EGI Conference 2024 – 02/10/2024

AI4 | eosc

# AI Inference Pipeline Composition with AI4Compose and OSCAR

Amanda Calatrava, Diego Aguirre, Vicente Rodríguez, Caterina Alarcón, Miguel Caballer and Germán Moltó

Universitat Politècnica de València (UPV)

Co-funded by the European Union

AI4 eosc
Co-funded by the European Union

# AI4EOSC Overview

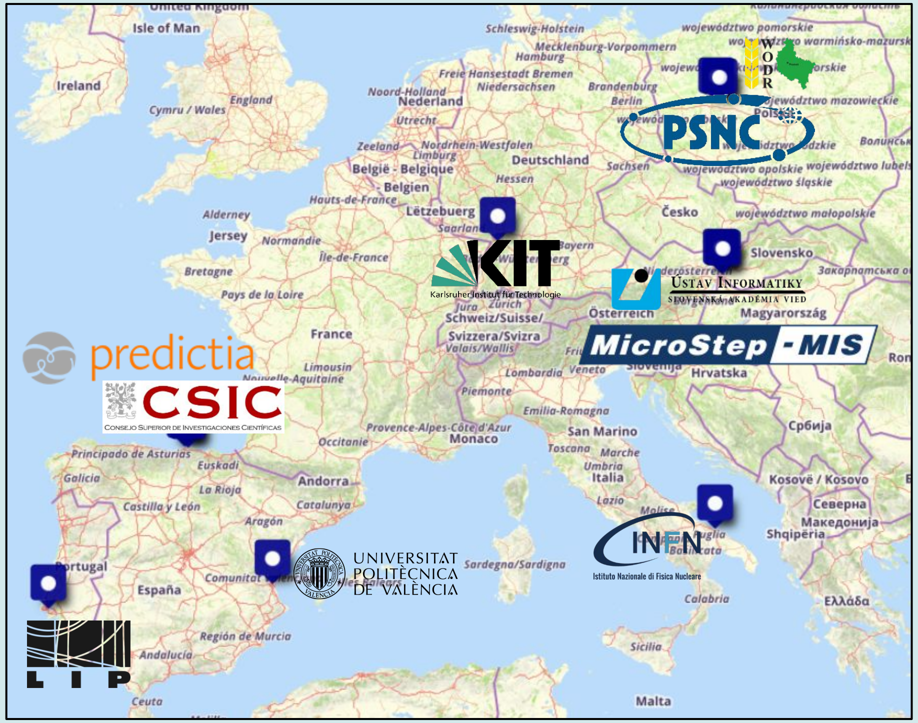
Artificial Intelligence for the #EOSC

- Evolution of  the DEEP Hybrid DataCloud platform.
- HORIZON-INFRA-2021-EOSC-01-04 call.
- Runs September 1st 2022 – August 2025 (36 months).
- 7 academic + 2 SME + 1 non-profit organization.

Advanced features for distributed, federated, composite learning, metadata provenance, MLOps, event-driven data processing, and provision of AI/ML/DL services.

**Objective 3**
Services to compose AI tools, enabling the development of complex data-driven applications

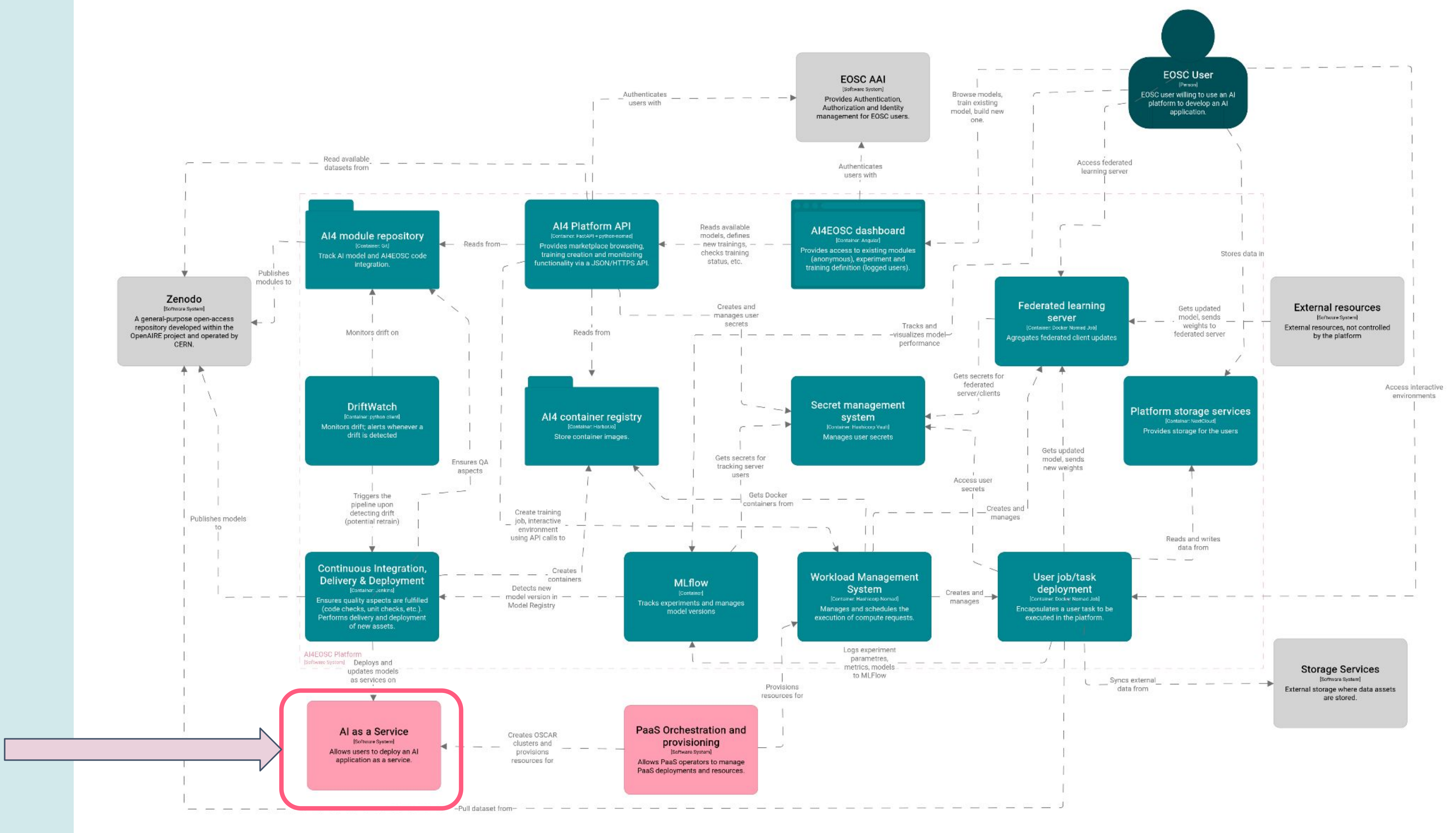
Graphical Composition of **AI Inference Pipelines**



