OBIS Infrastructure, Lessons Learned, and Vision for the Future

Eduardo Klein[1], Ward Appeltans[2]*, Pieter Provoost[2], Hanieh Saeedi[3,4], Abigail Benson[5], Lenore Bajona[6], Ana Carolina Peralta[1] and R. Sky Bristol[5]

[1] Departamento de Estudios Ambientales, Universidad Simón Bolívar, Caracas, Venezuela, [2] IOC Project Office for IODE, Intergovernmental Oceanographic Commission, UNESCO, Ostend, Belgium, [3] Senckenberg Research Institute and Natural History Museum, Frankfurt, Germany, [4] Institute for Ecology, Evolution and Diversity, Goethe University Frankfurt, Frankfurt, Germany, [5] United States Geological Survey, Lakewood, CO, United States, [6] Ocean Tracking Network, Dalhousie University, Halifax, NS, Canada

This mini-review paper analyses the achievements of the Ocean Biogeographic Information System (OBIS), as a distributed global data system and as a community of data contributors and users. We highlight some issues and challenges and identify ways OBIS is trying to address these with developing community standards, protocols and best practices, applying new innovative technologies, improving human capacity through training, and establishing beneficial partnerships. With the release of the second generation of OBIS (OBIS 2.0), we now have a more solid foundation to build improved data processing/integration workflows, new data synthesis routines that add value to OBIS data, and new types of products and applications for scientific and decision-making. The future of OBIS will be in working toward an open and inviting process of co-developing OBIS as a global networked open-source data system that will enable the community to organize, document, and contribute analytical codes that interface directly with OBIS, provide analyses, and share results. The main challenges will be in mobilizing and organizing the scientific community to publish richer and high quality data more rapidly in support of developing robust and timely indicators of status and change on Essential Ocean Variables and Essential Biodiversity Variables.

Keywords: ocean biodiversity, biogeography, research infrastructure, open-access, data and information, science-policy

INTRODUCTION

Playing a central role in fostering data sharing of marine species observation data since 1999, the Ocean Biogeographic Information System (OBIS) has built the world's most comprehensive database on the diversity, distribution, and abundance of life in the ocean. The OBIS Network is made up of thousands of scientists and data managers employed by hundreds of institutions around the world, who ensure that scientifically researched, collated and published data adhere to FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al., 2016). OBIS has built a platform for robust near real-time data integration and curation. It also provides powerful data access and analytical services that streamline the contribution of integrated, quality-controlled datasets into models and forecasts. OBIS has pioneered a solution for managing combined biological and environmental data, including details about sampling effort and methods

(De Pooter et al., 2017). OBIS is now extending beyond species occurrence data, embracing ecosystem Essential Ocean Variables (EOV) in support of the Global Ocean Observing System (GOOS) and the Marine Biodiversity Observation Network (MBON). The identification of the taxonomic components of the EOVs, and the capacity to assimilate and integrate non-taxonomic entities into the global database (i.e., community types like macroalgae) will contribute to their efforts to build a sustained, globally coordinated observing system on the status and trends of marine biodiversity and habitats (Benson et al., 2018).

The world agreed on the Aichi biodiversity targets (2011–2020) to better understand and predict biodiversity dynamics, that is how biological diversity underpins ecosystem function and how the provision of ecosystem services are essential for human well-being. With the recent announcement by the Intergovernmental Panel on Climate Change (IPCC) indicating that climate change must be addressed by 2030 to avert major catastrophic changes to global and marine ecosystems, it is clear that time is limited to more adequately understand and protect marine biodiversity.

The success of bringing millions of marine species observations into the public domain is a major achievement. Through FAIR access to data, OBIS provides equitable access and benefits to research, biodiversity conservation management and policy making, and also enhances international collaboration, for which OBIS is recognized by many global organizations including the United Nations General Assembly. OBIS is requested to support several international processes such as those under the UN Regular Process (World Ocean Assessment), the Convention on Biological Diversity (CBD) and the Intergovernmental science-policy Platform on Biodiversity and Ecosystem Services (IPBES). OBIS will have an active role providing open and timely integrated biodiversity and habitat records for the United Nations Educational, Scientific and Cultural Organization (2017), that will be crucial to its success.

