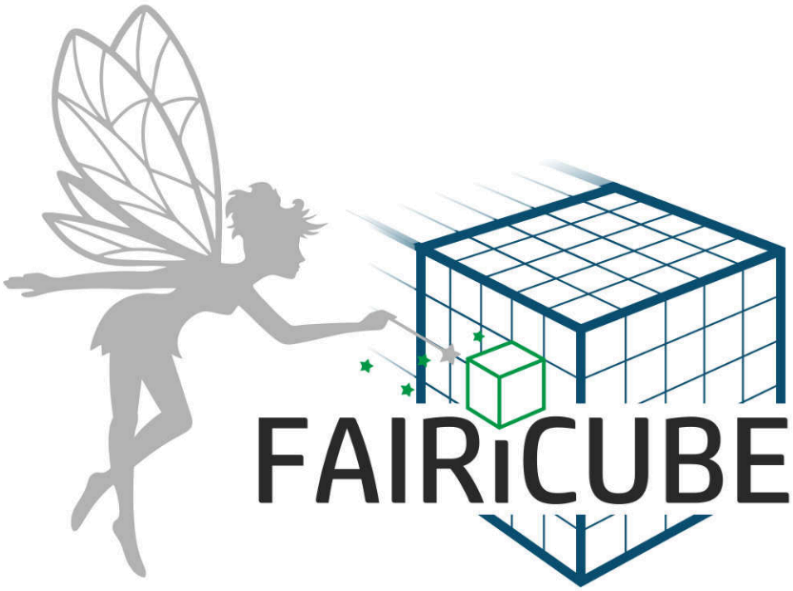

FAIRICUBE – F.A.I.R. INFORMATION CUBES

Work Package 2: Use
Deliverable 2.5: Use Case Validation

Deliverable Lead: space4environment
Deliverable due date: 30/06/2024

Version: 1.3
2024-06-26

Document Control Page

Document Control Page	
Title	Deliverable D2.5 Use Case Validation
Creator	space4environment
Description	The deliverable describes the different validation steps that UC must undertake to get reliable information on the output quality.
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	S4E, 4SF, NIL, NHM, WER
Date of delivery	30/06/2024
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	22/03/2024	Mirko Gregor	Creation of template and structure
0.2	21/05/2024	Mirko Gregor	First draft
0.3	07/06/2024	Maria Ricci	Updated 3.2; added 3.2.1
1.0	21/06/2024	Jaume Targa	Internal review and format checking
1.1	24/06/2024	Maria Ricci	Added chapter 3.2.6
1.2	25/06/2024	Jaume Targa	Further review and format check
1.3	26/06/2024	Maria Ricci	Final version

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 oft Contents

List of Figures & Tables

1 Context

1.1 Overall objective of WP2

The overall objective of Work Package 2: Use (WP2) is to ensure efficient execution of the Use Cases (UCs), assuring those potential synergies pertaining to both data and processing are identified and leveraged.

1.2 Description of WP2 work

WP2 focuses on the execution and cross-coordination of the different UCs on the FAIRiCUBE. Acting with an "outsider" role, the supervision crosscuts through all Use Case (UC) activities ensuring harmonisation with both upstream (data sources, ingestion, and processes) and downstream (results, promotion, and distribution of outputs) activities.

1.3 Description of Task 2.6

Controlling the quality of each UC is essential to ensure the success of the WP. This task will assure that all UCs address quantifiable goals, and that the interpretation of the results are in line with scientific methodology. This task will ensure that the outcomes of each UC are validated along the different steps from potential data pre-processing and ingestion over data processing to the final results and their representation. The task will validate that the UC objectives are met, the approach was followed, and the researched questions have been attempted to be answered.

This deliverable will summarise the various single validation steps that are planned in the context of the FAIRiCUBE UCs and provide a general protocol for their overall quality control.

