EUREMAP Data Management Plan

Human readable & machine actionable

Deliverable report

Project title	**EUREMAP - European Research Infrastructure for Marine Bioprospecting**
Grant Agreement No	101131663
Topic	HORIZON-INFRA-2023-DEV-01-04
Call	HORIZON-INFRA-2023-DEV-01 - Developing, consolidating and optimising the European research infrastructures landscape, maintaining global leadership (2023)
DG/Agency	DG Research and Innovation / Research Executive Agency
Deliverable number	14
Deliverable title	The EUREMAP DMP
Lead beneficiary	CCMAR
Contributing beneficiaries	All partners
Author(s)	Bruno Louro, Cymon J. Cox
Type of deliverable	Data management plan
Work package No	11
Work package title	Open science policy and project data management
Dissemination level	PU

Due date (in months)	M6
Delivery date (actual)	09-07-2024
Description of deliverable	The initial version of the EUREMAP Data Management Plan.

Legal Disclaimer

Funded by the European Union

Funded by the European Union (under grant agreement No 101131663 (HORIZON-INFRA-2023-DEV-01-04). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.

Table of contents

1. Introduction

The EUREMAP project aims to optimise the marine bioprospecting workflow within the European research community, enabling state-of-the-art research and enhanced capacity for academia and industry. This will be achieved by building upon the existing expertise available within the consortium members (EU-OPENSCREEN, EMBL, EMBRC, ELIXIR) to synergise capacities and resources to enhance marine bioprospecting for natural products. The EUREMAP platform to be developed will consist of a new multidisciplinary marine bioprospecting workflow that integrates the existing technical capacities, thereby increasing the competitiveness of marine natural products in drug discovery and strengthening the blue bioeconomy in line with the European blue growth strategy. The functioning EUREMAP marine bioprospecting workflow will support the discovery of novel bioactive natural products from marine organisms by utilising novel technology to improve the capacity of European Research infrastructures (RIs).

The multidisciplinary nature of the marine bioprospecting research space necessarily requires integrating data flows between disparate research fields. As such, while data management components within disciplines may be in place, they may not be adequately formulated to provide the basis for a cohesive and integrated data workflow. Consequently, the current Data Management Plan (DMP) is a living document that is continuously revised as the project proceeds, and the envisaged marine bioprospecting workflow is clarified and finalised. The DMP distinguishes internally produced project data (e.g. deliverables, meeting notes, agendas, etc.) and the data directly applicable to the functioning of the final workflow in production (e.g. sample metadata, DNA sequence data, liquid chromatography-mass spectrometry (LC-MS), etc.). Because the marine bioprospecting workflow is seen as having multiple entry and exit points, data management practices must be aligned between adjoining workflow technical components and service providers. The first draft of the DMP circumscribes the proposed data management practices based on an initial interpretation of results gathered from the beneficiaries in a "Survey of current data management practices" (M11.1). The survey results will ultimately provide the basis for the Deliverables 11.2 Status of the EUREMAP Open Science policy and 12.1 Data Interoperability Roadmap. At all stages of the project and throughout the implementation of the marine bioprospecting workflow, Open Science practices will be adhered to, and FAIR data practices will be followed whenever possible, subject to the requirements implied by ethically and commercially sensitive data.

2. Project data

2.1. Project management data

As described in D1.1, the EUREMAP project handbook, a variety of internal documentation will be generated by the project, both for publication and internal consumption only. Deliverables will be made available through the EU Open Research Repository[1] at Zenodo, while internal documents will be stored in a Microsoft 365 SharePoint EUREMAP[2] installation. Academic articles will be published in subject-appropriate Gold Standard Open Access journals. The use of open formats for documentation (e.g. Open Document Format) will be encouraged where practicable to increase the accessibility of project output. Any novel programming code developed during the project will be made publicly available through a project-specific Github repository. At the same time, any additions or enhancements to existing vocabularies, ontologies, and other metadata standards will be fed-back to the respective communities for their consideration. Further details on Licensing are provided below.