CHALLENGES AND OPPORTUNITIES, A VISION FOR 2030

After two decades of OBIS we can reflect on its achievements, where it performed well and where it needs continuous improvement. We can look ahead and identify challenges and opportunities for OBIS to respond to new demands for ocean data and information services. We plan to develop OBIS as a critical component in accelerating the pace of scientific exploration and discovery in response to the urgent needs imposed by our changing planet.

One critical aspect of OBIS is that despite two decades, including the Census of Marine Life, and integrating over 50 million records, widespread gaps in taxonomy, space, and time still remain (Müller-Karger et al., 2018a). This reflects the bias of sampling effort worldwide, but still half of the records added to OBIS after 2015 date before 2000, illustrating significant effort spent on data archeology – digitizing data from

before OBIS existed (Müller-Karger et al., 2018b). Establishing robust baselines against which current and future change can be detected is important but detecting the most recent changes using OBIS is currently difficult due to the typical delay in processing and reporting recent observations (<5 years) as shown by an overall decline in data in OBIS after 2010 (Müller-Karger et al., 2018b). It is vital to accelerate data availability through OBIS by building direct connections with marine biodiversity monitoring networks and other near real-time observation activities.

The scientific publishing track record of OBIS is growing with over 100 scientific papers per year referring to OBIS. Three chapters of the first United Nations World Ocean Assessment (The Group of Experts of the Regular Process, 2017) used data from OBIS, and several recent high-impact papers provide new insights in global marine species richness patterns (Chaudhary et al., 2016, 2017; Saeedi et al., 2017; Menegotto and Rangel, 2018), biogeographical classifications (Costello et al., 2017; Sutton et al., 2017), and projected biodiversity change linked to climate (Dornelas et al., 2014; Bowler et al., 2017; Griffiths et al., 2017).

State of the Framework

Due to the different nature and interests of the potential users, the value proposition from OBIS could vary, from data producers (scientific institutions) to data analysts (research scientists) to information product end users (policy makers and managers). We work to characterize and understand our global customer base through system monitoring, user engagement, and regular independent review, developing plans for action through the OBIS Steering Group. The following sections present the principles of the core design of OBIS.

International Data Management Best Practices

The breadth of the OBIS Network across such a wide swath of the marine biodiversity science community has been a core driver in adoption of and contribution to international standards and practices in research data management. The FAIR Principles represent a culmination of many years of work to suggest a core set of practices that OBIS ascribes to and has been implementing for some time. The Group on Earth Observations, an organization with which OBIS is closely aligned, adopted 10 Open Data Principles to promote and to encourage best practices and maximize the potential for appropriate re-use and combination with other data sources.

Following these and other relevant guidelines and best practices, OBIS provides the following core values:

- Data standards developed through the Biodiversity Information Standards (TDWG) body that help data producers decide on data attributes and parameters for their data collection efforts.
- A data repository framework consisting of distributed nodes and a central hub to submit data, cite it, and advertise it for use, achieving compliance with national or institutional requirements for releasing open data.

- Support and training in data management techniques, methods, and tools developed by the network for data management and analysis.
- A thriving research community developing biodiversity data analysis techniques.

Evolving Data Types

New and innovative methods for observing marine biodiversity such as the Imaging FlowCytobot (Olson and Sosik, 2007) and genomics techniques for assessing biodiversity (Bourlat et al., 2013) are currently being used to increase the observation and measurement coverage of the marine environment. Where possible, OBIS seeks to develop alliances leading to intelligent data interchange methods such as a developing partnership with the Global Coral Reef Monitoring Network. OBIS serves as the nexus where these new methodologies can integrate their data in a rich and standard format, and the OBIS community can help provide mapping across diverse variables to calculate metrics and produce biodiversity indicators.

Ocean Biogeographic Information System has provided leadership in this area through enhancements to the Darwin Core Standard to allow documentation of qualitative or quantitative information about both the sampling event and the species observation within that sampling event (De Pooter et al., 2017). This work has broadened the capacity of integrating information from a variety of sampling methods and instruments, enabling OBIS to leverage richer datasets containing other environmental observations and measurements. Previously, when measurements such as water temperature, salinity, and wind speed were collected in association with species observations, those data were difficult to link together. Now, it is possible to ingrate these datasets syntactically and enhance the connection between them through focused attention on the development and implementation of robust vocabularies. This new schema will provide the ability to understand the context of the origin of the data and explore the relationship of the biodiversity records with accompanying environmental variables.

