MARCO-BOLO

STRENGTHENING BIODIVERSITY OBSERVATION IN SUPPORT OF DECISION MAKING

Deliverable 1.1

Joint Data and Information Management Plan

1

Date of delivery

Dan Lear (MBA), Katrina Exter (VLIZ), Paolo Tagliolato (CNR), Pieter Provoost (UNESCO), Rob Barry (AWI)

PUBLIC

Funded by the European Union

UK Research and Innovation

Funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101082021 (MARCO-BOLO). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

UK participants in MARCO-BOLO are supported by the UKRI's Horizon Europe Guarantee under the Grant No. 10068180 (MS); No. 10063094

Document Information

Grant Agreement	101082021
Project Acronym	MARCO-BOLO
Project Title	MARine COastal BiOdiversity Long-term Observations

Deliverable Number	D1.1
Work Package Number	WP1
Deliverable Title	Joint Data and Information Management Plan
Lead Beneficiary	MBA, Partner Number
Author(s)	Dan Lear (MBA), Katrina Exter (VLIZ), Paolo Tagliolato (CNR), Pieter Provoost (UNESCO), Rob Barry (AWI)
Due Date	30.11.2024
Submission Date	29.05.2025
Dissemination Level	Public[1]
Type of Deliverable	Report[2]

Version 1	29.05.2025, Dan Lear (MBA)

[1] Dissemination level (DELETE ACCORDINGLY): **PU:** Public, **SEN:** Sensitive, **CL:** EU Classified, information as referred to in European Commission Decision 2015/844

[2] Type of deliverable (DELETE ACCORDINGLY): **R: Document,** Report, **DEM:** Demonstration, pilot, prototype, **DEC:** Website, patent filing videos, **DMP:** Data Management Plan, **Ethics:** Ethics deliverable

Executive Summary

The MARCO-BOLO project (MARine COastal BiOdiversity Long-term Observations) aims to enhance the integration, accessibility, and interoperability of marine biodiversity data across Europe and globally. This deliverable, D1.1, outlines the project's Joint Data and Information Management Plan, which serves as a foundational framework for managing data generated and mobilised by the project.

The plan is built around the principles of **FAIR data** (Findable, Accessible, Interoperable, Reusable), and where appropriate the **CARE**, and **TRUST principles**, additionally alignment with the **UN Ocean Decade Data and Information Strategy** is planned. The plan promotes the use of **Linked Open Data (LOD)** and **semantic web standards** such as JSON-LD to ensure broad accessibility and machine-readability of metadata.

Key components of the plan include:

- **Alignment with global standards** such as Essential Ocean Variables (EOVs) and Essential Biodiversity Variables (EBVs), ensuring data relevance and interoperability.
- **Support for data-generating work packages** through training, tools, and engagement activities to improve data literacy and standardisation.
- **Use of persistent identifiers (PIDs)** and community-standard vocabularies to ensure long-term traceability and reuse of data.
- **Integration with global and regional repositories** like OBIS, GBIF, EMODnet, and Zenodo to ensure long-term preservation and discoverability.
- **Provenance tracking** and metadata transformation workflows to document data origins and processing steps.
- **Community engagement** with global infrastructures and linked-data communities to align practices and share innovations.

The plan also addresses challenges such as digital literacy gaps, metadata standardisation, and the need for sustainable data infrastructure. It recommends the development of a long-term Persistent Identifier Service to support future projects.

Overall, this deliverable sets the stage for a robust, interoperable, and sustainable marine biodiversity data ecosystem, supporting evidence-based ocean governance and conservation efforts.

Contents

1. Objective

The overarching ambition of MARCO-BOLO (MARine COastal BiOdiversity Long-term Observations) is to demonstrate an enhanced, robust, and stakeholder-driven approach to aligning, integrating, and delivering biodiversity data and observing capacity. This approach will promote broad access and (re)use by connecting existing capability through innovation across the marine biodiversity value chain: from observation and data collection to data management. These advancements will be key to building the biological component of the coastal and marine Earth Observation Infrastructure in Europe, delivering mapping, monitoring and data access to support integrated ecosystem assessments in Europe. The heart of the biodiversity data challenge is the sheer heterogeneity of the data itself. "Biodiversity data" is not a single data type or drawn from a fixed set of sources: any data in which the presence of a lifeform, or traces thereof, can be recorded or detected can be classified as biodiversity data. (Bio)chemical data, sequence information, acoustics, remotely-sensed ocean colour, temperature, imagery, and videography are just a few sources of data which may be used to assess biodiversity. Consequently, biodiversity data are multi- and transdisciplinary, stewarded by diverse organisations, and widely scattered. High fragmentation of data acquisition, handling, and storage inevitably create problems in data management and delivery, restricting interoperability (at different levels/scales) and (in the marine realm) limiting opportunities to advance knowledge on coastal processes and resource management. Furthermore, the sustainability of isolated or fragmented coastal monitoring systems is very fragile. The MARCO-BOLO (MBO) ambition is to demonstrate how biodiversity monitoring assets can reduce fragmentation of coastal and marine biodiversity observations and further enable the use of agreed international standards towards a truly interoperable coastal and marine biodiversity data ecosystem at both European and global levels.

2. Key Principles

The project partners of MARCO-BOLO are committed to engage and work with the wider community of data generators, custodians and users to co-develop truly FAIR (meta)data systems. In order to facilitate sustainable and ongoing data flow, such systems must align with the standards and protocols of global data infrastructures. The key principal of 'Collect and describe once, publish and use many times' is core to the MARCO-BOLO project and the overarching principles are laid out in the Project Data Management Plan (https://doi.org/10.5281/zenodo.8208410).

The MARCO-BOLO project is the first EU project to align its data management activities with the UN Ocean Decade Data and Information Strategy Implementation plan and as such supports the three main principles of Ethics, Competency and Multilateralism, along with the aforementioned FAIR principles and the CARE and TRUST principles where relevant

3. Open Data Approaches

A proactive approach to the promotion of Linked Open Data (LOD) and interoperability mediated through established Web technologies ensures that MARCO-BOLO derived and mobilised (meta)data

can be accessible by the widest range of end-users and at a variety of points along operational data pipelines.

Leveraging those standards endorsed by the World Wide Web Consortium provides the widest potential interoperability for MARCO-BOLO (meta)data. For example JSON-LD is a lightweight, JSON-based serialisation format of RDF designed to structure and link data broadly across the web. It helps developers represent data in a machine-readable way that is easily understood by search engines, applications, and other systems beyond the environmental domain.

By utilising JSON-LD MARCO-BOLO is supporting and promoting the advancement of academic data with respect to the adoption of the Linked Open Data approach.

