

# A software ecosystem for provenance management in large-scale AI workflows

---

Prof. Sandro Fiore

Department of Information Engineering and Computer Science

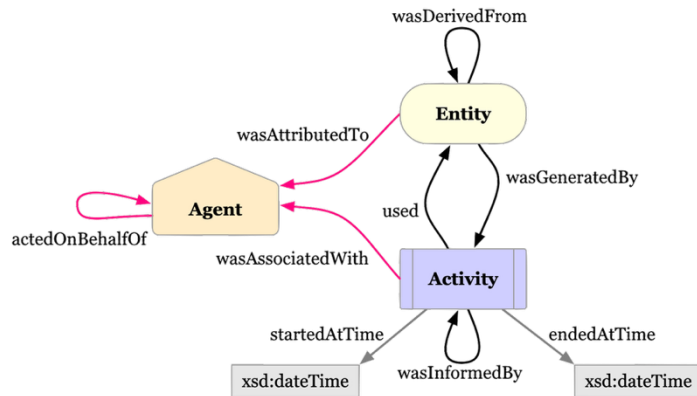
University of Trento

# Introduction: why

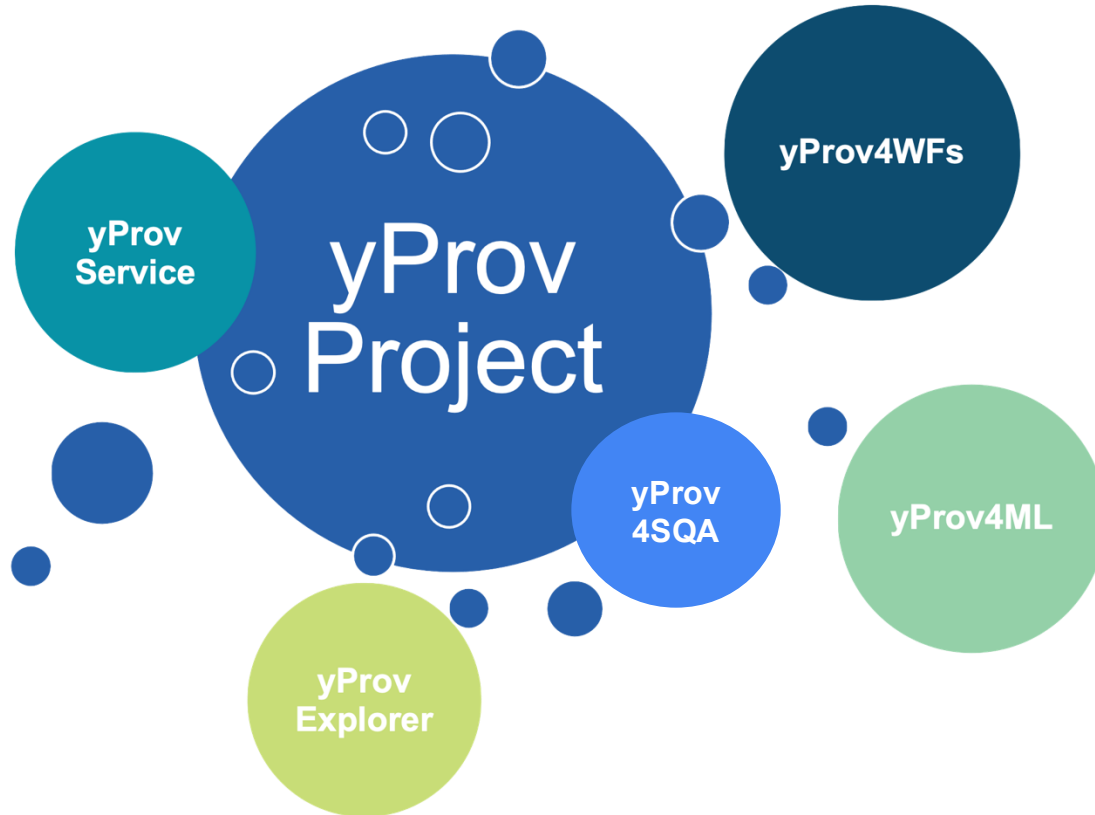
- Provenance provides the **historical record of data** from its original sources
- Provenance represents **accompanying documentation** for scientific data
- It represents the “**other face of the medal**” w.r.t. scientific experiments & workflows
- Provenance is a key enabling factor for **reproducibility, trust, source attribution**
- It is an **integral part of the experiment's output** — a research object in its own right — and as such, it is also subject to provenance tracking (i.e., *the provenance of provenance*).
- It helps address **end-to-end challenges**—through the use of Persistent Identifiers (PIDs)—across teams, departments, organizations, and national boundaries.

# W3C PROV (family of) standards

«**Provenance** is information about **entities**, **activities**, and **people** involved in producing a piece of data or thing, which can be used to form assessments about its **quality**, **reliability** or **trustworthiness**»



