Prepared for the 2022 National Ocean Exploration Forum

Ocean Exploration in a Data-Rich World

By Dr. Vicki Ferrini (Lamont-Doherty Earth Observatory) and Colleen Peters (Saildrone)

Vision Statement

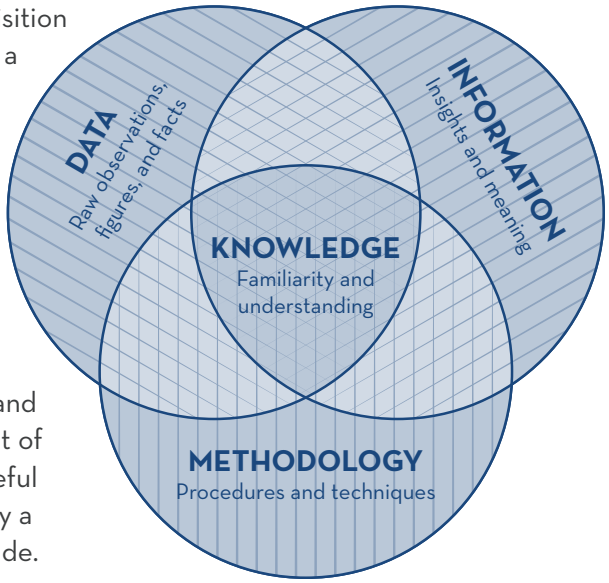
By 2032, ocean exploration will be enabled by a thriving community of practice that fosters networks of expertise for data gathering, management, access, and use.

Leading the Pack

The U.S. exclusive economic zone (EEZ) is the largest in the world, covering 3.4 million square nautical miles of ocean — an area larger than the combined land area of all 50 states.[1] The U.S. ocean exploration community is poised to explore, map, and characterize this massive territory armed with new technologies, in response to pressing societal needs, including the emerging blue economy, the quest for offshore resources, and a planet whose climate is changing.[2] Successfully and efficiently navigating this new era in ocean exploration, and the opportunities that doing so brings for discovery and direct benefit to the American people, relies on advancing our practices at the intersection of ocean data and communities of stakeholders. Data must be at the core of building a truly convergent community of practice that is focused on, and benefits from, ocean exploration. Collaborations of active and experienced organizations that seek efficiencies, support mutually beneficial goals, and set the stage towards building a proficient and efficient workforce will be the foundation to execute long-term ocean exploration strategies and goals successfully.

Advances in technology allow us to acquire larger volumes of higher quality data more rapidly, creating both opportunities and challenges. Recognizing ways that we can remove bottlenecks to data sharing and barriers to data access will accelerate our use of data for basic discovery, provide training opportunities for the next generation of STEM professionals, and allow broader engagement of the American public in ocean exploration.

With increasing numbers of research and exploration vessels, as well as autonomous, uncrewed, and remotely operated platforms operated by government, academia, philanthropic organizations, and industry, there is tremendous potential to coordinate practices around the acquisition, storage, and dissemination of ocean data to ensure that expensive, unique observations can benefit broad and diverse communities of stakeholders. As more data acquisition platforms come online, and the community of contributors to a national ocean exploration program extends and diversifies, there is a growing need to expand practices to ensure integration and sharing of methodology, information, and knowledge.



Significant progress has been made with respect to data and metadata standards, data sharing, and growth of national data archives, but sharing is not enough — transformation of data into useful, relevant, and understandable information is a key target. Many challenges and bottlenecks that currently inhibit the full return on investment of ocean data resources, and the transformation of data into useful information that can be accessed, queried, and understood by a broad variety of users, have barely advanced in the past decade.

1 *www.gc.noaa.gov/documents/2011/012711_gcil_maritime_eez_map.pdf*
2 *oeab.noaa.gov/wp-content/uploads/2020/12/2020-12-17_oeab_advice.pdf*

On Mission: Data Acquisition and Quality Standards

National resources that are central to ocean exploration already include data sourced from some academic, government, and industry contributors. Despite several successfully implemented data and metadata standards,[3,4,5] each contributing organization, individual, and/or platform operator is effectively, at present, a silo. Data quality standards remain variable, both within and between organizations, as do levels of documentation provided with data acquired on a given acquisition mission, which can result in inefficiency with respect to data transfer and consequent increases in the costs of data management, archival, and dissemination.

Expecting all data acquirers to use the same tools and workflows for data acquisition, annotation, and processing is unrealistic and would be largely unachievable. Each group has its own data acquisition and management tools, protocols, and practices, based on its operational history and needs, as well as staff, budget, knowledge, and priorities. Various platforms use commercial or in-house software suites to provide an efficient workflow, while others may be limited to open-source or legacy tools due to financial constraints or staffing limitations. Therefore, imposing requirements on platforms to use a particular software, method, or technology is impractical. What is feasible, however, is the documentation of practical information, such as metadata that fully describes the contents of a final data distribution. Regardless of what software or method is used, adopting a common approach and format will enable human-readable dashboards reporting on data acquired to require less human intervention and thus make data parsing and access more efficient. An example of proposed metadata that would meet this need is described in Appendix A of the *Draft OECI Data Management Plan (v1.0.0)*[6].