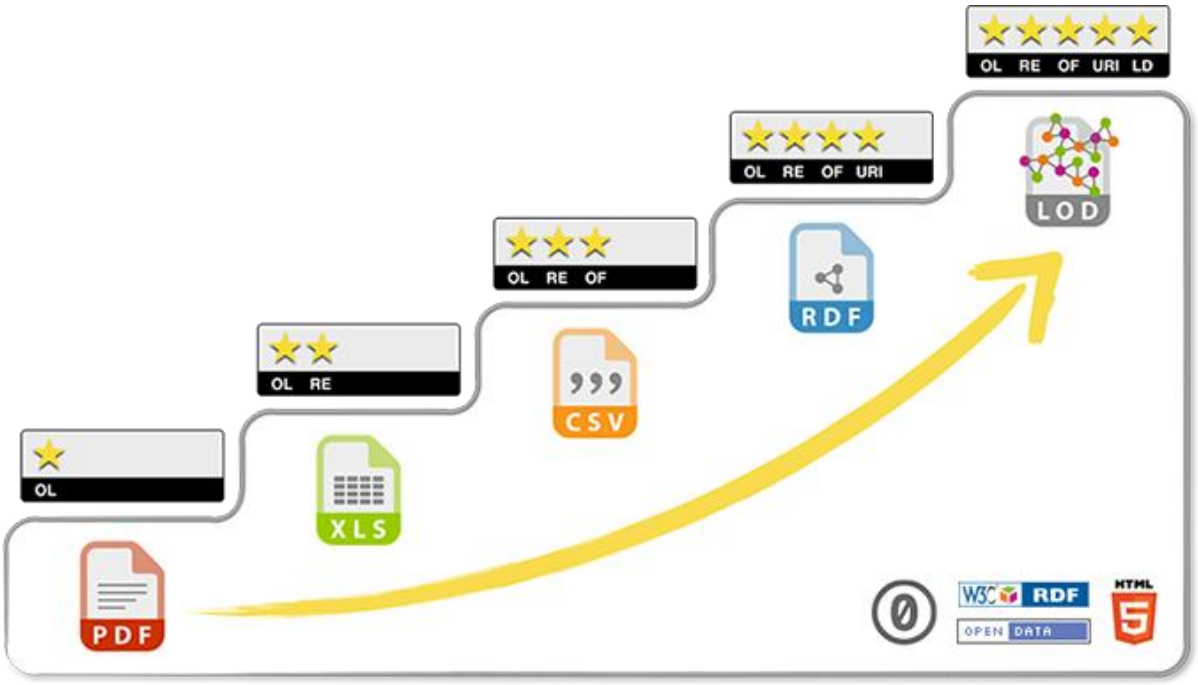

Fig 1: The 5 Star Open Data model

Tim Berners-Lee's **Five-Star Open Data** approach is a simple framework designed to encourage organisations and governments to make their data more open and accessible, especially on the web. Each "star" level represents an increasing degree of openness and usability:

★ : Data is available online in any format, even if not structured (e.g., a PDF or image). This level makes the data accessible to anyone, but it might not be easy to use.
★★ : Data is available in a structured format, like an Excel sheet, which allows for some basic sorting and filtering.
★★★ : Data is shared in a non-proprietary, open format (e.g., CSV instead of Excel), making it easier for anyone to use without needing specific software.

★★★★ : Data uses URIs (Uniform Resource Identifiers) for individual items, making each item uniquely addressable on the web. This approach enables people to link directly to data elements.

★★★★★ : Data is linked to other datasets, creating a web of interlinked, open data. This level maximises the data's utility by allowing connections across datasets, making it much easier to integrate, analyse, and find patterns.

Berners-Lee's model emphasises gradually improving data quality and accessibility, ultimately aiming to create a linked, open, and richly interconnected data ecosystem on the web.

JSON-LD supports this model as follows

★ : JSON-LD supports basic data availability on the web, as it is easily publishable online in a simple JSON format. This makes the data accessible and usable through a widely accepted format that's simple for most developers.

★★: JSON-LD is a structured format, allowing data to be organise in key-value pairs. This structure is machine-readable, making it easier to parse and process than unstructured data, such as PDFs or plain text.

★★★: JSON-LD is an open, non-proprietary format, fitting the requirement for an open format at the three-star level. This means anyone can use it without specific software constraints, and it is compatible across platforms.

★★★★: JSON-LD leverages **URIs** to identify entities within the data, which means that each element can be uniquely identified on the web. This feature aligns with the four-star approach, making data elements addressable and allowing them to be referenced or linked externally.

★★★★★: JSON-LD's main advantage is its focus on **linked data**, enabling the integration of data with other datasets on the web. JSON-LD provides a standardised way to link data elements to other datasets, fostering a web of interlinked, open data that fulfils the five-star approach. By using vocabularies like Schema.org, JSON-LD allows data to be understood within a broader context and easily integrated with other datasets.

4. Alignment with Essential Variables & Indicators

The MARCO-BOLO project aims to enable technologies for accurate biodiversity observations and improve data acquisition, focusing on Essential Ocean Variables (EOVs), Essential Biodiversity Variables (EBVs), and other relevant indicators, eg MSFD.

