DATA MANAGEMENT PLAN

PROJECT	
Project number:	101082130
Project acronym:	EVOLAND
Project name:	Evolution of the Copernicus Land Service portfolio (integrating novel EO data and latest machine learning algorithms to continuously monitor the status, dynamics and biomass of the land surface

DATA MANAGEMENT PLAN	
Date:	31/12/2025
Version:	V1.0

1. Data Summary

EvoLand generates, reuses, and integrates a wide range of Earth Observation (EO), in-situ, and ancillary datasets to support the development of next-generation Copernicus Land Monitoring Service (CLMS) products. The project covers multiple thematic domains—forest, agriculture, water, urban, and general land cover—and relies on dense satellite time series, multi-modal datasets, and in-situ reference information to develop, train, and validate advanced machine learning and continuous-monitoring workflows.

Re-use of Existing Data

EvoLand makes extensive use of established EO missions and in-situ datasets, including:

- Sentinel-1, Sentinel-2 and Sentinel-3 for radar and optical time series
- EnMAP, hyperspectral acquisitions, and other novel EO missions
- GEDI, ICESat-2, TanDEM-X, COPDEM10m/30m, ALOS national LIDAR data for elevation, structure, and biomass-related inputs
- HRL Vegetated Land Cover Characteristics (VLCC), CORINE Land Cover, and LUCAS for reference and validation
- LPIS and other national inventories where licensing permits
- Climate and meteorological datasets such as AGERA5
- Pleiades 1A/1B VHR data for model training and validation

The re-used datasets support the development of continuous forest monitoring, biomass modelling, land cover characterization, crop type mapping, and water/urban dynamics assessment. Licensing conditions vary: open-access datasets are preferred, while restricted data—such as VHR imagery or certain national datasets—are used under controlled access.

Types and Formats of Data

EvoLand handles multiple data types across a broad technological spectrum:

- EO imagery: optical, SAR, thermal, hyperspectral
- Derived products: vegetation indices, fAPAR, Tasseled Cap components, water/imperviousness masks
- In-situ and field-based measurements: forest inventory plots, LUCAS, crop information
- Training datasets: curated SSL4EO, MALICE, MMDC, SpectralEarth, and prototype-specific annotations
- Ancillary datasets: DEMs, meteorological grids, land use vectors

To ensure compatibility and scalability, data is managed in interoperable formats including Cloud-Optimized GeoTIFF (COG), Zarr, GeoParquet, NetCDF, HDF5, CSV, and common geospatial vector formats. Prototype outputs are delivered as cloud-ready COGs accompanied by detailed metadata.

Purpose and Alignment with Project Objectives

All re-used and newly produced datasets contribute directly to EvoLand's scientific and operational objectives:

- enabling continuous monitoring of the land surface
- supporting weakly supervised and self-supervised learning for land cover and forest structure
- improving biomass and tree height estimation
- enhancing detection of disturbances, urban imperviousness, water bodies, and small landscape features
- testing scalable, cloud-native data processing workflows
- informing the evolution of future CLMS products

Data across the workflow—from raw measurements to annotated training samples and prototype outputs—is structured to ensure transparency, reproducibility, and future integration into operational Copernicus services.

Expected Data Volume

Given the high spatial and temporal resolution of the datasets and the multi-year scope (2020–2022) across multiple demonstration sites, data volumes reach multi-terabyte scale. Examples include:

- Several terabytes of training data (e.g., SpectralEarth hyperspectral >3 TB)
- Tens to hundreds of gigabytes per prototype for annual candidate maps
- Large intermediate archives generated by temporal compositing, cloud masking, and feature extraction workflows

Cloud infrastructure ensures scalable storage, efficient retrieval, and parallel processing.

Data Origin and Provenance

Data originates from a combination of:

- established EO missions (ESA, NASA, CNES, DLR)
- European data services (e.g., Copernicus Data Space Ecosystem)
- national and regional authorities providing in-situ datasets
- project-generated training datasets and prototype outputs from WP2 and WP3
- cloud-native data catalogues hosted by IIASA, structured as STAC Collections

To ensure transparent provenance, all datasets—both re-used and newly created—are accompanied by metadata documenting acquisition parameters, processing levels, algorithms applied, versioning, quality assessment, and licensing constraints.

