30/06/2025
Version 1.0

MS19 Initial data space design with implementation plan

Author: Sharif Islam

Contributors: Mathias Dillen (MeiseBG), Ayco Holleman (Naturalis), Antonio José Sáenz (LifeWatch ERIC), Taimur Khan (UFZ), Ingolf Kühn (UFZ), Claus Weiland (SGN)

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
Milestone title:	Initial data space design with implementation plan
Milestone n°:	MS19
Means of verification:	Document
Work package:	WP4
Nature of the milestone:	Document
Contribution to deliverable n°:	D4.1
Licence of use:	This document is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. You are free to share and adapt the material, provided appropriate credit is given.
Lead beneficiary:	Naturalis
Recommended citation:	Islam, S., (2025). *Initial data space design with implementation plan*. BMD project deliverable MS19.
Due date of milestone:	Month 4
Actual submission date:	Month 4
Quality review:	Yes

Milestone status:

Version	Status	Date	Author(s)	Actions
0.1	First Draft	14 May 2025	Sharif Islam	Presented at WP4 meeting
0.2	Second Draft	30 May 2025	Sharif Islam	Send to internal reviewer for comments
0.3	Third Draft	02 June 2025	Sharif Islam	Received feedback from Mathias Dillen (reviewer) with

				incorporation of feedback from reviewers
0.4	Fourth Draft	20 June 2025	Sharif Islam	Final Edit
0.4	Fourth Draft	24 June 2025	Sharif Islam	Sent for Coordinator verification
0.4	Fourth Draft	30 June 2025	Vânia Ferreira, Niels Raes	Verified
1.0	Final	30 June 2025	Sharif Islam	Submitted

Table of contents

Summary

The [Biodiversity Meets Data](#) (BMD) project is creating a data space using cloud-native open infrastructure to support biodiversity monitoring, conservation, and policy across terrestrial, freshwater, and marine environments. This document (MS19) presents the initial architecture design for the BMD data space, led by Work Package (WP) 4 and developed in coordination with WP2 (data mobilisation), WP3 (harmonisation), WP5 (Virtual Research Environments), and WP6 (visualisation and Single Access Point).

The BMD Data Space will host harmonised, FAIR-aligned data cubes derived from high-throughput and legacy sources. These cubes will support scalable workflows, reproducible analysis, and dynamic policy reporting. Whenever possible, data will be accessed directly from source data providers, with local replication, caching used only when needed for transformation, performance, or reliability and always preserving metadata and provenance. The architecture aligns with the Green Deal Data Space (GDDS) and EOSC (European Open Science Cloud) Interoperability Frameworks, and adopts open lakehouse principles (such as composability and the separation of storage, metadata, and compute layers) to maximise flexibility and interoperability. It builds on cloud-native formats like Parquet[1], Zarr and GeoParquet, which support scalable, reusable data infrastructure across analytical environments. BMD's design adopts a dual-catalogue model (GeoNetwork for public metadata, open table catalogues for internal tracking) supports both transparency and operational efficiency.

Integration across WPs ensures a cohesive backend for user-facing services, with stakeholder needs informing cube design and access patterns. A flexible approach to data quality will allow permissive integration initially, with tighter validation introduced over time.

While infrastructure will be hosted for five years post-project, sustainability strategies remain under discussion. The architecture supports modular growth, cloud portability, and long-term FAIR data stewardship. This milestone reflects the first four months of development and sets a foundation for future implementation and stakeholder collaboration.

---

[1] [Apache Parquet](#) is an open-source, column-oriented data storage format designed for efficient data processing and analytics. It stores data in a compressed, columnar format that enables fast query performance and reduces storage costs compared to traditional row-based formats like CSV.

List of abbreviations

API	Application Programming Interface
EOSC	European Open Science Cloud
DEIMS	Dynamic Ecological Information Management System
FAIR	Findable, Accessible, Interoperable, and Reusable
GBIF	Global Biodiversity Information Facility
GDDS	Green Deal Data Space
MS	Milestone
SAP	Single Access Point
TDWG	Biodiversity Information Standards
OTF	Open Table Format
VRE	Virtual Research Environment
Web-GIS	Geographic Information Systems that employ the World Wide Web
WP	Work Package

1. Purpose and scope of the design

The MS19 document provides the initial architecture design and implementation plan for the BMD data space. It is intended to guide the technical development within WP4 while supporting cross-WP integration with WP2 (data cataloguing and mobilisation), WP3 (harmonisation), WP5 (VREs/workflows), and WP6 (Single Access Point and visualisation). It incorporates FAIR and data space design principles to ensure data governance, interoperability, scalability, and reusability. The implementation also aligns with WP4 tasks to manage dependencies on compute provisioning, visualisation infrastructure, FAIR workflow management and European Data Space interoperability concerns. BMD's design draws from ongoing work on the Green Deal Data Space development (see Blueprint - Communities - Data Spaces Support Centre), where a key motivation is enabling loosely federated systems through minimal but meaningful agreement on FAIR and open standards. This approach is also echoed in recent scholarly work (see Curry 2020, p.49), which advocates for a "loose integration" or "good enough" model where data sources coexist and are integrated only as needed, avoiding the overhead of rigid, upfront harmonisation.

The project proposal includes a commitment to host the BMD infrastructure for five years beyond the project's end, covering all core technical aspects. However, the long-term sustainability and funding model (especially for infrastructure operation and service delivery) remains an open and active area of discussion. Our architectural approach deliberately supports future migration, extension, and adaptation, keeping flexibility and modularity at its core to ensure the infrastructure can evolve with emerging needs, policies, and partnerships.