2 Introduction

2.1 FAIRiCUBE and its Use Cases

Implementing FAIRiCUBE leans on two main pillars. The first pillar is the development and provisioning of the FAIRiCUBE Hub which is a crosscutting platform and framework for data ingestion, provision, analysis, processing, and dissemination. The second pillar consists of the five Use Cases (UCs) that have been designed to illustrate how data driven projects can benefit from cube formats, infrastructure, and computational benefits.

The main role of the UCs in the context of the FAIRiCUBE project and the development of the FAIRiCUBE Hub is to test the system (from data ingestion and registration over data storage, manipulation and the development of the ML application to the visualisation of the outcomes), demonstrate what works and what does not and, thus, help to finetune the different elements of the FAIRiCUBE Hub during their development. While details of the UCs are already described in other deliverables (see list at the end of the Chapter), it is nevertheless beneficial for the further understanding of this report to provide a short overview of the objectives, data analysis plan and outcome expectations.

- UC 1: Urban adaptation to climate change.
    - Cities face numerous challenges in combating climate change, including mitigating the Urban Heat Island effect, adapting to shifting precipitation patterns, and addressing urban biodiversity loss exacerbated by human activities.
    - Efforts are underway at both the European and local levels to address these challenges through comprehensive data collection and tailored strategies.
    - On the European scale, data-driven analyses such as cluster analysis help identify cities with similar characteristics and inform decision-making on adaptation strategies.
    - At the local level, cities prioritize the implementation of concrete actions guided by reliable data to mitigate climate impacts effectively.
    - Initiatives like data cubes offer promise in consolidating diverse datasets and providing stakeholders with customized information, with platforms like the FAIRiCUBE Hub poised to support experts in generating tailored solutions for immediate implementation.

- UC 2: Agriculture and Biodiversity Nexus
    - Investigation of farming activities' impact on biodiversity within agricultural landscapes by using the concept of the Dutch Biodiversity Monitor (DBM) for standardized biodiversity assessment. Enhance understanding of the relationship between agricultural practices and biodiversity at a large scale
    - Focus on identifying correlations and causal relationships between farm activities and biodiversity changes.
    - Application of interpretable AI and Causal Machine Learning to attribute biodiversity changes to specific agricultural practices.
    - Implementation on the FAIRiCUBE hub for data collection, analysis, and accessibility, raising awareness among stakeholders in smart agriculture and biodiversity domains about data cubes and AI.

- Exploration of data cube-based infrastructure for improved access to biodiversity-related information. Refine biodiversity estimates within a spatial context and inform decisions on nature-inclusive practices.
- Provision of analysis tools within FAIRiCUBE for extracting causal relationships across different locations and questions.

- UC 3: Environmental Adaptation Genomics in Drosophila
  - Aims to provide insights into evolutionary dynamics and inform conservation strategies by understanding how organisms might respond to ongoing environmental changes. Intersects quantitative environmental and genomic datasets to understand how climatic and human-induced variations impact genetic diversity in *Drosophila melanogaster*.
  - Utilizes *D. melanogaster*, a widely studied genetic model organism with a global distribution, facilitating quantitative analysis of environmental influences on genetic variation. Integrates genomic data from diverse populations across different environments with high-resolution geospatial data to study ecological factors affecting local adaptation and identify genomic targets of selection.
  - Applies established population genetics theories and approaches to infer evolutionary history and predict future responses of *Drosophila* populations to changing environments.
  - Investigates the interplay between environmental factors and genetic diversity within *D. melanogaster* populations to uncover mechanisms driving local adaptation.

- UC 4: Spatial and temporal assessment of neighbourhood building stock
  - Buildings account for approximately 40% of the EU's energy demand and 36% of its greenhouse gas emissions, highlighting the importance of reducing their environmental impact.
  - Policy initiatives such as the "Renovation wave strategy" and "fit for 55" aim to enhance the energy efficiency and sustainability of European building stocks.
  - A transition to circular use of building materials is crucial to mitigate environmental impacts and ensure resilience to supply chain disruptions.
  - Varied levels of data clarity exist for buildings, with newly constructed ones often having detailed information compared to older structures.
  - UC 4 aims to develop models using FAIR-compliant data to estimate material use intensity, energy performance, and greenhouse gas emissions of building stocks.
  - These models will facilitate informed decision-making at the national level, prioritizing investments and promoting sustainable building practices.