# yProv: a software ecosystem for provenance management

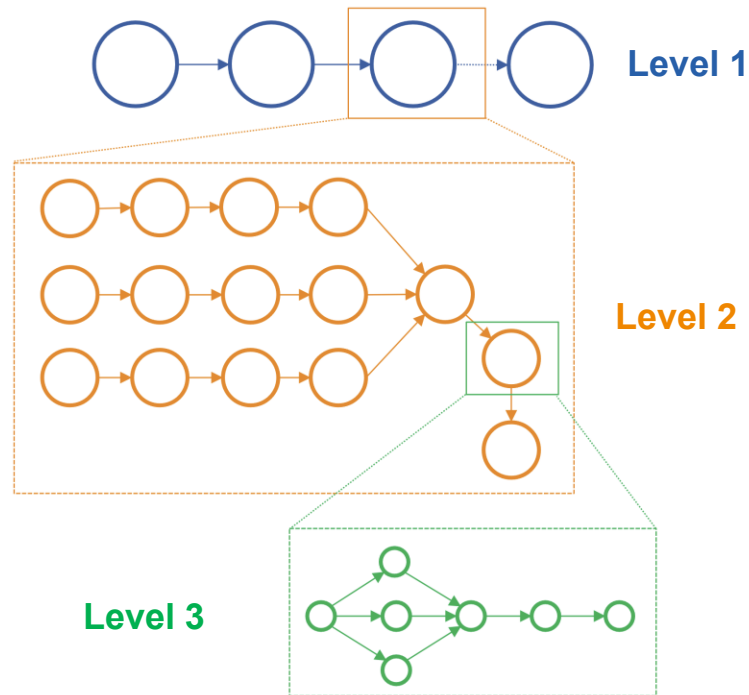


# yProv service: an interoperable component for provenance management

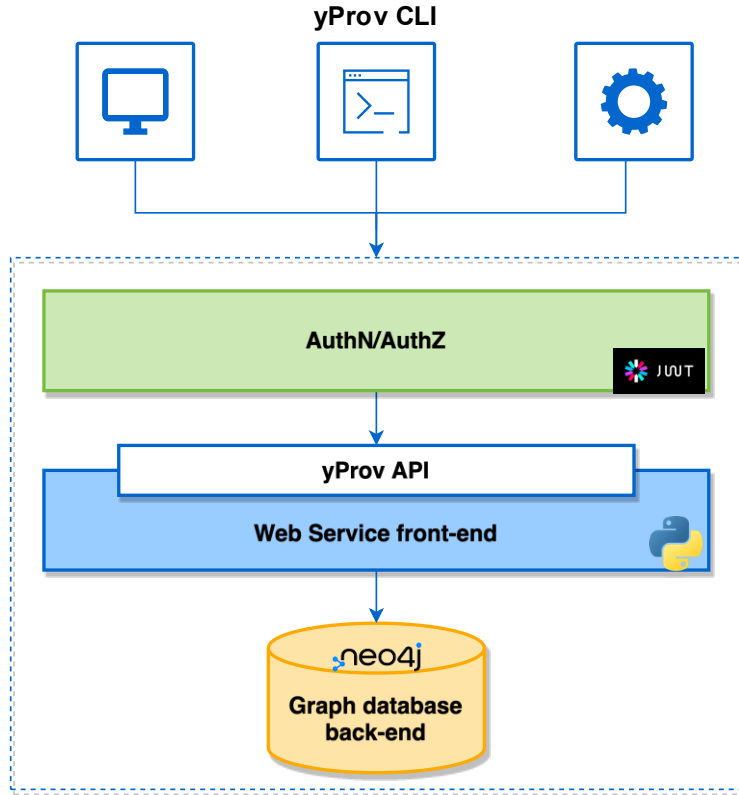
**yProv service is a lightweight and interoperable service for provenance documents management**

- Multi-level provenance management support
- Back-end based on graph data model
- RESTful interface and W3C PROV compliance

<https://github.com/HPCI-Lab/yProv>

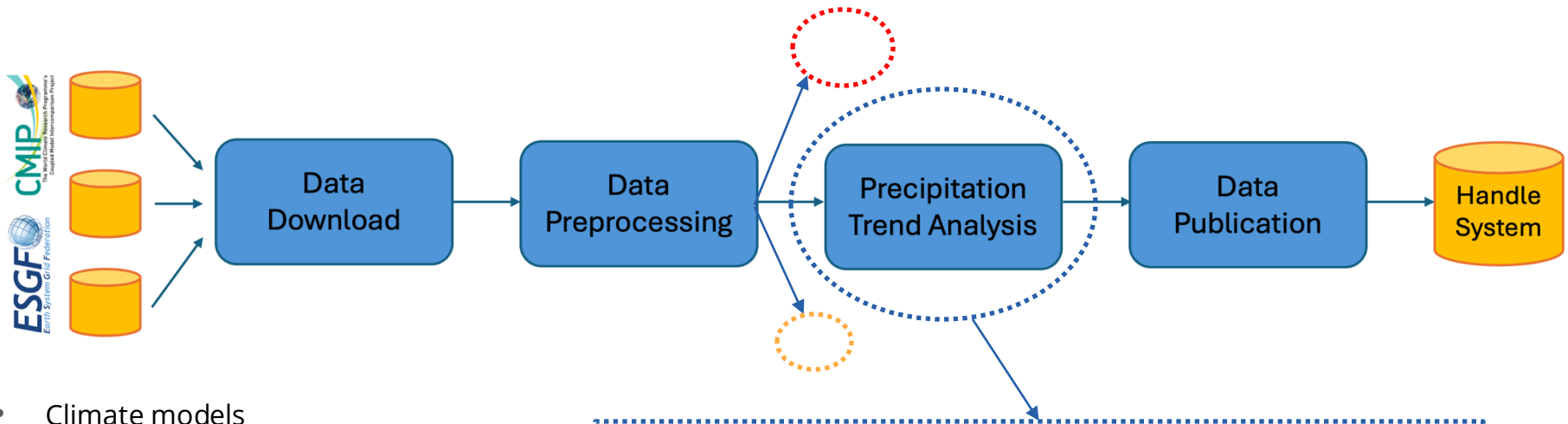


# yProv service architecture and the role of the API

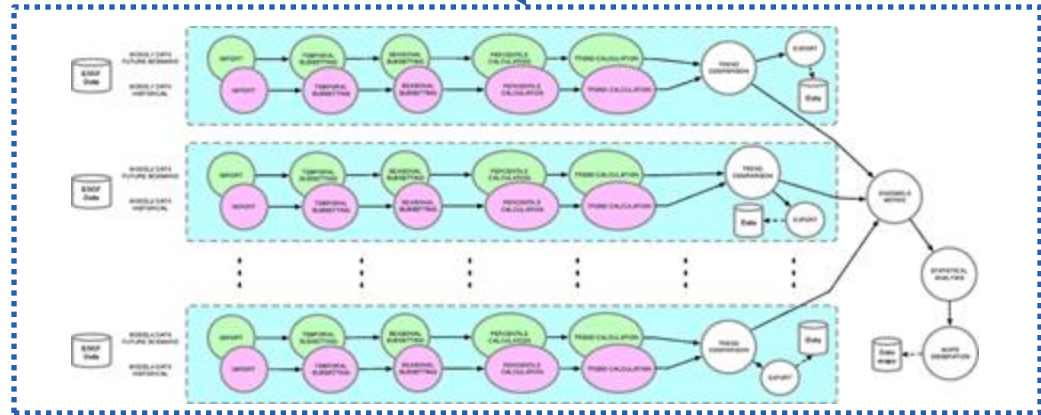


- **3 components**
  - Web Service front-end
  - Graph database engine back-end (Neo4j)
  - Command Line Interface
- **Authentication/Authorization**
  - Based on JSON Web Token (JWT)
- **RESTful API (OpenAPI)**
  - Easy way to interact with the service and manage PROV information

