Evoland
LAND MONITORING EVOLUTION

D5.8 Data Management Plan - Interim version

Work Package 5 - Communication, Dissemination & Exploitation

Authors: Wai-Tim Ng (VITO)

Date: 10.06.2023

Funded by the European Union

Full Title	**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)			
Project number	101082130		**Acronym**	EVOLAND
Start date	01.01.2023		**Duration**	36 months
Granting authority	European Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Vlaamse Instelling voor Technologisch Onderzoek N.V. (VITO)			
Date of delivery	**Contractual**	M6	**Actual**	M6
Type	R - Document, report		**Dissemination level**	PU - Public
Lead beneficiary	VITO			
Lead author	Wai-Tim Ng		**Email**	wai-tim.ng@vito.be
Other authors	Name and surname (Affiliation), Name and surname (Affiliation)			
Reviewer(s)	Ruben Van De Kerchove (VITO)			
Keywords	Data storage, security, preservation, documentation , FAIR principles,			

Document Revision History				
Version	**Issue date**	**Stage**	**Changes**	**Contributor**
0.1	10.06.2023	Draft	Setup template and draft version	VITO
0.2	22.06.2023	Draft	Completion of DMP	VITO
1.0	30.06.2023	Final	Internal review and finalization	VITO

Contents

Table of Tables

1 Introduction

The Data Management Plan – Interim version document is deliverable item 5.8 of the 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 land surface (EvoLand) project (Proj. nr: 101082130). It is presented by VITO on behalf of the project consortium and this document will be updated at M36 as D5.9 Data management plan - final version.

The purpose of this data management plan is to outline the strategies and procedures for managing EvoLand data. The document addresses the entire data lifecycle, including data collection, storage, documentation, quality assurance, security, access control, preservation, sharing, and long-term accessibility.

2 Data collection

2.1 EO Data sources

The following **EO data** (Table 1) have been identified and a number of these are summarized in the deliverable D2.1: EO missions catalogue relevant for EvoLand - interim version.

Table 1. List of EO data, their accessibility, licensing and providers

Name	Accessibility	Licensing	Source/provider
Sentinel-1	Copernicus Hub	ESA – User License	ESA
Sentinel-2	Copernicus Hub	ESA – User License	ESA
ConstellR	Through company	Commercial	ConstellR
OroraTech	Through company	Commercial	OroraTech
Pléiades 1A/1B	DINAMIS service	Licence restrictions	CNES
Pléiades-Neo 3/4	AirbusOneAtlas	ADS - User License	Airbus Defence & Space
SPOT 6/7	AirbusOneAtlas	ADS - User License	Airbus Defence & Space
TRISHNA	Geo Data Hub service	N/A	CNES
CO3D	AirbusOneAtlas	ADS - User License	Airbus Defence & Space
SWOT	hydroweb.next	N/A	CNES / NASA
Planet	Planet data	N/A	Planet Labs
GEDI	LP DAAC	no licensing restrictions	USGS, NASA
TanDEM-X	https://tandemx-science.dlr.de/	User License	DLR
NISAR	nisar.jpl.nasa.gov/data	open-source license	NASA
EnMap	planning.enmap.org	User License	DLR
ALOS-2 PALSAR-2	alos-pasco.com	End User License Agreement	JAXA
ICESat-2	nsidc.org/	free and accessible to scientists	NASA

2.2 Derived parameters

The following **EO derived parameters** have been identified. This list will be further elaborated once additional parameters are identified over the course of the project.

- Sentinel-2 fAPAR

- Sentinel-1 Surface Soil Moisture
- Sentinel-1 Soil Water Index
- Sentinel-2 Seasonal trajectories

2.3 In situ and ancillary data

The following table (Table 2) presents in situ and ancillary data, its means of access, licensing and provider.

Table 2. Table of in situ and ancillary data.

Name	Accessibility	Licensing	Source/Provider
LUCAS	Eurostat	Open access	European Commission
LPIS	Country dependent	N/A	European country governments
HRL forest dominant leaf type	CLMS	Open access	EEA
HRL forest tree cover density	CLMS	Open access	EEA
Swedish National Forest Inventory	slu.se	Via request	The Swedish National Forest Inventory, SLU
Spain National Forest Inventory	miteco.gob.es	Open access	Ministry for the Ecological Transition and the Demographic Challenge (MITECO)
TropiSAR	earth.esa.int/eogateway/campaigns/tropisar	Open access	ESA
Sweden LIDAR data	lantmateriet.se	open data license Creative Commons	Lantmäteriet
Spain LIDAR data	centrodedescargas.cnig.es/	Open access	Centro Nacional de Información Geográfica
CORDA	N/A	Access restricted	Various
Open Street Map	openstreetmap.org	Open Data Commons Open Database License	OpenStreetMap Foundation
Fluxdata GPP (iCOS)	icos-cp.eu/observations/carbon-portal	Open access	FLUXNET
AGERA5 Meteo data	openEO	Open access	ECMWF

3 Data Storage and Organization

3.1 Storage Infrastructure

Effective storage infrastructure is essential for managing and preserving the vast amount of EvoLand data. The storage infrastructure should be designed to accommodate the volume, velocity, and variety of data. Here are some key considerations for the storage infrastructure of earth observation data:

- **Storage Capacity and Scalability**: EO data can accumulate rapidly, especially with the increasing number of satellites and the higher spatial and temporal resolution of sensors. Therefore, the storage infrastructure should provide sufficient capacity to accommodate the growing data volumes. It is crucial to estimate the expected data growth and plan for scalable storage solutions to ensure long-term storage capabilities.
- **Redundancy and Data Resilience**: To safeguard earth observation data against hardware failures, data loss, or natural disasters, redundancy and data resilience measures should be implemented. This can include deploying redundant storage systems, utilizing data replication or mirroring techniques, and employing backup strategies to create multiple copies of the data.
- **Performance and Access Speed**: Given the large size of EO datasets, the storage infrastructure must deliver high-performance access to enable efficient data retrieval and analysis. Utilizing storage technologies such as solid-state drives (SSDs), parallel processing, and distributed file systems can help enhance the speed and performance of data access.
- **Data Transfer and Network Bandwidth**: Considering the size of EO data, transferring the data between different storage systems or locations can be time-consuming. It is important to have adequate network bandwidth to facilitate efficient data transfer and ensure timely access to the data, especially when collaborating with remote researchers or institutions.
- **Data Tiering and Hierarchical Storage Management**: To optimize storage resources and cost, a tiered storage approach can be implemented. Frequently accessed or critical data can be stored in high-performance storage tiers, while less frequently accessed or archival data can be stored in lower-cost storage tiers. Hierarchical storage management (HSM) techniques can be employed to automatically move data between different tiers based on usage patterns.
- **Data Integrity and Data Validation**: Maintaining data integrity is crucial for earth observation data. The storage infrastructure should implement mechanisms for data validation, including checksums or error-detection codes, to ensure the integrity of stored data. Regular data integrity checks should be performed to identify any errors or corruptions.
- **Data Format and Interoperability**: The storage infrastructure should support a range of file formats commonly used in EO data, ensuring compatibility and interoperability with different analysis tools and software.

It is advisable to use standard and open formats to maximize data accessibility and prevent format obsolescence.

- **Metadata Management**: Metadata, including descriptive, structural, and administrative metadata, plays a crucial role in organizing and discovering earth observation data. The storage infrastructure should provide mechanisms to capture, store, and manage metadata associated with each dataset, facilitating effective data search, retrieval, and understanding.

In conclusion, a well-designed storage infrastructure is fundamental for managing and preserving EO data. It should be scalable, redundant, and resilient to ensure data availability and integrity. Furthermore, it should provide efficient data access, support various file formats, and facilitate metadata management.

3.2 File formats

Earth Observation data is generated from various satellite missions, airborne sensors, and ground-based systems. To effectively manage and analyse EO data, it is crucial to use appropriate file formats that ensure data integrity, interoperability, and long-term accessibility. Here are some commonly used file formats:

- **Georeferenced Tagged Image File Format** (GeoTIFF) is a widely used file format for storing geospatial raster data, including satellite imagery and digital elevation models (DEMs). It supports embedded metadata, such as spatial reference information, coordinate systems, and projection details. GeoTIFF files are self-contained and can be easily opened and analysed using various geospatial software tools.
- **Network Common Data Form** (NetCDF) is a data format commonly used for multidimensional scientific data, including EO data. It is designed to store large datasets with multiple variables and dimensions. NetCDF files provide self-describing metadata, enabling the storage of geospatial and temporal information along with the data. This format is widely supported by scientific analysis software and libraries.
- **Hierarchical Data Format 5** (HDF5) is a flexible file format that supports the storage of large and complex datasets. It provides a hierarchical structure to organize data into groups, datasets, and attributes. HDF5 supports multidimensional arrays, making it suitable for storing EO data with multiple dimensions, such as time series data or hyperspectral imagery. The format allows for efficient data compression and provides options for chunking and parallel I/O operations.
- **Keyhole Markup Language** (KML/KMZ) is an XML-based file format used for visualizing and sharing geospatial data, including EO data, in applications like Google Earth. KML files can store georeferenced points, lines, polygons, and raster overlays. KMZ is a compressed version of KML that bundles multiple files into a single archive, making it easier to share and distribute geospatial data.
- **Shapefile** is a popular vector data format used for representing geospatial features, such as points, lines, and polygons. While it primarily stores vector geometries, it can also include attribute information associated with each

feature. Shapefiles are widely supported by Geographic Information System (GIS) software and can be used to represent EO data-derived features or ground truth data.

- **Comma-Separated Values** (CSV) is a simple and widely supported file format used for tabular data. While it does not inherently support geospatial information, it is often used to store attribute data associated with geospatial features. CSV files can be easily opened and manipulated using spreadsheet software or programming languages.
- **Cloud-Optimized Geotiff** (COG) is an optimized version of the GeoTIFF format that enables efficient access to EO data stored in cloud storage systems. It utilizes cloud-native capabilities, such as partial file reading and parallel access, to enhance data retrieval and analysis performance. COG files can be seamlessly accessed by geospatial analysis tools, making them suitable for cloud-based processing environments.

When selecting an EO file format, it is important to consider factors such as data structure, required metadata, compatibility with analysis tools, and long-term preservation. Adhering to open and standard formats enhances interoperability, accessibility, and future-proofing of EO.

3.3 Data organization

Proper organization of the EvoLand data is essential for efficient data management, retrieval, and analysis. An organized data structure enables easy navigation, understanding, and sharing of the datasets.