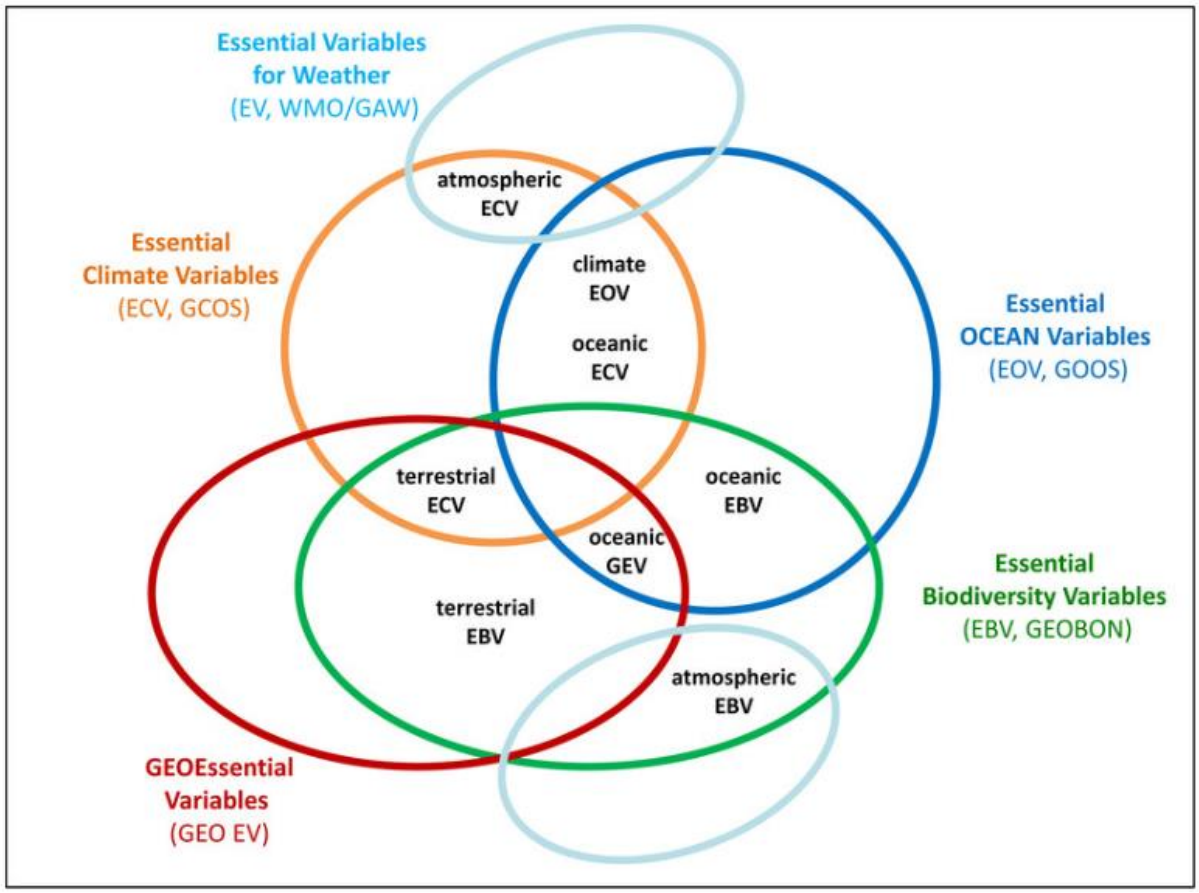

Figure 2- Conceptual overlap of Essential Variables

Essential Ocean Variables

The Global Ocean Observing System (GOOS) has defined a set of *Essential Ocean Variables* (EOVs) to monitor and understand ocean biodiversity, focusing on key biological and ecological aspects alongside those EOVs for Physics and Biochemistry[3]. The biodiversity-related EOVs are designed to capture information about the abundance, distribution, and health of marine species, ecosystems, and habitats. Key biodiversity-focused EOVs cover:

1. **Plankton**: This includes both phytoplankton and zooplankton communities, as these organisms form the base of the oceanic food web and are sensitive indicators of ocean health and climate changes.
2. **Fish**: The monitoring of fish biomass and distribution helps to understand marine ecosystem health, assess fisheries' sustainability, and track the impacts of climate change on fish populations.

[3] https://goosocean.org/what-we-do/framework/essential-ocean-variables/

3. **Marine Mammals, Birds, and Reptiles**: These higher trophic level species serve as indicators of ecosystem health, providing insight into food web dynamics and the impacts of human activities on marine biodiversity.
4. **Benthic Invertebrates and Seagrass**: Tracking species composition and abundance within these groups highlights ecosystem productivity and habitat health, including the resilience of coral reefs, kelp forests, and other key marine habitats.
5. **Habitat-forming Species**: This EOV monitors species that create or significantly modify habitats, such as corals, mangroves, and seagrasses. Changes in these habitats signal broader ecosystem shifts, including habitat degradation or recovery.

Each biodiversity EOV under GOOS supports monitoring of marine life responses to human pressures and natural variability, informing marine conservation and management efforts globally. The GOOS community have established specifications for the majority of the BioEco elements, based on published templates[4]. The EOV specifications describe sub-variables to be measured to estimate the EOV, derived products calculated from the EOV and its sub-variables, and observing requirements in terms of technology, platforms, resolution, and uncertainty. A data and information management section is currently being developed. Through the adoption of the EOV specifications and standards we increase the interoperability of data at a global scale, greatly increasing the available evidence base for ocean assessments and increasing the global FAIRness of marine biodiversity data.

Essential Biodiversity Variables

The GEO BON (Group on Earth Observations Biodiversity Observation Network) Essential Biodiversity Variables (EBVs) are a standardised framework to monitor global biodiversity changes. They provide a structured set of key biodiversity metrics across six categories: Genetic Composition, Species Populations, Species Traits, Community Composition, Ecosystem Structure, and Ecosystem Function. These variables are designed to be globally relevant, scientifically rigorous, and feasible to monitor, making them suitable for informing conservation policies, tracking progress on biodiversity targets, and guiding sustainable management.

Strengths:

1. **Global Relevance**: EBVs offer a consistent approach for monitoring biodiversity across regions and ecosystems, supporting international conservation goals like the Convention on Biological Diversity (CBD) targets.
2. **Integration**: They facilitate data harmonisation and integration across diverse datasets, which is crucial for comprehensive, multi-scale biodiversity monitoring.
3. **Policy Support**: By focusing on measurable changes over time, EBVs are aligned with policy needs, helping guide evidence-based decision-making at national and global levels.

[4] https://oceanexpert.org/document/34010

Issues:

1. **Implementation Challenges**: Many regions lack the infrastructure, technical capacity, or funding to consistently collect and analyse EBV data, making global implementation difficult.
2. **Data Gaps**: Some EBVs, such as those on genetic diversity or certain ecosystem functions, are challenging to measure at scale, leading to gaps in data quality and availability.
3. **Complexity**: The diverse nature of EBVs can make it challenging for non-specialists to implement or interpret, potentially limiting their accessibility and use in local-level decision-making.

Overall, while the EBV framework is a strong conceptual foundation for biodiversity monitoring, its effectiveness hinges on increased investment in capacity-building, improving the underlying specifications and alignment with other, more established and mature frameworks including the biological components of the GOOS EOVs.

The definition of EBVs is an ongoing work taken by diverse communities. The GeoBON website (https://geobon.org/ebvs/what-are-ebvs/) structures the framework in a single hierarchy of EBV classes and names. In the GeoBON EBV Data Portal, the hierarchical structure of EBVs features some minor variations (e.g., the addition of two classes: Ecosystem Services and Cross-cutting EBVs, see the machine readable structure at https://portal.geobon.org/api/v1/ebv).