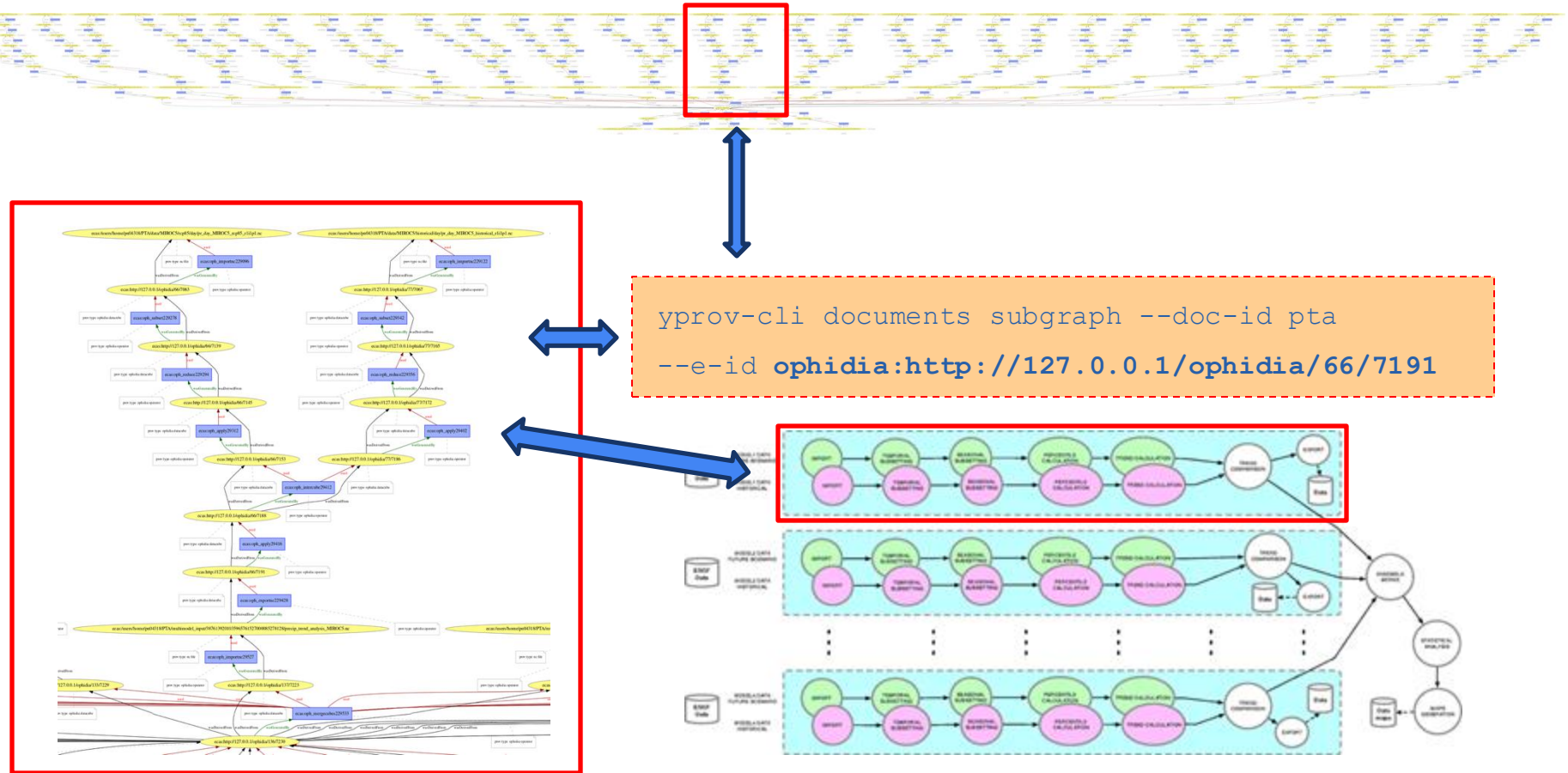
# End-to-end provenance management: a climate analytics wf



- Climate models intercomparison wf consisting of hundreds of tasks
- Multiple inputs and multiple outputs
- Both intermediate and final outputs
- Multiple *agents*



# Single output provenance: sub-graph feature





```

graph LR
    A[Data Download] --> B[Data preprocessing]
    B --> C[Precipitation Trend Analysis]
    C --> D[Data Publication]
    D --> E[Handl e System]
  
```



## Level-1

# yProvExplorer



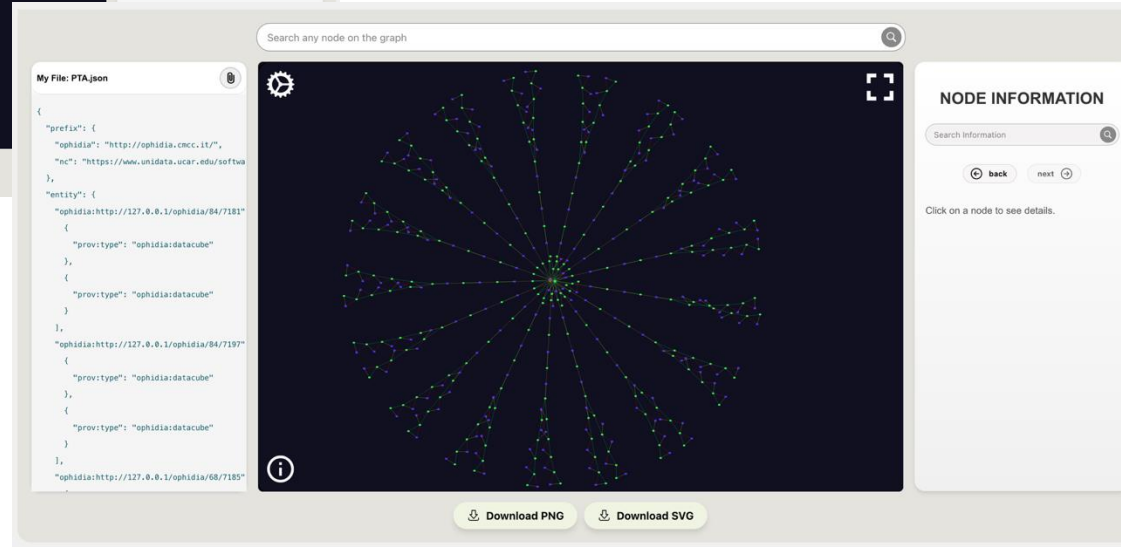
Multi-model  
experiment in  
climate



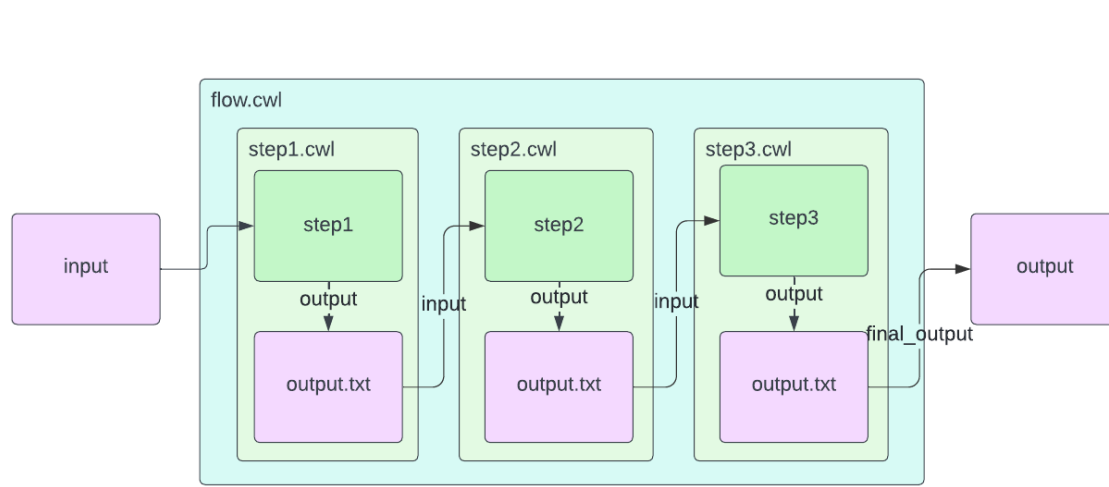
<https://explorer.yprov.disi.unitn.it/>

Workflow in the EO  
domain

<https://explorer.yprov.disi.unitn.it/?file=http%3A%2F%2Fyprov.disi.unitn.it%3A3000%2Fapi%2Fv0%2Fdocuments%2Fyprov4wfs>



