29/08/2025
Version 1.1

BMD
Biodiversity Meets Data

D8.2 BMD Data Management Plan

Author(s): Sharif Islam, Niels Raes (Naturalis)

Co-funded by
the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

BMD (Biodiversity Meets Data) receives funding from the European Union's Horizon Europe Research and Innovation Programme and the Swiss State Secretariat for Education, Research and Innovation (SERI) (ID No 101181294). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, the European Research Executive Agency (REA) or SERI. The EU, REA and SERI cannot be held responsible for them.

Prepared under contract from the European Commission
Grant agreement No. 101181294
EU Horizon Europe Research and Innovation Action

Project acronym:	**BMD**
Project full title:	**Biodiversity Meets Data**
Project duration:	01/03/2025 – 28/02/2029 (48 months)
Project coordinator:	Stichting Naturalis Biodiversity Center (Naturalis)
Call:	HORIZON-CL6-2024-BIODIV-01
Deliverable title:	Data Management Plan
Deliverable n°:	D8.2
Work package:	8
Nature of the deliverable:	DMP
Dissemination level:	Public
Licence of use:	Creative Commons Attribution 4.0 International (CC BY 4.0)
Lead beneficiary:	Naturalis
Recommended citation:	Islam, S. & Raes, N. (2025). BMD Data Management Plan. BMD project deliverable D8.2.
Due date of deliverable:	31/08/2025 (M6)
Actual submission date:	29/08/2025 (M6)
Quality review:	Yes

Deliverable status:

Version	Status	Date	Author(s)	Actions
1.0	Release	28/07/2025	Sharif Islam, Niels Raes (Naturalis)	Sent for review
1.0	Release	15/08/2025	Claus Weiland (SGN), Taimur Khan (UFZ)	Reviewed
1.1	Release	28/08/2025	Sharif Islam, Niels Raes (Naturalis)	Finalised, with incorporation of feedback from reviewers
1.1	Final	28/08/2025	Niels Raes (Naturalis)	Submitted

Table of contents

Executive summary

The Biodiversity Meets Data (BMD) project is a 4-year EU Horizon Europe project (2025-2029) led by Naturalis Biodiversity Center that aims to deliver a Single Access Point (SAP) that provides natural resources managers and policy makers access to: a) metadata catalogue that provides access to all relevant EU biodiversity data and spatial data that determine the distribution of species and/or represent drivers of biodiversity changes, b) high-throughput biodiversity monitoring tools, c) tools to mobilise historical baseline data, d) Virtual Research Environments (VREs) for the terrestrial, freshwater and marine domains to analyse and predict the distribution of species and to identify drivers of biodiversity change, and e) a web-GIS data viewer to visualise and consult the data and results of the VREs. By integrating historical baseline data with near real-time biodiversity monitoring data and providing standardised access to data and tools to analyse the data and monitor and manage natural resources across Europe, BMD will support more effective implementation of EU Nature Directives[1], the Biodiversity Strategy for 2030[2] and the EU Green Deal[3]. This document, the BMD Data Management Plan (DMP), required under Horizon Europe, includes policies on the publication, storage, persistence and accessibility of all data generated or reused by project partners, under the practices of FAIR and open science. The document follows EU Grants: Data management plan (HE):V1.1 – 01.04.2022 template as provided in the EU Funding & Tenders Portal.

Keywords
Data Management Plan, FAIR data, FAIR workflows, Metadata, Biodiversity, Open Access, Natura 2000

[1] https://environment.ec.europa.eu/topics/nature-and-biodiversity_en
[2] https://environment.ec.europa.eu/strategy/biodiversity-strategy-2030_en
[3] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