Using this new schema, OBIS nodes are able to document data that were previously difficult to integrate. For example, animal tracking and telemetry data managers are using the new OBIS schema on machine recorded observations and have developed and documented requirements specific to this type of data using the new schema. Similarly, data from FlowCytobots are being assessed and documented. Current research and development efforts are beginning on the most effective ways to encode observations and measurements where the focal entities are operational taxonomic units or habitats made up of groups of organisms rather than a single, identifiable taxon.

Technical Infrastructure

The OBIS infrastructure is fundamentally distributed in that it consists of source data repositories at OBIS Nodes where core data management is conducted. OBIS Node data is harvested into a central index where final data assessment processes are conducted, and third-party information is integrated. OBIS

follows an "API-first mindset," meaning that everything starts with a robust Application Programming Interface (API) that then serves the needs of a web portal, mapping tool, and open source/free scientific programming packages (R and Python). The power of this approach is in its ability to mobilize and enable the broader global community in building their own tools and capabilities on an open framework.

Recent developments and the release of OBIS 2.0 enables the following major capabilities:

- Near real-time data integration from OBIS Nodes – as soon as OBIS Nodes advertise the availability of new or updated data, the OBIS system begins automated processing to ingest data and completes integration routines in a timely manner.
- The ability to scale the system to hundreds or thousands of new datasets and millions of new records while providing timely results for queries, data access (download and streaming), and products (analytics and indexes).
- The ability to integrate environmental variables in addition to biological observations and to fully leverage this new type of data in queries, data access, and products.
- Data management support is provided at a foundational level for OBIS Node Managers through the obistools R Package (Bosch et al., 2018).
- Improvements on the real-time analytics capability of OBIS to directly convert raw data into EOV and into values of biodiversity indices and indicators, starting at the API level to support science end-users through the rOBIS package (Provoost and Bosch, 2018) and custom apps/portals along with a flexible and ever evolving set of reports through the OBIS portal.
- Improvements on the full visibility of the OBIS Network by persistently linking OBIS Nodes, data provider institutions, and individual contributors (data providers, authors, etc.) together such that each entity is cited and their full contribution to OBIS easily accredited.
- New portal and map explorer capabilities built to scale as data volumes continue to expand.

Continuous Data Assessment and Enhancement

Data generated through so many different campaigns and data collection efforts through time and changing practices are inevitably of varying overall quality and completeness. Part of the role of OBIS is to continually assess inbound data for alignment with standards and compliance to particular conventions used within the community to aid users in data selection and analytical uses (Vandepitte et al., 2015). OBIS performs a battery of tests on new and updated data, records metrics from these processes, and uses the values for intelligent filters in the portal, API, and other interfaces. The adopted approach leads toward evaluating data as to fitness for purpose, keeping track and exposing all QC tests performed on every record, rather than making judgments about quality.

Part of the real power in integrating hundreds of disparate datasets with a common standard is an ability to reach out

to related information systems linkable because of the points of integration in areas such as taxonomy, space, time, and other variables. These variables allow OBIS to write intelligent information linking processes that pull in useful information from third party sources from the World Register of Marine Species to regulatory context and information about marine regions. This information is linked and incorporated into the OBIS system in real time and continuously as data are added and updated.

Capacity Building

The organizational structure of OBIS and its placement within UNESCO's IOC provides a mandate and institutional framework for continual capacity building as a core function. More effort and as many resources as possible are put into developing the core capacity of the OBIS Nodes than building the central infrastructure. The OBIS secretariat at the IOC Project Office for

IODE in Belgium allies closely with the Flanders Marine Institute and other scientific institutions in leveraging core technical capabilities such as central source data hosting, provided free of charge to OBIS Nodes who do not have these capabilities in their host institutions. Close alignment with the OceanTeacher Global Academy (OTGA) provides advantages in developing and providing training courses in data management, data analysis, decision support, and other areas of importance to OBIS.

