NAVIGATING THE DATA DELUGE

A COLLABORATIVE VISION FOR A UNIFIED SUBSEA DATA ECOSYSTEM

By Vicki Ferrini, Christina Haskins, and Corinne Bassin

ABSTRACT. The marine science community is generating unprecedented volumes of data from advanced sensors, submersibles, and observational platforms, but fragmentation across the data ecosystem limits their discovery, value, and use. Inconsistent formats, variable at-sea workflows, and incomplete metadata, coupled with disconnected repositories and limited interoperability, hinder operational decision-making, data discovery, integration, and long-term reuse, creating inefficiencies for researchers. This paper synthesizes outcomes from the Data at Sea workshop, which brought together a cross-disciplinary community to address technical, operational, and institutional barriers spanning subsea data acquisition, near-real-time use, delivery, and stewardship. We present a collaborative framework envisioned by workshop participants for creating a more unified subsea data ecosystem built on shared tools, interoperable workflows, standardized products, improved practices, and community-driven knowledge sharing. Proposed actions include developing centralized knowledge resources, standardizing integration workflows, sharing of data processing plans, supporting data specialists, and expanding training. Reducing friction will improve operational efficiency, accelerate scientific insight, and broaden access to subsea data tools and knowledge.

INTRODUCTION

The marine science community is generating data at an unprecedented rate, driven by the rapid expansion and accessibility of advanced sensors and observational platforms. The full value of these data remains unrealized due to persistent friction points that hinder discovery, access, integration, and insight. Our ability to generate data has outpaced our ability to manage and operationalize it, limiting its potential to advance scientific discovery. This challenge is especially acute during at-sea operations, where large data volumes must be rapidly processed and interpreted to inform real-time decisions. Inconsistent data formats, products, and workflows further fragment this data ecosystem, highlighting an opportunity for community action.

The cross-disciplinary Data at Sea workshop brought together seagoing scientists, submersible operators, shipboard technicians, and data stewards to address challenges associated with data acquired by submersibles and generated in support of submersible operations. A multi-stage community engagement process informed workshop design. An initial interactive virtual panel discussion and participant survey identified key challenges, bottlenecks, and priority topics that shaped the agenda. The two-day workshop, held in October 2025 in Victoria, Canada, brought together 24 participants from eight countries and more than 15 organizations to address challenges in managing data related to submersible operations, referred to here as subsea data. The workshop progressed from defining problems to developing solutions. Participants discussed the current state of data practices, identified foundational datasets, shared experiences, and discussed challenges related to data stewardship, interoperability, and access. Topical breakout sessions then evaluated existing tools and resources, identified needs, and developed recommendations for community standards and tools.

This paper synthesizes workshop outcomes, including consensus on key challenges and opportunities for a more unified subsea data ecosystem. Participants agreed that a shift away from siloed efforts toward a collaborative, community-driven approach built on shared tools, interoperable workflows, open-source development, and common best practices is the best path forward. No single organization can solve the fragmentation of the current data ecosystem alone; meaningful progress requires coordinated investment, shared expertise, and collective action across the community. By improving the flow of data along all steps in the process from acquisition to archive, the community is poised to accelerate scientific insight while broadening access to data and knowledge.

THE FRACTURED DATA ECOSYSTEM

The ability to find, access, and effectively use the growing volume of observations is fundamental to advancing marine science. However, current practices result in inefficient, ad hoc approaches for data use. Workshop participants described technical, operational, and institutional barriers that limit scientific integration and data usability and create unnecessary barriers for new researchers (Figure 1).