List of abbreviations

ABCD	Access to Biological Collection Databases
AI	Artificial Intelligence
API	Application Programming Interface
BMD	Biodiversity Meets Data
CA	Consortium Agreement
CORDIS	Community Research and Development Information Service
DestinE	Destination Earth
DCAT	Data Catalog Vocabulary
DMP	Data Management Plan
DwC-A	Darwin Core Archive
EC	European Commission
EML	Ecological Metadata Language
ENA	European Nucleotide Archive
EO	Earth Observation
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
FDO	FAIR Digital Objects
GBIF	Global Biodiversity Information Facility
GDPR	General Data Protection Regulation
GDDS	Green Deal Data Space
INSPIRE	Infrastructure for Spatial Information in Europe
OBIS	Ocean Biodiversity Information System
OGC	Open Geospatial Consortium
PID	Persistent Identifiers
SAP	Single Access Point
SDM	Species Distribution Model
STAC	SpatioTemporal Asset Catalogs
TDWG	Biodiversity Information Standards - formerly known as the Taxonomic Databases Working Group (TDWG)
VRE	Virtual Research Environment
Web-GIS	Geographic Information System (GIS) that uses web technology to make spatial data accessible

1. Introduction

In line with the objectives of the EU Biodiversity Strategy for 2030 (European Commission 2021), the EU-funded Biodiversity Meets Data (BMD) project will deliver a Single Access Point (SAP) for high-throughput biodiversity monitoring tools, mobilisation of historical baseline data, a spatial environmental data catalogue, and a suite of Virtual Research Environments (VREs) supporting terrestrial, freshwater, and marine domains. These resources, co-designed with stakeholder communities, (primarily Natura 2000 site managers and relevant policy makers[4]) will enable integrated monitoring, analysis of drivers of change, and projections of climate and land cover change impacts on species and habitats. By providing harmonised and standardised access to data and tools, BMD will support more effective implementation of the EU Nature Directives.

Achieving these objectives depends on the integration, mobilisation, and FAIR management of diverse data sources, including historical biodiversity records, real-time sensor data, and spatial environmental layers. Transparent, interoperable, and well-documented data practices are essential to ensure that BMD's outputs can be reused, cited, and trusted. This Data Management Plan (DMP) sets out how BMD will handle data according to FAIR principles and Horizon Europe requirements, detailing practices for data reuse, metadata management, and publication. It also explains how federated access, a GeoNetwork-based metadata catalogue, and established standards (e.g. TDWG, INSPIRE, OGC) will underpin BMD's design and ensure sustainability of the project's data outputs.

