Reinforcing the AI4EU Platform by Advancing Earth Observation Intelligence, Innovation and Adoption

D4.1: Integration report on AI4EU tools with DIAS platforms

Grant Agreement ID	101016798	Acronym	AI4COPERNICUS
Project Title	Reinforcing the AI4EU Platform by Advancing Earth Observation Intelligence, Innovation and Adoption		
Start Date	01/01/2021	Duration	36 Months
Project URL	https://ai4copernicus-project.eu/		
Contractual due date	30/6/2022	Actual submission date	30/06/2022
Nature	REPORT	Dissemination Level	PU
Author(s)	Marcin Ziolkowski (CF)		
Contributor(s)	Martin Welß (Fraunhofer IAIS), Antonis Troumpoukis, NCSR-D George Stamoulis, UoA		
Reviewer(s)	Vasileios Baousis (ECMWF)		

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 101016798.

Document Revision History *(including peer reviewing & quality control)*

Version	Date	Changes	Contributor(s)
1	15/03/22	Introduction and initial structuring	Marcin Ziółkowski, CF
2	22/3/22	Section 2	Antonis Troumpoukis, NCSR-D George Stamoulis, UoA
3	2/6/22	Section 3	Iraklis Klampanos, NCSR-D
4	15/6/22	Section 4	Marcin Ziółkowski, CF
5	25/6/22	Document Review	Vasileios Baousis (ECMWF)
6	27/6/22	Section 2 (QA comments resolved)	Antonis Troumpoukis, NCSR-D
7	29/6/22	Finalising the deliverable (pending review by EC)	Marcin Ziółkowski, CF

Executive Summary

This is the first deliverable of WP4 (Implementation, customisation, integration and testing) and, more specifically, Task 4.1 (Integration of AI4EU platform with CREODIAS/WEkEO) . In this deliverable we present the architecture of AI4Copernicus and discuss how the software that we are developing interacts with various components of the AI-on-demand platform, and the DIAS targeted by the project, CREODIAS and WEkEO. This deliverable also specifically reports on the integration progress of the overall AI4Copernicus platform, i.e. between AI4EU and DIAS, as well as with regards to the integration of tools for transformation, querying, interlinking and federating big linked geospatial data.

Table of Contents

List of Figures

List of Terms & Abbreviations

Abbreviation	Definition
ML	Machine Learning
EO	Earth Observation
KG	Knowledge Graph
CM	Collaborative Management
AEM	AI4EU Experiments Management
k8s	Kubernetes container orchestration system
CF	Cloudferro
DIAS	Data and Information access services
PoC	Proof of Concept

1 Introduction

This is the first deliverable of WP4 (Implementation, customisation, integration and testing) and, more specifically, Tasks 4.1 (Integration of AI4EU platform with CREODIAS/WEkEO) . In this deliverable we present the architecture of AI4Copernicus platform and discuss how the software that we are developing interacts with various components of the AI-on-demand platform, and with the DIAS platforms targeted by the project, i.e. CREODIAS and WEkEO.

1.1 Purpose and Scope

The purpose of this deliverable is to report on the integration progress of the overall AI4Copernicus platform, i.e. between AI4EU and DIAS, as well as with regards to the integration of tools for transformation, querying, interlinking and federating big linked geospatial data.

1.2 Approach for Work Package and Relation to other Work Packages and Deliverables

Work package WP4 (Implementation, customisation, integration and testing) started on M4 and ends on M24 of the project. It is led by partner CF with the collaboration of partners CSR-D, UoA, TAS, ECMWF, UNITN . WP4 positions technically AI4Copernicus in the European AI and Copernicus ecosystems. In addition, it implements the software architecture of the project described in WP3.

The technical contribution of this task is the development of the software implementation of the project with a specific emphasis to interfacing with the AI-on-demand platform and CREODIAS and WEkEO.

WP4 has the following five tasks:

- Task 4.1 Integration of AI4EU platform with CREODIAS/WEkEO [M4-M12, lead: CF, contributor: TAS]. Configuration of the environment to accommodate the requirements identified in the WP2.
- Task 4.2 Integration of tools for transformation, querying, interlinking and federating big linked geospatial data [M4- M12, lead: UoA] This task will integrate the linked data suite (developed by UoA) to the platform.
- Task 4.3: Implementation of the semantic catalogue and the semantic search and discovery functionality [M4-M12, lead: UoA, contributor: NCSR-D] This task will implement the semantic catalogue designed in Task 3.2.
- Task 4.4 Machine learning models for EO [M4-M24, lead: UNITN, contributors: NCSR-D, ECMWF] Different supervised machine learning techniques and models will be identified and integrated taking into account the inputs from WP2.
- Task 4.5 Testing and operation of bootstrapping services [M7-M18, lead: CF, contributors: NCSR-D, UoA

The present deliverable D4.1 is the first deliverable of WP4 and contains the contributions of the project in Task 3.1 and Task 3.2 from WP3.

1.3 Organization of the Deliverable

The rest of the deliverable is organized as follows.