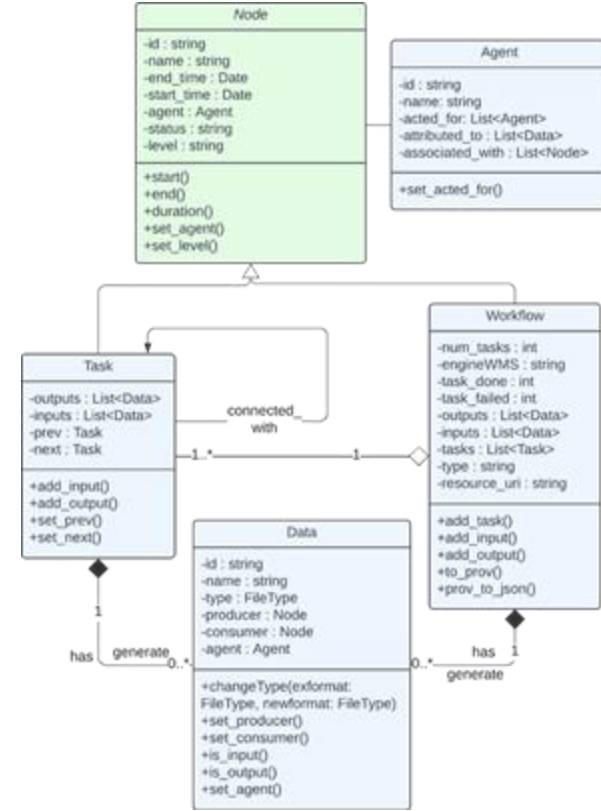
# Prov4WFs: a library for workflow provenance tracking



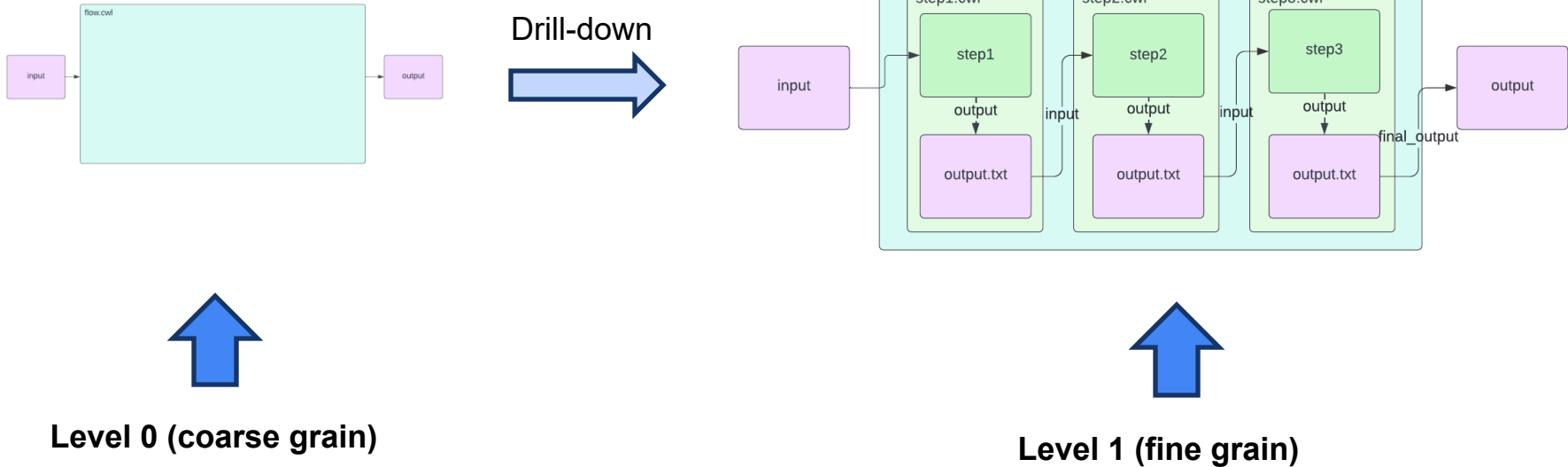
## Objectives:

1. navigability (task by task)
2. ease of understanding the workflow structure
3. comply with W3C PROV family of standards

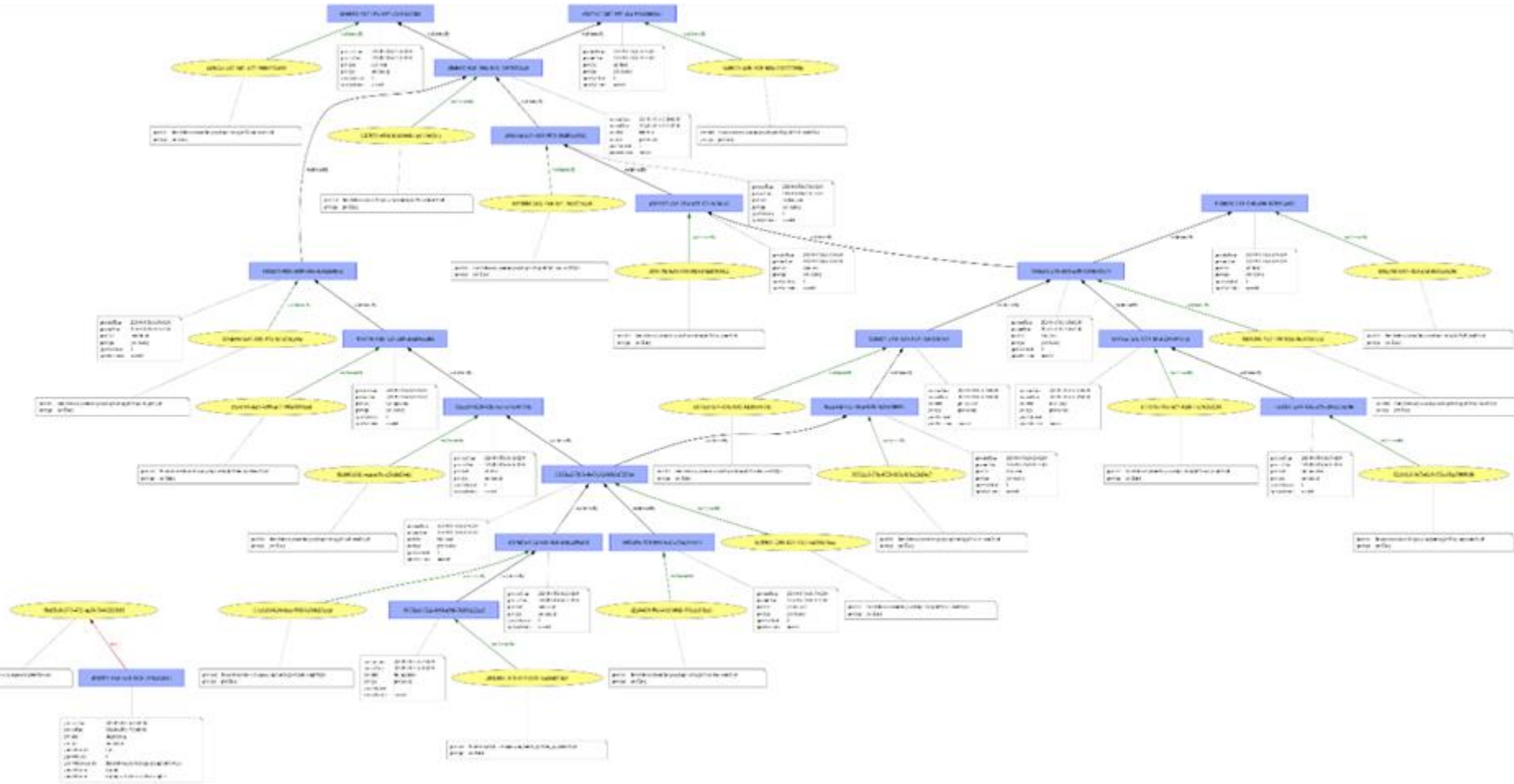
Target WfMSs: **Streamflow** (done), **openEO** (done), **cylc** (in-progress)



