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

Project Number: 101059238

WP 2 Use

D2.3 UC Ingest/Process Synergy Report

Deliverable Lead: s4e
Deliverable due date: 30/06/2024

Version: 2.0
2024-09-24

Document Control Page

Document Control Page	
Title	D2.3 UC Ingest/Process synergy report
Creator	s4e
Description	This deliverable describes the data ingestion experiences of the 5 UCs and will also highlight synergies that emerged between UCs.
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	WER, 4SF, NIL, NHM
Date of delivery	30/06/2023
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	22/05/2023	Mirko Gregor, s4e	Draft setup, headings, and partner / contributor assignments
		Rob Knapen	Provided UC2 related content
0.2	08/06/2023	Mirko Gregor, s4e	Updating content of chapters
0.3	13/06/2023	All	Finalisation of first draft
0.4	27/06/2023	Jaume Targa	Full review
1.0	29/06/2023	Mirko Gregor	Integration of review comments, finalization of first version
1.1	17/05/2024	Mirko Gregor, Susanna Ioni, Marian Vittek, Rob Knapen, Martin Kapun, Sonja Steindl	Update of first version towards v2, integration of reviewers' comments from first periodic report
2.0	21/06/2024	Jaume Targa, Stefan Jetschny	Full internal review and format checking

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

1 Introduction

WP2 ensures the efficient execution of the FAIRiCUBE use cases (which are described in more detail in deliverable D2.3 – Use Case Analysis Plans), assuring those potential synergies pertaining to both data and processing are identified and leveraged.

The meanwhile five use cases are the following:

- UC1 - Urban adaptation to climate change
- UC2 - Biodiversity and agriculture nexus
- UC3 - Drosophila Genetics
- UC4 - Spatial and temporal assessment of neighborhood building stock
- UC5 – Validation of Phytosociological methods through Occurrence Cubes

Acting with an "outsider" role, the supervision crosscuts through all UC activities ensuring harmonisation with both upstream (data sources, ingestion, and processes) and downstream (results, promotion, and distribution of outputs) activities. This deliverable, D2.4, is the result of the Task 2.4 Ingestion/Provision/Processing, which focuses on the coordination of the ingestion and processing from each Use Case (UC).

If the processing yields satisfactory results, the partners involved in the UC will document the analysis methodology, the reasons for the success, and identify potential alternative applications of this approach. In the case of unsatisfactory results from processing, these will also be documented, before UC partners can go back to Task 2.3. They will then work on developing an alternative analysis approach, potentially incorporating additional data sources.

In the case that the UC is still receiving unsatisfactory results despite diverse approaches, this "failure" should be documented, detailing why the UC could not be satisfactorily supported. This process will be documented, providing information on the common tools/processes, as well as the success/failure of the UC.

2 Ingestion and processing

Data ingestion plays a crucial role in data-centric processes, serving as the initial step in getting data from one location to another. It is crucial to ensure that the information is obtained and delivered accurately and timely. The most important thing about data ingestion is knowing what kind of information will be needed by your target environment and understanding how that environment will use that information once it arrives there.

Concretely, data ingestion is the process of importing and loading data that is coming from various sources into a system. Data processing is the method of collecting raw data and translating it into usable information.

Both ingestion and processing as well as analysis resources are described in various deliverables, coming from WPs 3, 4 and 5. These deliverables contain many very detailed descriptions of the ingestion pipelines and the processing and analysis resources and tools, separated according to the platform system used (EOX or rasdaman), hence they won't be repeated here.

This report is an update of the first version of it that was provided a year ago at the end of the first year of the project. The update now aims to provide a more mature view at the potential synergies that emerged during the first two years of the project, particularly after all use cases have been able to understand and get acquainted with the procedures applicable to the two main entry points, namely EOX and rasdaman. It will at the same time also take into account the reviewers' comments during the periodic reporting and update the report accordingly.

In this chapter, we will focus on the data inventory. We will explore the overlaps between data sets that have a certain priority for the different use cases. Considering the overall ambitions of FAIRiCUBE, the objective is to ingest a data set once and make it reusable by all other use cases that require the same data set. However, due to the current setup of the two systems, it can be challenging to achieve these synergies in practice. The report will include the insights gained from the initial cases as they relate to this issue.