2. High-Level Overview

The goal of BMD is to deliver data, tools, and services through a Single Access Point (SAP) to support the planning and management of biodiversity and protected areas across terrestrial, freshwater, and marine realms. The process and data flow from a project and WP-specific perspective is as follows (see Figure 1): different data sources (catalogued and mobilised via WP2) are harmonised in WP3 and transformed into BMD-specific, standardised data cubes. These cubes are stored and managed in the Biodiversity Data Space (WP4), enabling computation, access, and visualisation through:

- Virtual Research Environments (VREs – WP5)
- The Web-GIS Mapviewer and SAP interface (WP6)
- APIs and downloadable endpoints as needed

While each WP provides specific focus and delivers essential components, the overall architecture is designed to go beyond individual WPs. It provides an integrated, modular, and cloud-native infrastructure that supports cross-cutting stakeholder needs. The data space acts as a unifying layer where the technical components converge, enabling a shared vision of open, FAIR, and reusable biodiversity infrastructure. WP3 plays a key role in harmonisation by aligning taxonomic, temporal, and spatial dimensions, however, data quality assessment remains an open area of work. Given the scale and heterogeneity of data involved, we anticipate starting with a more flexible integration model, gradually introducing stricter semantic validation where needed.

Key technical principles described later in the document include:
- Adoption of open lakehouse patterns (composability and the separation of storage, metadata, and compute layers), open table formats (e.g., Delta Lake), and cloud-native data formats (Parquet, GeoParquet, Zarr, etc.).
- A dual-catalogue model for metadata (GeoNetwork for public, Delta Table/SQL-based for internal)
- Use of open standards and frameworks like RO-Crate, INSPIRE STAC, and ODRL for workflows, discovery, and access policies
- Interoperability with external infrastructures and initiatives (e.g., GBIF, Copernicus, EOSC, Green Deal Data Space)
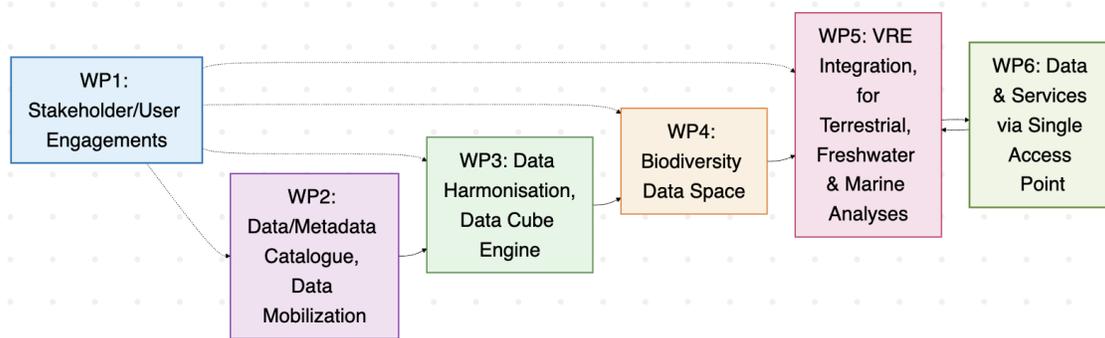

Figure 1: A WP-based look at the process and data flow

3. Key Design Principles

The proposed design principles underpin the BMD infrastructure strategy. They aim to ensure the system is efficient, secure, scalable, and responsive to user needs while aligning with FAIR and open scholarly infrastructure values. These principles also balance **functional requirements** (what the system must do, e.g., running analyses in VREs or accessing structured cubes) with **non-functional requirements** (how the system performs, e.g., speed, scalability, interoperability, or policy compliance).