- Section 2 presents tools for transformation, querying, interlinking and federating big linked geospatial data.
- Sections 3 and 4 present Integration with the AI4EU/AIonD.
- Section 5 presents an integration roadmap for the DIAS platforms.
- Section 6 presents conclusions.

2 Tools for transformation, querying, interlinking and federating big linked geospatial data

Linked Data lies at the heart of the Semantic Web and allows large scale integration over RDF resources. However, to make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but relationships (i.e. links) among data should be made available, too, to create a Web of Data (as opposed to a sheer collection of datasets).

AI4Copernicus offers a linked data suite as part of Task 4.2 that allows users to handle linked data at every step of the linked data pipeline (cf. Figure 1). Regarding the linked data pipeline, data are *transformed* from their original format into RDF using *GeoTriples*, *interlinked* using *JedAI-spatial*, and *stored* in the geospatial RDF store *Strabon* in order to be *queried*. *Semagrow* can be used to *federate* data that are served by Strabon with other external endpoints that are available in remote servers, and finally, since both Strabon and Semagrow expose GeoSPARQL endpoints, *Sextant* can be used for posing GeoSPARQL queries and *visualising* the result using an intuitive WebUI.

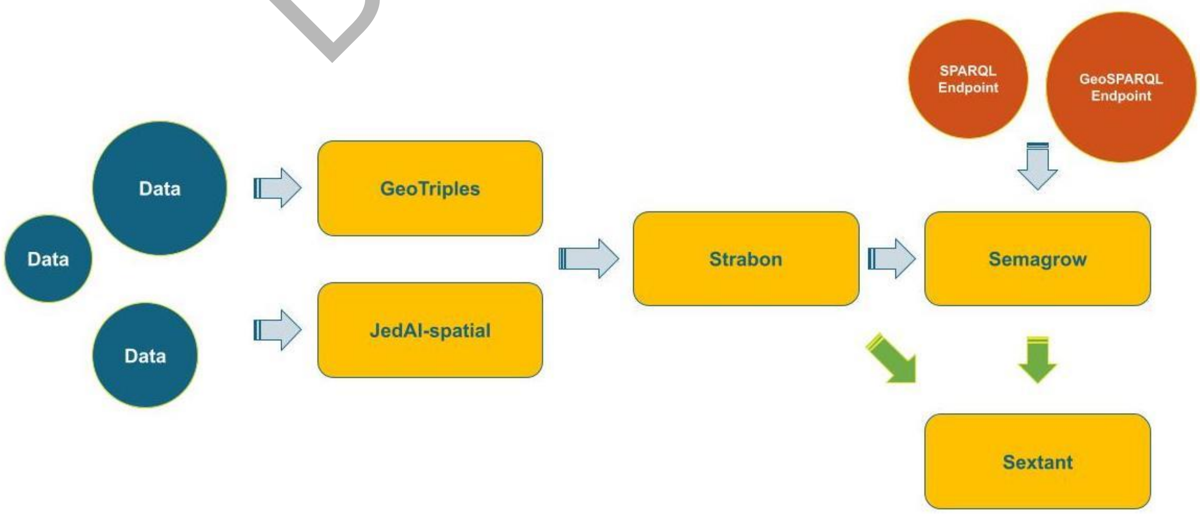


Figure 1. Linked Data pipeline

These tools are available in the AI4EU Research Catalogue and are distributed as open source through the AI-on-Demand platform. UoA and NCSR-Demokritos are responsible for providing support to the users of the open calls should they plan to use any of the proposed tools. In the following sections we present all relevant tools in detail.

The tools are available in the AI-on-Demand platform[1].

2.1 GeoTriples

GeoTriples[2] allows the user to transform geospatial data from their original formats into RDF. The software itself is GDPR compliant. Documents are processed locally and all data remains on the user's local computer. However, the users must ensure that they have the authority to store and process the documents, for example if they contain personal data or other sensitive GDPR-relevant information.

GeoTriples is a semi-automated tool that enables the automatic transformation of geospatial data into RDF graphs using state of the art vocabularies like GeoSPARQL, but at the same time, it is not tightly coupled to a specific vocabulary. The transformation process comprises three steps. First, GeoTriples generates automatically extended R2RML or RML mappings for transforming data that reside in spatially-enabled databases or raw files into RDF. As an optional second step, the user may revise these mappings according to her needs e.g., to utilize a different vocabulary. Finally, GeoTriples processes these mappings and produces an RDF graph. The input formats supported by GeoTriples are spatially-enabled relational databases (PostGIS and MonetDB), ESRI shapefiles, XML documents following a given schema (hence GML documents as well), KML documents, JSON and GeoJSON documents and CSV documents.

GeoTriples is used in the first step of the Linked Data pipeline and assists users in the transformation of data in the RDF format. In the context of ExtremeEarth, we have used the tool to transform the data extracted from the Satellite images in the RDF format using ontologies that we design for each use case.

2.2 Strabon

Strabon[3] is a spatiotemporal RDF store. You can use it to store linked geospatial data that changes over time and pose queries using two popular extensions of SPARQL. Strabon supports spatial datatypes enabling the serialisation of geometric objects in OGC standards WKT and GML. It also offers spatial and temporal selections, spatial and temporal joins, a rich set of spatial functions similar to those offered by geospatial relational database systems and support for multiple Coordinate Reference Systems. Strabon can be used to model temporal domains and concepts such as events, facts that change over time etc. through its support for valid time of triples, and a rich set of temporal functions.