Ocean Biogeographic Information System Nodes provide training courses in English, French and Spanish at the OTGA regional training centers in Belgium, Colombia, Kenya, Malaysia, and Senegal, with course materials available online for use under CC-BY-NC-SA license terms. Training has focused on working with students who have a basic knowledge of data management practices and tools from basic spreadsheet methods to R scripting such that skills can be immediately applied to the most critical areas of learning this particular
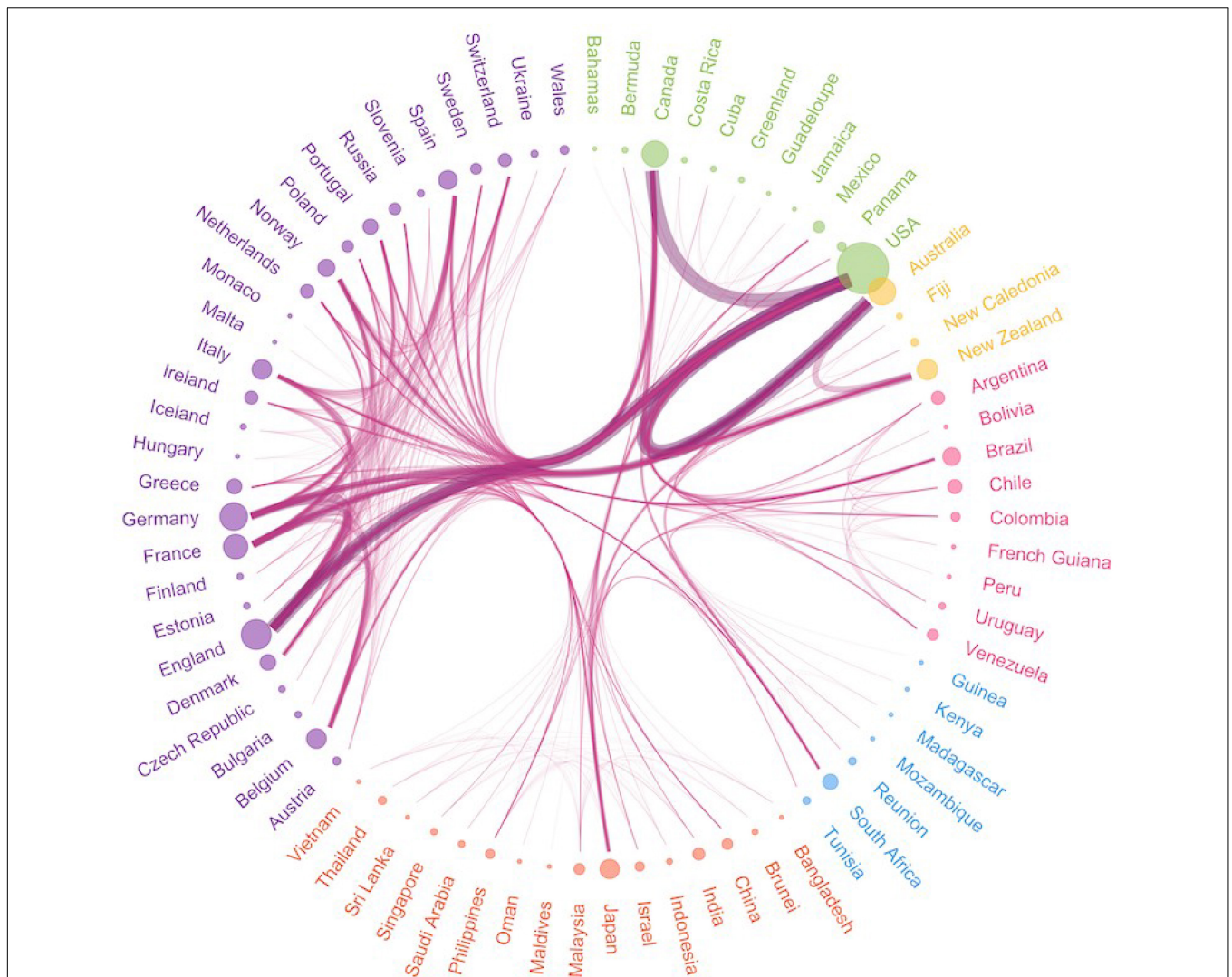
FIGURE 1 | Based on a bibliographic study in collaboration with the library of the Flanders Marine Institute, 2700 scientists from 73 countries collaborated on >1000 papers citing OBIS. This figure shows the connections between the countries in co-authoring these papers.

domain. Course modules cover both data management and the basics of data analysis such that trainees can either use the skills on their own or assist institutional research scientists and other users in working with the data in scientific analysis.