- **Establish domain-specific federation and governance:** As in other domain-oriented data spaces (e.g., health, agriculture), the biodiversity domain requires specific federation and governance mechanisms. BMD will explore establishing a domain-specific data space operator role, possibly within a pan-European research infrastructure. However, this requires policy alignment and working data federation governance models, which might require support from the wider European data space initiatives.
- **Use existing data standards** and controlled vocabularies: Reuse domain ontologies, terms, and controlled vocabularies (e.g., Darwin Core, EML, INSPIRE, DCAT, Schema.org, etc.) as much as possible. This encourages cross-data space interoperability through shared technical and semantic standards.
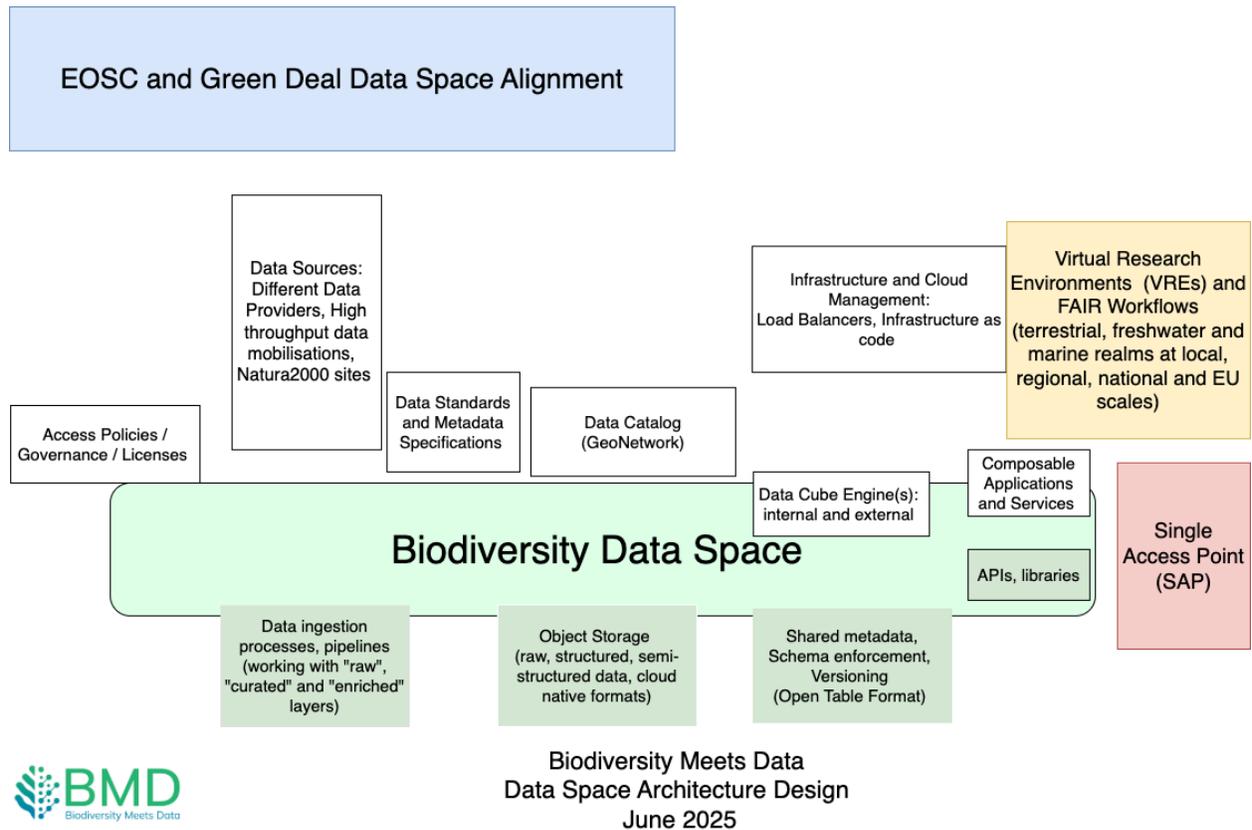
Figure 2: A component view of the design

- **Design for distributed infrastructures:** The BMD system spans several independently managed but interconnected components. For example, while the Data Space (WP4) hosts data and compute services, the SAP (WP6) and VREs (WP5) may be deployed by different partners using distinct cloud infrastructures. This requires clear agreements around hosting requirements, service level agreements, API integration, authentication, and deployment consistency. It is also important to note that biodiversity data itself is fundamentally decentralised – originating from different institutions, national agencies, and thematic initiatives that curate and maintain their own datasets. While centralised infrastructure can support coordination and shared services, the system must be designed to accommodate decentralised data governance, enabling federation and interoperability without imposing rigid central control (Sterner et al., 2020).
- **Align concepts and definitions:** BMD will adopt common definitions for key concepts like "site", "event", or "device", leveraging community standards such as TDWG's [Camera Trap Data Package](#) and [Humboldt Core Extension](#). These definitions will not only support semantic alignment but also ensure consistent implementation across components.
- **FAIR and open by design:** The implementation of FAIR principles is an evolving process varying across data providers and shaped by context, legal frameworks, and technical maturity. BMD promotes FAIRness by prioritising machine-readable metadata, persistent identifiers, and

semantic interoperability wherever possible. Achieving this requires ongoing collaboration and alignment with data providers, research infrastructures, aggregators, and repositories. Our commitment is to make datasets and services [as open as possible, and only as closed as legally necessary](), ensuring transparency and usability while respecting data sensitivity and access restrictions.

- **Support Cloud-native storage and compute:** Use formats and systems designed for horizontal scalability and efficient access (e.g., Zarr, Parquet, GeoParquet). This cloud-native strategy also accommodates auto-scaling methodologies, which can minimise the carbon footprint and save cloud resource costs.
- **Provide separation of public vs. internal metadata catalogues:** We are adopting GeoNetwork for public discovery; open table catalogue (e.g., Delta Tables) for internal tracking and metadata brokering. This separation ensures that stakeholders can easily discover relevant datasets through standard catalogue interfaces, while internal workflows benefit from version control, schema tracking, and operational flexibility.
- **Federated by default, replicated when needed:** Access source data whenever feasible; store local copies only when transformation, performance, or reliability demands it. Transformed products and outputs should be stored as needed, while always preserving metadata and provenance regardless of the storage approach.
- **Allow for differentiated access patterns:**
  - Static: For rapid download/reporting use (e.g., Natura 2000 officers, national agencies, SAP users)
  - Dynamic: VRE-based workflows with dynamic querying and data cube generation.
- **Design for Data Space Interoperability:** Conform to Green Deal Data Space and EOSC architecture guidelines (see section 7 for more).
- **Security and Trustworthiness:** Adhere to cybersecurity best practices and align with the security policies of participating organisations. This is crucial in ensuring the infrastructure is secure and resilient against attacks or misuse.
- **Align WP task integration:** The BMD architecture is designed to align with and support the responsibilities and deliverables of multiple technical tasks across the project. These integrations ensure coherence across the data space infrastructure, workflow execution, visualisation, and interoperability layers:
  - Task 4.2: Provides on-demand compute provisioning using Infrastructure-as-Code (IaC) tools (e.g., Terraform, AWS CloudFormation) and load-balanced access to scalable cloud resources to support VREs and dynamic processing needs.
  - Task 4.3: Delivers the geospatial visualisation engine (e.g., Web-GIS MapViewer), ensuring that data cubes and VRE results can be rendered interactively via RESTful APIs, GeoJSON, and OGC-compliant services (e.g., [WMS]()).
  - Task 4.4: Focuses on aligning BMD metadata and access governance with broader European frameworks (e.g., EOSC, DEDL, GDDS). This includes embedding ODRL policies for usage control and preparing for interoperability with Common European Data Spaces using specifications such as [Simpl]() and potential EuroHPC integration.
  - Task 5.1: Develops FAIR-aligned workflow packages (based on RO-Crate, Common Workflow Language, and WorkflowHub standards). These workflows are tightly coupled

