

Publishing Large Datasets: Experiences with Data Scaling and Hybrid Model Publication

Big Data, Small Uploads: A Hybrid Workflow for Large Datasets

Tobias Kerzenmacher, Sabine Barthlott, Matthias Schneider — KIT, Karlsruhe, Germany

Felix Bach, Kerstin Soltau, and Stefan Hofmann — FIZ Karlsruhe, Eggenstein-Leopoldshafen, Germany

© ESA

1. The Challenge

- Multi-terabyte satellite datasets (25 TB MUSICA IASI v3.2.1) exceed upload limits due to archival restrictions on bwArchive (~600 GB)
- Need FAIR, citable, and accessible solutions
- Ensure reproducibility, metadata compliance, and long-term preservation



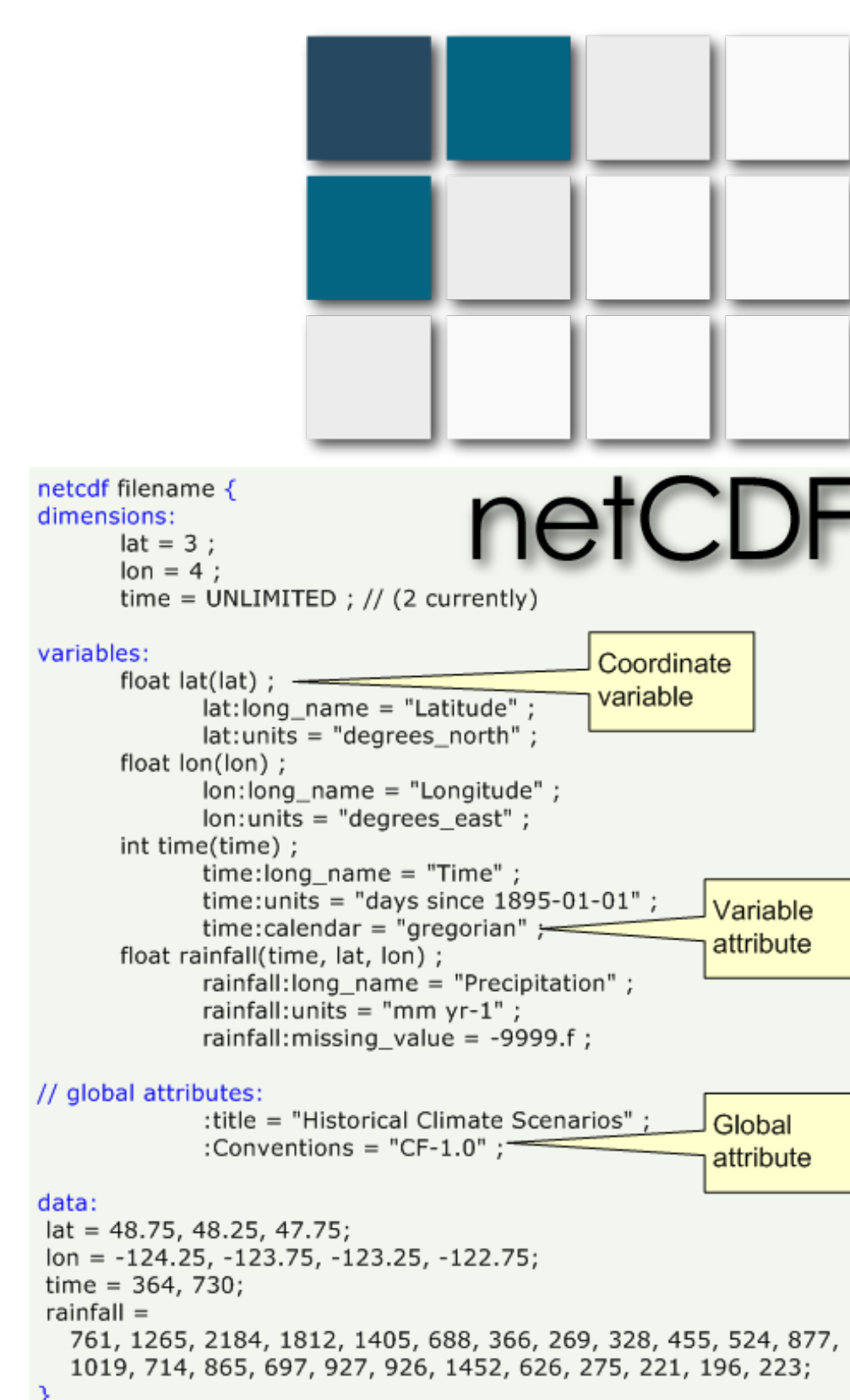
2. Hybrid Workflow

- Representative subset → RADAR4KIT (citable, DOI: 10.35097/408)
- Full dataset → LSDF (Large Scale Data Facility), data accessed via THREDDS data server
- Long-term archive → bwDataArchive