# AI4EOSC Platform architecture

AI4 | eosc

Co-funded by the European Union



**EOSC User**
[Person]
EOSC user willing to use an AI platform to develop an AI application.

**EOSC AAI**
[Software System]
Provides Authentication, Authorization and Identity management for EOSC users.

**Zenodo**
[Software System]
A general-purpose open-access repository developed within the OpenAIRE project and operated by CERN.

**AI4 module repository**
[Container: Git]
Track AI model and AI4EOSC code integration.

**AI4 Platform API**
[Container: FastAPI + python-nomad]
Provides marketplace browseing, training creation and monitoring functionality via a JSON/HTTPS API.

**AI4EOSC dashboard**
[Container: Angular]
Provides access to existing modules (anonymous), experiment and training definition (logged users).

**Federated learning server**
[Container: Docker Nomad Job]
Agregates federated client updates

**External resources**
[Software System]
External resources, not controlled by the platform

**DriftWatch**
[Container: python client]
Monitors drift; alerts whenever a drift is detected

**AI4 container registry**
[Container: Harbor.io]
Store container images.

**Secret management system**
[Container: Hashicorp Vault]
Manages user secrets

**Platform storage services**
[Container: NextCloud]
Provides storage for the users

**Continuous Integration, Delivery & Deployment**
[Container: Jenkins]
Ensures quality aspects are fulfilled (code checks, unit checks, etc.). Performs delivery and deployment of new assets.

**MLflow**
[Container]
Tracks experiments and manages model versions

**Workload Management System**
[Container: Hashicorp Nomad]
Manages and schedules the execution of compute requests.

**User job/task deployment**
[Container: Docker Nomad Job]
Encapsulates a user task to be executed in the platform.

AI4EOSC Platform
[Software System]

**AI as a Service**
[Software System]
Allows users to deploy an AI application as a service.

**PaaS Orchestration and provisioning**
[Software System]
Allows PaaS operators to manage PaaS deployments and resources.

**Storage Services**
[Software System]
External storage where data assets are stored.

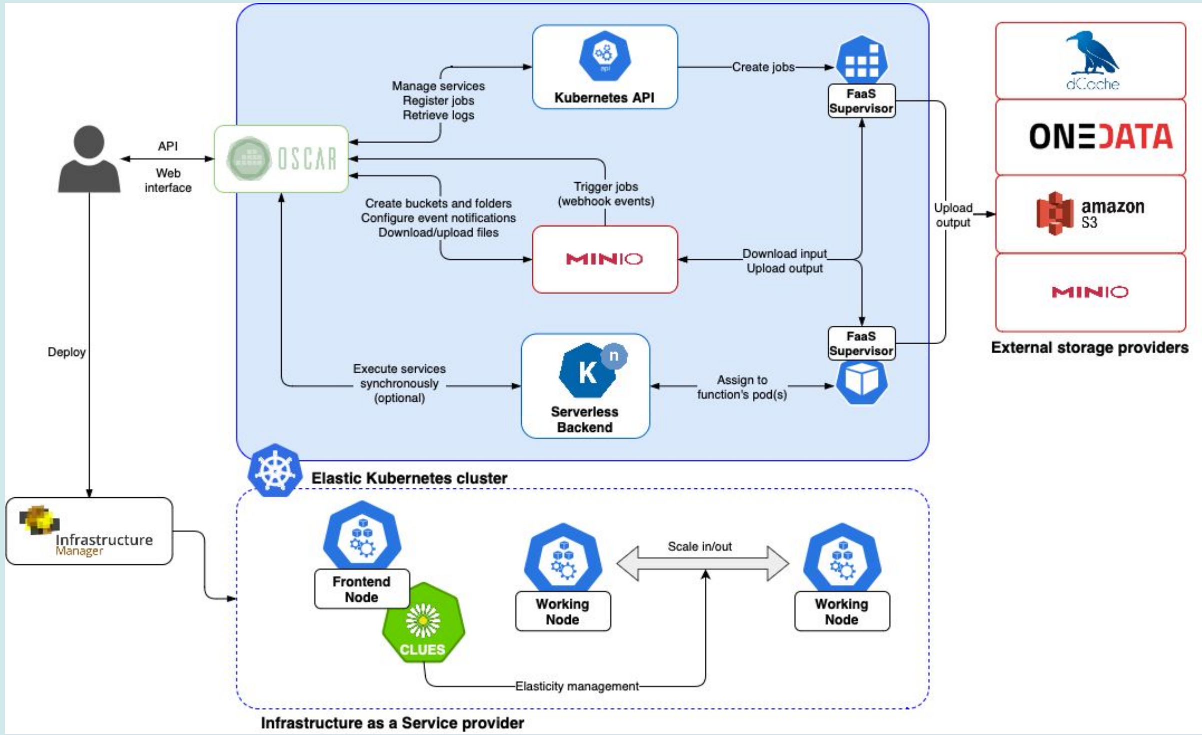
Authenticates users with
Read available datasets from
Authenticates users with
Browse models, train existing model, build new one.
Access federated learning server
Stores data in
Publishes modules to
Reads from
Reads available models, defines new trainings, checks training status, etc.
Creates and manages user secrets
Tracks and visualizes model performance
Gets updated model, sends weights to federated server
Monitors drift on
Reads from
Gets secrets for federated server/clients
Access interactive environments
Ensures QA aspects
Gets secrets for tracking server users
Gets updated model, sends new weights
Access user secrets
Triggers the pipeline upon detecting drift (potential retrain)
Gets Docker containers from
Publishes models to
Create training job, interactive environment using API calls to
Creates and manages
Reads and writes data from
Creates containers
Detects new model version in Model Registry
Creates and manages
Deploys and updates models as services on
Logs experiment parametres, metrics, models to MLFlow
Provisions resources for
Syncs external data from
Creates OSCAR clusters and provisions resources for
Pull dataset from

