#EUSpace

DaFab: AI Driven Metadata Generation Applied to Copernicus Data

EUSPA AI week 2026

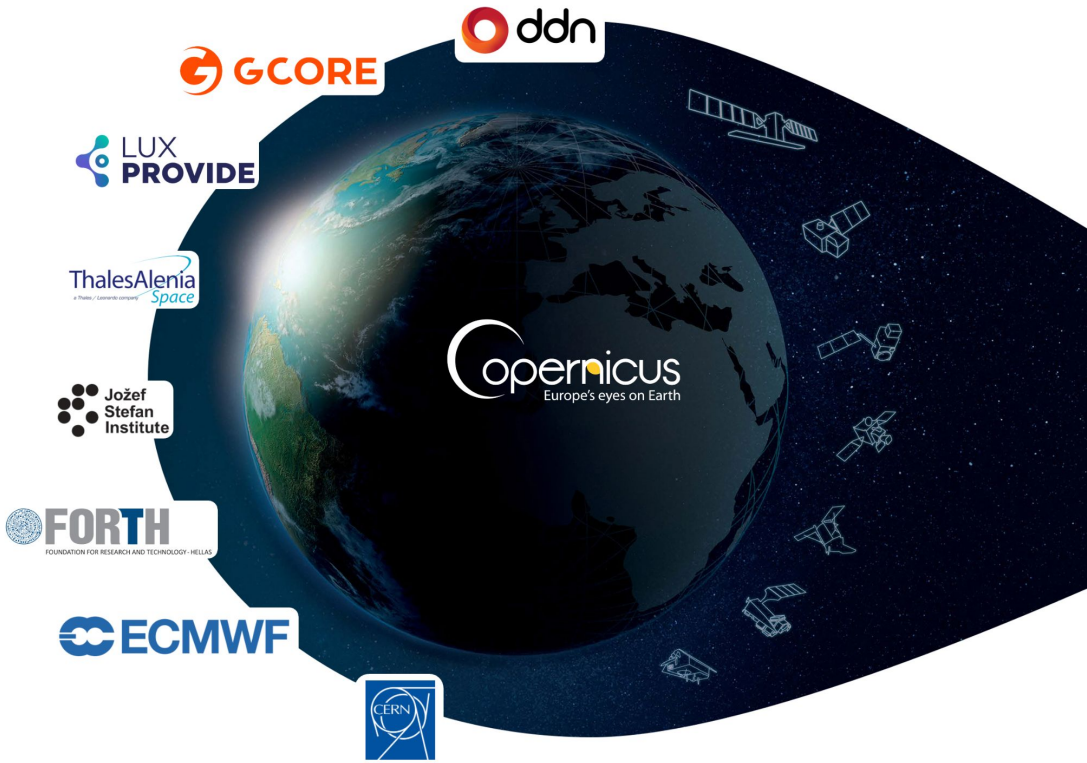
Jean-Thomas Acquaviva, DDN France
Farouk Mansouri, LuxProvide, Luxembourg