To better identify the EBVs in relation to the actual practices in European observing systems and environmental research, the H2020 project EuropaBON (https://europabon.org/ has proposed EBV specifications resulting from consultations with experts and research infrastructures. The ongoing work is documented in project documents ([EuropaBon – EBV Workflow template – https://doi.org/10.5281/zenodo.10971094],[D4.1. List and specifications of EBVs and EESVs for a European wide biodiversity observation network https://doi.org/10.3897/arphapreprints.e102530]) and within the project's GitHub wiki pages (https://github.com/EuropaBON/EBV-Descriptions/wiki/). EuropaBON is proposing to subdivide the EBVs with respect to the measured entity (ie species and ecosystems) [D4.1]. Both in the cited deliverable and in the wiki pages several **EBV fact sheets** in the form of synthetic tables are presented. Dimensions of EBV definitions are, among others, the metrics, i.e. what has to be measured, the spatial and temporal resolution needed for specific EBVs, and the taxonomic/ecosystem focus group. The inclusion of "metrics" definitions within the framework of the EBV concept is of particular interest in MARCO[-BOLO: Their inclusion enables the encoding of semantic relationships between the EBV concept itself and the observed property concepts already described in semantic web resources (e.g., ontologies and controlled vocabularies) used for the metadata annotation of scientific datasets. The presence of such a layer of relationships in the semantic web can be leveraged to assess the suitability of a dataset for various reuse scenarios. This approach will be further explored within MARCO-BOLO to facilitate EBV data discovery and integration.

The required level of standardisation at the variable level for EBVs is still in a considerable state of flux. Therefore, at this time the recommendation from MARCO-BOLO is when mapping to EBVs only the general name, as provided by GEOBON is used. For an example of the state of EBV descriptions see here - https://github.com/EuropaBON/EBV-Descriptions/wiki/Freshwater-Species-abundances-of-selected-wetland-bird-species

MARCO-BOLO Data Generating Work Packages

Validating the use of eDNA

Work Package 2 (WP2) focuses on validating the use of environmental DNA (eDNA) as a transformative tool for biodiversity monitoring in marine and coastal environments. eDNA involves analysing genetic material from environmental samples, such as water or sediment, to detect the presence and abundance of species. WP2 aims to refine and standardise methodologies for eDNA sampling, processing, and analysis, ensuring they are reliable, cost-effective, and applicable across diverse settings. This includes addressing challenges like sample contamination, optimising laboratory techniques, and developing robust protocols for data interpretation.

Metadata catalogue for WP2

Linking Land and Sea Biodiversity Observation

Work Package 3 (WP3) focuses on linking biodiversity observations across land and sea ecosystems. This integration aims to address the interconnected impacts of human activities and environmental changes on coastal and marine biodiversity. WP3 emphasises harmonising methodologies for data collection and interpretation to create a seamless observation network spanning terrestrial, coastal, and marine systems. By integrating various datasets and methodologies, the project seeks to identify cross-system interactions and dependencies, offering a comprehensive understanding of how land-based activities influence marine ecosystems and vice versa.

Metadata catalogue for WP3

Mapping Biodiversity with Autonomous Systems

Work Package 4 focuses on enhancing biodiversity mapping using autonomous systems. It leverages advanced robotics, optical sensors, and acoustic technologies to conduct unmanned monitoring of marine and coastal ecosystems. These systems provide high-resolution spatial and temporal data, critical for understanding biodiversity patterns and changes over time.

Metadata catalogue for WP4

Modelling and Mapping Coastal and Marine Biodiversity

Work Package 5 focuses on developing innovative methods for modelling and mapping coastal and marine biodiversity. Its objective is to create models that combine data from various sources, such as autonomous systems, environmental DNA (eDNA), and existing biodiversity databases, to predict biodiversity distribution patterns and assess ecosystem health. The work package emphasises understanding causal relationships between biodiversity and environmental factors, including human-induced stressors like pollution and climate change. This modelling effort will facilitate targeted conservation efforts and enhance the predictability of biodiversity trends in marine and coastal systems.

Metadata catalogue for WP5

Data support services

The aim of WP1 is to facilitate the adoption of current best practice in biodiversity data management within the MARCO-BOLO project. Through the provision of support, tools and the development of data literacy within the project we aim to ensure the long-term discoverability and accessibility of project generated and mobilised data. Additionally we will create data legacy whereby MBO data are aligned with global standards to ensure future integration into frameworks including EOVs and EBVs, increasing the available evidence base in support of effective ocean governance.

To aid the data generating WPs of MBO, WP1 has undertaken a range of engagement activities to provide support and guidance. These activities have included online and in person 'data surgeries' where the overall approach has been demonstrated. Additional follow-up one-to-one meetings have also been key in partner engagement.

Additionally the questions and experiences of partners engagement with data generation and mobilisation tasks has been integrated into a Frequently Asked Questions document, hosted on the MARCO-BOLO project github.

5. (Meta)data Transformation

Use of semantic web standards (JSON-LD)

Interoperability is at the heart of MARCO-BOLO across multiple platforms and actors. The use of JSON-LD, which is a widely-used, machine understandable, semantic web standard, to describe MARCO-BOLO research assets (principally, their metadata), provides the mechanism for optimal interoperability and FAIRness with key partner infrastructures including those forming part of the UN data architecture to facilitate integration at the global scale.

It is important to recognise the variability of data literacy within the project and the wider communities. Familiarity with formats like JSON-LD is patchy at best and the ability to engage with,

edit and update JSON-LD formatted text files presents a barrier to the effective capture of rich and machine-understandable metadata.

In addition to ensuring that the metadata of the MARCO-BOLO datasets are provided using semantic-web-standard serialisations and formats, it is important that the content of those metadata records align with established and open community-standard vocabularies and include community-standard terms where they exist, including those from the NERC Vocabulary Server (NVS) already used in SeaDataNet, EMODnet & OBIS and the ontologies from the OBO Foundry and Library.

At a minimum all MBO datasets will include content on their spatial and temporal coverage, on the parameter and taxonomic information included in the data, on the protocols used, and on the agents (both human & machine) involved in and responsible for the data collection and processing.

In addition, we will collect provenance metadata (again using web-standard serialisations) relevant to datasets originating from biological material and ending with E(OB)Vs, including the information about any source data or samples and all processing steps and involved software/workflows. Work will take place to align and describe the collected provenance information with established standards wherever applicable.

Work Package 1 will require and check that the data generating work packages provide PIDs (e.g. DOIs and URIs) , and appropriate community-standard vocabularies will be used for all relevant fields including the E(OB)Vs within their (meta)data.

By promoting and following the ODIS templates, i.e. thematic profiles encoding the definition of object “types” (classes), for the metadata (dataset metadata, software metadata, agents and provenance metadata) for all MARCO-BOLO assets, we will ensure the availability of MARCO-BOLO outputs to the IODE programme component, via Ocean Info Hub .

Where MARCO-BOLO utilised data are already published with appropriate levels of metadata we aim to harvest existing metadata. In the case of datasets being newly generated or lacking critical levels of metadata, the minimum level of metadata will be created and standardised by the responsible task owner within MARCO-BOLO.

