



Let's talk data archiving: practical insights into FAIR and long-term data management from early-career perspectives

Joana Kristin Steffens^{1,2}, Lukas Monrad-Krohn^{1,3}, Amelie Driemel⁴, Nicole Greco⁵, and Johannes Röttenbacher⁶

¹German National Committee of the Association of Polar Early Career Scientists (APECS Germany), Germany

²Institut des Sciences de la Mer de Rimouski (ISMER), Université du Québec à Rimouski (UQAR), Rimouski, QC G5L 2Z9, Canada

³Department of Geosciences, University of Oslo, 0316 Oslo, Norway

⁴Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27570 Bremerhaven, Germany

⁵National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, 93101, United States of America

⁶Institute of Environmental Physics, University of Bremen, 28359 Bremen, Germany

Correspondence: Joana Kristin Steffens (joanakristin.steffens@uqar.ca)

Received: 8 June 2026 – Revised: 11 September 2026 – Accepted: 23 September 2026 – Published: 27 October 2026

Abstract. Research data provide the basis for scientific discovery, yet their long-term value depends on careful curation and documentation. Data archiving is often treated as a final administrative task rather than an integral part of research. This report summarizes key insights from an interactive seminar organized by the Association of Polar Early Career Scientists (APECS) Germany, which explored practical approaches to making data findable, accessible, interoperable, and reusable (FAIR) from early-career perspectives. Speakers from PANGAEA, the Arctic Data Center, and the (AC)³ project highlighted the importance of high-quality metadata, persistent identifiers, editorial curation, secure long-term storage, and integration of data publication into everyday workflows. The discussion emphasized that incomplete documentation remains a major barrier to reuse and that planning data management from the start of a project greatly reduces later effort. The seminar demonstrated that data archiving is a shared responsibility and a strategic investment that strengthens reproducibility and scientific impact. Building on the seminar discussions, we recommend expanding practical FAIR training opportunities, promoting community metadata standards, recognizing data archiving within research funding and project planning, and strengthening a research culture in which high-quality data stewardship is valued as an integral component of scientific excellence.

1 Why talk about data archiving?

Research data are the foundation of scientific discovery, yet what happens to them once a project ends is often unclear. Data archiving is frequently treated as a final administrative step, carried out shortly before publication or project completion. However, research has shown that well-curated and accessible data provide substantial long-term scientific value, enabling reuse far beyond the scope of the original study and yielding high returns on relatively modest investments in data infrastructure (Colavizza et al., 2024; Piwowar et al., 2011).

This value, however, is not automatic. Over time, information content fades if contextual knowledge is lost, particularly when metadata are incomplete or inconsistent. Community-driven efforts such as the Paleoclimate Community Reporting Standard (PaCTS) highlight that data reusability depends critically on clear documentation of what was measured, how, where, and why (Khider et al., 2019). These insights emphasize that decisions made during a research project – rather than at its end – determine whether data remain findable, understandable, and reusable in the future. In this context, persistent identifiers (PIDs) are a cornerstone of effective data publication (<https://www.pid-network.de/>, last access: 12 October 2026, Genderjahn and Schrader, 2026). By uniquely and permanently linking datasets, researchers, instruments, institutions and publications, they make the entire research workflow traceable – from data collection to

data publication. This not only enables transparency and reproducibility but also ensures the context of the data is unambiguously described.

Ensuring this value is preserved and made accessible is the task of research data management, which encompasses the planning, organization, documentation, storage, sharing, and preservation of research data throughout the entire research lifecycle. Data archiving is one component of this broader process and focuses specifically on the long-term preservation and publication of datasets in trusted repositories so that they remain findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al., 2016). Although this article focuses primarily on data archiving, many of the practices discussed – including metadata creation, documentation, and planning – are integral elements of good research data management.