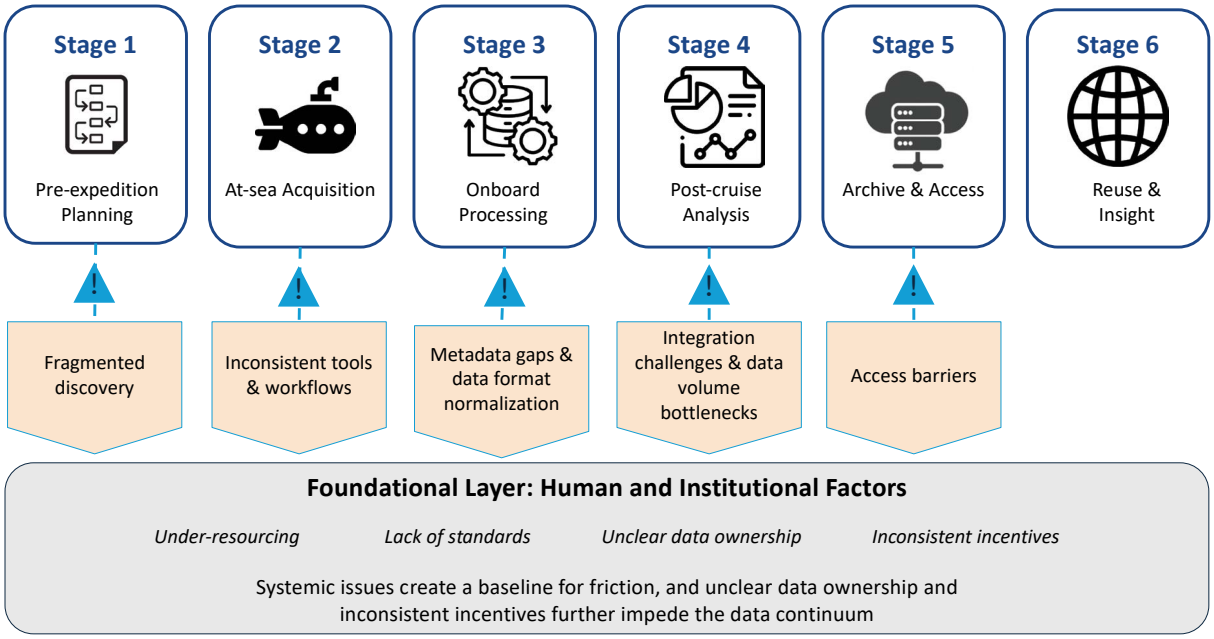

FIGURE 1. Friction points disrupt continuity throughout the data life cycle, creating inefficiencies in at-sea operations, limiting the ability to rapidly extract insight from data, and preventing observations from reaching their full scientific and societal value.

DATA DISCOVERY AND ACCESS

Challenges for researchers include identifying and accessing the data needed to plan at-sea operations, assessing knowledge gaps, and assembling critical datasets from online data systems such as the Marine Geoscience Data System (MGDS) and the Global Multi-Resolution Topography (GMRT) synthesis (Ferrini et al., 2025a, 2025b) for offline use at sea. Navigating the patchwork of repositories, websites, and personal hard drives often requires specialized knowledge, making data discovery inefficient and heavily reliant on personal networks and improvised methods. As a result, existing datasets may be overlooked, leading to redundant data acquisition that consumes limited at-sea time and constrains opportunities for new observations. This problem is especially acute for historical dive, sample, and mapping data, which can be difficult to locate, spatially contextualize, or integrate across data providers. Technologies to address these challenges exist but are not widely adopted, and opportunities for user feedback are often missed.

AT-SEA STANDARDS, TOOLS, AND WORKFLOWS

The absence of community-wide, at-sea standards, tools, and workflows creates an inefficient experience for researchers, especially for those moving between platforms. Inconsistencies across vessels, institutions, and expeditions create operational inefficiencies and steep learning curves for new users and force researchers to continually adapt their workflows. Successful standardization will require balancing community-wide consistency with the flexibility needed by individual programs.

Metadata, the foundation of FAIR (findable, accessible, interoperable, reusable) data practices (Wilkinson et al., 2016), remains a critical point of failure. Responsibility for metadata creation at sea is often distributed without clear ownership and accountability, resulting in incomplete or inconsistent results that limit interoperability and long-term usability. Without adhering to shared metadata standards, datasets remain isolated within individual expeditions and research groups, limiting their broader value. Embedding robust and user-friendly metadata assembly and validation systems into expedition workflows is therefore essential.

The need to integrate data from multiple sensors and formats quickly at sea to support ongoing science and decision-making is another common issue. Data and metadata are frequently generated in proprietary or bespoke formats, creating barriers to data documentation, integration, and use. As a result, even basic products, such as linked navigation, observations, seafloor photos, and situational awareness data require specialized software, custom processing, or substantial technical expertise.