Name	**schema.org equivalent property**	**Description**
DatasetIdentifier	identifier	An internal ID for the dataset, provided to link the datasets in this google sheet
DatasetDescriber	sdPublisher	The name of the agent (filled in in the Agents tab) who filled in this dataset row, and whom we can contact with any questions
DatasetTitle	name	A title for the dataset

DatasetDescription	description	A brief description of the dataset
DataLandingPageURL	url	A dataset landing page is a dedicated web page that serves as the entry point to access comprehensive information about a specific dataset, including its contents, structure, and metadata: for example, the DOI or URL of a metadata record of the dataset. If the same data are published in multiple places, use "|" to separate them.
NotAccessibleData	conditionsOfAccess	If the data do exist but are not published, please explain here why that is and how WP1 could help publish these data
InProgressDataDate	additionalProperty	For data that are still in progress, please fill in an approximate date (YYYY-MM) you expect them to be published
EVDescription	additionalProperty	If the dataset contains EOVs/EBVs, please describe the class of EOV/EBV in this dataset. WP1 will use this to choose an EOV/EBV MBO vocabulary, so please make the explanation clear and avoid the use of acronyms
IndicatorDescription	additionalProperty	If the dataset contains Indicators; please describe the class of indicators in this dataset. WP1 will use this to choose an EOV/EBV MBO vocabulary; so please make the explanation clear and avoid the use of acronyms

Table 1. Mandatory (or Conditional) dataset metadata elements

Name	schema.org equivalent property	Description
AgentKey	identifier	As used in the other sheets. Please just use A-Z and no spaces
Type	rdf:type	"Person", "Organization", or "Project". Marco-Bolo, for example, is a Project, while EMBRC is an Organization.
ID	identifier	Preferably ORCID for persons, ROR or EDMO for organizations. Enter the full URL for the ID
Name	name	First name last name(s) for persons (without a "," between the two), or Organisation or Project full name
AffiliationAgentKey	AffiliationAgentKey	For persons only, enter their organization affiliation here. Then make sure that organization is also listed in this sheet, and use their "AgentKey" here
URL	url	A URL for the agent. For organisations and projects this is required, for people it is optional
Email	email	Contact email address (for questions related to the dataset)
Country	workLocation	County of the agent's address. Use the ISO 2-letter code (https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2)

Table 2. Mandatory Agent metadata elements

Currently scripts created by MARCO-BOLO and hosted in the MBO GitHub repository[5] transform the spreadsheet metadata templates into JSON-LD for harvesting and discovery via ODIS.

Data and software produced within the project will also be required to be published in the most relevant repositories, and metadata for these will also be placed in OIH following our ODIS-based standards. For software, it is expected that the level of literacy relating to JSON-LD will be higher, and text templates for describing those software natively will be provided directly to the WPs to fill in, upon which they will be made available to ODIS via OIH.

[5] https://github.com/marco-bolo

6. Challenges

Digital literacy challenges

Data literacy within the marine biodiversity community faces several challenges that have limited effective data use, sharing, and analysis. Many researchers do not traditionally have strong backgrounds in data management, digital tools, or statistical software. Additionally, the diverse and often complex data formats historically used in marine science—such as genetic data, species occurrence records, or more general oceanographic measurements—have made it difficult to manage and standardise data for effective analysis and integration and wider adoption of globally interoperable FAIR data principles.

There have also previously been limitations on access to data science training and resources, which further hampers efforts to adopt data-driven methods and open data practices. As a result, these limitations have created barriers to collaborative research, data sharing, and broader use of marine biodiversity data for conservation and policy-making. In part this is mitigated by platforms such as the OceanTeacher Global Academy (OTGA). The OTGA is a global capacity-building initiative led by the Intergovernmental Oceanographic Commission (IOC) of UNESCO and contains free, online courses aimed at increasing digital literacy and familiarity with the key, UN endorsed infrastructures including ODIS, OBIS and EMODnet. Providing specialised training in oceanographic and marine science topics through online courses, in-person workshops, and a network of regional training centres. OTGA supports sustainable ocean management by improving knowledge and skills in areas such as marine biodiversity, data management, coastal zone management, and marine policy.

Due to a lack of familiarity with serialisation formats such as JSON-LD, it has been necessary for WP1 to create and support intermediate steps in the standardisation and transformation of (meta)data. The additional work has delayed the publication of MARCO-BOLO metadata whilst the development work took place. However challenges remain in the encouragement of data generating MBO WPs in the comprehensive and timely completion of the standardised spreadsheets.

Provenance Metadata Model

The MBO project aims to collect provenance metadata which should hold:

- **at minimum**: a high-level series of discrete steps which describe the processes and transformations which were applied to the data inputs and resulted in the creation of the published dataset; this should be expressed in a minimalist fashion, but contain sufficient detail that a reasonably informed member of the field could produce a close approximation to the published dataset without being privy to the specific tools and configurations used in its generation.
- **in most circumstances**: references linking each of the high-level steps to the specific applications, spreadsheets, scripts, and any relevant parameters or configuration which would enable a reasonably informed auditor to execute, with minimal additional work, each step used to generate the published dataset.

In order to meet these goals and at the same time ensure that the provenance metadata structure is well matched and sufficiently expressive for the approaches and tools used in the technical MBO work packages, WP1 has elected to consult with the technical work packages on the delivery of this element.

The consultative design process will take the form of an initial stage of user research which will elicit narratives describing the generation of datasets similar to those planned in MBO; these narratives will then be mapped to WP1's proposed schema.org provenance model; at this point the other work packages will be consulted to ensure that the provenance representation holds the information necessary to meet the goals of MBO's provenance model.

Aggregated Datasets

Several outputs from MARCO-BOLO derive from the aggregation of many (sometimes hundreds) of source datasets. In these cases it is neither feasible nor desirable to create or fully populate metadata records for each individual dataset. In these cases a bulk metadata record is recommended. One bulk record per source, or other pragmatic division should be created.

The Dataset (with an array linking out to composite datasets in the schema.org value space of hasPart) and DataCatalog types should be explored here, noting that the definition of DataCatalog is currently poorly crafted and too general. Not every collection of data sets is structured in a catalog.

For third-party datasets with PIDs already present on the Web, this should be straightforward. If any of the source data sets are not professionally managed (e.g. emailed, transiently cloud transferred from one scientist to another, kept on an FTP server with no sustainability plan or Linked Open Data approaches, or siloed in a poorly managed and/or Web-opaque institutional archive), we will need more details, especially of the responsible party and point of contact..