Strabon extends the well-known RDF store Sesame, allowing it to manage both thematic and spatial data expressed in stRDF and stored in the PostGIS spatially enabled DBMS. Strabon implements fully

[1] https://www.ai4europe.eu/ai-community/projects/ai4copernicus?category=ai_assets

[2] https://www.ai4europe.eu/research/ai-catalog/geotriples

[3] https://www.ai4europe.eu/research/ai-catalog/strabon

the Core, Geometry Extension and Geometry Topology Extension components of GeoSPARQL. It supports all three topological relation classes defined by GeoSPARQL (OGC-SFA, Egenhofer, RCC8), both geometry serialisations (WKT, GML) and multiple CRS.

Stravon sits at the core of the Linked Data pipeline and is used to store our data once they are made available in the RDF format.

2.3 JedAI

JedAI[4] comprises a set of domain-independent, state-of-the-art techniques that apply to any domain. At their core lies an approximate, schema-agnostic functionality based on blocking for high scalability. JedAI constitutes an open source, high scalability toolkit that offers out-of-the-box solutions for any data integration task, e.g., Record Linkage, Entity Resolution and Link Discovery. At its core lies a set of domain-independent, state-of-the-art techniques that apply to both RDF and relational data. These techniques rely on an approximate, schema-agnostic functionality based on (meta-)blocking for high scalability.

JedAI-WebApp is a GUI developed with Spring (boot+ MVC) and ReactJS that facilitates the execution of JedAI. It enables the user to construct its desired workflow by sequentially selecting the algorithm(s) of each step. Furthermore, JedAI-WebApp provides the following capabilities:

- Multiple data input interfaces
- Data (entities) Exclusion
- Data Exploration
- Automatic configuration of the algorithms' parameters. User can specify the values of the parameters or he can leave them to the system to detect which parameters produce the best results. The detection of the ideal parameters is performed by Grid Search or by Random Search.
- Detailed Results and display of the logs
- Exploration of the data and results.

Furthermore, it facilitates the benchmarking of different workflows or configurations over a particular dataset through the workbench window, which summarises the outcome of all runs and maintains details about the performance and the configuration of every step. JedAI can be used to discover links between different sources.

JedAI has been extended with new algorithms for interlinking big geospatial data sources. We developed a new module, called JedAI-spatial, which serves both as an open-source library of the state-of-the-art works in the field and as an open-source system that implements new methods that go beyond existing works in terms of efficiency and scalability. JedAI-spatial has the following unique characteristics:

- It organises the main algorithms for Geospatial Interlinking into a novel taxonomy that facilitates their use and adoption by practitioners and researchers.
- Its intuitive user interface supports both novice and expert users: they simply have to select one of the available methods per workflow step and optionally configure it. It also simplifies

[4] https://www.ai4europe.eu/research/ai-catalog/jedai

the benchmarking of the main algorithms in the field through the workbench window that summarises the performance of the algorithms executed so far. This is a crucial task for identifying the best approach for a particular task at hand, given that the experimental analyses in the literature are usually limited with respect to the variety in datasets or the baseline methods.

- Its modular and extensible architecture allows for easily incorporating improvements to all algorithms.
- It optimises the implementation of existing algorithms, some of which have not been applied to Geospatial Interlinking before.
- It conveys new techniques that achieve competitive performance.

JedAI can be used to discover links between different sources.

2.4 Semagrow

Semagrow[5] is a dynamic data integration system that presents multiple (syntactically or semantically) heterogeneous datasets as a unified, homogeneous virtual dataset. Semagrow provides a federated query processor that allows combining, cross-indexing and, in general, making the best out of all public data, regardless of their size, update rate, and schema. In this manner, it offers a single SPARQL endpoint that serves data from remote data sources and that hides from client applications heterogeneity in both form (federating non-SPARQL endpoints) and meaning (transparently mapping queries and query results between vocabularies).

Semagrow has been used to integrate diverse datasets in multiple domains and applications. Among others, meteorological, land-usage, water availability, and crops data for food security; meteorological, GIS, and dispersion modelling data for risk estimation, biology and pharmacology datasets for pharmacological research.

2.5 Sextant

Sextant[6] is a web based and mobile ready platform for visualising, exploring and interacting with linked geospatial data. The core feature of Sextant is the ability to create thematic maps by combining geospatial and temporal information that exists in a number of heterogeneous data sources, ranging from standard SPARQL endpoints, to SPARQL endpoints following the standard GeoSPARQL defined by the Open Geospatial Consortium (OGC), or well-adopted geospatial file formats, like KML, GML and GeoTIFF.

