PENGUIN: AI-Assisted Cross-Platform Discovery of Research Records Using the Polar Science Knowledge Graph

Qi Zhang[1,*], Masayoshi Kozai[1], Kunio T. Takahashi[1,2,3], Yoshimasa Tanaka[1,2,3], Akira Kadokura[1] and Masaki Kanao[1,2,3]

[1]*Polar Environment Data Science Center, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Tachikawa, Japan*
http://pedsc.rois.ac.jp/en/
[2]*National Institute of Polar Research, Tachikawa, Japan*
https://www.nipr.ac.jp/english/
[3]*The Graduate University for Advanced Studies, SOKENDAI, Tachikawa, Japan*
https://www.soken.ac.jp/en/

Abstract

PENGUIN is an interactive web application built on the Polar Science Knowledge Graph (PSKG) to support cross-platform discovery of heterogeneous polar research records and help users understand why these records may be scientifically related. PSKG provides a record-level graph index over distributed polar science resources. It represents records from source platforms as graph nodes with URLs linking back to their origins, associates them with scientific entities, and connects those entities through explicit scientific relations. PENGUIN turns this index into a graph exploration workspace in which users can search for records and entities, explore the graph, collect records of interest, inspect relation paths, and access original records on their source platforms. It also supports AI-assisted exploration, in which a research question expressed in natural language guides stepwise graph exploration through relevant entities and relations toward potentially useful records. By exposing the graph paths behind discovery while preserving links to original sources, PENGUIN enables users to move from a research question to related records distributed across platforms, while also allowing exploration to continue from records encountered on connected source platforms. This demo presents the interface and representative exploration workflows of PENGUIN.

Keywords

Knowledge graphs, Polar science, AI-assisted exploration, Graph-based data discovery, Cross-platform discovery

1. Introduction

To understand complex environmental phenomena and ecological processes, polar research relies on research records distributed across data platforms, domain-specific databases, archives, and non-database research materials. Previous efforts in the polar community have supported cross-repository discovery through federated search [1, 2] and enhanced semantic discovery of research data through semantic annotation [3]. These approaches improve the findability and accessibility of distributed resources, but a gap remains between a research question and the source records that may help address it. For example, a researcher investigating "what may influence changes in Adélie penguin populations" may need records on sea ice, ocean conditions, and prey species. These concepts are not explicitly mentioned in the question, while the term "Adélie penguin" may not appear in the metadata of those records. The scientific relations that make those records relevant to the question may therefore be overlooked in keyword-based discovery. To address this gap, a companion study introduces the Polar Science Knowledge Graph (PSKG) as a graph-based, record-level index over distributed polar

IJCKG 2026 Posters, Demos and Workshops, November 19–21, 2026, Bangkok, Thailand
*Corresponding author.
zhang.qi@nipr.ac.jp (Q. Zhang)
0000-0002-7551-5792 (Q. Zhang); 0000-0002-3948-3666 (M. Kozai); 0000-0003-1868-8176 (K. T. Takahashi); 0000-0002-6105-9562 (A. Kadokura); 0000-0001-6330-4691 (M. Kanao)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

