FAIR long term preservation of climate and Earth System Science data with a focus on reusability at the World Data Center for Climate (WDCC)

Karsten Peters[1], Heinke Höck[1], and Hannes Thiemann[1]

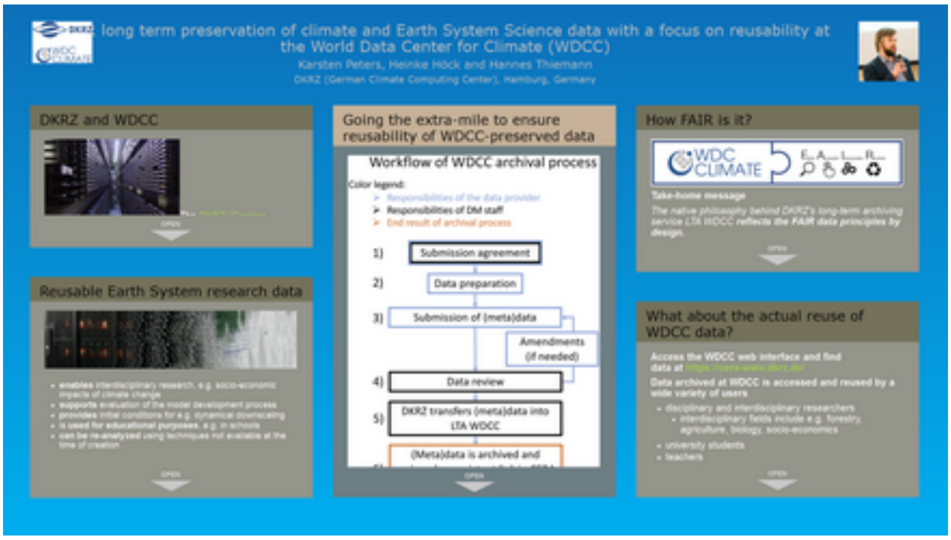
[1]DKRZ German Climate Computing Centre

November 22, 2022

Abstract

The full-featured and CoreTrustSeal certified long term archiving service LTA WDCC (World Data Centre for Climate) at DKRZ (German Climate Computing Center, Hamburg) offers long term preservation for datasets relevant for climate and Earth System research. The WDCC collects, stores, and disseminates Earth System data with a focus on climate simulation data and climate related data products. It has established itself as a staple infrastructure for the global climate modelling research community. Data preservation in LTA WDCC is preceded by a thorough technical quality control and provides intense data curation for storage periods longer than 10 years. During the preservation period, long term findability, searchability and reusability of the data are ensured. Long term findability of the curated data is enabled through the possibility of assigning DataCite DOI's to archived datasets. The data undergo additional quality checks before being eligible for DOI assignment. This process is performed in close collaboration with the data providers. The focus of these quality checks is to ensure the unambigous (inter-)disciplinary reusability of the preserved datasets and includes checking for proper documentation, adherence to domain specific (meta)data standards, uncertainty analysis and cross-referencing. Only then can a high level of reusability of the data be achieved, justifying the involved effort. The perceived need for research data repositories to comply with the 2016-published FAIR Guiding Principles has motivated us to perform an even-handed and systematic self assessment of LTA WDCC FAIRness. Due to lack of a standardised evaluation framework, this assessment reflects our specific, albeit objective, interpretation of the principles. Our assessment, published on the DKRZ webpages, shows that the native philosophy behind DKRZ's LTA WDCC service – especially the focus on reusability – reflects the FAIR Guiding Principles by design and even goes beyond them by ensuring very long-term (>10 years) preservation and therefore reusability of archived data.

FAIR long term preservation of climate and Earth System Science data with a focus on reusability at the World Data Center for Climate (WDCC)



long term preservation of climate and Earth System Science data with a focus on reusability at the World Data Center for Climate (WDCC)

Karsten Peters, Heinke Höck and Hannes Thiemann

DKRZ (German Climate Computing Center), Hamburg, Germany

DKRZ and WDCC

OPEN

Going the extra-mile to ensure reusability of WDCC-preserved data

Workflow of WDCC archival process

Color legend:

- Responsibilities of the data provider
- Responsibilities of DM staff
- End result of archival process

1) Submission agreement

2) Data preparation

3) Submission of (meta)data

Amendments (if needed)

4) Data review

5) DKRZ transfers (meta)data into LTA WDCC

(Meta)data is archived and

OPEN

How FAIR is it?

