Publishing datasets, using artificial intelligence to help with metadata, can enhance ocean sustainability research and management

Angel Borja*

AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Herrera Kaia, Pasaia, Spain

Introduction

Just 40 years ago I completed my PhD and, since then, my whole career has been dedicated to monitor the ocean and make accessible the information and knowledge generated by my research team to decision-makers and managers. Also, just 30 years ago, my research team started to monitor the Basque coast and estuaries (North Spain), for the Basque Water Agency (URA), collecting physico-chemical and biological data on water, sediments, biomonitors, phytoplankton, macroalgae, macroinvertebrates and fish (Borja et al., 2016). This information has served to assist in the implementation of the European Water Framework Directive (European Commission, 2000), by developing assessment tools (Borja et al., 2004), but especially in implementing measures to restore and recover marine degraded systems, through different hydrological plans in the basins, culminating in the third one for the period 2022–2027 (MITERD, 2023).

As in this monitoring network which we coordinate, in the last 50 years the number of marine monitoring networks has increased dramatically (Borja and Elliott, 2021). At the same time, the public availability of raw data from marine networks have been increasing (Míguez et al., 2019; Coumans, 2024). In fact, open data is a fundamental component of the broader open science process and publication (Beck et al., 2020). In this process, national research funding organizations and governments, together with research organizations, have an important role in setting conditions for open access publishing of research resulting from public funding, as it has done in the European Union (European Commission, 2024) and United States, after the Presidential executive order from August 2022 followed suit and mandated immediate public access to all articles published by the end of 2025 resulting from federally funded research (Franco-Santos, 2024). However, synthesizing heterogeneous data from different ecosystem components in a monitoring network, coding all data preparation, and creating standard formats and metadata, to make reproducible, collaborative and transparent science (Lowndes et al., 2017), could prevent scientists from publishing large open datasets.

Monitoring and innovation in accessing open datasets

During the 40 years of my career, although evolving toward more and better technologies, most of the methods used in marine monitoring can be considered as

traditional and standardized (Anonymous, 2002; Karydis and Kitsiou, 2013; UNEP, 2016). However, in the last 10–15 years, many innovative and practical tools for monitoring and assessing the marine status have been developed and have experienced a growing use (Borja et al., 2024). The most common types of emerging methods include, among others, portable eDNA sequencers, underwater cameras, modeling methods, drones, satellites and artificial intelligence assisted data processing (European Commission et al., 2023).

Regarding data, there is now a range of technologies emerging for processing large volumes of heterogeneous environmental data (Vitolo et al., 2015). In fact, one of the ten strategic areas to strengthen the European Union's global leadership, is the capacity in data management, artificial intelligence and cutting-edge technologies (European Commission et al., 2022). In the introduction, I have commented some facts that can prevent scientists to share datasets. Nowadays, the need for ever more sophisticated data processing makes it even harder to meet the open data standards, which are needed going forward to make data accessible and synoptic analyses possible (Addison et al., 2018). Hence, the increasing scope of data collected and the potential future purposes for which they can be used (e.g., different sectors of Blue Economy -fisheries, aquaculture, tourism, biotechnology, etc., as well as maritime spatial planning, conservation, management, protection, restoration, assessment, etc.), means that traditional and emerging tools and processes for collecting, storing and analyzing datasets may become increasingly bespoke, particularly if the trend for repurposing data continues (e.g., the use of artificial intelligence and machine learning to extract new information from existing open access databases) (Addison et al., 2018).

In the last decade, several scientific journals have been created to publish open data, e.g., *Data in Brief, Scientific Data, GigaScience, Biodiversity Data Journal*, etc. However, when I was contacted by Frontiers Media to attend the presentation of the idea of a new platform for publishing open data, using generative artificial intelligence to assist the authors in preparing the datasets and metadata, as well as in writing the text accompanying the data, I was impressed by the first tests undertook. Hence, I offered the developers of the tool to use the large database generated for the Basque Water Agency, challenging the tool with real data and a good knowledge of the environment. The fact that the tool can learn not only from the dataset itself, but also from the ORCID numbers of the authors or additional information, was an added value for the experience.