These considerations are particularly relevant for polar and Earth system sciences, where many datasets consist of unique observations collected under environmental conditions that cannot necessarily be reproduced. Long-term climate observations, field campaigns, and remote sensing products often document processes that change continuously over time, making historical datasets an irreplaceable scientific resource. Consequently, preserving not only the measurements but also their accompanying metadata is essential for future reuse, data rescue efforts, climate assessments, and policy-relevant research. In addition, polar research frequently relies on large international collaborations that generate diverse observational and modeling datasets, making interoperable data standards and coordinated archiving practices especially important.

Against this backdrop, early-career researchers in particular need practical entry points into archiving practice. On 5 November 2025, the Association of Polar Early Career Scientists (APECS) Germany therefore hosted an online “Let’s talk...!” seminar dedicated to data archiving. The event brought together participants from various countries in an interactive setting, reflecting the discussion-focused nature of the “Let’s talk...!” format. The seminar aimed to demystify data archiving, highlight its relevance for early-career researchers, and provide practical guidance on how to manage and publish data in a FAIR way.

Rather than focusing on abstract principles alone, the event emphasized concrete workflows, tools, and real-world experiences from data repositories and large collaborative research projects.

2 Let's talk data archiving

The seminar was structured around short introductory presentations followed by a Q&A session. Three complementary perspectives framed the discussion: a long-established international data repository, an Arctic-specific data repository,

and a large collaborative research project acting as an intermediary between researchers and repositories.

2.1 Data curation and long-term reuse with PANGAEA

Amelie Driemel (PANGAEA, Alfred Wegener Institute) introduced PANGAEA – a Data Publisher for Earth & Environmental Science. Rather than functioning as a simple storage space, PANGAEA is a curated information system, aimed at archiving, publishing and distributing georeferenced open access data from Earth system research. It was established more than 30 years ago and is jointly hosted by the Alfred Wegener Institute (AWI, Bremerhaven) and the Center for Marine Environmental Sciences (MARUM, Bremen) (Felden et al., 2023).

A key insight in this presentation was that data publication should be understood as a fundamental component of the research process rather than an optional add-on. Researchers naturally focus only on their own analysis and interpretation, yet the datasets they produce often hold untapped potential and can serve as the foundation for entirely new research questions – if archived properly. Over time, however, the information content of data fades if context is lost. Without sufficient metadata, datasets become difficult – or impossible – to interpret, turning potentially valuable data into “dead” data stored locally without public access.

PANGAEA addresses this challenge through the following:

- long-term archiving (with a storage horizon of at least 10 years),
- assignment and use of persistent identifiers (PIDs), including DOIs,
- adherence to FAIR principles,
- usage of official ontologies and controlled vocabularies,
- high metadata quality standards and interoperability,
- professional data curation and editorial review,
- integration with scientific publishing workflows.

The submission process includes initial upload, editorial review, curated processing, and dissemination via <http://www.pangaea.de> (last access: 10 October 2026), through portals, search engines, and international data infrastructures. During review, datasets are only visible to authors via private links; publication occurs only after author approval. PANGAEA supports a wide range of data types (e.g., tabular data, images, videos, shapefiles) but does not archive code or genetic sequences directly, instead allowing links to external repositories such as GitHub and ENA (European Nucleotide Archive) (see Fig. 1). Researchers can submit their data through the PANGAEA submission portal (see Further reading).