CEUR-WS.org/Vol-4316/pd-short-1.pdf

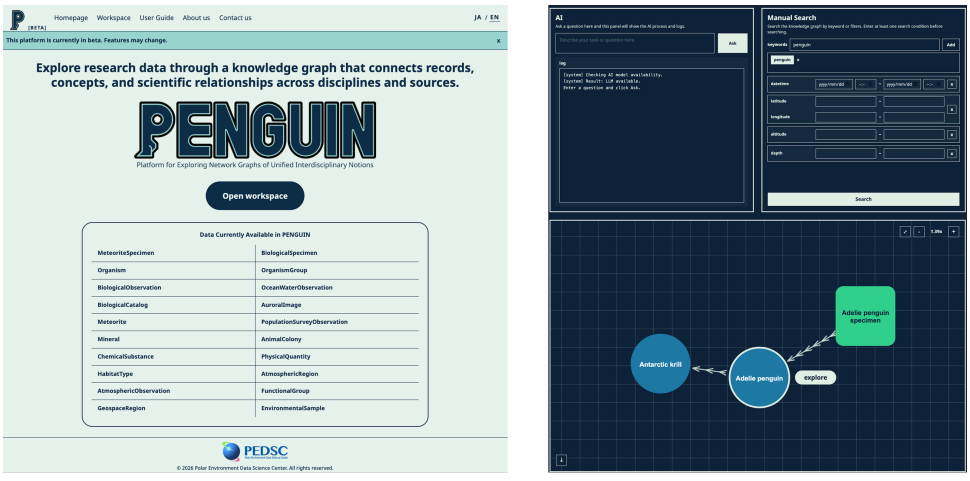
science resources [4]. Building on the general knowledge-graph model [5], the PSKG represents source records and the scientific context connecting them in a layered structure: the data layer consists of nodes representing research records, while the knowledge layer contains scientific entities and explicit relations that provide paths between those records.

PENGUIN (Platform for Exploring Network Graphs of Unified Interdisciplinary Notions) operationalizes PSKG through a web-based interface. It supports AI-assisted exploration that guides users from a natural-language research question toward potentially relevant entities and records, while retaining keyword-based search for users who know relevant terms or record attributes. The resulting nodes and relations are displayed in an interactive graph workspace, where users can expand the surrounding graph, inspect node and relation details, and access records on their source platforms. The companion work describes the PSKG architecture and construction workflow; this paper focuses on how PENGUIN turns the resulting index into a practical, interactive environment for exploring polar research records and their scientific context.

2. Overview of the PENGUIN Platform

PENGUIN is a web-based platform through which researchers and data managers access and explore the Polar Science Knowledge Graph (PSKG). PSKG provides the underlying infrastructure for integrating representations of polar research records, scientific entities, and their relations across source platforms. PENGUIN builds on this infrastructure to provide keyword search and interactive graph exploration. Its optional AI module operates at the application layer, assisting users in navigating the information integrated in PSKG. Figure 1 shows the home page and main workspace. The home page lists the principal node types currently represented in the PSKG, giving users an initial overview of the available content. The *Open workspace* button launches the interactive graph workspace.

The main interface consists of an AI-assisted exploration module, a keyword-based search module, and a graph workspace. The graph workspace maintains the visible subgraph assembled during the current session. Data-layer nodes are shown as green rounded rectangles, knowledge-layer nodes as blue circles, and edges are labeled by relation type. Users can inspect, remove, expand, and download the assembled subgraph for later use. The current implementation is available at https://penguin.nipr.ac.jp/. Table 1 summarizes the scope of PSKG v0.1.



(a) Home page (b) Main workspace

Figure 1: Overview of the PENGUIN interface. (a) The home page presents the principal research-record and knowledge-entity types represented in the PSKG. (b) The main workspace contains the AI-assisted exploration module, the search module, and the shared graph workspace.

Table 1
Scope of PSKG v0.1. Relation counts represent individual edges; record units vary by resource.

Component	Count	Component	Count
Source platforms	6	Integrated resources	10
Data-layer record nodes	16,129	Knowledge-layer nodes	1,042
Record–entity relations	71,390	Entity–entity relations	3,082
Node labels	20	Relation types	38

Source platforms: AMIDER; Arctic and Antarctic Data archive System (ADS); NIPR Science Database; Polar Fungi; Southern Ocean Zooplankton Monitoring Database; Syowa Station illustrated guide.
Resource types: observations, specimens, auroral images, long-term products, and illustrated scientific guides.