After some interactions and tests, the text created had some shortcomings, but the experience of the authors allowed to easily and quickly build a final manuscript which has been the first published as a new article type (FAIR[2] Data Article) in *Frontiers in Ocean Sustainability* (Borja et al., 2025). As main author of this manuscript, I'm fully engaged with the five principles of human accountability and responsibility to protect the integrity of science in the age of generative artificial intelligence, as proposed by Blau et al. (2024): (i) transparent disclosure and attribution of the work done with the artificial intelligence in handling the dataset and writing the paper; (ii) verification of the content and analyses generated by the artificial intelligence, ensuring as scientists the accuracy of the data, imagery, and inferences draw from the use of generative models in writing the paper; (iii) documentation of data

and metadata generated by the artificial intelligence; (iv) focusing on ethics and equity, to ensure that products (i.e. metadata, texts, figures, tables) are scientifically sound and provide socially beneficial results (in this case, datasets fully and freely available), and (v) continuous monitoring, oversight, and public engagement to evaluate the impact of artificial intelligence on the scientific process, to maintain integrity and reproducibility.

Use of open access datasets in ocean sustainability research and management

In 2022, member states asked the United Nations Environment Programme (UNEP) to examine how artificial intelligence could accelerate work in three areas: climate action, nature protection, and pollution prevention (Wilson, 2024). In response, the UNEP (i) launched the World Environment Situation Room (wesr.unep.org), a digital platform that is planning to leverage artificial intelligence capabilities to analyse complex, multifaceted data sets, and (ii) is committed to develop a Global Environmental Data Strategy by 2025, aiming to improve monitoring data standards and digital cooperation between countries, and finally contributing to drive new frontiers in ecological research and management (Wilson, 2024).

Most of the ecological and biodiversity monitoring data will be needed to take decisions on conservation and restoration, especially after the adoption of the Kunming–Montreal Global Biodiversity Framework of the Convention on Biological Diversity (CBD, 2022). Similarly, the European Biodiversity Strategy 2030 has as a main policy goal to halt the decline of biodiversity and promote its recovery by 2030 (European Commission, 2020). One way to achieve this goal is based in legally binding restoration targets of 30% of degraded ecosystems, by 2030, 60% by 2040, and 90% by 2050, as approved by the Nature Restoration Law (Hering et al., 2023).

In this context, after a standardized survey, undertook by Moersberger et al. (2024), European science and policy stakeholders identified four clusters of key policy questions related to biodiversity monitoring within the next decade: (i) "Assessing biodiversity and species trends", including biodiversity status and trends, indicators for the quality of habitats, and assessing the impact of invasive species on the environment; (ii) "Biodiversity policy impact and effectiveness", including the assessment of the effectiveness of biodiversity policies and the outcomes of conservation management and restoration; (iii) "Integrating biodiversity in other policy sectors", including agriculture, fisheries, water management, climate change, green and blue infrastructure projects, poverty, equity, and trade; and (iv) "Operationalization of monitoring", including ways to standardize and harmonize biodiversity monitoring programs and integrate novel technologies to meet policy targets. Among those novel technologies, artificial intelligence occupies a relevant position (Moersberger et al., 2024).

In the case of the ocean, the increasing threats to biodiversity, coming from human activities, as well as the effects of climate change, resulted in a Workshop between the Intergovernmental Panel on Climate Change (IPCC) and the Intergovernmental

Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (Pörtner et al., 2021). After that, to build synergies between strategies for climate, biodiversity, ocean and human health, a group of scientists proposed the establishment of an International Panel for Ocean Sustainability (IPOS) (Gaill et al., 2022). After these authors, IPOS could facilitate the implementation of a global, integrated and fit-for-purpose observing system, providing information for robust understanding, monitoring, predicting and projecting the state of the ocean, across requirements and scales (from global to local), in alignment with the Global Ocean Observing System.