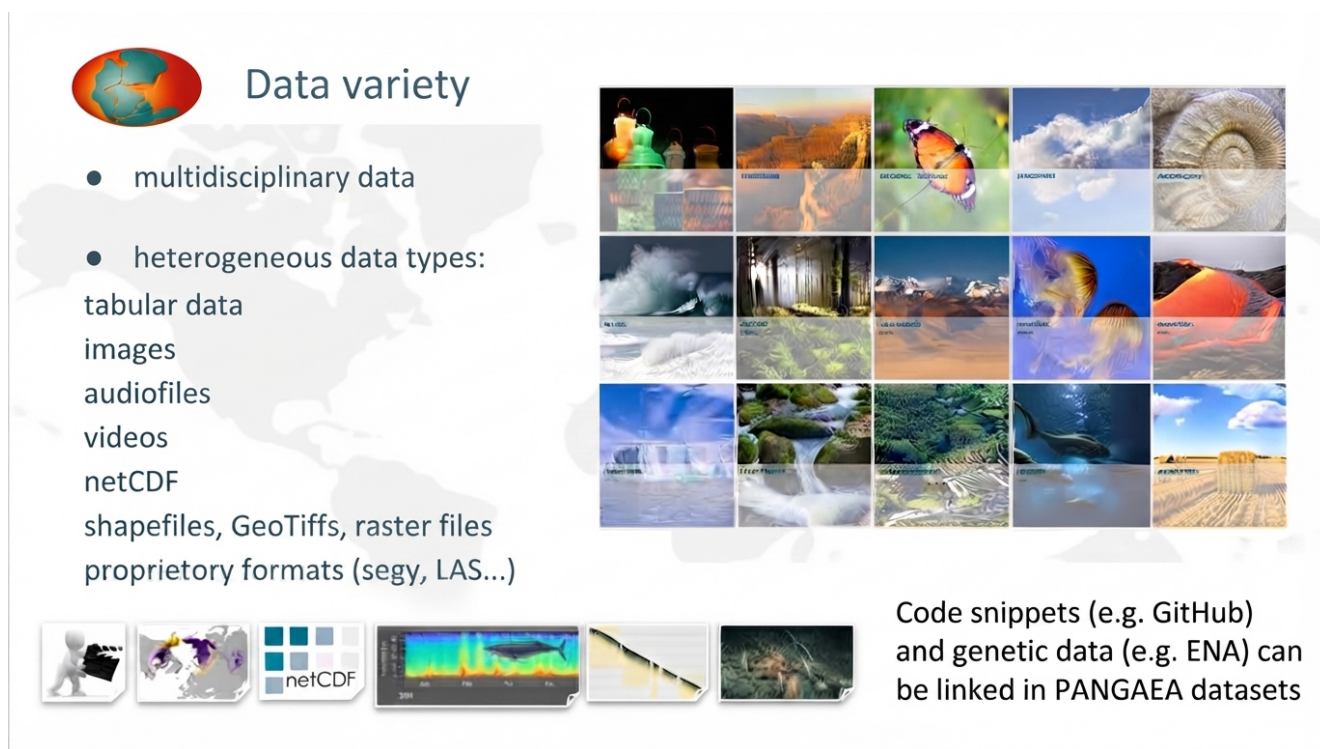


Figure 1. Screenshot from the online “Let’s talk data archiving” seminar showing the PANGAEA presentation. The slide illustrates the variety of data types that can be archived in PANGAEA, including tabular data, images, videos, and geospatial files, highlighting the repository’s versatility in supporting long-term data management and reuse.

The panelists emphasized that an extra pair of eyes is extremely valuable: experienced data managers, or repository support staff often identify missing metadata, unclear units, or undocumented abbreviations that would otherwise limit reuse. Units, abbreviations, and contextual explanations were repeatedly emphasized as essential components of reusable datasets.

2.2 The Arctic Data Center: data preservation, reuse, and training

Nicole Greco (Arctic Data Center, National Center for Ecological Analysis and Synthesis; greco@nceas.ucsb.edu) presented the Arctic Data Center, a repository serving as the primary data and software archive for Arctic research funded by the US National Science Foundation (NSF) Office of Polar Programs. Importantly, the Arctic Data Center also accepts data from projects beyond NSF funding.

In addition to archiving, the Arctic Data Center helps researchers

- learn data science best practices through training courses and workshops (e.g., reproducible research in R),
- submit and share data for free with expert curation support,

- upload code, data, and documents from their research,
- get DOIs for datasets to use in publications and grant reporting,
- track dataset use with detailed metrics and FAIR assessments,
- find and explore data using spatial and keyword-based tools (see Fig. 2).