- **Folder Structure and Naming Conventions**: aim to establish a well-defined folder structure and consistent naming conventions is crucial for maintaining a logical organization of EO data. Consider creating a hierarchical structure that reflects the project or mission and use descriptive folder and file names to provide clear identification and context. Organizing data by date, location, sensor, or thematic area can facilitate efficient data discovery and retrieval.
- **Metadata and Documentation**: Metadata plays a critical role in describing and contextualizing EO data. Develop a metadata schema that captures essential information such as acquisition details, processing steps, calibration parameters, and spatial-temporal coverage. Attach metadata files to each dataset or store them in a central repository. Additionally, maintain documentation that provides detailed descriptions of the datasets, data variables, units of measurement, and any relevant data processing workflows.
- **Data Versioning:** Implement a version control system to manage different versions of the datasets. Use a consistent and transparent approach for labelling and tracking changes or updates made to the data. This allows for reproducibility, traceability, and the ability to revert to previous versions if necessary.
- **Data Categorization and Classification**: Develop a system for categorizing and classifying EO data based on specific parameters, such as sensor type,

spatial resolution, temporal coverage, or thematic content. Assigning appropriate tags or keywords to datasets can enable quick searches and facilitate the discovery of relevant data.

- **Data Access Control**: Establish access control mechanisms to ensure appropriate levels of data security and privacy. Define user roles and permissions to regulate who can view, modify, or delete the data. Consider implementing user authentication systems, encryption protocols, and access logs to monitor and track data access activities.
- **Data Indexing and Searchability**: Implement indexing mechanisms or metadata catalogues to create searchable databases of EO data. This allows users to query and retrieve datasets based on specific criteria, such as location, date range, or thematic attributes. Indexing can be done using dedicated geospatial database systems or by leveraging metadata standards and search tools.
- **Data Replication and Distribution**: Replicating EO data across multiple storage locations or servers to enhance data availability and reduce latency for distributed access. Replication enables faster access to data and can improve resilience against hardware failures or network disruptions. Additionally, establish protocols for data distribution to external stakeholders or collaborators, ensuring secure and controlled data sharing practices.
- **Regular Data Maintenance**: Regularly review and update the data organization structure, metadata, and documentation as new data is acquired or changes are made to existing datasets. Perform periodic data integrity checks, error handling, and quality control procedures to identify and rectify any issues that may arise during data management processes.

A preliminary scheme for organizing EvoLand data and overview of the attributes is listed below. This will be adapted for the given sensor or data source.

Mandatory Metadata

- Title / Product (File) name
    - REGION_DAY(YYYYMMDD)_TIME(HHMMSS)_SENSOR_LEVEL_VERSION
- Project
    - EvoLand
- Institution
    - VITO, CLS, IIASA, UT3, INRAE, SIN, JR, CNES, DLR, GAF, EVF, IRD
- Contact
    - Email address for the respective sources
- Product version
- Time stamp
    - time coverage start YYYY-MM-DDThh:mm:ss.mms

    - time coverage end YYYY-MM-DDThh:mm:ss.mms
    - time coverage centre YYYY-MM-DDThh:mm:ss.mms
    - time coverage duration
    - time_coverage_resolution
- Geographic coordinate system / Projection
- GeoBoundary
    - geospatial_lat_min
    - geospatial_lat_max
    - geospatial_lon_min
    - geospatial_lon_max
- Raster Width
- Raster Height
- Pixel size
- Processing
    - date_created
    - creator_name
    - creator_url
    - creator_email
    - Processing step i
        - Processor name and version
        - Processor parameters
        - Source product(s)
            - Sensor / instrument
            - creator_name
            - File name
            - Relative and absolute orbit OR path/row
            - Sensing Start
            - Sensing Stop
        - Auxiliary data used including version number
    - Processing step i+1

4 FAIR principles

In today's data-driven world, where vast amounts of information are generated and collected across various domains, it has become increasingly important to ensure that data is Findable, Accessible, Interoperable, and Reusable (FAIR). The FAIR principles provide a framework for maximizing the value and impact of data by making it more discoverable, usable, and compatible across different systems and disciplines. The FAIR principles were first introduced in 2014 as a response to the challenges researchers and organizations face in managing and sharing data effectively. Developed by a diverse group of experts from academia, industry, and government, the principles offer a set of guidelines to enhance data management practices and promote collaboration and knowledge discovery.

- The first principle, **Findability**, emphasizes the importance of enabling easy and efficient discovery of data and associated metadata. To achieve findability, data should be assigned globally unique and persistent identifiers, making it citable and referable. Additionally, comprehensive metadata, including descriptions of data, should be provided in a machine-readable format, enabling automated discovery and linking.
- The second principle, **Accessibility**, focuses on ensuring that data is readily available to users. It emphasizes the need for data to be openly accessible, where possible, and at least accessible under specific conditions, such as through controlled access mechanisms. Accessible data should be retrievable through standardized protocols and have clear terms of use, licensing, and access restrictions.
- The third principle, **Interoperability**, highlights the importance of enabling data integration and interoperability across diverse systems and platforms. Data should be structured using standardized vocabularies, formats, and ontologies, enabling seamless exchange and integration with other datasets. This promotes data harmonization, enabling meaningful comparisons and integrative analyses across different sources and disciplines.
- The fourth principle, **Reusability**, addresses the ability to reuse data beyond its original purpose. To enhance reusability, data should be well-described, including information on its provenance, quality, and potential limitations. Clear and accessible usage licenses should be provided, specifying the conditions and constraints under which data can be reused. Moreover, data should be prepared in a way that is easily understandable and interpretable by humans and machines.