2.2. Workflow data

The proposed EUREMAP marine bioprospecting workflow will reuse many data from various sources, such as those from public (meta-)genomic databases, previous marine bioprospecting projects, and existing marine organism collections. The purpose of reusing data is to build upon existing knowledge and resources and to streamline the discovery workflow for novel marine bioactive natural products. Specifically, Pillar 2 Genomics reuses data for biosynthetic gene cluster (BGC) prediction that is available from public genomic and metagenomic assemblies in the EMBL European Nucleotide Archive (ENA)[3] repository. BGC predictions also leverage data from previous and ongoing projects such as the European Marine Omics Biodiversity Observation Network (EMO BON, EMBRC)[4] to avoid redundant genomic sequencing efforts and ensure efficient use of resources. However, data reuse cannot always and consistently be achieved because of incompatibility between data formats or insufficient contextual metadata, making integration with new data difficult. Biomass production datasets can be discarded when new biomass production and respective new datasets are required to meet stringent requirements for new high-throughput screening and bioassay development.

Given the multidisciplinary nature of the marine bioprospecting workflow, the project will necessarily generate and reuse a variety of data types and formats. The data types will vary in their encoding (binary or text), in their type (tabular, image, proprietary instrument output readings, etc.), and in their syntax (FASTA, JSON, XLSX, etc.). Various data formats are generated by the different technical components of the EUREMAP workflow. In Pillar 2 Genomics, GFF3 formatted gene annotation files of BGC predict data are derived from FASTA, FASTQ, SQN and BAM files of genomic and metagenomic data. In Pillar 3 Green

[1] EU Open Research Repository (zenodo.org): https://zenodo.org/communities/eu/

[2] EUREMAP - ÎNFRA-2023-DEV-01-04 SharePoint: https://euopenscreeneu.sharepoint.com/sites/EUREMAP

[3] European Nucleotide Archive (ENA): https://www.ebi.ac.uk/ena/browser/home

[4] European Marine Omics Biodiversity Observation Network: https://www.embrc.eu/services/emo-bon

Chemistry and Pillar 4 Marine Natural Compounds, different types and formats of chemistry-related datasets, such as open data “mzXML/mzML” formats for storage and exchange of mass spectroscopy data and proprietary instrument formats, like Thermo’s “.raw” format, are used. On tasks such as biomass production, bioassays, biological profiling, and natural product synthesis, experimental results or instrumental readings are typically recorded in XML, CSV, or TAB for tabular formats. Experimental data in the form of images is also generated using various instruments, including microscopes (e.g., Operetta High Content Imaging System, Leica SPE confocal) and in-house well/plate image capture systems that output high-resolution TIF and JPG files. These data file formats are generally well suited to or deliberately designed to cater to the generated data types. These data files are most often complemented by descriptive metadata in JSON, XML, and CSV formats to ensure compliance with FAIR principles and are linked to primary datasets through unique identifiers.

All newly generated data from the marine bioprospecting workflow will be stored by the technical service provider before deposition in an appropriate data repository by the service user. The use of electronic laboratory notebooks by workflow users is encouraged but not mandated. The expected amount of genomic and metagenomic data to be generated or re-used by the project is in the range of several terabytes. Most genomic data will be reused and is publicly available in repositories (i.e. EMBL ENA, MGnify-EBI[5]); however, a large amount of data is likely to be produced in the form of intermediate analysis files that are not likely to need data management initially. Data created from chemical analyses will be in the range of hundreds of gigabytes: chromatographic and spectroscopic methods used in the workflow generate several types of proprietary files that range in size from a few megabytes to several gigabytes, differing based on factors such as the number of scans, resolution, and duration of the experiment. Bioactivity testing, biological profiling, and bioprocessing also generate heterogeneous data types, ranging from simple and small (bytes-wise) tabular files resulting from experimental recording to many huge high-resolution image files produced by high-throughput screening and bioassay experiments. Several terabytes of data are expected, mainly due to imaging analysis, which can be compressed and filtered for handling.