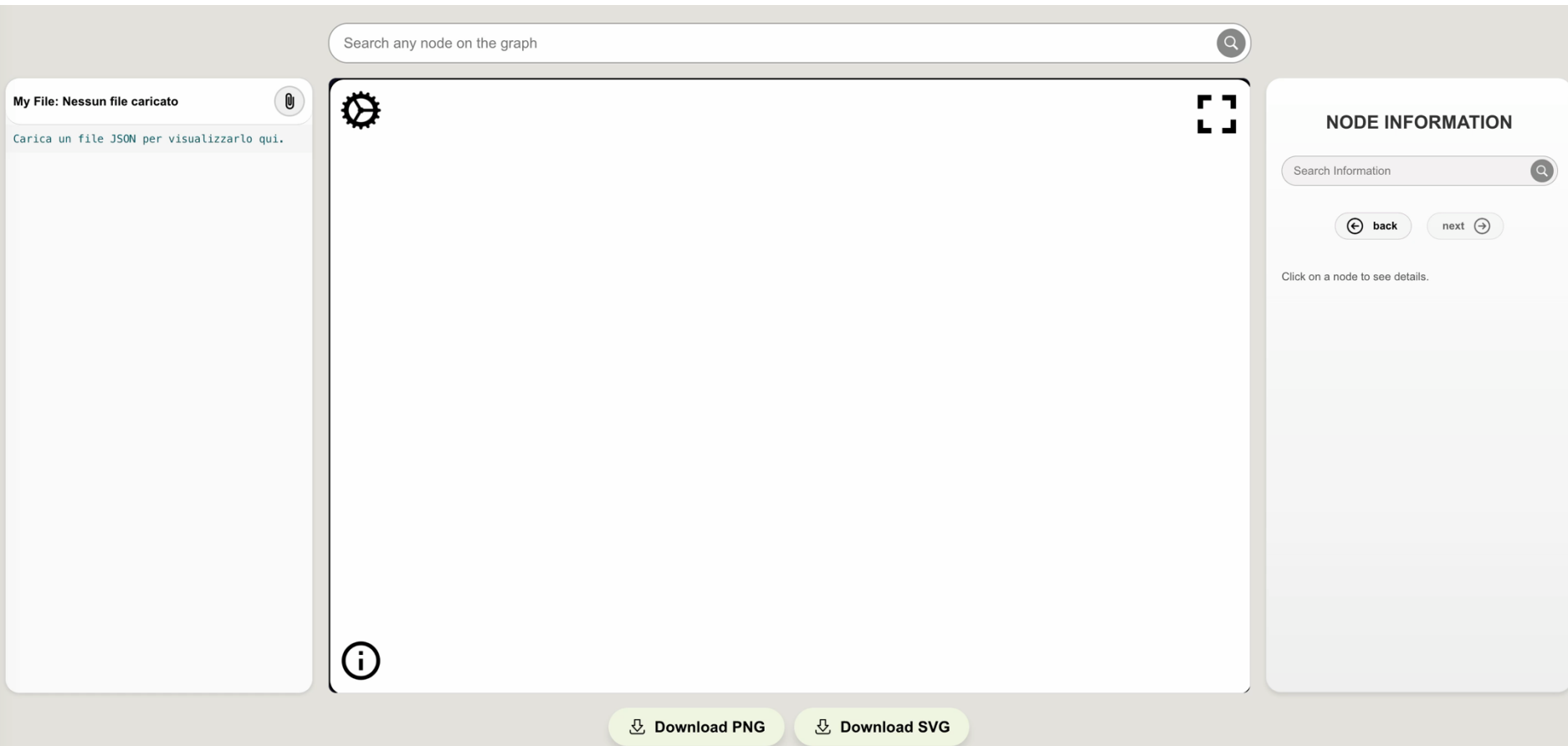
# Simple workflow multi-level test case



## Prov4WFs: a simple test case with multi-level outputs



## Prov4WFs: a production workflow use case

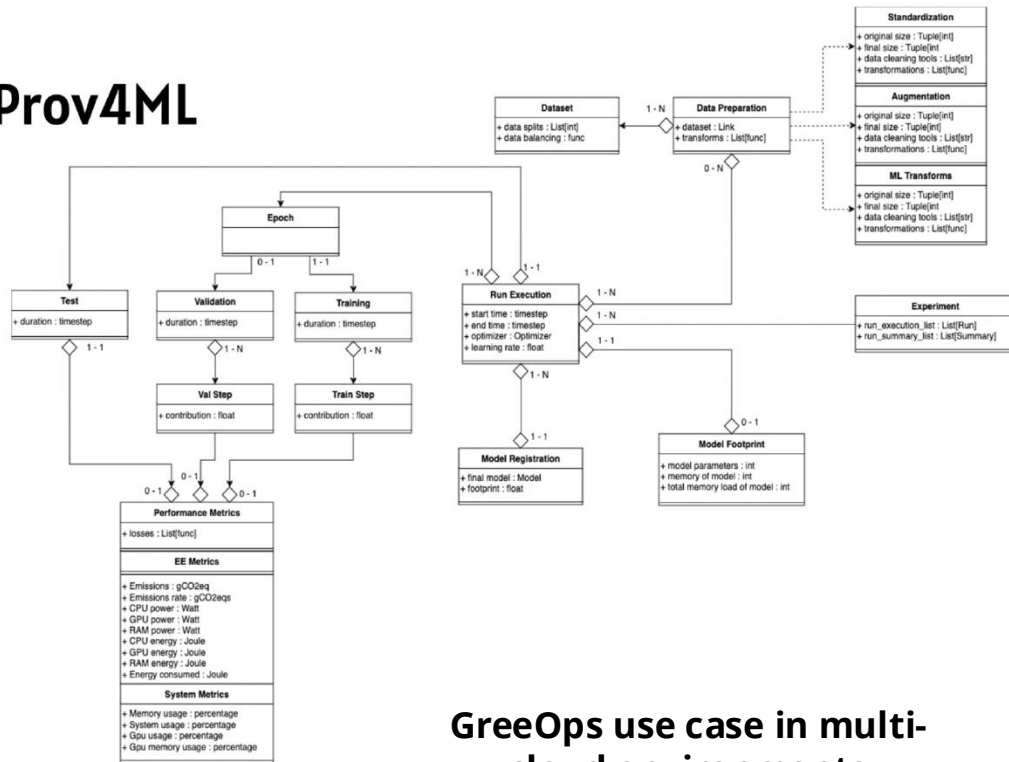


# Prov4ML: A library for tracking provenance in ML processes

**Prov4ML goal:** library able to track W3C PROV-compliant provenance within AI processes jointly with a set of metrics, across runs and epochs