2.1 Initial data inventory

At the project's beginning, the four designated UCs initiated the systematic compilation of a comprehensive list of geospatial data sets intended to be used in their analysis. This data inventory has been updated during the course of the second project year and been stored on the common shared Teams workspace and described in more detail in deliverable "D2.2 – Report on UC data source synergies" providing detailed descriptions. To get rid of the cumbersome way of working with an Excel file, data requests were handled via GitHub issues within a dedicated repository (https://github.com/FAIRiCUBE/data-requests). Recently, a data catalog GUI has been finalised via which data requests will now be handled in a standardised way including all required metadata (https://catalog-editor.eoxhub.fairicube.eu/).

Additional information about the UCs can be found in "Deliverable D2.3 – Use Case Analysis Plans". In the meantime, a fifth UC has been added to the portfolio, dealing with occurrence cubes.

One main element of the common data inventory was the identification of data sets that were relevant to more than one UC while also indicating the respective priority. Analysing this list makes clear that several of the pan-European data sets (such as the Copernicus High-Resolution Layers) are required by more than one UC only, even though not necessarily with the same priority. Also a few climate datasets (e.g., temperature) are required by three or four UCs.

Full synergy of the UCs would then mean that certain data sets only need to be ingested once, assuming that the EOX and rasdaman systems are able to communicate with each other. This is not yet the case but will be explored in the next steps.

In order to enhance the visibility of these synergies among all UCs, monthly synergy meetings (that now also include WP3 representatives to align use case data with use case data science) are being organised. These meetings serve two primary purposes: firstly, to discuss the overall progress and data requirements of the UCs, and secondly, to explore the data science techniques foreseen to be applied to the research questions specific to each UCs' data. These collaborative meetings facilitate knowledge sharing and foster a deeper understanding of the interconnections and potential collaborations among the UCs.

2.2 UC experiences with data ingestion

UC1: Urban adaptation to climate change

To facilitate data management and processing, raster datasets are registered to the SentinelHub service. Therefore, the ingestion routine for raster datasets is based on the SentinelHub documentation. These are the steps in short:

- raster file is tiled and converted in cloud optimized geotiff (COG) locally with GDAL utility (link to a/p resource: https://catalog.eoxhub.fairicube.eu/collections/no-ML%20collection/items/26YU5NATNB)
- COG tiles are uploaded to the UC1 or to the fairicube-common object storage S3 bucket (configured to properly connect to SentinelHub).
- COG tiles are registered to SentinelHub using this Jupyter notebook template https://github.com/FAIRiCUBE/uc1-urban-climate/blob/master/notebooks/f01_ingestion/ingestion_00_template.ipynb.

Until now, we followed this method to ingest Urban Audit, NUTS3 (city boundaries), Urban Atlas (land cover/land use), Environmental zones, CLMS Tree cover density, CLMS Imperviousness.

Defining and troubleshooting the ingestion process proved to be time-consuming, primarily due to incomplete documentation of SentinelHub and the occasional lack of technical support. Currently, the ingestion workflow has been fine-tuned, but there is still room for better automation and subsequent reduction of ingestion time.

To fulfill our requirements, vector and tabular datasets should be ingested into a geodatabase. However, setting up a geodatabase for the project has not been feasible thus far. As a temporary solution, the vector and tabular data are stored in shapefiles and SQLite files within the common object storage. The deployment of a test geodatabase is, however, agreed upon and ongoing.

Finally, some datasets are fortunately already freely available through cloud services, e.g. CLMS Corine Land Cover in EuroDataCube, ERA5 reanalysis model in Google Cloud. These datasets are accessed through API from the working environment.