3. FAIR (meta)data

To facilitate integration into the EOSC[6] Common European Data Space, data should be published following FAIR principles such that the data are Findable, Accessible, Interoperable, and Re-usable. The importance of FAIR data principles will be emphasised throughout the project, with precise

[5] MGnify – EBI: https://www.ebi.ac.uk/metagenomics

[6] EOSC Association: https://eosc.eu/

recommendations for different data types produced by the technical components of the workflow, as detailed in D12.1 of the Data Integration Roadmap. To ensure the interoperability of EUREMAP data in the wider research community, wherever possible, data management will adhere to established data and metadata standards, formats, and methodologies that facilitate data exchange and reuse within and between disciplines. This approach will ensure that the data generated by the EUREMAP project will be re-usable and replicable by researchers in academia and industry involved in marine bioprospecting. This, in turn, will help demonstrate the effectiveness of the new marine bioprospecting workflow and the state-of-the-art tools and methods employed and aid the development of new techniques for the pharmaceutical and biotechnology industries. All data and metadata generated and reused in the EUREMAP project will be identified by persistent unique identifiers (PIDs). To ensure the standardised assignment of PIDs to each dataset, data will be deposited in thematic and stable repositories as soon as possible after the data are generated. (As an aside, from a data management perspective, and subject to restrictions imposed by the handling of sensitive data, it is highly beneficial to have data and its associated metadata published to appropriate repositories in the earliest possible timeframe: this ensures data are not lost, and places the emphasis on providing the appropriate contextual metadata at the time of data generation when it is most likely to be apparent.) Assigned PIDs can vary from the commonly used Digital Object Identifiers (DOIs) for research outputs (e.g., publications) to accession numbers for sequence data assigned in repositories such as NCBI GenBank[7] and EMBL-ENA. Descriptive metadata of core governance features, for example, creator (i.e. ORCID[8]) and institutional (i.e. ROR[9]) identifiers, will be included to meet the minimal metadata requirements of the EU Open Research Repository at Zenodo but also include as rich and extensive as possible metadata on provenance and methodology of the data generation procedure. Data and metadata associated with research output datasets will adhere to community standards, controlled vocabularies, and ontologies that facilitate machine-to-machine harvesting and disambiguation of meaning, such as Dublin Core, schema.org, and specific repository standards. Having the data and metadata structured, identified with PIDs, described with controlled vocabularies, and deposited in repositories with API and query services to index and harvest data will ensure the findability and reusability of project data.

[7] GenBank (nih.gov): https://www.ncbi.nlm.nih.gov/genbank/

[8] ORCID: https://orcid.org/

[9] Research Organization Registry (ROR): https://ror.org/

3.1. Open, community-driven, (meta)data standards

The EMBL-EBI BioSamples[10] repository will act as the main infrastructure for the deposition of biological sample metadata records, enabling interoperability and connection to experimental data throughout the workflow. This interoperability extends not only to internal EUREMAP data but also facilitates integration with external data housed in EBI databases. To effectively use EMBL-EBI BioSamples as a data management mechanism for "FAIRification" of sample metadata, metadata should be uploaded to the repository as soon after data collection as possible to facilitate validation and metadata brokering. It is anticipated that, if necessary, workflow users would register and create accounts on the EMBL-EBI BioSamples portal, thereby gathering and standardising all the required information about a biological sample and ensuring the use of controlled vocabularies and ontologies. This will enable the integration of sample metadata in the BioSamples repository with the data outputs of technical components of the workflow, while maintaining alignment with FAIR principles by ensuring the existence of comprehensive metadata, standard formats, and data accessibility.