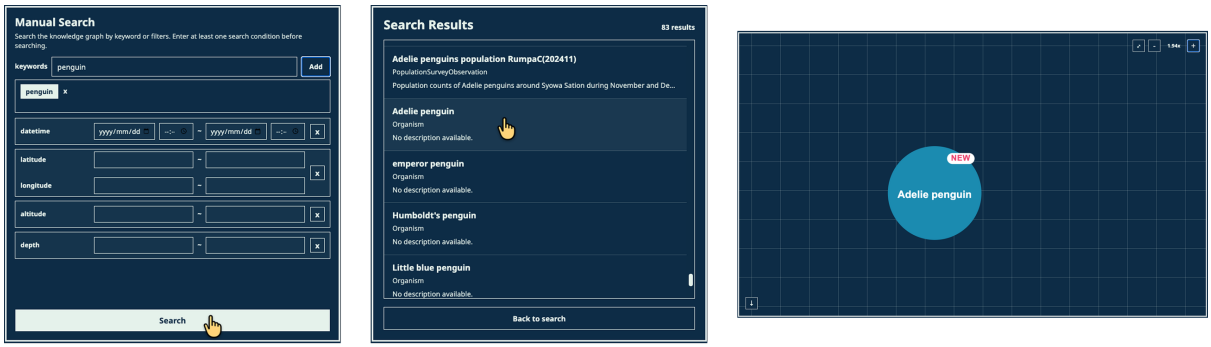
3. Demonstration Workflows

The following workflows demonstrate how PENGUIN supports progressive, source-traceable discovery through a shared graph workspace. We first describe the keyword search and interactive graph functions available when users know relevant terms or record attributes, and then show how AI-assisted exploration guides users from a natural-language research question toward potentially relevant records. A separate final scenario demonstrates how PENGUIN can connect independent source platforms by serving as a cross-platform discovery hub.

3.1. Keyword Search and Interactive Graph Exploration

For users who know a relevant term or record attribute, PENGUIN provides keyword-based search as a direct way to locate a seed node. The interactive workflow then combines seed selection, graph-neighborhood expansion, and inspection of graph elements and their supporting information. Shared scientific entities connect records across platforms, while explicit relations between these entities support discovery beyond the terms in the initial query. For example, following the prey relation from Adélie penguins to Antarctic krill can lead to krill specimen records without requiring the user to include krill in the query.

Users can combine keywords with optional ranges for date and time, latitude, longitude, altitude, and depth. After a query is submitted, the search results panel displays candidate nodes. Selecting a result adds the corresponding node to the graph workspace as an entry point for further exploration (Figure 2).

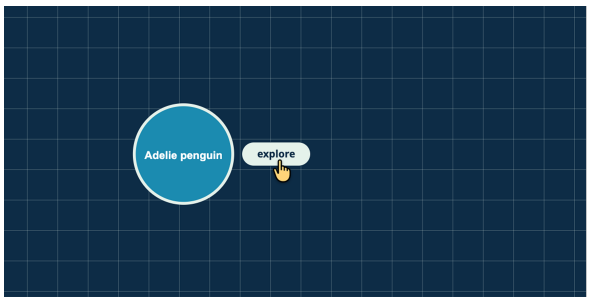


(a) Search criteria (b) Candidate nodes (c) Selected node in graph workspace

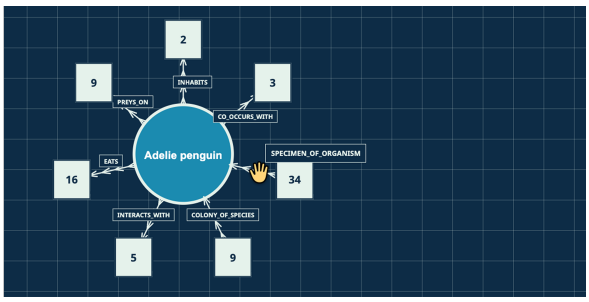
Figure 2: Keyword search in PENGUIN. (a) The user specifies a keyword and optional date/time, latitude, longitude, altitude, and depth filters. (b) The search module returns candidate nodes. (c) A selected node is added to the graph workspace as the exploration seed.

Once a seed node is visible, the user can invoke *Explore* to retrieve adjacent nodes. Because a node may have many neighbors, PENGUIN groups candidates by relation type and displays the number

available in each group. The user can select a relation type, review the associated candidates, and add chosen nodes together with their connecting relations (Figure 3). Repeating this process progressively extends the visible subgraph, allowing users to discover potentially related records together with the typed relations and intermediate entities that connect them.