Interactive C4 model: https://structurizr.com/share/73873/2f769b91-f208-41b0-b79f-5e196435bdb1/diagrams#ai4eosc_container_view

# What is OSCAR?

OSCAR is…

- Open-source: https://github.com/grycap/oscar
- Serverless (event-driven + automated elasticity) computing model for data-processing Docker-based applications on Kubernetes.
  - Users do not need to manage the capacity provisioning of the cluster.
- Easily deployed on multi-Clouds via the Infrastructure Manager (IM).
- OSCAR allows to execute AI models packaged as Docker images.
- Users can upload files to an OSCAR cluster to automatically trigger the inference.
- OSCAR also allows programmatic access to trigger the inference of AI models.
- OSCAR can help you to expose your trained AI model to the community.

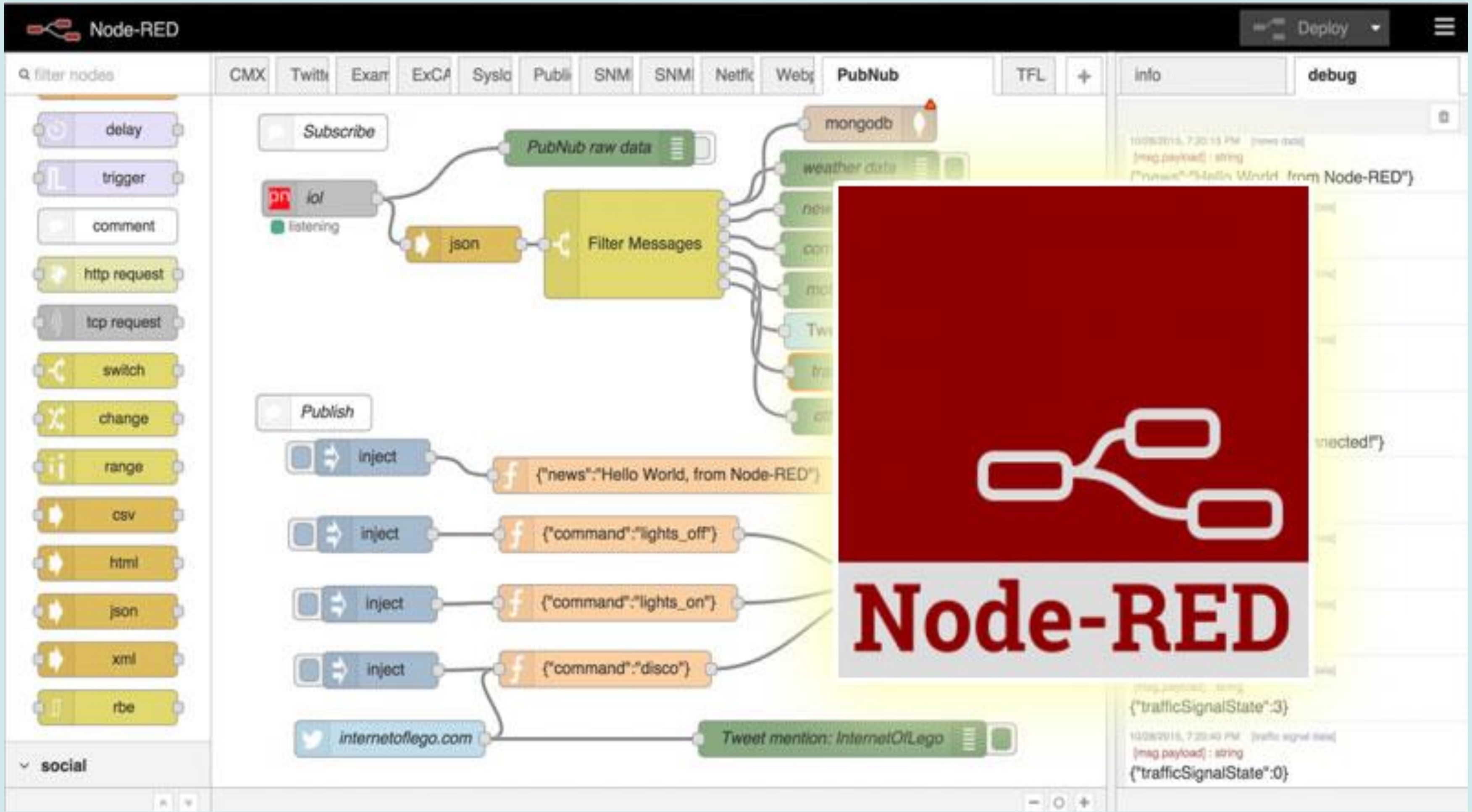


https://oscar.grycap.net

AI4 eosc
Co-funded by the European Union

# What is Node-RED?

- Node-RED is a flow-based programming tool for connecting hardware devices, APIs, and services.
- Built on NodeJS and the D3.js library.
- The minimal structure are the nodes.
- Nodes are organized in flows that connect nodes.