- UC 5: Validation of Phytosociological Methods through Occurrence Cubes
  - Phytosociology classifies vegetation communities based on species cover but lacks full explanation of community formation due to limited consideration of environmental conditions.
  - The main objective is to validate traditional phytosociological methods by linking distribution data of plant species from sources like GBIF and botanical collection platforms.
  - Additionally, the aim is to develop a new phytosociological approach using satellite and occurrence data to predict the presence of plant communities in unknown areas.
  - Integration of distribution data with environmental factors enhances understanding of vegetation community formation and classification.
  - This UC contributes to improved conservation and management strategies by advancing the ability to characterize and predict plant communities.
  - FAIRiCUBE offers a framework for ground-truthing by comparing known environmental factors with areas where plant communities occur.

More details on the UCs can be explored in the deliverables of WP2 mainly dealing with their data needs, data analysis plans and synergies between UCs regarding ingestion and processing, and WP3 looking at the Machine Learning approaches and their implementations. Key deliveries to provide context to this report are:

- D2.1 Report on UC data sources
- D2.2 Report on data analysis plan
- D3.1 UC Exploratory data analysis
- D3.2 Machine learning strategy specific for each use case.

2.2 Quality control and validation under UCs' context

Quality control (QC) refers to the application of methods or processes that determine whether data, processing outputs or infrastructure tools meet overall quality goals and defined quality criteria. To determine whether something is 'good' or 'bad' - or to what degree they are so - one must have a set of quality goals and specific criteria against which data and outputs are evaluated.[1]

In general, three different types of quality can be distinguished:

- Thematic quality guarantees that the thematic correctness of the results of the data processing chains will meet the quality requirements. The principial indicators used to assess thematic quality of the results are standardized quality measures (such as overall, user's and producer's accuracy etc.) based on a comparison of the results and independent reference data summarized in a form of validation protocol. This is the standard implemented in EO processing projects.
- Technical quality guarantees that the technical characteristics of the processing outputs agree with the technical specification of the given product (e.g. pixel or grid size, coordinate system, acceptable ranges of the thematic variables, etc.). Technical characteristics of any output data layer produced during operational use of the given data processing chain are finally compared with the corresponding product specification list to check that the output layer meets the technical specifications.
- Scientific quality assesses the developed processing chains from the scientific perspective. It guarantees that the used workflows and processing methods agree with the current state-of-the-art.

Quality control can be qualitative or quantitative, whereby the qualitative systematic accuracy assessment consists of a systematic qualitative survey conducted as a preceding step to the statistically rigorous quantitative accuracy assessment. The qualitative systematic accuracy review can already provide feedback to the production or development team and help to improve the processing chain at an early stage. In addition, quality control should also be able to answer the question whether the products are fit-for-purpose or whether there are any limitations of the products with respect to their intended uses. This last step can be complemented by qualitative checks by user organisations.

[1] https://www.usgs.gov/data-management/quality-control-qc-detecting-and-repairing-data-issues-recommended-practices#:~:text=By%20Data%20Management,quality%20criteria%20for%20individual%20values

3 Use Case Validation

3.1 Context

This deliverable will summarise the various single validation steps that are planned in the context of the FAIRiCUBE UCs and provide a general protocol for their overall quality control. The FAIRiCUBE project foresees several validation steps along the way, i.e., from smaller validation tasks (e.g., data ingestion validation or processing validation) to bigger ones (e.g., the validation of the FAIRiCUBE Hub or the validation of the entire FAIRiCUBE project as such). The UC validation sits somewhere in the middle between the smaller and larger validation activities as it subsumes the smaller tasks that cover the QC of single steps that the UCs must undertake. Strictly speaking, a full UC validation will only be possible once all single steps have been implemented and carried out, i.e., it also includes the validation of the FAIRiCUBE Hub which contains, e.g., the visualisation or final data provision to the users.