(a) Invoke Explore



(b) Neighbor groups by relation type



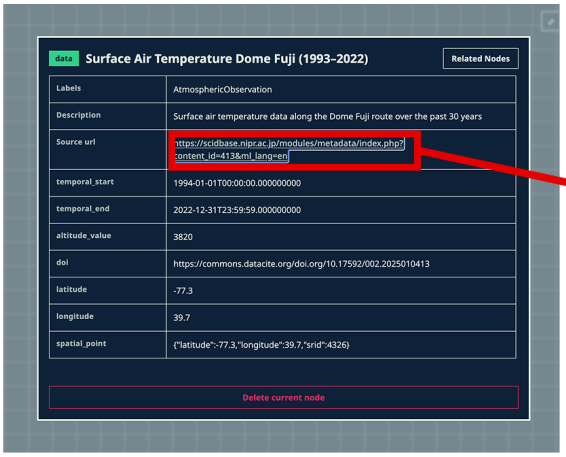
(c) Candidate neighboring nodes



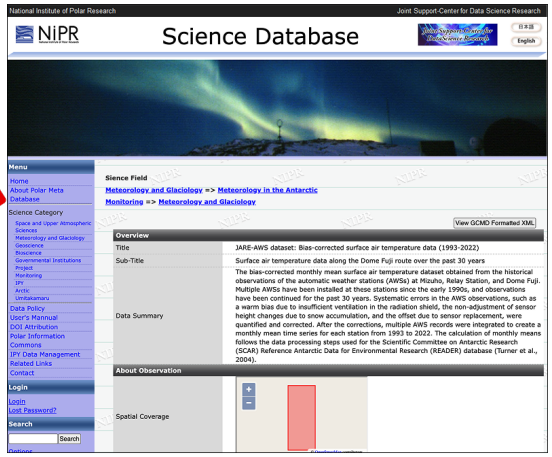
(d) Selected neighbor added

Figure 3: Progressive graph-neighborhood expansion. (a) The user invokes *Explore* on a visible node. (b) Neighboring nodes are grouped by relation type. (c) The user reviews candidates in a selected group. (d) A chosen node and its connecting relation are added to the workspace.

Users can double-click any visible node or relation to inspect its properties and supporting information. For a data-layer node, the source URL normally points to the corresponding record page on its authoritative platform. For a knowledge-layer node or relation, an evidence URL may instead identify an external reference supporting the represented scientific entity or relation. Figure 4 shows an atmospheric-observation data-layer node representing surface air temperature observations along the Dome Fuji route. Its source URL points to the corresponding metadata record in the NIPR Science Database, which serves as the authoritative source platform in this example [6].



(a) Data-layer node details



(b) Authoritative source record

Figure 4: Inspection of a data-layer node and navigation to its source record. (a) The detail panel displays node properties and a source URL. (b) Following the URL opens the corresponding record page on its authoritative source platform, which in this example is the NIPR Science Database. This source page corresponds to multiple data-layer nodes with different start dates; panel (a) shows one of these nodes. Knowledge-layer nodes and relations may likewise include links to supporting references where available.

3.2. AI-Assisted Exploration from Research Questions to Records

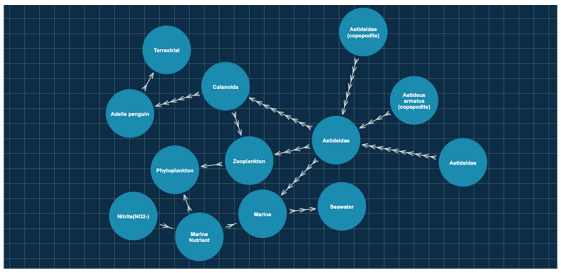
A central function of PENGUIN is to help users move from a research question toward potentially relevant research records without requiring them to know the keywords, domain terminology, classification schemes, or record types used by relevant data sources. Research on data discovery has likewise emphasized the need for repository interfaces to account for users' search contexts and diverse search behaviors [7]. To support such question-driven and interdisciplinary discovery, users can initiate AI-assisted exploration by submitting a research question in natural language.