Again, innovative digital tools that use observation, advanced modeling and data management can be integrated into a digital twin of the ocean (an open source of combined ocean observations, artificial intelligence, and advanced modeling providing a consistent, high resolution, multi-dimensional and near real-time virtual representation of the ocean) (Gaill et al., 2022). This will be a multidisciplinary endeavor, involving the acquisition, integration and analysis of an increasing amount of ocean data. For completing this, Sagi et al. (2020) identified the key missing tools, with a focus on *"(i) development of artificial intelligence-based tools for assisting ocean scientists in aligning their schema with existing ontologies when organizing their measurements in datasets; (ii) extension and refinement of conceptual coverage of—and conceptual alignment between—existing ontologies, to better fit the diverse and multidisciplinary nature of ocean science; (iii) creation of ocean-science-specific entity resolution benchmarks to accelerate the development of tools utilizing ocean science terminology and nomenclature; (iv) creation of ocean-science-specific schema matching and mapping benchmarks to accelerate the development of matching and mapping tools utilizing semantics encoded in existing vocabularies and ontologies; (v) annotation of datasets, and development of tools and benchmarks for the extraction and categorization of data quality and preprocessing descriptions from scientific text; and (vi) creation of large-scale word embeddings trained upon ocean science literature to accelerate the development of information extraction and matching tools based on artificial intelligence."*

Hence, one of the main lessons learnt during these years is that building on adequate knowledge architecture is essential for sustainability transitions (Oliver et al., 2021). For assisting in this endeavor, *Frontiers in Ocean Sustainability* has included this new article type (FAIR[2] Data Article), making data available, which can benefit the ocean scientific community by providing the necessary information to take informed decisions on marine management, for a sustainable use of the ecosystem services. As pointed out by Borja (2023), this can benefit also multiple international initiatives needing data available, and taking place around the sustainability of the planet and, specifically, the ocean: (i) the United Nations (UN) Sustainable Development Goals (SDGs), including SDG14, to conserve and sustainably use the ocean, seas and marine resources for sustainable development; (ii) the UN Decade of Ocean Science for Sustainable Development, which will increase the international collaboration on scientific research; (iii) the UN Decade on Ecosystem Restoration, including marine degraded ecosystems; (iv) the "30by-30" from the "High Ambition Coalition for Nature and People", a worldwide initiative for governments to designate 30% of Earth's land and ocean area as protected areas by 2030; and (v) the Agreement under the United Nations Convention on the Law of the Sea on the conservation and sustainable use of marine biological diversity of areas beyond national jurisdiction.

Of course, artificial intelligence can be used also in an unethical way, e.g., by creating fake datasets, or creating new patterns of overexploitation and unforeseen interactions between human activities and marine ecosystems. This presents a paradox: while generative artificial intelligence can enhance sustainability through better data management, it may also drive the depletion of marine resources, creating new environmental costs and sustainability challenges. Hence, as editors of the journal, we must be attentive to any misuse of these technologies, verifying the content and analyses generated by the artificial intelligence, and ensuring the accuracy of the data and accompanying information and explanations.

We encourage the whole ocean scientific community to provide and use data from surveys, monitoring networks, PhD and master thesis, national and international projects, etc., on the benefit of the sustainability of the ocean, through an informed management decision process.

Author contributions

AB: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

Acknowledgments

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Addison, P. F. E., Collins, D. J., Trebilco, R., Howe, S., Bax, N., Hedge, P., et al. (2018). A new wave of marine evidence-based management: emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES J. Marine Sci.* 75:941–952. doi: 10.1093/icesjms/fsx216

Anonymous (2002). "JAMP guidelines on quality assurance for biological monitoring in the OSPAR area," in *OSPAR Commission, Ref. No. 2002-15* (OSPAR), 1–38. Available at: https://www.ospar.org/work-areas/cross-cutting-issues/cemp (accessed January 10, 2025).