Some of the validation steps are documented in dedicated deliverables. These are, next to this overall description of the UC validation, the following:

- D5.3 Validation of data ingestion routines
- D3.6 Validation of processing and ML applications
- D4.6 Validation of sharing
- D6.11 AI ethics assessment

Since UCs go through most or all steps of the FAIRiCUBE chain (i.e., ingestion, data processing/machine learning, output/sharing), this UC validation protocol can be understood as the overarching document for the entire QC chain when executing a UC via the FAIRiCUBE Hub. It might, however, be more correct to call it a UC auditing, as the protocol itself does not deal with the validation of the UCs, but rather controlling that the different steps have been implemented and the UC is conformed to its main questions and the user requirements.

3.2 Elements to be validated

Each FAIRiCUBE UC follows its own logical flow of processes to convert data into information. However, several of the processing steps fall under the same headings which are, by consequence, applicable to all UC implementations. The following subsections will dive deeper into those more general elements and provide an overview of what those steps entail (including references to the deliverables in which they are explained in detail, which have already been shortly mentioned in the previous chapter) and which approach can be used to control their quality. Chapter 4 presents a draft schematic protocol (i.e., checklist) how the UC validation could be implemented in an operational setting. The following steps are key during the implementation of each UC.

3.2.1 Use case specification

The first step for a UC is to define the specifications about objectives, target users, resources (data and processing), methods, and final product. Validation of the UC specifications ensures that UCs have a clearly defined plan before starting to collect, process and analyse the data. A priori statement of these specifications enhances trustability in the UC results. The following requirements should be met for a successful UC:

- **Clear goal defined**: the problem to be addressed and its relevance is stated in a precise way. The objective is specific, measurable, achievable, realistic and time-bound (SMART principles)
- **Target users identified**: the users or user groups who are going to benefit from the outcomes of the UC have been identified and contacted; the goals have been adjusted to the user needs and feedback.
- **Required datasets identified**: the most suitable datasets to reach the goals have been identified; an estimate of the required resources for acquiring and storing them has been made.
- **Required processing and ML/AI approaches identified**: the best available processing and ML/AI approaches to reach the defined goals have been identified; an estimate of the required processing resources has been made.
- **Workflow designed**: how the processing and ML/AI resources are applied to the datasets have been outlined; these steps are optimally documented by means of a workflow diagram; regular updates of the workflow are carried out.

Finally, the UC specifications should be transparently documented and communicated through the designated outlets (e.g. Project website, Github repository).

3.2.2 Data pre-processing and ingestion

Related deliverable: D5.3 Validation of data ingestion routines[2].

Short description: Data pre-processing consists of actions that are undertaken by users before or during the data ingestion process, oftentimes on their own machines, e.g., resampling or calculation of indices from raw imagery. From a quality check standpoint, preprocessing does not differ from any further processing undertaken after the data is ingested in the system. Hence the validation steps pertaining data processing (algorithm implementation validation, benchmarking and comprehensive documentation, see below) outlined in deliverable D3.6 should be applied during pre-processing.
The ingestion of new datasets into FAIRiCUBE Hub depends on the dataset nature and on the target platform (rasdaman or EOX). For a detailed explanation of the ingestion pipeline refer to deliverable D5.2 Ingestion Pipelines. The validation of ingestion is however independent of the target platform. The proposed method includes the following key aspects:

[2] https://nilu365.sharepoint.com/:w:/r/sites/Horizon2021_CUBE/Shared%20Documents/General/deliverables_milestones_inprep/D5_3_Validation%20of%20ingestion_WORKING_OFFLINE.docx?d=w89d2a9bb56154b14b355c3d192f4d518&csf=1&web=1&e=6agGzM

- A list of characteristics to be checked after ingesting the dataset.
- Calculation of descriptive statistics to support checking the defined characteristics.
- Spatial validation.
- An automatic anomaly detection method to identify deviations from previously ingested data, by means of attribute validation.
- Comparison of source and ingested metadata.
- Error labelling and data incorporation.
- Reporting and logging.

