



# EOOSC-Pillar

*Coordination and Harmonisation of National & Thematic Initiatives to support EOOSC*

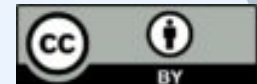
## Boosting the findability of data: a search engine for semantic concepts

José Manuel Domínguez

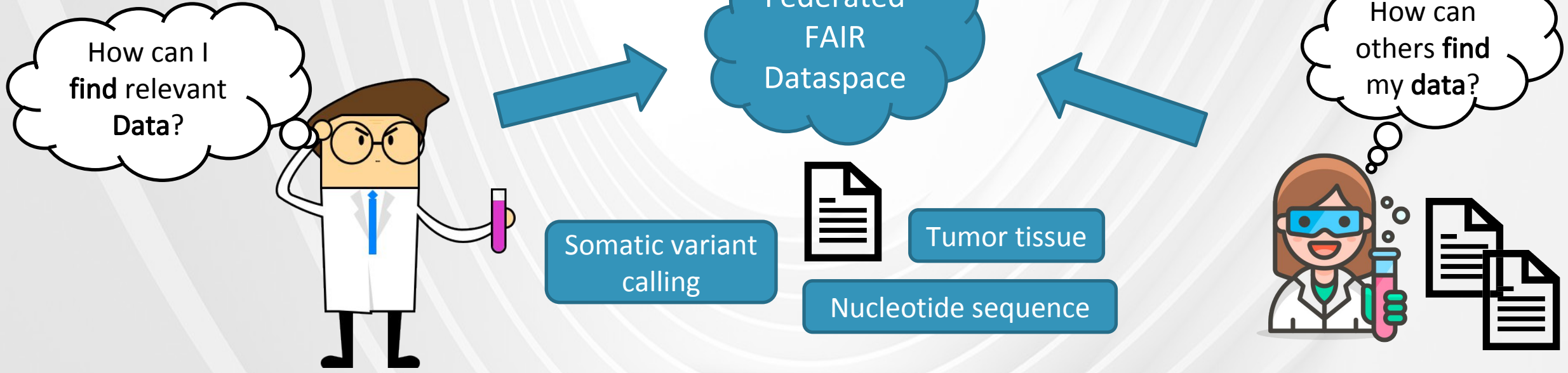


EOOSC-Pillar has received funding from the European Union's Horizon 2020 research and innovation Programme under Grant Agreement No. 857650.

This material by the EOOSC-Pillar is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)



# Use Case



## Situation:

Two scientists work in the same domain, but it is difficult to provide data in a way that the others can find it easily.

## Solution:

Federated FAIR Data Space

Datasets described with title, description and keywords (strings).

## Enhancement:

Use concepts from [semantic artefacts](#).

## Problem:

Find suitable concepts from existing semantic artefacts.

# Semantic artefacts

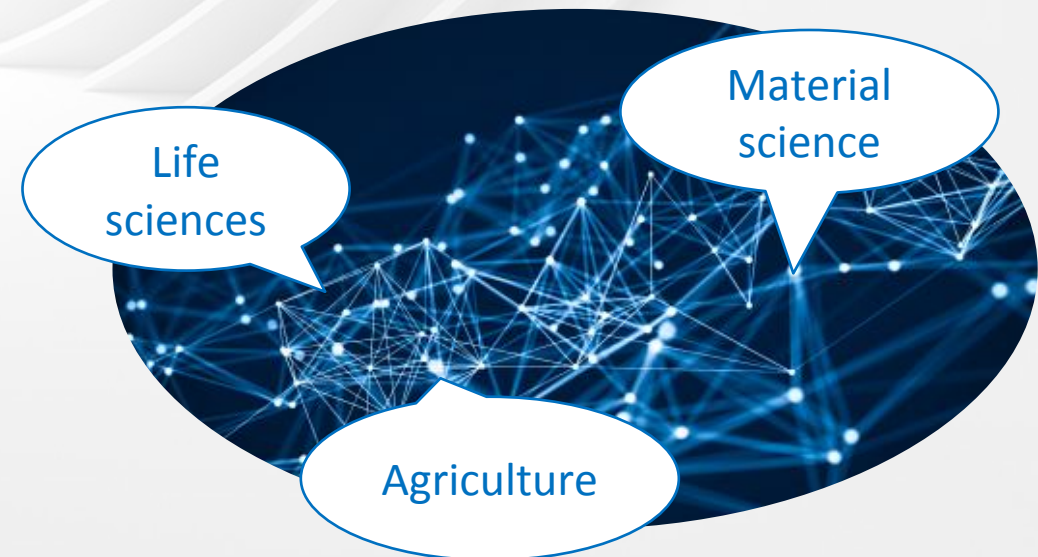
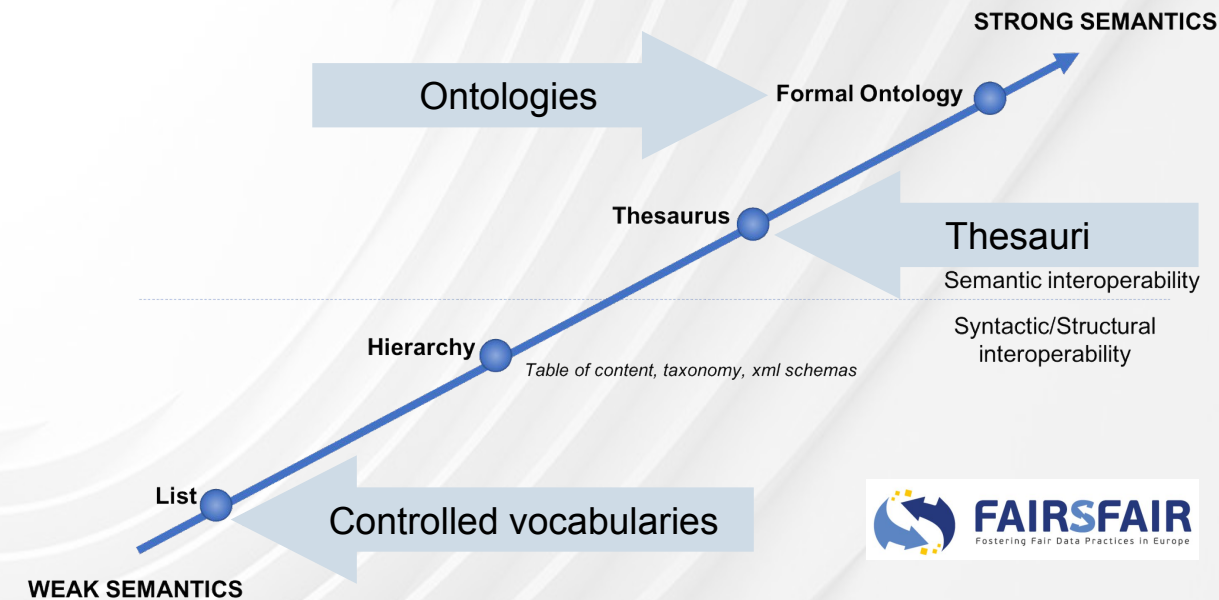
## What are semantic artefacts?