In the latter case, above, MARCO-BOLO partners should archive "wild" data in the MARCO-BOLO Zenodo space if permitted by the copyright holder.

Licensing

MBO data will, by default, be made openly accessible in public repositories under licences compatible with CC0 or CC-BY. Restrictions on data access are generally discouraged; however, if partners require restrictions (such as embargoes) to maintain competitive academic standing, these will be documented. Access to the data will be limited accordingly, with publicly available metadata detailing the nature of the restrictions, their limitations, and plans for eventual public release.

It is important to recognise that data accessed in the creation of MARCO-BOLO derived data products will respect and reflect the licensing attributed by the original creators.

Embargoed Data

MBO data should in general not be subject to embargo periods, and should be made available at or before the point of publication of any associated papers.

The conditions in which data may be embargoed are generally limited to circumstances where earlier publication would preclude any mandatory commercial exploitation, or where publication would lead to an unnecessary risk to endangered species. It is expected that any and all efforts will be taken to obviate or reduce the need for, and length of, any data embargo period.

Where an embargo is permitted, or negotiated with the CoP, the following conditions must be met:

- The embargo period should not last for longer than 12 months from the date of data generation or harvesting.
- Metadata must be published within 1 month of the generation or harvesting of data, and must include as a minimum:
    - Full metadata describing the data to the same extent as would be published alongside a dataset not subject to any embargo period.
    - Additional metadata describing the reasoning for the need for an embargo period, the length of the embargo period, and the expected publication date of the data.
    - The dataset's metadata must include a link which, upon completion of the embargo period, resolves to a location the data can be publicly accessed.

Any embargoed data should be uploaded to a suitable repository with an embargoing mechanism which supports denial of public access to the data prior to the embargo being lifted. Recommended repositories include the European Nucleotide Archive for nucleotide sequence information, and the MBO community on Zenodo for all other data types.

Where it is reasonably necessary to publish embargoed data to a repository which is not compatible with the above requirements, community members should make contact with WP1 to request support in the generation of a persistent URL (PID). This persistent URL can be listed as the data access location; it will initially display an embargo holding page but will redirect to the live data at the end of the embargo period. Responsibility for the timely provision and communication of a live data access URL at the end of the embargo period will lie with the data producer.

Persistent Identifiers

The use of globally unique PIDs is key to ensuring the *Findability* of data in the FAIR data approach; they ensure that it is possible to search for and unambiguously determine the statements which apply to a given entity. To meet this requirement, WP1 requires that every entity described in metadata which could reasonably be reused in another context must have a globally unique PID; this is to say that the anonymous definition of entities without PIDs is accepted so long as these entities are related to a single parent entity which has a globally unique PID, and either:

- the child entity exists solely to link to or wrap another globally identifiable term, or
- it is only possible to intelligibly interpret the meaning of the child entity given the context of the parent; i.e. it would not be possible to make sense of the child entity independently of the parent.

To ensure that *Accessibility* of the data, FAIR mandates that all PIDs are *retrievable* over an open, free and universally implementable protocol. To meet this requirement WP1 mandates the use of URLs as identifiers using the *HTTP*, or preferably *HTTPs* protocols.

Persistent Identifier Services

In order to support the long-term validity and interpretability of the data published by the short-term funded projects, such as MARCO-BOLO, it is necessary to make use of third party Persistent Identifier Services which are hosted and maintained by organisations committed to the maintenance of these identifiers over the coming decades to centuries.

Without the use of such a system, the MBO project PIDs would cease to be dereferencable at the end of project funding. In essence, without use of a Persistent Identifier Service the URLs used as identifiers by MBO which return (by HTTP) information necessary for the correct interpretation of other MBO data would cease to function; this would result in a rift in the web of linked data generated by MBO, making it difficult to impossible for a third party to correctly interpret the data produced by this project.

The MARCO-BOLO project makes use of the w3id.org permanent identifier service run by a consortium of organisations which aims to support URL PIDs over the timescale of decades to centuries. Should the location of data behind a particular PID be moved, for instance at a point in the future when a data hosting solution used by the project ceases to function, it will be possible to redirect requests for a given PID to an updated location where the data can be accessed.

The w3id.org permanent identifier service requires a degree of technical literacy in order to operate; the use of git, creation of pull requests and modification of apache2 `.htaccess` files is necessary in order to manage PIDs. The WP1 team do not believe that such a service is accessible to the wider marine and oceanographic communities; further, the WP1 team is not aware of any other Persistent Identifier Service which meets the affordability, reliability, longevity, and technical accessibility requirements which would make maintaining PIDs accessible to such an audience. The MARCO-BOLO project therefore recommends that the Horizon Europe programme should consider commissioning, financing, and maintaining an *HTTPs-based Persistent Identifier Service* for all funded projects which would meet the accessibility, longevity, and reliability requirements necessary to support global researchers of all fields in the affordable creation and maintenance of permanent identifiers with the aspiration to support such a service for at least the next two decades.

MARCO-BOLO Persistent Identifiers

Persistent identifiers registered by MBO will be semantically opaque, i.e. it will not be possible to infer what a given PID URL represents without querying the underlying data. This will ensure that PIDs remain valid in circumstances where the name or otherwise identifying information of the underlying entity changes over the course of time.

MBO Persistent identifiers will be of the form https://w3id.org/marco-bolo/mbo_0000001 where the slug holds a numeric ID which may range from `mbo_0000001` to `mbo_9999999`.

Delegation of PID Ownership

To support the independent work of MBO work packages PIDs will be partitioned into numeric ranges. Ownership of these numeric ranges will be delegated to each work package so that decisions on their use can be made without central control.

Individual work packages are free to reference their delegated PIDs at any time during the process of metadata creation and are responsible for deciding how and where they are assigned. At the point of metadata publication each work package will be required to communicate with WP1 to ensure that their assigned PIDs are resolvable to suitable linked data resources.

Reuse of OceanExpert Identifiers

Two key components of the FAIR principles are the interoperability and reusability of data; to this end the MBO project plans to, where possible, reuse existing identifiers relating to Organisations and Individuals collected in the IOC's OceanExpert system. This obviates the need to define and maintain an up-to-date registry of the individuals and institutions which collaborate with or provide data to the MBO project. Further, it promotes the interoperability of MBO data with other datasets which make use of OceanExpert identifiers and the wider IOC data infrastructure.

One drawback to relying on the IOC's OceanExpert system is that registration requires the input and open publication on the web of every individual contributor's name, nationality, work address, and email address, amongst other data. Publication of such personal information may represent an unnecessary risk to some contributors who are concerned about data privacy. This may lead to some individuals refusing to register for an OceanExpert PID; the impact of this would be to reduce the interoperability of some of the MBO data by requiring MBO to coin and maintain new PIDs and associated metadata for these contributors.