Data Utility and Target Users

The data generated and organised under EvoLand is intended for a broad range of users, including:

- CLMS service developers and operational teams
- policy stakeholders (e.g., EEA, JRC, national agencies)
- the scientific community working on EO, machine learning, and environmental monitoring
- developers of downstream applications requiring harmonised, cloud-accessible EO datasets

The combination of open data, reproducible workflows, and interoperable metadata maximises long-term reusability beyond the duration of the project.

2. FAIR data

The EvoLand project adheres to the FAIR (Findable, Accessible, Interoperable, Reusable) principles to ensure that all datasets, training data, prototype outputs, and metadata products can be effectively located, accessed, and reused during and after the project. FAIR implementation applies to both re-used third-party data and data generated within the project, including machine learning training datasets, annotations, intermediate outputs, and prototype candidate products.

Making data findable, including provisions for metadata

Persistent identifiers

All EvoLand-produced datasets are assigned globally unique and persistent identifiers (e.g., DOI or stable URLs) to ensure long-term traceability and citability. STAC Collections and Items are published under permanent links within trusted cloud repositories maintained by project partners.

Metadata standards and richness

Metadata follows established standards such as ISO 19115, OGC, and GEOSS profiles. EvoLand further adopts the SpatioTemporal Asset Catalog (STAC) specification as its core metadata structure, providing machine-readable JSON descriptions for all datasets.

STAC metadata includes:

- spatial and temporal coverage
- sensor and acquisition metadata
- processing levels and algorithms applied
- asset links to cloud-optimised formats (COG, Zarr, GeoParquet)
- provenance, quality metrics, and versioning
- licensing and access conditions

STAC extensions such as EO, SAR, and Label are used where appropriate, enabling consistent descriptions of optical time series, radar backscatter, hyperspectral data, and training/validation labels.

Keywords for discovery

Metadata records use controlled vocabularies aligned with Copernicus, GEOSS, and INSPIRE terminology. Keywords enable efficient search and semantic indexing across themes such as forest monitoring, biomass, land cover characteristics, crop type, imperviousness, and water bodies.

Harvesting/indexing

Metadata is openly harvestable through:

- STAC API endpoints
- static STAC catalogues (https://stac.iiasa.ac.at/)
- OGC-compatible catalogue services

This ensures interoperability with cloud platforms, Copernicus infrastructures, and external research data portals. Metadata remains accessible even when underlying datasets require restricted access.

Making data accessible

Data Repositories

EvoLand stores datasets in trusted, secure, and scalable repositories, including institutional cloud infrastructures (e.g., IIASA, VITO) and S3-compatible object storage. The STAC catalogue provides unified access to all project datasets through structured collections.

Data access conditions

- **Open access** applies to datasets derived from publicly funded missions (e.g., Copernicus Sentinel, HRL, LUCAS).
- **Restricted access** applies to licensed or confidential datasets (e.g., VHR imagery, national inventories requiring permission), and selected training datasets with provider-specific terms (e.g., SpectralEarth, SSL4EO-EU-Forest).
- STAC metadata explicitly states all licensing terms, restrictions, and authentication requirements.

If embargo periods apply, they are reflected in metadata and repository access policies.

Protocols for access

Data is accessible through a range of interoperable protocols:

- HTTPS downloads
- cloud-native access to COGs, Zarr, and GeoParquet
- OGC services such as WMS and WCS
- STAC queries, enabling selective and scalable retrieval

These mechanisms support both GIS-based workflows and cloud computing environments (e.g., openEO, Python APIs).

Identity verification

Restricted datasets require user authentication in line with provider agreements. Access control follows the procedures defined in WP5's security framework and repository-specific policies.

Metadata accessibility

Metadata is openly available under CC0 and includes all necessary information to locate, interpret, and retrieve the data. Metadata remains permanently accessible even if datasets are archived or require authorization.