Beck, M. W., O'Hara, C., Stewart Lowndes, J. S., Mazor, R., Theroux, S., Gillett, D. J., et al. (2020). The importance of open science for biological assessment of aquatic environments. *PeerJ* 8:e9539. doi: 10.7717/peerj.9539

Blau, W., Cerf, V. G., Enriquez, J., Francisco, J. S., Gasser, U., Gray, M. L., et al. (2024). Protecting scientific integrity in an age of generative AI. *Proc. Nat. Acad. Sci.* 121:e2407886121. doi: 10.1073/pnas.2407886121

Borja, A. (2023). Grand challenges in ocean sustainability. *Front. Ocean Sustainab.* 1:1050165. doi: 10.3389/focsu.2023.1050165

Borja, A., Adarraga, I., Bald, J., Belzunce-Segarra, M. J., Cruz, I., Franco, J., et al. (2025). Marine biodiversity and environmental data: an AI-ready, open dataset from the long term (1995–2023) Basque Country monitoring network. *Front. Ocean Sustain* 2:1528837. doi: 10.3389/focsu.2024.1528837

Borja, A., Berg, T., Gundersen, H., Hagen, A. G., Hancke, K., Korpinen, S., et al. (2024). Innovative and practical tools for monitoring and assessing biodiversity status and impacts of multiple human pressures in marine systems. *Environm. Monit. Assessm.* 196:694, doi: 10.1007/s10661-024-12861-2

Borja, Á., Chust, G., Rodríguez, J. G., Bald, J., Belzunce-Segarra, M. J., Franco, J., et al. (2016). 'The past is the future of the present': learning from long-time series of marine monitoring. *Sci. Total Environm.* 566–567, 698–711. doi: 10.1016/j.scitotenv.2016.05.111

Borja, A., and Elliott, M. (2021). From an economic crisis to a pandemic crisis: the need for accurate marine monitoring data to take informed management decisions. *Adv. Marine Biol.* 89, 79–114. doi: 10.1016/bs.amb.2021.08.002

Borja, Á., Franco, J., Valencia, V., Bald, J., Muxika, I., Belzunce, M. J., et al. (2004). Implementation of the European water framework directive from the Basque country (northern Spain): a methodological approach. *Marine Pollut. Bullet.* 48, 209–218. doi: 10.1016/j.marpolbul.2003.12.001

CBD (2022). *Kunming-Montreal Global biodiversity framework: Decision adopted by the Conference of the Parties of the Convention on Biological Diversity CBD/COP/DEC/15/4. 19 December 2022*. Montreal: Convention on Biological Diversity. Available at: https://www.cbd.int/conferences/2021-2022/cop-15/documents (accessed December 9, 2024).

Coumans, F. (2024). Meet the European marine observation and data network. *Hydro Int.* 28, 23–25. Available at: https://www.hydro-international.com/content/article/meet-the-european-marine-observation-and-data-network (accessed February 12, 2025).

European Commission (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for community action in the field of water policy. *Offic. J. European Union* L327, 1–72.

European Commission (2020). "Communication from the Commission of the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, EU Biodiversity Strategy for 2030," in *Bringing Nature Back into Our Lives*. Brussels: COM(2020), 27.

European Commission (2024). *Study on Scientific Publishing in Europe: Development, Diversity, and Transparency of Costs*. Luxembourg: Publications Office of the European Union, 102.

European Commission, Directorate General for Maritime Affairs and Fisheries, Addamo, A., Calvo Santos, A., Guillén, J., Neehus, S., et al. (2022). *The EU Blue Economy Report 2022*. Luxembourg: Publications Office of the European Union. 232 pp. doi/10.2771/793264

European Commission, Directorate General for Research Innovation, Jessop, A., Chow, C., Dornelas, M., Pereira, H., et al. (2023). *MarBioME: Overview and Assessment of the Current State of Marine Biodiversity Monitoring in the European Union and Adjacent Marine Waters*. Luxembourg: Office of the European Union, 63.