with WP4 infrastructure and feed processed outputs and provenance metadata back into the data space.

- ○ Task 3.1: Supplies harmonised, multidimensional data cubes built from biodiversity and environmental data sources. These cubes serve as core analytical inputs to VREs and the SAP.
- ○ Task 6.2: Integrates the visualisation engine (developed in Task 4.3) into the SAP to provide stakeholders with interactive access to relevant datasets, overlays, and derived products.

This task alignment ensures that the data space does not function as a siloed component but as a federated, interoperable foundation that ties together data access, compute, visualisation, and policy relevance.

4. Example user journey

The BMD infrastructure is planning to serve a broad spectrum of users. The following user journey examples illustrate how different user types might engage with the data space: one through exploratory analysis using Virtual Research Environments (VREs), and the other through structured GUI interaction via the SAP. While the modes of access differ, both journeys depend on a shared backend: harmonised data cubes, consistent metadata, and policy-aware governance models. The VREs act as a bridging layer (abstracting complex data pipelines and harmonisation logic) so different types of users can benefit from a unified, reusable, modular infrastructure.

Note: The following user journeys are **illustrative examples**, developed to help contextualise how different user groups might interact with the BMD infrastructure. While grounded in realistic use cases, they are currently **hypothetical** and will be further refined through ongoing stakeholder engagement led by WP1. As user requirements and priorities become clearer, these scenarios will evolve to better reflect actual workflows.

Example User journey 1:
A Natura 2000 stakeholder ("user") in a researcher capacity is investigating climate change impacts on bird range shifts as a consequence of climatically induced dispersal. Working from their VRE environment, the user begins the analysis by accessing the data space's integrated species occurrence and climate data cubes.

The user loads a comprehensive species cube containing "x" years of bird observation data, combined with corresponding climate variables, including temperature and precipitation patterns from present and recent past conditions. Users can apply filters such as a priory spatial filtering of occurrence data results in truncated niche space models which do not reflect the species niche.

Using the pre-configured analytical tools, the user runs a model to identify patterns (such as receiving projections of projected range shifts under scenario conditions). The system generates derived products, including trend maps and assessments. Throughout the workflow, the system captures provenance data. Uncertainties and warning notes, if applicable, can also be provided.

Upon completion, the user exports results as a comprehensive RO-Crate package, ensuring the research is reproducible and FAIR-compliant for future collaboration and publication.

Example User journey 2:
A Natura 2000 stakeholder in a policy officer capacity is preparing annual protected area network reporting required under the EU Nature Directives. Using the SAP, the officer begins by browsing the latest relevant data to assess the conservation status of protected species within their territories.
A policy officer then would like to assess whether the site adequately protects the species and habitats for which it was designated.

Using the SAP's reporting tools, the policy officer generates summary statistics, downloads visualisations, summary statistics within protected areas. They download PDF reports and maps that comply with reporting standards, including the required checklist subset for directive compliance.

Example User question 1:
Which protected areas of the Annex 1 Habitats Directive ("Oligotrophic to mesotrophic standing waters with vegetation of the Littorelletea uniflorae and/or of the Isoëto-Nanojuncetea") are exceeding a critical load for nitrogen?

Example User question 2:
How has the national population status of the Annex 1 Birds Directive species Great Bustard (*Otis tarda*) changed over the past three decades (1995-2025)?

Example User question 3:
Which areas of the Mediterranean Spanish coast will remain most suitable for restoration of the Annex 1 Habitats Directive European native oyster beds (*Ostrea edulis*), under climate change (by 2080)?

5. Architecture Components
5.1.Data Ingestion & Provisioning

- Landing Zone: Object storage, such as S3 buckets or compatible alternatives, serves as the initial staging area for raw and incoming datasets.
- Ingestion Tools: Pipelines developed will manage the transformation and loading of data. The pipeline should be capable of performing quality control steps that can include schema validation, taxonomic name lookup, georeferencing, etc., if needed.
- Audit logs will be generated to keep track of things, so the errors can be triaged.
- Indexing/full text support: Elasticsearch can be used for indexing and fast retrieval support. These pipelines will interact with WP2, WP3, and WP5 components. Tools like PostGres and PostGIS are also being tested.
- Source Definition: The data sources and relevant metadata are being identified through ongoing catalogue work in WP2 and will ultimately reflect stakeholder priorities. This includes both legacy datasets and high-throughput monitoring sources. The data space will rely on this catalogue as the authoritative source.