EUSPA European Union Agency for the Space Programme

DaFab HORIZON-EUSPA-2022-SPACE

Grant Number: 101128693

Getting EU data AI Ready



Earth Observation

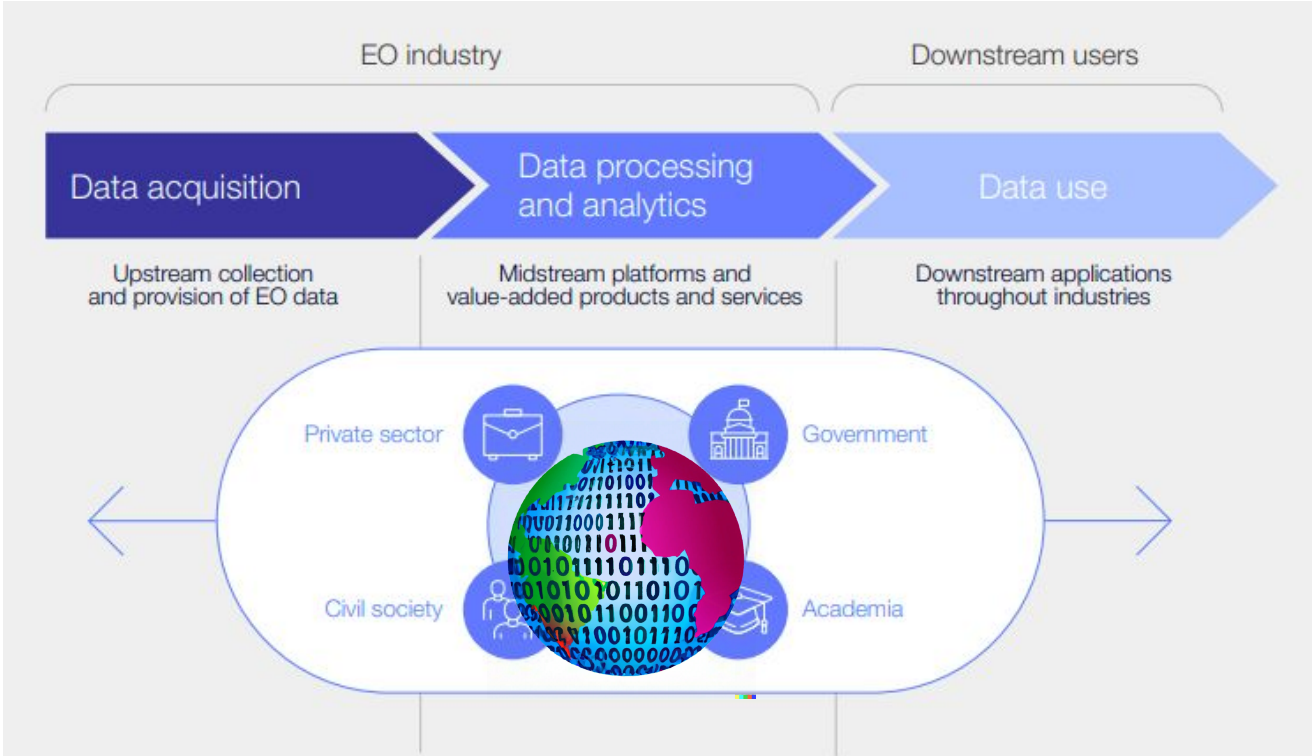
Data Intensive field with major economic and societal impact

- x3 in the next 5 years
- Saving 2 GT of carbon emission

Data are available it is now a computational problem

Amplifying the Global Value of Earth Observation
INSIGHT REPORT MAY 2024
World Economic Forum / Deloite

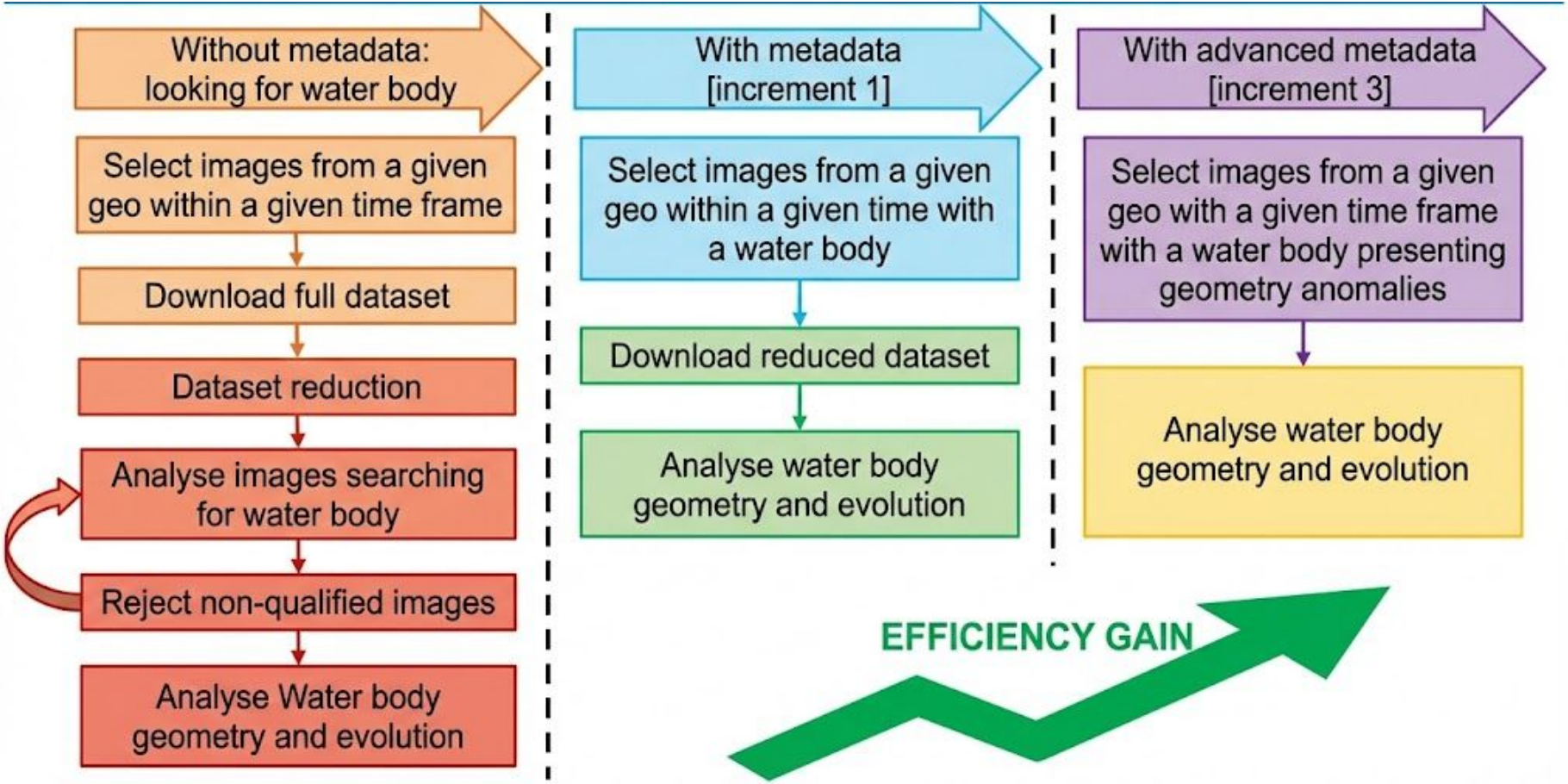
Value Chain: Data Broker from Acquisition to Exploitation