The current implementation uses Google's Gemma 3 (12B) large language model (LLM), deployed locally through Ollama (model tag: `gemma3:12b`) with default generation settings. After a question is submitted to the AI-assisted exploration module, the system expands the graph step by step. At each step, the backend supplies the language model with the question, a summary of the current workspace subgraph, and candidate graph elements derived from the PSKG. The model selects a promising exploration direction from these candidates, after which the backend retrieves the corresponding nodes and relations. An exploration log displays the process and allows the user to pause it. Exploration stops when the model judges that further graph expansion would not provide useful new information relevant to the question. For a source containing many similar relevant records, the system retrieves only a few examples, without enumerating all matching records.

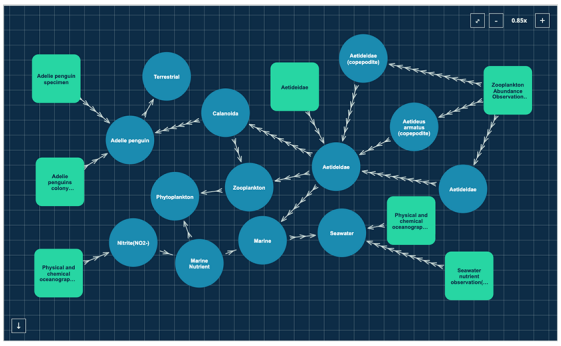
The workflow first identifies potentially relevant knowledge-layer entities and expands the graph through their relations to construct a contextual subgraph. It then continues from selected knowledge-layer entities to connected data-layer nodes representing source records. Figure 5 shows this process for a question concerning algal productivity and Antarctic coastal food webs.



(a) Query and exploration log



(b) Contextual knowledge-layer subgraph



(c) Expanded subgraph with data-layer nodes

Figure 5: AI-assisted exploration in PENGUIN. (a) The user submits a natural-language research question and monitors the exploration log. (b) The system constructs a contextual knowledge-layer subgraph. (c) Connected data-layer nodes are subsequently added to the same workspace.

The language model does not directly add graph elements to the workspace or provide a standalone final answer. All nodes and relations added during the process are retrieved from the PSKG and remain visible for inspection. Users can examine connection paths and supporting sources, remove elements, and continue manually from the resulting subgraph. Keyword-based exploration starts from a user-supplied term, with users guiding subsequent graph expansion. AI-assisted exploration uses the question, current workspace, and PSKG-derived candidates to help select exploration directions. The

LLM thus serves as an application-layer navigation assistant, while PSKG supplies the integrated record representations and scientific relations used by both workflows. This demonstration focuses on how AI uses PSKG to locate research records. A task-based comparison with conventional keyword-based search, including retrieval effectiveness and user experience, remains future work.

3.3. PENGUIN as a Cross-Platform Discovery Hub

Beyond these search and exploration functions, PENGUIN is designed to serve as a graph-based discovery hub across distributed source platforms. Source platforms remain responsible for publishing, describing, and providing access to their records, while PENGUIN exposes the PSKG as a record-level graph index through which records maintained by different platforms can be explored together.

A connected source platform can include a dedicated PENGUIN link on the page of a record represented in the PSKG. Following the link opens PENGUIN with the corresponding data-layer node already loaded in the workspace. From this node, users can explore the PSKG to reach related data-layer nodes representing records from other platforms. Because each data-layer node retains its source URL, users can then open the corresponding record page on its authoritative source platform.

Figure 6 demonstrates the complete pathway. Starting from a record in the Southern Ocean Zooplankton Monitoring Database [8], the user enters PENGUIN at the corresponding data-layer node, follows a graph path through knowledge-layer entities related to Hydrozoa to a specimen record, and then opens the corresponding authoritative record on AMIDER [9]. This platform-to-PENGUIN-to-platform pathway supports cross-platform discovery without requiring participating systems to centralize their underlying records or adopt a common internal schema.