5.2. Data and Metadata Storage

Modern data stack and Lakehouse 2.0 ideas:
To ensure data management and metadata consistency, the BMD data space incorporates emerging concepts from the modern data stack and the open [lakehouse](#) architecture model (See Figure 3). As this milestone has been delivered at month four of the project, we still have more details to gather on requirements that will provide more details on capacity and volume. But the design can accommodate several TBs of raw data, and assumption of growth. These are based on Naturalis's experience with high-throughput media and DNA data integration in the lakehouse and some initial GIS data testing. The storage requirement and growth also require considering the cost and long-term sustainability options.

This design choice reflects both the rapid evolution of cloud-native tooling and a commitment to open, community-driven infrastructure. BMD design adopts open table formats such as Delta Lake, which extend common file types like Parquet by adding transactional control and metadata layers that support interoperability and versioning. These open table formats store structured tabular data in distributed file systems, while managing metadata separately (typically in formats like JSON or Avro) to enable schema evolution, efficient querying, and performant access across analytic engines. Open lakehouse components like [Delta Lake](#) are supported by the Linux Foundation Projects, an independent open-source project and not controlled by any single company.

By following these open data lakehouse principles, BMD ensures that its backend infrastructure keeps pace with the broader changes in data stacks while avoiding lock-in to proprietary systems. This approach also aligns with FAIR and Open Science values (particularly around transparency, modularity, and reusability) and enables easier integration with other initiatives such as EOSC and Destination Earth. Rather than enforcing a single technology stack, the BMD Data Space is designed for interoperability. By decoupling metadata, storage, and compute engines, BMD enables diverse analytical tools and engines (e.g., Databricks, Clickhouse, VREs, notebooks) to interact without format lock-in. This decoupling is critical for compositing workflows involving cube generation, reprocessing, and sharing of intermediate data products across the VREs and SAP.

From a FAIR perspective, interoperability encompasses more than just technical access: it also involves semantic alignment and workflow reproducibility. In the semantic layer, BMD plans to support "pragmatic" mappings (allowing loose integration between heterogeneous vocabularies and classifications where strict formalisation is not feasible. For more details (see Broeder et al. 2021). However, in the area of workflow execution, especially for the VREs, stricter compliance is essential. Aligning workflow metadata with FAIR standards (e.g., RO-Crate + CWL) allows both humans and machines to understand, validate, and reproduce analysis pipelines consistently across contexts.

Adoption of Cloud-Native Formats
GBIF and OBIS have already adopted Parquet to optimise large-scale biodiversity data delivery. [Microsoft's Planetary Computer](#) exposes climate and environmental datasets via Parquet/Zarr and STAC APIs.

Amazon's [AWS Open Data](#) platforms and platforms like Destination Earth Data Lake (DEDL) use similar design patterns to combine open data access with scalable, trustworthy infrastructure.

A brief note here to illustrate the diversity of data handled in the BMD data space: structured, tabular biodiversity data from sources such as GBIF and OBIS are stored and queried using Parquet, a format well-suited for schema evolution and efficient read performance. In contrast, satellite and Earth observation data from providers like Copernicus often take the form of large, multi-dimensional arrays (e.g., NDVI or temperature time series), which are best represented using formats like Zarr or NetCDF, both designed for spatio-temporal gridded data and cloud-native, chunked access. This variation in data types and formats highlights the need for a flexible, format-aware infrastructure that can harmonise heterogeneous sources into integrated, analysis-ready cubes while still preserving usability across diverse tools and platforms.
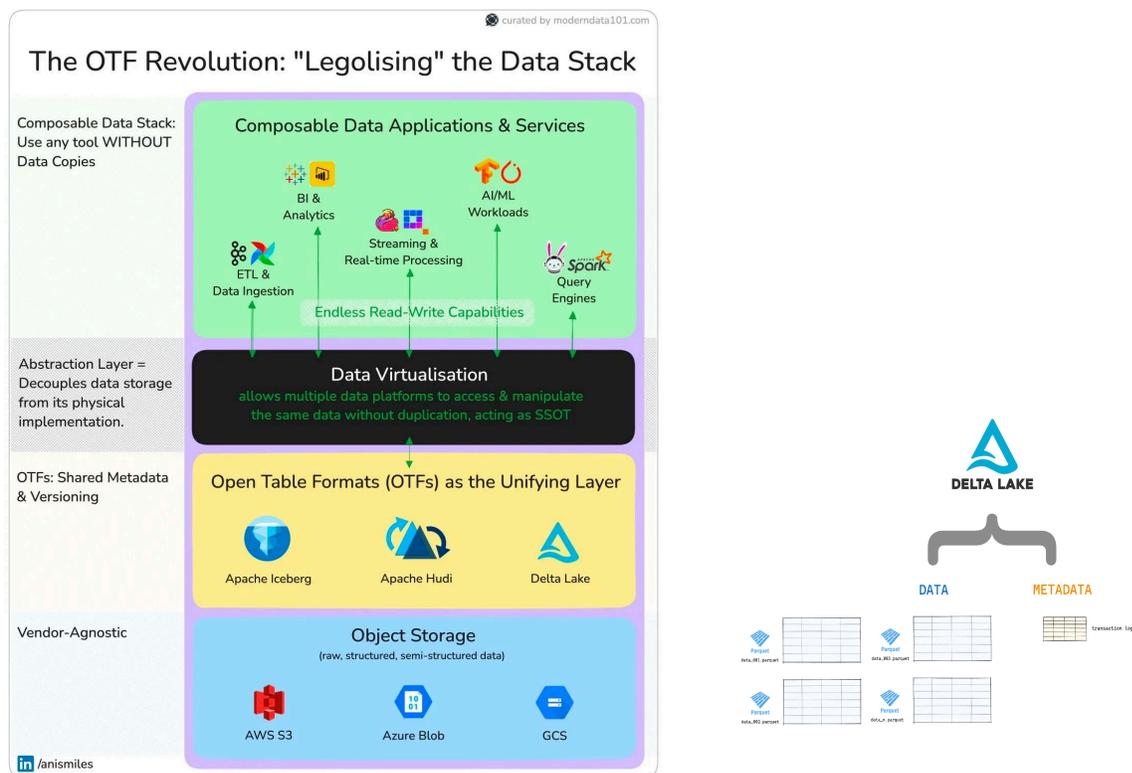

