Data services at DKRZ

CLIMATE MODELLING - A DATA INTENSIVE SCIENCE

Climate models produce extremely large amounts of data. In order to manage these data production rates, DKRZ operates data lifecycle services. With the ICSU World Data Center Climate (WDCC), DKRZ runs a fully documented long-term data archive with a size of currently 3 Petabytes. The entire tape archive is currently equipped to handle data production rates of up to 10 Petabytes per year. With the upcoming HLRE-3 system it will even manage annual rates up to 75 Petabytes. Beside data storage resources, a seamless end-to-end workflow from data production over data processing, data dissemination to data storage is applied to make optimal use of the huge but nevertheless limited HPC resources at DKRZ.

1. Data management plan

The data time line as well as volumes, structures, access patterns and storage locations have to be defined as accurate as possible for each DKRZ HPC project in order to realize a seamless workflow and efficient use of DKRZ resources.

2. DKRZ storage

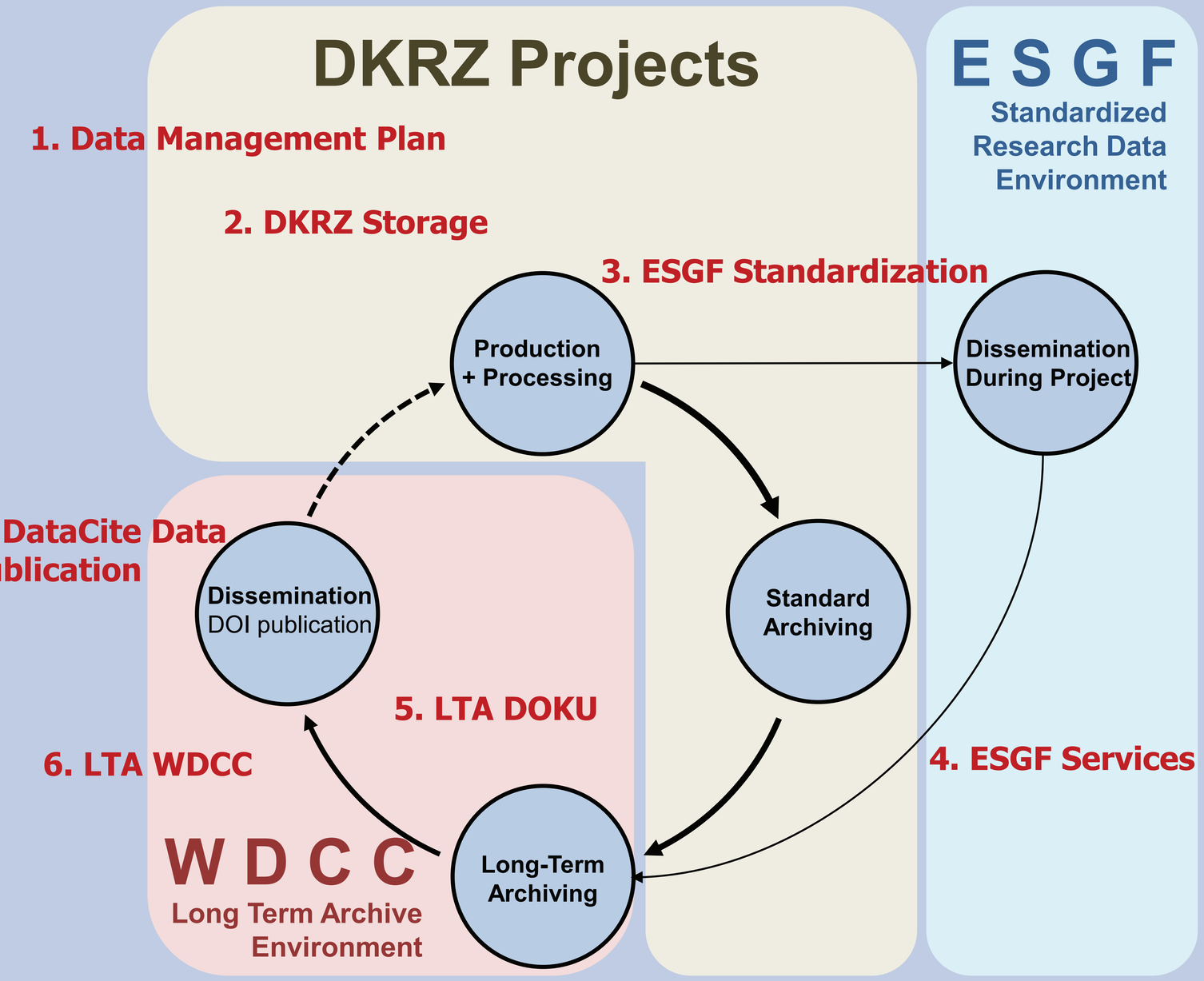
Each DKRZ HPC project has to specify and to apply for compute and storage resources on an annual basis. Storage resources contain disc and tape storage. All resources are monitored on the basis of DKRZ HPC projects.

3. ESGF standardization

Climate data integration into ESGF (Earth System Grid Federation) requires standardization in order to make data intercomparable within the federation. This data preparation process includes project specifications as well as adaptation of data and metadata to the ESGF data publication interface.

4. ESGF services

DKRZ offers a number of services to integrate („publish"), manage, discover and access climate data in the international ESGF. The data allocation includes definition of project specific publication and access policies, adaptation of data check routines and the data publication on the ESGF data node at DKRZ. User support for data publication and data access is included as well.



5. LTA DOKU

LTA DOKU stands for in-house longterm archiving in the DOKU(mentation) section of the tape archive at DKRZ. This service offers longterm archiving for data of DKRZ HPC projects as internal reference data only. A minimum set of metadata has to be supplied by data providers in order to characterize and identify them in the longterm archive of DKRZ. No additional information on data interpretation is provided. Focus here is set on internal data access from data providers.

6. LTA WDCC

LTA WDCC stands for longterm archiving in the World Data Center for Climate (WDCC). This service is open for data from DKRZ HPC projects, data from ESGF but also for data from outside DKRZ. These data are fully integrated in the database system of the WDCC. The full set of metadata is provided in order to allow data interpretation even after ten years or more without contacting the data author. Connected to this archiving service is a fine granular data storage whi ch allows field based data access (CERA container files) in distinction to the file based data access in the LTA DOKU service. Focus here is set on interdisciplinary data access.

7. DataCite data publication

DataCite is an international organization which aims to establish easier access to research data, increase the acceptance of research data as legitimate contribution in the scholarly record, and support data archiving to permit results to be re-used. Scientists are enabled to give and to get credit for the preparation of data products by formal data citations.

All data from the LTA WDCC service, that have passed a final quality assurance procedure are suitable for a DataCite data publication, i.e. citation metadata are published and a DOI (Digital Object Identifier) is minted. After receiving a DOI the data and key metadata remain unchanged, and the data is persistently accessible via its DOI.

DKRZ

www.dkrz.de