By adhering to the FAIR principles, researchers, institutions, and organizations can unlock the full potential of data. FAIR data enables more efficient collaboration, accelerates scientific discovery, facilitates evidence-based decision-making, and supports the development of innovative applications and services. Furthermore, FAIR data promotes transparency, reproducibility, and accountability, reinforcing the foundation of trustworthy and open science. As the digital landscape continues to evolve and data volumes grow exponentially, the FAIR principles provide a robust framework for ensuring that data remains a valuable and reusable asset, driving progress and innovation across diverse domains and disciplines.

4.1 Findability

Making EvoLand data findable is crucial for ensuring its discoverability and accessibility by researchers, scientists, policymakers, and other stakeholders. To achieve findability, EO data should be assigned unique and persistent identifiers, such as Digital Object Identifiers (DOIs) or Uniform Resource Identifiers (URIs), which allow for easy referencing and citation. These identifiers enable users to locate and access specific datasets reliably.

Comprehensive metadata is a vital component of findability for Earth observation data. Metadata should provide detailed information about the data, including its acquisition parameters, temporal and spatial coverage, resolution, and processing methods. Standardized metadata schemas, such as those defined by the ISO or the Global Earth Observation System of Systems (GEOSS), facilitate consistency and interoperability, making it easier to discover and compare datasets from different sources.

In addition to rich metadata, appropriate indexing and cataloguing mechanisms are essential for enhancing the findability of Earth observation data. Data repositories and catalogues, both centralized and distributed, play a crucial role in curating and organizing the data, ensuring it is searchable and discoverable through appropriate search engines or discovery platforms. These repositories can be discipline-specific, thematic, or encompass a broader scope, providing users with a centralized access point to locate and retrieve relevant Earth observation data.

Standards and protocols for data access, such as Open Geospatial Consortium (OGC) standards like Web Map Service (WMS) and Web Coverage Service (WCS), promote the interoperability and accessibility of Earth observation data. By adopting these standards, data providers ensure that their datasets can be accessed and utilized across different software tools, platforms, and applications, further enhancing the findability and usability of the data.

4.2 Accessibility

The Accessibility principle of the FAIR principles is essential for ensuring that EvoLand data is readily available and accessible to users. EO data holds immense value for a wide range of applications, from scientific research to policymaking and disaster management. By adhering to the Accessibility principle, data providers can enable seamless access to EO data, thereby fostering its effective use and maximizing its societal impact.

Open data licenses, such as Creative Commons licenses, can be applied to clarify the terms of use, ensuring that users understand the permissions and restrictions associated with the data. In cases where open access is not feasible due to privacy, security, or licensing concerns, controlled access mechanisms can be implemented. Controlled access ensures that sensitive or restricted EvoLand data is accessible to authorized users while protecting privacy and ensuring compliance with legal and ethical frameworks. Access can be granted through secure data portals or platforms that require authentication or authorization processes, allowing only approved individuals or organizations to access the data.

Furthermore, to enhance accessibility, EO data should be provided in formats that are compatible with a wide range of software tools, applications, and platforms. Commonly used formats such as GeoTIFF, NetCDF, or standard geospatial formats like Shapefiles or Keyhole Markup Language (KML) facilitate interoperability and ease of use across different software environments. Additionally, the provision of Application Programming Interfaces (APIs) and web services, following standardized protocols such as OGC standards, allows users to programmatically access and integrate Earth observation data into their own systems and workflows.

Efforts should also be made to ensure the long-term accessibility of Earth observation data. This involves considerations for data preservation, including data archiving, backup strategies, and long-term storage solutions. Data repositories, data centers, and data infrastructures play a crucial role in preserving and providing ongoing access to Earth observation data, ensuring its availability for future generations of researchers and users.

4.3 Interoperability

The Interoperability principle of the FAIR principles is of paramount importance when dealing with EvoLand data. Achieving interoperability ensures that these diverse datasets can be integrated, combined, and analysed effectively, enabling comprehensive understanding and meaningful insights into the Earth system.

To promote interoperability, EvoLand data should adhere to standardized formats, vocabularies, and ontologies. The use of internationally recognized standards, such as those developed by OGC, ensures compatibility and facilitates seamless exchange of data across different systems and software tools. Standardized formats like GeoTIFF, NetCDF, or raster and vector formats enable interoperability between different software applications and platforms, allowing for efficient data sharing and analysis.

In addition to standardized formats, consistent and well-defined metadata play a crucial role in facilitating interoperability. Metadata should adhere to commonly accepted schemas, such as those defined by the ISO or the OGC, ensuring that essential information about the data, such as its temporal and spatial coverage, resolution, and acquisition parameters, can be understood and interpreted consistently by users and software systems. This facilitates the integration and fusion of diverse datasets, enabling researchers to combine data from multiple sources and perform meaningful cross-analysis.

Standardized interfaces and protocols, such as those defined by the OGC, facilitate the interoperable access and use of Earth observation data. Protocols like Web Map Service (WMS) and Web Coverage Service (WCS) enable the retrieval of geospatial data over the web, promoting interoperability by allowing data to be accessed and integrated into various applications and platforms. Application Programming Interfaces (APIs) provide programmatic access to data services, enabling developers to build applications that can seamlessly interact with and utilize Earth observation data.

