Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

# Building AI-Ready Data Ecosystems

**Arnaud Ceol, arnaud.ceol@supercomputing-icsc.it**
**ICSC - Italian Research Center on HPC, Big Data and Quantum Computing**

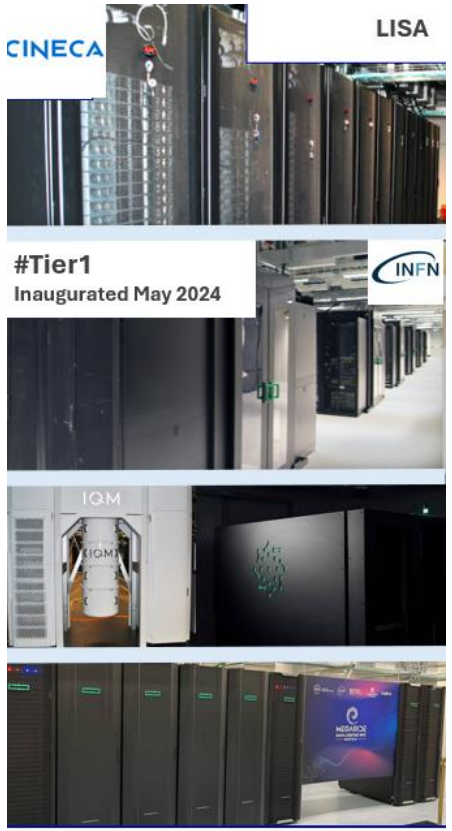
**IT4LIA AI Factory | IBERGRID 2026, 28 Oct**

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

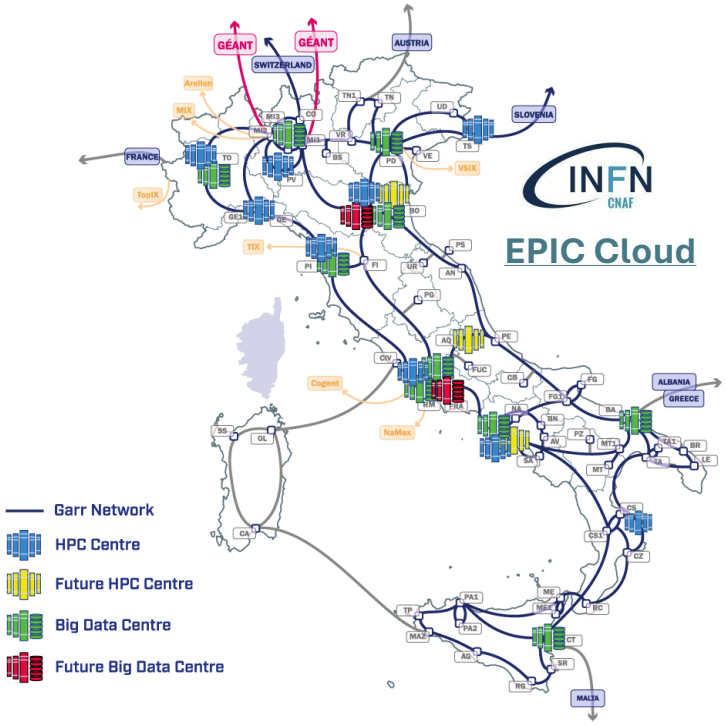
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing




#MEGARIDE
Optimized for AI and Cybersecurity

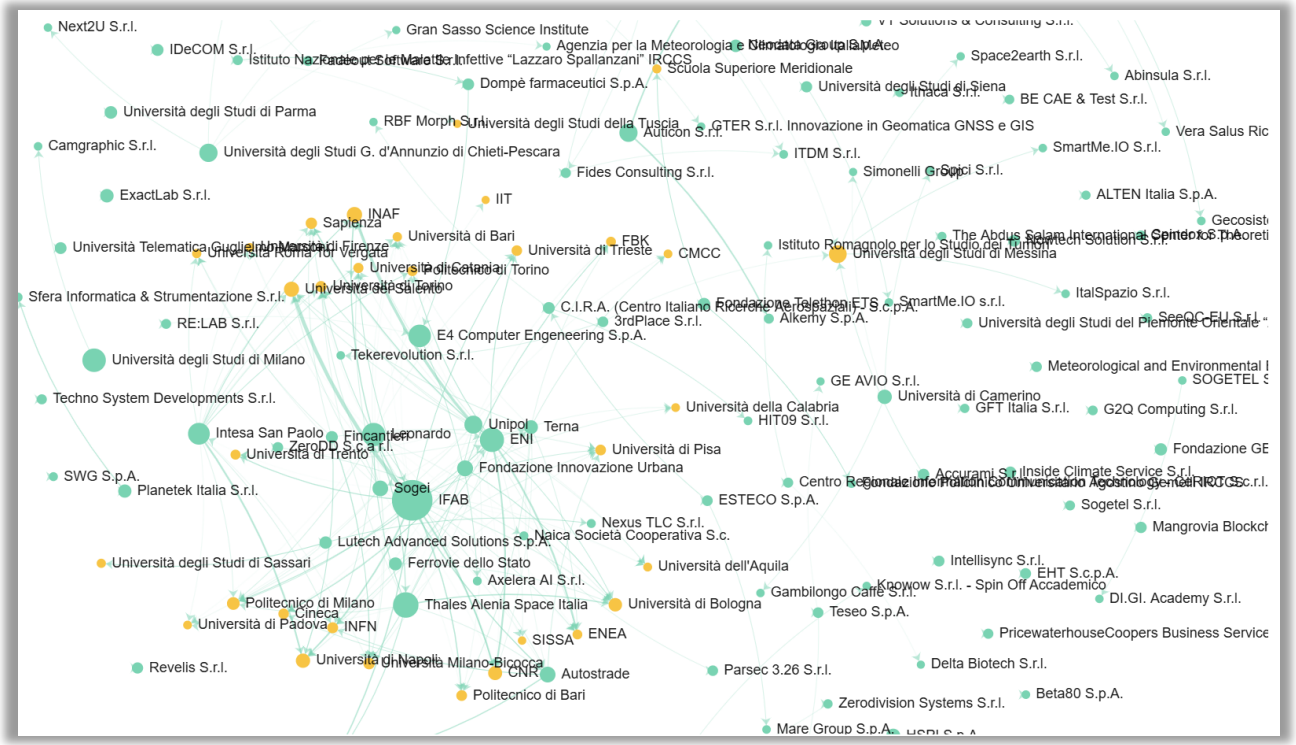


**INNOVATE**: The First INDUSTRIAL EURO HPC Supercomputer




INFN CNAF

## EPIC Cloud




## National Research Institutes

**13** Research Institute

**25** Universities

## Private partners

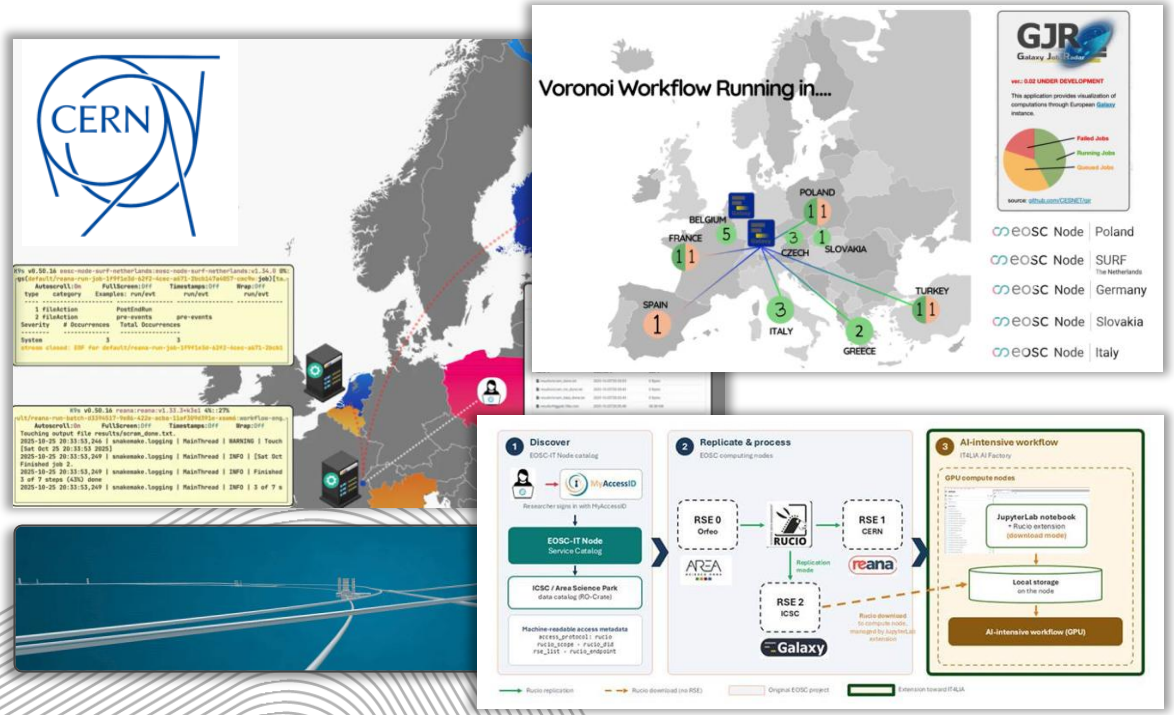
**14** Private Companies

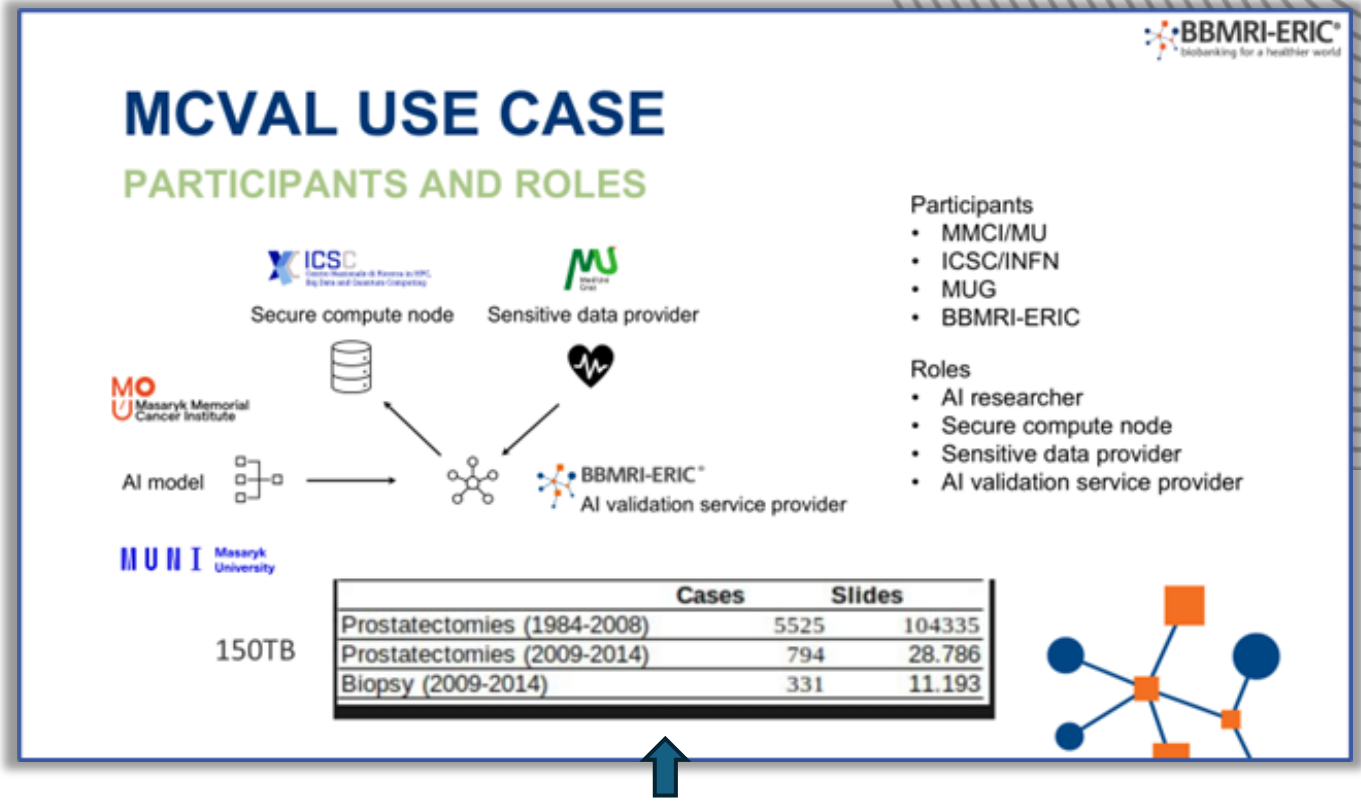
# eosc




**Bologna, February 20** – Today, it has been confirmed that ICSC – Italian Research Center on HPC, Big Data and Quantum Computing has been officially selected as a National Node of the **European Open Science Cloud**

https://eosc.eu/building-the-eosc-federation/eosc-node-italy




**Main goals:**

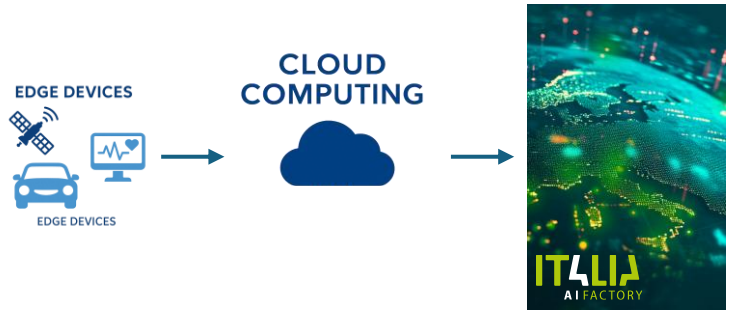
- Provide a cloud-enabled platform for the multi-centric validation of AI models for prostate cancer screening on sensitive health data in compliant frameworks

**Main contribution of EOSC Node Italy:**

- **Exploit sensitive data storage and analysis** facilities leveraging the INFN ISO-certified cloud infrastructure
- Provide compute, storage and AI capacities in EHDS (European Health Data Space Regulation) compliant manner for data-users to the research community leveraging INFN resources and competencies

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

# ECHO-TWIN: extending to the edge




**An integrated digital ecosystem to design, run and transfer AI-enabled digital twins across the edge–cloud–HPC continuum.**

## Why

- AI and real-time data applications span edge devices, cloud and HPC
- Companies and public administrations need validated infrastructure, labs and use cases to adopt them

## What

- Federated, secure and interoperable edge–cloud–HPC infrastructure
- Digital twins and AI pipelines running along the continuum
- Thematic labs: **health**, **environment**, **toxicology**, **mobility**, **cybersecurity**

**Goal**
Develop an integrated monitoring and AI framework to collect real-time air pollutant measurements in poorly monitored, urban and high-traffic areas, supporting high-resolution pollution mapping, exposure assessment and short-term forecasting.

**Sniffer sensor**
Portable system for real-time pollutant measurements and georeferenced data collection.

**Drones**
Airborne monitoring to extend spatial coverage and identify pollution hotspots.

**Workstations**
High-performance computing for data analysis, AI model development and forecasting.




Università degli Studi di Bari Aldo Moro
Coesione Italia 21-27 Scuole e competenze
Cofinanziato dall'Unione europea
Ministero delle Imprese e del Made in Italy
Ministero dell'Università e della Ricerca