UC2: Biodiversity and agriculture nexus

Until now two datasets have been "formally" ingested. The initial dataset (for internal project use only) is the test input data used for the development of the deep learning UDF for Rasdaman (for more details see the UC2 related section in D3.3). This was an already existing dataset that has been ingested as-is, while keeping the original coordinate reference system and content, to make sure model inference on the test data should produce the same (or at least very similar) results as in the original situation (when using the deep learning model not embedded into Rasdaman). The dataset consists of two raster files in TIFF format:

(1) A set (667 MiB) of stacked Sentinel-2 images (R, G, B, and NIR bands, 10m grid cells) from 2018, consisting of 7 monthly images covering the main crop growing season in The Netherlands. All images were clipped to a study area containing both representative agricultural and more nature rich areas, with some small cities are inland lakes and rivers.
(2) A 10m x 10m rasterized version (24 MiB) of all registered crop parcels in 2018 in the same region, taken from the open data provided by the Dutch Land Parcel Information System (LPIS).

Selected parts ('chips' in computer vision parlance) of (1) have been used for training the deep learning model, with (2) providing the ground truth data (labels) and added for validation and testing purposes.

The second ingested dataset is the Dutch National Land Use Database (LGN). Time series of six geospatial layers from years 2012, 2018, 2019, 2020, 2021 and 2022 were stacked together. Since the number and definition of classes were not identical, a common set had to be established. Also, the spatial resolution of 2012 layer was resampled from 25 to 5 meters to match the remaining layers.

Currently, the ongoing work for this use case focuses on handling local datasets. This is primarily due to privacy concerns related to the precision of agriculture data at the farm-specific level, which need to be addressed first. Additionally, the initial rasterization of the biodiversity occurrence data is being addressed to enable its ingestion into the gridded data cubes for subsequent processing and use in the planned ML tasks.

UC3: Drosophila Genetics

Thus far, only a limited number of datasets have been identified as being relevant to all use cases (Corine Land Cover, Dominant Leaf Type, EU demography, Forest Type, Near Surface Air Temperature), but no dataset specific to the needs of UC3, like the genomic data of *Drosophila*, has been ingested to date. However, for most of these datasets, the reference years of the layers are very limited and do not cover the collection dates of the *Drosophila* samples in the genomic dataset processed in UC3 to the same extent. The only ingested environmental dataset available to date that covers most of the temporal range of the genomic dataset on a daily basis and also provides sufficient metadata for interpretation of values is near surface air temperature (NSAT) from Copernicus. We are using this dataset to optimize our analysis pipeline for the application of additional, to be ingested, gridded datasets. We further experienced occasional unannounced changes of data formats (e.g., projection types) on the rasdaman server which required regular adjustments of our highly customized script to extract point data for point coordinates from rasdaman. As a consequence, it remains difficult to validate the ingestion process and we suggest setting up a more formal and more detailed changelog documentation of rasdaman services to facilitate the adoption and implementation of updates and new features in our analysis pipeline. Although ingestion of layers is formally making progress, the practicality and easy accessibility for UC3 could be improved by the abovementioned updates in the future.

In parallel to the data ingestion through rasdaman, UC3 currently employs the EOX hub for data ingestion using local data for the city of Vienna as part of a new collaboration with UC1 (see below). Additionally, Copernicus data is accessible via the Climate Data Service API from Copernicus provided at the EOX hub.

UC4: Spatial and temporal assessment of neighbourhood building stock

Formally, no data has been ingested during the execution of this use case, i.e., no data ingestion routine has been used. Data was either accessed through an API (Open Street Map data) or by manually downloading data to local disc and later uploads to EoXHub (a very good demonstrator of how to upload the data has been found and documented). Due to the storage on EoXHub, the data is accessible for everyone within FAIRiCUBE. As a consequence, synergies may be established by other UCs. So far, the uploaded data to the EoXHub platform is limited to a few datasets, namely digital elevation model, building tags from openstreetmap, orthoimagery, building rooftop geometry, and tabular data. And these datasets correspond to the cities of Barcelona, Luxemburg, Oslo, and Vienna. Some of the processing outcomes from UC4 (e.g., building energy performance) are in vectorized form. Here, learning outcomes from UC1 in storing vectorized datasets into grided format will be tested.

