guardtime

SEPTEMBER 2025

# Disruptive Technologies for EO Data Provenance

# Use Case Definition

ID: GT-TRACE4EO-UC-0001

Version: 1.0

Status: Released

Date of Issue: 09/09/2025

Classification: ESA UNCLASSIFIED – For Official Use Only

GUARDTIME.COM

# Contents

# Document versioning

| Date (MM.YYYY) | Version | Author | Changes |
|---|---|---|---|
| 06.2025 | 0.1 | Tuuli Lõhmus (Guardtime) | Initial draft |
| 09.2025 | 1.0 | Patryk Grzybowski (CloudFerro) Ludovic Augé (Alt239) | First release |

# 1. Introduction

## 1.1. Purpose and scope

The purpose of the Use Case Definition document is to provide a practical overview of how traceability can be applied in Earth observation data workflows. It demonstrates this through two complementary use cases, one domain-specific (agriculture) and one general (AI/ML), to highlight critical requirements for tracking data and model processes.

In particular, the document aims to:

- Demonstrate applicability - show how traceability works in both specific and general AI/ML contexts.
- Highlight critical requirements - emphasize systematic tracking of input data, metadata, and model lifecycle.
- Ensure reproducibility and trustworthiness - enable validation and reliable comparison of results.
- Provide a future-proof reference - present a solution resilient to technological changes and vendor lock-in.

By offering a vendor-independent approach, the document supports reproducibility, transparency, and long-term trustworthiness of AI applications in Earth observation.

## 1.2. Structure of the document

The document consists of the following chapters.

- **Introduction** – this section provides an introduction and a summary of what will be presented in the document. It also includes requirements related to the implementation of activities and references the baseline documents on which this work is based.
- **Potential Use Cases for EO Data Provenance** - This chapter presents two use cases in the context of traceability. The first one is domain-specific but

broadly refers to typical machine learning workflows. The second one addresses AI data and models in a more generic way. Within this chapter, the selection of these use cases is explained, also taking into account feedback from potential stakeholders. It presents which data will be traced and describes the processes by which traceability was applied by the users.

- **Conclusion** – contains the conclusion.

# 2. Potential Use Cases for EO Data Provenance

## 2.1. Use Case: Forecasting Cereal Yields Using Traceability combined with Sentinel and Meteorological Data

Traceability in agriculture is increasingly important for optimizing resources, improving productivity, and ensuring food security under climate change. Earth Observation (EO) data from the European Space Agency (ESA) Sentinel satellites provides high-resolution, multi-spectral imagery that enables detailed monitoring of crop growth and environmental conditions. When combined with meteorological data, these inputs support reliable crop yield forecasting by offering a continuous, verifiable record of the factors influencing productivity.

Such traceable workflows enhance transparency and credibility, as every step of the forecasting process can be verified, ensuring that results are based on reliable EO and weather data. This not only aids government agencies in policy-making and funding decisions but also gives farmers confidence in the recommendations provided.

Additionally, traceability improves risk management by documenting extreme events, such as droughts or pest outbreaks, which helps validate insurance claims and strengthens trust between insurers and farmers. The approach is also highly scalable, allowing forecasting systems to be adapted to new regions, crops, or technologies while maintaining methodological integrity, enabling consistent application by both local agencies and multinational companies.

By integrating EO data, traceability, and advanced analytics, this use case fosters resilience, transparency, and sustainability in agricultural systems, while laying the groundwork for future innovations in data-driven farming.

### 2.1.1. Use case identification

The crop yield forecasting use case was developed through a multi-step process. Insights from Earth Observation conferences highlighted the sector's demand for reliable, data-driven yield prediction, while feedback from clients and stakeholders

revealed the specific needs of agricultural agencies. The project team's expertise in EO and agriculture allowed them to refine potential applications, and consultations with institutions such as the Joint Research Centre (JRC) and the Polish Space Agency (POLSA) confirmed the high relevance of yield forecasting for operational users. Based on these inputs, crop yield forecasting was chosen as a priority use case, and its methodology, particularly in data processing, is designed to be transferable to other EO-based applications beyond agriculture.