2. Data Summary

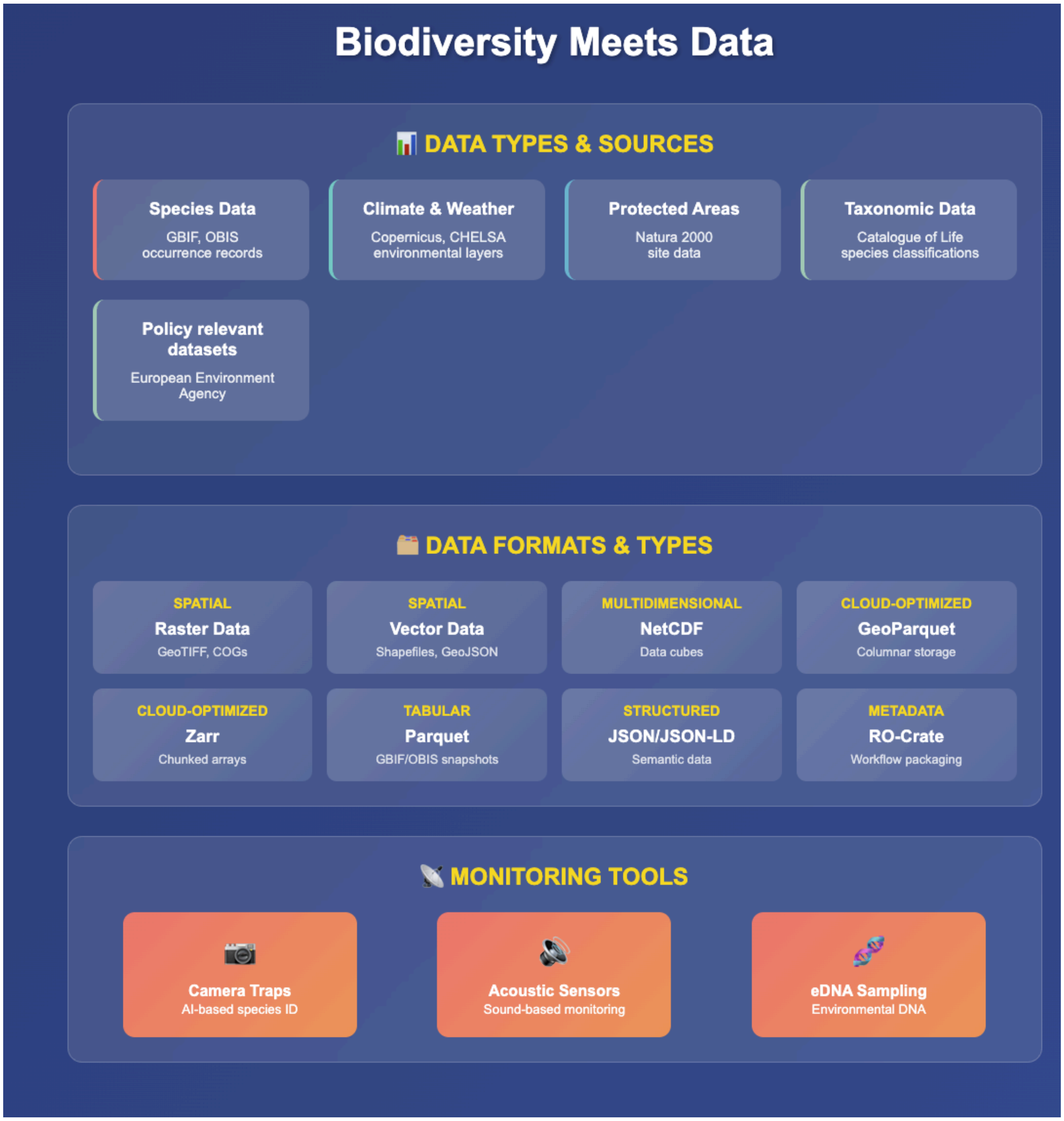
BMD will re-use species occurrence data (primary sources are GBIF- and OBIS-mediated datasets), climate and weather data (e.g., Copernicus, CHELSA), and existing datasets from Natura 2000 sites and the European Environment Agency (EEA). Re-use is essential for building data cubes, and the project will strive to maximise re-use wherever possible. BMD Work Package 2 (WP2) is developing a GeoNetwork data catalogue that will list all existing data sources and related metadata to support FAIR workflows (Leo et. al 2024). This catalogue will be linked to the project's SAP.

Existing data will primarily be accessed via standard APIs, Python libraries, and cloud-native, S3-compatible object stores. Re-use will depend on data access modes: some datasets will be consumed on-demand via federated querying or APIs, while others may be mirrored or copied locally when required for performance, integration, or reproducibility. This follows a “federation-first, mirror-if-needed” design principle. Federation is preferred to respect authoritative data stewardship and to reduce duplication. In any case, attribution of data is guaranteed through FAIRification via the GeoNetwork data catalogue.

Re-use is also critical for integrating and harmonising fragmented biodiversity data sources. It enables the enrichment of biodiversity datasets with geospatial and environmental variables and supports the development of interoperable infrastructures aligned with GBIF, EU research infrastructures (RIs), and initiatives such as GDDS, DestinE, and EOSC. The primary data formats BMD will work with include raster

[4] For more details see Wooldridge, T., Mroz, W., Alonso Vizcaino, E., Vidal, M., & Hollingsworth, P. (2025). Biodiversity Meets Data (BMD) MS2 Project-wide published stakeholder engagement plan (1.3). Zenodo. https://doi.org/10.5281/zenodo.16985228

and vector data, NetCDF for multidimensional data cubes, GeoParquet and Zarr for cloud-optimised storage, as well as tabular formats, GBIF/OBIS Parquet snapshots, JSON, JSON-LD for semantic serialisation, and RO-Crate for metadata, workflow packaging, and provenance. See Figure 1 for a visual schema of the data types and file formats.