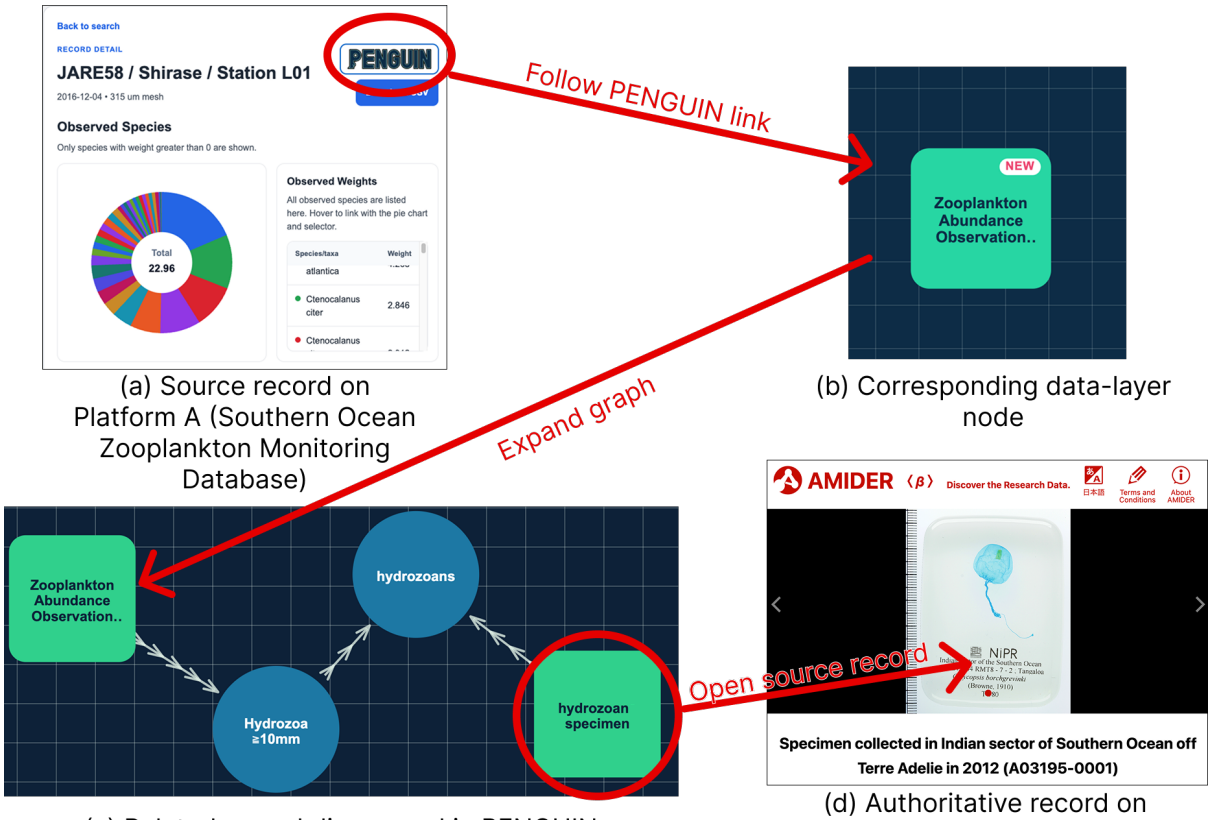


Figure 6: Cross-platform discovery through PENGUIN. (a) A record in the Southern Ocean Zooplankton Monitoring Database (Platform A) provides a PENGUIN link. (b) The corresponding data-layer node is loaded as the initial node in the workspace. (c) Graph expansion through related knowledge-layer entities and typed relations reveals a hydrozoan specimen record. (d) Following the discovered node's source URL opens the authoritative record on AMIDER (Platform B).

