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

Tobias Kerzenmacher[1,*], Sabine Barthlott[1], Matthias Schneider[1], Felix Bach[2], Kerstin Soltau[2], and Stefan Hofmann[2]

[1]Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing (IMKASF), Karlsruhe, Germany
[2]FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, Germany
[*]Corresponding author: kerzenmacher@kit.edu

Abstract

The publication of large datasets poses significant challenges for research data management, particularly regarding storage, access, citability, and adherence to FAIR principles [1]. We present our experience publishing a 25 TB dataset from the Infrared Atmospheric Sounding Interferometer (IASI) covering October 2014 to June 2019 (MUSICA IASI v3.2.1) [2] using a hybrid approach. A representative one-day subset was deposited in the institutional repository RADAR4KIT [3] with a DOI, while the complete dataset is archived at the Large Scale Data Facility (LSDF) [4] provided by the Scientific Computing Center (SCC) at KIT and accessed via a THREDDS Data Server (TDS) hosted by IMKASF at KIT [5,6]. We describe the workflow, metadata harmonisation, persistent linking strategy, and lessons learned from this publication process.

Introduction

Large satellite datasets are essential for advancing climate change research because they provide globally consistent, high-resolution atmospheric observations [2]. However, their enormous size—often reaching tens of terabytes—creates major challenges for storage, management, and dissemination within institutional repositories. To address these challenges, this paper presents a practical publication strategy for the 25 TB MUSICA IASI dataset, which combines repository-based citation for credit with domain-service access for bulk retrieval [3,7]. This hybrid model ensures both citability and scalable data access.

Dataset and Context

The dataset corresponds to the MUSICA IASI full-retrieval product, version 3.2.1 [2]. It provides full information on retrieval input (a priori and constraints) and output (inverted trace gas profiles, error profiles and covariances, averaging kernels, and kernel metrics) and covers the period between October 2014 and June 2019. The full dataset, comprising about 25 TB, is archived at bwDataArchive for long-term preservation and accessible via LSDF and THREDDS [4,6].

Because RADAR4KIT imposes upload size limitations (approximately 600 GB) [7], a representative subset corresponding to one full day of IASI observations (all orbits on 2018-08-01) was deposited as the citable entity [3]. This subset captures the complete retrieval configuration and data structure, making it suitable for demonstration, testing, and reproducibility, while the complete multi-year dataset must be accessed via the THREDDS server for full-scale analyses. Details on data format, access, and authenticity are provided in the equally uploaded `ReadMe.pdf`, and MD5 checksum files enable verification of data integrity. Additionally, a direct link to the full dataset is present on the DOI landing page.

Publication Workflow and Technical Approach

Our workflow balances citability, scalable access, and long-term usability.

The first stage, *metadata harmonisation*, aligned core elements such as temporal coverage, spatial bounds, variable lists, and retrieval version with the CF (Climate & Fore-

cast) metadata conventions [8]. Missing variable names were proposed to, and subsequently accepted by, the CF Metadata Conventions Committee. Using these conventions, the data were stored in netCDF files compliant with CF standards, ensuring comparability, reusability, and transparency.

The second stage, *sampling and subsetting*, involved selecting the one-day representative subset for RADAR4KIT [3] while maintaining the full dataset on LSDF for scalable access. DataCite compliance was achieved by manually adding metadata and links to the full dataset on the DOI landing page.

The third stage, *persistent linking*, ensured that the RADAR4KIT DOI page includes stable links to the LSDF archive via the THREDDS catalogue [3, 4, 6]. Citation, storage, and remote access responsibilities are clearly separated across platforms. The complete dataset is also archived at bwDataArchive [9], and MD5 checksums of archived files were published alongside the sample dataset to allow authenticity verification in case of data restoration.

Finally, *access strategies* include direct HTTP download of the subset from RADAR4KIT, remote access to the full dataset via THREDDS, and LSDF support for large-volume transfers. This decoupling enables scalable access while maintaining a persistent, citable record. To enhance findability and interpretability, a companion publication describing the dataset in detail was published in *Earth System Science Data* (ESSD) [2].

Challenges and Lessons Learned

Publishing a multi-terabyte dataset introduces both technical and conceptual challenges. Repository limits required selecting a representative subset, which adequately reflects data structure and retrieval characteristics but does not substitute the full dataset for comprehensive analyses. Clear communication of such limitations is essential for transparency.

Ensuring metadata consistency across RADAR4KIT, LSDF, and THREDDS demanded close coordination. While the DOI assigned to the subset remains stable, the long-term accessibility of the full dataset depends on institutional infrastructure, which must be explicitly stated to users. Remote access via THREDDS significantly improves usability but requires dedicated technical support and maintenance. Finally, balancing citability with bulk access was a key consideration to enable reliable citation while supporting large-volume retrieval [3].

The hybrid model combining repository-level citation with domain-level access proved effective and reproducible. It highlights the trade-off between persistence and scalability that repositories must address as dataset volumes grow. Providing multiple access pathways (HTTP, OPeNDAP, compute-to-data) and clearly documenting data retention and access guarantees enhance both usability and user trust.

Conclusion and Outlook

Publishing multi-terabyte satellite datasets is feasible using a hybrid approach that combines repository citation with domain infrastructure for scalable access. Establishing this workflow informed ongoing improvements to RADAR4KIT, including WebDAV upload support, extended metadata fields, and integration of the Terminology Service 4 NFDI (TS4NFDI) for improved keyword selection. Features under development—such as single-file downloads and Zenodo-style versioning—will further facilitate large-data publication.

Future work within DiTraRe's use case *Publication of Large Data Sets* focuses on developing repository-side solutions for handling multi-terabyte datasets and exploring cross-repository standards for hybrid data publication. As research increasingly relies on large-scale observational and model data, specialised repositories optimised for big data publication are likely to emerge.

Acknowledgements

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).

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016. doi: 10.1038/sdata.2016.18 .

[2] M. Schneider, B. Ertl, C. J. Diekmann, F. Khosrawi, A. Weber, F. Hase, M. Höpfner, O. E. García, E. Sepúlveda, and D. Kinnison, "Design and description of the musica iasi full retrieval product, version 3.2.1," *Earth System Science Data*, vol. 14, pp. 709–742, 2022. doi: 10.5194/essd-14-709-2022 .

[3] M. Schneider, B. Ertl, and C. Diekmann, "Musica iasi full retrieval product standard output (processing version 3.2.1)," 2021. doi: 10.35097/408 .

[4] Scientific Centre for Computing (SCC), KIT, "LSDF Online Storage @ KIT," https://wiki.scc.kit.edu/lsdf/index.php/Category:LSDF_Online_Storage, 2024, accessed October 2025.

[5] KIT, IMKASF, "MUSICA IASI v3.2.1 project page," https://www.imkasf.kit.edu/english/musica-iasi-v321.php, 2024.

[6] Unidata, "THREDDS Data Server (TDS)," https://www.unidata.ucar.edu/software/tds, 2025, accessed October 2025.

[7] RADAR Repository, FIZ Karlsruhe/KIT Library, "RADAR Repository FAQ and Technical Description," https://www.bibliothek.kit.edu/radar-faq.php, 2025, accessed October 2025.

[8] CF Metadata Conventions Committee, "Climate and Forecast (CF) Metadata Conventions," https://cfconventions.org, 2023, accessed: October 2025.

[9] Karlsruhe Institute of Technology (KIT), "bwdataarchive – long-term data archive for scientific data," https://www.rda.kit.edu/english/, 2025, accessed: October 2025.