{"news":"Hello World, from Node-RED"}
{"trafficSignalState":3}
{"trafficSignalState":0}

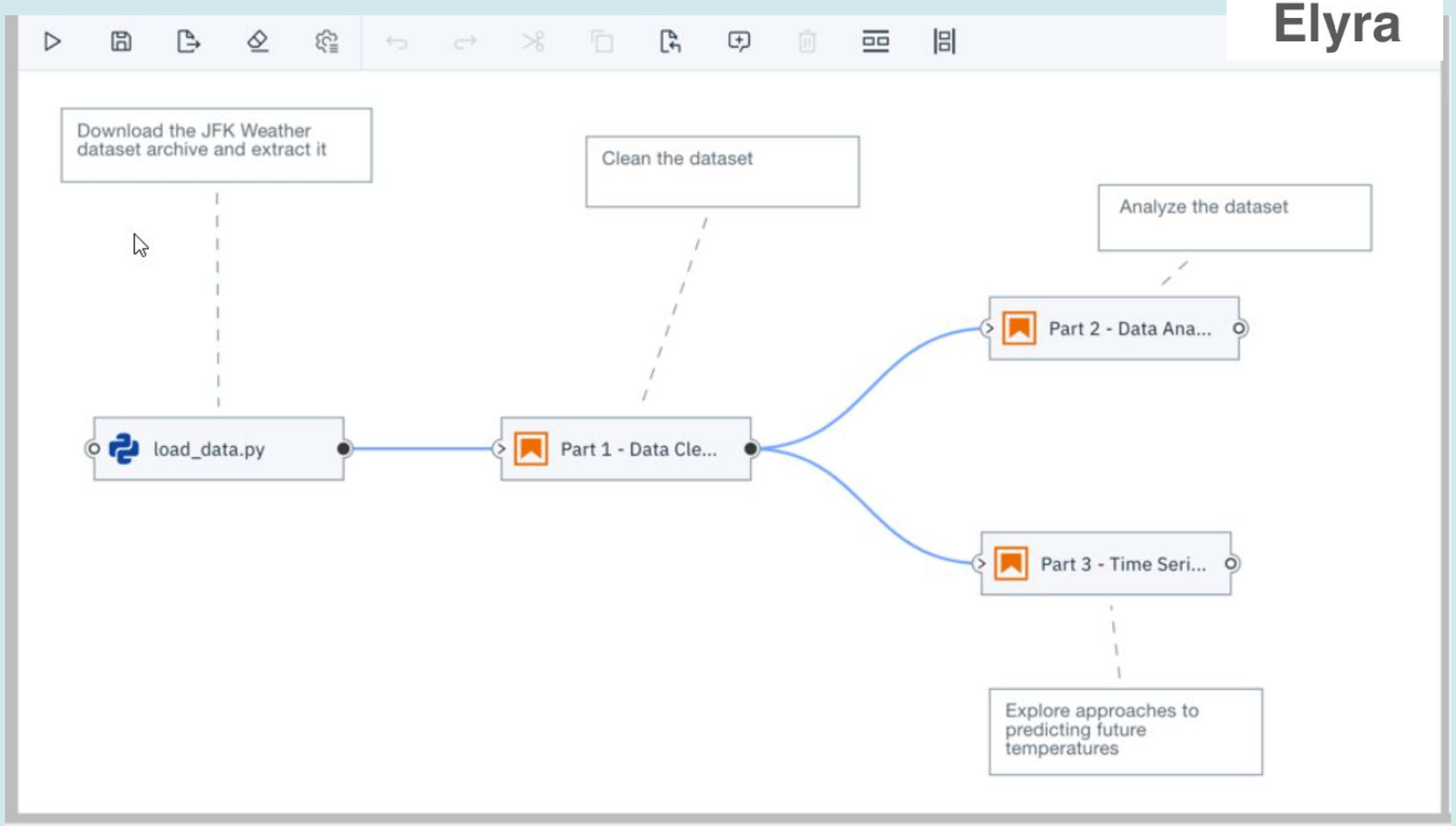
https://nodered.org

# What is Elyra?

AI4 | eosc Co-funded by the European Union

- Elyra is an open-source project for developing and running machine learning workflows in JupyterLab.
- Provides tools for users to create visual pipelines for machine learning workflows.
- Elyra is programmed in Python and allows the use of many libraries focused on data analysis.
- The minimum structure for a workflow are the nodes (Python script, R script, and Jupyter notebook).