A second data type for which a community-driven data framework is readily identifiable and currently used by EUREMAP bioprospecting service providers is the ReDU[11] (meta)data sample framework designed for mass spectrometry-based data. ReDU sample information metadata will be created and used to find and reuse public tandem mass spectrometry data. Such rich metadata with a community-driven controlled vocabulary will allow interoperability between the Global Natural Product Social Molecular Networking Analysis Platform (GNPS)[12] and MassIVE[13], a public data repository for mass spectrometry data. ReDU sample information metadata enables users to select public data for re-analysis in GNPS or to conduct co-analysis with their own data sets. Mass spectrometry data of the various proprietary formats will be converted to open formats such as ".mzXML/.mzML" (metadata embedded) and deposited in the MassIVE repository as an open data asset. The ReDU sample information template will be filled with rich and structured metadata descriptions, validated, and uploaded to the corresponding MassIVE accession. For mass spectrometry datasets of proteomics components, ProteomeXchange (PX)[14] data submission guidelines are recommended to provide globally coordinated standard data submission

[10] BioSamples < EMBL-EBI: https://www.ebi.ac.uk/biosamples/

[11] Reanalysis of Data User Interface for MS2 (ReDU): https://redu.ucsd.edu/

[12] GNPS - Analyze, Connect, and Network with your Mass Spectrometry Data: https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp

[13] MassIVE (ucsd.edu): https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp

[14] Proteomexchange PX: https://www.proteomexchange.org/

and dissemination pipelines involving the main proteomics repositories and adhere to open data policies in the proteomics field.

As an example of community-endorsed best practices, the EUREMAP project will use the "minimum amount of information about X" set of sequence metadata specifications required during ENA data registration that conform to a defined checklist of expected metadata values. The most suitable checklist for data registration depends on the type of data. Checklists are derived from the Genomic Standards Consortium minimum information about any (x) nucleotide sequence (i.e. MIxS) which is the core standard with sub-field checklists for describing genomic (i.e. MIGS), metagenomic (i.e. MIMS) data, and marker sequences (MIMARKS).

3.2. Provenance, controlled vocabularies, and interoperability

An entire chain of provenance (i.e., the origination of the data in question) should always be readily identifiable from the descriptive metadata of a particular data asset. In all component steps of the marine bioprospecting workflow, not only will minimum standards for metadata regarding provenance and interoperability, as defined by the community, be adhered to, but additional record descriptors will be included that enhance and provision for the downstream workflow analyses. These extra fields will be added to the metadata checklist to enrich the data, improve interoperability, and add value to the bioprospecting pipeline. Discussion between Pillar participants will be promoted to highlight relevant metadata for downstream procedures that might not seem relevant for a particular component service provider to report but for which metadata may enhance a downstream workflow analysis and ensure reproducibility and disambiguation of data provenance.

Using standardised controlled vocabularies for all metadata and data formats is essential for maintaining consistency and enabling interoperability, especially when it relates to machine-based discovery. A controlled vocabulary is an "organised arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching"[15] within a defined scope or domain. Preferred and variant terms are also usually specified. A controlled vocabulary can be a common "broad sense" specification to enable machine-readability, such as those used in information systems to organise concepts into hierarchical and/or associative relationships. Alternatively, a controlled vocabulary can be more domain-specific, such as the BioAssay Ontology (BAO)[16] that describes chemical biology screening assays and their results, including high-throughput screening (HTS) data to categorise

[15] Harpring, P., 2010: ISBN 978-1-60606-027-8

[16] BioAssay Ontology | NCBO BioPortal (bioontology.org): https://bioportal.bioontology.org/ontologies/BAO

assays and data analysis. A catalogue of relevant ontologies for the purpose of EUREMAP bioprospecting metadata and data product annotations will be constructed and adopted where relevant.

The interoperability of EUREMAP data assets will be enhanced through developing new data access endpoints facilitating machine-to-machine discovery and re-use. These will feed directly into the EOSC data ecosystem and provide the relevant entry points to the EUREMAP data catalogue for domain-specific data brokers such as Blue-Cloud[17] and FAIR-EASE[18]. This work will include the development of new data descriptors using RDF ontologies and triples stores, thereby providing SPARQL query endpoints to data brokers that are made available through data access APIs. Gaps, deficiencies, and opportunities identified through detailed analysis of the M11.1 “Survey of the current data management practices”, combined with additional direct consultation with EUREMAP service providers, will identify the precise data assets that a thorough and enhanced interoperability plan could provide the most direct benefit to the marine bioprospecting community. Provision of these interoperability enhancements will be scheduled within the Task 12.1 Generation of data services roadmap.