UC5: Validation of Phytosociological Methods through Occurrence Cubes

For what concern the UC5, a first source of data is based on occurrence data of taxa and Eunis habitat types. The former are retrieved by the Global Biodiversity Information Facility, while the second is a raster file. A second source of data is instead represented by Earth Observation data coming from the Copernicus services. In particular, these data include the datasets Copernicus DEM and several climatic datasets from the reanalysis of the model ERA5.

Up to this point, only the Copernicus DEM dataset has been ingested by EOX and is available from the Sentinel Hub. However, the UC5 does not have a subscription to access the Sentinel Hub. Regarding the rest of the required data, UC5 is in the process of requesting the ingestion of 15 datasets from the ERA5 model and to upload the occurrence data and EUNIS habitat raster in the EOXHub.

3 Synergies of ingestion and processing routines

3.1 Synergy example: occurrence cubes in UC2 and UC5

When working with biodiversity-related data, it typically involves obtaining species occurrence data from relevant organizations that take care of the collection and initially harmonising the information. This 'raw' input data then needs to undergo processing to transform it into a suitable format for further analytics. In the FAIRiCUBE project, our objective is to streamline this process by defining and producing occurrence cubes, which can be shared as data cubes, enabling reuse both by other project partners as well as potential users beyond FAIRiCUBE.

For UC2 this relates to the intended calculation of a 'biodiversity index' from species observations, that then can be related to agricultural activities. UC5 intends to use such occurrence cubes for validating phytosociological methods, comparing EUNIS Habitat types and their main vegetation associations with the recorded occurrence of their individual component species.

An occurrence cube is a gridded dataset calculated by aggregating or distributing the available occurrences data (i.e., the species field observations, typically of presence) across a selected grid, with a certain grid cell size and coordinate system. Space, time, and taxonomy are evident dimensions of such hypercubes, with other, more thematic dimensions, still under consideration. There may also be multiple occurrence cubes derived from the same data, e.g., a species occurrence cube with aggregated proportional abundance, and a species group occurrence cube with boolean presence values.

The 'cubing approach' described in Oldoni et al.[1] serves as the basis of this work, whereby we are currently refining the described algorithm for transforming the vector (point or polygon) based observation data into a grid format. One critical aspect concerns the correct interpretation of the attributes available in the datasets (which might differ per data provider), including the various radii and geometries used to indicate observation accuracy, species range of motion or territory, and possibly additional ranges for protected species. In addition, we are investigating allocation of individuals to grid cells, where in contrast to the approach selected by Oldoni et al., we propose a more proportional mode of individual allocation.

3.2 Synergy example: possible impacts of environmental factors in cities on genetic variations of Drosophila (UC1 and UC3)

One of the central questions in climate change research in the face of the accelerating biodiversity crisis is how organisms can cope with changing environments in the long term. This applies particularly to urban areas where, for example, soil sealing, increased air pollution, and habitat fragmentation results in increased challenges to wildlife. Moreover, urban areas are often characterized by higher ambient temperatures compared to natural habitats, which further influences wildlife with specific and narrow thermal niches. It is thus important to better understand how urbanization affects the survival and adaptation of organisms.

In this synergy project we want to tackle unresolved questions concerning the evolutionary potential of organisms living in urban or natural environments. We will therefore compare evolutionary trajectories, i.e., changes in allele frequencies and genetic variation, of the fruit fly *Drosophila melanogaster,* a long-standing genetic model organism, in urban and natural environments through space and time. The recently extended DEST dataset (http://dest.bio) contains genomic data of hundreds of fly populations from both environments that have been collected and sequenced during the last 10 years.

[1] doi: https://doi.org/10.1101/2020.03.23.983601

The team at the NHM Vienna will provide the allele frequency data which can be used to estimate genomic change due to environmental stress. The team of space4environment will provide high resolution environmental data (such as land use/land cover changes, soil sealing and its changes, or urban tree cover as well as urban green infrastructure), which can be used to trace for critical environmental factors that may influence genetic variation, specifically in urban areas. However, given that *D. melanogaster* is a human commensal, we speculate that urbanization could be even beneficial to *Drosophila* given that flies have long adapted to co-existence with humans, which may be reflected in higher amounts of genetic variation close to human settlements compared to rural areas.