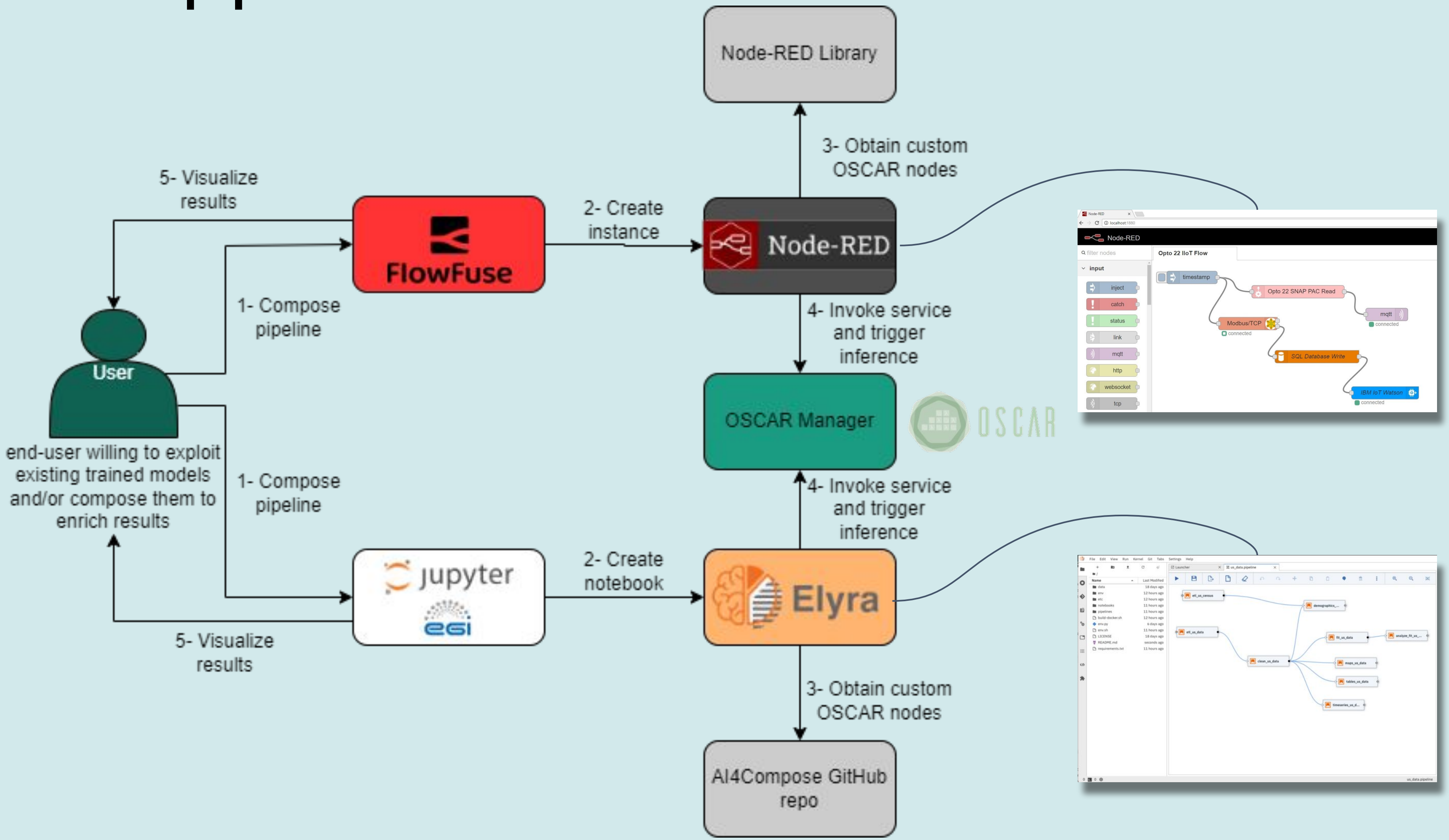
Elyra



https://elyra.readthedocs.io/en/stable/

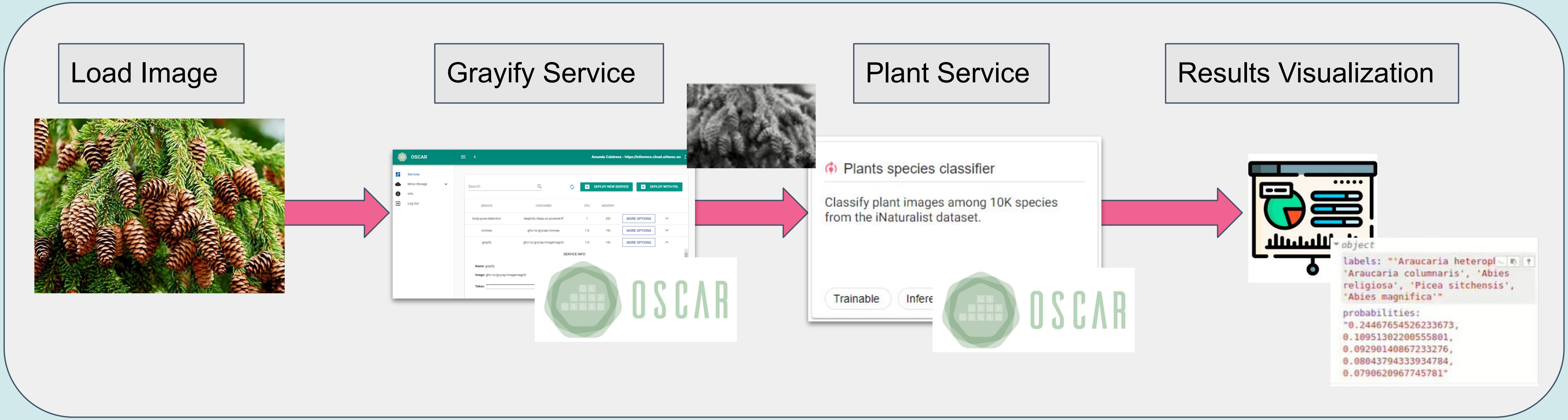
Co-funded by the European Union

# **AI4Compose:** Low-code composition of AI inference pipelines.



https://github.com/ai4os/ai4-compose/tree/main

# Composing AI pipelines: an example



OSCAR — Amanda Calatrava - https://inference.cloud.ai4eosc.eu

| SERVICE | CONTAINER | CPU | MEMORY | |
|---|---|---|---|---|
| body-pose-detection | deephdc/deep-oc-posenet-tf | 1 | 2Gi | MORE OPTIONS |
| cowsay | ghcr.io/grycap/cowsay | 1.0 | 1Gi | MORE OPTIONS |
| grayify | ghcr.io/grycap/imagemagick | 1.0 | 1Gi | MORE OPTIONS |