3.3. Licensing and persistent identifiers

Following FAIR principles and to enable machine-actionable discovery, reusability, and reproducibility, data generated during the workflow will be promptly deposited in the appropriate designated data repository and adhere to the respective repository's license. All other non-sensitive data and research outputs produced by the EUREMAP consortium will be accessible through a free and standardised access protocol and placed under the most recent version of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)[19] or a license with equivalent rights, in line with the principle of data being 'as open as possible as closed as necessary'. The metadata of deposited publications will be open under a Creative Commons Public Domain Dedication (CC 0)[20], complying with the research data management requirements under Article 17 of the Grant Agreement. While this licensing is designed to ensure open access through free and standardised access protocols, the EUREMAP consortium GA may determine that providing open access to sensitive data would contravene the beneficiary's legitimate interests due to commercial partnerships aimed at commercial exploitation, as foreseen in the Pillar 6 Valorisation & Industry Engagement.

[17] Blue-Cloud 2026: https://blue-cloud.org/

[18] Fair-Ease: https://www.fairease.eu/

[19] CC BY-SA 4.0 Deed | Creative Commons: https://creativecommons.org/licenses/by-sa/4.0/

[20] CC0 1.0 Deed | Creative Commons: https://creativecommons.org/publicdomain/zero/1.0/

Microsoft 365 SharePoint will be used to organise, share, and archive EUREMAP internal support documentation and collaborative documents, making them accessible to the whole consortium. The EU-OS partner is responsible for monitoring and maintaining the EUREMAP SharePoint store. Finalised documents (deliverable reports, presentations, posters, etc.) will be converted from modifiable formats to PDF to be deposited in a EUREMAP community profile in the EU Open Research Repository (zenodo.org). The EUREMAP community will provide a space to share research outputs (e.g. data, software, posters, presentations, publications) and project deliverables. This simplifies compliance with open science requirements and ensures free publishing and access to research outputs, sharing data and materials, fostering transparency and reproducibility in the EUREMAP outcomes. Scientific peer-reviewed publications will be published in open-access publishing peer-reviewed journals with the requirement of meta(data) standards that align with FAIR principles (e.g. Open Research Europe).

3.4. Core data repositories

Having rich structured metadata available in repositories, with access via open, standard communication protocols (or via authentication when necessary), will help ensure the findability and accessibility of EUREMAP data. The experimental datasets produced by the EUREMAP bioprospecting workflow will be deposited in thematic, trusted repositories that will ensure that the meta(data) are assigned with PIDs and URIs. Mirroring the heterogeneity of the data types that the bioprospecting workflow can create, several thematic and field-specific repositories could be used for data deposit: below are listed and briefly described several repositories central to EUREMAP data management:

The BioSamples database at EMBL-EBI is the central European repository for biological sample metadata storage and provides connectivity to additional EMBL-EBI resources. All EUREMAP biological sample metadata will be deposited in an EUREMAP-specific BioSamples project catalogue as soon as reasonably possible after data collection.

The PRIDE PRoteomics IDEntifications (PRIDE)[21] Archive database is a centralised, standards-compliant, public-facing data repository for mass spectrometry proteomics data. Mass spectrometry-based proteomics datasets are submitted to ProteomeXchange via PRIDE and are handled by expert bio-curators. All PRIDE public datasets can also be searched in ProteomeCentral, the portal for all ProteomeXchange datasets. Other Mass spectrometry-based omics, such as metabolomics, lipidomics, etc., are to be deposited in MassIVE, a mass spectrometry data repository designed to store, browse, redistribute, re-analyse, and integrate all publicly available mass spectrometry data.