Parameter	Description	Unit
Emissions	Emissions of the system	gCO2eq
Emissions rate	Emissions rate of the system	gCO2eq/s
CPU power	Power usage of the CPU	W
GPU power	Power usage of the GPU	W
RAM power	Power usage of the RAM	W
CPU energy	Energy usage of the CPU	J
GPU energy	Energy usage of the GPU	J
RAM energy	Energy usage of the RAM	J
Energy consumed	Energy consumed by the system	J

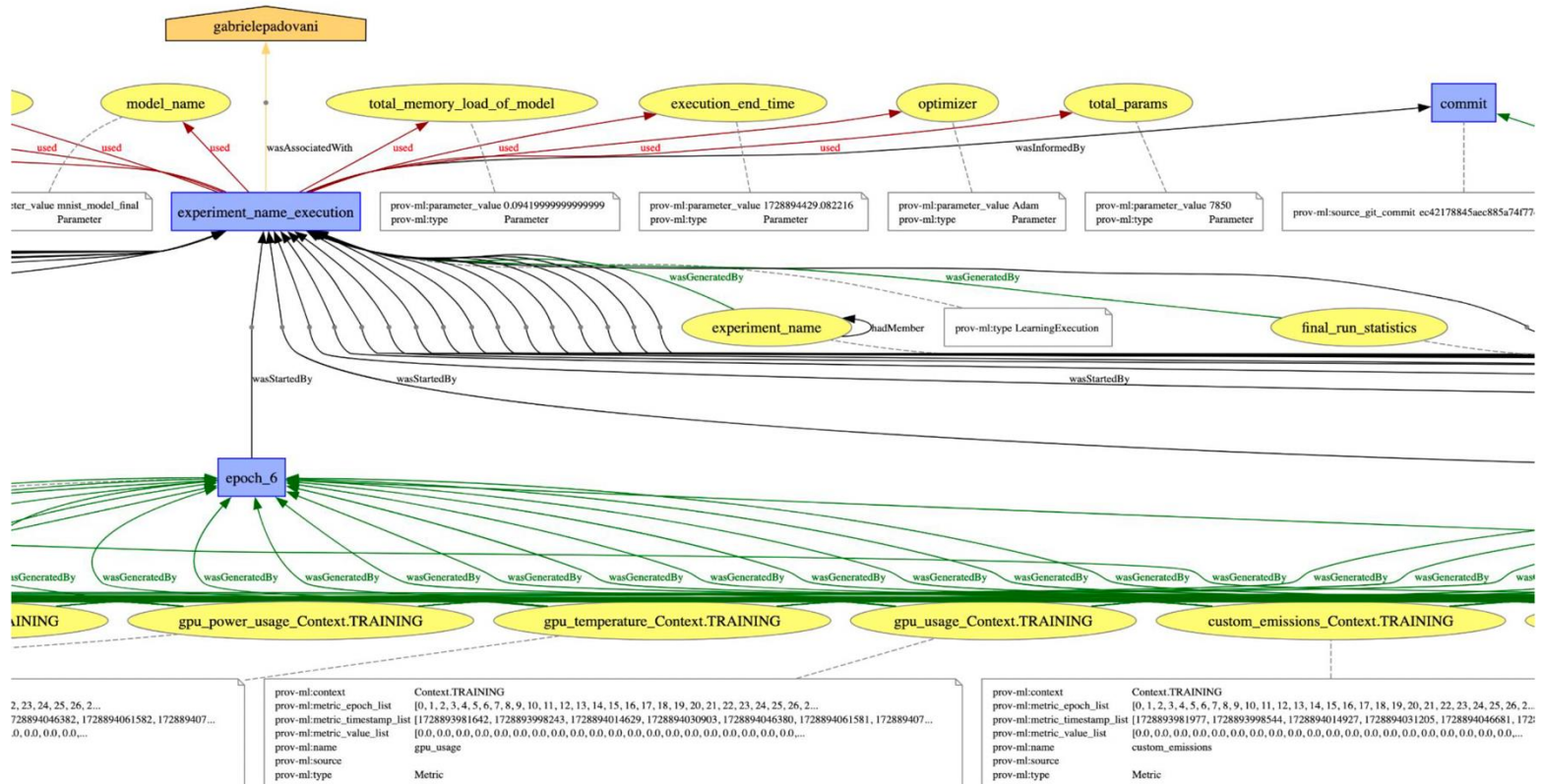
## Prov4ML



**GreeOps use case in multi-cloud environments**



# Prov4ML: A library for tracking provenance in ML processes



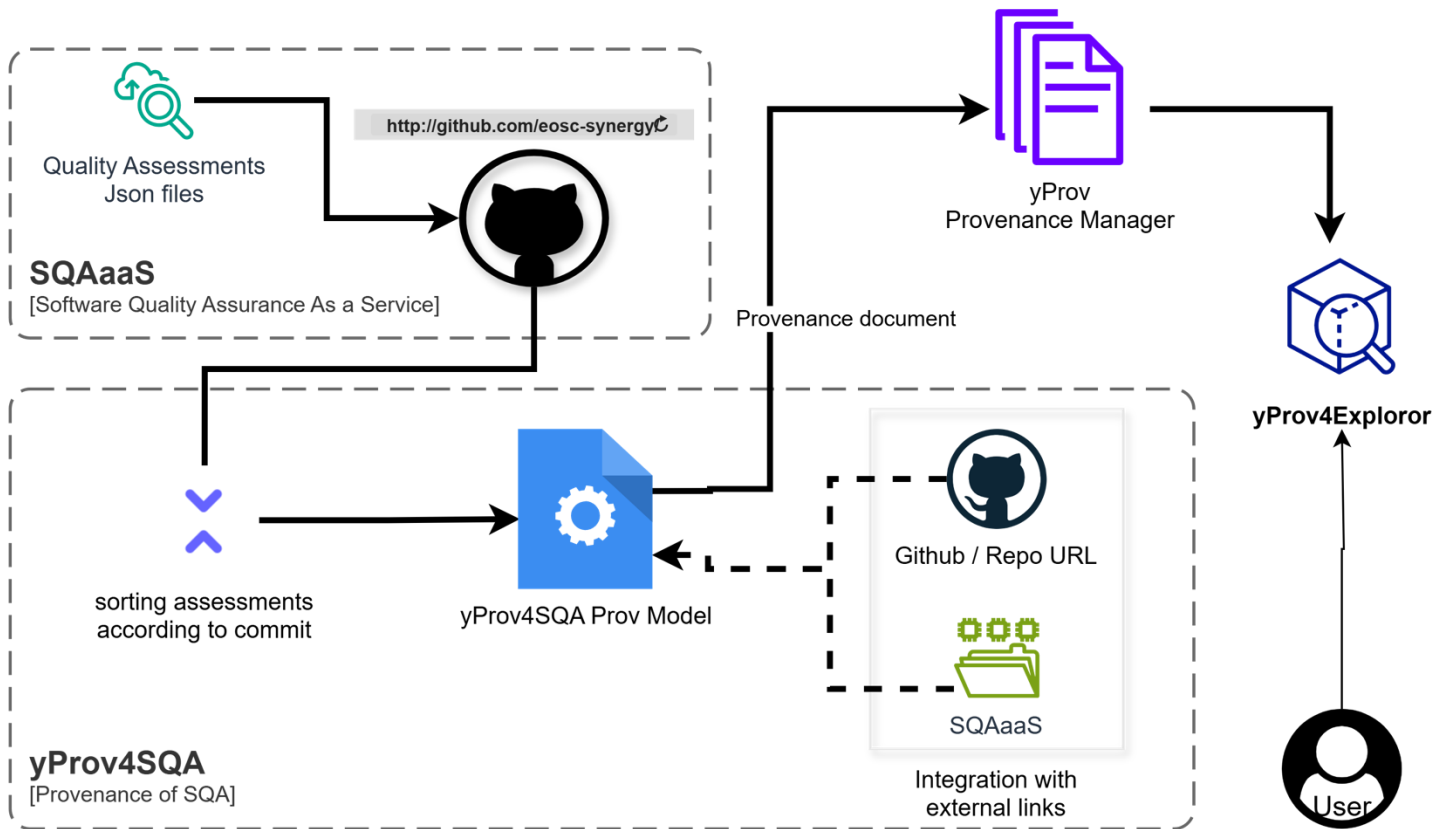