SERVICE INFO

**Name:** grayify

**Image:** ghcr.io/grycap/imagemagick

**Token:** ••••••••••••••••••••••••••••••••••••

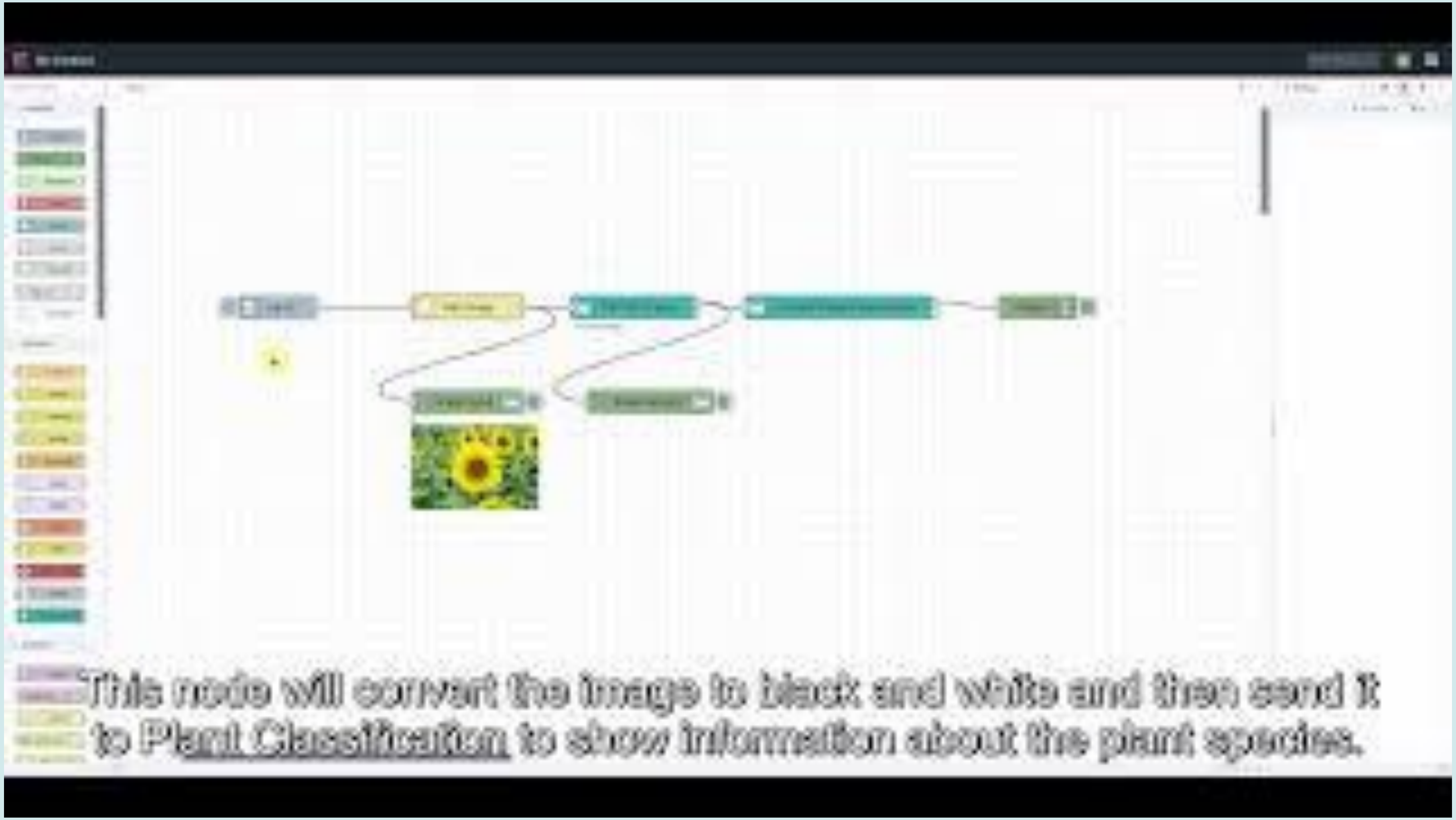
Plants species classifier

Classify plant images among 10K species from the iNaturalist dataset.

```
object
labels: "'Araucaria heteropl
'Araucaria columnaris', 'Abies
religiosa', 'Picea sitchensis',
'Abies magnifica'"
probabilities:
"0.24467654526233673,
0.10951302200555801,
0.09290140867233276,
0.08043794333934784,
0.0790620967745781"
```

# Composing AI pipelines: Node-RED demo

Node-RED



This node will convert the image to black and white and then send it to Plant Classification to show information about the plant species.

https://www.youtube.com/watch?v=9a019SA5GW4&t=1s

# Composing AI pipelines: Elyra demo

## Notebooks

Notebooks is an environment based on Jupyter and the EGI cloud service that offers a browser-based, scalable tool for interactive data analysis. The Notebooks environment provides users with notebooks where they can combine text, mathematics, computations and rich media output.

Access requires a valid EGI account and enrolling to one of the supported VOs.

Default environment provides 2 CPU cores, 4 GB RAM and 20GB of personal storage space per user.

User communities/advanced users can have their customised EGI Notebooks service instance. EGI offers consultancy and support, as well as can operate the setup. Order a community notebooks instance via the EOSC Marketplace.

