FAIRICUBE –
F.A.I.R. INFORMATION CUBES

Work Package 2: Use
Deliverable 5.3: Validation of ingestion

Deliverable Lead: 4sfera
Deliverable due date: 29/02/2024

Version: 2.0
2024-11-06

Document Control Page

Document Control Page	
Title	Validation of data ingestion
Creator	4sfera
Description	Deliverable D5.3 Validation of data ingestion
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	NIL, S4E
Date of delivery	30/06/2023y
Type	R — Document, report
Language	EN-GB
Rights	Copyright "FAIRICUBE – F.A.I.R. information cubes"
Audience	☒ Public ☐ Confidential ☐ Classified
Status	☐ In Progress ☐ For Review ☒ For Approval ☐ Approved

Revision History			
Version	Date	Modified by	Comments
0.1	01/06/2023	Jaume Targa and Cristina Carnerero	First draft
0.2	15/06/2023	Jaume Targa	Second draft
0.3	29/06/2023	Jaume Targa and Lorena Banyuls	Draft for review
1.0	30/06/2023	Jaume Targa	Incorporation of comments from reviewer and ready for submission
1.1	07/02/2024	Jaume Targa	Updating with work done M13 to M20
1.2	18/03/2024	Jaume Targa	Further updates on document based on bilateral meetings
1.2	05/04/2024	Jaume Targa and María Colina	Further improvements
1.3	18/04/2024	Jaume Targa and María Colina	Further updates as a formal protocol are written
1.4	19/04/2024	Mirko Gregor	Reviewed
1.5	06/05/2024	Jaume Targa and María Colina	Further improvements following internal review
1.6	12/06/2024	Jaume Targa	Further improvements, re-solve comments and final review
1.7	14/06/2024	Kathi Schleidt	Further review
2.0	18/06/2024	Jaume Targa and María Colina	Solve final comments to be ready for submission

Disclaimer

This document is issued within the frame and for the purpose of the FAIRICUBE project. This project has received funding from the European Union's Horizon research and innovation programme under grant agreement No. 101059238. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

This document and its content are the property of the FAIRICUBE Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the FAIRICUBE Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the FAIRICUBE Partners. Each FAIRICUBE Partner may use this document in conformity with the FAIRICUBE Consortium Grant Agreement provisions.

Table of Contents

List of Figures & Tables

1 Context

1.1 Overall objective of WP5

The overall objective of Work Package 5: Validation (WP5) is to assure that all required input data is available in an aligned gridded data format, and that non-gridded resources (point and vector data) have been transformed to gridded formats as required.

1.2 Description of WP5 work

WP5 focuses on making all required input data available in a data format understandable by the two datacube stacks used in FAIRiCUBE, EOX and rasdaman. The data will either be made available directly via the EarthServer datacube federation (rasdaman) and EuroDataCube/SentinelHub (EOX) or be provided by the use case owners (e.g. gridded products from Copernicus, Eurostat, as well as required ancillary vector and metadata). Where necessary, point and polygon data will be rasterized and transformed into gridded data.
This process is called ingestion.

1.3 Description of Task 5.5

Task 5.5 (Validation of ingestion) focuses on checking the quality of all data being ingested. To avoid bias, checks must be performed by partners different from the ones executing the ingestion tasks.

2 Ingestion

The process of ingesting data is required for the effective use of any data resource under FAIRiCUBE. This process varies depending on the nature of the data involved. FAIRiCUBE-specific data undergoes a careful ingestion process into either rasdaman data cubes or EOX systems, depending on where they are required. Specific attention is paid to ensuring the FAIR principles (Findable, Accessible, Interoperable, and Reusable) throughout.

Ingestion validation (including data pre-processing) is an initial part of the User Case implementation steps under FAIRiCUBE. **Figure 1** illustrates the four general steps of UC implementation validation.

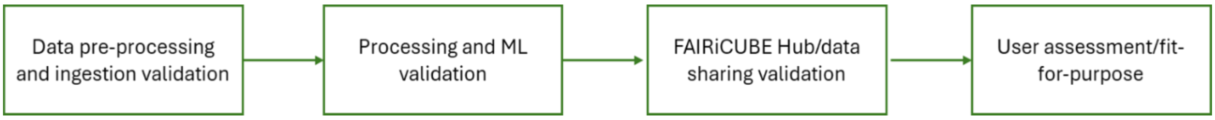
Data pre-processing and ingestion validation
Processing and ML validation
FAIRiCUBE Hub/data sharing validation
User assessment/fit-for-purpose

Figure 1: FAIRiCUBE User Case implementation flow

The data cube ingestion pipelines are comprehensively described within Delivery D5.2. These pipelines represent the careful attention and accuracy needed to handle data effectively, highlighting FAIRiCUBE's dedication to making geospatial information easily findable, accessible and usable for a variety of purposes and stakeholders. This section includes some information on the ingestion process by rasdaman and EOX which are important for the validation of this process.

In terms of validation of ingestion, it is key that validation covers all data preparation steps that may be required. Pre-processing may be an initial step before any ingestion is done, and it is often performed outside the platform, or on the other hand, pre-processing may happen within the ingestion phase. Regardless of where this step happens, validation is required.

2.1 Ingest process by rasdaman

rasdaman's wcst_import.sh utility simplifies the import of raster data in formats like TIFF, JPEG2000, netCDF, and GRIB. Internally it makes WCS-T requests to the rasdaman geo service to either *ingest* the data permanently into the rasdaman database or *register* the data files without modifying or copying them into rasdaman. The *wcst_import.sh* tool is based on two concepts:

- **Recipe** - A recipe defines how a set of data files can be combined into a well-defined coverage (e.g., a 2-D mosaic, regular or irregular 3-D timeseries, etc.);
- **Ingredients** - A JSON file that configures how the recipe should build the coverage (e.g., the server endpoint, the coverage name, which files to consider, whether to ingest or register the files in situ, etc.).