Figure 3: A generic schematic example of lakehouse 2.0 implementation and Delta Lake Open Table Format (OTF). Source: https://moderndata101.substack.com/p/lakehouse-20-the-open-system-that

5.3.Data Processing & Harmonisation

Cubing Engine Deployment:

- Given the federate-first approach in BMD, the cubing engine will make use of existing cubing services offered by data providing infrastructures as much as possible. Some post-hoc standardisation may still be needed, in particular to harmonise cubes generated by different data

providers. Cubing can be a very computationally intensive process, so most of these services will still be asynchronous, necessitating a mechanism to soften the latency for the user, such as caching certain scenarios or use cases.

- Cube generation may still occur internally within the BMD cloud infrastructure, particularly for data types that cannot be externally processed because cubing services are unavailable or found to be insufficient for the requirements of some VREs. This requires a landing zone for the raw data, which can be substantial, and sufficient computing power to cube the data. The potential latency issue also applies for internal cubing.

The architecture must accommodate both internal cubing and external cube API integration.

5.4.Metadata and Discovery

Metadata is a critical enabler of both discoverability and interoperability in the BMD Data Space. To support a range of user needs, the system separates public and internal metadata layers while ensuring they are consistently linked.

Public Discovery Layer:
GeoNetwork-based implementation will serve as the primary public-facing metadata catalogue, aligned with INSPIRE and ISO standards. It enables users to discover datasets via spatial and thematic filters. For spatial-temporal assets such as raster layers and data cubes, STAC (SpatioTemporal Asset Catalogs) endpoints also provide lightweight, machine-actionable metadata for efficient querying by Web-GIS platforms and VRE tools. There are some overlaps in metadata between INSPIRE and STAC: INSPIRE/GeoNetwork is mostly used for regulatory compliance and comprehensive metadata management, and STAC for modern, cloud-native geospatial data discovery and access.

Internal Metadata Layer:
Internally, metadata is also maintained through SQL-based or Lakehouse-compatible catalogues such as Postgres, PostGIS, or Delta Table schemas. These serve operational roles: tracking versions, schema changes, and cube configurations used in VRE workflows.

Persistent Identifiers (PIDs) and UUIDs:
To support reproducibility and long-term reference identifiers for datasets and derived products will be used. Internally, UUIDs can be used for traceability of workflow components, transformations, and intermediate artefacts. These identifiers can be embedded into metadata records for machine-actionable provenance (e.g., within RO-Crate packages or STAC Items).

Site-Level Identifiers:
Where applicable, data related to monitoring locations, Natura 2000 sites, or other study sites is cross-referenced and linked using persistent identifiers. In INSPIRE specification, there are different identifiers. With the help of the GeoNetwork catalog we will need to ensure these identifiers are linked and mapped. There are also DEIMS IDs (used in eLTER and related RI networks) that are used in some monitoring sites. These identifiers and mappings support integration with domain-specific monitoring infrastructures and data.

```
<ps:inspireId>
                <base:Identifier>
                        <base:localId>ProtectedSite.BE1000001_IB10</base:localId>

<base:namespace>http://databrussels.be/BELB/PS/ProtectedSite/</base:namespace>
                </base:Identifier>


….

<gmd:fileIdentifier>
  <gco:CharacterString
xmlns:gco="http://www.isotc211.org/2005/gco">9B2A50A8-E419-4536-9771-AD2470D66E12</gco:C
haracterString>
 </gmd:fileIdentifier>
 <gmd:language>
  <gmd:LanguageCode codeList="http://www.loc.gov/standards/iso639-2/"
codeListValue="dut">Nederlands; Vlaams</gmd:LanguageCode>
 </gmd:language>
 <gmd:characterSet>
  <gmd:MD_CharacterSetCode
codeList="http://schemas.opengis.net/iso/19139/20060504/resources/Codelist/gmxCodelists.xml#M
D_CharacterSetCode" codeListValue="utf8">utf8</gmd:MD_CharacterSetCode>
 </gmd:characterSet>
```

Box 1. Excerpts from an INSPIRE file.

Access and Policy Metadata:
Usage conditions are specified using ODRL (Open Digital Rights Language) statements embedded in metadata. These define allowable uses (e.g., attribution, no-commercial use) and support policy-aware data delivery to the SAP and VREs.

Cross-Registry Interoperability:
BMD aligns with EOSC interoperability practices by including references to the Metadata Schema and Crosswalk Registry (MSCR) and Data Type Registry (DTR) where appropriate. This ensures that metadata can be translated or mapped across systems, supporting reuse beyond the immediate project scope.

Together, these layers form a modular, extensible metadata framework that supports FAIR principles while also enabling compliance with EU data space and Green Deal architecture standards.