DATA VOLUME AND CURATION

The volume of high-resolution video and photos being collected presents a major curation and accessibility challenge. Managing these high-value data on personal hardware or within lab environments is no longer feasible. While repositories exist for final published products, few platforms support collaborative access for analysis and interpretation. Some programs, such as FathomNet (Katija et. al, 2022), have established platforms for depositing and accessing annotated underwater still imagery through a publicly available repository, but there is a growing need for accessible video infrastructure capable of supporting petabyte-scale datasets without reliance on individually managed physical hard drives.

HUMAN AND INSTITUTIONAL BARRIERS

Responsibility for at-sea data management is often diffuse and under-resourced, with scientific and operational priorities taking precedence over data stewardship. Although responsibilities are shared between scientists and marine technicians, real-time logging of observations frequently falls to the least experienced participants. Data management is commonly treated as a matter of individual effort or diligence rather than as a core objective. Although funding agencies often require data management plans (DMPs), these requirements frequently lack specificity, accountability, and enforcement. Even within a relatively closed system, where a single sponsor funds the ship, submersible, and science activities, adherence to data management policies is often inconsistent. Without clear requirements, defined responsibilities, and accountability, data management at sea remains underprioritized, leaving the broader data ecosystem fragile. Human and systemic barriers underlie these challenges, and without dedicated resources, they perpetuate a lack of standardization that limits data reuse and access. This, in turn, exacerbates technical challenges in data discovery and curation, creating a cycle of inefficiency that can only be addressed through coordinated, community-wide action. Workshop participants identified chronic underfunding as the underlying challenge, leaving insufficient time and resources to support effective stewardship. As marine data volumes and complexity grow, embedded data specialists are increasingly essential for managing at-sea workflows and delivering complete, validated data packages.

A COLLABORATIVE PATH FORWARD

Workshop participants identified foundational, technical, systemic, and cultural actions to strategically address challenges and build a more cohesive collaborative infrastructure that operates across the data life cycle, from acquisition to archive. These recommendations form a roadmap for a more efficient, accessible, and interoperable subsea data ecosystem.

FOUNDATIONAL: A COLLABORATIVE KNOWLEDGE HUB

Recurring needs described during the workshop included reliable ways to find information, enable discovery across distributed systems, and leverage existing tools. As a first step, participants proposed the creation of a centralized community-sourced knowledge base. Analogous to the Multibeam Community Wiki, this should be a simple, easily editable website that supports community contributions and updates. Because tools, workflows, and code evolve iteratively, the knowledge base would serve as a living resource that grows through community contributions, discussion, and requests for new tools and solutions.

The Subsea Data Collective was created during the workshop to serve as the desired central knowledge hub linked to distributed data resources, at-sea submersible operation guides, expedition planning tools, workflows, software libraries, and community-developed best practices. Built using GitHub, the site provides built-in support for version control, collaborative editing, wikis, and reusable code. Participants co-designed the initial structure and a table of contents for this knowledge base during the workshop. Recognizing that many relevant resources already exist or are under construction within the broader community, the wiki is intended to link to existing resources (e.g., COBRA, 2026), compiling a set of relevant practices, lessons learned, and recommendations that bridge multiple tools and organizations. It enables researchers, especially newcomers, to quickly locate relevant information without needing to navigate multiple systems and sites.

TECHNICAL: COMMON TOOLS AND PRODUCTS

With a central hub in place, the next step is to build the shared technical infrastructure it supports. To reduce the challenges of moving between vessels, workshop participants concluded that the community should work toward a common toolbox of software and data products that researchers can expect regardless of data acquisition platform. Many vessels already use open-source tools for data and event logging such as OpenRVDAS, OpenVDM, and Sealog (Pinner and Cohn, 2025), as well as CORIOLIX and Elog. Establishing a community of practice around the most commonly used at-sea data systems and tools will lead to the delivery of higher-quality data products that are more consistent and are easier to navigate and integrate.

In addition, there is an opportunity to create standardized "operational products" at sea by integrating and time-synchronizing datasets such as sensor data, navigation, and event logs at consistent intervals. Providing a clean, integrated baseline dataset would eliminate the need for science teams to recreate these products themselves while also supporting real-time decision-making and post-expedition analysis. Although custom tools exist within many research groups, consistent cross-platform practices are needed to enhance efficiency and minimize the data-wrangling burden on researchers. The Subsea Data Collective is an ideal venue for aggregating these workflows and tools. As multiple solutions are brought together, shared efficiencies will emerge, increasing the likelihood of developing common cross-platform approaches. Workshop participants identified an initial set of priority tools needed to address key gaps across the subsea data life cycle (Table 1).