The current ingest process can be schematised as follows:

1. Check the inventory list to find the next dataset to ingest,
2. Download the dataset and extract it on the FAIRiCUBE VM,
3. Prepare an "ingredients file", a JSON file defining the parameters required to start the ingest process (see example below),
4. Start the ingest.

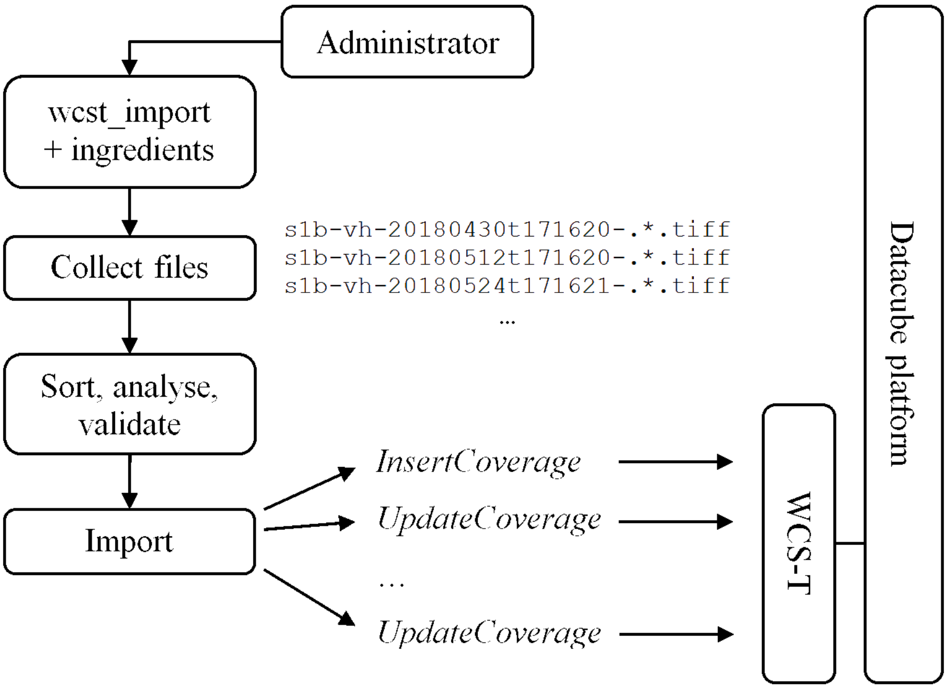

Figure 2 - Data ingestion pipeline workflow using rasdaman platform.

2.2 Ingest process by EOX

The pipeline in the EOX deployment focuses on data registration rather than ingestion. Each Use Case gets a dedicated bucket for storing datasets, with access credentials automatically provided in the user's EOxHub workspace and integrated services like Sentinel Hub. **Figure 3** illustrates the workflow for "ingesting" data to the FAIRiCUBE Hub using the EOX platform.

The data uploaded to the object storage only needs to be registered in services as required and not ingested. EOX deployment is provided with a dedicated bucket on the object storage to store their relevant datasets. Detailed information is available in Deliverable "*D5.2 The description of the data cube ingestion pipelines*".