Making data interoperable

EvoLand ensures interoperability through:

- use of widely adopted EO formats: **GeoTIFF/COG**, **NetCDF**, **HDF5**, **Zarr**, **GeoParquet**
- harmonised spatial reference systems and consistent data structures across prototypes
- adoption of STAC extensions for optical EO, radar, and labelled datasets

Project-specific vocabularies (e.g., prototype class definitions) are documented and mapped to recognised standards to ensure consistency across systems and users.

Interoperability extends to machine learning workflows. Training datasets include:

- standardised label schemas
- clear band definitions (e.g., Sentinel-2 spectral order)
- masks, flags, and acquisition angles when relevant
- common metadata structures enabling model benchmarking and reproducibility

All datasets include references to their source files, processing lineage, and dependencies.

Increase data re-use

Documentation Supporting Reuse

Each dataset—whether training data, prototype outputs, or intermediate layers—includes comprehensive documentation:

- descriptions of processing workflows and algorithms
- definitions of variables, units, and classes
- uncertainty and quality information

- model hyperparameters (for ML-based outputs)
- version histories and update logs

This ensures that users can fully understand how datasets were generated and under which assumptions.

Licensing for Reuse

Licenses applied to EvoLand datasets follow these principles:

- CC-BY or similarly open licences for products derived from open data
- provider-specific licences for restricted datasets (e.g., commercial or VHR data)
- explicit licensing statements embedded in STAC metadata for transparency

Reusability Beyond the Project Lifetime

EvoLand designs its data and metadata to remain accessible and reusable after project completion. The STAC catalogue—lightweight, static, and widely supported—ensures that metadata and links to assets remain stable over time.

Open datasets will be preserved in public repositories, and prototype outputs are made available in interoperable, cloud-optimised formats suitable for integration into future Copernicus workflows.

The results that were uploaded to the Results Portal (Copernicus Browser Theme) will remain available in the Portal after the project ends for interactive visualization and for data download of the visualised as well as the raw data.

Provenance and Quality Assurance

All datasets include:

- detailed provenance chains
- algorithm versions and software references
- processing-level descriptions
- QA metrics and validation summaries (e.g., accuracy metrics for land cover and disturbance prototypes)

These details support scientific reproducibility and facilitate integration into operational services.

3. Other research outputs

EvoLand generates a variety of digital research outputs in addition to datasets. These outputs play a critical role in ensuring transparency, reproducibility, and long-term scientific and operational value. All outputs are aligned with FAIR principles wherever applicable.

Software and Processing Workflows

EvoLand develops a suite of processing workflows covering EO pre-processing, data fusion, machine learning model training, continuous monitoring, biomass estimation, and prototype generation. These include:

- openEO-based workflows for large-scale processing of Sentinel time series
- ML pipelines for forest disturbance detection, imperviousness mapping, cover crop typing, and continuous land surface classification
- Data-fusion methods for multi-modal EO integration (Sentinel-1 / Sentinel-2 / DEM / meteo grids)
- STAC metadata generation tools for automated catalogue creation

Software is documented with usage instructions, parameters, and dependencies. Where licensing allows, workflows are released under open-source licenses in repositories with long-term persistence (e.g., GitHub linked to Zenodo DOIs).

Models and Algorithms

Machine learning and statistical models developed within EvoLand include:

- SSL and weakly supervised models trained on datasets such as SSL4EO-EU-Forest, MALICE, and MMDC
- Biomass and tree height estimation models using EO and in-situ inputs
- Temporal classifiers for continuous monitoring of land cover dynamics
- Deep learning models for tree type classification, crop type mapping, urban land use change detection, and small landscape feature identification

Model metadata includes architecture descriptions, training datasets, hyperparameters, performance metrics, and version histories to support reproducibility.

Standard Operating Procedures (SOPs)

EvoLand maintains SOPs that standardise technical practices across partners, covering:

- EO image pre-processing
- temporal compositing, cloud screening, and harmonisation
- QA/QC procedures for datasets and annotations
- validation methodologies for prototype products
- cloud-processing best practices from WP2 (e.g., optimised tiling, chunking, caching strategies)

These SOPs ensure coherent and reproducible processes across the consortium.

Physical Outputs

The project does not produce significant physical outputs. All relevant outputs are digital and handled according to data governance requirements.