Take-home message

The native philosophy behind DKRZ's long-term archiving service LTA WDCC reflects the FAIR data principles by design.

OPEN

Reusable Earth System research data

- enables interdisciplinary research, e.g. socio-economic impacts of climate change
- supports evaluation of the model development process
- provides initial conditions for e.g. dynamical downscaling
- is used for educational purposes, e.g. in schools
- can be re-analyzed using techniques not available at the time of creation

OPEN

What about the actual reuse of WDCC data?

Access the WDCC web interface and find data at https://cera-www.dkrz.de/

Data archived at WDCC is accessed and reused by a wide variety of users

- disciplinary and interdisciplinary researchers
  - interdisciplinary fields include e.g. forestry, agriculture, biology, socio-economics
- university students
- teachers

OPEN

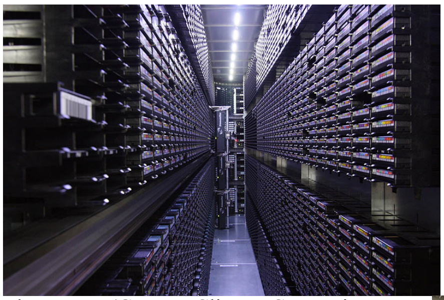
Karsten Peters, Heinke Höck and Hannes Thiemann

DKRZ (German Climate Computing Center), Hamburg, Germany

PRESENTED AT:

AGU 100
ADVANCING EARTH AND SPACE SCIENCE
FALL MEETING
San Francisco, CA | 9–13 December 2019

DKRZ AND WDCC



The DKRZ (German Climate Computing Center) (https://www.dkrz.de/dkrz-partner-for-climate-research?set_language=en&cl=en) is the national IT service provider for the German Earth System (ES) research community and also plays an important role for enabling globally connected ES research.

Scientists are supported with a combined portfolio of high performance hardware as well as discipline-specific application and data management services.

Hardware

HPC System *Mistral* (top right):

- #80 in TOP500 (https://www.top500.org/list/2019/11/) (November 2019), to be replaced in 2020
- 52 PiB disk space on parallel file system (#4 worldwide (https://www.vi4io.org/hpsl/start))

Tape archive (top left):

- >200 PiB total capacity
- 5 PiB disk cache for quick access of high-demand data

Long-term archiving service at DKRZ

DKRZ's Data Management (DM) department (https://www.dkrz.de/about-en/staff/data-management?set_language=en&cl=en) offers dedicated and comprehensive **support for all stages of the research data life cycl**e of ES research data.

Long-term archiving (LTA) of climate data has a long tradition at DKRZ - the oldest datasets in the archive have been preserved **since 1995**.

(https://cera-www.dkrz.de/WDCC/ui/cerasearch/)In 2003, DKRZ LTA service was approved as domain-specific World Data Center - the **World Data Center for Climate (WDCC)**. (https://cera-www.dkrz.de/)

WDC CLIMATE

(https://www.coretrustseal.org)

Since 2018, LTA WDCC is CoreTrustSeal certified and regular member of the World Data System.

CORE TRUST SEAL

Data preservation at the WDCC is focused on long-term reusability of ES reserach data. In-house data management services are especially tailored towards dealing with output from climate model simulations, e.g. those required for the preparation of the IPCC's Assessment Reports (https://www.ipcc.ch/reports/).

See the box below for some illustrative examples of reusable climate model data in action

REUSABLE EARTH SYSTEM RESEARCH DATA

- **enables** interdisciplinary research, e.g. socio-economic impacts of climate change
- **supports** evaluation of the model development process
- **provides** initial conditions for e.g. dynamical downscaling
- is **used for educational purposes**, e.g. in schools
- **can be re-analyzed** using techniques not available at the time of creation

Climate model output stored at the WDCC provides the basis for e.g. products of public interest, such as

Example animation 1:

CMIP5 multi-model mean temperature change by scenario

[VIDEO] https://www.youtube.com/embed/LylIi_dSQZI?feature=oembed&fs=1&modestbranding=1&rel=0&showinfo=0

Example animation 2:

CMIP5 multi-model mean precipitation change for RCP8.5

[VIDEO] https://www.youtube.com/embed/cHZeUvKLecM?feature=oembed&fs=1&modestbranding=1&rel=0&showinfo=0

Example animation 3:

CMIP3 multi-model mean 2m temperature change for SRES A1B

[VIDEO] https://www.youtube.com/embed/Xjn-AHh2Caw?feature=oembed&fs=1&modestbranding=1&rel=0&showinfo=0
Interactive visualization of climate projection data:

- https://www.dkrz.de/webvis/ (https://www.dkrz.de/webvis/)

Scientific importance of reusable ES data

Evaluating the simulation of Earth's climate system across climate model generations is essential.