# IT4LIA AI Factory

**Promoters**

**Hosting Entity**

CINECA

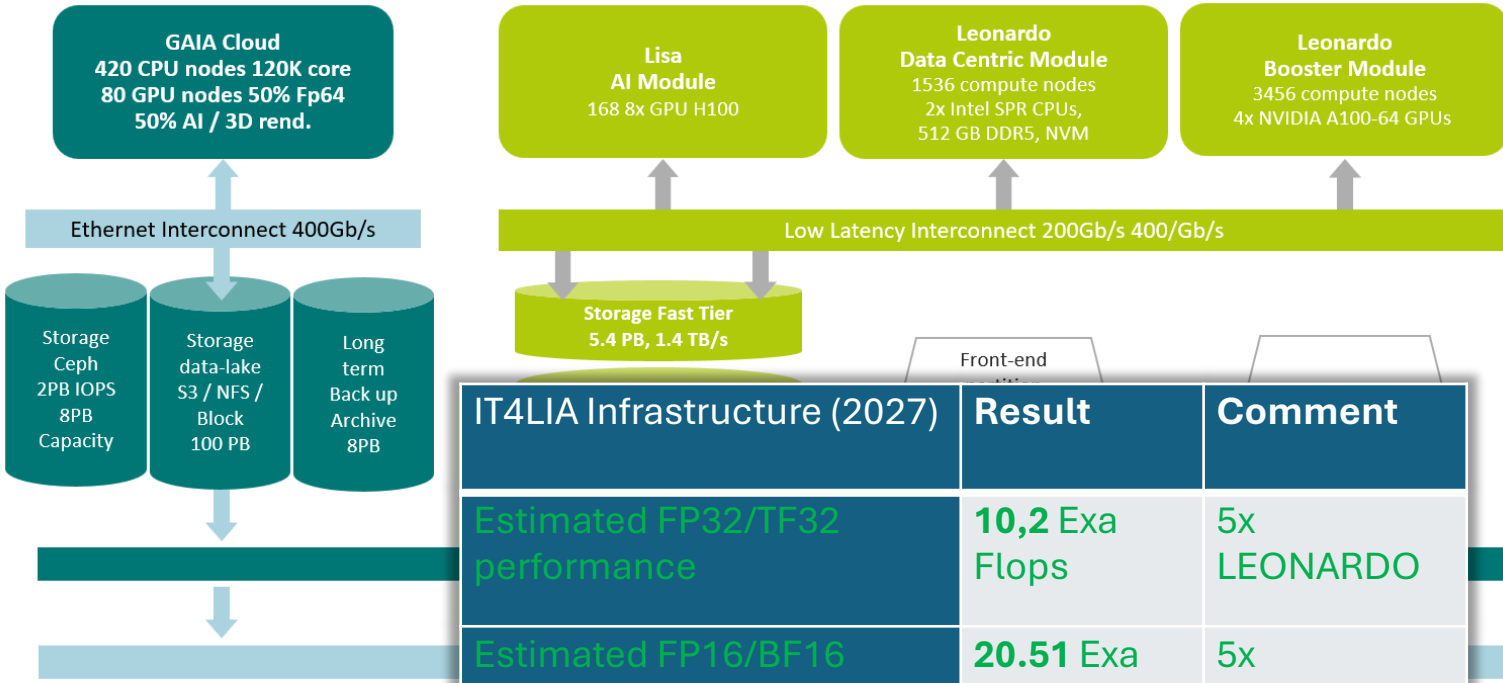
**Other Participating Member States**

Federal Ministry Republic of Austria Climate Action, Environment, Energy, Mobility, Innovation and Technology