SYSTEMIC: SHARED RESPONSIBILITY AND COMMUNITY-LED STANDARDS

A key outcome of the workshop was recognition that a sustainable data ecosystem requires a long-term shift in how data management responsibilities are defined and supported. This includes embedding dedicated data specialists at sea to ensure high-quality data capture, integration, and metadata completeness. These roles would improve efficiency at sea and accelerate delivery of data into open archives.

However, clearly defining roles alone is not enough. These responsibilities must be supported by planning frameworks and shared expectations. Although DMPs are widely required, they are

TABLE 1. Tools and applications identified by workshop participants, with associated requirements for improving at-sea data management.

FUNCTION	REQUIREMENT
Data Stitching	Simplify merging multiple data streams based on timestamp, with user control over variables and output frequency.
Header Mapping	Map variable column headers used across different ships and platforms to a standardized, controlled vocabulary.
Image Metadata Embedder	Automatically embed geospatial metadata (e.g., latitude, longitude, depth) directly into image file metadata using IPTC/EXIF standards, making images self-documenting.
Lightweight Visual Annotation	Simplify image tagging and annotation without renaming original files to support downstream machine learning workflows.
Operational Products	Generate near-real-time post-dive maps showing vehicle tracks and basic sensor data (e.g., temperature, salinity) with outputs in common formats (KML, GeoJSON, CSV).
Subsampling	Downsample high-frequency submersible sensor data (e.g., CTD) to usable frequencies (e.g., 1 Hz).
Optimized Event Logging	Streamline event logging to reduce human error during at-sea operations (e.g., voice to text).

often difficult to operationalize in practice. Accessible templates and plain-language guidance are needed to make them easier to create, understand, and implement. DMPs must also be enforceable and regularly updated to reflect evolving data acquisition practices and stewardship plans.

Workshop participants also proposed a complementary data processing plan (DPP) to fill a gap in planning for post-collection workflows. Unlike a DMP, which focuses on planning for data collection, the DPP would explicitly address the handling, processing, integration, and delivery of data after acquisition.

Key components of a DPP include:

- PROCESSING STEPS – A clear, sequential outline of how data will be processed following acquisition
- HANDLING PROCEDURES – Detailed protocols for managing data throughout each stage of processing
- DEFINED RESPONSIBILITIES – Explicit identification of roles and responsibilities across the data processing workflow
- TOOLS AND METHODS – Specification of the software, tools, and methodologies used to transform raw data into final products
- DATA PRESERVATION – Dedicated planning for preparing and delivering data to repositories in accordance with FAIR principles, including validating metadata completeness and quality prior to submission

These components will prevent data from languishing on personal hard drives or within institutional silos and will facilitate the migration of data products into open access data repositories.

Participants emphasized that DPPs should address community-defined expectations with consensus-driven guidance—cognizant of the needs of scientists and downstream data systems. These resources would promote standardization through shared accountability while also serving as practical guides for new researchers, clarifying expectations and requirements. Implementation will require dedicated funding to support personnel, technical infrastructure, and tools needed for execution.

CULTURAL: A CULTURE OF DATA STEWARDSHIP THROUGH TRAINING

Finally, sustaining these technical and systemic solutions requires investment in the people who use them and the practitioners who support the underlying infrastructure. Workshop participants identified a critical need to improve data management literacy, so researchers are better prepared to work with data effectively at sea and prepare it for open archives. Proposed solutions include developing a microcredentialing program or a series of pre-expedition training courses, with initial content drawn from data managers and research groups that have robust workflows. By equipping scientists with foundational skills and best practices *prior to deployment,* the community can foster a more resilient and consistent culture of data stewardship from the ground up.

A FRAMEWORK FOR INTERVENTION

To effectively implement these solutions, we situate the subsea data workflow (Figure 1) within the general scientific data life cycle (Figure 2). This model identifies where targeted intervention, standardization, and collaboration can have the greatest impact. Based on the workshop discussions, this framework outlines key phases: acquisition, near-real-time products, expedition distribution, and long-term stewardship, where the challenges and opportunities identified by participants converge.