Franco-Santos, R. M. (2024). A journey through open access publishing in aquatic sciences. *Limnol. Oceanog. Bullet.* 33, 52–59. doi: 10.1002/lob.10628

Gaill, F., Brodie Rudolph, T., Lebleu, L., Allemand, D., Blasiak, R., Cheung, W. W. L., et al. (2022). An evolution towards scientific consensus for a sustainable ocean future. *NPJ Ocean Sustainab.* 1:7. doi: 10.1038/s44183-022-00007-1

Hering, D., Schürings, C., Wenskus, F., Blackstock, K., Borja, A., Birk, S., et al. (2023). Securing success for the Nature Restoration Law. *Science* 382, 1248–1250. doi: 10.1126/science.adk1658

Karydis, M., and Kitsiou, D. (2013). Marine water quality monitoring: a review. *Mar. Pollut. Bullet.* 77, 23–36. doi: 10.1016/j.marpolbul.2013.09.012

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., et al. (2017). Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* 1:0160. doi: 10.1038/s41559-017-0160

Míguez, B. M., Novellino, A., Vinci, M., Claus, S., Calewaert, J. B., Vallius, H., et al. (2019). The European Marine Observation and Data Network (EMODnet): Visions and roles of the gateway to marine data in Europe. *Front. Marine Sci.* 6: 313 doi: 10.3389/fmars.2019.00313

MITERD (2023). *Real Decreto 35/2023, de 24 de enero, por el que se aprueba la revisión de los planes hidrológicos de las demarcaciones hidrográficas del Cantábrico Occidental, Guadalquivir, Ceuta, Melilla, Segura y Júcar, y de la parte española de las demarcaciones hidrográficas del Cantábrico Oriental, Miño-Sil, Duero, Tajo, Guadiana y Ebro. Boletín Oficial del Estado, 35, de 10 de febrero de 2023*, 19510–21315. Available at: https://www.boe.es/eli/es/rd/2023/01/24/35 (accessed February 12, 2025).

Moersberger, H., Valdez, J., Martin, J. G. C., Junker, J., Georgieva, I., Bauer, S., et al. (2024). Biodiversity monitoring in Europe: user and policy needs. *Conserv. Lett.* 17:e13038. doi: 10.1111/conl.13038

Oliver, T. H., Benini, L., Borja, A., Dupont, C., Doherty, B., Grodzińska-Jurczak, M., et al. (2021). Knowledge architecture for the wise governance of sustainability transitions. *Environm. Sci. Policy* 126, 152–163. doi: 10.1016/j.envsci.2021.09.025

Pörtner, H. O., Scholes, R. J., Agard, J., Archer, E., Arneth, A., Bai, X., et al. (2021). *Scientific Outcome of the IPBES-IPCC Co-Sponsored Workshop on Biodiversity and Climate Change* (IPBES). Available at: https://www.ipbes.net/events/ipbes-ipcc-co-sponsored-workshop-biodiversity-and-climate-change (accessed January 9, 2025).

Sagi, T., Lehahn, Y., and Bar, K. (2020). Artificial intelligence for ocean science data integration: current state, gaps, way forward. *Elementa. Sci. Anthrop.* 8:21 doi: 10.1525/elementa.418

UNEP (2016). *Integrated Monitoring and Assessment Guidance*. Copenhagen: UNEP(DEPI)/MED IG.22/Inf.7, 282.

Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J., and Buytaert, W. (2015). Web technologies for environmental Big Data. *Environm. Model. Softw.* 63, 185–198. doi: 10.1016/j.envsoft.2014.10.007

Wilson, N. (2024). Artificial intelligence helps drive new frontiers in ecology. *BioScience* 74, 306–311. doi: 10.1093/biosci/biae016