4. Conclusion

This paper presented PENGUIN, an interactive web application for exploring the PSKG, a record-level graph index over polar research records distributed across databases and platforms. Rather than replacing source platforms or limiting discovery to a static result list, PENGUIN exposes data-layer nodes representing source records, knowledge-layer entities, and typed relations so that users can expand the graph, inspect relation paths, and follow links to authoritative source records or supporting evidence. While retaining keyword-based search for known terms and record attributes, PENGUIN uses AI-assisted exploration to guide users step by step from a natural-language research question through relevant scientific entities and relations toward potentially useful records, helping bridge the gap between research questions and relevant records distributed across platforms. Overall, this design supports inspectable and source-traceable cross-platform discovery while preserving the authority and independence of the source platforms.

The companion PSKG study reports seven competency-question checks and three discovery cases demonstrating cross-platform access and relation-based discovery from predefined seed entities [4]. These results support the feasibility of the graph mechanisms used by PENGUIN; task-based evaluation of retrieval effectiveness and user experience remains future work.

Acknowledgments

This work was supported by the Strategic Research Projects program of the Research Organization of Information and Systems (ROIS) and JSPS KAKENHI Grant Number JP24K14966. We also thank AMIDER (https://amider.rois.ac.jp/), ADS (https://ads.nipr.ac.jp/), the NIPR Science Database (https://scidbase.nipr.ac.jp/), and the NIPR Bioscience Group (https://www.nipr.ac.jp/biology/) for their support and resources.

5. Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

The authors used ChatGPT to assist with spelling and grammar checks, wording revisions based on author-provided content, LaTeX template migration, typesetting, and format checking. The authors reviewed and edited the suggested changes and take full responsibility for the final manuscript.

References

[1] Chantelle Verhey, Melinda Minch, Karen Payne, Polar Federated Search: New infrastructure to support the polar community, Polar Science 36 (2023) 100947. doi:10.1016/j.polar.2023.100947.

[2] Lynn Yarmey, Siri Jodha Singh Khalsa, Building on the International Polar Year: Discovering interdisciplinary data through federated search, Data Science Journal 13 (2014) PDA79–PDA82. doi:10.2481/dsj.IFODA-13.

[3] Steven S. Chong, Mark Schildhauer, Margaret O'Brien, Bryce Mecum, Matthew B. Jones, Enhancing the FAIRness of Arctic research data through semantic annotation, Data Science Journal 23 (2024) 2. doi:10.5334/dsj-2024-002.

[4] Qi Zhang, Masayoshi Kozai, Kunio T. Takahashi, Yoshimasa Tanaka, Yasuyuki Minamiyama, Yoshizumi Miyoshi, Akira Kadokura, Masaki Kanao, The Polar Science Knowledge Graph: A Scientific Index for Cross-Platform Data Discovery, 2026. doi:10.22541/essoar.15005050/v2, preprint, version 2.

[5] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, Antoine Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 71:1–71:37. doi:10.1145/3447772.

[6] Masaki Kanao, Masaki Okada, Julie E. Friddell, Akira Kadokura, Science metadata management, interoperability and data citations of the National Institute of Polar Research, Japan, Data Science Journal 17 (2018) 1–6. doi:10.5334/dsj-2018-001.

[7] Mingfang Wu, Felicitas Löffler, Brigitte Mathiak, Fotis Psomopoulos, Uwe Schindler, Amir Aryani, Jordi Bodera Sempere, Antica Culina, Andreas Czerniak, Chris Erdmann, Kathleen Gregory, Nick Juty, Allyson Lister, Ying-Hsang Liu, Samantha Pearman-Kanza, Bridging the data discovery gap: User-centric recommendations for research data repositories, Data Science Journal 25 (2026) 6. doi:10.5334/dsj-2026-006.

[8] Polar Environment Data Science Center, Southern Ocean Zooplankton Monitoring Database, Available at: https://penguin.nipr.ac.jp/zooplankton/, 2026.

[9] Masayoshi Kozai, Yoshimasa Tanaka, Shuji Abe, Yasuyuki Minamiyama, Atsuki Shinbori, Akira Kadokura, AMIDER: A multidisciplinary research database and its application to promote open science, Data Science Journal 24 (2025) 7. doi:10.5334/dsj-2025-007.