Semantic interoperability is another aspect of the Interoperability principle. It involves the use of shared and well-defined vocabularies and ontologies, ensuring a common understanding of data across different disciplines and communities. By employing standardized vocabularies, such as those defined by the OGC or other domain-specific ontologies, data providers can enhance the integration and exchange of Earth observation data, supporting cross-disciplinary research and analysis.

4.4 Reusability

The Reusability principle of the FAIR principles plays a vital role in maximizing the value and impact of EvoLand data by enabling its reuse beyond its original context. Ensuring the reusability of this data promotes efficiency, collaboration, and knowledge advancement.

To enhance reusability, EO data should be well-described and accompanied by comprehensive metadata. The metadata should provide information about the data's provenance, collection methods, processing steps, quality assessment, and any limitations or caveats associated with the data. By providing clear and transparent documentation, data users can assess the suitability and reliability of the data for their specific needs and make informed decisions about its reuse.

The availability of standardized data formats and structures contributes to the reusability of Earth observation data. Adhering to widely accepted formats, such as GeoTIFF, NetCDF, or standard geospatial formats like Shapefiles or KML, ensures that data can be easily accessed and interpreted by a variety of software tools and platforms. This enables seamless integration and analysis of Earth observation data with other datasets, promoting its broader usability and reusability.

Documentation of data processing workflows and methodologies is crucial for ensuring reproducibility and enabling data reuse. By providing clear instructions and descriptions of data processing steps, including algorithms used, parameters applied, and any necessary pre-processing or calibration procedures, data users can replicate and validate the results obtained from the data. This promotes scientific rigor and fosters collaboration and transparency within the research community.

The provision of usage licenses and clear terms of use is essential for facilitating the reusability of Earth observation data. Data providers should specify the conditions under which the data can be reused, including any restrictions, attribution requirements, and potential commercial or non-commercial use limitations. By providing this information upfront, data users can understand and comply with the licensing terms, ensuring responsible and ethical data reuse.

Data accessibility is closely tied to reusability. By making Earth observation data easily accessible through open repositories, data portals, or web services, data providers enable researchers and practitioners to discover, access, and utilize the data for various purposes.

5 Data documentation

5.1 Data Description

Accurate and comprehensive data description is crucial for effective data management, interpretation, and sharing of Earth Observation (EO) data. A well-documented data description provides essential information about the data's characteristics, origin, and processing steps. Here are key components to consider when describing EvoLand data:

- **Dataset Identification**: Clearly identify each dataset, including a unique identifier, title, and a brief summary that outlines the purpose and scope of the data. Include relevant information such as the mission or project name, acquisition platform (satellite, sensor), and any specific data version or release.
- **Spatial and Temporal Coverage**: Define the spatial extent and resolution of the data, indicating the geographic area covered and the level of detail captured. Specify the coordinate reference system used, such as latitude-longitude or UTM (Universal Transverse Mercator). Additionally, specify the temporal coverage of the dataset, including the acquisition period and frequency of data collection.
- **Data Variables and Parameters**: Provide a comprehensive list of the variables or parameters included in the dataset. For each variable, describe the data type (e.g., numerical, categorical), units of measurement, and any associated uncertainties or error estimates. Include detailed definitions and explanations of the variables to ensure a common understanding among data users.
- **Data Format**: Specify the file format used to store the data, ensuring compatibility and accessibility. Common formats for EO data include GeoTIFF, NetCDF, HDF5, or proprietary formats specific to certain sensors or platforms. State any compression or encoding techniques applied to the data and provide information on how to access and open the files.
- **Data Processing Steps**: Document the processing steps applied to the raw EO data, including pre-processing, calibration, or transformation procedures. Provide an overview of the algorithms, software tools, and parameters used during data processing. Detail any specific corrections, atmospheric corrections, or geometric transformations performed to derive the final dataset.
- **Metadata**: Provides essential information about the dataset, facilitating data discovery, understanding, and interoperability. Include metadata elements such as data source, acquisition date and time, sensor characteristics, spatial resolution, and spectral bands. Document processing history, quality assessment measures, and any known limitations or caveats associated with the data.
- **Data Documentation**: Provide detailed documentation, such as user guides, technical manuals, or data dictionaries, to support data

interpretation and analysis. Describe the data structure, file naming conventions, variable naming conventions, and any specific data organization schemes used. Include information on data units, coordinate systems, and any transformations applied to the data.
- **Data Access and Usage**: Specify the terms and conditions for data access and usage. Determine if the data will be freely available, restricted to specific users or organizations, or subject to licensing agreements. Define the citation guidelines for referencing the dataset in publications or research outputs. Consider ethical and legal considerations regarding data usage, intellectual property rights, and privacy concerns.
- **Data Updates and Versioning**: Establish a plan for data updates, including periodic updates or revisions. Define the versioning system to track changes and updates made to the dataset. Clearly indicate any updates, modifications, or corrections to previous versions, and provide documentation on the reasons for the changes.

5.2 Data processing steps

Data processing is a critical component of EvoLand's data management, enabling the transformation of raw sensor data into useful information products. A well-defined data processing workflow ensures consistency, accuracy, and reproducibility of data analysis.