### 2.1.2. Introduction and overall view

Accurate and timely yield forecasts are critical for food policy, agricultural planning, and government decision-making. Machine and deep learning approaches, combined with EO data and meteorological indicators, allow for more precise and frequent forecasts.

In this use case, a forecasting workflow will be demonstrated for selected NUTS-2 regions in the European Union, using freely available datasets. Key inputs include:

- Vegetation indices from Sentinel-3 OLCI (optical) and SLSTR (thermal),
- Agro-meteorological data from ERA-5,
- Crop yield statistics from Eurostat.

The trained machine learning models will produce yield predictions for specific crops in the selected regions. Importantly, the approach is designed to be transferable, allowing application to other crops and geographic areas.

### 2.1.3. Methods

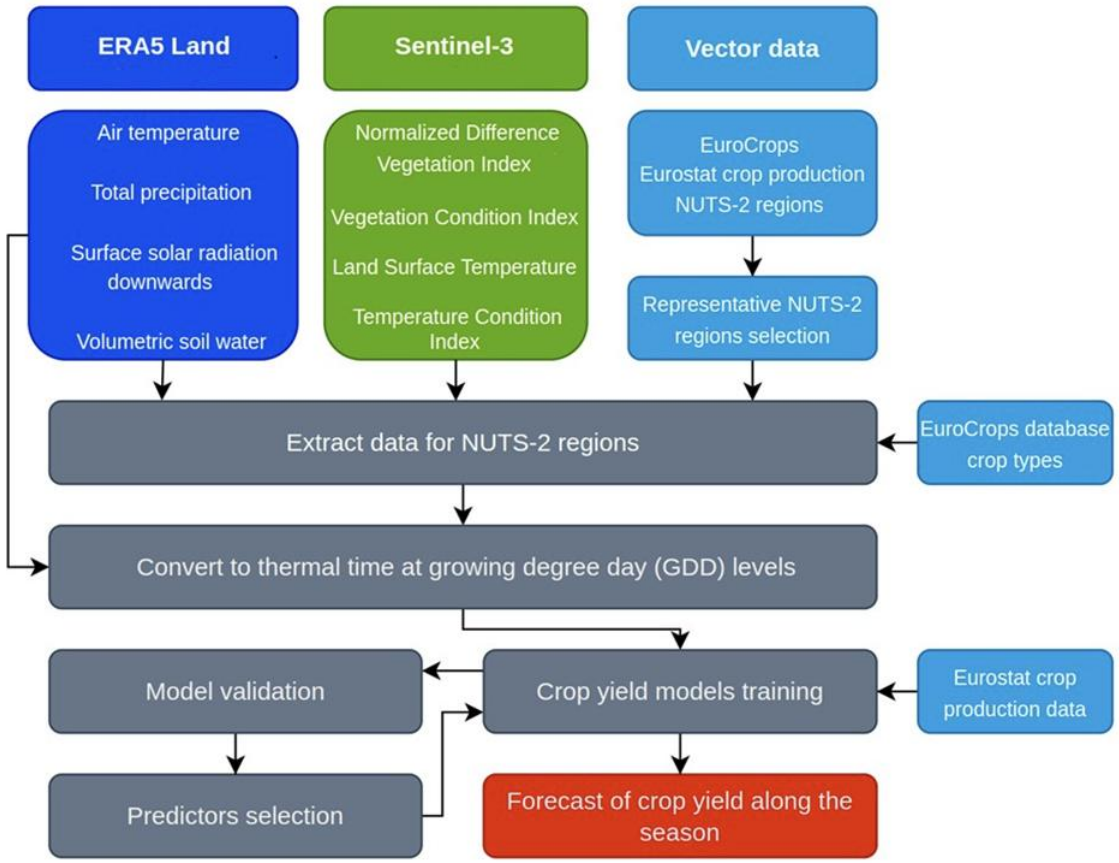
The solution will be built based on machine learning algorithms that leverage data fusion of satellite and climate reanalysis data including: daily vegetation condition indicators provided by Sentinel-3 OLCI (optical) and SLSTR (thermal) imagery, as well as air temperature, total precipitation, surface radiation, and soil moisture derived from ERA-5 Land climate reanalysis generated by the European Centre for Medium-Range Weather Forecasts (ECMWF).