Long-term preservation and curation of ES research data in the WDCC **enables such analyses** and thus fundamentally contributes to the advancement of climate science.

How do we reliably achieve the reusability of ES data archived in the WDCC?

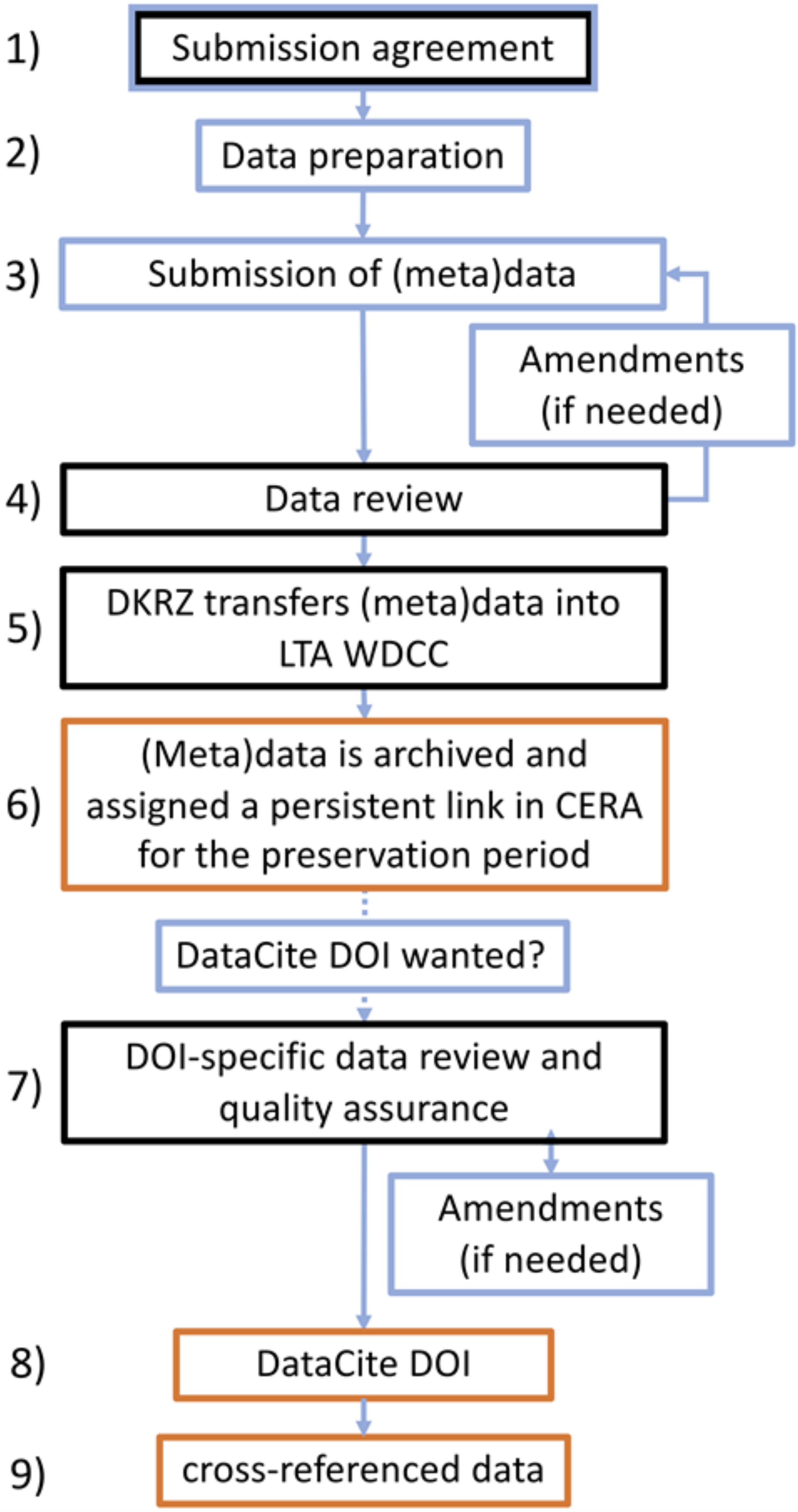
...continue with the middle box

GOING THE EXTRA-MILE TO ENSURE REUSABILITY OF WDCC-PRESERVED DATA

Workflow of WDCC archival process

Color legend:

- ➢ Responsibilities of the data provider
- ➢ Responsibilities of DM staff
- ➢ End result of archival process



1) Submission agreement

2) Data preparation

3) Submission of (meta)data

Amendments (if needed)

4) Data review

5) DKRZ transfers (meta)data into LTA WDCC

6) (Meta)data is archived and assigned a persistent link in CERA for the preservation period

DataCite DOI wanted?

7) DOI-specific data review and quality assurance

Amendments (if needed)

8) DataCite DOI

9) cross-referenced data

Close co-operation between data providers (ES researchers) and DM staff at DKRZ is essential for successful completion of the archival process.

Achieving reusability of WDCC-archived ES data

- the essentials

- **dedicated support and one-on-one contact** between DM staff and the data providers during the entire process
- (meta)data are strictly required to comply with ES research **domain-specific file and data formats**, e.g. netcdf-CF or CMIP conventions
  - **automated and manual technical quality assurance** procedures are applied
  - data providers are provided with **detailed instructions** for (meta)data amendments in case of non-compliance
- if possible, **metadata contain PID-based references** to associated publications
- **DOI-assignment requires additional information** to increase reusability, e.g. uncertainty analysis and provenance information

For more details, please see the WDCC-How-To Guide (https://www.dkrz.de/up/services/data-management/LTA/how-to-use-lta-wdcc) online.

HOW FAIR IS IT?

Take-home message

*The native philosophy behind DKRZ's long-term archiving service LTA WDCC **reflects the FAIR data principles by design.***

The details

Findable ✅

- **DataCite DOI's** assigned to archived (meta)data
- **rich metadata** are required for archival in WDCC
- **indexed** in various external catalogues
- **machine-readable** (meta)data

Accessible ✅

- access to (meta)data via **standard protocol**, i.e. web-interface
- **metadata are open**
- data access requires **authentication** and is mostly unrestricted
- **metadata remain accessible** after data has been retracted
- **machine-readable** (meta)data

Interoperable ✅

- **domain-specific vocabularies** (machine-readable), e.g. CF, CMIP
- **DataCite metadata schema**
- **open file formats** (NetCDF, GRIB)

Reusable ✅

- **rich metadata**, including scientific documentation and cross-references to associated publications
- **domain-specific file and data format** standards
- **clear licenses for data reuse**, i.e. CC-BY is the default

For more information and details regarding our even-handed self-evalution along the lines of the published FAIR principles, please see the WDCC web-pages (https://www.dkrz.de/up/services/data-management/LTA/fairness).

WHAT ABOUT THE ACTUAL REUSE OF WDCC DATA?

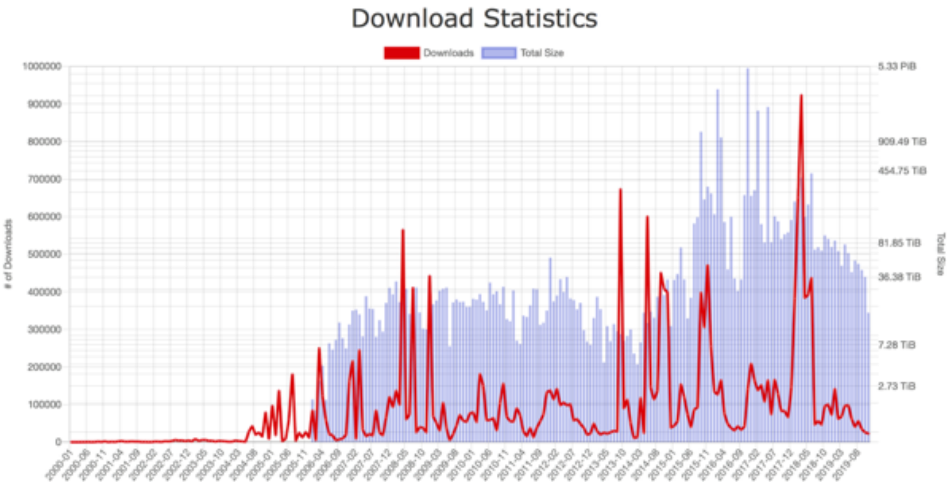
Access the WDCC web interface and find data at https://cera-www.dkrz.de/ (https://cera-www.dkrz.de/)