At the time of the seminar, the Arctic Data Center hosted data from more than 5400 creators, comprising over 7000 datasets and 170+ TB of content. Submissions are free and unlimited, and a wide range of file formats are accepted, including model output, tabular data, text, images, videos, and observational data. All datasets are fully open access, with special conditions for sensitive social science data subject to privacy regulations.

Metadata guidelines at the Arctic Data Center focus on answering key questions: what was measured, why, by whom, when, where, and how the data are structured and can be reused. Curation substantially improves FAIRness, and all published datasets receive a DOI, can be updated with strict versioning standards, and are discoverable through the DataONE network, which connects dozens of environmental and Earth science data repositories and improves cross-platform discoverability. Interactive data viewers, map-based

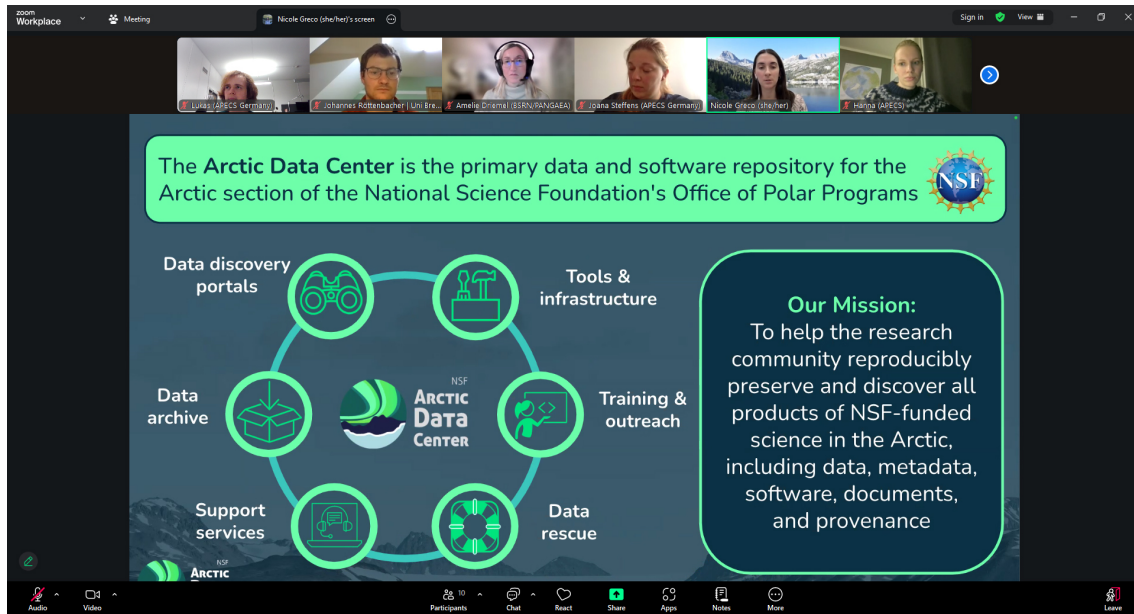


Figure 2. Screenshot from the online “Let’s talk data archiving” seminar showing the Arctic Data Center presentation. The slide highlights the services offered by the Center, including data archiving, training workshops, data rescue, user support, and data discovery tools, emphasizing its role in supporting FAIR and reusable data practices for Arctic research. Permission to publish the photographs of all individuals shown was obtained.

discovery tools, and custom data portals further support data reuse, sharing, and reporting to funders.

More information on data submission and the support services offered by the Arctic Data Center is provided in the Further reading section.