Biodiversity Meets Data

📊 DATA TYPES & SOURCES

Species Data
GBIF, OBIS occurrence records

Climate & Weather
Copernicus, CHELSA environmental layers

Protected Areas
Natura 2000 site data

Taxonomic Data
Catalogue of Life species classifications

Policy relevant datasets
European Environment Agency

📁 DATA FORMATS & TYPES

SPATIAL
Raster Data
GeoTIFF, COGs

SPATIAL
Vector Data
Shapefiles, GeoJSON

MULTIDIMENSIONAL
NetCDF
Data cubes

CLOUD-OPTIMIZED
GeoParquet
Columnar storage

CLOUD-OPTIMIZED
Zarr
Chunked arrays

TABULAR
Parquet
GBIF/OBIS snapshots

STRUCTURED
JSON/JSON-LD
Semantic data

METADATA
RO-Crate
Workflow packaging

📡 MONITORING TOOLS

📷
Camera Traps
AI-based species ID

🔊
Acoustic Sensors
Sound-based monitoring

🧬
eDNA Sampling
Environmental DNA

Figure 1:BMD data ecosystem highlighting some major data sources. The GeoNetwork catalogue that will be used in the project will list all the datasets actively used within BMD.

BMD directly addresses the expected outcome of the call by advancing high-throughput biodiversity monitoring, essential for the effective implementation of the EU Nature Directives. The project will provide Natura 2000 site managers with plug-and-play solutions for deploying camera traps, acoustic sensors, and eDNA sampling tools, all supported by AI-based taxon identification services. Data pipelines will be tailored to terrestrial, freshwater, and marine realms, ensuring biodiversity data shared with global infrastructures such as GBIF, OBIS, and ENA, while respecting restrictions or encryption for sensitive species.

In parallel, BMD will develop practical guidelines for mobilising historical baseline biodiversity data to FAIR-aligned repositories, integrating both new and historical datasets within its VREs. This will enable trend analysis, identification of biodiversity loss drivers, and support science-based conservation actions. By linking the WP2 data catalogue of biodiversity and environmental resources with interoperable, space-time-taxonomy harmonisation from WP3, and delivering these datasets via scalable, compute and workflow management resources from T4.2, BMD will provide a unified processing environment. As visualisation and mapping are central to the BMD data ecosystem, the SAP and VREs will offer user-friendly, interactive data viewers (similar in concept to Jupyter Notebooks or R Shiny apps) enabling intuitive exploration and analysis of multidimensional datasets for a diverse user base.

Expected size of the data will depend and refined based on the use cases and stakeholders requirements. Based on the initial investigation a full integrated data space, including species occurrence, geospatial and high throughput data and derived data, EO products and derived models, is expected to reach several hundreds of terabytes (100-200TB range)[5], depending on the resolution and number of regions processed. Again, for how long we will store these data will also depend on the use case and policy.

The data used and generated in the BMD project originates from a wide range of authoritative European, and national sources. These include biodiversity data aggregators, Earth observation platforms, in-situ monitoring infrastructures, and outputs from VREs. In addition to reused datasets, BMD will produce internally derived data such as integrated data cubes.

To ensure transparency and traceability, BMD will maintain a central metadata catalogue using GeoNetwork (Task 2.1.2). This catalogue will:

- Document the provenance of all reused and generated datasets
- Align metadata records with INSPIRE, EML, and DCAT standards
- Support keyword indexing, descriptions, and metadata harvesting
- Serve as a discoverability layer for internal and external users

This approach ensures that all data sources (whether de novo generated, reused, derived, or aggregated) are properly described, findable, and connected to their original context for:

- Managers of natural resources across Europe
- EU and national policy actors (e.g., EEA, ministries)
- Biodiversity researchers and conservation planners
- Earth observation and digital twin communities

[5] The current GBIF occurrence snapshot is around 200GB. However, earth observation and climate data sizes are large. For instance, data from CHELSA (see https://doi.org/10.48364/ISIMIP.836809) is 1.2TB.