Data archived at WDCC is accessed and reused by a wide variety of users

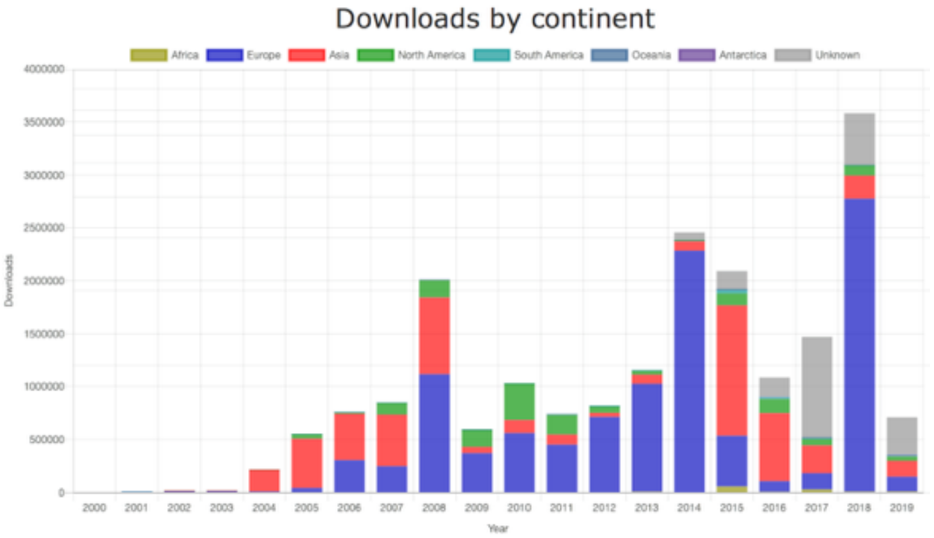
- disciplinary and interdisciplinary researchers
  - interdisciplinary fields include e.g. forestry, agriculture, biology, socio-economics
- university students
- teachers

Routine statistics collected at DKRZ showcase the national as well as international demand for datasets archived at WDCC. Citation statistics are not tracked yet but is planned for the near future.

Overall download statistics / month

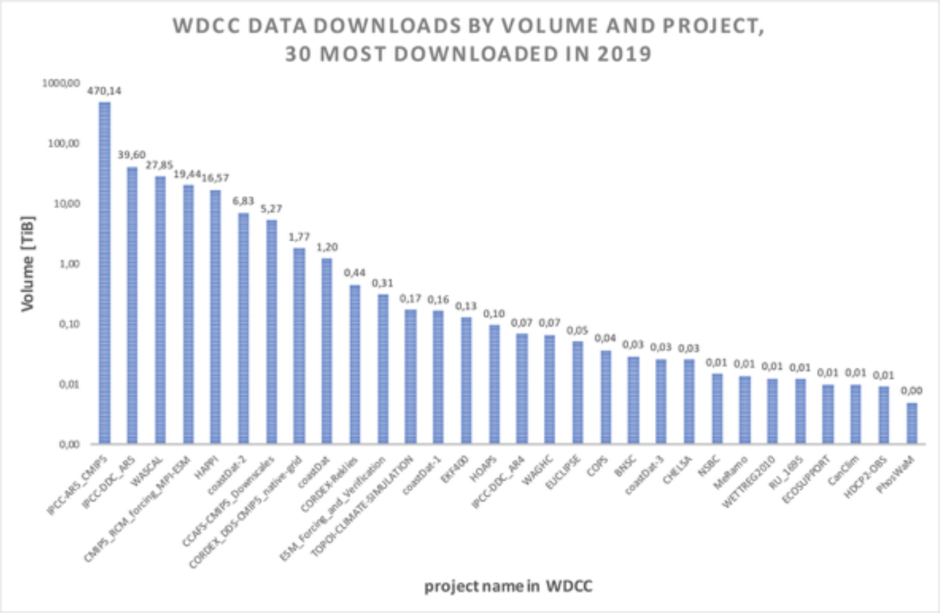


Download statistics / continent and year



Downloaded data volume per project in 2019

see below for description of some highly accessd projects



IPCC AR5, CMIP5, IPCC-DDC CMIP5:

datasets produced in the framework of CMIP5 - these data were also the basis for the 5th IPCC Assessment Report (published in 2013).