We see multiplication of training efforts catalyzed through the OceanTeacher online Moodle platform by various institutions linked to the national OBIS nodes which in the last 2 years organized OBIS training courses in Chile, Ecuador, Brazil, Germany, Kenya, Mexico, Russia, Iran, and the United States. These efforts represent direct return on investment to the OTGA and its funders, increasing the distribution and development of knowledge on marine data management and use worldwide.
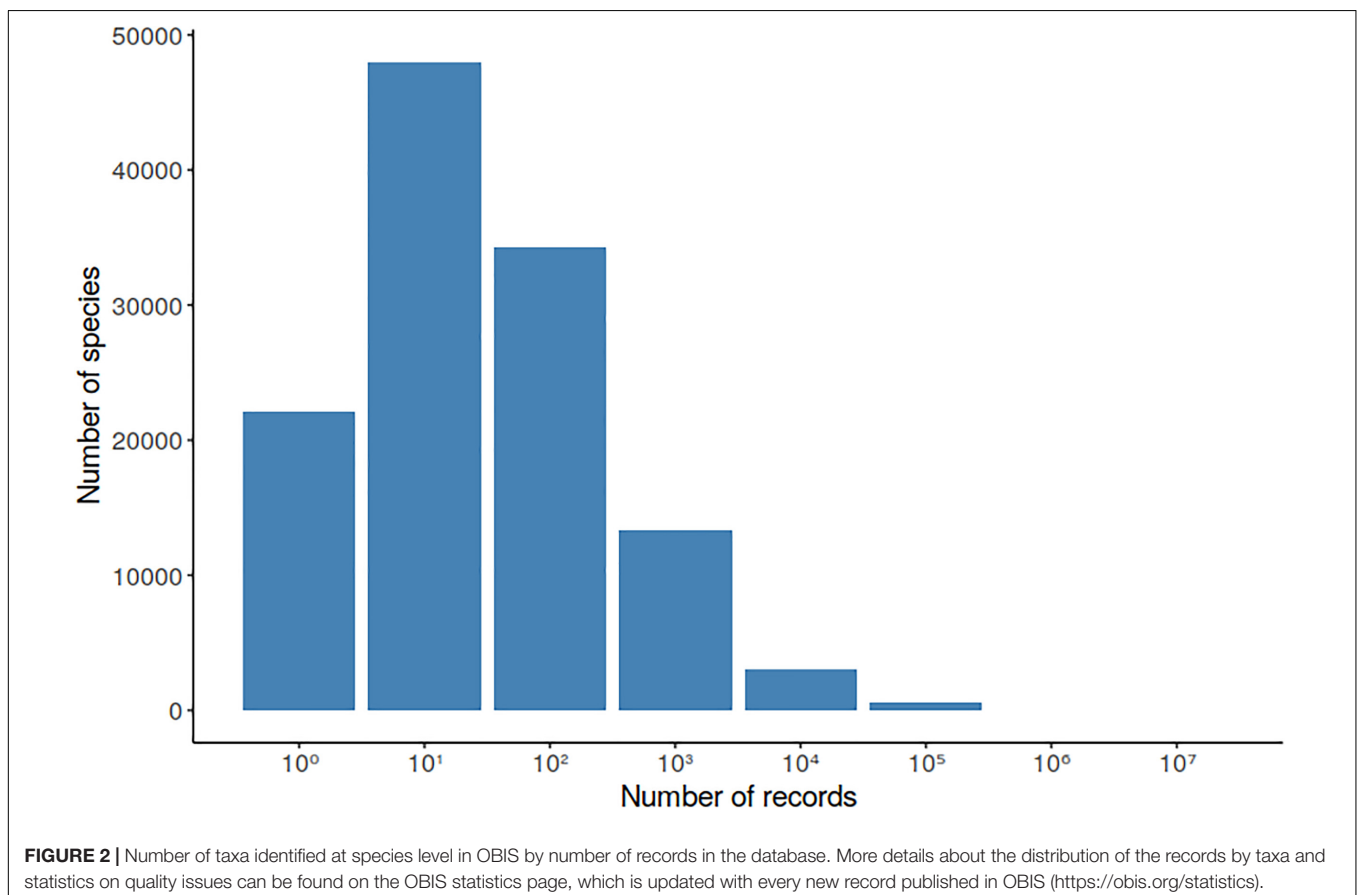
In addition to formal training, the OBIS Network is constantly working to build capacity through a variety of other services to the community. These often include relatively low cost and simple steps such as providing guidance on ways to leverage third party and readily accessible tools to publish source data and receive Digital Object Identifiers for citation purposes. While readily known in many developed parts of the world, these types of services are often not as well known in developing countries, and the OBIS Network provides an opportunity for knowledge sharing in many subjects.