The reference data for the crop yield forecasting model will consist of official yield statistics provided by Eurostat at NUTS-2 level database and EuroCrops, which is

a dataset linking all publicly available crop reporting datasets from European Union countries. Due to the varying availability of yield data and crop locations across EU countries, NUTS-2 regions with complete information on agricultural production culture and typical land use characteristics will be selected for this use case.

The crop yield forecasting algorithm will be based on thermal time (growing degree days derived from ERA-5 data) to more accurately track the crop development stages. The predictors will be extracted and calculated based on vector locations of the types of crop types from the EuroCrops database. These predictors will include the following variables transformed to thermal time at several growing degree day (GDD) levels serving as a proxy for crop development stage: total precipitation, soil moisture, surface radiation, air temperature, Normalized Difference Vegetation Index (NDVI), Vegetation Condition Index (VCI), Land Surface Temperature (LST), Temperature Condition Index (TCI), all based on GDD, as well as the maximum NDVI during the growing season. Primarily NDVI computed from Sentinel-3 data will be applied to the model. However, given its higher spatial resolution and capability to more accurately determine the state of vegetation growth, NDVI derived from Sentinel-2 data will also be tested. Moreover, analyses of plant growth conditions during the season will be calculated based on air temperature-based growing degree day (GDD).

A fusion of satellite data with agrometeorological information will be then implemented to serve as input to the machine learning model. Based on multi-year yield reference data for selected crops in NUTS-2 units, along with multi-year predictors, machine learning model training will be conducted to develop a yield forecasting model for each region. During the training process, the recursive feature elimination will be used to derive an optimal set of yield predictors for each administrative unit, which will ultimately be employed by the Extreme Gradient Boosting regressor to forecast yields using official yield statistics as a reference. The model will predict crop yields and generate the final outputs (tabular, graphical maps), which can be used for dashboard visualization and time series analyses to determine the variability of yields by year for selected crop types in NUTS-2 regions. In addition, crop growth conditions based on thermal time will be determined for each selected region along with the most important predictor variables. Model validation will be performed using a cross-validation method with a leave-one-year-out approach and the model's prediction accuracy metrics will be calculated.




## 2.1.4. Data

A comprehensive list of input data used in the system, including satellite, agrometeorological, and ancillary materials:

- Satellite data:
  - Sentinel-3 (300 m) NDVI, TCI, VCI,
  - Sentinel-3 (1000 m) LS;
- Climate reanalysis ERA5-Land hourly data:
  - 2m air temperature,
  - Total precipitation,
  - Surface solar radiation downwards,
  - Volumetric soil water layer 1 (0-7 cm) and 2 (7-28 cm) ;
- Crop production related data:
  - Statistical regions NUTS-2 (nomenclature of territorial units for statistics) polygon vector,
  - EuroCrops vector database for delineation of extent and distribution of agricultural fields with a specific crop type,

- Crop production information: Eurostat - Crop production in EU standard humidity by NUTS 2 regions - database.

### 2.1.5. Explanation and Development of Traceability in the Context of the Use Case

**Why is this important?**

- In agricultural planning, yield forecasts directly influence food security policies, economic strategies, and operational decisions. Errors in forecasting can lead to economic and logistical challenges.
- In terms of EU policies, yield forecasting could serve as a basis for subsidies and other agricultural financing across countries.

Regarding the points above, it is necessary to provide a reliable and consistent system that allows for verification at each stage of activity. In this respect:

- Traceability ensures compliance with scientific and regulatory standards, fostering trust in the predictions.
- It allows practitioners to revisit and refine the process in response to audits, disputes, or evolving needs.