This approach does not require domain experts to define data quality constraints or provide valid examples. The implementation of the data ingestion validation can be partially automated by hooking "validation routines" into the ingestion pipeline. A first draft of such automated validation routine within FAIRiCUBE is available at https://github.com/FAIRiCUBE/common-code/blob/main/quality_check/. There is currently a Python file that performs quality checks on ingested raster data and a readme file: https://github.com/FAIRiCUBE/common-code/blob/main/quality_check/quality_check_readme.md.
The data quality control workflow is available via the project Sharepoint: quality_control_data_workflow.vsdx

3.2.3 Data processing, machine learning applications

Related deliverable: *D3.6 Validation of processing and ML applications*.

Short description: The validation of processing and ML applications can be subdivided into data, model, performance, and ethical/bias validation, hence covers part of the value adding chain of FAIRiCUBE, i.e., algorithm implementation validation, machine learning validation, benchmarking, comprehensive documentation. The above-mentioned deliverable concludes with a validation checklist supporting users in their validation efforts.

3.2.4 Machine learning output/data and information sharing

Related deliverable: *D4.6 Validation of sharing*.
Short description: Validation of FAIRiCUBE Hub assures that all components of FAIRiCUBE Hub work correctly, both individually as well as in interaction with each other. To this purpose, validation steps have been described for all FAIRiCUBE Hub components:

- **Information**: within FAIRiCUBE Hub, information is provided both in the form of documentation (using read-the-docs) as well as via the Knowledge Base.
- **Data**: while validation of individual datasets is described in D5.3, here the focus is on the systems describing and serving data. This section includes metadata editors, catalogs and their search functionality as well as web services and APIs for data access.
- **Processing**: while validation of individual processing routines is covered in D3.6, here the focus is on the systems enabling the processing. These approaches must be verified, to assure that they allow for correct execution of scripts and tools developed within FAIRiCUBE.

- **Portrayal**: different tools are utilized by different UC for portrayal of the results of their work. These tools must be validated to assure they correctly display the data generated by the UCs.

In addition, validation of the interaction of these individual components is foreseen.

3.2.5 Assessment of usability/fitness-for-purpose

Product quality[3] is not an absolute measure, but a relative one because it depends on the intended use of the product. The following collection of criteria helps to increase the general understanding on the users' perspective on product quality, and product design should aim at covering as much of it as is feasible and realistic:

- Products should support the users' work, i.e., they should clearly address the policy or thematic area within which the respective user operates. This is a prerequisite for products to be included in the user's working practices, e.g., support decision making or be applicable within (compulsory) monitoring.
- Production should be service-oriented, i.e., put in place a transparent service chain to be able to completely, and in detail, retrace all the steps of alteration that were applied to the original data. It should be possible for users to be involved in their development if they are technically capable to do so.
- Products must be reliable, i.e., make the production method publicly available and attach complete meta-data; moreover, the QC of the final products should be independently executed, and results published.
- Products must be applicable, i.e., they should be fit-for-purpose and appropriate within the users' work environment; this relates to both spatial and temporal coverage, the completeness, scaling, timeliness, resolution, as well as quality and the balance between quality and costs.
- Data systems with which the products are created and shared should be stable and reliable, including being interoperable.

Based on those criteria, a final step of the validation process should be a qualitative assessment of the usability and fit-for-purpose of the output, both when serving internal (i.e., other UCs) or external users:

- In case of a UC synergy, the UC that integrates data from another UC recognises immediately whether the data is of use (i.e., fit for purpose).
- If the processing chain is executed to serve an external user, this user should be able to provide feedback on the product usability as well.

The collection of such feedback should be standardised and formalised as much as possible, e.g., by using a questionnaire that is provided to each user. In general, the quality control and validation of the UC work will be largely qualitative as oftentimes reference data are lacking to implement a qualitative validation of the output products, only some of the processing steps can be quantified and compared to benchmarks (e.g., use of computational resources).