- **Data Pre-processing**: The first step in data processing is data pre-processing, which involves preparing the raw sensor data for further analysis. This includes data calibration, sensor corrections, and data formatting to ensure data quality and compatibility across different sensors or platforms. Pre-processing may also involve the removal of noise, atmospheric corrections, geometric corrections, or spatial resampling to align data from different sources.
- **Data Fusion and Integration**: involve combining EO data from multiple sensors, platforms, or time periods to create composite or derived datasets. This step aims to enhance the information content and improve the spatial or temporal resolution of the data. Fusion techniques may include image registration, data fusion algorithms, or multi-sensor data integration methods.
- **Feature Extraction**: involves identifying and extracting relevant features or characteristics from the EO data. This may include vegetation indices, land cover classes, water bodies, urban areas, or other specific features of interest. Feature extraction techniques may involve spectral analysis, texture analysis, object-based image analysis (OBIA), or machine learning algorithms to automatically classify or extract features from the data.
- **Data Analysis and Modelling**: Once the relevant features have been extracted, data analysis and modelling techniques are applied to derive meaningful information from the EO data. This may include statistical analysis, trend analysis, change detection, classification algorithms, or predictive modelling approaches. The analysis aims to reveal patterns,

relationships, or trends in the data, enabling scientific research, decision-making, or operational applications.

- **Validation and Quality Assurance**: are crucial steps to ensure the accuracy and reliability of the processed EO data products. This involves comparing the processed data with ground truth or reference data, assessing the data quality, and quantifying uncertainties or errors. Validation may include field measurements, independent sensor data comparisons, or accuracy assessments using established reference datasets or ground-based observations.
- **Metadata Documentation**: is essential to provide information about the processed EO data, including the processing steps, algorithms used, parameters, and assumptions made during processing. It includes information on data sources, data processing software, versions, and any relevant processing history. Metadata documentation helps ensure data reproducibility, understanding, and proper attribution of the processed data.
- **Version Control**: Implement a version control system to manage different versions of the processed data. Maintain records of changes made during the data processing workflow, enabling traceability and reproducibility of data products. Version control ensures that updates or modifications to the processing steps or algorithms are properly documented and tracked.
- **Documentation and Standard Operating Procedures** (SOPs): Document the specific steps, parameters, and software tools used in each processing stage. SOPs provide guidance to data processors, ensure consistency in data processing, and facilitate knowledge transfer within the organization or research community.

6 Data Security and Access Control

6.1 Data Security Measures

Earth Observation (EO) data often contains sensitive information and valuable intellectual property that requires robust security measures to protect against unauthorized access, data breaches, and misuse. A comprehensive data management plan should include appropriate data security measures to ensure the confidentiality, integrity, and availability of EvoLand data.

- **Access Control**: Implement access control mechanisms to restrict data access to authorized individuals or user groups. Define user roles and permissions based on job responsibilities and the need-to-know principle. Use strong authentication methods, such as passwords, two-factor authentication, or biometric authentication, to verify user identities. Regularly review and update access privileges as needed.
- **Data Encryption**: Apply encryption techniques to protect data during storage and transmission. Encrypt sensitive data at rest using methods such as full-disk encryption or file-level encryption. Implement secure protocols when transmitting data over networks to prevent eavesdropping or tampering. Encryption ensures that even if unauthorized individuals gain access to the data, it remains unreadable and unusable.
- **Secure Storage**: Store EO data in secure and controlled environments. Implement physical security measures, such as access controls, surveillance systems, and fire suppression systems, to protect data storage facilities. Use reliable and redundant storage systems to prevent data loss due to hardware failures or disasters. Regularly back up data and test data restoration procedures to ensure data recoverability.
- **Data Transfer and Exchange**: Implement secure protocols and encryption mechanisms when transferring or exchanging EO data. Use secure file transfer methods, such as Secure File Transfer Protocol (SFTP) or secure cloud storage, to protect data during transit. Consider the use of secure virtual private networks (VPNs) when accessing or transferring data over public networks to establish secure and encrypted connections.
- **Data Audit and Monitoring**: Establish logging and auditing mechanisms to monitor data access, modifications, and system activities. Maintain audit trails that record user actions, system events, and data changes. Regularly review and analyse audit logs to detect any unauthorized access attempts or suspicious activities. Implement intrusion detection and prevention systems to identify and respond to security breaches promptly.
- **Data Backup and Disaster Recovery**: Implement robust backup and disaster recovery procedures to ensure data availability and minimize downtime in the event of data loss or system failures. Regularly back up data to secure off-site locations or cloud storage providers. Test backup and recovery processes to verify data recoverability and establish recovery time

objectives (RTO) and recovery point objectives (RPO) for different types of data.
- **Staff Training and Awareness**: Provide regular training and awareness programs to educate staff on data security best practices, policies, and procedures. Train employees on secure data handling, password hygiene, phishing prevention, and incident reporting. Foster a culture of data security by promoting vigilance and emphasizing the importance of protecting sensitive EO data.
- **Incident Response and Contingency Planning**: Develop an incident response plan to handle security incidents, including data breaches or unauthorized access. Establish a clear escalation process, incident reporting channels, and communication protocols. Regularly test and update the incident response plan to ensure its effectiveness. Develop a contingency plan to address potential disruptions, including cyber-attacks, natural disasters, or system failures, and establish procedures for data recovery and resumption of operations.

By implementing robust data security measures, researchers and organizations can safeguard the confidentiality, integrity, and availability of EvoLand data, maintaining trust and compliance with data protection standards. These procedures are already in place within the consortium.