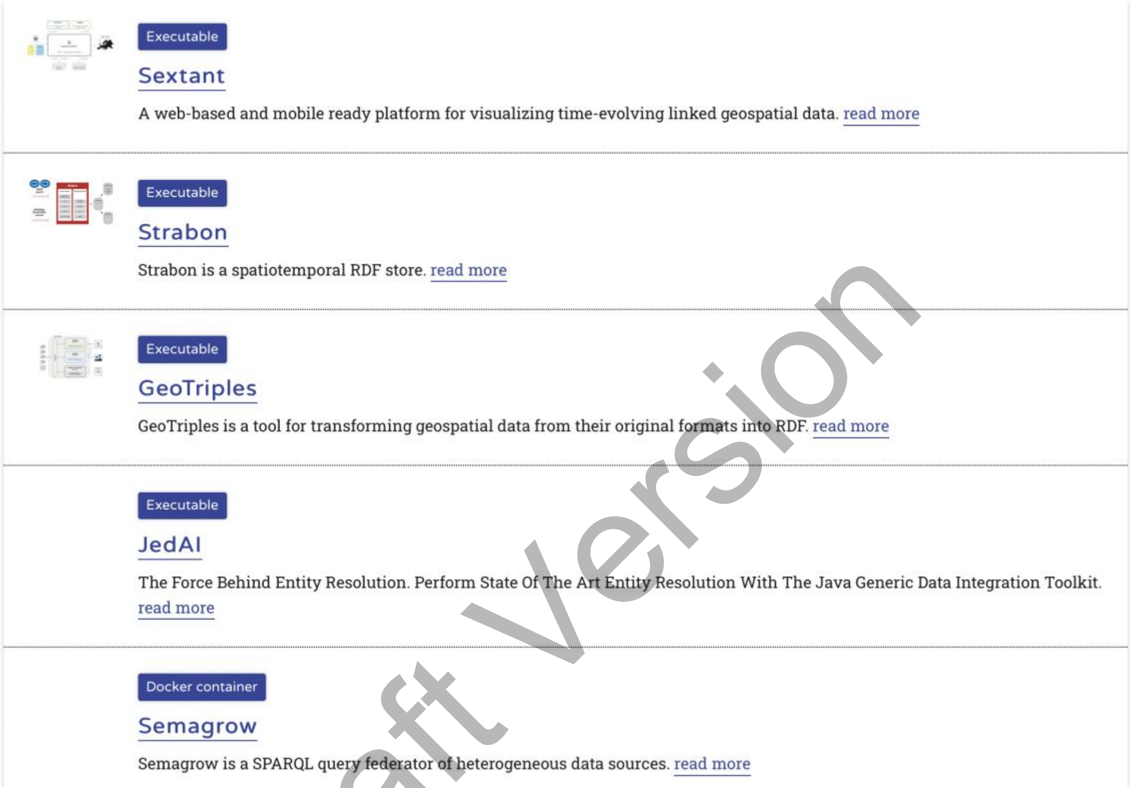
3 Integration with the AI4EU/AIonD catalogue

Integration with the AI4EU/AIonDemand catalogue is one of the integration pathways that AI4Copernicus embraces and utilises. To this end, both internally developed services and tools as well as products developed as part of the cascade funding activities are being catalogued and made available on the AI4EU catalogue.

[5] https://www.ai4europe.eu/research/ai-catalog/semagrow
[6] https://www.ai4europe.eu/research/ai-catalog/sextant

At the time of writing this report, a number of internally developed tools have been catalogued. The catalogue will be further enriched when the awarded projects will start delivering their tools and solutions. As part of the AI4Copernicus Open Calls, awardees are required to create the corresponding entries on the Catalogue as soon as it is practical.

At the time of writing, the AI assets being registered on the Catalogue are: Sextand, Strabon, GeoTriples, GedAI and SemaGrow:



4 Integration with the AIonD Experiments service

Ongoing work and discussions with the AI4EU technical working group, as well as with the ICT-49 technical integration working group, led to the preparation of the following concept of integration. Together, in cooperation with WP3, the foundations for the architecture of the software solution were established within deliverable 3.1. Integration assumes the use of elements such as image respository, dedicated models for AI4EU Experiments and the runtime environment provided by DIAS

WP3 Task 3.1 is the main provider of architecture specification, tools and used components. The technical contribution of this task led to the development of the software architecture of the project with a specific emphasis to interfacing with the AIonD platform, CREODIAS and WEkEO.

The following goals have been described as a preferable user experience:

- User can access EOdata
- User can upload model to the platform to implement he's own solution
- Builded pipeline has elements that are related to EOdata
- User can download created model and use it in dedicated environment
- DIAS provide dedicated environment for the user to run created solution

Activities mentioned above have been presented in the form of Proof of Concept (PoC) during the hackathon presented by the DIAS. What has already been achieved:

- WP3 Deliverables and WP meetings provided concept for solution
- Solution of integration AIonDemand was presented
- Initial Proof of concept presented data download possibility through the DataBroker
- Initial Proof of concept presented data analyse possibility through the DataAnalyzer
- User could use dedicated CREODIAS environment to run AIonD solution
- Dedicated k8s has been setup on CREODIAS
- Possibility to provide results has been show

Conclusions from the conducted research within the presented PoC showed that the solution itself requires more sophisticated customization which has to be developed on DIAS side.