**What needs to be tracked?**

**Input Data**

- Identifiers of Sentinel-3 images (e.g., file name, acquisition date, orbit, baseline).
- Image metadata (e.g.: observation angle, processing level, spatial resolution, calibration information).
- Meteorological data (e.g.: source, acquisition time, resolution, model version used for computation).
- Eurostat yield statistics by NUTS-2 region (e.g.: name, reporting year, update frequency)
- EuroCrops database (e.g.: vector file identifiers, attributes)

### Processing Parameters

- Algorithms and configurations used for processing satellite data (e.g., NDVI calculation formula, thresholds for VCI or TCI).
- Parameters for data fusion, such as spatial interpolation of ERA-5 data to match Sentinel-3 resolutions.
- Transformation of variables (e.g., conversion of climatic variables to thermal time using GDD calculations).

### Indices

- Mathematical formulas for vegetation indices (e.g., NDVI = (NIR - Red) / (NIR + Red)) and GDD.
- Parameters for growing degree day (GDD)-based transformations of soil moisture, precipitation, and radiation.
- Selection and calibration of predictors for each NUTS-2 region.

### Predictive Model

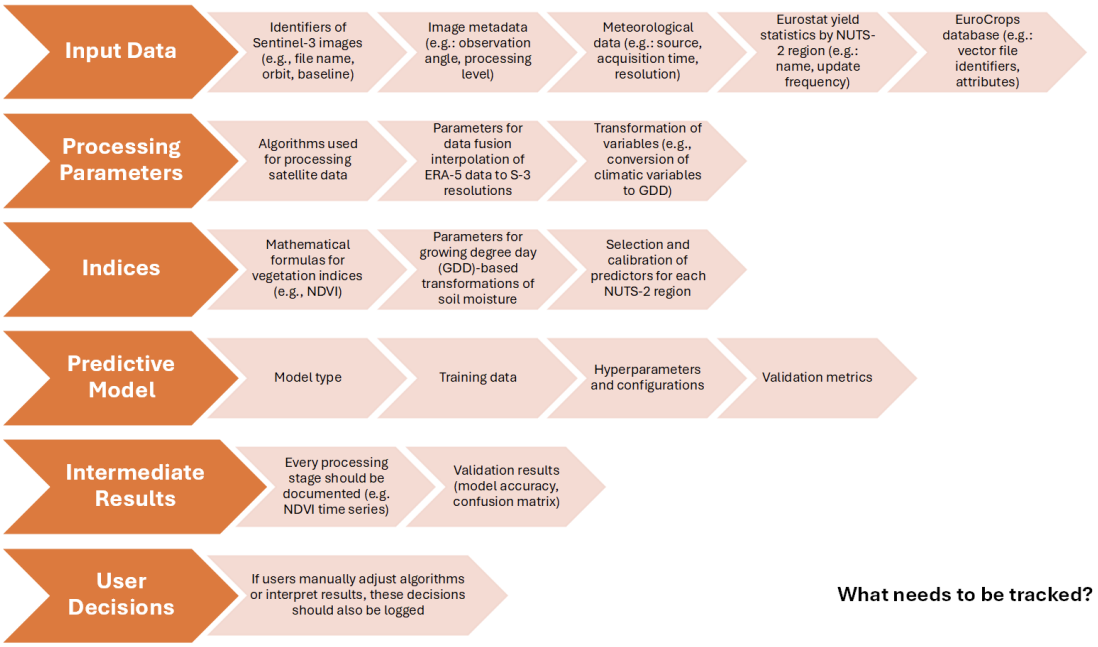
- Model type (e.g.: Extreme Gradient Boosting (XGBoost) regressor).
- Training data (e.g. Multi-year crop yield statistics and corresponding predictors for the selected regions).
- Hyperparameters and configurations (e.g., learning rate, tree depth).
- Validation metrics (e.g.: RMSE)

### Intermediate Results

- Every processing stage should be documented (e.g., soil moisture maps, NDVI time series).
- Validation results (model accuracy, confusion matrix).

### User Decisions

- If users manually adjust algorithms or interpret results, these decisions should also be logged.




