

From 25 TB to a DOI

Publishing large research datasets at KIT – a case study from IMKASF, and the wider IMK data landscape

Sabine Barthlott und Tobias Kerzenmacher · IMKASF, KIT · on behalf of IMKASF, IMKAAF, IMKTRO, IMKIFU · DiTraRe

Open Science at KIT: Show us your data · 21 Oct 2026



Leibniz ScienceCampus
Digital Transformation
of Research

25 TB do not fit into a repository

RADAR4KIT: maximum size per dataset – 600 GB



$\approx 40\times$

Satellite data: MUSICA IASI v3.2.1 – 25 TB, October 2014 to June 2019



The KIT Library recommends 50–100 GB per dataset for easy reuse. Very large datasets can be split and linked – or published in a hybrid way.

One citable entry point, two access layers, three storage systems

Publish small subset with Link

RADAR4KIT: 17.5 GB,
netCDF (CF-1.7),
README with access
information.



Central access to whole data set

THREDDS data server



Complete record on secured data store

Hosted by LSDF (SCC):
Only accessible via THREDDS
(not directly)

Backup to guarantee long-term availability

bwDataArchive (complete data set - 25 TB); checksum; requirement for journal publication



Users start with a real sample they can open in minutes – then follow persistent links to everything.

Dataset description: Schneider et al., ESSD 14, 709 (2022), doi:10.5194/essd-14-709-2022 · DiTraRe poster: doi:10.5281/zenodo.17815953

Small data: the QBO winds – a living dataset

1953–now

monthly record, 4
equatorial stations

15

pressure levels, 100–10
hPa

30+

versions, updated regularly

~8,000

downloads across all
versions

Why it works

- CC0 licence, DOI, part of the DataHub community on Zenodo
- Formats: netCDF, ASCII, plot
- Continues a FU Berlin record; now processed at IMKASF
- A service for the global QBO and trend community

The data world at IMK is big – and diverse

IMKASF

MIPAS satellite retrievals ·
GLORIA airborne data ·
IAGOS-CARIBIC · NDACC/
MUSICA, TCCON, COCCON
· QBO winds · model
results

IMKAAF

AIDA aerosol chamber ·
ice-nucleating particle
data (PINE) · analysis
software (PIA)

IMKTRO

ICON model releases ·
KITCube · Swabian MOSES
campaign data · dust and
weather-extreme studies

IMKIFU

TERENO · regional climate
and hydro-meteorological
data

Published in RADAR4KIT, Zenodo, WDCC, TCCON/EVDC and more – collected in the ATMOHub catalogue

Many repositories, one catalogue

Data	Repository	DOI
MIPAS v8 retrievals (satellite)	RADAR4KIT	10.35097/1948 (CH4)
GLORIA airborne Level 2 (HEMERA-2022)	RADAR4KIT	10.35097/t2hqj3ef41x9nb55
IAGOS-CARIBIC MS files collection	Zenodo	10.5281/zenodo.15573499
NDACC/MUSICA ground-based water isotopes	Zenodo	10.5281/zenodo.48902
TCCON Karlsruhe	TCCON	10.14291/tccon.ggg2020.karlsruhe01.R2
COCCON Karlsruhe	EVDC	10.48477/COCCON.PF24.KARLSRUHE.SN037.R04
ICON model releases	WDCC	10.35089/WDCC/IconRelease2025.10
PINE ice-nucleation data (aerosol)	RADAR4KIT	10.35097/mdxcqd7p646cwc0u

Source: ATMO Hub data publications, atmohub.kit.edu/english/507.php – examples, not a complete list

Metadata is the bridge: NetCDF guidelines

1

Standard from the start

CF conventions and other standards do not cover all fields, solution is combination/adaptation of various standards

2

Community guidelines

Helmholtz Metadata Guidelines for NetCDF, v1.0.0 (April 2026), by Hereon and KIT

3

Continually updated

Growing community, further development of guidelines and metadata checker, cornerstone for AI readiness, cooperation with HMC

Challenges, costs and lessons learned

Challenges

- Volume: 25 TB, ~40×the repository limit
- Keeping DOI, sample and full data persistently linked
- Transparent versioning
- Metadata harmonisation across groups and formats

Costs

- RADAR4KIT: publication free of charge, KIT guarantees retention of at least 25 years
- SCC: LSDF, bwDataArchive
- institutional service: THREDDS hosting

What helped

- KIT infrastructure: SCC, LSDF, RADAR4KIT
- Versioning for living datasets
- Close collaboration with SCC, RADAR4KIT, KITopen
- Metadata standards: community, exchange across different scientific fields
- RDM team across IMKs

Take-aways

- 1 Size decides the route – not whether to publish
- 2 Large data: DOI + representative sample + persistent access
- 3 Small data: direct deposit, open licence, versioning
- 4 Standardised metadata makes data reusable across fields
- 5 Constant improvement of RADAR4KIT through feedback and discussions (DiTraRe)