Dataset Metadata Interoperability

There are a number of information sources on the web which publish metadata about datasets, organisations, researchers, software applications, and other entities relevant to the MBO project's published metadata which it would be helpful to integrate into the MBO data-graph. For instance it is anticipated that many datasets used as inputs to MBO tasks will be accessible from institutional websites and data repositories which contain existing metadata about who published the data, when it was published, which geographic area it covers, the time period it applies to, alongside other helpful information such as licensing. This metadata is present in a variety of formats: some being only human readable, others being machine readable but expressed in formats or vocabularies incompatible with the ODIS/MBO schema.org representation profile, and some data which is already stored in ODIS.

Work Package 1 initially agreed to perform a limited amount of mapping of existing but incompatible metadata into ODIS in order to reduce the burden of data input on the data delivering work packages. Unfortunately, due to the previously unanticipated high number of existing datasets referenced across the MBO project, it is no longer practical for WP1 to provide help in this way. As a

result metadata input for input datasets is the responsibility of the data producing WPs; the minimal metadata input requirements are as follows:

Input Dataset Condition	Minimum WP Metadata Input Requirements	Resulting Metadata Quality
Already in ODIS (likely from EMODnet or OBIS)	Once referenced by the URI no further metadata describing the dataset is required.	High quality human and machine-readable metadata in the ODIS schema.org representation profile.
Not in ODIS, but metadata is available in an open format on the web.	A dataset definition or aggregate dataset definition is required which should define: • the name of the dataset, • a URL where the data and metadata can be openly accessed, and • the licensing conditions. Additional metadata may be provided where desired by the WP.	Reasonable quality human readable metadata. A very limited amount of machine-readable metadata in the ODIS schema.org representation profile.

Input Dataset Condition	Minimum WP Metadata Input Requirements	Resulting Metadata Quality
Not openly accessible on the web.	A dataset definition or aggregate dataset definition is required which should define: • the name of the dataset • an email address or other contact information of who and where the data was requested from, • a description of the data that was requested by the MBO WP, • a description of the data that was received by the MBO WP, and • the licensing conditions. Additional metadata may be provided where desired by the WP. The dataset itself should be uploaded to the MBO Zenodo community and made publicly available if permission to do so is granted by the data owner.	Limited human-readable metadata. A very limited amount of machine-readable metadata in the ODIS schema.org representation profile.

Table 3 Minimal metadata element examples

We see that the best results with the least effort on the part of the MBO work packages comes when using input datasets which are already defined in OBIS or EMODnet; this results in both high-quality human and machine-readable metadata with minimal effort.

The primary benefit to data producers and data consumers when marine and oceanographic metadata is published in an agreed machine-readable community standard format is that the provenance of data is more transparent; meaning that data can more easily be well interpreted, mistakes in upstream datasets can be contested sooner, inappropriate uses of data can be surfaced faster, and data becomes more trustworthy as a result. Where metadata is not published in an agreed and consistent machine-readable community standard this benefit is simply not present and it can be seen in MBO that when the responsibility of mapping any metadata into the community standard rests on the data consumer the effort required quickly adds up to something unsustainable and likely leads to inaccuracies of representation.

7. Community Engagement

The existing MARCO-BOLO partnership benefits from a number of individuals already being embedded within key global infrastructures in a variety of roles. This, coupled with the power of the Community of Practice established through WP6, ensures optimal engagement and facilitates alignment of process and standards.

WP1 is also coordinating with other Horizon Europe funded projects including BioEcoOcean in order to discuss and share approaches to ensure alignment around global standards and the infrastructures and processes as defined in the UN Decade Data Implementation Plan.

As part of MARCO-BOLO's goal to deliver interoperable metadata to downstream services, we have engaged with our anticipated primary data consumer ODIS to ensure that the JSON-LD generated by MBO can be ingested without issue [https://github.com/marco-bolo/csv-to-json-ld/issues/3].

The Wider RDF & Linked-data Communities

WP1 makes use of a number of existing tools and resources from the wealth of prior work done by those in the RDF and linked-data communities. Not all of the approaches trialled in MBO have worked out, but these approaches have led to WP1 contributing knowledge to the wider community; this includes a number of bug reports on the linkml GitHub repository, the discovery of and communication of an oversight in the W3C CSV on the Web specification and a contribution towards the maintenance of an open source tool called csvw-check.

8. Data Storage, Preservation, and Long-term Accessibility

The MARCO-BOLO project promotes the engagement with and utilisation of established, long term repositories for environmental data and data products generated and mobilised within the project. By leveraging existing infrastructures we ensure the long term availability of MBO data and transparency in our derived data products.

In order for a repository to be recommended within the MARCO-BOLO project, it is necessary for it to demonstrate a commitment to the adoption and promotion of open data principles and demonstrable and actionable implementation of the FAIR principles. By working with a federated set of thematic data centres, the MARCO-BOLO project is not dependent on or limited by project-based systems. The recommendations take into account factors such as repository longevity, obsolescence policies, persistent identifier (PID) handling, and embargo mechanisms.

Whilst not exhaustive an initial list of recommended repositories can be found below. WP1 of MARCO-BOLO will further define the requirements for long-term data storage in subsequent outputs.

ODIS

The United Nations Ocean Data and Information System (ODIS) is an integrated digital platform designed to support the sharing, accessibility, and use of ocean data worldwide. Developed by the

Intergovernmental Oceanographic Commission (IOC) of UNESCO, ODIS is part of the UN's Decade of Ocean Science for Sustainable Development (2021–2030) and aims to enhance global collaboration on ocean data by bringing together diverse sources of marine and oceanographic information. MBO is currently in discussion with the European Nucleotide Archive, Lifewatch and Zenodo with the intention of integrating these repositories into the ODIS federation.

OBIS

The United Nations' Ocean Biogeographic Information System (OBIS) is a global platform for collecting and sharing information about marine biodiversity. OBIS was developed to support the scientific understanding of ocean ecosystems by providing open access to data on the distribution and diversity of marine species. Managed under the Intergovernmental Oceanographic Commission of UNESCO, OBIS brings together data from various sources, including research institutions, government bodies, and citizen scientists, to create a comprehensive, centralised database.

GBIF

The Global Biodiversity Information Facility (GBIF) is an international organisation that provides open access to data on biodiversity around the world. It was established in 2001 and is supported by governments, research organisations, and other partners. GBIF's goal is to make data about life on earth freely available to anyone, which includes species occurrence records, specimen data, and observations from various sources such as research institutions, government agencies, and citizen scientists.

