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

Work Package 3: Process
Milestone 8: Processing and ML applications for each use cases released

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

Version: 1.1
21/11/2024

Document Control Page

Document Control Page	
Title	Processing and ML applications for each use cases released
Creator	Stefan Jetschny
Description	M08: Processing and ML applications for each use cases released
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Stefan Jetschny, Mohamed-Bachir Belaid
Date of delivery	31/12/2024
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	01/12/2023	Stefan Jetschny	Initial draft structure
1.0	10/01/2024	Stefan Jetschny	Final version for review
1.1	15/01/2024	Jaume Targa	Review and format check

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

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

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

List of Tables

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.

1 Introduction

The use cases executed under the FAIRiCUBE Hub and during the project duration are not formally a key delivery from the project but nevertheless serve a significant role: to test, develop, improve, and co-define the FAIRiCUBE hub which bundles all cloud-based services to be executed while solving typical data science research questions. As the use cases (UCs) are truly diverse, operate on different scale length & time ranges and cover various scientific fields, we thereby cover a wide range of potential data science tasks. Providing all online services such as accessing, ingesting, storing, processing, and sharing data, in a FAIR, efficient, streamlined, and user-friendly way is a key project component and only by executing our UCs, we can determine which components of the FAIRiCUBE Hub are already sufficiently working and where is room for improvement.

As part of the data science work in WP3 and in close collaboration with the WP2 domain experts, we have first studied the nature and characteristics of the data identified by the UC owners as being relevant to solve the UC research question(s). This was achieved as part of M5 "*Use cases exploratory data analysis released"* and is effectively the basis of M6. Driven by this exploratory data analysis, a machine learning strategy was developed for each use case which efficiently balances the UC needs, the appropriate machine learning algorithm and the computational efforts and resources required. This was achieved as part of M6 "*Machine learning strategy specific for each use case released*".

As the UC work can be very specific due to the nature of the specific scientific domain, non-standard processing steps as well as tools and functionality are required to address and solve some unique problems. This comprises specific ML models and approaches to tackle specific UC prediction solutions or standalone applications to e.g. retrieve data across data cubes. All exploiting the unique possibilities of the FAIRiCUBE Hub architecture and highlighting the benefits of analysing/processing rasterized data cube data. An important aspect of FAIRiCUBE is not only to document and make available the developed processing and ML applications for each use case but also to publish the computational resources that are consumed during runtime. This shall provide detailed knowledge for future project planning as well as assess the environmental impact of executing ML and data processing methods in terms of consumed energy and CO_2 equivalents.

Funded by
the European Union

This project has received funding from the Horizon Europe research and innovation programme under grant agreement No. 101059238.

2 Deliverables contributing to M8

In contrast to previous milestones, there are no direct deliverables formally contributing to the M8 Milestone, instead, they are spread out over several related deliverables as listed in Table 1. Each UC has specific application needs and therefore focuses on a diverse range of ML and processing applications:

- UC1 has developed several urban classification models, published under https://github.com/FAIRiCUBE/uc1-urban-climate and described in D2.2 and D3.2
- UC2 has focused on a feasibility study to make use of user-defined functions (UDF) close to the rasdaman database core to demonstrate the benefits of inferring an ML model close to a large amount of data. This is covered in D3.2
- UC3 has developed a user-friendly tool to retrieve a single vector of data through several data cubes and data layers for a given geospatial coordinate, called the *Wormpicker*. This allows the correlation of species properties and occurrence with environmental data for each of the sspecies locations. This is available under https://github.com/FAIRiCUBE/uc3-drosophola-genetics/tree/main/projects/WormPickerOOP and is described in D2.2. UC3 has further developed a ML strategy to gap-fill genomic data which is published under https://github.com/FAIRiCUBE/uc3-drosophola-genetics/tree/main/projects/gap_filling as well as described in D3.2.
- UC4 has employed several ML modes to provide essential input to the description of buildings for the energy performance assessment, e.g. building height, age and type, as these properties are not widely available. This is available under https://github.com/FAIRiCUBE/uc4-building-stock and described in D3.2.
- UC5 started later than the other UCs has not yet entered the processing and ML application stage.

Some typical examples of consumed resources while running the processing and ML application as described above are listed in D3.3. A full overview is envisioned to be available through the FAIRiCUBE knowledge base (KB, https://fairicube-kb.dev.epsilon-italia.it). Deliverable D3.1 indirectly contributed to the M8 milestone by providing insights into the data, describing the characteristics or shortcomings (with e.g. respect to the completeness of data) and suggesting the optimal machine learning methods and processing steps.

Table 1: Deliverables related to M8

Description	Lead Beneficiary	Type	Dissemination level	Due dates
D2.2: UC Analysis Plans	4SF	R	PU	31.12.2023
D3.1 UC exploratory data analysis	NIL	R	PU	30.06.2023 / 29.02.2024
D3.2 Machine learning strategy specific for each use case	NIL	R	PU	30.06.2023 / 29.02.2024
D3.3 Processing and ML applications	NIL	DATA	PU	31.12.2023

Funded by the European Union

This project has received funding from the Horizon Europe research and innovation programme under grant agreement No. 101059238.

3 Summary

Each use case (UC) has defined domain-specific research questions, documented as use case analysis plans, and they are being addressed as data science tasks. As part of the data science work and following the thorough analysis of the suitability and characteristics of the input data (exploratory data analysis) a use case-specific machine learning strategy has been developed, and partly executed which made available machine learning models that are targeted for specific prediction and classification applications.

- UC1: classification of cities according to land use land cover, gap filling of socio-economic data
- UC3: gap filling of genomic data through unsupervised classification and auto-encoder deep neural networks
- UC4: building height estimation through a supervised ML model

In addition, UC2 explored the suitability of UDFs for ML model inference on large amounts of data and the performance benefits of executing this as close to the data source as possible. Finally, UC3 developed a standalone "*Wormpicker*" tool that can extract values from several data cubes and data cube layers for a given point location. This can be a very illustrative tool and suitable for a wide audience. All documentation of ML and processing applications are documented as processing meta data, accompanied by monitored compute resource information and subject to further improvements.

Funded by
the European Union

This project has received funding from the Horizon Europe research
and innovation programme under grant agreement No. 101059238.