This synergy allows us to optimally use domain expertise and resources by the involved Use Cases when working on different stages of the synergy project. Tasks which require such different domain expertise have already been completed and are explained in the following section.

Data selection and processing

The first requirement in this project is an assessment of available and useful data for the question posed. Datasets that are potentially relevant for the organism *Drosophila melanogaster* that specifically cover an urban area (in this case Vienna) and surrounding areas have been investigated and preselected by UC3. Most datasets are provided by the City of Vienna and offer very dense information on a variety of urban features. The technical properties of these data like projection systems, resolution, data format and other aspects concerning suitability to UC1-specific workflows and infrastructures, have been assessed and processed by UC1.

Harmonization of the available data was carried out by UC1 and attributes relevant for the project have been newly conceptualized together with UC3 to suit research questions centred around the influence of urban environment on genetic variation. This was achieved by either categorizing already existing attributes (such as daily average temperature) as potential drivers of genetic variation or genetic homogeneity in the population, or by recalculation and accumulation of certain specific attributes (such as density of fruit trees) that potentially influence the genetic variation in Viennese *Drosophila melanogaster* populations. These attributes were categorised based on their expected impact on the biology of fruit flies, e.g. the presence of fruit trees (e.g. *Prunus sp.*), may have a more positive influence on genetic diversity of *Drosophila* than the presence of industrial buildings. We aggregated different attributes into a newly created "Fly indicator", which can evaluate areas of interest as optimal or non-optimal habitats for the *Drosophila melanogaster* flies, based on the expected cumulative effect of the independent attributes on the fly biology. The generation of this aggregated indicator and integration into coverage data has already been started by UC1.

Using processing results to assist sampling

This Fly indicator will assist in choosing sampling locations for *Drosophila melanogaster* samples within and in the suburbs of Vienna and will also be used for data analysis and postprocessing. Accordingly, UC3 started collecting fly populations originating from different locations in the city of Vienna and will generate allele frequency data through pooled resequencing of these flies. These data will be used for association analysis. A maximum of 100 populations will be sequenced and included into the Landscape Genomics Analysis pipeline, which was set up by UC3 in the scope of FAIRiCUBE

Additional analysis of urban *Drosophila* data

The Landscape Genomics Pipeline established for world-wide *Drosophila* samples will be adopted for statistical analysis of the Viennese Drosophila samples using distinct environmental measures, like temperature and precipitation, which are directly associated to the physical location of the samples. This synergy projects carries the potential to extend these analyses by using combinations of environmental variables, presence-absence-data of certain features on a very fine-grained geographical resolution.

4 Conclusions and outlook

While not all objectives related to workflow setup and implementation have been fully achieved, there has been promising progress in the area of ingestion and processing synergies. Regarding data ingestion, the procedures for ingesting data into the EOX and rasdaman systems appear to be well-defined. However, there is a lack of clear interfaces for direct interaction. It is necessary to develop and implement improvements that enable the direct retrieval of ingested data, irrespective of the system in which it was ingested.

On the processing and analytical side, the synergies regarding occurrence cubes are an interesting playing field for technological development of a common processing of dedicated cubes (how are they created, i.e., processed), but also at a later stage for the conception of synergetic analytical tools or machine learning applications.

Another interesting synergy has developed with the collaboration of UC1 and UC3 on a project that UC3 has conceived for Vienna. The aim of this new project is to create a network of fly collectors by which flies should be collected to sequence their genomic variations and relate them to environmental conditions. Both during the process of identifying ideal places to position fly traps (the "Fly indicator") and later during the analytical phase, UC1 provides urban data on various environmental topics.

Another synergy is related to the application of urban areas as reference units for the analysis. Both UC1 and UC4 work on cities, but with a different focus. Nevertheless, we have merged the two UCs during the preparation of the first ever FAIRiCUBE UC Seminar that is part of the stakeholder consultation process, organized by WP6. But so far, there are only very limited processing synergies, so there is a lot of room for further exploration.

Compared to the first version of the report, the UCs have made progress regarding the ingestion of data sets, also further synergies could be identified. It will nevertheless be important to deepen those further which will now happen during the coming months.