One of the main issues to follow is provision of the dedicated deployment possibilities (a user can downloadthe created model and use it in a dedicated environment). For further development it would be advisable to create a pipeline between AIonDemand and CREADIAS to provide users with some automation deployer. From a user perspective, a dedicated k8s environment for the project purpose could be connected with deployment options located in AIonDemand Experiments.

Some of the possibilities have been presented within the Playground Deployer solution presented by AIonDemand Acumos Team.

According to the PoC, interested parties will be able to use the above AI4Copernicus architecture for small-scale experimentation as well as for demanding applications that require big EO data and high computational power. In the first case, the users will be able to create their desired Acumos pipelines using the AI4EU Experiments component. There, users will have access to the AI4EU resources and other EO data. This EO data can be either some small-scale datasets, accessed externally from CREODIAS via commercial S3 block storage interface or users can upload their own data. This implies that users can experiment while creating their Acumos pipelines. On the other hand, if a user's goal is to create an application using big EO data, it can be achieved by the option to use AI4EU Experiments Playground in order to create a deployable Kubernetes docker, which user can deploy on CREODIAS and use internal S3 API to access the data close to the source (Input data should be accessed through data brokers in the Acumos pipeline. So if the data are already in a DIAS, it would be best for the deployment to take place in the same DIAS, to avoid moving the data. Other than that, Acumos pipelines could be deployed in any Kubernetes installation.). In this way, the users will be able to take advantage of the big EO data provided by CREODIAS, without the need to download it locally.

A number of steps need to be taken in order to access the experiments ecosystem. These include user registration, preparation of execution environment and composition of the desired pipeline. Such a customer journey is provided below, with each step documented in accordance with WP3 Deliverable D3.1.

1. A customer registers in CREODIAS and sets up a Kubernetes cluster in order to prepare an execution environment for experiments.
2. Additionally the user registers in the AI4EU experiments platform (Acumos) in order to access solutions and collaborate with other scientists.
3. One can compose a solution either from existing components or create and share new models using Acumos onboarding and publication processes. It is also possible to use existing solutions published by members of the AI4EU community (this step is optional). In case of new components, docker images need to be stored in a preferred docker image repository. Address of the repository and image identifiers are pointed during component onboarding process.
4. Once the solution is chosen it can be downloaded from Acumos. Solution file contains a complete set of scripts allowing AI4EU users running them in a chosen execution environment.
5. Solution scripts deploy the pipeline to the Kubernetes cluster created in step 2.
6. Solutions run in CREODIAS Kubernetes cluster may access EO data. It is however also possible to execute other types of solutions if needed.

Initial Guidelines for this activity have been provided in Technical Documentation[7]. It describes the necessary steps needed to be implemented for AI4Copernicus users with access to DIAS platforms and AI4EU resources in a streamlined way (with support along the way).

Technical Documentation procedure is based on PoC where a user wants to use a self created solution using Data Broker and Data Analytics mechanism/tools.

At the beginning the user needs to prepare a data broker to find relevant data products, and an analyzer to read data from product files. Data Broker is a tool developed by CF and it supports downloading data collection (based on URL address taken from CREODIAS Finder). The user then needs to publish the relevant docker images in the image registry. Data Broker, itself needs a lot more customization. For example: it may need to be rebuild to make use of parameters tool. After putting tool on the docker registry it may be used for creating the pipeline/solution. At this point there are few things that User needs to fulfil. Configuration of kubectl to point execution environment in CREODIAS is one of them. Deployment and execution of the solution pipeline for particular use-case.

7 https://ai4copernicus-project.eu/wp-content/uploads/2022/03/Technical-Documentation_March2022.pdf

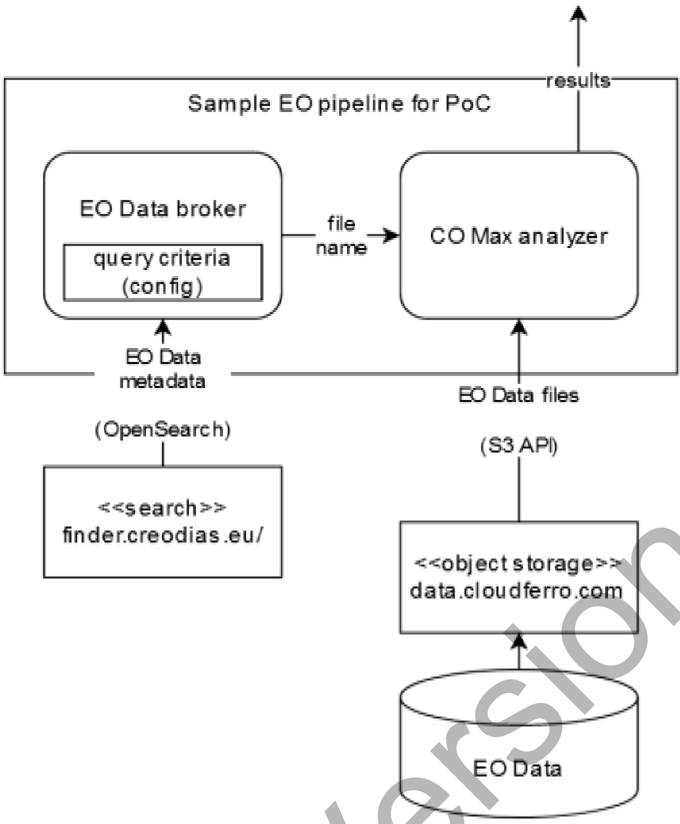


Figure 3. Data Broker and Data Analyzer UML pipeline

The initial guidelines have already been provided in the Technical Documentation[8].