2.3 Project-level data management: the (AC)³ perspective

Johannes Röttenbacher (University of Bremen; johannes.roettenbacher@uni-bremen.de) presented the perspective of the (AC)³ project (Arctic Amplification: Climate Relevant Atmospheric and Surface Processes and Feedback Mechanisms), a large collaborative research initiative investigating Arctic climate processes across multiple spheres and their interactions with mid-latitudes (see Fig. 3).

In (AC)³, dedicated data managers act as intermediaries between researchers and repositories such as PANGAEA. Manuscripts and datasets are reviewed internally before submission to ensure completeness and metadata quality, reducing the workload for repository curators and increasing reproducibility. At the time of the seminar, (AC)³ had published around 2900 datasets.

This approach highlights the importance of building good data habits early, supported by project-level standards, workshops, and training. Streamlined workflows help integrate FAIR data management into everyday research practice rather than treating it as a final administrative burden.

Another step the project takes to make data easier to work with is by asking its scientists to share short code snippets (e.g., Jupyter notebooks, ...) showcasing the read in process, necessary preprocessing steps and first plotting steps for the data. The researchers are further encouraged to publish their analysis code and any programs, which could be reused on similar datasets. The code publishing also includes an internal peer review by the data steward to improve quality and help early career researchers to write more understandable code.

Further information about the (AC)³ project is provided in the Further reading section.

3 Discussion and key takeaways

The discussion addressed practical challenges frequently encountered by early-career researchers.

- Incomplete or poorly documented datasets were identified as a major barrier to reuse. If gaps are too large or contextual information is missing, repositories may decline publication. Legacy datasets with missing metadata illustrated how time-consuming reconstruction can be.
- Planning FAIR data management from the start is crucial. Panelists emphasized the use of data management plans, which should be updated as projects evolve. Technical solutions, such as self-describing formats (e.g., NetCDF), help keep data and metadata together.