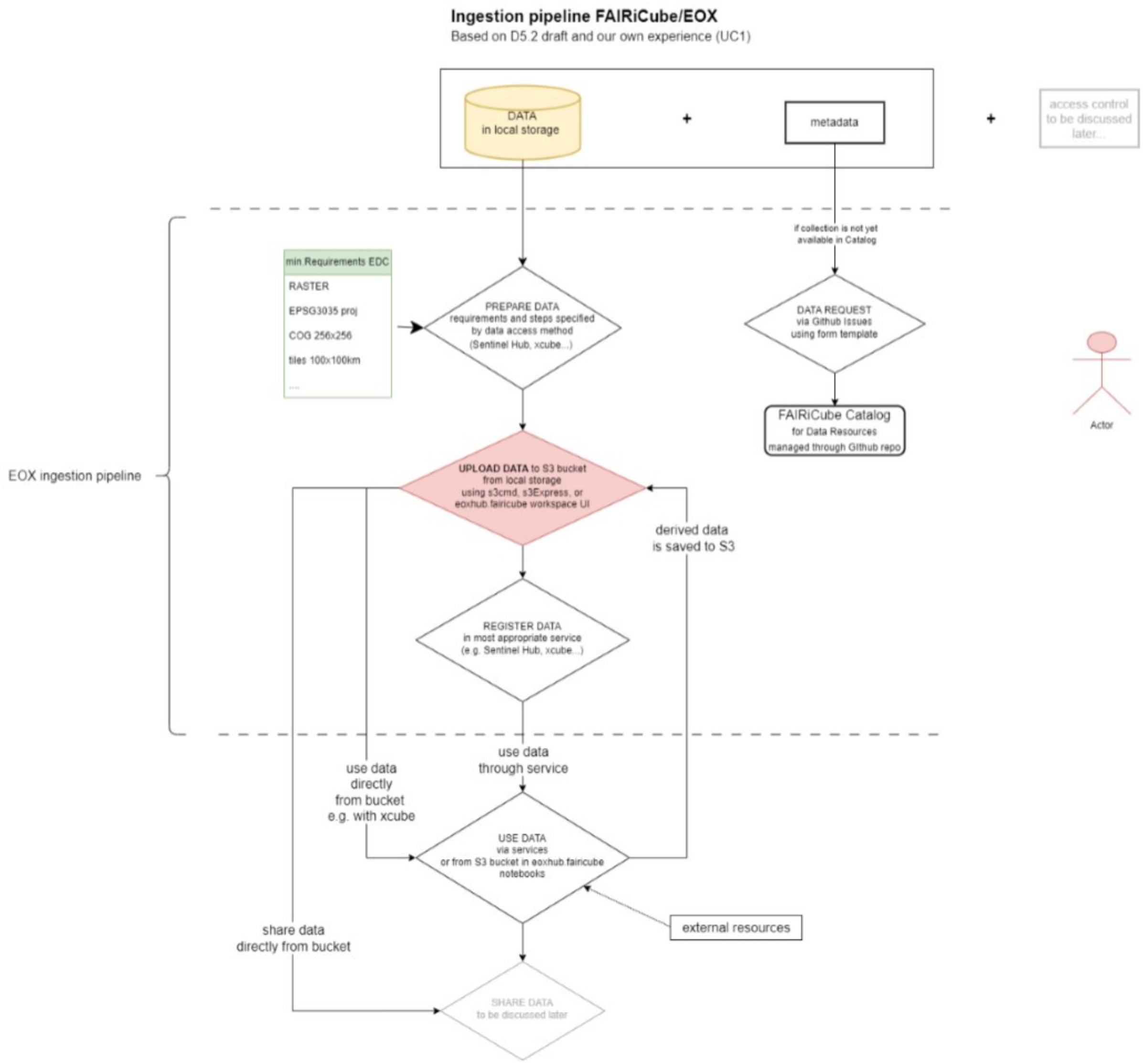

Figure 3 : Data ingestion pipeline workflow using EOX-platform.

3 Data Validation for Ingestion

Errors can occur at every step of a process, which also applies to the ingestion process. Therefore, it is extremely important to check after each ingestion whether it was performed correctly. There are several ways to check a process, which are presented in more detail in the following chapter. The methodology to validate ingestion is under development. The methodology as described in this document is currently being tested on the large set of data that has been ingested into the FAIRiCUBE system. Hence, there could be further evolutions or adaptations to the validation scheme after the tests have been finalised and evaluated.

When ingesting data, errors can occur that will affect the result. Data validation is crucial to prevent this. Some examples on issues that can go wrong are:

- Mixing of CRS, geodetic values despite projection
- Invalid CRS transformation without documentation
- Truncation of values, e.g. one- or two-digit CLC codes
- ...

3.1 Key aspects

The implementation of data validation for ingestion involves a series of checks to ensure the quality and integrity of the ingested datasets. Here is an overview of the key aspects of the implementation:

- **List of Characteristics:** This check is to define a comprehensive list of characteristics that need to be validated for each ingested dataset. This list encompasses various aspects such as data completeness, correctness, consistency, and conformity to predefined standards. It serves as a reference for the subsequent validation process.

- **Spatial Validation:** In addition to the characteristics-based validation, spatial validation is performed to ensure the spatial integrity of the ingested data. This involves verifying that the spatial attributes have been preserved accurately during the ingestion process. It also includes checks for proper re-projection, if applicable, to ensure the spatial relationships and alignments are maintained correctly.

- **Descriptive Statistics Calculation**: Once the dataset is ingested, descriptive statistics are computed for each attribute. These statistics capture important characteristics of the data, including measures such as mean, standard deviation, minimum, maximum, and other relevant statistical indicators. These statistics provide valuable insights into the distribution and properties of the data. The computed descriptive statistics need to be compared to the statistics derived from original datasets prior to ingestion.

- **Anomaly detection of existing datasets:** A detection process should be applied to identify any significant deviations or anomalies in the newly ingested data, compared to existing datasets. This process uses the original statistics of acceptable data as a baseline for detecting inconsistencies or unexpected patterns. This is important when digesting updates and new versions of similar datasets.

- **Source Data/Metadata Comparison**: It is crucial to compare the ingested data with the original source data and metadata. Any transformations or modifications made during the ingestion process must be clearly documented and justified. This check ensures that the data maintains its integrity and reliability, and that users have a transparent understanding of any changes from the original source.

- **Error Labelling and Data Incorporation**: Based on the results of above checks, the ingested data is labelled either as acceptable or erroneous. If significant discrepancies are detected, the data is flagged as erroneous, indicating the presence of potential data quality issues. However, if no discrepancies are found, the data is labelled as acceptable, and its descriptive statistics are incorporated into the ensemble of reference statistics for future comparisons.

- **Reporting and Logging:** Throughout the validation process, comprehensive reports and logs are generated to document the validation results. These reports provide detailed information about the validation outcomes, including any detected errors or anomalies. This documentation facilitates further analysis, investigation, and troubleshooting of data quality issues. This information should be part of the meta-data catalogue of each dataset.

3.1.1 General Data quality control workflow

Considering the key aspects necessary to perform a correct data validation for ingestion, the proposed method has been developed and it is based on:

- first, a list of characteristics to be checked after ingesting the dataset,
- secondly, validation checks to ensure reliable data, by means of data validation, attribute validation and spatial variation check, and
- thirdly a re-processing of data (after rectification of the errors).

The aim is to implement this method in an automated manner as far as possible to turn the data quality assessment more efficient and less reliant on manual intervention. However, this is not possible in all the steps/platforms, where some manual procedures are still required. The automated nature of the approach (at least in some parts, currently only available on EOX) allows for consistent and reliable anomaly detection, even without extensive domain expertise or predefined quality constraints. However, experience from FAIRiCUBE UC reflect the importance of both manual/visual inspection, at some degree, of any ingested data.

Figure 4 shows a general data quality control workflow for data validation. This has been produced based on UC experience during initial phases of data ingestion. For each data ingestion process, a quality control procedure is performed. The following describes each step in the data quality control workflow:

- **Requirements**: Before initiating the data quality control process, it is essential to gather requirements from various sources. These requirements include both the technical specifications of the data source and the specific needs of the end-users who will be using the data. This information is available in the ingestion metadata request. (https://github.com/FAIRiCUBE/data-requests)

- **Definition of Acceptance Criteria**: Based on the gathered requirements, a set of acceptance criteria is defined. These criteria outline the standards that the data must meet to be considered acceptable. The acceptance criteria may vary depending on factors such as the nature of the data source and the specific objectives of the data analysis.