Copernicus DataHub: 47 PB

2 EU SMEs already using DaFAB data broker services

Unified Metadata Catalog as an Enabler



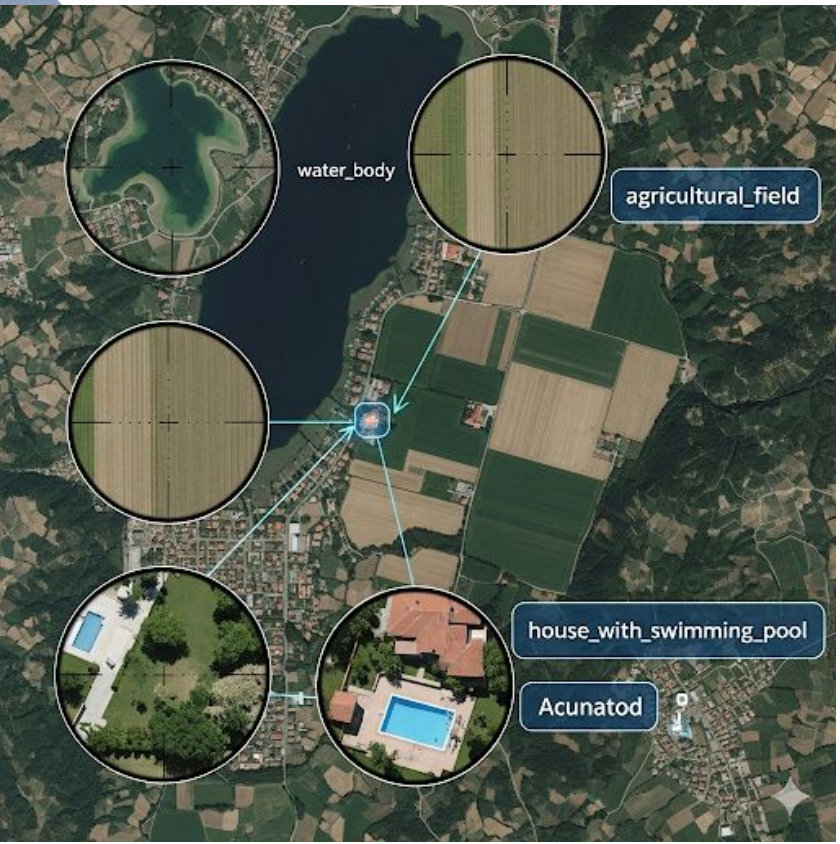
Unified Metadata Catalog Key Properties

- At a given **scale,** metadata becomes data
  - Requires tool: **Scientific Data Management**
- Ability to host a sufficiently large volume of information
  - images annotation can be verbose (GeoJSON)
- De-coupling of metadata catalog from data hub
  - Ingest annotation from distributed processes across multiple datahubs
  - Mandatory to manage public/private data sources
- Queryable
- Long-term support
  - [sept. 2024] RUCIO selected by the SKA project