Republic of Slovenia Ministry of Higher Education, Science and Innovation

**Partners**



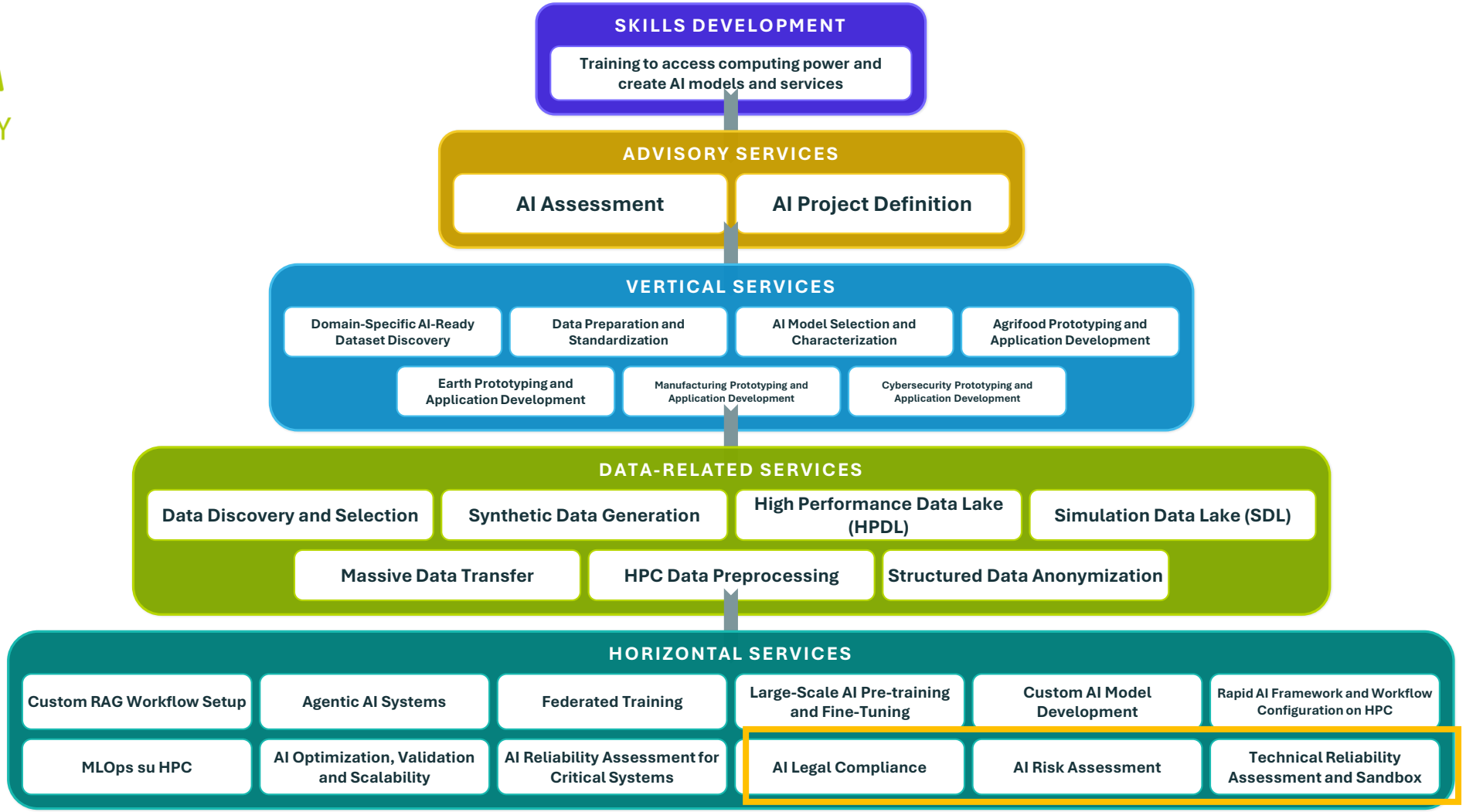







| IT4LIA Infrastructure (2027) | Result | Comment |
|---|---|---|
| Estimated FP32/TF32 performance | **10,2** Exa Flops | 5x LEONARDO |
| Estimated FP16/BF16 performance | **20.51** Exa Flops | 5x LEONARDO |
| Estimated FP8 performance | **41.02** Exa Flops | 15x LISA |
| Estimated INT8 performance | **41.02** Exa POPs | 15x LISA |
| AI science performance (FP64) | 327 Peta Flops | 2.5x LEONARDO |

ITALIA AI FACTORY



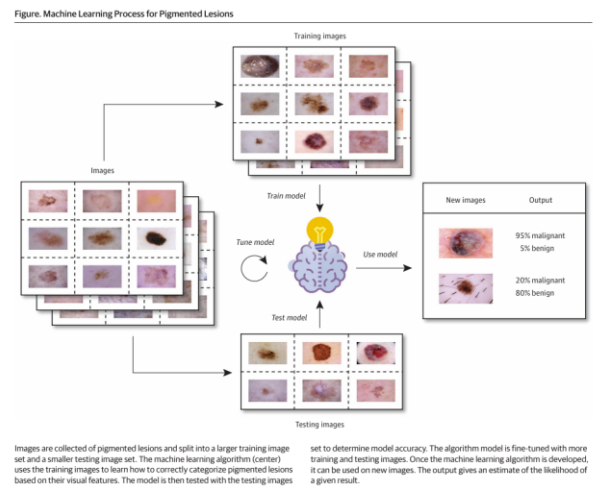

From capability building to the technology platform: 29 connected services in one integrated pathway

# Embedding AI Trust into professional training

## Example: AI for skin-cancer detection

- Trained mostly on light skin
- Performs poorly on dark skin
- Consequence: diagnostic inequalities

The model looks accurate on average – the bias stays hidden until patients are affected.




**Data not representative**

↓

**Bias hidden in the model**

↓

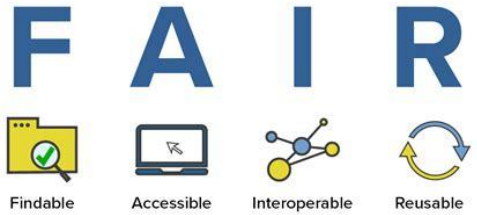
**Errors surface only in clinical use**

↓

**Oversight has to start with the data**

*Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatology. 2018 Nov;154(11):1247-1248. DOI: 10.1001/jamadermatol.2018.2348. PMID: 30073260.*

# FAIR is the bases (not only for research)




The value of **FAIR data** beyond academia: it creates value beyond the original project, accelerating research, innovation and AI adoption.

## For researchers

**Greater Visibility and Impact**
Datasets become discoverable, citable and reusable

**More Reproducible Science**
Clear metadata and provenance improve trust and validation

**Faster Data Reuse**
Less time spent rediscovering, cleaning and understanding data

**More Collaboration Opportunities**
Easier participation in national and international projects

## For companies

**Alignment with Industry Priorities**
FAIR data practices improve efficiency, innovation speed, and reduce risks in industrial settings

**Cost Reduction and AI Acceleration**
Shared vocabularies and standardized metadata lower integration costs and accelerate AI development

**Regulatory Readiness and Traceability**
FAIR data ensures traceability and provenance critical for audits in regulated sectors

**Data Collaboration and Long-term Value**
FAIR enables safe partner collaboration and maintains data usability over time

**ICSC and IT4LIA work with researchers, public administrations, startups and SMEs: FAIR, AI-ready data matter to all of them.**

# AI-ready data is the prerequisite for trustworthy AI




AI data readiness: can a dataset be used to train, validate and deploy models reliably and safely?

## Quality Art. 10(3)

**Is the data clean and complete?**

✗ Missing values, corrupted images

## Representativeness Art. 10(3)–(4)

**Does it reflect reality fairly?**

✗ Skin-cancer model trained on light skin

## Annotations Art. 10(2)(c)

**Can AI understand the labels?**

✗ Two radiologists label the same scan differently

## Coverage Art. 10(2)(e),(h)

**Is there enough data?**

✗ 99% healthy patients, 1% rare disease

## Structure Art. 11 · Annex IV

**Is it standardised and documented?**

✗ Every hospital exports a different format

## Traceability Art. 10(2)(b) · 11

**Can we reconstruct its history?**

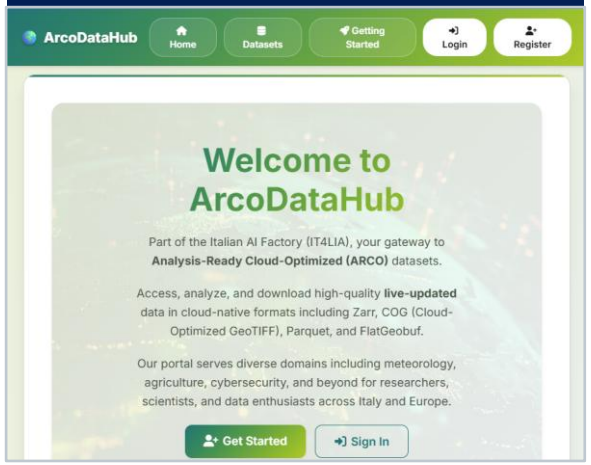
✗ Nobody knows which dataset version trained the model

**The AI Act turns data readiness into a compliance requirement**. FAIR data principles help, but AI compliance additionally requires **evidence of quality, bias management, documentation, and provenance**.