- AI model training
- Institutions developing biodiversity indicators and scenarios

3. FAIR data

3.1. Making data **F**indable - including provisions for metadata

All publishable datasets generated by the BMD project will be assigned persistent identifiers (PIDs) such as DOIs, issued through DataCite, Zenodo, GBIF, or OBIS, depending on the data type and domain. Intermediate workflow outputs will be internally versioned and tracked to ensure provenance and reproducibility throughout the data lifecycle.

Metadata will comply with established standards, including the INSPIRE Metadata Implementing rules based on ISO 19115/19119, DCAT, and domain-specific schemas such as Darwin Core and EML. Additional contextual and provenance metadata will be packaged using RO-Crate, supporting rich, machine-actionable discovery and linking across workflows, datasets, and software components.

All metadata records will include searchable keywords based on domain vocabularies and INSPIRE specifications, enhancing thematic and cross-disciplinary discoverability. Metadata will be published and maintained through a GeoNetwork-based catalogue, which will be integrated with BMD's SAP for seamless dataset discovery.

The system architecture has been designed to support metadata harvesting, full-text indexing, and SQL-style query capabilities, enabling advanced search, filtering, and interoperability.

3.2. Making data **A**ccessible

3.2.1. Repository

The BMD project will deposit data in trusted, community-recognised repositories, depending on the nature and type of the data. Wherever possible, existing domain-specific repositories will be used to ensure interoperability, visibility, and long-term preservation.

Examples include:

- GBIF/OBIS for species occurrence and biodiversity datasets
- The European Nucleotide Archive (ENA) for sequence-based data
- Zenodo general research data, software, deliverables and documentation
- Hugging Face for AI models and datasets
- Based on the use case and stakeholder feedback, DestinE Data Lake or Copernicus Data Space Ecosystem for EO and derived environmental data where integration is appropriate

Repository selection will be based on data type, licensing, disciplinary standards, and sustainability of access. All repositories used will support persistent identifiers and metadata harvesting in line with FAIR principles.

BMD will use established, trusted repositories where some technical workflows are already in place that can be re-used. These repositories, such as GBIF, ENA, and Zenodo, ensure that deposited datasets are assigned persistent identifiers (e.g. DOIs) and that these identifiers resolve to accessible, citable digital objects with appropriate metadata.

3.2.2. Data

Most datasets produced by the BMD project will be made openly available under permissive licenses such as CC0 or CC-BY. Where reused datasets are subject to more restrictive licensing (e.g. CC-BY-NC), this will be clearly indicated in the metadata and accompanying documentation. All licensing terms from third-party providers will be fully respected. In cases where legal or regulatory constraints apply (such as the protection of sensitive sites or threatened species) BMD will follow relevant national and international guidelines. In such instances, access to data may be restricted, encrypted, blurred, replaced by metadata-only records, or made available under authenticated or conditional access.

This approach aligns with established best practices in biodiversity data publishing, where sensitive datasets remain discoverable through metadata, while access to the actual records is controlled. Where applicable, technical mechanisms will be implemented to flag restricted datasets, ensuring compliance, transparency, and discoverability.

To facilitate the use of sensitive species data in the BMD VRE workflows, the species occurrence data can be added to the workflows as long as biodiversity data standards are adhered to. These include the Darwin Core or ABCD data formats.

Based on current knowledge, no embargo period is foreseen for data publication. All datasets will be accessible via HTTP/HTTPS, APIs, or S3-compatible protocols. An Authorisation and Authentication Infrastructure (AAI) will be implemented to provide users of the BMD SAP with personalised access to their own data and VREs. User profiles will be linked to this AAI system.

At present, there is no need for a data access committee. The project has finalised a Project handbook (Deliverable 8.1) which outlines project guidelines, including reference to an Ombudsperson for Ethics, Diversity, Equality and Inclusivity (DEI Champion) and good practice principles for engaging with stakeholders.

3.2.3. Metadata

All metadata will be published under the CC0 license. Metadata will contain dataset references and access information. BMD data and metadata will remain accessible for a minimum of 5 years after finalising the project. The extent of this of course will depend on the feasibility and capacity of the organisation involved. Code, scripts, and documentation will be published via GitHub, WorkflowHub, Zenodo (with DOIs). Software dependencies and library details will be documented as part of best practice.