### Auditing Predictive Processes

Auditing predictive processes allows for detailed examination of how and why specific forecasts were generated. This ensures the ability to answer questions like:

- Where did the input data come from?
- What processing steps were applied to the data?
- Why did the model produce this particular forecast?

## 2.1.6. Summary of stakeholders meetings

Discussions with potential users confirmed both the relevance and the broader applicability of the traceability approach. While JRC emphasized the critical importance of metadata traceability for ensuring reproducibility of results, POLSA highlighted the need for practical tools to verify the quality and credibility of delivered products. Both institutions see strong potential for applying the methodology beyond agriculture, underlining its transferability and scalability.

JRC

- Traceability of input data identified as the most challenging and critical aspect (frequent changes in algorithms, baselines, and methods for Sentinel data).

- Need to distinguish between the input data layer and the model/user activities layer ("human layer").
- Ensuring reproducibility requires accounting for differences in baselines and processing versions (e.g., verifying forecasts months later when input data have been reprocessed).
- Managing evolving metadata in input datasets is a fundamental requirement.
- These challenges affect not only agriculture but all EO-based use cases.

POLSA

- Strong interest in traceability as a key recipient of agricultural maps and layers.
- Concern about discrepancies between declared and actual product accuracy; manual verification is too time-consuming.
- High interest in testing the approach on a specific use case, with emphasis on transferability to other cases.
- Key requirement: verification of input data and model details (parameters, hyperparameters) provided by contractors.
- Transferability: methodology applicable to processes beyond agriculture (input → indicators → ML model → results).
- Scalability: desirable to scale analysis to national level, or alternatively validate random 10–20% of the surface area.

### 2.1.7. User story (Audit example)

As a European agricultural policy maker, researcher, or farmer,

I want crop yield information that is fully traceable and verifiable throughout the data collection, processing, and modeling workflow,

So that I can rely on the forecasts for decision-making, policy design, subsidy allocation, and operational planning, knowing that the data and results are accurate, reproducible, and free from unintended manipulation.

Acceptance Criteria:

- All input data sources, versions, and processing steps are recorded and auditable.
- Machine learning models used for yield prediction are fully documented, including parameters, training data, and validation metrics.
- Results and published outputs (maps, statistics) can be traced back to their original inputs and processing steps.
- The workflow is adaptable to different regions, crops, or EO-based applications while maintaining methodological integrity.

Traceability in crop yield forecasting ensures that data, models, and results are fully verifiable, improving trust, reproducibility, and decision-making for policy makers, researchers, and farmers. The presented workflow is a scalable and adaptable framework that can be applied to other EO-based applications, supporting reliable, high-quality information across different regions and crops.

# 2.2 Traceability for AI Modelling

## 2.2.1 The needs and benefits of Traceability in AI

In order to comply with AI users and the related Regulation, Traceability has now become a de facto necessary tool in order to comply with AI's act in particular on the Quality, Documentation and Preparation process of the input Data or input Models (in the case of model refinement), not to forget on the Explainability need which requires to be able to relate its inferences with its training material.

As per a Traceability Service, we could then summarize the rationale of the needs and expected outcomes according to the following families:

Trust and Transparency:

Trusting an AI model calls for transparency and traceability of its inputs. From a quality point of view, self-declaration of an AI model provider cannot comply with a transparency need, hence this service has to be provided by a 3rd party, acting independently from the AI provider.

Quality and Genuineness Assessment

In a world where AI models are becoming an obvious and daily tool, leveraging services, a natural question for the consumers of those services will be about the genuineness of those models, in particular about the input data used, in order to know which of them come from an authenticated source, or which of them are the result of synthetic processings from initial data or even which of them are the result of generative AI used to increase the training population of data.

Expected outcomes and value :

Ethics and Fairness of an AI model are currently natural and obvious concerns, being partly addressed by Traceability. It shall be noted that, from the considerations above, as Traceability can contribute to Quality assessment of an AI model, it has an indirect impact on the economic viability of AI-based services.

Indeed, Quality concerns might become also of economical importance as a differentiator between several usages derived from different AI models.

