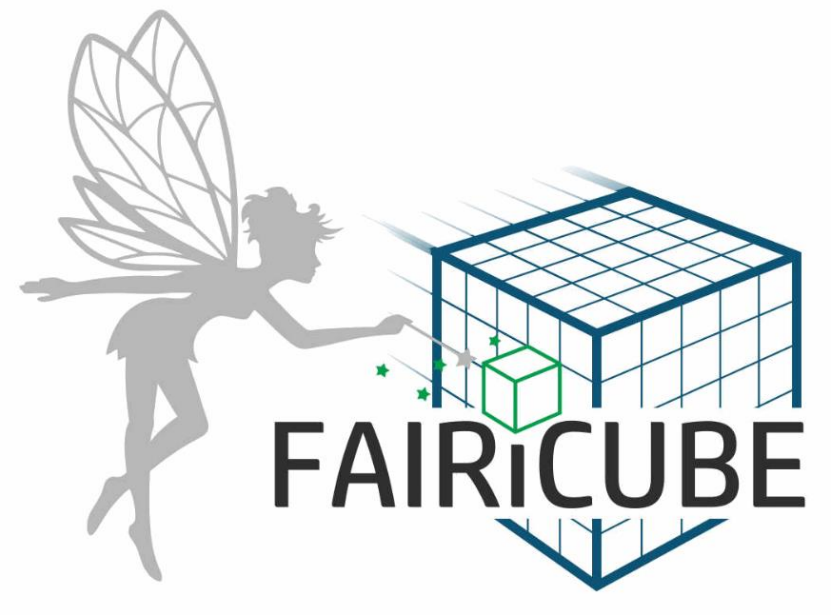

FAIRICUBE – F.A.I.R. INFORMATION CUBES

Work Package 4: Progress Report
Milestone 9: Apps to support community collaboration platform

Deliverable Lead: EOX
Deliverable due date: 31/12/2023

Version: 1.2
2023-11-24

Document Control Page

Document Control Page	
Title	Use cases exploratory data analysis released
Creator	Christian Schiller
Description	M9: Apps to support community collaboration platform
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Christian Schiller
Date of delivery	31/12/2023
Type	Text
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	22/09/2023	Christian Schiller	Initial release
1.0	23/11/2023	Stefan Jetschny	Review, final edits
1.1	24/11/2023	Christian Schiller	Final edits
1.2	01/12/2023	Stefan Jetschny	Final review

Disclaimer

This document is issued within the frame and for the purpose of the FAIRICUBE project. This project has received funding from the Horizon Europe 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.

Table of Contents

Document Control Page 2
Disclaimer 3
Table of Contents 4
List of Tables 5
1 Introduction 6
2 Deliverables contributing to M9 7
3 Progress summary 8

List of Tables

Table 1: Deliverables related to M9_____ 7

1 Introduction

This document provides an overview of the operational status of the implemented applications enabling the development of use case specific solutions and supporting the collaboration on the FAIRiCUBE Hub community platform.

The core mission of FAIRiCUBE is to enable players from beyond classic Earth Observation (EO) domains to provide, access, process and share gridded data and algorithms in a FAIR and TRUSTable manner. A further goal is to leverage power of Machine Learning (ML) operating on multi-thematic datacubes for a broader range of governance and research institutions from diverse fields, who are at present cannot easily access and utilize these potent resources. Therefore, when providing data, processing functionality and data products to relevant stakeholders all these aspects must be considered.

Following the **FAIR principles**, while data is becoming increasingly **findable**, **accessible**, and **interoperable**, true **reusability** depends on the availability and functionality of suitable processing mechanisms. Especially the interoperability and reusability part of data management still causes much trouble and has frequently been found to be highly complex and very time consuming. FAIRiCUBE aims to advance the FAIRness of both data and data analysis and subsequent products by enhancing the reusability of existing data, as well as show, and hopefully overcome, short comings in this area.

The broader potential of machine learning applied to multi-thematic datacubes has rarely been demonstrated elsewhere. Most ML applications on datacubes focus on a limited number of data sources or often just temporal steps within one data set. FAIRiCUBE aims to deliver the power of datacubes and ML to decision makers and data scientists. To achieve this the FAIRiCUBE Hub acts as the community collaboration platform allowing sharing and concurrent work on specific scientific questions. Additionally, the provisioning of two pillars to create and manage datacubes enables users a wider spectrum of data analysis possibilities. The provisioning of diverse ML-Toolkits allows users to choose the best tool for solving their research questions.

2 Deliverables contributing to M9

There are several formal deliverables contributing to M9 Milestone as listed in Table 1. This milestone represents a compilation of the already implemented workflows and the implemented ML-tools.

Table 1: Deliverables related to M9

Description	Lead Beneficiary	Type	Dissemination level	Due dates
D3.4 Processing knowledge base services	EPS	DATA	Public	29.02.2024
D4_1_FAIRiCUBE-Hub-Architecture	EOX	DEM	Public	31.06.2023
D4.5_FAIRiCUBE_Apps_supporting_community_collaboration	EOX	DEM	Public	31.12.2023
D5_2_FAIRiCUBE ingestion pipelines	JUB	R	Public	31.12.2023

The deliverable D3.4 Processing knowledge base services" presents the FAIRiCUBE Knowledge Base providing the community with a set of tools, documents, algorithms, code, tips and tricks, mistakes to avoid, and examples of use.

The deliverable "D4_1_Deliverable_FAIRiCUBE-Hub-Architecture" provides a detailed description of the components comprising the FAIRiCUBE Hub, its deployment and operations strategy based on control and worker plane as well as on-boarding requirements and processes for service and app providers. Although there is no formal planning to re-release this deliverable, we intend to keep this document up to date reflecting all changes and additions applied to the FAIRiCUBE Hub.