5.5.Data and service consumer interfaces

The SAP serves as the primary interactive access point for stakeholders. Through its embedded Mapviewer, users can explore harmonised biodiversity data, visualise trends, overlay protected habitats, and generate policy-ready and summary maps, for instance. VREs developed in this context offer programmatic access to the same underlying data space. These environments (similar to Jupyter notebook, RStudio) support data filtering, modelling, and workflow execution using structured cube data and linked metadata. VREs abstract the complexity of ingestion and harmonisation while ensuring provenance and reproducibility through RO-Crate enabled pipelines.

In addition, RESTful APIs allow integration with external tools and platforms, enabling more advanced use cases and interconnection with national or thematic data infrastructures.

Finally, a download service can also provide access to static products (pre-computed maps, tables, and datasets), ensuring that insights from the BMD infrastructure remain easily accessible. The design decisions and implementation reflect these use cases.

6.Implementation Timeline

This implementation timeline represents a tentative internal planning tool for WP4 activities and coordination across WPs (especially WP2, WP3, WP5, and WP6). It reflects the current understanding of technical dependencies and available resources. While the overall project milestones and deliverables as defined in the Grant Agreement remain unchanged, the scope of specific development activities may evolve as the project progresses. This is in line with an agile and iterative approach, which allows flexibility to respond to emerging requirements and stakeholder feedback. Regular coordination meetings will ensure alignment with both technical goals and stakeholder needs.

Table 1: WP4's current internal planning

Date	Output	WP Involved	Notes
June 2025	**MS19**: Initial Data Space Design & Implementation Plan Submitted.	WP4	Finalisation of this design document.
June 2025, July 2025	Start building internal cloud infrastructure for testing.	WP4	This is handled by Naturalis and involves setting up S3 buckets, Delta Lake tables, ingestion logic, Data Cube storage, etc.
July 2025	Align with task **5.1** and **MS 23** (MS23: Workflow concept for FAIR workflow into VREs available and disseminated in in-person workshop Improving	WP4 + WP5	Initial setup of metadata needed for the RO-Crate files and how/where to store these. Gather feedback from the BMD Workshop: Improving

Date	Output	WP Involved	Notes
	FAIRability of research with RO-Crates and Bioschemas).		FAIRability of research with RO-Crates and Bioschemas.
Aug 2025	Align with **MS16** requirements (MS16: First Data Cubes Ready by Aug 31).	WP3 + WP4	Initial testing of ingestion, data cube storage, and serving.
Aug 2025	Metadata Interoperability Checkpoint (GeoNetwork ↔ Open Table in data space).	WP2 + WP4	Metadata schema alignment, support for automated harvesting. decision on minimum metadata. There are potential risks around changes in accessibility of data source, Interoperability mismatch between the Biodiversity Data Space and GDDS, EOSC. These risks have been documented in the BMD project risk register.
Sept 2025	Cube prioritisation feedback (variables, sources, resolution).	WP3 + WP1 + WP4	Use stakeholder priorities to guide early cube builds and storage strategy.
Oct 2025	Towards first Cubing Engine & Metadata Registry Deployment (**D3.1** coming up in July 2026).	WP3 + WP4	RO-Crate metadata embedded; initial cube pipeline tested with sample GBIF/Copernicus data.
Nov 2025	VRE Integration (FAIR workflows, cube access).	WP4 + WP5	Connect internal cube API to RO-Crate-based workflows in the VREs. We need to understand the basic functionality for cube access, metadata queries, and SAP integration. We also need to understand how the queries will be sent and handled by the data space.
Nov 2025	Allocate resources for initial testing, bug fixing sprint before the functional release.	WP4	This is to get ready for the MS20 deadline.
Dec 2025	SAP/Mapviewer API and Visualisation Design.	WP4 + WP6	Define visualisation endpoints (REST, GeoJSON); confirm support for protected areas overlays.
Dec 2025	**MS20** readiness check.		In order to be ready for the Feb 2026 release of MS20 by Dec 2025, we need to have several things in place.

Date	Output	WP Involved	Notes
Jan 2026	Stakeholder feedback incorporation (via SAP/VRE Demos).	WP1 + WP4 + WP5	Test user journeys; adjust interface parameters and visualisation priorities.
Feb 2026	**MS20:** First Functional Release of BMD Data Space.	WP4	MS 20 delivered.
Mar – Jun 2026	Agile iterations and continuous Integration.	WP2–6 + WP4	Ongoing enhancements: versioning, Add more data types (like eDNA metadata), licensing information, different cube types.
July 2026	Align with **D3.1 Data Cube Engine** (D3.1: Software protocols available for use in WP4 to build any cubes necessary for the VREs).	WP3 + WP4	By now, we should know what the data cube engine will entail and how it will interact with the data space.
July 2026 onwards	Further refinement and improvements in order to be ready for D4.1.	WP4	
28 Feb 2027	**D4.1 First version of BMD data space released** (prototype of the BMD data space with first data operational).	WP4	This is a prototype of the BMD data space with an operational element supporting initial datasets.