- **Extraction of Data Attributes**: Once the acceptance criteria are established, the next step is to extract relevant attributes from the incoming data. These attributes encompass various aspects of the data, such as numerical values, categorical variables, timestamps, and spatial coordinates.

- **Comparison with Acceptance Criteria**: The extracted data attributes are then compared against the defined acceptance criteria. This comparison evaluates whether the data meets the specified standards and aligns with the expectations outlined in the acceptance criteria.

- **Quality Control Evaluation**: If the comparison indicates that the data attributes match the expected attributes as per the acceptance criteria, the quality control process is deemed successful, and the data passes the quality control check.

- **Generation of Non-Matching Attributes List**: In cases where discrepancies are identified between the extracted data attributes and the expected attributes defined in the acceptance criteria, a list of non-matching attributes is generated. This list provides insights into areas where the data does not meet the desired standards.

- **Error Rectification**: To address the discrepancies and ensure data quality, errors identified in the non-matching attributes are rectified. This may involve cleaning, transforming, or correcting the data to bring it into alignment with the acceptance criteria.

- **Re-processing of Data**: After the necessary corrections have been made, the data undergoes a re-processing step to apply the quality control procedures once again. This iterative process continues until the data meets the specified standards and fulfils the acceptance criteria.

- **Visual Inspection Check**: In addition to the automated comparison of data attributes, a visual inspection check is incorporated into the workflow. This manual review allows for the identification of any anomalies or patterns that may not be captured through automated processes alone.

By following this comprehensive data quality control workflow, UCs can ensure that the data ingested into their systems meets the required standards of accuracy, completeness, and consistency.

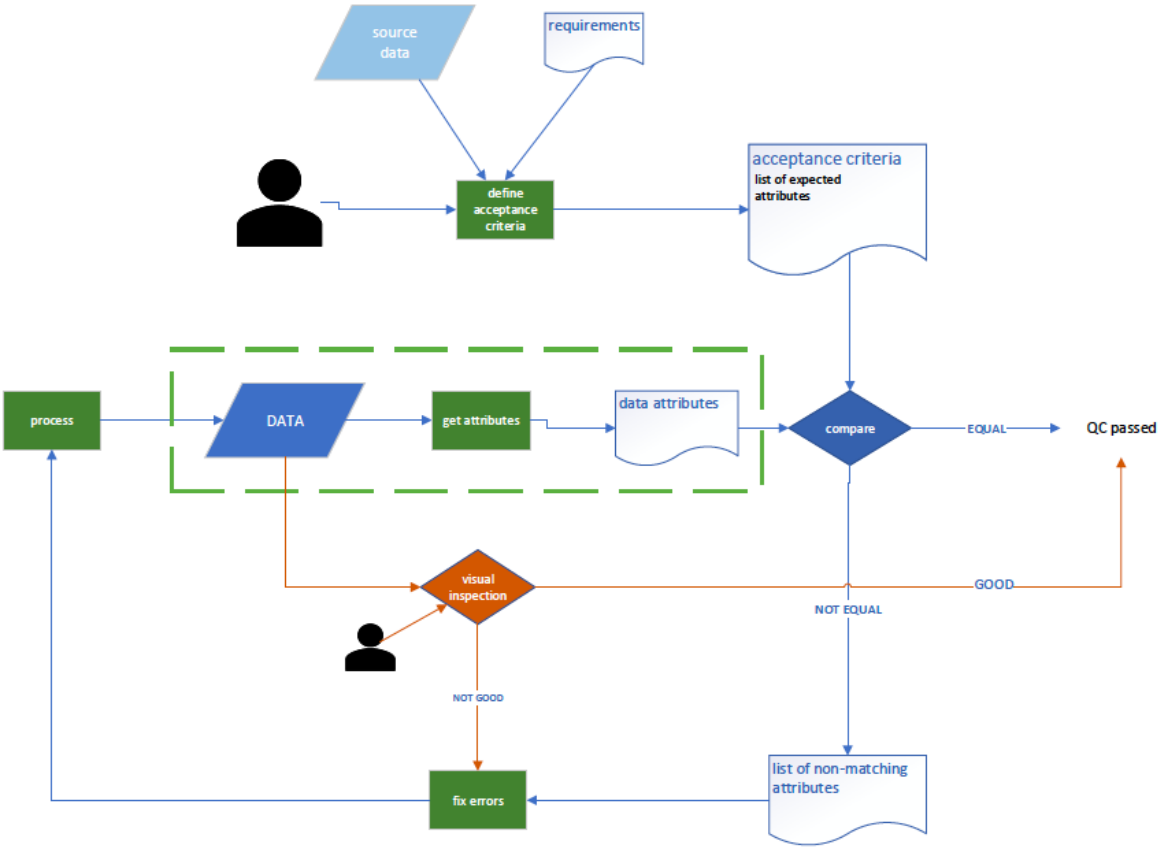

Figure 4 - Data quality control workflow

3.1.2 Extraction of Data Attributes

The attributes from the incoming data encompass various aspects such as numerical values, categorical variables, time stamps and spatial coordinates. The extraction of these attributes is an essential step to perform a correct data validation of ingestion as this information will be used to compare it with the Acceptance Criteria (using for example the statistics obtained from the extracted data attributes) and to perform the validation checks.

Figure 5 illustrates the specific workflow for the extraction of data attributes. The key steps for the extraction of Data Attributes are the following:

1. **Sampling** involves selecting a subset of data from the entire dataset for analysis. This subset should be representative of the larger dataset. Sampling helps in efficiently processing large volumes of data by working with manageable portions, reducing computational overhead, and facilitating quicker analysis.