We can imagine that in the near future, Traceability information might yield or contribute to a Quality ranking of AI models, hence having an important role in scoring the underlying services, making it a marketing and pricing argument.

### 2.2.2 About Dataionics and Traceability :

About Dataionics

Dataionics is a newly created startup, aiming at offering seamless and federated access to satellite imagery for data consumers, particularly for AI training use cases involving large geospatial datasets.

In addition to delivering raw or preprocessed imagery, Dataionics proposes a traceability service dedicated to the input data used in AI model development. This service allows AI model providers to rely on an external and trustworthy Traceability Manifest describing:

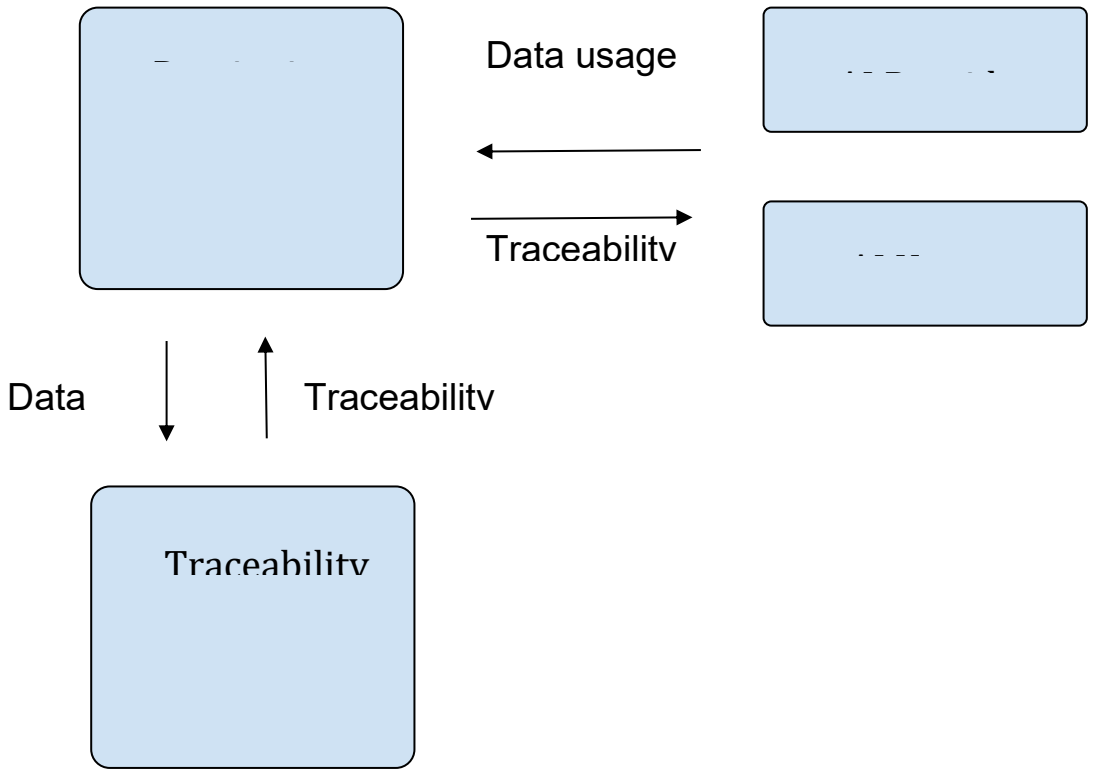
- the full list of input data used (IDs, origin, date, location, processing history),
- and any transformation pipeline applied prior to model training.

Such traceability is not only valuable for transparency and quality control—it is increasingly a regulatory requirement. The European AI Act, especially Articles 10, 12, and 13, mandates that providers of high-risk AI systems and foundation models:

- document the origin, characteristics and processing of training datasets;
- maintain logbooks of training and inference activities;
- and enable external auditability of the data sources involved.