The architecture of AI4Copernicus related to CREODIAS and its integration with the AI-on-demand Platform. This part of the AI4Copernicus architecture is shown in Figure 4.

8 https://ai4copernicus-project.eu/wp-content/uploads/2022/03/Technical-Documentation_March2022.pdf

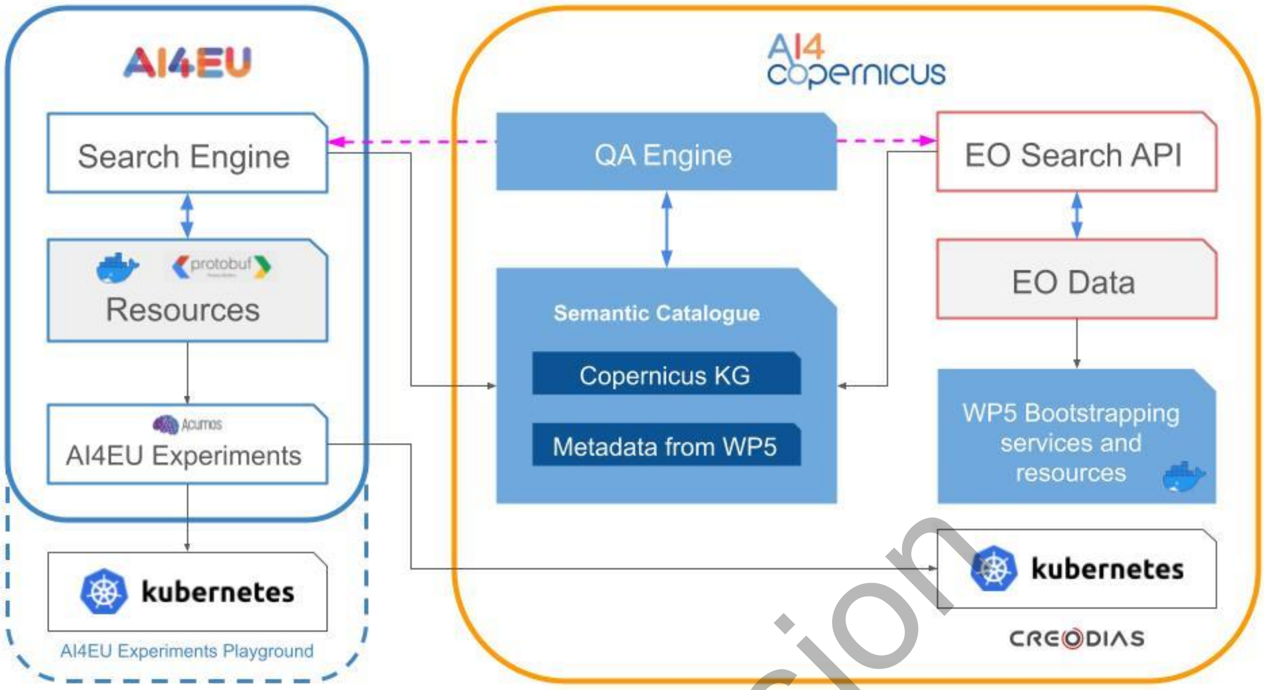


Figure 4: AI4Copernicus architecture

On the left side of the figure above we can see the AI-on-demand platform. The platform has a search engine, which can be used to find resources that are available in the AI-on-demand platform. For example, a resource can be a dataset, a model or a software tool (e.g., the implementation of a machine learning (ML) algorithm). Once the desired resources are found using the AI-on-demand platform search engine, a user can use Acumos to build a pipeline using these resources. In a future version of the AI-on-demand platform, the AI4EU Experiments Playground will also be provided, where users will be able to deploy Kubernetes and run their pipelines in order to develop their applications. The AI Playground will be based on standard servers without GPU acceleration to allow users the run of small pipelines that can be deployed by one click from AI4EU Experiments.

• Further integration ideas

The biggest challenge for further integration would be to propose a viable implementation schedule.

Topics needed to be discussed would include:

- Access to EO data outside CREODIAS vs CREODIAS as a playground environment for AI4EU?
- Scaling options (currently manual adjustments in deployment configs which are part of solution.zip)
- Serial orchestrator & other options
- Where to publish/store experiment results?
- Development process for AI4EU components accessing EO data

The architecture of AI4EU Experiments is designed to support the integration of many different deployment services for execution environments. Each deployment service is a separate microservice running inside the AI4EU Experiments cluster transforming the pipeline definition into deployment artifacts. There is for example the standard deployer for local Kubernetes (called internally kubernetes-client). It generates several kubernetes service.yaml and deployments.yaml files from the pipeline definition and creates a zip file for download containing them and adding a python script to finish the deployment on the target kubernetes cluster. If the deployment is not a single model but a pipeline, it automatically adds the orchestrator to the deployments and handles the shared folder node.