The EMBL-EBI European Nucleotide Archive (ENA) is an open platform for managing, sharing, integrating, archiving, and disseminating biological sequence data. It is both a database of biological

[21] PRIDE - Proteomics Identification Database (ebi.ac.uk): https://www.ebi.ac.uk/pride/

sequence records and a platform for data management of sequence data. Sequence data created by EUREMAP, namely bacterial genome sequencing, will be submitted to ENA and linked to the respective Biosamples metadata records in the project profile. Besides submitting data, EUREMAP sequence-based studies for bioprospecting purposes will also utilise nucleotide sequence data available in ENA or secondary service providers (such as UniProt, RNAcentral, EBI Metagenomics, Ensembl, Ensembl Genomes, ArrayExpress) that are built on ENA services and content. Pillar 2 Genomics (EMBL) reuses data for BGC prediction from public genomic and metagenomic assemblies in the EMBL-EBI ENA and MGnify repositories and, in turn, experimentally characterised BGCs downstream in the pipeline in pillar 3 and pillar 4 are to be deposited in the MIBiG[22] repository.

4. Other research outputs

Besides appropriate data management aligned with FAIR principles, a key aspect is the link between digital and physical objects (e.g., samples, biological collection, chemical compounds, etc.). EUREMAP participants with microbial and other biological collections, biobanks, and natural products libraries have their physical object managing system with sample key IDs recorded on the metadata of any eventual analysis that creates a dataset. Some standardised sampling identifiers initiatives exist, such as ToLID - Tree of Life Identifiers (sanger.ac.uk)[23]; registered ToLIDs identifiers provide species recognition, differentiate between specimens of the same species and add some taxonomic context. Having structured and standardised sample identifiers facilitates EUREMAP internal records and helps internal and external communication about the samples. Any new marine-derived product produced will be maintained and managed within EUREMAP participants' biological and compound collections with appropriate records to track provenance up to the sampling from produced metadata/data deposited in the repositories.

5. Allocation of resources

The costs for making data and other FAIR research outputs in the EUREMAP project will encompass direct and indirect expenses. Direct costs will include fees for data storage solutions, archiving systems, security measures to protect data integrity, and long-term data preservation, which the majority will be ensured through a combination of institutional repositories and established data archives as public service. Several participants will create, use, and reuse the EUREMAP data, which have in-house IT

[22] MIBiG repository: https://mibig.secondarymetabolites.org/repository

[23] ToLID - Tree of Life Identifiers: https://id.tol.sanger.ac.uk/

support and resources necessary for storage solutions and secure backup systems to prevent data loss of any EUREMAP data being handled temporarily in in-house infrastructures. Indirect costs will involve personnel time for data management activities, staff training on FAIR principles, and potential software development/upgrades to facilitate data handling and preservation. These indirect costs will be funded as part of the EUREMAP grant, under the grant budget for person/months for data management activities described in the Pillar 5 Data Management & tools work packages.

6. Data security

The repositories where data are stored are responsible for their long-term safe storage and security. The SharePoint access and EUREMAP EU open research repository will be accessed with authentication and access through HTTPS.

7. Ethics

The data generated by the participants' implementation of the EUREMAP pipeline should not raise ethical or legal issues. The ethical committee will discuss potential ethical issues. Any legal issue not covered by the general grant agreement, non-disclosure agreement, or consortium agreement and in conflict with the participants' declaration of honours will be handled by the general assembly of the EUREMAP consortium.

Personal data is processed by Directive 95/46/EC and Regulation (EU) 2016/679 of the European Parliament and European Council. No questionnaires regarding personal data will be conducted within the scope of EUREMAP. Each participant's personal contact information and name will be included in the EUREMAP contact list and mailing list and described in planned surveys to obtain capacity roadmaps as responsible participant contact details.

8. Other issues

This document was adapted from the EU Grants: Data management plan (HE): V1.1 – 01.04.2022 template and complies with the research data management requirements under Article 17 of the Grant Agreement. In addition to the initial deliverable, updated versions will be regularly uploaded to the Portal Grant Management System.

Table 1. History of changes.

Version	Publication date	Change
1.0	08-07-2024	Initial Version