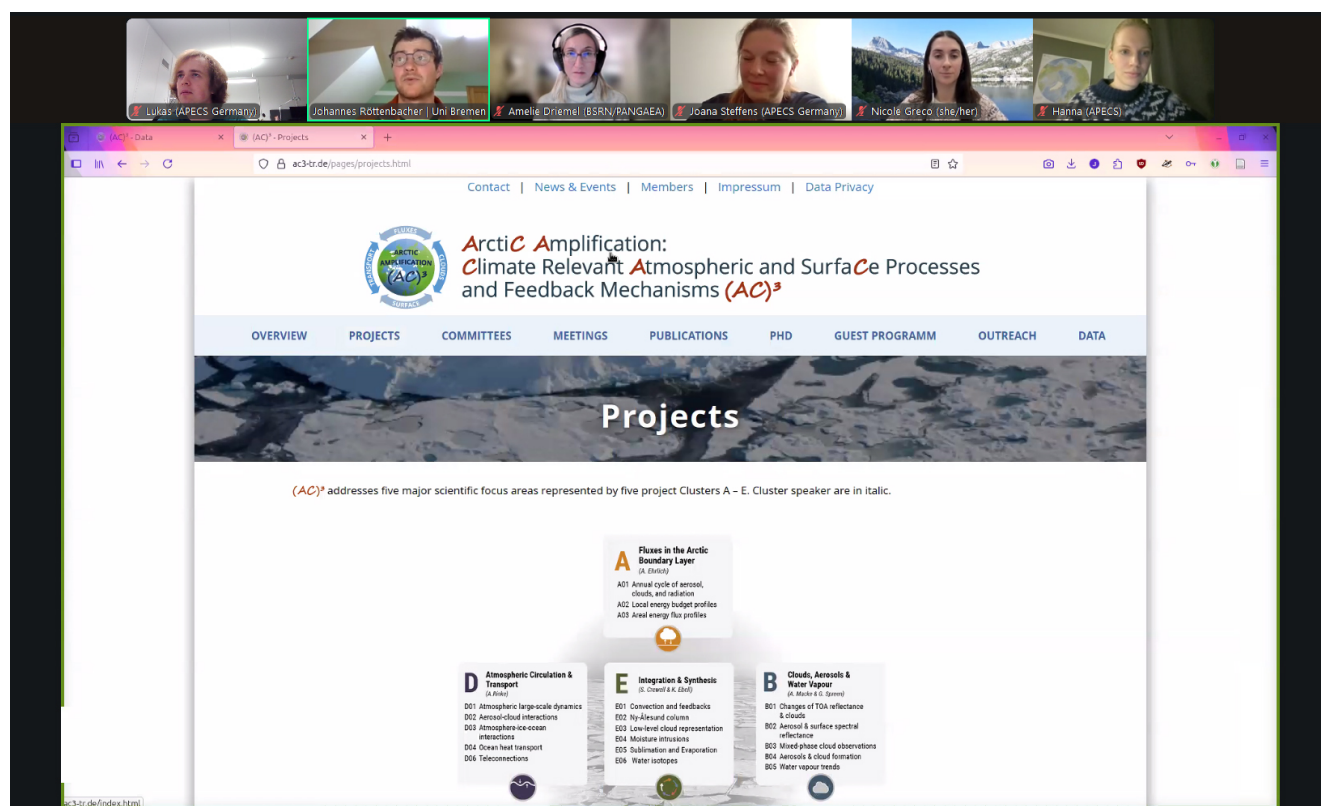


Figure 3. Screenshot from the online “Let’s talk data archiving” seminar showing the (AC)³ project presentation. The slide illustrates the project’s main scientific focus areas, including fluxes in the Arctic boundary layer, atmospheric circulation and transport, clouds, aerosols and water vapor, and integration and synthesis, highlighting how multiple research clusters work together to study Arctic climate processes. Permission to publish the photographs of all individuals shown was obtained.

Even simple formats like CSV can be improved by structured comments and clear documentation.

- Responsibility for FAIR data management was described as both top-down and bottom-up. Principal investigators play a key role in setting expectations and ensuring continuity when team members change institutions, while early-career researchers benefit from learning FAIR principles early and seeking support.
- Data security and backup: repositories rely on distributed storage and multiple physical locations. PAN-GAEA follows CoreTrustSeal requirements and stores data redundantly in Hamburg and Bremerhaven. ADC physically hosts data at the University of California, Santa Barbara, with multiple backup systems.
- Repository cooperation: initiatives such as DataONE connect platforms, provide unified discovery services, and improve cross-repository findability: <https://www.dataone.org/> (last access: 10 October 2026).
- Community-driven standards: the PaCTS 1.0 reporting standard for paleoclimate data (Khider et al., 2019) provides a framework for including essential metadata el-

ements – such as site information, measurement methods, chronology, and uncertainties – to maximize data reuse and comparability across studies. Likewise, the Climate and Forecast (CF) Metadata Conventions provide a widely adopted standard for describing Earth system and climate datasets stored in NetCDF format, improving interoperability across software tools and repositories (Eaton et al., 2017). Together, such community-developed standards reduce ambiguity, facilitate automated data processing, and strengthen FAIR data archiving.

- Archiving as an investment: archiving not only is about compliance but also produces scientific return on investment. Piwowar et al. (2011) note that funding data repositories yields substantial benefits by enabling reuse of datasets across many independent studies, often exceeding the impact of the original research. This reinforces the seminar message that investing effort in FAIR and long-term archiving strengthens both individual and community research outcomes.

4 Recommendations for strengthening FAIR data archiving

While the seminar primarily aimed to introduce practical aspects of data archiving, the discussions also highlighted several broader challenges that extend beyond individual repositories. From an early-career perspective, one recurring theme was that data archiving is still too often viewed as an administrative task rather than an integral part of scientific research. This perception can discourage researchers from investing sufficient time in metadata preparation and documentation, particularly when publication deadlines or project deliverables take priority.