Figure 1 shows the amount of collaboration between nearly 3,000 scientists from 70 countries in publishing over 1,200 papers citing OBIS. This clearly illustrates strong connections between countries to develop new science making use of this open data resource, but it also demonstrates emerging North-South and South-South collaboration, which are areas OBIS wish to improve further.

State of the Data

Of all described marine species today (243,000 in the World Register of Marine Species), about half of them have a distribution record in OBIS. Furthermore, 56% of the species in OBIS have less than 10 records, which indicates that the bulk of the data in OBIS represents well-known, easily monitored species (**Figure 2**). Also, there is an important bias toward the shallow coastal areas and the surface layers of the ocean, creating a gap in the midwater pelagic environments and deep-sea communities (Webb et al., 2010). Despite the fact that rare or under-reported species can occupy important local ecological niches and therefore should not be neglected in biodiversity assessments, incidental observations rarely end up in big monitoring or research datasets and hence are missing in global data systems such as OBIS. At best, new species records or invasions are published in the literature (often gray literature). Those few observations can sometimes provide early indicators of change, and often fill crucial gaps which can influence distribution models and forecasts or provide important insights in the distributions of marine species. There is growing interest from

FIGURE 2 | Number of taxa identified at species level in OBIS by number of records in the database. More details about the distribution of the records by taxa and statistics on quality issues can be found on the OBIS statistics page, which is updated with every new record published in OBIS (https://obis.org/statistics).

individual scientists who wish to share biodiversity observation records through OBIS directly, be able to immediately validate and publish those records, and subsequently retrieve a more complete and robust dataset for analysis. OBIS is developing a simple online data entry and editing tool that will allow individual scientists to submit these incidental records without the need to worry about whether they follow agreed standards or the need to organize data in the right formats which is identified as the main challenge of data sharing in a large survey published by Nature (Stuart et al., 2018). This will enable the network of data contributors to grow by lowering the barrier of entry.

RECOMMENDATIONS AND CONCLUSION

Modern society requires robust and timely information in order to generate a better understanding of the marine environment and make better informed decisions in management and conservation. Having a system that dynamically navigates the relationships between disparate data, synthesizing a set of useful information products is a vital asset in analyzing decisions. But the effective use of the system is only achieved by a carefully planned training program that brings together different levels of users and different needs into a common platform.

Countries should openly communicate on their policies regarding the public release of environmental data and increase the implementation of data publication mechanisms through active encouragement and if necessary, enforcement. In some countries, project funding depends on data released in the public domain, or a scholar's performance is measured taking into account dataset citations along with peer-reviewed papers. Measuring performance in this paradigm has well-established methods and practices. Measuring the impact of science on policy and resource management decisions is less clear. New metrics must be developed considering the impact of contributions of data and products on the sustainable use and conservation of our marine environment. OBIS promotes the establishment of a virtuous cycle between data and knowledge generators, data integrators and synthesizers, and policymaking stakeholders.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS

REFERENCES

Benson, A., Brooks, C. M., Canonico, G., Duffy, E., Muller-Karger, F., Sosik, H. M., et al. (2018). Integrated observations and informatics improve understanding of changing marine ecosystems. *Front. Mar. Sci.* 5:428. doi: 10.3389/fmars.2018. 00428

Bosch, S., Provoost, P., and Appeltans, W. (2018). *iobis/obistools: version 0.0.6. (Version v0.0.6) Zenodo.* Available at: http://doi.org/10.5281/zenodo.1489937 (accessed November 16, 2018).