SCIENTIFIC DATA MANAGEMENT
RUCIO

CERN

Metadata Standard and Ontologies



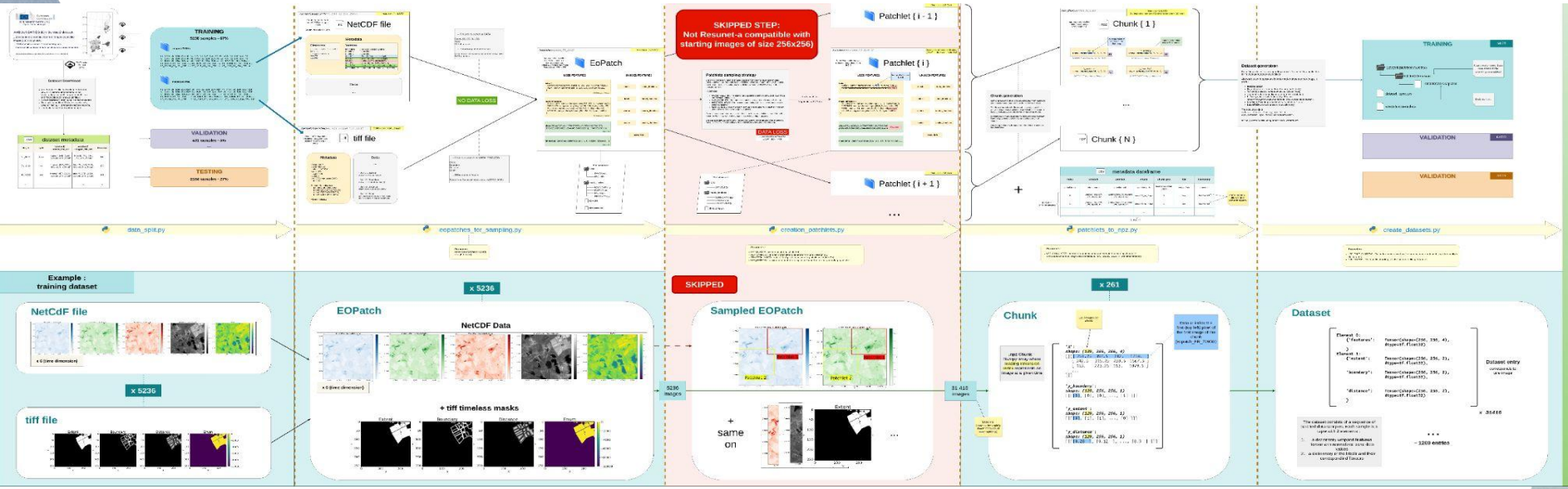
metadata format follow standard

- *GeoJSON*
- *STAC structure*

```
catalog.json  (Catalog)
├─ sensors/                (Catalog)
│  ├─ sentinel-2/          (Catalog)
│  │  ├─ sentinel-2_l2a.json (Collection)
│  │  │  ├─ 2024/06/       (Catalog partition)
│  │  │  │  └─ S2A_20240611…json (Item)
│  │  └─ sentinel-2_l1c.json …
│  └─ landsat-9/…
└─ derived/…
```

- Clients follow **self** → **child** → **item** to crawl downward.
- Every object also offers **root** to jump straight to the top.

Efficient Infrastructure for End-to-End AI Pipeline



On-going effort of multistage characterization *Workflow Roofline Model*

Performance = f(time, energy)

I/O pattern	idle Power (Watt)	Peak Power (Watt)	BW (GB/s)	IOPS
Write sequential 4M	430	564	40	10K
Write sequential 4KB	430	565	35	9260K
Write Random Write 4BK	430	470	1.2	320K
Write Single Shared File Random 4KB	430	470	0.03	7500
Read sequential 4M	430	510	48	12K
Read sequential 4k	430	510	25	6.5M
Read Random 4k	430	455	1.7	430K