The acquisition stage is the most consequential opportunity to prevent downstream failure. Lack of metadata at the point of acquisition is rarely recoverable, making early-stage standardization essential. Establishing consistent baseline fields for event logging and enforcing core metadata capture and validation ensures that datasets are properly documented from the outset, preventing cascading issues later in the workflow. Near-real-time products represent a key opportunity to integrate disparate data streams into unified, time-synchronized operational products. Providing this capability as a standard service reduces the burden on individual science teams and enhances real-time decision-making, allowing researchers to fully leverage the data they are collecting while still

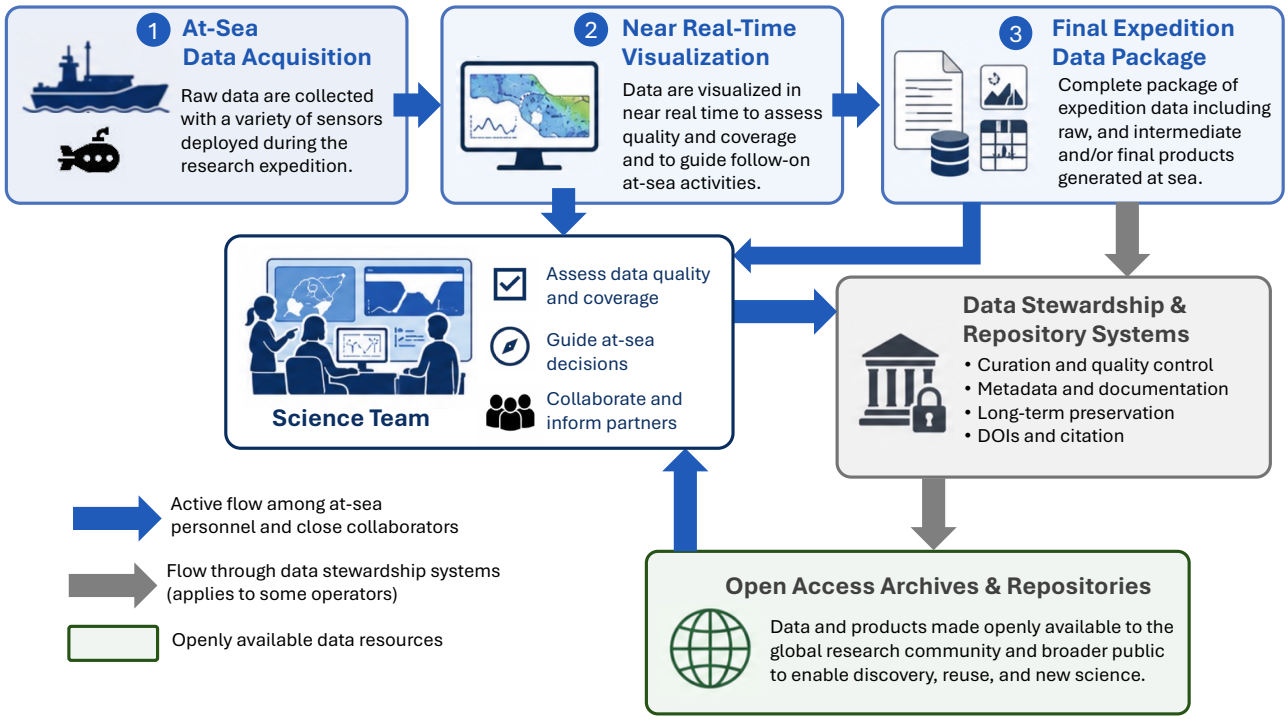

FIGURE 2. Scientific data life cycle from acquisition to distribution. Data flow from (1) at-sea acquisition through (2) near-real-time visualization, and (3) final expedition data distribution. Blue arrows indicate active data exchange among at-sea personnel and close collaborators. Gray arrows indicate flow through data stewardship programs and repository systems used by some operators. The green box represents openly available data resources populated by both platform operators and individual scientists. Boxes highlight stages for intervention, while arrows indicate opportunities for improving data flow across the life cycle.

at sea. A major friction point occurs at the conclusion of an expedition, when ambiguity in responsibility often leads to breakdowns in data handoff. Clear data management plans, standardized delivery formats, and agreed-upon protocols function not as administrative overhead but as essential agreements that ensure data are preserved, accessible, and not lost within institutional gaps or confined to local storage.