2. **Computing Statistics** involves calculating various descriptive measures and metrics from the sampled data. These statistics provide insights into the characteristics, distribution, and trends present within the dataset. Commonly computed statistics include mean, median, maximum, minimum, standard deviation...

3. **Get Spatial Attributes**: Spatial attributes refer to data elements that are associated with geographical or spatial information. These attributes may include coordinates, shapes, boundaries, or other spatial identifiers. Getting spatial attributes may involve extracting and analysing spatial data to understand spatial relationships, patterns, and distributions within the dataset. This step is crucial for geospatial analysis, mapping, and visualisation.

4. **Get Temporal Attributes**: Temporal attributes pertain to time-related data elements such as timestamps, dates, durations, or intervals. These attributes capture the temporal aspects of events, processes, or phenomena. Getting temporal attributes involves parsing and interpreting temporal data to uncover temporal patterns, trends, and dependencies. This step enables temporal analysis, forecasting, and trend identification over time.

5. **Metadata** refers to descriptive information about the dataset, including its structure, format, content, and context. Metadata provides valuable insights into the data's origin, quality, usage, and relevance. Getting metadata involves retrieving and documenting metadata attributes such as
   - data source,
   - data schema,
   - data lineage,
   - data quality indicators, and
   - any relevant annotations or annotations.

Metadata plays a crucial role in data governance, data management, and data integration by facilitating data discovery, understanding, and interpretation.

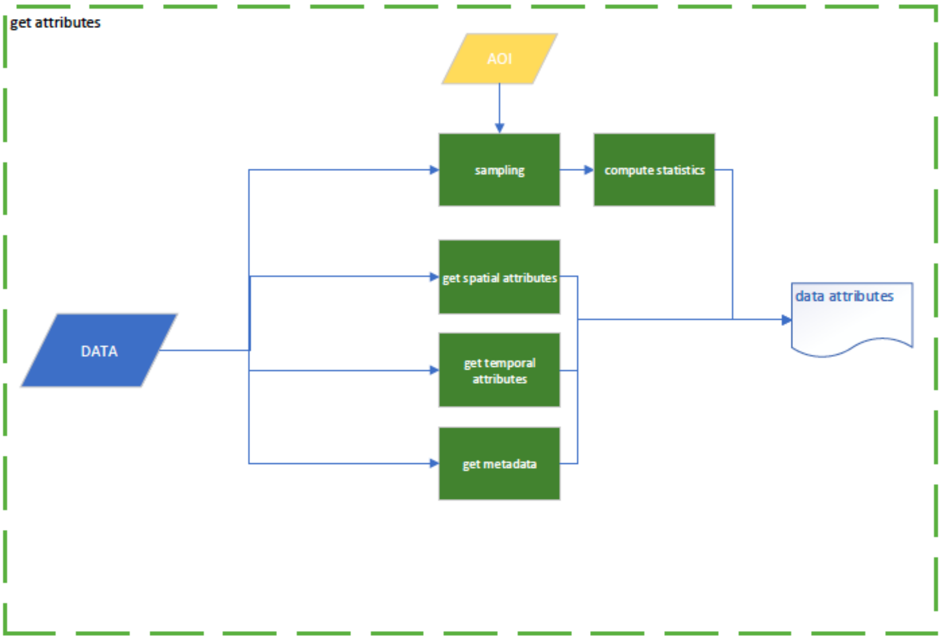

Figure 5 - Extraction of Data Attributes

These steps collectively contribute to the extraction and characterisation of data attributes during the data ingestion validation process. By systematically sampling, computing statistics, capturing spatial and temporal information, and documenting metadata, UCs can gain a comprehensive understanding of their data assets, enabling informed decision-making, analysis, and interpretation.

3.2 Validation checks

In the data ingestion process, specific validation checks are vital to uphold data integrity and quality. Several validation processes are required to ensure that the ingested data meets all established Acceptance Criteria and is fully valid for its intended use.

3.2.1 Data validation checks

The characteristics outlined in Table 1 provide a clear description of the items that require validation. It becomes possible to systematically review and verify the quality of the ingested datasets, ensuring their accuracy, completeness, consistency, and adherence to the predefined standards and requirements. This characteristics-based analysis will ensure an effective data quality validation process.

In the data ingestion process, specific validation checks are vital to uphold data integrity and quality. This section outlines crucial validation checks essential for ensuring reliable data:

1. **Duplicates Detection**: Detecting and eliminating duplicate records safeguards data cleanliness and prevents redundancy in analysis. By identifying duplicate entries, organizations can maintain the accuracy and integrity of their datasets.
    - **Objective**: The objective is to identify and eliminate duplicate records to maintain data cleanliness and prevent redundancy in analysis
    - **Approach:** Utilise algorithms or methods to detect duplicate entries based on key attributes or fields. Implement procedures to resolve duplicate records, such as merging or removing duplicates while preserving data integrity.

2. **Temporal Overlap and Gaps**: Checking for date overlaps and gaps ensures temporal consistency within the dataset. Addressing overlaps and filling gaps in temporal data enhances the reliability of temporal analyses and facilitates accurate trend identification.
    - **Objective:** To ensure temporal consistency within the dataset by detecting and addressing date overlaps and gaps
    - **Approach:** Analyse date ranges and timestamps to identify overlaps and gaps. Implement strategies to resolve inconsistencies, such as adjusting date ranges or filling gaps with interpolated values.

