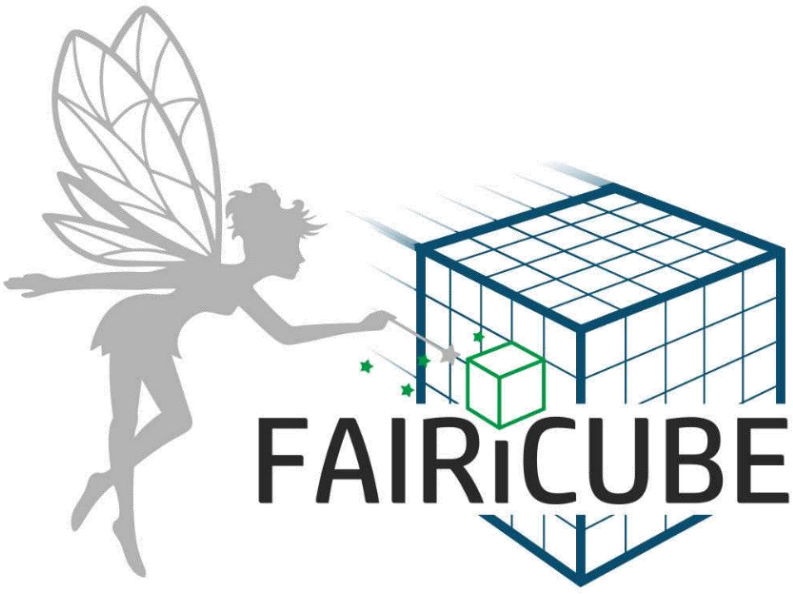

FAIRICUBE – F.A.I.R. INFORMATION CUBES

Work Package 5: Ingest
Milestone 10: Description of datacube ingestion pipelines released

Deliverable Lead: JUB
Deliverable due date: 30/06/2023

Version: 1.0
2024-11-21

Document Control Page

Document Control Page	
Title	Description of datacube ingestion pipelines released
Creator	Mohit Kumar Basak
Description	M10 Use cases exploratory data analysis released
Publisher	"FAIRICUBE – F.A.I.R. information cubes" Consortium
Contributors	Peter Baumann
Date of delivery	23/11/2023
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	22/08/2023	Mohit Kumar Basak	Initial draft with headlines
0.2	28/08/2023	Mohit Kumar Basak	Created a new introduction, added a section for ingestion
0.3	13/09/2023	Peter Baumann	edited
0.4	21/11/2023	Mohit Kumar Basak	Added more EOX datasets; wrote about GUI-based ingestion
1.0	31/11/2023	Stefan Jetschny	Review, editing and clean up

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 2020 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

List of Tables

1 Introduction

The FAIRiCUBE project aims to provide access to gridded data beyond Earth Observation (EO) domains according to the FAIR principle. Central to this endeavour is the role of work package 5, entrusted with the responsibility to make gridded data available for use case owners. This process is referred to as "ingestion". In the FAIRiCUBE project, two key partners, Rasdaman and EOX, collectively contribute their expertise to help with ingestion tasks.

2 Deliverables contributing to M10

There are several formal deliverables contributing to M7 Milestone as listed in Table 1

Table 1 : Deliverables related to M10

Description	Lead Beneficiary	Type	Dissemination level	Due dates
D5.1 List of data cube resources made available (M6, M12)	JUB	R	Public	31.03.2023, 31.09.2023
D5.2 Description of the datacube ingestion pipelines (M9, M18)	JUB	R	Public	28.02.2023, 31.10.2023
D5.3 Validation of ingestion (M12, M20)	4SF	R	SEN	30.06.2023, 29.02.2024

D5.1 describes the data cube resources that are already available on the EOX and the Rasdaman platforms. This deliverable provides an overview of all data that have been ingested using the data ingestion pipeline

D5.2 describes the ingestion pipelines, which define the steps that are carried out to ensure that the dataset requested by the UC owners end up in the EOX or Rasdaman systems in a form that can be used by the Use Cases.

D5.3 describes the validation steps for ingestion.

3 Ingestion and validation

Within FAIRiCUBE vocabulary, "ingestion" refers to the process by which datasets are made available for use by the Use Case (UC) owners. This process requires a "Validation" step which is dealt later in this section. The ingestion process is more complex than just copying the datasets from some source. Among the challenges are:

- Diversity of data source: datasets have different sources, there is no single comprehensive source of information.
- Accessibility: Access methods are highly individual for each source; sometimes these are even assuming human interaction for downloading, which makes it hard to automatically update timeseries.
- Formatting: The datasets have different formats, each coming with its own CRS, resolution, null values, etc.
- Inconsistencies: Datasets are incoherent in themselves; for example, resolution of sometimes changes over time.
- Each dataset is different, and data needs to be made available in a way that it becomes convenient for UC partners to use.

This illustrates the status quo in (Earth) data exploitation in a symptomatic manner. The above challenges are one of the reason why non-experts have no chance to utilize such data and experts need to invest a high overhead before they can perform their analyses. This has led to the quest for Analysis-Ready Data (ARD) where data are homogenized and otherwise made easier to handle. Achieving ARD involves several aspects, a central one being to prepare data upfront in a better way towards ARD. This is the main focus of WP5, contributing to the FAIRiCUBE objective "*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*". Using the ingestion pipeline established, the process becomes transparent and reproducible.

Two different technology stacks are deployed in FAIRiCUBE, with different architecture, functionality, and consequently ingestion workflows, object storage by EOX and datacubes by rasdaman. One of the research goals of FAIRiCUBE is to determine a common ingestion wrapper unifying both approaches to a degree that administrators adding new data always use the same workflow and only select the underlying data store in a final step.