WASCAL:

Regionalized climate change projections for the West African Region, produced in the framework of the WASCAL project (https://wascal.org) (West African Science Service Centre on Climate Change and Adapted Land Use).

HAPPI:

Datasets produced in the framework of the HAPPI (Half a degree Additional warming, Prognosis and Projected Impacts) project. HAPPI data contributed to the IPCCs Special Report "Global Warming of 1.5C (https://www.ipcc.ch/sr15/)"

coastDat:

Datasets providing high-resolution information on marine enrivonments in data sparse regions, e.g. the North Sea. The project has been ongoing for >10 years and the archived data is extended on a regular basis (Weblink).

Sorry but time is up!

CV

Dr. Karsten Peters

Data Management Service Communication and Development

Deutsches Klimarechenzentrum (DKRZ), Hamburg, Germany

peters@dkrz.de

Personal webpage (https://www.dkrz.de/about-en/staff/dr-karsten-peters)

Work Area

- Communication, mediation and integration of DKRZ research data management services for the Earth System Science Community
- Design and execution of courses/workshops
- Development of DKRZ's portfolio of research data management services according to scientific-community demands

Background

Meteorology and climate science:

- Diploma of Meteorology (equivalent to MSc, University of Hamburg), 2008
- PhD in Meteorology (University of Hamburg, Max Planck Institute for Meteorology), 2011
- Post-Doc in Meteorology (Monash University, Melbourne and Max Planck Institut for Meteorology), 2012-2018

Profiles

- Google Scholar (https://scholar.google.de/citations?user=TQf9PhQAAAAJ&hl=de&oi=ao)
- ReserachGate (https://www.researchgate.net/profile/Karsten_Peters)
- ORCID (https://orcid.org/0000-0003-0158-2957)
- LinkedIn (https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&cad=rja&uact=8&ved=2ahUKEwjXv7O4x6DmAhUKuqQKHesPA_sQFjAFegQIBhAB&url=https%3A%2F%2Fde.linke peters-352649163&usg=AOvVaw0JZQ8Dyhmi7Ej-yw4cFa-4)

ABSTRACT

The full-featured and CoreTrustSeal certified long term archiving service LTA WDCC (World Data Centre for Climate) at DKRZ (German Climate Computing Center, Hamburg) offers long term preservation for datasets relevant for climate and Earth System research. The WDCC collects, stores, and disseminates Earth System data with a focus on climate simulation data and climate related data products. It has established itself as a staple infrastructure for the global climate modelling research community. Data preservation in LTA WDCC is preceded by a thorough technical quality control and provides intense data curation for storage periods longer than 10 years. During the preservation period, long term findability, searchability and reusability of the data are ensured.

Long term findability of the curated data is enabled through the possibility of assigning DataCite DOI's to archived datasets. The data undergo additional quality checks before being eligible for DOI assignment. This process is performed in close collaboration with the data providers. The focus of these quality checks is to ensure the unambigous (inter-)disciplinary reusability of the preserved datasets and includes checking for proper documentation, adherence to domain specific (meta)data standards, uncertainty analysis and cross-referencing. Only then can a high level of reusability of the data be achieved, justifying the involved effort.

The perceived need for research data repositories to comply with the 2016-published FAIR Guiding Principles has motivated us to perform an even-handed and systematic self assessment of LTA WDCC FAIRness. Due to lack of a standardised evaluation framework, this assessment reflects our specific, albeit objective, interpretation of the principles. Our assessment, published on the DKRZ webpages, shows that the native philosophy behind DKRZ's LTA WDCC service – especially the focus on reusability – reflects the FAIR Guiding Principles by design and even goes beyond them by ensuring very long-term (>10 years) preservation and therefore reusability of archived data.

SWITCH TEMPLATE