# What AI-ready data looks like in practice

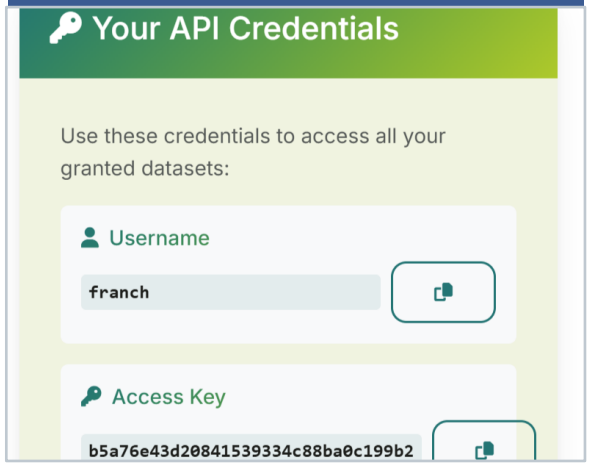
FONDAZIONE BRUNO KESSLER

## 1 Register



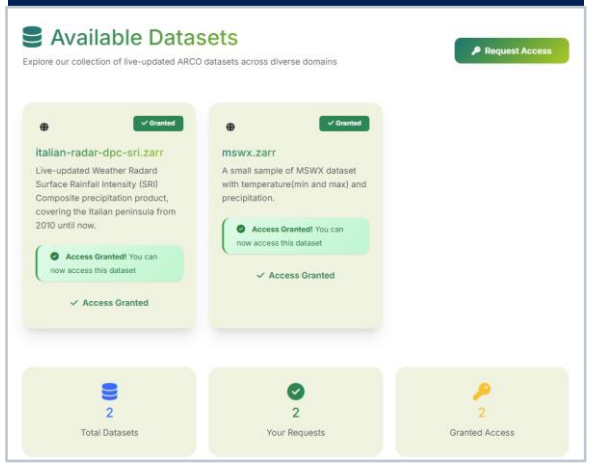

create an account

## 2 Get your key



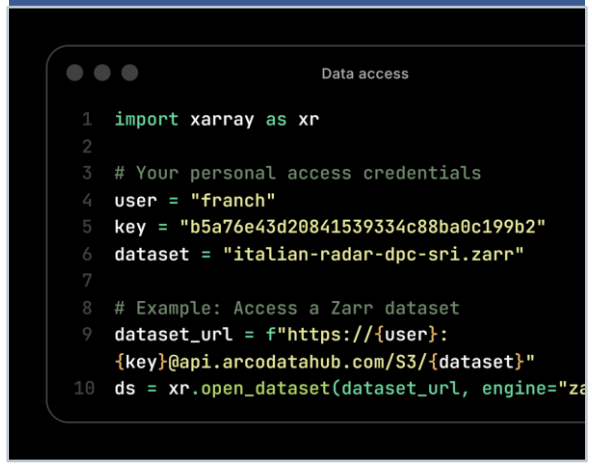

personal API credentials

## 3 Browse datasets




ARCO data, Zarr / Parquet

## 4 Start using it




open it in one line of code

**From:** CSV exports, local column names, no data dictionary, unknown version.
**To:** documented **ARCO** (**A**nalyze **R**eady, **C**loud **O**ptimized) datasets that load in one line

Screens from ArcoDataHub (https://arcodatahub.com/), the IT4LIA portal for analysis-ready, cloud-optimised datasets (with WP5)

# Data stewardship

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing CINECA

## What IT4LIA put in place

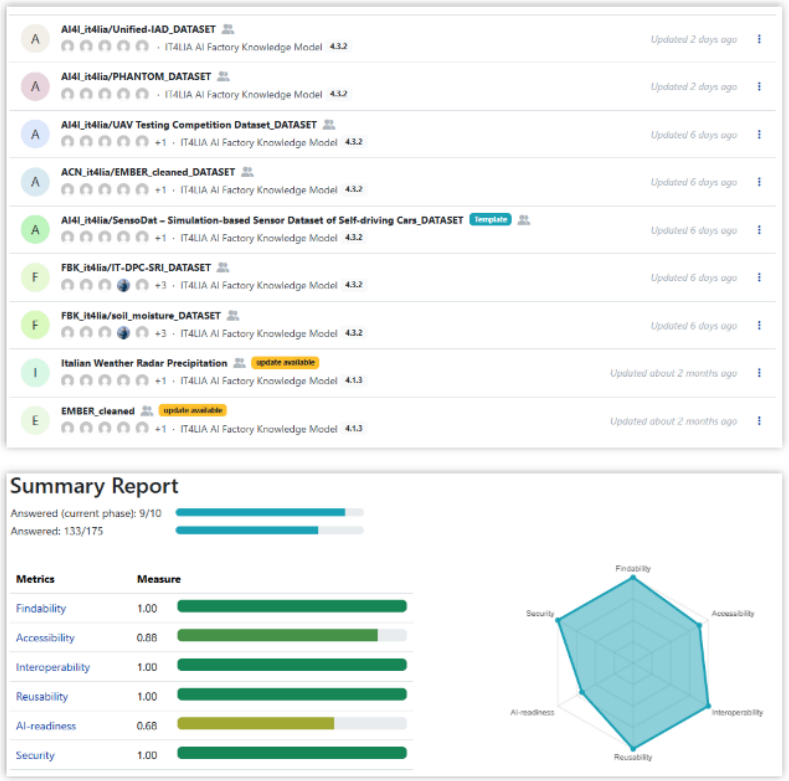
- **Chief Librarian appointed**
- **Data Steward network: 4 stewards trained**
- **Onboarding procedures**
- **Governance reviews before data enters the ecosystem**