[3] Description including the criteria have been taken from the Deliverable Number 2.4a - Quality criteria for GMES products prepared in the context of the FP6 project GMES Network of Users (GNU) that ran from 2007-2010 (see https://www.copernicus.eu/en/gmes-network-users).

3.2.6 AI ethics assessment

Related deliverable: *D6.11 OEI - Requirement No. 2 Ethics Board review*.

Short description: In the rapidly evolving field of artificial intelligence (AI) and data science, ethical considerations have become increasingly critical. The EC mandates the use of the Assessment List for Trustworthy AI (ALTAI) to assure that any ML/AI tools created in Horizon projects align with the requirements of ethical AI. Within FAIRiCUBE, a socio-technical scenario template was created using a predefined set of questions:

- Aim of the system
- Actors
- Actors' Expectations and Motivation
- Actors' Concerns and Worries
- Context where the AI system is used
- Interaction with the AI system
- AI Technology used
- Clinical studies /Field tests/ Other Evidence
- Intellectual Property
- Legal framework
- Ethics oversight and/or approval.

These questions have been answered for all FAIRiCUBE UCs, assuring that none of the principles for trustworthy AI have been broken.

4 Validation protocol checklist

The protocol mainly serves the purpose to remind UCs (and later everyone who runs a process via the FAIRiCUBE Hub) of the different quality control and validation steps they need to take to assure that the output products are the best quality possible. The main aim is to increase trust in the output products, make sure that the products deliver what was requested and that they are scientifically sound. This checklist needs to be read together with the more detailed checklists of the single validation steps for each of the sub-processes as described in the previous chapter. To avoid repetition, the protocol/checklist below will be high-level and without too much detail.

Figure 1 UC implementation steps that must be validated

As for now, this checklist has only been conceptionally verified with the UCs under FAIRiCUBE and is mostly meant to provide guidance on which steps to follow. Further, the guidance shall be applicable to all current and future UCs that will be executed under FAIRiCUBE and in principle every data science framework. No UC specifics are therefore included and the checklist as the simplified and potentially main output of this deliverable will not contain results of the UC validation. It is foreseen to make a webservice form from *Table 1* with an optional field for comments for each checkbox item and the guidance results with comments can be harvested and provisioned as part of the Knowledge Base services.

UC Specifications (Section 3.2.1) and User assessment / fitness-for-purpose (Section 3.2.5) are validation checks described and developed within this delivery. These required under UC validation. However, the others are validation processes developed and described in other deliverables. These, however, are key and will form part of the overall UC validation which is overarching this deliverable. These are:

- D5.3 – Validation of ingestion
- D3.6 - Validation of processing and ML applications
- D4.6 - Validation of sharing
- D6.11 - AI ethics assessment

Moreover, this report (D2.5 UC validation), together with the four reports above, are part of the overall FAIRiCUBE validation document D1.2 .

Table 1 : UC validation checklist.

UC implementation step	Check type	Check
UC specifications	Clear Goal defined Required datasets identified Required ML/AI approaches identified Workflow designed Visualisation of outputs designed All details in Section 3.2.1	☐
Data pre-processing and ingestion	List of characteristics Descriptive Statistics Calculation Spatial Validation Anomaly detection Error Labelling and Data Incorporation Reporting and Logging All details in D5.3 – Validation of ingestion	☐
Processing and Machine Learning	Algorithm implementation validation Machine learning validation Benchmarking Comprehensive documentation All details in D3.6 – Validation of processing and ML applications	☐
Data sharing	Information Data Processing Portrayal All details in D4.6 – Validation of sharing	☐
User assessment / fitness-for-purpose	Support the users' work Service orientation Reliability Applicability Data systems stability, reliability, and interoperability All details in Section 3.2.5	☐
AI ethics assessment	Ethics (Trustworthy AI) GDPR applicability All details in D6.11 – OEI - Requirement No. 2 Ethics Board review	