To meet these needs, Dataionics positions itself as a neutral Recording Entity: a third-party infrastructure component collecting and recording the data lineage of satellite imagery used in AI workflows. This ensures that the provenance, structure, and usage of each dataset can be proven at any time—independently of the AI model provider—and exported through standardized formats to support regulatory compliance, model validation, or risk mitigation.

By combining sovereign access to spatial data with continuous metadata logging, Dataionics helps organizations align with the AI Act's data governance requirements while strengthening trust in Earth Observation-based AI systems.

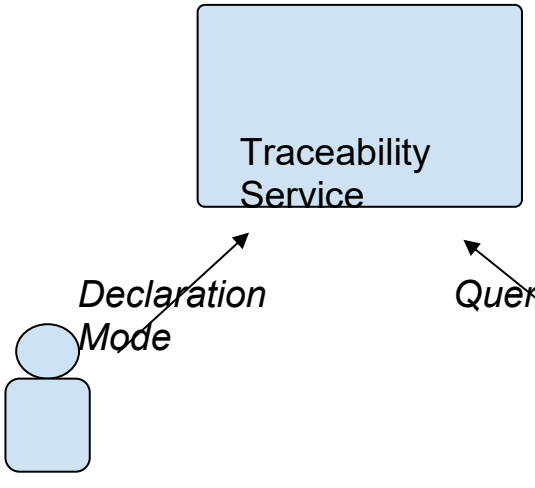



Typical user for Dataionics :

A user of Dataionics can be a data/AI consumer, or a data/AI provider.

He can either use the Traceability Service according 2 modes, with a different role:

- As a Traceability Contributor, in Declaration mode, to Provide traceability information about :
  - Every piece of imagery (full image or tiles) he is using for further image processing or to train or specialize an AI model
  - An already existing Model (like a Foundation Models for instance), he is using as a base for further refinement
- As a Traceability Inquirer, in Query Mode, to Request traceability information related to :
  - 1 to N images,
  - 1 to N tiles from a given Image
  - An existing AI model




<<Traceability Contributor>>, typically a data scientist building an AI model

<<Traceability Inquirer >>, Typically, a user of an AI model, looking for evidences of the AI model training material quality

### 2.2.3 Requirements

Note: The following requirements are expressed from the point of view of a typical User of Dataionics, as a Traceability Contributor or as a Traceability Enquirer:

User needs in Declaration Mode:

At any time, a user shall be able to declare to the Traceability 1 or N inputs by stating for each input:

- an ID,
  - for an image or a model
  - optionally, when the declared data is a tile, its footprint within the original image
- data traceability elements like:
  - geographical footprint (for an image or a tile)
  - date of creation,
  - date of acquisition (for an image or a tile)
  - Data Origin, describing the Data Provider

- Licensing conditions (Licensing name, URL of the IP owner, ...)
- Storage Location (URL, Coordinates or Country Name of the physical location of the hosting servers)
- Applicable Regulation (e.g Cloud Act, EU AI Act, ...)

User needs in Query Mode:

- In Query Mode, the Traceability shall be requestable thanks to either :
  - an AI model Id
  - a single image Id, optionally completed by list of tile bounding box within that image , and expressed in geographical coordinates
  - a list of Image Id, with their own optional list of tile bounding boxes
- in Query mode, the Traceability Service shall return, for every contributing input Data from their previous step of traceability :
  - the ID of data
  - Details about the owner and authors if applicable
  - Date of the data generation
  - Storage origin and current location.
  - If it is an image :
    - Acquisition date, geographical location, sensor, provider
    - Processing levels and transformations applied
  - If it is a tile from an image : details about its encompassing image and frame, its processing steps
  - If the Data is a model : its own traceability record

Filtering the results of the Query Mode:

When performing a Query to the Traceability Service thanks to the Query Mode, it shall be possible to reduce the outputs by filtering theme thanks to :

- a Time Range.
- a Geographical Area of Interest
  in order to scope the spatial coverage of the data used (e.g., bounding box, region code).
- a Data Origin, by specifying for instance a Data Provider
- a Licensing conditions (Licensing name, URL of the IP owner, ...)