4. Allocation of Resources

EvoLand dedicates appropriate human, computational, and financial resources to the management, preservation, and dissemination of data and research outputs. These resources enable full compliance with FAIR principles and ensure operational readiness of developed prototypes.

Costs for Making Data and Outputs FAIR

Costs arise from:

- cloud storage and compute for EO and ML processing
- maintaining the STAC catalogue and metadata infrastructure
- staff time for QA/QC, documentation, and versioning
- repository fees where applicable
- security monitoring and compliance

These activities fall within eligible Horizon Europe research-support activities.

Coverage of Costs

Costs are covered through:

- personnel allocations in WP2, WP3, and WP5
- contributions from infrastructure-providing partners (e.g., VITO, CNES, DLR, IIASA, Sinergise)
- dedicated budget for dissemination, exploitation, and data management under WP5

Responsibility for Data Management

Data management is coordinated by:

- VITO as project coordinator
- WP5 lead for DMP implementation
- data stewards appointed by each partner institution

Responsibilities include dataset documentation, quality control, metadata creation, adherence to FAIR and security policies, and maintaining the STAC catalogue.

Long-Term Preservation

Long-term preservation uses trusted repositories and institutional infrastructures, ensuring:

- persistent metadata
- redundancy and backups
- compliance with licensing and embargo rules
- long-term findability and accessibility for open datasets

Decisions on preservation duration follow consortium agreements, regulatory obligations, and scientific value.

5. Data Security

EvoLand implements robust data security measures to protect datasets, metadata, and digital research outputs throughout their lifecycle.

Security Measures

Security practices include:

- role-based access control and authentication for restricted datasets
- encrypted data transfer (HTTPS, SFTP, VPN) and encrypted storage
- secure cloud infrastructure with physical and logical safeguards
- access logging, monitoring, and auditing
- partner-wide training on secure data handling, credentials, and GDPR compliance

Security procedures apply to both shared project spaces and partner-internal infrastructures.

Backup and Recovery

Backups follow structured schedules:

- periodic incremental and full backups
- off-site storage for disaster recovery
- routine restore tests to ensure backup integrity

Backup retention periods follow institutional policies and legal requirements.

Long-Term Archiving

Data is archived in trusted environments that guarantee:

- redundant storage
- integrity verification
- controlled access aligned with licensing terms
- durability of metadata for future discovery

Sensitive Data

Most EvoLand datasets are EO-based and do not contain personal data. Some in-situ or national datasets may carry access restrictions, which are handled through secure environments and documented agreements. No direct personal data processing is anticipated.

6. Ethics

EvoLand complies with all applicable ethical, legal, and regulatory frameworks governing data use, sharing, and protection.

Ethical and Legal Compliance

The project adheres to:

- EU and national legislation
- GDPR principles (where applicable)
- licensing terms of data providers
- ethics requirements outlined in the Grant Agreement

EO datasets generally present no ethical risks. Any dataset with potential sensitivities is handled under strict conditions.

Informed Consent

If any ancillary activities involve surveys or interviews (e.g., stakeholder consultations), informed consent governs the collection, use, storage, and disposal of personal information.

Ethical Data Sharing

EvoLand ensures:

- no data is shared in violation of licensing or confidentiality
- transparency regarding restrictions
- ethical reuse aligned with open science principles

Ethics Documentation

Ethical considerations are documented in:

- ethics self-assessments
- internal data-handling guidelines
- deliverables addressing ethical compliance

7. Other Issues

Where beneficial, EvoLand aligns with established national, institutional, or funder-level data management procedures. These include:

- IIASA's STAC governance and cloud data stewardship framework
- VITO's operational data security, validation, and metadata management procedures
- ESA and Copernicus data governance best practices for operational EO services
- internal partner guidelines ensuring high-quality, reproducible machine learning workflows

These complementary procedures enhance robustness, interoperability, and sustainability of project outputs.

HISTORY OF CHANGES		
VERSION	PUBLICATION DATE	CHANGE
0.1	01.07.2023	Completion of interim version
0.2	05.12.2025	Review of interim version
1.0	22.12.2025	Final review and consolidation