3. **No Data Values**: Verifying the absence of no data values is crucial for identifying missing or null entries. Handling these values appropriately prevents errors in analysis and ensures the completeness and accuracy of the dataset. As no data values may exist in the original dataset, it is crucial to check and identify these.
    - **Objective**: The objective is to verify the absence of no data values, ensuring the completeness and accuracy of the dataset

- **Approach**: Develop validation rules or scripts to check for missing or null entries in the dataset. Implement procedures to handle and rectify missing values, such as imputation or flagging for further investigation. Moreover, null value must be checked by to original dataset as these could be correct.

4. **Value Types and Encoding**: Validating value types and encoding ensures that data values are accurately represented and encoded. This check prevents misinterpretation or corruption of data during processing, contributing to the reliability of analysis results.
   - **Objective**: To validate value types and encoding, ensuring accurate representation and interpretation of data values
   - **Approach**: Verify the data type and encoding of each value against predefined standards or specifications. Implement encoding conversion or validation routines to ensure consistency and compatibility across systems.

Characteristics	Description
Duplicates	Check for duplicate entries
Temporal overlap	Check for overlaps in the date column, i.e., repeated dates with different values (only for timeseries)
Temporal gaps	Check for gaps between start and end date (only for timeseries)
No data values	Verification of the correct use of no data
Value types	Check if data types are correct (string, integer, float, datetime format)
Value encoding	Check if the encoding of the data is correct (e.g., character encoding is utf-8; point (.) is used as decimal separator)

Table 1: Proposed characteristics to be checked after the ingestion of a dataset.

3.2.2 Characteristics validation

The proposed approach for data quality assessment and anomaly detection involves a systematic way to identify discrepancies in the ingested data. A set of descriptive statistics, outlined in Table 2, is computed for each observed partition within the dataset. These statistics capture various characteristics and properties of the data, providing valuable insights into its distribution and behaviour. The computed statistics are then combined into a vector that represents the data. To validate the correctness of the ingestion, these attributes must be comparable to the original dataset.

Characteristics	Description
Completeness	The ratio of not-NULL values
Count of distinct values	Number of distinct values in the dataset
Ratio of the most frequent value	Number of occurrences for the most frequently repeated value, normalised by the batch size
Maximum	Maximum value of the dataset
Mean	Mean value of the dataset
Minimum	Minimum value of the dataset
Standard deviation	Standard deviation of the dataset
Date range	Start and end date (only for timeseries)

Table 2: Proposed characteristics to be calculated and compared after the ingestion of a dataset.

Characteristics validation aims to ensure the integrity and reliability of data. This section outlines the key elements of the dataset attribute validation from a data ingestion standpoint:

1. **Completeness**: The completeness of the dataset attributes is assessed to verify the presence of data values within each attribute. By meticulously checking for missing or null values, data completeness is ensured, laying a strong foundation for downstream analyses.
   - **Objective**: Evaluate the presence of data values in each dataset attribute.
   - **Approach**: Check for missing or null values within the dataset attributes to guarantee data completeness upon ingestion.

2. **Count of Distinct Values**: Assessing the uniqueness of values within the dataset attributes is essential to identify potential duplicates or irregularities in the ingested data. Calculating the count of unique values for each attribute provides insights into the diversity and richness of the dataset.
   - **Objective**: Assess the uniqueness of values within each dataset attribute.
   - **Approach**: Calculate the count of unique values for each dataset attribute to identify potential duplicates or irregularities in the ingested data.

3. **Ratio of the Most Frequent Value**: Understanding the prevalence of the most frequently occurring value within the dataset attributes is critical for comprehending data distribution patterns. Computing the ratio of records containing the most frequent value offers valuable insights into the dominant trends present in the ingested data.
   - **Objective**: Determine the prevalence of the most frequently occurring value within the dataset attributes.
   - **Approach**: Compute the ratio of records containing the most frequent value in each of the dataset attribute to understand data distribution patterns during ingestion.

4. **Maximum, Mean, and Minimum Values**: Establishing the range and central tendencies of values within the dataset attributes provides a comprehensive understanding of the data landscape. By calculating metrics such as maximum, mean, and minimum values, organizations gain insights into the breadth and typicality of data values ingested.
   - **Objective**: Establish the range and central tendencies of values within the dataset attributes.
   - **Approach**: Calculate the maximum, mean, and minimum values for each dataset attribute to capture the breadth and typicality of data values ingested.

5. **Standard Deviation**: Measuring the variability of values within the dataset attributes is essential for assessing data consistency and variability. Computing the standard deviation for each attribute helps identify outliers and anomalies, ensuring data quality and reliability.
   - **Objective**: Measure the variability of values within the dataset attributes.
   - **Approach**: Compute the standard deviation for each dataset attribute to assess data consistency and variability during ingestion.

6. **Number of Records**: Validating the total count of records ingested is crucial for ensuring data completeness and consistency. By confirming the number of records within the dataset, organizations can verify data volumes align with expectations and requirements.
   - **Objective:** Confirm the total count of records ingested.
   - **Approach**: Validate the number of records within the dataset to ensure consistency with expected data volumes during the ingestion process.

7. **Date Range**: Analysing the temporal span covered by date-related dataset attributes provides insights into the completeness and validity of temporal data. Assessing the earliest and latest dates present in date-related attributes ensures the integrity of time-related information ingested.
   - **Objective**: Determine the temporal span covered by date-related dataset attributes
   - **Approach**: Analyse the earliest and latest dates present in date-related dataset attributes to verify the completeness and validity of temporal data during ingestion

3.2.3 Spatial validation

Validating the spatial integrity of the ingested data involves checking whether the spatial information associated with the datasets has been preserved accurately during the ingestion process. Ensuring the spatial integrity of ingested data involves a comprehensive validation process to guarantee accuracy and reliability in spatial information preservation. This section outlines key aspects of spatial validation, including considerations for grid boundaries, data completeness, projection/coordinate reference system (CRS), cell size, number of bands, data type, data format, centre coordinates, and total area, which are checked against the original source dataset.