A first recommendation is therefore to better recognize data archiving within research funding and project planning. Funding agencies and research institutions could explicitly acknowledge the time and expertise required for high-quality data management by incorporating dedicated resources for data stewardship, metadata preparation, and repository submission into research proposals. Embedding data management activities throughout a project rather than treating them as an end-of-project obligation would reduce workload and improve data quality.

From the repository perspective, many metadata issues encountered during editorial curation could be avoided if researchers engaged with repositories earlier in the research process. Early communication reduces the need for extensive revisions and improves both metadata quality and long-term data usability.

Second, training opportunities should be expanded. Although awareness of the FAIR principles has increased considerably in recent years, many early-career researchers receive limited practical instruction on preparing reusable datasets. Dedicated workshops at conferences, graduate schools, and institutional training programs could complement repository-specific guidance and provide hands-on experience with metadata standards, persistent identifiers, and data publication workflows.

From the repository perspective, training is particularly effective when it focuses on real datasets and submission workflows rather than only introducing FAIR principles conceptually. Repository-led workshops and consultation services therefore complement institutional research data management training.

Third, continued development and adoption of community standards will further facilitate data reuse. Standards such as PaCTS and the CF Metadata Conventions provide common frameworks for documenting datasets in consistent and interoperable ways. Their broader adoption reduces ambiguity, improves machine readability, and lowers the effort required for future users to understand and integrate published datasets.

Finally, fostering a research culture that values data publication alongside traditional research outputs will benefit both individual researchers and the scientific community. Assign-

ing persistent identifiers, encouraging citation of datasets, and recognizing high-quality data publications in research assessment can provide tangible incentives for investing in FAIR data practices. Experiences from (AC)³ further demonstrate that dedicated project data managers can substantially reduce the workload for researchers by coordinating internal quality assurance before repository submission.

Early-career researchers have an important role in driving this cultural change by adopting good practices early in their careers, while supervisors, institutions, repositories, publishers, and funding organizations all share responsibility for creating an environment in which FAIR data management becomes routine rather than exceptional.

5 Thoughts and feedback

The seminar demonstrated that data archiving is not merely a technical requirement but a shared responsibility and an opportunity to increase the longevity and impact of scientific work. Participants valued practical insights, open discussion, and the chance to ask questions in a non-judgmental environment.

For early-career researchers, understanding data archiving early can prevent data loss, reduce stress at the end of projects, and support reproducible science. The strong engagement during the seminar reinforced the importance of creating spaces like “Let’s talk ...!” for peer-to-peer exchange on topics often underrepresented in formal training.

APECS Germany looks forward to continuing this seminar series and supporting the polar research community in navigating data management challenges together. We hope that discussions such as these contribute not only to increasing awareness of FAIR data principles but also to fostering a research culture in which high-quality data stewardship is recognized as an essential component of excellent science.

6 Further reading

- PANGAEA – Data submission and guidance: <https://www.pangaea.de/submit/> (last access: 10 October 2026);
- Arctic Data Center – Data archive, training, and support: <https://arcticdata.io/> (last access: 12 October 2026);
- (AC)³ project – Arctic climate system research and data management: <https://www.ac3-tr.de/> (last access: 12 October 2026);
- DataONE – Environmental data discovery across repositories: <https://www.dataone.org/> (last access: 10 October 2026).

Data availability. No datasets were used in this article.

Author contributions. JKS conceptualized the article, wrote the text based on her notes and contributions from LMK, and integrated feedback from all co-authors. LMK contributed to the text through notes and feedback on manuscript drafts. AD, NG, and JR provided expert content and insights from the seminar, which were incorporated into the article. All authors reviewed and approved the final version.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We thank Amelie Driemel (PANGAEA), Nicole Greco (Arctic Data Center), and Johannes Röttenbacher ((AC)³ project) for sharing their expertise and practical experiences, and all participants for their thoughtful questions and engagement. This seminar was organized as part of the "Let's talk . . . !" series by APECS Germany and relies on voluntary contributions from early-career scientists.