© Simon Raffener/SCC/KIT

3. Representative Subset

- One day (17.5 GB) of the large satellite data (25 TB, 1734 days) selected.
- Captures retrieval configuration and structure.
- Enables testing, validation, and reproducibility.
- Includes MD5 checksums, ReadMe
- Includes stable links to the LSDF archive via the THREDDS catalogue



Example netCDF header

4. Lessons Learned

- Hybrid model balances persistence and scalable access
- Metadata harmonization and clear documentation are essential

RADAR4KIT

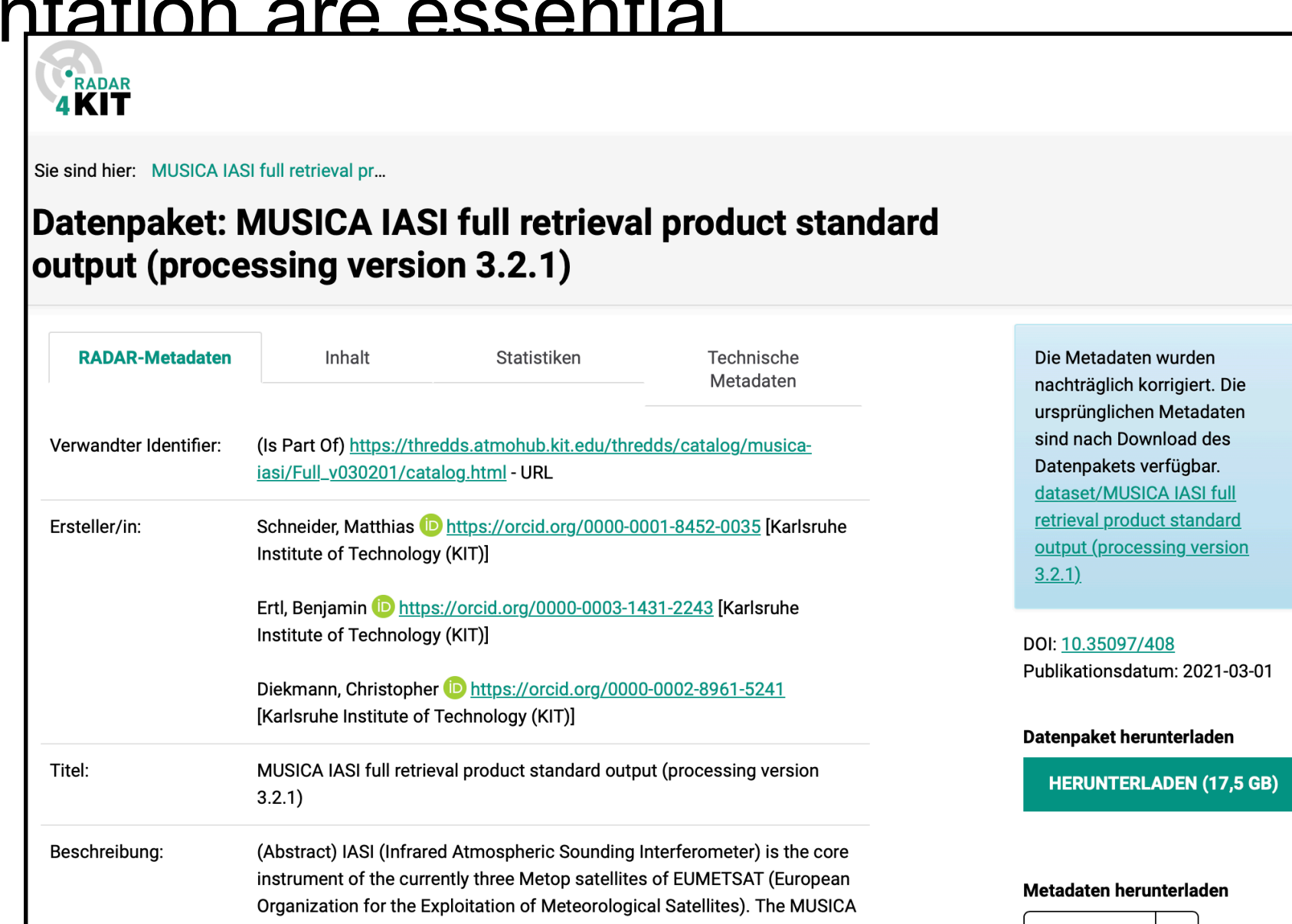


THREDDS

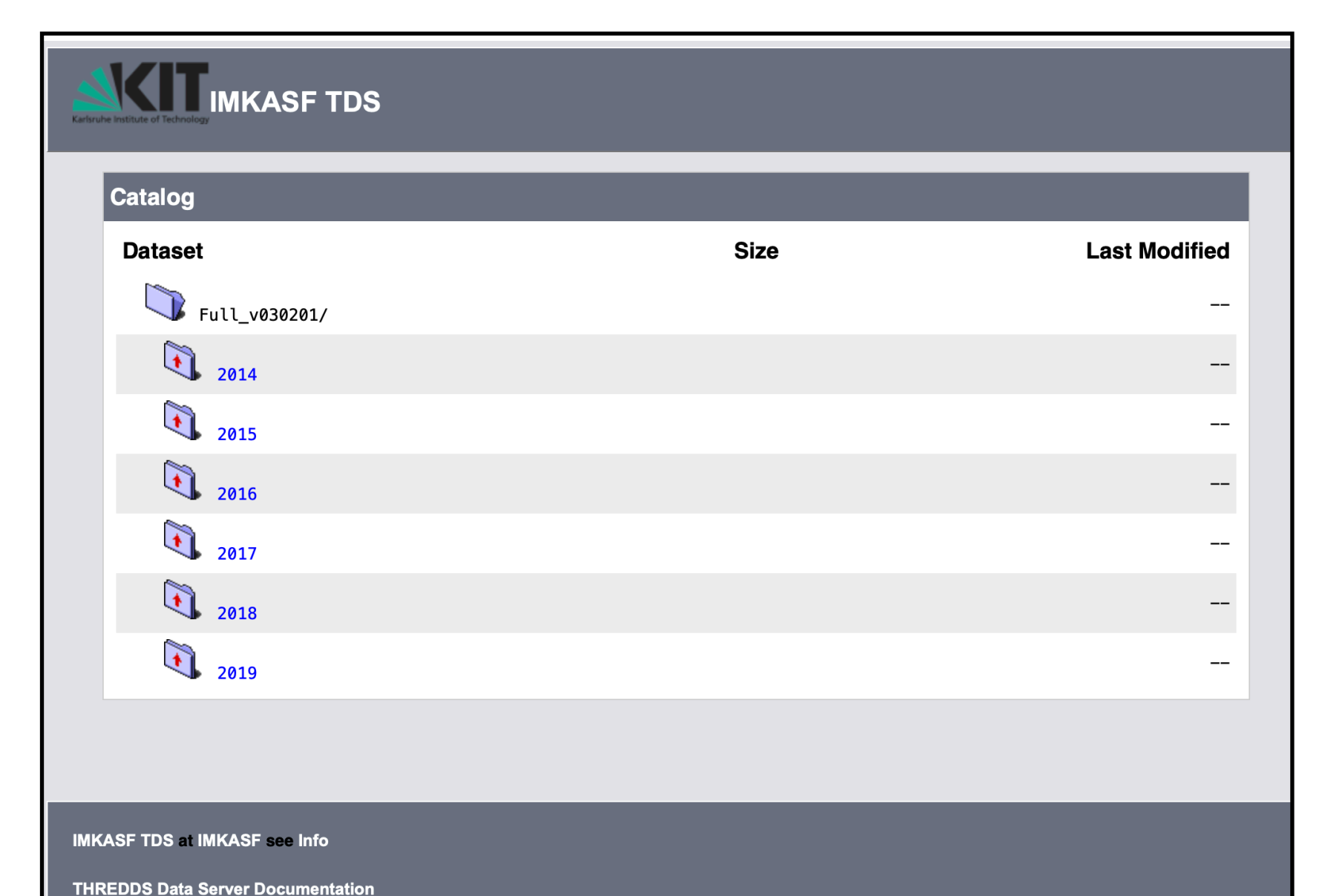


5. Outlook

- Repository improvements: Terminology Service 4 NFDI (TS4NFDI), WebDAV upload
- Additional metadata fields, single-file downloads, Zenodo-style versioning
- Specialized repositories for multi-terabyte datasets may emerge



Landing page for the published data on RADAR4KIT.



THREDDS server at IMKASF to the full data set on the LSDF.

The authors thank the data infrastructure staff at FIZ Karlsruhe, IMKASF, and KIT, as well as the supporting funding agencies. The Leibniz Science Campus Digital Transformation of Research (DiTraRe) is funded by the Leibniz Association (W74/2022).

Contact: Tobias.Kerzenmacher@kit.edu