In contrast to the kubernetes deployer, the playground deployer does not offer the solution.zip for download, but instead sends it to the playground server which handles then the further steps of local deployment on the playground kubernetes cluster.

A deployment service must take the following cases into account:

- the solution can be just a single model (= only one node)
- the solution can be a pipeline
- the pipeline can contain a shared folder node

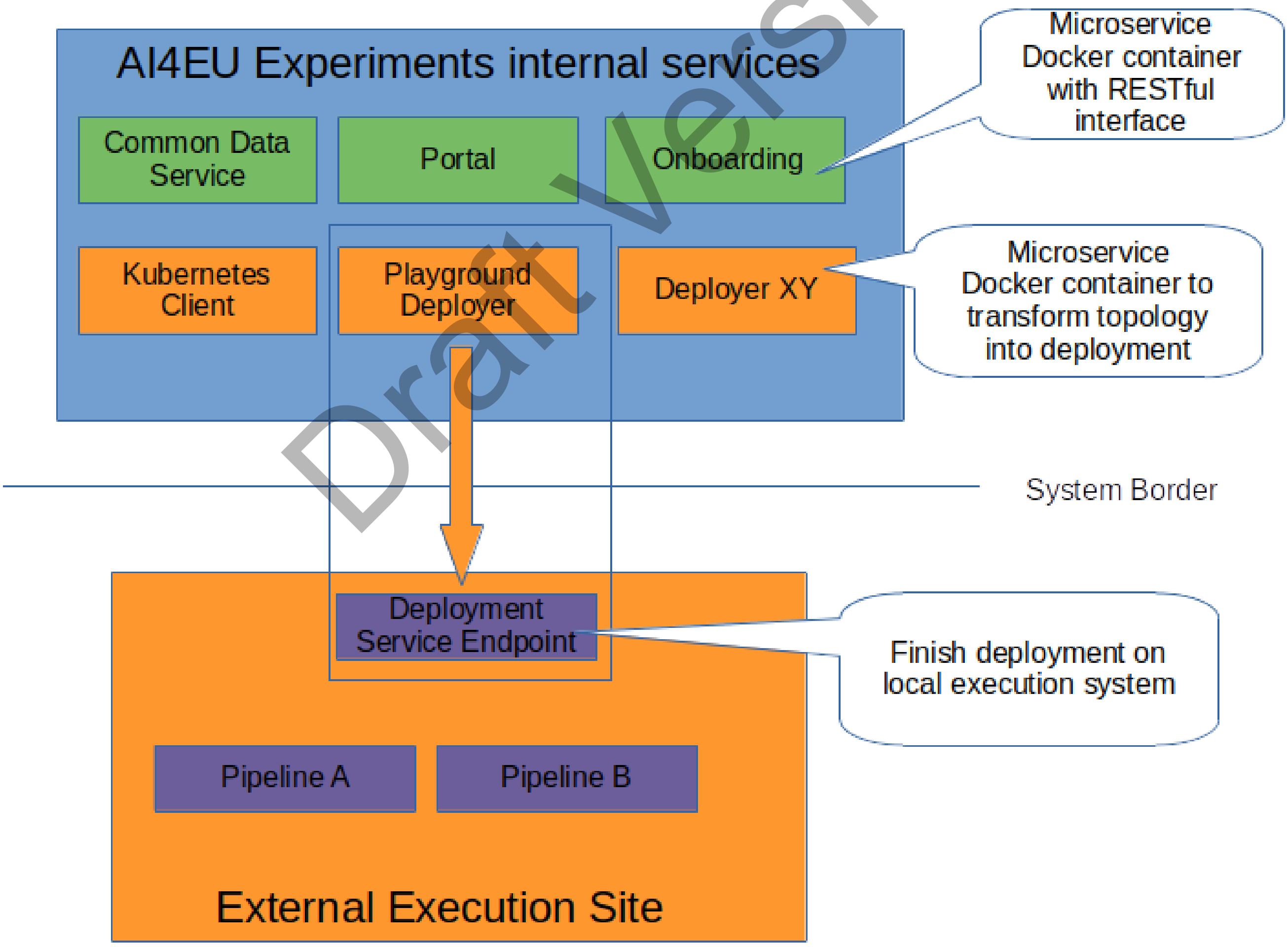


Figure 5. Playground deployer architecture

A deployment service can be composed of two parts: the first running internally in the AI4EU Experiments cluster and the second part runs on the target execution system to finish the deployment and start the actual execution.