6.2 User Access Levels

Effective data management involves controlling user access to EvoLand data to ensure data security, privacy, and compliance with data sharing policies. Implementing different access levels helps to ensure that users only have access to the data they need for their specific roles and responsibilities.

- **User Categories**: Define different user categories based on roles and responsibilities within the organization or research community. Common categories may include data owners, data managers, data analysts, researchers, external collaborators, and public users. Each category should have a clear description of the roles and privileges associated with it.
- **Access Levels**: Establish different access levels that align with the user categories. Determine the permissions and privileges granted to each access level. For example, data owners and managers may have full control over data, including the ability to modify metadata or grant/revoke access, while researchers may have read-only access for data analysis purposes.
- **Data Classification**: Classify EO data based on sensitivity, confidentiality, or data sharing restrictions. Identify different data classification levels (e.g., public, internal, confidential) and assign appropriate access levels to each category. This classification helps to determine the level of access granted to users and ensures compliance with legal and ethical requirements.
- **Access Requests and Approvals**: Define a process for users to request access to specific data or access level upgrades. Establish a review and approval system to assess the legitimacy of access requests. The process

may involve verifying user credentials, evaluating the purpose of access, and obtaining any necessary approvals from data owners or project leads.

- **User Authentication and Authorization**: Implement secure authentication mechanisms to verify user identities. This may include usernames and passwords, two-factor authentication, or single sign-on (SSO) systems. Upon authentication, authorize users based on their access levels to determine what data they can access and what actions they can perform on the data.
- **Data Sharing Agreements**: For external collaborators or partners, establish data sharing agreements that outline the terms and conditions of data access and usage. These agreements should specify the access levels granted, the duration of access, and any restrictions or obligations associated with the data. Ensure that users agree to adhere to these agreements before accessing the data.
- **Data Access Monitoring**: Implement monitoring mechanisms to track user access and actions performed on the data. Maintain audit logs that record user activities, including data access, modifications, and downloads. Regularly review these logs to identify any suspicious or unauthorized activities. Monitoring helps ensure compliance, detect potential security breaches, and support data governance and accountability.
- **Data Access Documentation**: Maintain detailed documentation of user access levels, including the access control policies, data sharing agreements, and any modifications or revocations. Document the process for granting or changing access levels, as well as any justifications or approvals associated with access requests. This documentation helps ensure transparency, accountability, and compliance with data management policies.

6.3 Data Sharing Policy

A well-defined data sharing policy ensures that EvoLand data is shared in a responsible and transparent manner, while protecting data integrity, privacy, and intellectual property rights.

- **Data Sharing Objectives**: Clearly state the objectives of data sharing in the EO context. This may include facilitating scientific research, supporting policy-making, enabling innovation, promoting data reuse, or fostering international collaboration. Align the data sharing policy with the broader goals and mission of the organization or research community.
- **Data Sharing Scope**: Define the scope of data sharing, specifying the types of EO data that are eligible for sharing. This may include raw sensor data, processed datasets, derived products, or metadata. Clarify any restrictions or limitations on sharing, such as proprietary data, sensitive information, or data subject to legal or ethical considerations.
- **Data Sharing Principles**: Establish guiding principles that govern data sharing activities. These principles may include openness, transparency, fairness, attribution, respect for intellectual property rights, and adherence

to relevant data protection and privacy regulations. Emphasize the importance of responsible and ethical data use by data recipients.

- **Data Sharing Formats and Standards**: Specify the preferred data sharing formats and standards to ensure interoperability and ease of data integration by data users. Use open and widely adopted file formats whenever possible to promote accessibility and avoid vendor lock-in. Encourage the use of standard metadata formats to facilitate data discovery and understanding.
- **Data Access Mechanisms**: Outline the mechanisms for accessing and obtaining EO data. This may include data portals, web services, application programming interfaces (APIs), or direct data transfers. Define any access restrictions, licensing agreements, or user agreements that data recipients must adhere to when accessing the data.
- **Data Attribution and Citation**: Establish guidelines for data attribution and citation to ensure proper recognition and acknowledgment of data contributors. Specify the preferred citation format or provide examples of how to cite the data. Encourage data recipients to acknowledge the source of the data and to provide appropriate references in publications or derived works.
- **Data Embargo and Release Policies**: Address the issue of data embargo and release timing. Determine whether there are specific periods during which data may be subject to limited access, such as for validation, quality control, or exclusive research purposes. Clearly communicate the embargo periods and outline the process for releasing the data to the broader community.
- **Data Governance and Compliance**: Outline the roles and responsibilities of data providers, data managers, and data users in adhering to the data sharing policy. Establish a data governance framework to monitor and enforce compliance with the policy. Identify any legal, ethical, or privacy considerations that data providers and users must be aware of and comply with.
- **Data Sharing Agreements**: Consider the need for formal data sharing agreements when sharing EO data with external collaborators, partners, or data users. These agreements should outline the terms and conditions of data sharing, including access rights, restrictions, data usage limitations, intellectual property rights, and confidentiality provisions.

7 Data Preservation and Long-Term Accessibility

7.1 Data Archiving

Data Archiving is a critical component of data management for EvoLand data, ensuring long-term preservation and accessibility. This subchapter outlines the procedures and considerations for archiving EvoLand data.