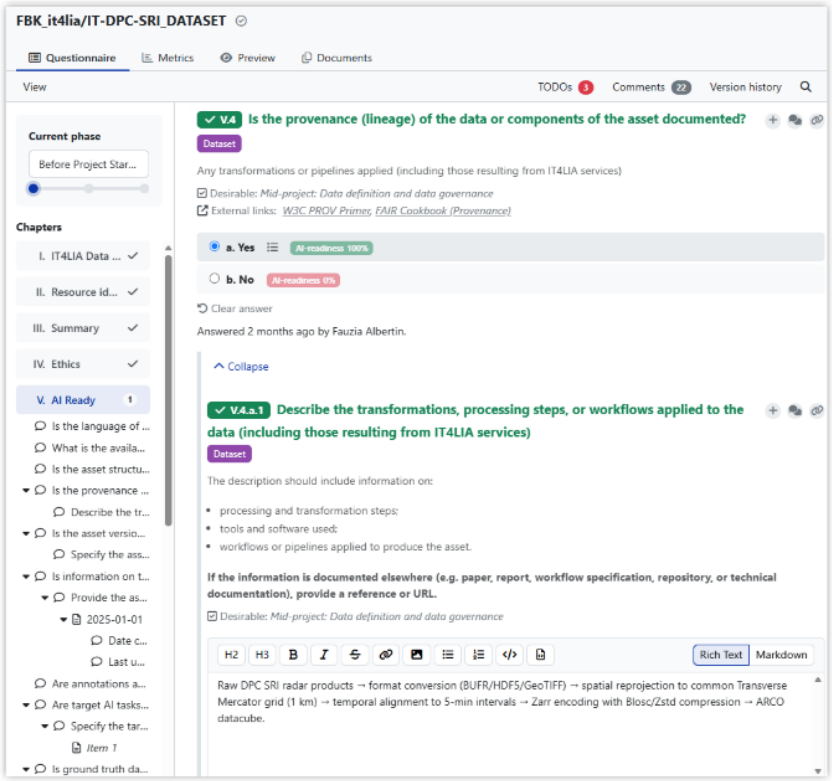
## What a steward does

- Reviews datasets
- Assesses governance
- Checks documentation
- Evaluates AI readiness
- Monitors compliance

**Stewardship turns data governance into practice and makes AI-readiness checkable before the data is used.**

# Embedding AI Trust into professional training




## Extended knowledge model

AI-ready questions added to the standard questionnaire

## AI Readiness score

Summary report with a readiness score for each dataset

Data Stewardship Wizard: https://ds-wizard.org/

# Traceability: a pillar of trustworthy AI

## Accountability is impossible if nobody can answer:

- Which data trained the model?
- Which version?
- Under which licence?
- With which transformations?
- Who approved and who modified it?

## Data Services building blocks:

**National Data Registry + Metadata Engine**

| | |
|---|---|
| Metadata management | Provenance |
| Lineage | Versioning |

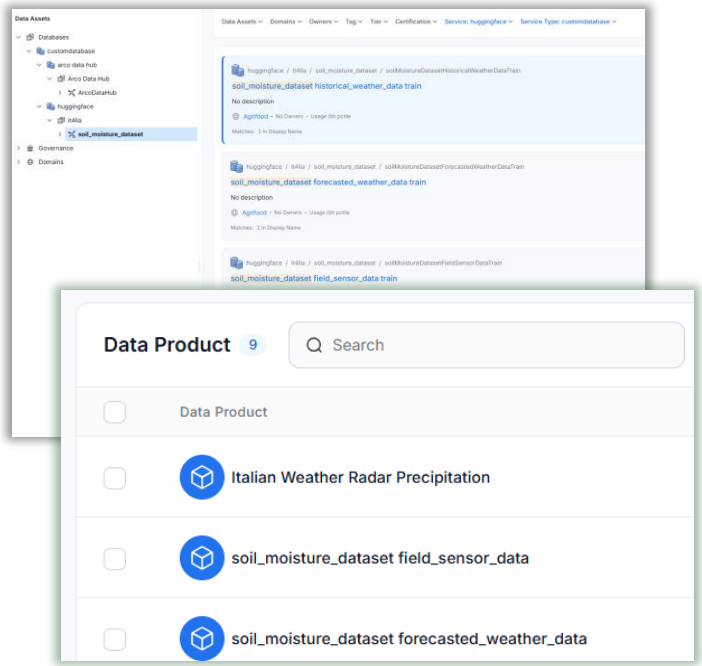
## AI Act hooks

- Art. 10(2)(b) – origin and collection of data
- Art. 11 + Annex IV – dataset documentation
- Art. 12, 19 – logs of system operation