Bourlat, S. J., Borja, A., Gilbert, J., Taylor, M. I., Davies, N., Weisberg, S. B., et al. (2013). Genomics in marine monitoring: new opportunities for assessing marine health status. *Mar. Pollut. Bull.* 74, 19–31. doi: 10.1016/j.marpolbul. 2013.05.042

Bowler, D. E., Hof, C., Haase, P., Kröncke, I., Schweiger, O., Adrian, R., et al. (2017). Cross-realm assessment of climate change impacts on species' abundance trends. *Nat. Ecol. Evol.* 1:67. doi: 10.1038/s41559-016-0067

Chaudhary, C., Saeedi, H., and Costello, M. J. (2016). Bimodality of latitudinal gradients in marine species richness. *Trends Ecol. Evol.* 31, 670–676. doi: 10.1016/j.tree.2016.06.001

Chaudhary, C., Saeedi, H., and Costello, M. J. (2017). Marine species richness is bimodal with latitude: a reply to fernandez and marques. *Trends Ecol. Evol.* 32, 234–237. doi: 10.1016/j.tree.2017.02.007

Costello, M. J., Tsai, P., Wong, P. S., Cheung, A. K. L., Basher, Z., and Chaudhary, C. (2017). Marine biogeographic realms and species endemicity. *Nat. Comm.* 8:1057. doi: 10.1038/s41467-017-01121-2

De Pooter, D., Appeltans, W., Bailly, N., Bristol, S., Deneudt, K., Eliezer, M., et al. (2017). Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. *Biodivers. Data J.* 5:e10989. doi: 10.3897/BDJ.5. e10989

Dornelas, M., Gotelli, N. J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., et al. (2014). Assemblage time series reveal biodiversity change but not systematic loss. *Science* 344, 296–299. doi: 10.1126/science.1248484

Griffiths, H. J., Meijers, A. J. S., and Bracegirdle, T. J. (2017). More losers than winners in a century of future Southern Ocean seafloor warming. *Nat. Clim. Chang.* 7, 749–754. doi: 10.1038/nclimate3377

Menegotto, A., and Rangel, T. F. (2018). Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. *Nat. Commun.* 9:4713. doi: 10.1038/s41467-018-07217-7

Müller-Karger, F. E., Hestir, E. L., Ade, C., Turpie, K., Roberts, D. A., Siegel, D. A., et al. (2018a). Satellite sensor requirements for monitoring essential biodiversity variables of coastal ecosystems. *Ecol. Appl.* 28, 749–760. doi: 10.1002/eap. 1682

Müller-Karger, F. E., Miloslavich, P., Bax, N. J., Simmons, S. E., Costello, M. J., Sousa Pinto, I., et al. (2018b). Advancing marine biological observations and data requirements of the complementary Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs) frameworks. *Front. Mar. Sci.* 5:211. doi: 10.3389/fmars.2018.00211

Olson, R. J., and Sosik, H. M. (2007). A submersible imaging-in-flow instrument to analyze Nano-and microplankton: imaging FlowCytobot. *Limnol. Oceanogr. Methods* 5, 195–203. doi: 10.4319/lom.2007.5.195

Provoost, P., and Bosch, S. (2018). *iobis/robis: version 1.0.2. (Version v1.0.2) Zenodo.* Available at: http://doi.org/10.5281/zenodo.1489949 (accessed November 16, 2018).

Saeedi, H., Dennis, T. E., and Costello, M. J. (2017). Bimodal latitudinal species richness and high endemicity of razor clams (Mollusca). *J. Biogeogr.* 44, 592–604. doi: 10.1111/jbi.12903

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., et al. (2018). *Whitepaper: Practical Challenges for Researchers in Data Sharing*. LONDON: Figshare. doi: 10.6084/m9.figshare.597 1387

Sutton, T. T., Clark, M. R., Dunn, D. C., Halpin, P. N., Rogers, A. D., Guinotte, J., et al. (2017). A global biogeographic classification of the mesopelagic zone. *Deep Sea Res. I* 126, 85–102. doi: 10.1016/j.dsr.2017. 05.006

The Group of Experts of the Regular Process, (ed.) (2017). *The First Global Integrated Marine Assessment: World Ocean Assessment I*. Cambridge, MA: Cambridge University Press.

United Nations Educational, Scientific and Cultural Organization (2017). *United Nations Decade of Ocean Science for Sustainable Development (2021-2030)*. Available at: https://en.unesco.org/ocean-decade (accessed April 29, 2019).

Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., et al. (2015). Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. *Database* 2015:bau125. doi: 10.1093/database/ba u125

Webb, T. J., Vanden Berghe, E., and O'Dor, R. (2010). Biodiversity's big wet secret: the global distribution of marine biological records reveals chronic under-exploration of the deep pelagic ocean. *PLoS One* 5:e10223. doi: 10.1371/journal. pone.0010223

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18