3.3. Making data Interoperable

BMD will follow community-endorsed interoperability standards and vocabularies to ensure that data can be exchanged and re-used both within biodiversity research and across related domains such as geospatial analysis and Earth Observation.

We will adopt established standards developed and maintained by TDWG, including:

- Darwin Core (DwC) for species occurrence and sampling event data

- Camera Trap Data Package (CamTrap DP) for sensor-based wildlife observations
- GBIF and Catalogue of Life vocabularies for taxonomy, identification, and metadata alignment
- Ecological Metadata Language (EML) for dataset-level metadata

We will use standards from the INSPIRE Directive, particularly for spatial data themes, and align with relevant specifications from the Open Geospatial Consortium (OGC), such as: GeoTIFF, NetCDF, and Cloud Optimized GeoTIFFs (COGs) for spatial raster data. OGC API, WMS/WFS, and STAC for data access and discovery.

To enable broader interoperability and FAIRness, especially for machine-actionable workflows and metadata packaging, BMD will adopt: RO-Crate for encapsulating datasets, software, and workflows. Schema.org, SKOS to support structured, linkable metadata. Persistent identifiers (PIDs) and machine-readable metadata aligned with FAIR Digital Object (FDO) approach.

Custom mappings between biodiversity, EO, and monitoring ontologies will be generated and published. If new terms are needed, these will be submitted to relevant standards bodies.

3.4. Increase data **R**e-use

All datasets will be accompanied by structured metadata, README files, and additional documentation as needed to ensure transparency and reusability. These will include information on:

- Methodology and data collection context
- Variable definitions and units of measurement
- Known limitations or caveats

For internally generated data and workflows, outputs will be packaged using RO-Crate, which encapsulates not only the data and metadata but also links to associated software specifications, workflow definitions, and environment configurations.

VRE outputs will be documented with input data, linked provenance, workflow details, and code dependencies to support reproducibility and validation by third parties. These outputs (depending on the type) can also be submitted to existing trusted repositories. Datasets will use CC0 or CC-BY whenever possible. Third-party data will retain original licensing. Project outputs will follow standard citation formats and attribution practices (see GBIF citation guidelines). Data will be retained in public repositories with persistent identifiers. Outputs will be referenced in publications and available via metadata catalogues.

BMD will ensure that data provenance and lineage are thoroughly documented using appropriate, machine-readable standards. RO-Crate will serve as the primary packaging mechanism for workflows, allowing datasets, workflow definitions, and software dependencies to be bundled together with structured metadata and provenance information. Provenance, as defined in the W3C PROV model, records the entities, activities, and agents involved in producing or modifying data, supporting transparency and reproducibility. By adopting PROV-aligned metadata, BMD will make it easier to trace data origins, transformations, and usage.

Git-based versioning (e.g. via GitHub or institutional Git repositories) will be used to track changes in source code, configuration files, and pipeline logic, supporting transparency and reproducibility throughout the development process. Once complete, curated workflows (along with their RO-Crate metadata) will be published via platforms such as WorkflowHub (Gustafsson et. al 2025), which are designed to support FAIR sharing and indexing of computational workflows. This combination ensures that both the evolution and execution context of BMD workflows are clearly documented and openly accessible.

4. Other research outputs

In addition to data, BMD will generate key digital outputs such as software (Python/R scripts, notebooks), RO-Crate-packaged workflows and results, and cloud-optimized data cubes. These outputs will be managed following FAIR principles wherever possible, ensuring they are discoverable, reusable, and citable. Documentation, versioning, and licensing practices will be applied consistently, with workflows and code shared via repositories like GitHub and Zenodo to support transparency and reuse across the broader research community.

4.1. Product specification documentation

For major outputs of the BMD project, including BMD’s metadata catalog, the datacubing engine, the VREs and SAP, product specification documents will be developed. The product specification documents serve as comprehensive, blueprint-like documents that detail the product's features, data requirements, functionality, design, technical requirements, and performance criteria.