# Metadata Engine: tracking data, quality and lineage

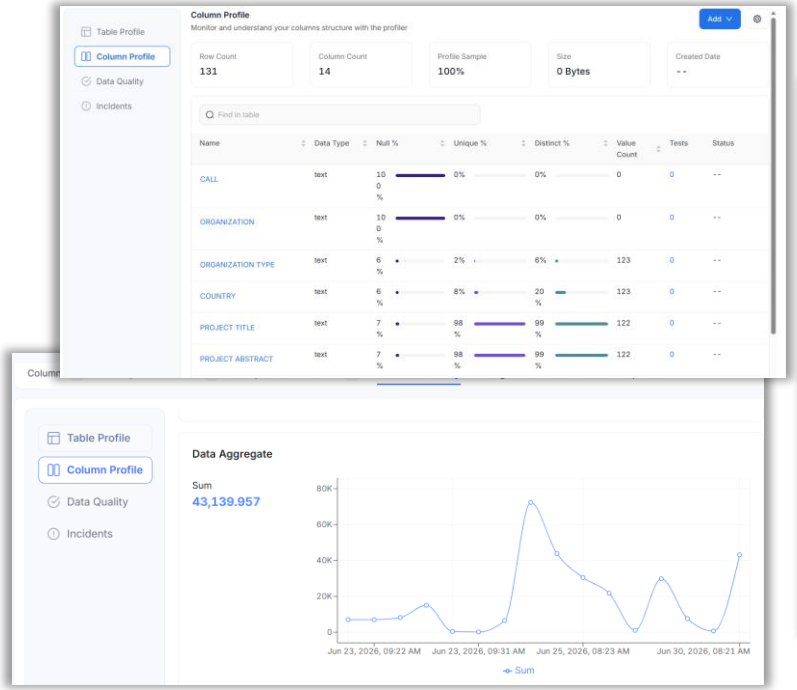
## What we hold

Catalogue of datasets and data products



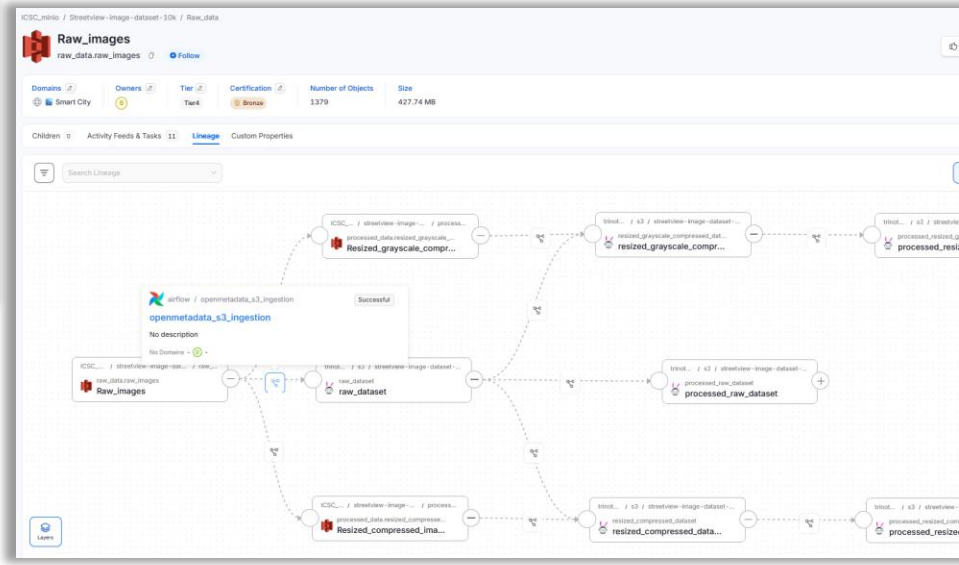

## How good it is

Column profiles and data observability



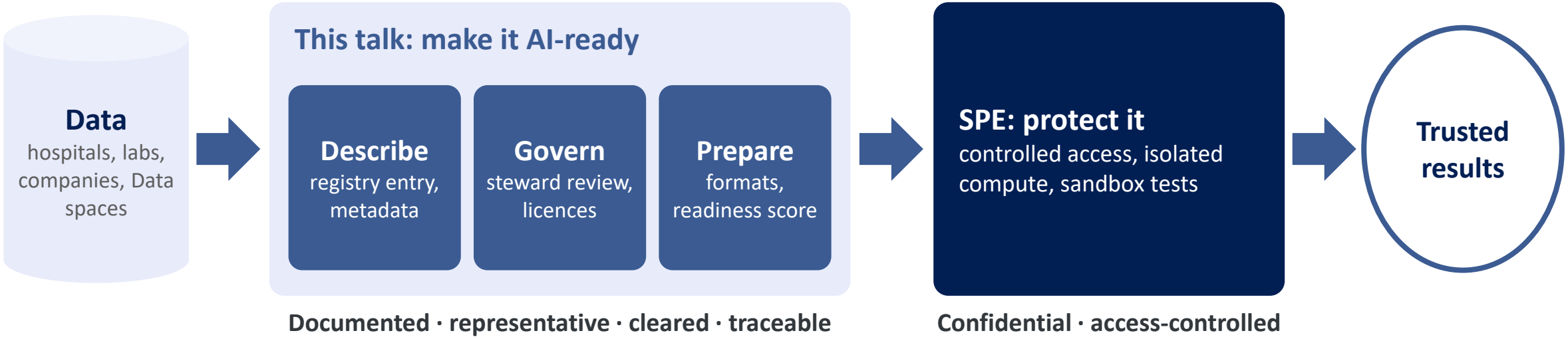

## Where it comes from

Lineage from source data to processed datasets




**Screens from OpenMetadata: one place to see what data we hold, how good it is and where it comes from.**

# Secure is not enough: make the data AI-ready



Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

# ICSC

Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

arnaud.ceol@supercomputing-icsc.it

https://www.supercomputing-icsc.it/

IT4LIA AI FACTORY

www.it4lia-aifactory.eu

IT4LIA AI Factory