*Metrics tracking and comparison analysis but also forecasting*



# yProv4SQA: traceability in SQA workflows

- Integrate automated provenance tracking into automated software quality assurance workflows.
- Enable traceability of software assessments across versions/releases.
- Provide tool to understand and compare software quality over time.
- Facilitate decision-making through historical analysis of code and quality evolution.
- Support integration with platforms like GitHub and SQAaaS for real-time assessments comparison.

# Architecture of yProv4SQA

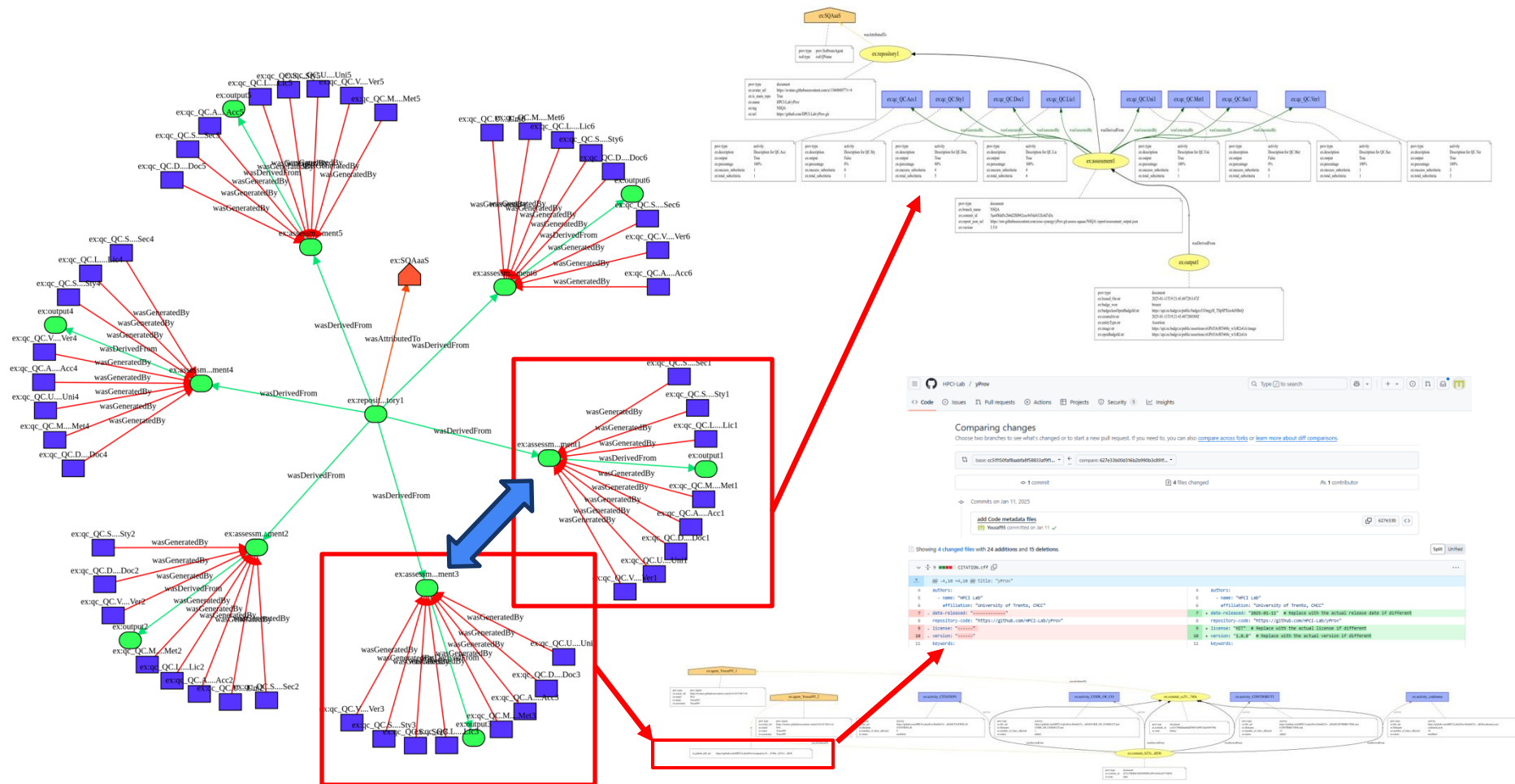


(UNITN/CSIC collaboration)