The EOX workflow is more about registration and less about ingestion. The datasets are retrieved by the UCs and uploaded to the object storage provided by the EOX deployment. The uploaded datasets are then registered with the EOX services to become available for python-based analysis tools.
Further description of this pipeline is provided in delivery D5.2.

The Rasdaman ingestion process creates datacubes based on the OGC WCS-T standard which defines request types for inserting, updating, and deleting datacubes. As this is too low level for day-to-day work, an ETL suite has been built around these requests to automate the process as much as possible, relying on user input only where necessary. The datacube configuration is based on "recipes" readily provided for all common situations plus concrete, individual "ingredients" containing all relevant information about the particular dataset to be ingested. Once started, the ingestion process is fully automated for building and updating timeseries, for example.

Along with ingestion of the datasets, augmenting them with the corresponding metadata is important. This is addressed by the FAIRiCUBE catalog, established by WP4. Metadata helps users to identify and access available resources. Both pillars, EOX and rasdaman, are connected to the common FAIRiCUBE catalog which is based on STAC. Datasets get connected to their corresponding catalog record through a bidirectional reference: The catalog references the datasets, and each dataset in turn contains a back reference to its corresponding catalogue entry. To this end, in the ingestion process catalogue and datastores are coupled:

- Use Cases request a dataset by creating an issue in the dedicated GitHub repository.
- For capturing all necessary information, such issues offer a form to be filled in.
- From this formalized input, an ingestion script gets generated.
- The script creates the dataset, the catalogue record pointing to the dataset, and inserts a backlink from the dataset to the catalogue entry.
- After completion, the result can be checked and then used by the UCs.

With ingestion, another process that becomes important is validation – data can be incomplete, inconsistent or faulty, the process can fail (e.g., due to otherwise unnoticed spurious power failures overnight). Establishing a common validation process is described in D5.3.

4 Progress and Status

Capturing information required for ingestion. Early on, the WP5 lead proposed a Datacube Management Plan document which collected all the information necessary for building up datacubes from external sources. Details included geographic extent and coordinate system used, temporal extent and resolution, pixel type, and more relevant information. The intention was that Use Case partners, when requesting ingestion, at the same time provide the necessary detail information based on the idea: if they request the data, they know the data.

This actually proved wrong, despite an introductory webinar on the information contents. For the sake of smooth progress WP5 lead started ingestion based on incomplete data, which however led to hiccups due to missing knowledge. As a remedy the WP5 lead established "data kick-off" sessions where the use case partners were supported in determining parameters. This turned out useful – for example, questions like "what is the validity in time of a dataset published with resolution one year, but with data provided every 3 years, and irregularly" – because different interpretations are possible which lead to different datacubes and subsequently different statistics results. Experience was that such meetings proved rather efficient and avoided that parties are waiting for each other to react. Learnings from these meetings will provide input to establish a formalized process that can capture standard issues and the "data kick-off" sessions can focus on specific issues.

Towards a unified ingestion process. The Datacube Management Plan, in the end was used as a blueprint for setting up a central spreadsheet containing, per data set, the above information plus extra details, such as information for corresponding catalogue entry. This was planned as an interim solution, to be replaced by an approach based on GitHub issues where predefined forms should allow the use case partners to conveniently fill in all items required. From this input, a script can be triggered which creates the corresponding ingest procedure individually per dataset and datacube.

One problem encountered was that more and more metadata items were felt necessary or at least useful so that the spreadsheet is in continuous evolution. To not wait until a possible convergence and finalization of the spreadsheet datacube ingestion actually proceeded to have at least preliminary data available for the use case partners, bearing in mind that at a later stage data may need to be ingested again, once new insight is gained. Therefore, the WP5 lead insists on having such kick-off meetings, despite the extra effort for all partners, and also a final check by the partners to confirm that data are useful as ingested.

Another problem is that the ingestion procedure shall both collect the required information for the technical ingestion phase (making the data available to the user) as well as for the complete meta data description – currently this comprises of 74 data descriptor fields. Discussion is ongoing on the right balance between comprehensiveness and documentation effort. While we are making the unified ingestion process operational and execute several data ingestion requests, we will collect experience on how many fields are mandatory or optional to further streamline this procedure. Aim is to reduce the time effort for users to fill out the data request form while still provide complete set of information for the data and meta data description.

A final solution was now proposed by EOX, which contains of a user-friendly GUI. Use Case partners can use it to fill a form, and it creates a Pull Request(PR) and a STAC file. The PR acts as a central point for further discussion about the datacubes. When the PR is merged, it means that the dataset is already present in the FAIRiCUBE catalog, signifying that the ingestion is complete. This will be completed soon and will become the final way in which the ingestion process is tracked.

New technology requirements. A side effect from these conversations was the spotting of new requirements on the datacube technology. One example is the refinement of temporal queries like "average over every January for the last 10 years" which led to an enhanced calendar handling which currently is being implemented. Another example is the addition of extra information about the pixel type and establishing conventions for feeding it.

Vision. While data access and analytics is finding much attention in standardization, this is much less the case for the ingestion side; the single existing standard in the field, OGC WCS-T, is being used in the project but experience shows that much more support is needed. The ultimate vision and scientific contribution, therefore, is to have unified, easy-to-handle, quality-driven processes for data ingestion.

5 Summary

Ingestion plays a vital role for achieving Analysis-Ready Data (ARD), and hence all involved partners in WP5 devotes particular attention to it, also involving WP4 and its catalogue work. Ingestion is ongoing as further datasets are requested by the UCs. Further, work on the common ingestion process wrapping the different technology pillars is continued further.