Post Mission: Bridging the Gap between Acquisition and Archives

Thanks to consistent internet access, web services, and social media, it is significantly easier for the ocean exploration community, and ultimately the general public, to follow real-time expeditions on a wide variety of platforms, including, but not limited to, oceanographic vessels and autonomous surface vehicles.[7,8] Posts and updates about current operations, live-streamed video, active ship tracks, and live connections provide an on-mission camaraderie and public engagement.[9] However, continuity of engagement at the conclusion of an expedition, as data are processed, integrated, and prepared for public access, can prevent disconnection, duplication of effort, and acquisition inefficiencies. Providing rapid access to usable information about the existence of data and its evolution, along a continuum from acquisition to archive, will ensure that data delivers maximum value to the broadest community of stakeholders to maintain engagement and enable vessels of opportunity to fill critical data gaps. A short-term strategy for openly sharing select preliminary data products and practical metadata across all acquisition platforms does not require that all data resources are consolidated in a single repository, but instead envisions use-case-driven information and web service standards based on the specified operational needs of ocean explorers. This will help optimize new data acquisition, aid in sharing information about the prioritization of areas for further exploration, and offer new opportunities for more efficiently exploring, mapping, and characterizing the nation's undersea territory.

3 *www.ncei.noaa.gov/resources/metadata/*
4 *www.rvdata.us/about/technical-details/metadata/*
5 *www.oceanbestpractices.org/repository/*
6 *web.uri.edu/oeci/files/2021-07-21_OECI_DMUWG_DMP-Plan-v1.0.0.docx.pdf*
7 *nautiluslive.org/blog/2021/06/14/mapping-great-lakes-thunder-bay-national-marine-sanctuary/*
8 *sanctuaries.noaa.gov/live/2020/whoi.html*
9 *deepoceaneducation.org/explore-resources/*

While well-established outreach activities that engage the public in ocean exploration have often focused on at-sea activities, there are varied possibilities for ongoing shoreside engagement, including citizen science, that have barely been realized.

With the right tools, communication, and coordination mechanisms, a broad, distributed, and diverse community of stakeholders can be engaged in the post-expedition phases of exploration — data processing and interpretation. These activities can serve as a bridge for students in STEM to careers in ocean exploration and will help to diversify and extend public participation nationally. The coordination of a distributed community of students, faculty, and researchers, all working on specific data types, will help to accelerate the creation and dissemination of data products by harnessing the power of the crowd.

Post Mission: Long-term Data Accessibility and Discovery

In an evolving data landscape, with a growing number of data acquisition platforms operated by an increasingly diverse and distributed community, there is a rapid increase not only in data volumes but also in data management skills, standards, support infrastructure, and business models, adding increasing complexity to an already complicated data problem. Understanding variable data directory structures, naming conventions, and formats takes time and represents a complex bottleneck to serving the data quickly. While standard directory structures and naming conventions are not easily achievable across any diverse community, standards with respect to how data distributions are described can be achieved. This would help sea-going scientists and technicians more easily navigate data and lower the costs and effort associated with broadening access to data from multiple providers.

The commercial geophysical survey industry, for example, stewards massive repositories of data that are difficult to make available due to contractual restrictions and the business implications of sharing data funded by a private entity and otherwise viewed as intellectual property.[10] Small site surveys conducted by colleges and universities or by state regulators may be online, but can be difficult to discover and integrate. Uniting such disparate data resources will be beneficial but requires significant effort.

Data attribution, to acknowledge efforts from acquisition to product creation, is critical for providing credit for both data acquisition and processing efforts and is necessary for building trust. As more data become available and more individuals engage in the creation of data products, attribution becomes more complex and necessary. An example would be raw swath data acquired by a ship in the U.S. academic fleet that is sent to NOAA's National Centers for Environmental Information (NCEI) by *Rolling Deck to Repository (R2R)*[11] and then processed and redistributed as processed swath and grids by an individual scientist or the *Global Multi-Resolution Topography Data Synthesis (GMRT)*[12] project. This data stewardship hand-off includes multiple institutions, projects, and individuals, and all should be acknowledged accordingly. Accurate attribution will further strengthen the willingness to participate in making data accessible and highlight the achievements and collaborations within the broader community. In addition, it is often the most direct method of requesting data that may not otherwise be publicly available.