- a Storage Location (URL, Coordinates or Country Name of the physical location of the hosting servers)
- an Applicable Regulation (e.g Cloud Act, EU AI Act, ...)

Output Formatting

The outputs of a request to the Query Mode shall be made available in standardized and downloadable formats, including :

- JSON → For integration in automated systems
- CSV → For analysis in spreadsheets or databases

Each output should be timestamped and carry a unique query ID for reproducibility and audit trail.

Missing or Partial Records Handling

- If no traceability information exists for a given image or model ID, return a clear message, stating that no records were found.
- If only partial information exists (e.g., missing metadata), return a clear message, stating that only partial records were found, along with the available fields and indicate missing ones by tagging them as null or undefined.

Access, Interface & Performance

- Access to the Traceability Service shall be provided through:
  - A secure, authenticated REST API
  - A lightweight Graphical User Interface (GUI) for testing and demo
- The Traceability Service shall provide Export formats: JSON, CSV, Dump
- Both in Declaration and Query mode, the Traceability Service must respond to queries with the following latency guarantees:
  - instantly for queries involving 1 entries
  - < 5 seconds for typical queries involving ≤ 1,000 entries
  - ≤ 1 hour for heavy queries involving ≥ 1 million records, or requiring deep federation scans

- Service status
    - At any time, a service health-check endpoint should return live information on the service availability and uptime.

### 2.2.4 Typical scenario and expected outcomes

Note: The following scenarios are subject to further change, from additional requirements found necessary or any technical constraint raised by the Engineering team.

Data Declaration Scenario

Based on the usage of a Dataionics user, a 1st scenario would be about Querying a Trace Record, as follow :

1- A user authenticates on the Traceability platform (1st connection ; the user is invited to create an account)

2- A user selects the object whose Traceability Record should be enriched. This object can be an AI model, an Image, a subset of an image.

3- The user can select a list of data to be declared as contributing to the traceability, which are ingested by batch into the Traceability service in order to update the current Traceability Record:

- /* As the idea is not to upload the contributing data, we must address how the relevant attributes are extracted from the input and by who, in order to be safely sent to the traceability record. */

Querying a Trace Record Scenario

Based on the usage of a Dataionics user, a 1st scenario would be about Querying a Trace Record, as follows:

1- A user authenticates on the Traceability platform (1st connection; the user is invited to create an account)

2- A user selects the object to be verified thanks to its unique Id within the traceability environment, resulting in the loading of the Trace Record of that object. This object can be an AI model, an Image, a subset of an image.

An alternative way is that the Traceability service allows the upload of a Trace record previously exported.

In that case, the Trace Record is read and checked for genuineness and integrity and permissions.

3- Once the Trace Record is made available, the user is entitled to perform queries about the data used to train the model, such as:

- The number of data used
- The Spatial Coverage of the input data, to be displayed in the Traceability Tool GUI, and exportable in GeoJSON format

# 3. Conclusion

The document presents two complementary use cases of traceability in the context of Earth observation data.

The first use case focuses on the application of machine learning models (e.g., Random Forest) in agriculture. While rooted in a specific domain, it represents a typical ML workflow, covering data acquisition, processing, as well as model training and validation. The second use case is more general and refers to AI solutions at large, addressing both the models themselves and the processes of feeding them with data. Together, the two use cases illustrate the full spectrum of traceability needs in practice.

Both scenarios highlight common critical aspects: the necessity of tracking input data and their metadata, including processing levels, baselines, and transformations. In particular, it is essential to maintain traceability of key input attributes such as time range, geographical coverage, and data origin, as they directly impact the validity and reproducibility of results.

Moreover, both use cases emphasize two complementary layers of traceability: the data input layer (covering the characteristics and provenance of input data) and the model layer (covering design, training, and performance). Ensuring consistency and transparency across both layers is a prerequisite for trustworthy AI and ML applications.

The proposed solution provides traceability independent of specific ML and AI vendors. This makes it reliable, repeatable, and resilient to technological changes, thereby enhancing its practical value in the context of Earth observation data.