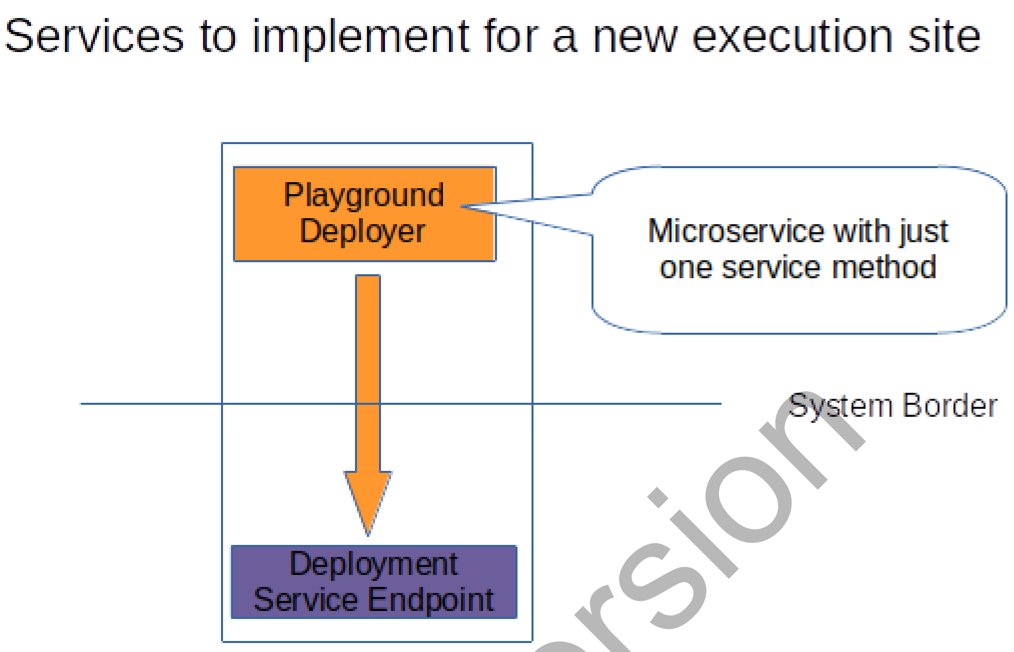


Figure 5. Implementation for Playground Deployer

That kind of integration would benefit the platform users experince an accesibility.

Few steps are needed to be performed for any platform to participate in the integration:

1. One of the fundamentals to implement all the solutions for the goal of automation is to set up proper infrastructure on the provider side. Acumos platform itself will have to access (through) the deploer k8s environment
2. AI4EU Acumos Credentials are needed to be somehow expanded to give Acumos user possibility to access k8s cluster or dedicated namespace of the client
3. The last but not least is to rewrite deploy solution dedicated to DIAS, according to the instructions provided in Acumos Team Github[9]

5 Integration roadmap for the DIAS platforms

The proposed integration between AI4EU and DIAS is presented in the section above. It contains information:

- What has been achieved so far

[9] https://github.com/ai4eu/playground-deployer

- What are the plans for the future in terms of further development of the system

This section ("Integration Roadmap for DIAS platforms") is intended to indicate the potential directions of development and integration for DIAS (Data and Information access services) platforms.DIAS are cloud-based platforms developed to facilitate and standardise the access to Copernicus data and information.

Depending on their internal architecture, the way of integration may differ significantly. CREODIAS gives users access to quite broad permissions in terms of access to the environment. The user can access the cloud space where he can develop his own applications, e.g. using EOData data. The project consortium involves technically experienced users who deliver in-house services and models for inclusion in the AI4EU ecosystem, and who also mentor successful external projects.

AI4Copernicus already has significant feedback from AI4EU. A number of meetings where all project stakeholders could present their needs have been performed. This resulted in quite specific implementation requirements. Meanwhile, a lot of information could be found on the deliverables inputs. Based on the information provided during the meetings it was important to understand the user journey, expectations and compare them to the technical possibilities. This comparison could affect the whole effectiveness of the solution implementation.

The identified main conditions for correct integration are:

- Enabling connection between AI4EU and EOdata resources
- Collaboration in the provision of cloud infrastructure for the project

The next milestones for CREODIAS are:

- PoC presentation for the correct operation of the DataBroker module needed for the connection to EOdata (M20 of the project),
- PoC for the consumption of cloud resources for the project (M20 of the project)
- Docker repository for the Bootstrapping services(PoC M18 of the project; M20 of the project)
- Providing a cloud environment with a dedicated Kubernetes to support computational models (M12 of the project)

These PoCs will drive the evolution of DIASs from data and information access services, to knowledge producing services. Based on experience CREODIAS could then provide a clear pathway to leverage Copernicus data and services in AI workflows using the AI4EU platform for other DIASes. At the time of writing, the exact scheduling and extent of this integration is being worked out, also based on prior AI4EU experience and documentation.

6 Conclusions

In this deliverable, we presented the activities we carried out in M1-M18 of AI4Copernicus in order to provide an integration on AI4EU tools with DIAS platforms. We have tested the most important elements of the infrastructure.

Integrating AI4Copernicus technologies and DIAS platforms with core AI4EU structures and procedures is a core objective of this project. We have achieved to:

(1) Improve, adapt and integrate a number of software tools for transforming, interlinking and federating big linked geospatial data. A number of tools have been documented and made available on the AI4EU catalogue.
(2) Provide basic integration between the AI4Experiments (Acumos-based) and CREODIAS platforms, albeit involving technically involved manual steps.
(3) Create a roadmap and maintain an ongoing discussion which will lead to a better integration to benefit multiple projects and initiatives.