A product specification typically includes:

- Product Vision & Goals: What the product is, why it's being built, and what it needs to achieve for the stakeholders/users.
- Features & Functionality: Detailed descriptions of the product's capabilities and how it will work.
- Design Specifications: How the product should look and user interface details.
- Technical Requirements: The technology stack for software, or materials for physical products.
- Performance Criteria: Standards for how the product should perform and its quality parameters.
- Compliance & Data standards: GDPR, sensitive species policy, spatial data and biodiversity data standards
- Target Users: Information about the specific user/stakeholder personas the product is designed for.

5. Allocation of resources

The costs for making BMD data and other outputs FAIR are difficult to precisely estimate at this stage, particularly given the modular, federated architecture and the variety of data types and workflows involved. Direct costs for repository deposition (e.g. DOIs, storage in Zenodo) are relatively easy to predict; however, costs associated with continuous metadata generation, RO-Crate packaging, and provenance tracking across distributed VREs are less clear, especially as volumes and stakeholder needs evolve.

Institutional contributions, varying repository policies, and support from existing infrastructures (e.g. LifeWatch ERIC, GBIF, ENA, EOSC nodes) will also affect total costs, making initial projections approximate. To address this uncertainty, BMD will track data mobilisation, processing, and deposition activities throughout the project, allowing post-hoc estimation of actual data management costs for core workflows.

6. Data security

Data security in BMD follows a distributed responsibility model, where each participating organisation is responsible for providing and managing its own cloud resources, including compute and storage infrastructure. All partners apply their institutional data security policies, ensuring compliance with national and EU regulations regarding data protection and storage.

For sensitive or access-restricted data, secure storage and controlled access can be managed at the institutional level, using encrypted storage, role-based access controls, and secure transfer protocols if needed (e.g. HTTPS, S3 with token-based authentication). Data recovery and backup strategies follow institutional standards, typically including regular automated backups and multi-region redundancy.

For long-term preservation, BMD relies on trusted, domain-recognised repositories (e.g. GBIF, ENA, Zenodo), each of which maintains its own established policies for data curation, security, and preservation. Deposited datasets will inherit the data protection and access control frameworks of these repositories, ensuring their integrity and availability over time. By adopting a federation-first architecture, BMD ensures that data sovereignty and security responsibilities remain with the original providers or authoritative repositories, while still enabling controlled integration and reuse across the project ecosystem.

7. Ethics

Overall, no major ethical or legal issues are expected to arise in the BMD project, as the data will predominantly originate from open-access databases. However, BMD fully recognises the importance of handling sensitive species data with care to prevent misuse and ensure ethical, legal, and biodiversity safeguards. Such data will be blurred, encrypted or generalised, following established protocols such as those developed by GBIF (Chapman 2020).

Sensitive species data may still be included in VRE workflows through the secure upload of DwC-A files. In all cases, data protection remains the responsibility of the original data providers. Additionally, the BMD project includes a planned deliverable 'D8.7 - Ethics Assessment and Guidelines' dedicated to identifying potential ethical concerns, which will outline a strategy and action plan with appropriate mitigation measures to address any issues that may arise.

8. References

European Commission: Directorate-General for Environment, EU biodiversity strategy for 2030 – Bringing nature back into our lives, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2779/677548

Chapman, A. D. 2020. Current Best Practices for Generalizing Sensitive Species Occurrence Data. Version 1. https://doi.org/10.15468/doc-b02j-gt10

Gustafsson, O.J.R., Wilkinson, S.R., Bacall, F. et al. WorkflowHub: a registry for computational workflows. Sci Data 12, 837 (2025). https://doi.org/10.1038/s41597-025-04786-3

Leo, S., Crusoe, M.R., Rodríguez-Navas, L., Sirvent, R., Kanitz, A., De Geest, P., Wittner, R., Pireddu, L. Garijo, D., Fernández, J.M. and Colonnelli, I., 2024. Recording provenance of workflow runs with RO-Crate. PLoS one, 19(9), p.e0309210. https://doi.org/10.1371/journal.pone.0309210