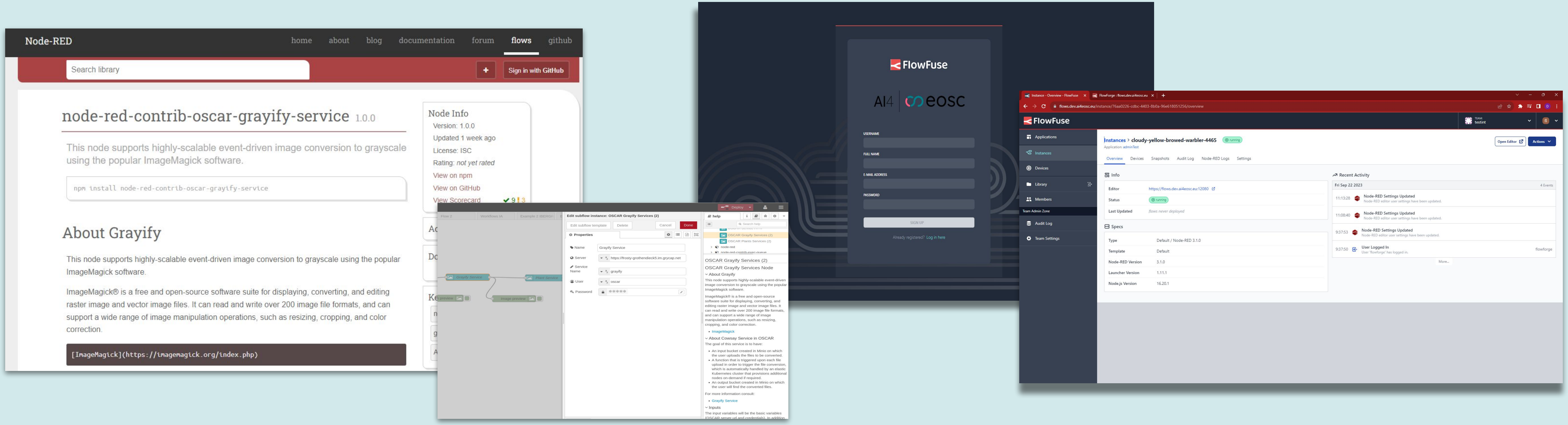
https://www.youtube.com/watch?v=1pFjs0LND4E

# **AI4Compose:** Innovation in AI4EOSC

AI4 | eosc

Co-funded by the European Union

- New modules to interact with OSCAR, published in the official Node-RED library (*more coming soon*):
  - https://flows.nodered.org/search?term=AI4EOSC
- Flowfuse: to manage users and multiple instances of Node-RED.
  - Dedicated instance for the project.
  - OSCAR templates ready to offer a pre-configured Node-RED instance with OSCAR support.



node-red-contrib-oscar-grayify-service 1.0.0

This node supports highly-scalable event-driven image conversion to grayscale using the popular ImageMagick software.

```
npm install node-red-contrib-oscar-grayify-service
```

Node Info

Version: 1.0.0
Updated 1 week ago
License: ISC
Rating: *not yet rated*
View on npm
View on GitHub
View Scorecard ✓ 9 ! 3

About Grayify

This node supports highly-scalable event-driven image conversion to grayscale using the popular ImageMagick software.

ImageMagick® is a free and open-source software suite for displaying, converting, and editing raster image and vector image files. It can read and write over 200 image file formats, and can support a wide range of image manipulation operations, such as resizing, cropping, and color correction.

```
[ImageMagick](https://imagemagick.org/index.php)
```

Edit subflow instance: OSCAR Grayify Services (2)

Edit subflow template | Delete | Cancel | Done

Properties

Name: Grayify Service
Server: https://frosty-grothendieck5.im.grycap.net
Service Name: grayify
User: oscar
Password: *****

OSCAR Grayify Services (2)
OSCAR Plants Services (2)
node-red
node-red-contrib-exec-queue

OSCAR Grayify Services (2)

OSCAR Grayify Services Node

About Grayify

This node supports highly-scalable event-driven image conversion to grayscale using the popular ImageMagick software.

ImageMagick® is a free and open-source software suite for displaying, converting, and editing raster image and vector image files. It can read and write over 200 image file formats, and can support a wide range of image manipulation operations, such as resizing, cropping, and color correction.

- ImageMagick

About Cowsay Service in OSCAR

The goal of this service is to have:

- An input bucket created in Minio on which the user uploads the files to be converted.
- A function that is triggered upon each file upload in order to trigger the file conversion, which is automatically handled by an elastic Kubernetes cluster that provisions additional nodes on-demand if required.
- An output bucket created in Minio on which the user will find the converted files.

For more information consult:

- Grayify Service

Inputs

The input variables will be the basic variables (OSCAR server url and credentials). In addition

Instances > cloudy-yellow-browed-warbler-4465 running
Application: adminTest

Info

Editor: https://flows.dev.ai4eosc.eu:12080
Status: running
Last Updated: *flows never deployed*

Specs

Type: Default / Node-RED 3.1.0
Template: Default
Node-RED Version: 3.1.0
Launcher Version: 1.11.1
Node.js Version: 16.20.1