Acknowledgement of the use of ChatGPT (<https://chatgpt.com/>, last access: 26 September 2026) for assistance with partial rephrasing of text segments, which supported the clarity and refinement of certain sections.

Review statement. This paper was edited by Lena Nicola and reviewed by Erik Loebel.

References

- Colavizza, G., Cadwallader, L., LaFlamme, M., Dozot, G., Lecorney, S., Rappo, D., and Hrynaszkiewicz, I.: An analysis of the effects of sharing research data, code, and preprints on citations, *PLOS ONE*, 19, e0311493, <https://doi.org/10.1371/journal.pone.0311493>, 2024.
- Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., and Horne, R.: NetCDF Climate and Forecast (CF) Metadata Conventions, <http://cfconventions.org/> (last access: 12 October 2026), 2017.
- Felden, J., Möller, L., Schindler, U., Huber, R., Schumacher, S., Koppe, R., Diepenbroek, M., and Glöckner, F. O.: PANGAEA – Data Publisher for Earth & Environmental Science, *Sci. Data*, 10, 347, <https://doi.org/10.1038/s41597-023-02269-x>, 2023.
- Genderjahn, S. and Schrader, A.: Persistente Identifikatoren als Fundament unserer Wissenslandschaft: Die PID-Roadmap für Deutschland, 114. BiblioCON, Berlin, Zenodo [data set], <https://doi.org/10.5281/zenodo.20426735>, 2026.
- Khider, D., Emile-Geay, J., McKay, N. P., Gil, Y., Garijo, D., Ratnakar, V., Alonso-Garcia, M., Bertrand, S., Bothe, O., Brewer, P., Bunn, A., Chevalier, M., Comas-Bru, L., Csank, A., Dassié, E., DeLong, K., Felis, T., Francus, P., Frappier, A., Gray, W., Goring, S., Jonkers, L., Kahle, M., Kaufman, D., Kehrwald, N. M., Martrat, B., McGregor, H., Richey, J., Schmittner, A., Scroton, N., Sutherland, E., Thirumalai, K., Allen, K., Arnaud, F., Axford, Y., Barrows, T., Bazin, L., Pilaar Birch, S. E., Bradley, E., Bregy, J., Capron, E., Cartapanis, O., Chiang, H.-W., Cobb, K. M., Debret, M., Dommain, R., Du, J., Dyez, K., Emerick, S., Erb, M. P., Falster, G., Finsinger, W., Fortier, D., Gauthier, N., George, S., Grimm, E., Hertzberg, J., Hibbert, F., Hillman, A., Hobbs, W., Huber, M., Hughes, A. L. C., Jaccard, S., Ruan, J., Kienast, M., Konecky, B., Le Roux, G., Lyubchich, V., Novello, V. F., Olaka, L., Partin, J. W., Pearce, C., Phipps, S. J., Pignol, C., Piotrowska, N., Poli, M.-S., Prokopenko, A., Schwanck, F., Stepanek, C., Swann, G. E. A., Telford, R., Thomas, E., Thomas, Z., Truebe, S., von Gunten, L., Waite, A., Weitzel, N., Wilhelm, B., Williams, J., Williams, J. J., Winstrup, M., Zhao, N., and Zhou, Y.: PaCTS 1.0: A crowdsourced reporting standard for paleoclimate data, *Paleoceanogr. Paleoclimatol.*, 34, 1570–1596, <https://doi.org/10.1029/2019PA003632>, 2019.
- Piowar, H. A., Vision, T. J., and Whitlock, M. C.: Data archiving is a good investment, *Nature*, 473, 285, <https://doi.org/10.1038/473285a>, 2011.
- Wilkinson, M., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hoof, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 3, 160018, <https://doi.org/10.1038/sdata.2016.18>, 2016.