In some cases, data ingestion may involve re-projection, where the data is transformed from one coordinate system to another; if this is even possible is dependent on the nature of the data being provided, e.g., continuous, discrete, ordinal, nominal, as well as what the values represent. A dataset where the numerical values contained represent elevation can be resampled to a different CRS with great accuracy whereas correctly resampling the counts provided by a population grid is not numerically feasible. It is crucial to ensure that this re-projection process is executed correctly and does not introduce any spatial distortions or errors. The approach includes verifying that the re-projection, if required and applicable, has been accurately performed and that the spatial relationships between the dataset are maintained as expected. Moreover, it is checked that the integrity of the contained data is retained after re-projection.

By conducting these spatial validations, the approach helps ensure the integrity and accuracy of the spatial information in the ingested data. It contributes to maintaining the reliability and consistency of the dataset, enabling subsequent analyses and applications to rely on accurate spatial representations.

Characteristic	Description
Grid boundaries	Top-left and bottom-right coordinates (only for gridded datasets)
Data completeness	Check to verify that the data set is complete (total area, total number of features or cell)
Projection/CRS	Verification of the correct use of the projection/CRS.
Cell size	Check to verify that the cell size is correct
Number of bands	Verification that all channels have been transmitted correctly.
Datatype	Check if the data type is OK.
Data format	Check if the ingested data follows the desired standard format (e.g., for raster, cloud optimized Geo Tiff)

Table 3: Proposed characteristics to validated.

The characteristics to be validated against the original source data as listed in table 3 include:

1. **Grid Boundaries** must align with expected boundaries and cover the entire study area without gaps. Consistency in grid resolution and alignment with spatial reference frameworks is essential to ensure accurate spatial representation.
   - **Objective**: Validate the boundaries and extent of the spatial grid
   - **Approach**: Verify that grid boundaries align with expected boundaries and cover the entire study area. Ensure consistency in grid resolution and alignment with spatial reference frameworks.

2. **Data Completeness:** Spatial data should cover the entire extent of the study area comprehensively, without gaps or missing areas. Ensuring completeness in spatial attribute information and grid coverage is vital for comprehensive data representation.
   - **Objective**: Ensure the completeness of spatial data coverage
   - **Approach**: Validate that spatial data covers the entire extent of the study area without gaps or missing characteristics. Check for completeness in spatial attribute information and grid coverage to ensure comprehensive data representation.

3. **Projection/CRS:** Validating the applicability, accuracy and consistency of spatial data projections ensures that spatial data is correctly referenced to the specified coordinate reference system (CRS). This involves verifying CRS metadata and transformation parameters to ensure accurate spatial referencing.
   - **Objective**: Confirm the applicability, accuracy and consistency of spatial data projections
   - **Approach**: Validate that spatial data is projected correctly, with a resampling technique suited to the contained data, onto the specified CRS. Verify CRS metadata and transformation parameters to ensure accurate spatial referencing and alignment with intended coordinate systems.

4. **Cell Size:** The spatial resolution and cell size of raster data must be validated to ensure uniformity and accuracy in spatial data representation. Calculating cell dimensions and resolution helps verify consistency and appropriateness for the intended application.
   - **Objective**: Validate the spatial resolution and cell size of raster data
   - **Approach**: Verify that cell size is consistent and appropriate for the intended application. Calculate cell dimensions and resolution to ensure uniformity and accuracy in spatial data representation.

5. **Number of Bands:** Confirming the number of bands as well as the correct description of the individual bands within raster datasets is essential for data analysis and processing. Validating the presence and count of spectral bands ensures compatibility with data processing requirements.
   - **Objective**: Confirm the number of spectral bands in raster data. Confirm the correct description of the individual bands including the definition of the property that is represented by the band
   - **Approach**: Validate the presence and count of spectral bands within raster datasets. Check band metadata and specifications to ensure compatibility with data analysis and processing requirements.

6. **Data Type and Format:** Spatial data should be stored in appropriate data types and formats for efficient processing and analysis. Checking file formats, compression methods, and encoding schemes ensures compatibility with data processing workflows.
    - **Objective**: Validate the data type and format of spatial datasets
    - **Approach**: Verify that spatial data is stored in appropriate data types and formats for efficient processing and analysis. Check file formats, compression methods, and encoding schemes to ensure compatibility with data processing workflows.

By rigorously validating these, during the data ingestion process, organizations can ensure the accuracy, completeness, and reliability of spatial data for various applications, including geographic information systems (GIS), remote sensing, environmental modelling, and spatial analysis.

4 Data Ingestion Validation Protocol (DIVP)

The implementation of data validation for ingestion involves a series of steps to ensure the quality and integrity of the ingested datasets. Depending on the data type ingested, some key characteristics described above are important.

To facilitate the Data Validation for Ingestion and to provide a guide to enable it to be carried out in a unified and clear manner under the Data Ingestion Validation Protocol (DIVP). The DIVP is currently under development at https://github.com/FAIRiCUBE

The FAIRiCUBE DIVP aims to be a hands-on protocol for implementing data validation for ingestion. The repository aims to be a living document as close to the data ingestion process as possible. The protocol has been developed together with UC and platform experts. The DIVP includes the following sections:

- Introduction
- Data Ingestion Process Overview
- Data pre-processing steps
- Validation checks
- Error handling and reporting
- Documentation and traceability
- Validation case studies
- Conclusion

The content of each section is summarised below:

4.1 DIVP - Introduction

In the introduction section, a brief overview of the purpose of the protocol is provided. The importance of data ingestion validation is highlighted to ensure the quality and integrity of data used in various applications. The potential consequences of using erroneous or incomplete data, such as inaccurate analyses, faulty decision-making, or compromised research outcomes.