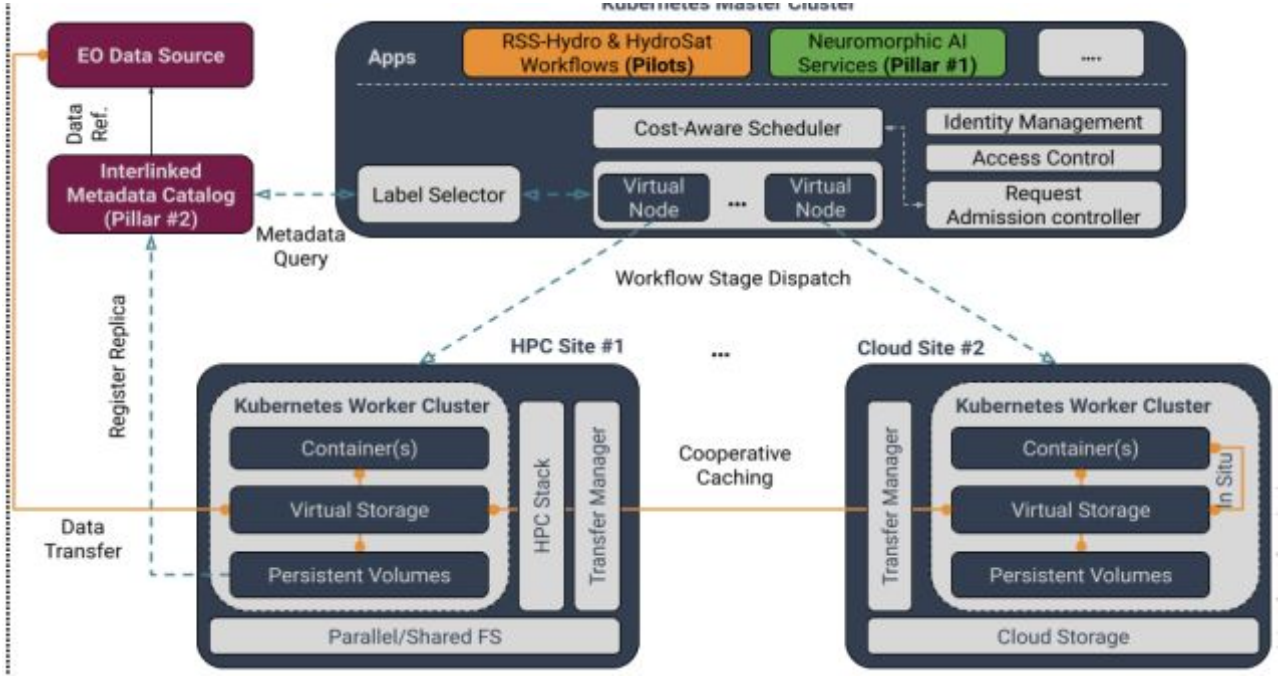
Metadata generation for full Copernicus archive

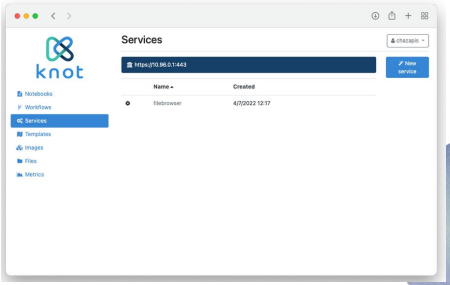
- 55 years of CPU
- 1+ year of A100 GPU

Metadata Locality HPC

- **Churn data at scale**: Efficient infrastructure to handle huge datasets
- **No Silos:** metadata compliant with standards
- **No Silos at scale**: workflows executed depending on data location
- **Foster Earth Observation Data Innovation**
  - Increase data informational density.
  - Reduce the volume of data to fetch by use case providers
  - Moving Copernicus archive over internet: between 3 and 50 GWh

Data Aware Workflow





Federated workflow management system for transparent access to distributed and multi-source data

Rational: AI pipeline require automated discovery of data from different sources with minimal data transfers

Work In Progress

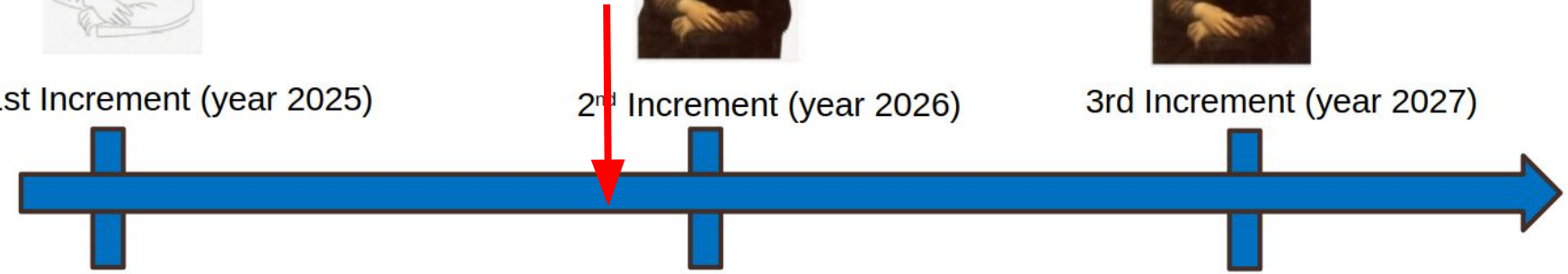
Agility is useful to build upon users feedback