- **Archiving Objectives**: specify the objectives of data archiving, which may include ensuring data integrity, long-term preservation, reproducibility of research, compliance with funding requirements, and facilitating data sharing and reuse.
- **Archiving Plan:** including the selection of appropriate archiving technologies and infrastructure. Consider factors such as data volume, format, metadata requirements, and data retention policies. Identify the responsible entity or data centre that will oversee the archiving process.
- **Archiving Procedures:** outline the procedures for archiving EO data. This includes preparing data for archiving, organizing and packaging data according to archival standards, and transferring the data to the designated data repository or archive. Specify any data format conversion, metadata documentation, or data transformation steps required for archiving.
- **Metadata Documentation**: emphasize the importance of documenting metadata for archived EO data. Specify the metadata standards and elements that need to be included, such as data collection parameters, processing methodologies, calibration information, and quality assessment results. Ensure that the metadata is properly formatted and complies with relevant metadata standards.
- **Data Retention Period:** define the data retention period for archived EO data. Consider the requirements of funding agencies, research communities, or legal obligations. Specify the duration for which the data will be retained and when data may be eligible for disposal or transfer to long-term preservation archives.
- **Data Accessibility:** Address the issue of data accessibility for archived EO data. Specify the conditions and procedures for accessing the archived data, including any restrictions, embargoes, or user agreements. Provide information on how researchers and users can request access to archived data, including the necessary documentation or approval processes.

7.2 Backup Procedures:

Backup procedures are essential for data management, ensuring data integrity and protection against data loss or system failures. This subchapter outlines the backup procedures for EvoLand data.

- **Backup Objectives:** Specify the objectives of backup procedures, including data protection, disaster recovery, minimizing data loss, and maintaining data availability. Emphasize the importance of regular backups to ensure data resilience.
- **Backup Schedule:** Define the backup schedule for EO data. Determine the frequency of backups based on data criticality and update frequency. Specify daily, weekly, or monthly backup schedules, or differentiate between full backups and incremental backups.
- **Backup Methods:** Describe the backup methods to be used for EO data. This may include full disk backups, file-level backups, or cloud-based backups. Determine whether backups will be stored on-site, off-site, or in cloud storage, considering factors such as data volume, security, and disaster recovery requirements.
- **Backup Testing and Verification:** Establish procedures for regularly testing and verifying backups. Conduct periodic tests to ensure data recoverability and integrity. Verify the backup files' accessibility and confirm that they can be restored successfully in case of data loss or system failure.
- **Backup Storage Security:** Address the security measures for backup storage. Ensure that backup storage systems are secure, protected from unauthorized access, and comply with relevant data protection regulations. Implement encryption for backup files to prevent unauthorized data access or data leakage.
- **Backup Retention:** Specify the retention period for backup data. Determine how long backup files will be retained before being overwritten or discarded. Consider legal requirements, data recovery time objectives (RTO), and data versioning needs.

7.3 Data Versioning

Data versioning is a crucial aspect of data management for Earth Observation data, allowing for the tracking and documentation of changes made to datasets over time.

- **Versioning Objectives:** Specify the objectives of data versioning, which may include maintaining a historical record of changes, enabling reproducibility of research results, facilitating data comparison and analysis, and ensuring data integrity and transparency.
- **Versioning Plan:** Describe the overall plan for data versioning, including the identification of key milestones or events that trigger a new version. Define the versioning scheme or naming convention to be used, ensuring consistency and clarity in identifying different versions of the data.
- **Version Control System:** Identify and implement a version control system for EO data. This may include using dedicated version control software or leveraging existing tools and platforms commonly used in software development, such as Git or Subversion. Outline the procedures for creating, managing, and documenting data versions within the chosen version control system.

- **Versioning Procedures:** Outline the procedures for creating and managing data versions. Describe the steps to be followed when making changes to the data, including data collection updates, data processing modifications, or metadata enhancements. Specify the documentation requirements and metadata elements that need to be updated with each new version.
- **Version Documentation:** Emphasize the importance of documenting each data version. Specify the metadata elements that need to be captured for each version, including version number, date, description of changes, and reasons for the modifications. Document any specific instructions or guidelines for referencing or citing different data versions.
- **Data Retrieval and Access:** Address the procedures for accessing and retrieving specific data versions. Define how users can identify and access different versions of the data, ensuring that the appropriate version is selected based on specific research or analysis needs. Specify any access restrictions or embargoes that may apply to certain versions of the data.
- **Versioning Governance:** Establish a governance framework for data versioning. Define the roles and responsibilities of data owners, data managers, and data users in managing and maintaining data versions. Identify the process for reviewing and approving changes that lead to new data versions, ensuring proper documentation and data integrity.
- **Versioning Auditing:** Implement periodic audits to ensure compliance with versioning procedures. Conduct reviews to verify that data versions are created and managed according to the defined versioning plan. Address any discrepancies or inconsistencies identified during the auditing process and take appropriate corrective actions.
- **Data Versioning and Archiving:** Integrate data versioning with the data archiving process. Specify how different versions of the data will be archived and retained over time. Ensure that data archiving procedures consider the preservation of all relevant versions, enabling the retrieval of specific historical versions if needed.

Evoland
LAND MONITORING EVOLUTION

EvoLand Consortium

vito remote sensing

International Institute for Applied Systems Analysis IIASA www.iiasa.ac.at

CLS COLLECTE LOCALISATION SATELLITES

CESBIO

SINERGISE

JOANNEUM RESEARCH

cnes

DLR

GAFAG an e-GEOS (ASI / Telespazio) Company

evenflow