THE PATH FORWARD: A CALL TO COLLABORATIVE ACTION

The marine science community stands at a pivotal moment. The rapid expansion of data presents an unprecedented opportunity and a significant challenge. Unlocking the full value of data will only come from a shared, community-wide commitment to collaboration, standardization, and mutual support. Friction points that currently limit progress can be addressed if we work collectively to build a more unified and efficient data ecosystem. The workshop concluded with a clear consensus on three key activities that will translate this vision into practice.

- **ESTABLISH A COMMUNITY-SOURCED KNOWLEDGE HUB**
  Such a publicly accessible, easily updatable resource will serve as a central directory for data tools, community standards, repository links, and best practices that will lower barriers to entry and improve discoverability.

- **SUSTAIN AND EXPAND COMMUNITY ENGAGEMENT**
  Leverage key scientific forums (e.g., Haskins et al., 2025) and future workshops to build consensus and incorporate diverse perspectives. Solutions related to growing volumes of underwater photos and video, for example, will require further community input and consensus on standards and specifications.

- **ADVANCE COLLABORATIVE DEVELOPMENT EFFORTS**
  Convene community hackathons and coordinated, multi-institution projects to develop the standardized tools and operational products identified during the workshop.

The ultimate goal is to democratize access to data and tools while ensuring that new users are not disadvantaged by a lack of familiarity. This approach is anticipated to enable a more efficient, integrated, and impactful future for ocean science, powered by data that are truly FAIR. By working together, the community can improve efficiency and outcomes of at-sea operations and accelerate the delivery of more insight from subsea data.

REFERENCES

COBRA. 2026. Crustal Ocean Biosphere Research Accelerator (COBRA) Early Career Deep-Sea Expedition Leadership Master Class and Fellowship. Bigelow Laboratory for Ocean Sciences, https://cobra.bigelow.org/master-class-and-fellowship/.

Ferrini, V.L., J.J. Morton, and A.M. Goodwillie. 2025a. Integrating the deep: MGDS services for access, discovery, and integration of submersible data. Poster presented at the fall 2025 meeting of the America Geophysical Union, New Orleans, LA, OS51C-0428.

Ferrini, V.L., J.J. Morton, and A.M. Goodwillie. 2025b. GMRT: A curated bathymetric synthesis for scalable, data-driven marine geoscience. Paper presented at the fall 2025 meeting of the America Geophysical Union, 2025, New Orleans, LA, OS24B-01.

Haskins, C., V.L. Ferrini, and C. Bassin. 2025. Strengthening the data lifecycle at sea: Community strategies for useable, shareable, and impactful submersible data. Poster presented at the fall 2025 meeting of the American Geophysical Union, New Orleans, LA, OS51C-0429, https://doi.org/10.22541/essoar.177160689.95323520/v1.

Katija, K., E.C. Orenstein, B.M. Schlining, K. Zurow, L.A. Kuhnz, H.A. Ruhl, and K. Barnard. 2022. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports* 12:15914, https://doi.org/10.1038/s41598-022-19939-2.

Pinner, J.W., and D.P. Cohn. 2025. *OpenRVDAS, OpenVDM, and Sealog* [Source code]. Ocean Data Tools, https://www.oceandatatools.org.

Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, and others. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018, https://doi.org/10.1038/sdata.2016.18.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Schmidt Ocean Institute for funding the workshop and providing logistical support, as well as the workshop participants whose time, ideas, and experience shaped the discussions and outcomes. NotebookLM and ChatGPT were used to assist with preliminary synthesis of workshop transcripts and notes, editorial revision for brevity and clarity, and refinement of illustrative mockups. The authors reviewed and verified all outputs and take full responsibility for content.

AUTHORS

Vicki Ferrini (ferrini@ldeo.columbia.edu), Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA. **Christina Haskins**, Woods Hole Oceanographic Institution, Woods Hole, MA, USA. **Corinne Bassin**, Schmidt Ocean Institute, Palo Alto, CA, USA.

ARTICLE CITATION

Ferrini, V., C. Haskins, and C. Bassin. 2026. Navigating the data deluge: A collaborative vision for a unified subsea data ecosystem. *Oceanography* 39(3), https://doi.org/10.5670/oceanog.2026.e303.

COPYRIGHT & USAGE