7.Alignment with FAIR and GDDS

The Green Deal Dataspace (GDDS) emerged as an initiative within the framework of the European Green Deal, which has the dual objective of addressing the challenges of climate change and fostering the prosperity of European society and the competitiveness of its economy. The GDDS is conceptually designed as an open ecosystem for trusted sharing and pooling of Earth and environmental data, together with additional datasets mobilised across domains. The BMD data space is strategically designed to align with the GDDS reference architecture, building upon shared agreements and leveraging distinct WP contributions to establish its logical components. Our main stakeholders, Natura 2000 site managers, will benefit directly from this integrated approach.
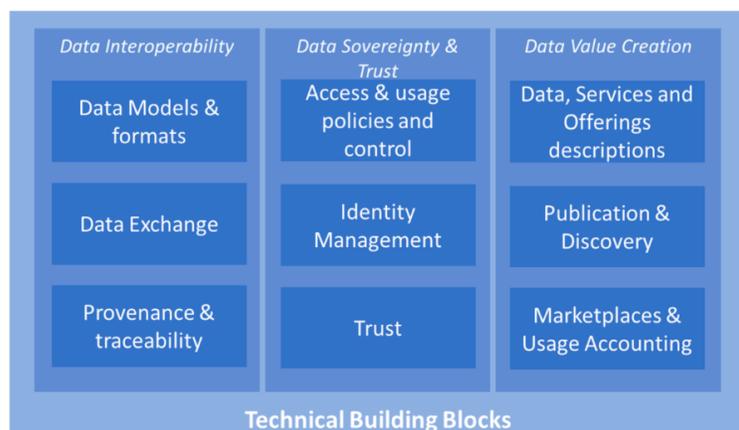
Figure 13 - Technical Building Blocks from DSSC Taxonomy Document

Figure 4: Technical Building block design. Source: GDDS Reference Architecture based on Data Space Support Center technical blueprint.
https://dssc.eu/space/bv15e/766066850/Technical+Building+Blocks

The architectural decisions described earlier in this document (such as dual catalogues, open table formats, machine-actionable metadata) directly support the technical building blocks outlined by the Data Spaces Support Centre (DSSC) technical building blocks blueprint. These building blocks are structured into three key pillars: Data Interoperability, Data Sovereignty and Trust, and Data Value Creation (see Figure 4).

Below is a brief mapping of how core BMD components correspond to these building blocks:

- **Shared Agreements (Metadata & Governance)** Supports: Data Sovereignty and Trust, Provenance & traceability. The BMD data space builds trust and interoperability through agreed-upon metadata standards and governance models. WP2 defines and implements metadata specifications (e.g., INSPIRE, ISO, EML), enabling consistent discoverability across diverse data sources. For sensitive datasets (e.g., protected species within Natura 2000), access and usage can be governed via machine-readable ODRL policies, supporting policy enforcement and audit trails.
- **Catalogue Layer** (Public + Internal) Supports: Data Value Creation, Discovery, Provenance & traceability. As outlined in earlier sections, BMD uses a dual-catalogue approach: A public-facing catalogue (GeoNetwork) enables stakeholders to discover datasets using spatial and thematic filters. An internal catalogue (e.g., Delta Table, Postgres) supports schema versioning, cube lineage, and operational tracking. This separation improves governance while ensuring datasets remain findable, reusable, and auditable.
- **Compute Engine and Harmonisation:** Supports Data Interoperability, Data Models & formats, Data Exchange. WP3 handles data harmonisation (standardising spatial, temporal, and taxonomic dimensions) while WP4 provides the compute layer for generating and serving data cubes. This supports both internal workflows (e.g., high-resolution raster ingestion) and

integration with external APIs (e.g., GBIF cube services). The infrastructure enables semantic integration possible, while also supporting "loose" interoperability to accommodate real-world heterogeneity.

- **API & Access Layer (SAP + VREs)** Supports: Data Exchange, Data Value Creation. BMD offers multiple entry points for data use: VREs (WP5) provide programmatic, model-ready access through APIs, supporting complex analysis and reproducible workflows. SAP (WP6) offers an intuitive, visual interface for non-technical users such as Natura 2000 site managers. Together, these tools allow users to interact with the same underlying data in different ways, reflecting the composable and reusable design of the architecture.

By structuring the system along these building blocks, BMD ensures that its infrastructure can not only meet current project needs but also scale and integrate into broader European data ecosystems such as EOSC and GDDS. These principles and components (introduced in previous sections of this document) ensure that the design is not only functional but also sustainable, interoperable, and extensible.

8.Conclusion/Future Work

This initial architecture provides the technical foundation for BMD's integrated and modular data space. As implementation progresses, future work will focus on aligning the infrastructure with evolving stakeholder needs, expanding support for diverse data modalities (e.g., sensors, eDNA), and enhancing interoperability with related European initiatives such as GDDS and EOSC. Sustaining the infrastructure beyond the project's lifecycle will require coordinated governance and resource planning, to be explored within the project scope and in collaboration with potential long-term service providers from the research infrastructure landscape.

9.Acknowledgements

We thank all colleagues who participated in the WP4 meetings and contributed through discussions, comments, and shared expertise. Special thanks to Mathias Dillen for his review and constructive feedback.

10.References

Broeder, D., Budroni, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Jeffery, K., Weiland, C., Wittenburg, P., & Zwolf, C. M. (2021). SEMAF: A Proposal for a Flexible Semantic Mapping Framework (1.0). Zenodo. https://doi.org/10.5281/zenodo.4651421

Curry, E. (2020). Dataspaces: Fundamentals, Principles, and Techniques. In: Real-time Linked Dataspaces. Springer, Cham. https://doi.org/10.1007/978-3-030-29665-0_3

Sterner BW, Gilbert EE and Franz NM (2020). Decentralized but Globally Coordinated Biodiversity Data. Front. Big Data 3:519133. https://doi.org/10.3389/fdata.2020.519133