The deliverable "D4_5 _Apps_supporting_community_collaboration" outlines the process of collecting and streamlining the requirements of the FAIRiCUBE use cases and provides a short introduction of first set of applications provided for the use cases.

The deliverable "D5_2_Deliverable_FAIRiCUBE-Ingestion-Pipelines" describes how data sets are ingested into the catalogue and into the two data management systems, comprising the FAIRiCUBE Hub.

3 Progress summary

The project's goal is to leverage power of Machine Learning (ML) operating on multi-thematic datacubes for a broader range of governance and research institutions from diverse fields, who are at present cannot easily access and utilize these potent resources. Machine Learning may be able to provide answers to question of highly complex nature driven by multi-parameter factors.
The apps provided in the FAIRiCUBE Hub shall foster finding solution to the questions ask by the UC scientists. The apps provide the basis where ML based algorithms can be developed, trained, applied, shared and discussed.

In addition, general information and lessons learned shall be collected and made available for other future users of FAIRiCUBE Hub. Currently there exist two partially overlapping portals / information. The FAIRiCUBE **Knowledge Base** (KB, https://fairicube-kb.dev.epsilon-italia.it/) is targeting the output and experience of the use cases executing on the FAIRiCUBE Hub. It provides a self-training library containing a set of links to web pages and project resources, appropriately selected, and organised into categories, with the aim of providing the user with a basic background on FAIRiCUBE Knowledge Base topics. An interactive query tool provides the user with an extensive search possibility inside the KB. Its core task is to provide the community with a set of tools, documents, algorithms, code, tips and tricks, mistakes to avoid, and examples of use.

The second information portal is currently focusing on easy to follow 'hands-on' type chapters, a la 'Getting Started', and 'How To…', for to gain understanding and usage of the FAIRiCUBE Hub associated tools (https://fairicube.readthedocs.io). It further allows to collect, describe, and present example Jupyter notebooks and other useful technical instructions and allows users to submit examples and descriptions via GitHub merge requests. As this documentation is stored in GitHub and users can therefore easily extend the content and supply additional examples. Note: At the time of writing, it is not finally decided if the two documentation portals/tool will stay 'stand-alone' or merged into the Knowledge Base.

An important pillar of the FAIRiCUBE Hub is to provide a platform where users can develop, test, and share their algorithms to analyse the respective data sources. This includes the versioning, sharing, and collaborative usage of the various ML artifacts like code, data, models, results, etc. . Based on an initial user consultation, during an early phase of the project, readily available Open-Source tools like **MLflow**, **TensorBoard**, **Data Version Control (DVC),** have, recently been integrated and are now available to be exploited by the use cases.

MLflow is an open-source platform for managing the end-to-end machine learning life cycle. It allows the user to track experiments, package code into reproducible runs, and share and deploy models. MLflow can be incorporated into Jupyter notebooks or other code and supports multiple programming languages. It is widely used in industry and academia and is constantly evolving to support the latest trends and technologies in the field of machine learning. At a high level, MLflow consists of four main components: tracking, projects, models, and registry. All components can be accessed via Python code in the FAIRiCUBE Lab.

To support model evaluation during training, the FAIRiCUBE Hub processing environment is also extended by a **TensorBoard** to support the tracking of individual experiments and training runs. This tool can be used with PyTorch and TensorFlow, and it provides a state-of-the-art toolset for data scientist to inspect the tuning and training process and compare metrics. In machine learning, to improve something you often need to be able to measure it. TensorBoard is a tool for providing the measurements and visualizations needed during the machine learning workflow.

DVC is a Data Version Control system for collaborative data management like ML artifacts. It is a free and open-source, platform-agnostic version system for data, machine learning models, and experiments. It is designed to make ML models shareable, experiments reproducible, and to track versions of models, data, and pipelines. DVC works on top of Git repositories and cloud storage. DVC's features can be divided into three categories: data management, pipelines, and experiment tracking.

More details about MLflow, TensorBoard and DVC can be found in the deliverable *D4.5_FAIRiCUBE_Apps_supporting_community_collaboration* as well as in the deliverable *D4_1_Deliverable_FAIRiCUBE-Hub-Architecture*.

A second round of user consultation will happen in the beginning of December 2023. It will include an interactive discussion group to collect and clarify the requirements of the various use cases. In order to better facilitate the usage of the already available ML tools, an introductory lecture will be given to reduce the barrier to use the ML tools. In the frame of this consultation round also a discussion about further needs will be initiated and will be used to collect additional user needs. Based on the outcome of this second user consultation round and the collected use case requirements upgrades to the FAIRiCUBE Hub, like additional ML tools could be performed.

During the course of the project quite a big effort was put into the collection, harmonization and preparation of metadata describing the datasets used in FAIRiCUBE. This resulted in so many fields to be covered in the metadata description that the initial process of data-metadata submission and ingestion could not be followed anymore due to some technical limitation by GitHub. Therefore, a new system to handle data-metadata was proposed, accepted, developed, and put in place. This new process is based on an interactive WebGUI, which allows data to be entered and edited and provides additional features like providing a drop-down-list for certain fields. This makes the collection much easier since e.g., Dates will always be correctly formatted as required. When the WebGUI is filled and submitted, the input from the WebGUI will be collected, translated into a STAC item and stored in GitHub. Further changes on any dataset can be made as Pull Request. The same WebGUI can be used to submit new datasets as well as to edit already available datasets. Further details can be found in the *D4_1_Deliverable_FAIRiCUBE-Hub-Architecture* and *D4_2 Public Listing (Catalog) of FAIRiCUBE data resources.*