Recent Activity

Fri Sep 22 2023 — 4 Events

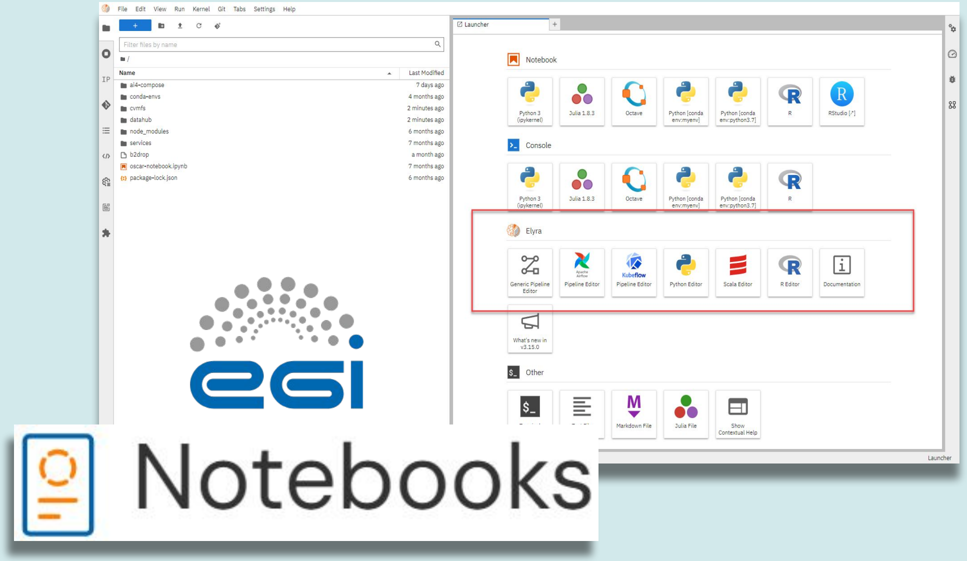
11:13:28 Node-RED Settings Updated — Node-RED editor user settings have been updated.
11:08:40 Node-RED Settings Updated — Node-RED editor user settings have been updated.
9:37:53 Node-RED Settings Updated — Node-RED editor user settings have been updated.
9:37:50 User Logged In — User 'flowforge' has logged in. flowforge

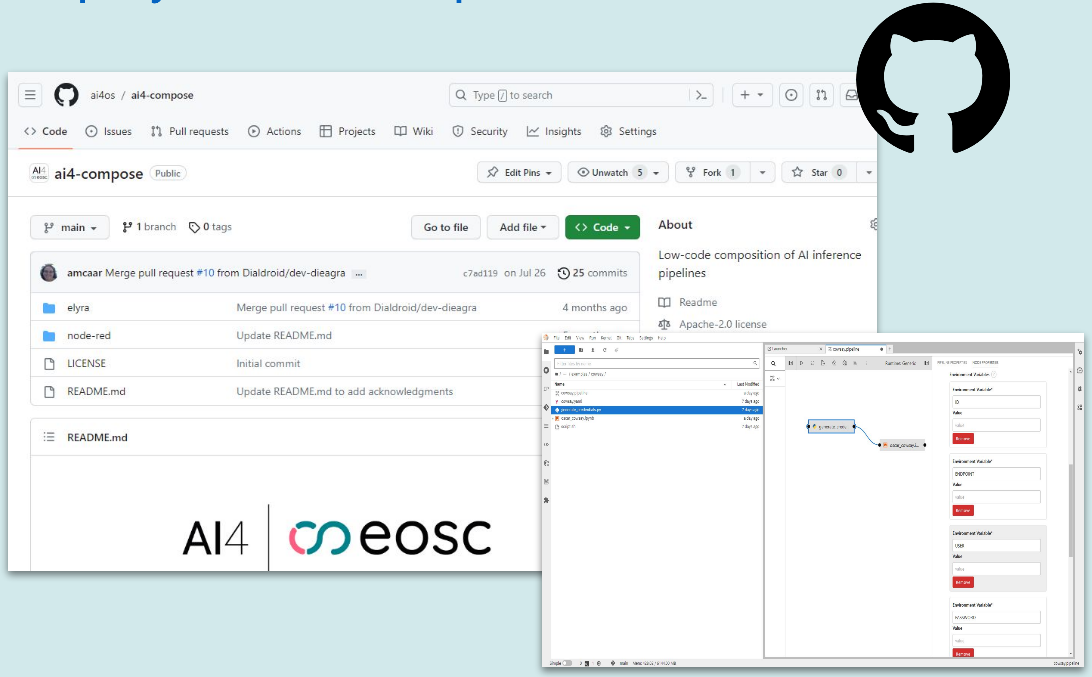
More...

# **AI4Compose:** Innovation in AI4EOSC

AI4 | eosc
Co-funded by the European Union

- Elyra in EGI Notebooks: we have contributed to have official support of Elyra inside the service, so users can easily access the tool.
- Notebooks to be used in Elyra for AI models of the DEEP Open Catalog:
  - https://github.com/ai4os/ai4-compose/tree/main/elyra/examples
- Documentation in the official AI4Docs repo: https://docs.ai4os.eu/en/latest/user/index.html#deploy-a-model-in-production





Low-code composition of AI inference pipelines

amcaar Merge pull request #10 from Dialdroid/dev-dieagra … c7ad119 on Jul 26 25 commits

| | | |
|---|---|---|
| elyra | Merge pull request #10 from Dialdroid/dev-dieagra | 4 months ago |
| node-red | Update README.md | |
| LICENSE | Initial commit | |
| README.md | Update README.md to add acknowledgments | |

# Conclusions

Co-funded by the European Union

- AI4Compose offers a visual support (drag & drop + customization) to compose AI inference pipelines.
- It minimizes the orchestration effort:
  - Multiple AI models can be triggered for inference and later aggregate the results for enhance accuracy.
- Reusable functions:
  - Pre-defined workflows can be created to facilitate interaction among the AI models in the DEEP Open Catalog.
  - Specific nodes can be created for the different AI models for a simpler definition of workflows.
- Workflows along the computing continuum can be supported (e.g. OSCAR clusters in disparate computing infrastructures).
- Each node can be configured to invoke an OSCAR service within a specific OSCAR clusters.
- Deployment of pre-defined Node-Red instances to facilitate AI composition of workflows (multi-tenant support thanks to FlowFuse).

EGI Conference 2024 – 02/10/2024

AI4 | eosc

Co-funded by
the European Union

AI4EOSC | amcaar@i3m.upv.es | ai4eosc.eu

# Thank you! Any questions?

Amanda Calatrava (on behalf of all the authors)