# YProv4SQA: Multi-Level Provenance Document

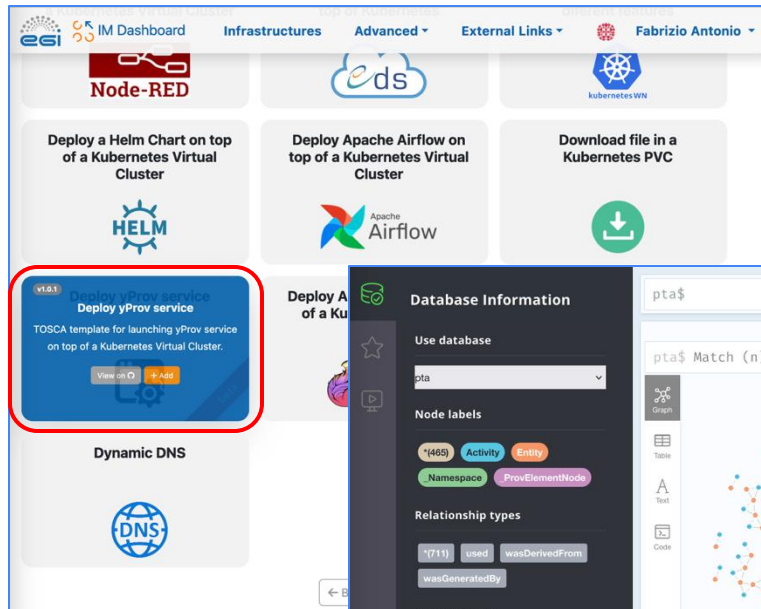
- YProv4SQA is a provenance library that records multiple assessments of a single repository within a single provenance document.
- These assessments are generated by the SQAaaS and all display in one place.
- Provenance document contains two levels:
  - **Level 1: Repository & Assessment Data**
    - Provides repository and assessment level Provenance graph
  - **Level 2: Code Change Comparison**
    - Shows the comparison of any code changes between two assessments

# Overall assessment view and assessments comparison



# Cloud-based yProv version in EOSC

<https://im.egi.eu/im-dashboard/>



The screenshot shows the IM Dashboard with a navigation bar at the top containing 'IM Dashboard', 'Infrastructures', 'Advanced', 'External Links', and a user profile 'Fabrizio Antonio'. Below the navigation bar, there are several deployment cards. A red box highlights the 'Deploy yProv service' card, which includes the text: 'TOSCA template for launching yProv service on top of a Kubernetes Virtual Cluster.' Other visible cards include 'Node-RED', 'eDS', 'kubernetes WNN', 'Deploy a Helm Chart on top of a Kubernetes Virtual Cluster', 'Deploy Apache Airflow on top of a Kubernetes Virtual Cluster', 'Download file in a Kubernetes PVC', and 'Dynamic DNS'.



### Deploy a Kubernetes Virtual Cluster + yProv

**Description:** Deploy a Kubernetes Virtual Cluster.  
TOSCA template for launching yProv service on top of a Kubernetes Virtual Cluster.

Infrastructure Name  
yProv service

**FE Features**   **WNs Features**   **Kubernetes Data**   **yProv**

**Cloud Provider Selection**

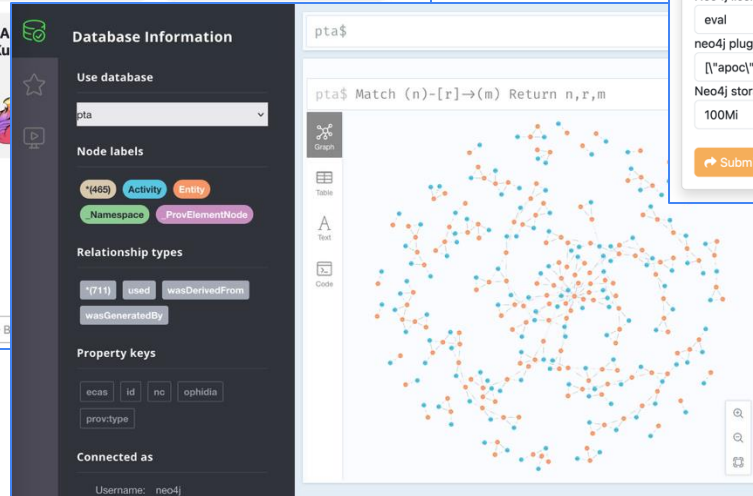
Neo4j username  
neo4j

Neo4j password  
.....

Neo4j license type  
eval

neo4j plugins to be enabled (remember to escape quotes with backslashes)  
["apoc", "graph-data-science"]

Neo4j storage capacity  
100Mi



The screenshot shows the Neo4j database interface. On the left, there is a sidebar with 'Database Information' and 'Node labels' (including 'Activity', 'Entity', 'Namespace', and 'ProvElementNode'). The main area displays a graph visualization with nodes and edges. A query is entered in the top bar: `pta$ Match (n)-[r]-(m) Return n,r,m`. The graph shows a complex network of nodes and relationships.



UNIVERSITAT  
POLITÀCNICA  
DE VALÈNCIA

**Acknowledgements**  
**M. Caballer and I. Blanquer**

# EOSC-Node Interoperability in EOSC-Beyond

marketplace.sandbox.eosc-beyond.eu/services/21.15124%  
Pre Production < **eosc BEYOND** About EOSC Discovery Hub Providers Hub Monitoring Status Contact us My EOSC  
Go to Search

in

o

W

z

✉

Service

Natural Sciences - Earth & Related Environmental Sciences

Data - Scientific/Research Data

Organisation: [ENES](#)

Provenance as a service

★ ★ ★ ★ ★

0.0/5 (0 reviews)

☐
[↗ Compare](#)

Access the service

Manage the service

OPEN ACCESS

Provides feedback

About

Details

Reviews (0)

The **Provenance Service** offers a comprehensive historical record of data, outlining its origins and transformations within **open science environments**. It captures and manages data **lineage information** across different levels of granularity, supporting the needs of large-scale scientific experiments.

By maintaining a complete provenance trail of experimental processes and data, the service facilitates the **reproducibility** and **verifiability** of scientific results, thereby reinforcing trustworthiness and integrity in analytical research.

The diagram illustrates three interconnected components, each represented by a rounded rectangle with a light blue border and a light gray background. The components are arranged horizontally and connected by thin gray lines.

- Target Users:** Contains a list of three items: "Research Communities", "Research Projects", and "Research Groups".
- Tags:** Contains a list of four items: "Provenance", "provenance-service", "open-science", and "data-analysis".
- Availability and Language:** Contains two sections: "Regions:" with the value "World", and "Languages:".

Thin gray lines connect the right side of the "Target Users" box to the left side of the "Tags" box, and the right side of the "Tags" box to the left side of the "Availability and Language" box.

### Provenance Service

The **Provenance Service** offers a comprehensive historical record of data, outlining its origins and transformations within **open science environments**. It captures and manages **data lineage information** across different levels of granularity, supporting the needs of large-scale scientific experiments.

By maintaining a complete provenance trail of experimental processes and data, the service facilitates the **reproducibility** and **verifiability** of scientific results, thereby reinforcing trustworthiness and integrity in analytical research.

Creation Date: Apr 28, 2025

Last Update: May 22, 2025

# Conclusions, ongoing activities and lots of people to thank



**Full  
provenance  
software  
ecosystem**



**Domain  
agnostic, it  
can serve any  
research  
infrastructure  
interested in  
traceability  
aspects**



**FAIRness of  
provenance  
(data &  
process)**



**A catchall  
service is  
running at  
UNITN, jointly  
with the  
yProvExplorer**



**It can enable  
reproducibility  
scenarios on  
top of the  
provenance  
information**



**It implements  
a vision  
aligned with  
the EU Data  
Strategy**

# Thanks

