long_term_archiving/metadata-interface

⁸ <http://cera-www.dkrz.de/docs/DKRZ-LTA-DepositorAgreement.pdf>

Related documents

- Depositor agreement⁸
- Storage Policy⁹
- Supported data formats³
- Archiving task list¹⁰
- CERA2 Metadata Submission Guide¹
- Help on metadata submission interface¹¹

Contacts

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

⁹ <http://cera-www.dkrz.de/docs/DKRZ-LTA-PreservationAndStoragePolicy.pdf>

¹⁰ http://cera-www.dkrz.de/docs/Archiving_Task_List.pdf

¹¹ http://cera-www.dkrz.de/LTA_metadata