19th Jan. 26

1st Increment (year 2025)

2nd Increment (year 2026)

3rd Increment (year 2027)



First demonstration

- One functionality
- One algorithm
- Two platforms
- First system workflow

Platforms connection

- Some functionalities
- Some algorithms
- All platforms
- Simple system workflow

Demonstrations

- All functionalities
- All algorithms
- All platforms
- Complex system workflows

Thank you!

#EUSpace

Linking space to user needs

Get in touch with us

www.euspa.europa.eu

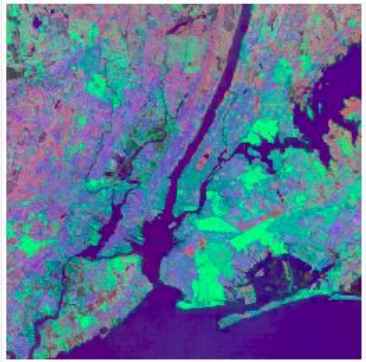
EUSPA EU4Space EUSPA space4eu EUSPA

EUSPA AI WEEK 2026

Position relatively to Google Earth July 30 announcement

End of July 2025, Google Earth and Alpha Earth release a 10mx10m 64 dimensions embeddings

Satellite Embedding V1



Dataset Availability

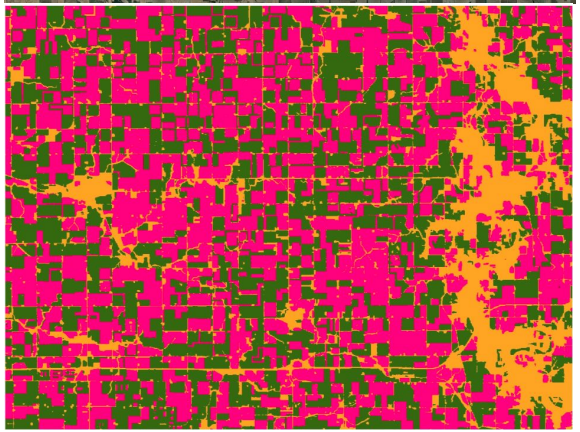
2017-01-01T00:00:00Z–2024-01-01T00:00:00Z

Dataset Provider

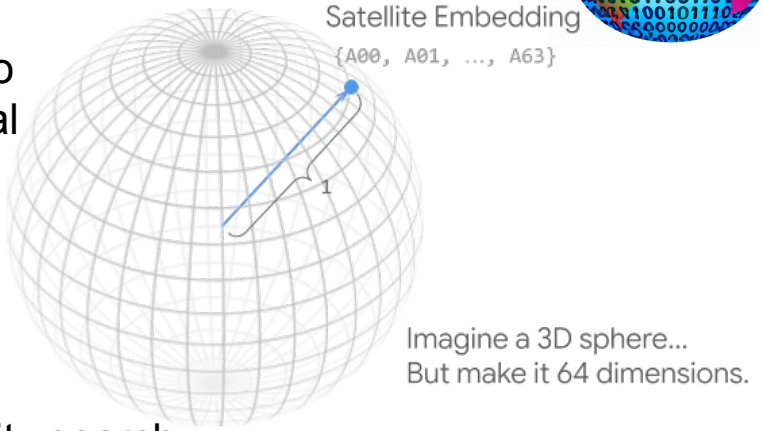
Google Earth Engine Google DeepMind

Earth Engine Snippet

```
ee.ImageCollection("GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL")
```



DaFab's Position to Google Earth July 30 announcement



Google claims "embedding layers are analysis-ready; no need for atmospheric correction, cloud masking, spectral transformations, speckle filtering, or other featurization techniques — just superior results at reduced effort and complexity"

DaFab does not produced embedding but **metadata**

embedding are model dependant

- usually stored in a vector database for fast similarity search

metadata are model independant

- usually stored in a structured database

DaFab can add embeddings as metadata in our Unified Metadata catalog

The ability from Google to process the full Copernicus database while we are at the stage of feasibility studies is worrying