10 *www.frontiersin.org/articles/10.3389/fmars.2019.00283/full/*
11 *www.rvdata.us/*
12 *www.gmrt.org/*

Blueprint for 2032: Building a Convergent Community of Practice

Exploration is the act of searching for the purpose of discovery of information or resources. To discover means to find, to explore means to search.

Data are a strategic asset for our nation. Ensuring that related ocean data resources are preserved, accessible, and usable establishes a strong foundation that can support exploration, discovery, and STEM training opportunities. Uniting the data not only preserves them but enables new and innovative analyses and discovery through the exploration of data. These national resources also offer material that can be used to help build a diverse and highly skilled American workforce that is ready to leverage the opportunities of the emerging blue economy. It is critical that data are not only made available for scientific research, education, and industry purposes, but must also be assembled in ways that they can be used for decision making and accountability for local, state, and federal entities.[13] Looking ahead, a world where the maximum number of people across the planet have access to and curiosity for ocean data will cultivate environmentally conscious generations of ocean stewards.

Managing increasing volumes of oceanographic data is time consuming and costly, and is historically underfunded, but is necessary to enable next-generation analysis and visualization.[14] Data management and integration efforts should be fully outlined in advance so they can be more easily integrated in the planning and budgeting of ocean exploration expeditions. Post-acquisition data processing and stewardship is more efficient when data acquisition activities are cognizant of post-cruise data pipelines. Well-documented workflows enable repeatability, increase return on investment, and provide a mechanism for the community to share ideas and lessons learned. This will ensure that the most effective tools and solutions are utilized to streamline efforts and that development of new tools is focused on features that don't yet exist. Coordinating methods and tools will benefit the ocean exploration community by providing a more cost-effective approach to managing distributed data resources.

Before the work and efforts of ocean explorers, engineers, and researchers across the nation and the world can be integrated and leveraged, the U.S. ocean exploration community first needs to get its house in order, ensuring that vessel operators and data centers are adequately prepared and resourced for increases in data volumes and data-informed decisions that lay ahead. In five years, there should be comprehensive data products that accurately describe the status of ocean exploration and characterization throughout the U.S. EEZ (and beyond) that is inclusive of data from stakeholders across all sectors. In 10 years, these data should routinely support training and development of the blue economy workforce as well as new business models that support the transformation of data to actionable information that can inform decisions that benefit the nation. The technologies and impetus to share and integrate is upon us; challenged to advance toward a new, open, and more inclusive model for ocean exploration and discovery that transcends sectors and stovepipes and results in a thriving community of practice.

13 *oeab.noaa.gov/wp-content/uploads/2020/Documents/FY-21-RD-Budget-Priorities.pdf*
14 *oceanexplorer.noaa.gov/national-forum/media/noef-2017-report.pdf*

CHALLENGES	5-YEAR OPPORTUNITIES	10-YEAR OPPORTUNITIES
Differing motivations for data acquisition resulting in variable data quality standards and data products.	Communicating and coordinating needs through workshops and annual planning discussions.	Developing standards and guidelines to capitalize on opportunities for data acquisition.
Variable data management resources such as staffing, hardware, workflows, and budget.	Communicating knowledge and experience, sharing existing efforts with tools, expertise, and solutions that can be leveraged with new partners.	Streamlining processes to make data documentation, management, and submissions more efficient.
Variable and complex post-mission data distribution formats, quality standards, and naming conventions.	Encouraging new perspectives, better use of existing tools, new tools, and knowledge sharing.	Implementing community-wide shared/common best practices, training, and tools for data management and interpretation.
Lack of consistency in transforming data into fit-for-purpose, easily integrable data products.	Developing community-wide shared/common methods and tools for data management, quality assessment, and interpretation.	Developing transformational best practices, used both nationally and (eventually) globally.
Variable data release timelines obscure knowledge of existing data which can yield redundant data acquisition effort.	Sharing basic information about the existence of data to eliminate redundant effort and provide new opportunities for collaboration.	Extending, embellishing, and integrating web services to ensure that users can easily find out what data exists, irrespective of where it is held.
Gaps in connectivity/discoverability between online data resources managed through different systems/repositories.	Enhancing web services and interoperability, promoting comprehensive inventories, and improving methodologies regarding how data distributions are described.	Aggregating information about data available through a variety of sources into a data discovery portal that will lower the barriers for users and provide a more holistic perspective.
Lack of recognition of effort or correct attribution of data, eroding trust among alliances and impeding data sharing.	Supporting all the varying U.S. outfits that help build this catalog of data with a more all-encompassing concept of 'exploration.'	Developing tools and best practices for data provenance and attribution. Routinely acknowledging both acquisition and product generation.
Instilling data management, data processing, and data science skills in the emerging workforce to leverage the opportunities of the growing blue economy.	Evolve into an equitable and inclusive community that is not reliant on personal connections for access to data and information.	Creating training programs focused on data management, including degree programs and certifications.