The objectives of the protocol are outlined including standardising validation procedures, minimising errors, and enhancing data reliability. It clearly defines the scope of the protocol, specifying the types of data and validation checks it covers, as well as any limitations or constraints.

4.2 DIVP - Data Ingestion Process Overview

The overview describes the end-to-end data ingestion process, from data acquisition to storage or analysis. It aims to identify the different sources of data, linking to data request forms and data catalogue. It will summarise the different formats, including raster formats (e.g., GeoTIFF, JPEG), vector formats (e.g., Shapefile, GeoJSON), and point data formats (e.g., CSV, Excel).

4.3 DIVP - Data Pre-processing Steps

Data pre-processing is key for many thematic data sources within FAIRiCUBE. This section details the pre-processing steps that may be necessary before conducting any ingestion. These steps may include data cleaning to remove errors or inconsistencies, normalization to standardise data formats or units, and transformation to reproject or rescale spatial data.

It should provide clear instructions and examples for each pre-processing step, along with relevant Python and R code snippets demonstrating how to implement them programmatically.

4.4 DIVP – Validation checks

The validation checks section in the DIVP is a crucial part of the document. This section delineates practical validation checks spanning data format validation, metadata validation, data quality validation, geometric validation, and attribute validation.

Through the application of automated validation scripts, we uncover potential discrepancies and anomalies embedded within our datasets, fostering informed decision-making and analytical rigor.

Within this section, key subsections will cover:

- Data Format Validation
- Metadata Validation
- Data Quality Validation
- Geometric Validation (for vector and point data)
- Attribute Validation

4.5 DIVP – Error Handling and Reporting

The error handling and reporting section will outline procedures for handling validation errors encountered during the data ingestion process. Some errors will require re-ingestion of data. However, some errors will require reporting and logging in the corresponding meta-data. This section will provide how to handle errors, including corrective actions or mitigation strategies.

4.6 DIVP – Documentation and Traceability

The final section of the FAIRiCUBE DIVP will highlight the importance of keeping comprehensive documentation for recording validation results, procedures, and outcomes. It will provide guidance on how maintain to link validation results back to source data, validation criteria, and validation checks performed. It will give examples of documentation formats for recording validation findings, observations, and resolutions.

5 Examples of validation processes

In this section, examples of data ingestion validation (including pre-processing) are provided. These are real examples from FAIRiCUBE use cases.

5.1 Pre-processing validation raster datasets ingested into EOX/SentinelHub

Use Case 1 has developed a python script[1] to validate the ingestion of raster files into EOX platform The script defines a QualityChecker class to perform quality control (QC) checks on raster datasets ingested. The QC process is performed as follows:

5.1.1 Overview of the QualityChecker Class

5.1.1.1 Initialisation

Various attributes are required for the quality control process:

- configuration settings,
- paths for original and transformed raster datasets, and
- placeholders for collection metadata and QC reports.

5.1.1.2 Connection to SentinelHub

A method sets up a connection to SentinelHub using credentials stored in environment variables.

5.1.1.3 Reading Metadata from Original Raster

Original raster metadata is read from a local raster file, such as the number of bands, bounds, coordinate reference system (CRS), data type, no-data value, and cell size. This information is printed and returned as a dictionary.

5.1.1.4 Setting DataCollection for SentinelHub

The SentinelHub data collection using a collection ID is defined. This data collection represents the ingested raster dataset.

5.1.1.5 Retrieving Metadata from Sentinel

The metadata for the ingested collection and its tiles from SentinelHub, respectively, is fetched.

5.1.1.6 Quality Control Report Initialization and Management:

The QC report, which is stored as a DataFrame, is created the report records various QC checks, appending new rows for each check performed and saving the report to a CSV file.

[1] https://github.com/FAIRiCUBE/uc1-urban-climate/blob/master/pre-processing/quality_check.py

5.1.2 Quality Control Checks

5.1.2.1 Metadata Checks

The check_metadata method performs several checks to compare the metadata of the original raster with the metadata of the ingested raster in SentinelHub:

1. **CRS Check**: Compares the CRS of the original raster with the ingested raster.
   Cell Size Check: Ensures the cell size matches between the original and ingested raster.
2. **Bounds Check**: Verifies that the spatial bounds of the original raster match those of the ingested raster.
3. **Number of Bands Check**: Confirms that the number of bands in the original raster matches the ingested raster.
4. **Data Type Check**: Compares the data type (e.g., uint8, uint16) of the original and ingested raster.
   No-Data Value Check: Ensures the no-data value is consistent between the original and ingested raster.
5. **Timestamp Check**: Verifies that the sensing times match, although in this example, it defaults to a fixed timestamp.

Each check appends a result to the QC report, indicating whether the check passed ("OK") or failed ("error") in a format which can be evaluated and summarized by a script.

5.1.2.2 Statistics Checks

The statistics between the original raster and the ingested raster over a specified area of interest (AOI) are compared:

- **Reading Statistics from Original Raster**: Extracts statistics (minimum, maximum, mean, standard deviation, and count) from a specified AOI in the original raster using rasterio.
- **Retrieving Statistics from SentinelHub**: Uses a SentinelHubRequest to retrieve similar statistics from the ingested raster for the same AOI.

1. **Comparison**: Compares the statistics of the original and ingested raster, appending the results to the QC report.

5.1.3 Running the QC

In summary the QC checker follows the following steps:

- Initialises the QualityChecker instance and connects to SentinelHub.
- Retrieves metadata from both the original raster and the ingested collection in SentinelHub.

1. Performs metadata checks to compare original and ingested metadata.
2. Performs statistics checks for a defined AOI, comparing statistics between the original and ingested raster.
3. Saves the QC report to a CSV file.

This method, in this example, ensures that the data ingested into SentinelHub matches the original dataset in terms of metadata and statistics, providing a robust method for verifying data integrity.