INSDC

The International Nucleotide Sequence Database Collaboration (INSDC) is a global partnership among three major bioinformatics organisations: GenBank, the European Nucleotide Archive (ENA), and DNA Data Bank of Japan (DDBJ). This collaboration enables the sharing and coordination of nucleotide sequence data (DNA and RNA) across the globe.

EMODnet

The European Marine Observation and Data Network (EMODnet) is an established European Commission marine in situ data service, providing seamless access to FAIR trusted marine data, metadata and products at pan-European scale. EMODnet provides free, standardised, and interoperable data across seven broad thematics: r.

- Bathymetry
- Geology
- Biology
- Chemistry
- Physics
- Human activities at sea

- Seabed habitats

The Marine Data Archive

VLIZ hosts a data archiving service, the Marine Data Archive (MDA) and associated data cataloging service, the Integrated Marine Information System (IMIS) that are open and as such can be used by all MBO colleagues. With the MDA, data files can be shared with colleagues, referees, and with the wider world. With IMIS, FAIR metadata records for any MBO datasets can be published. IMIS and the MDA are especially useful for data that cannot find a space in more specialised data systems. There are no constraints on data types or sizes, however we do not encourage using these VLIZ systems for datasets that are better placed in specialised archives (e.g. such as for omics data).

9. Climate Impact

The MBO team acknowledges that there will be a climate impact from the energy emissions necessary to process, store and query the metadata produced by each of the individual work packages. In order to minimise the impact of greenhouse gas emissions generated by the data processing recommended by WP1, the software and scripts have been designed using an incremental build system so that computationally intensive tasks are only performed on files which are known to have changed since the previous build; this will minimise computational runtime, and so reduce greenhouse gas emissions, at the same time as fitting in with the anticipated iterative approach of building metadata.

10. Next Steps

As we reach the midpoint of MARCO-BOLO it is clear that significant work remains in the communication and adoption of data-related standards within the technical work packages of the project. The role of WP1 has adapted to provide an interface to these standards, reducing the friction relating to data transformation and continuing to promote and increase data literacy within the MARCO-BOLO community. Further updates to this document, the organisational data management plan and through the regular project reporting will highlight the incremental improvements made, and we will utilise verification through the ODIS knowledge graph to highlight the publication successes for MBO datasets and other representable outputs.

Appendix

Acronym	Definition
ABS	Access and Benefit Sharing
API	Application Programming Interface
ARMS MBON	Autonomous Reef Monitoring Structure
	Marine Biodiversity Observation Network
AtlantECO	Atlantic Ecosystems Assessment, Forecasting and Sustainability
BBNJ	Biodiversity Beyond National Jurisdiction
BiCOME	Biodiversity of the Coastal Ocean: Monitoring with Earth Observation (project)
Bio Eco Panel	Biology and Ecosystems Panel
CA	Consortium Agreement
CC0	Creative Commons Zero
CC-BY	Creative Commons Attribution
CDIF	Cross-domain Interoperability Framework
CODATA	Committee on Data of the International Science Council
CoP	Community of Practice
CWL	Common Workflow Language
DCAT	Data Catalog Vocabulary
DITTO	Digital Twins of the Ocean
DMP	Data Management Plan
DWC	Darwin Core
DWC-A	Darwin Core Archive
EBI	European Bioinformatics Institute
EBV	Essential Biodiversity Variable
EC	European Commission
EMODnet	European Marine Observation and Data Network
EEA	European Environmental Agency

Acronym	Definition
EF	Ecosystem Function
EMBL	European Molecular Biology Laboratory
EMBR	European Marine Biological Resource Centre
EML	Ecological Markup Language
EMO BON	European Marine Omics Biodiversity Observation Network
ENA	European Nucleotide Archive
ENVRI	Environmental Infrastructure
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable
ERIC	European Research Infrastructure Consortium
ESA	European Space Agency
EU	European Union
Europa BON	Europa Biodiversity Observation Network
EV	Essential Variable
FAIR	Findable, Accessible, Interoperable, Reusable
FTP	File Transfer Protocol
GA	Grant Agreement
GBIF	Global Biodiversity Information Facility
GEO BON	Group on Earth Observations Biodiversity Observation Network
GOOS	Global Ocean Observing System
GPDR	General Protection Data Regulation
GSC	Genomic Standards Consortium
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IAB	International Advisory Board
IEC	International Electrotechnical Commission
IGSN	International Geo-Sample Number
IIIF	International Image Interoperability Framework

Acronym	Definition
INSDC	International Nucleotide Sequence Database Collaboration
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
IP	Intellectual Property
IPR	Intellectual Property Right
ISO	International Organisation for Standardisation
IT	Information Technology
IUPAC	International Union of Pure and Applied Chemistry
JSON-LD	Javascript Object Notation linked Data
LOD	Linked Open Data
MB	Mega Byte
MBO	MARCO-BOLO
MBON	Marine Biodiversity Observation Network
MIME	Multipurpose Internet Mail Extensions
MIxS	Minimum Information about any Sequence
MSFD	Marine Strategy Framework Directive
NERC	Natural Environment Research Council
NP	Nagoya Protocol
NVS	NERC Vocabulary Server
OBIS	Ocean Biodiversity Observation System
OBON	Ocean Biomolecular Observing Network
OBPS	Ocean Best Practices System
ODIS	Ocean Data and Information System
ODIS-Arch	Ocean Data and Information System Archive
ORCHID	Open Researcher and Contributor ID
PANGAEA	Publishing Network for Geoscientific and Environmental Data
PDEC	Plan for Dissemination, Exploitation and Communication Activities
PID	Persistent Identifier

Acronym	Definition
PROV	specification that provides a vocabulary to interchange provenance information
QA	Quality Assurance
QC	Quality Control
ROM	Requirement Oriented Model
ROV	Remote Operated Vehicle
SOSA	Sensor Observation Sample and Actuator
SSSOM	Simple Standard for Sharing Ontology Mappings
TB	Tera Byte
TDWG	Taxonomic Databases Working Group
UN	United Nations
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USV	Uncrewed Surface Vessel
W3C's Prov	World Wide Web Provenance
WFD	Water Framework Directive
WP	Work Package

MARCO-BOLO

STRENGTHENING BIODIVERSITY OBSERVATION IN SUPPORT OF DECISION MAKING

Project Coordinator
Nicolas Pade | nicolas.pade@embrc.eu

Project Manager
Giulia Vecchi | giulia.vecchi@embrc.eu

Press and Communications
Mathilde Vidal | mathilde@erinn.eu

Website: MarcoBolo-Project.eu
Twitter: @MARCOBOLO_EU
LinkedIn: MARCO-BOLO
BlueSky: @marco-bolo.bsky.social

Funded by the European Union

UK Research and Innovation