- Machine readable models of knowledge
- “Expressivity” of the knowledge can vary

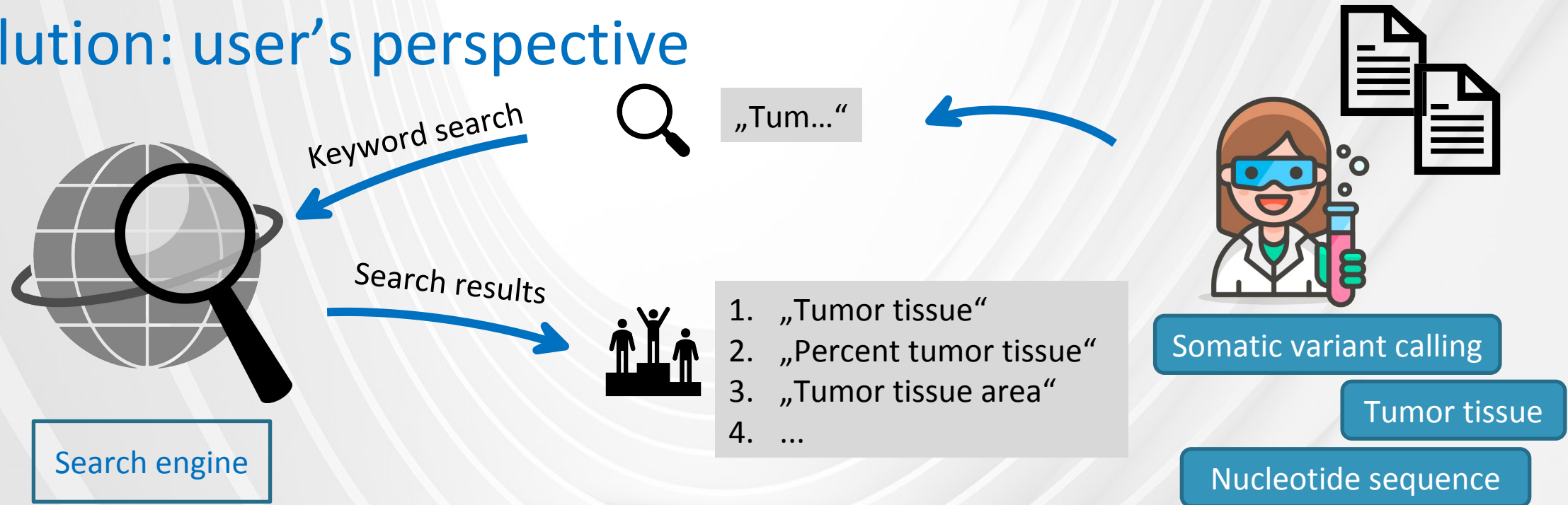
## What are semantic artefacts needed for?

- Capture domain knowledge
- “Speak” the same language
- Enable interoperability

→ Use vocabulary from Semantic Artefacts to annotate datasets



# Solution: user's perspective



Type the desired keyword in the search field

The search engine will immediately provide as-you-type suggestions

Select desired result from list of best matching concepts

# Solution: engineering perspective

## Landscape Analysis

Find out which SAs are relevant for EOSC-Pillar UCs

## EOSC-OntoPortal

Have all SAs accessible in repositories

## Harvesting concepts

Harvesting concepts from the various repositories

## Normalization

Common schema for concepts

## Search engine

Build an index and develop a scoring mechanism

## Integration in F2DS

Integrate search engine in F2DS

# Landscape Analysis

## Expert interviews

- Explain the needs
- Collect semantic artifacts

## Define evaluation criteria

- Machine-readability
- Availability in repositories
- Licenses

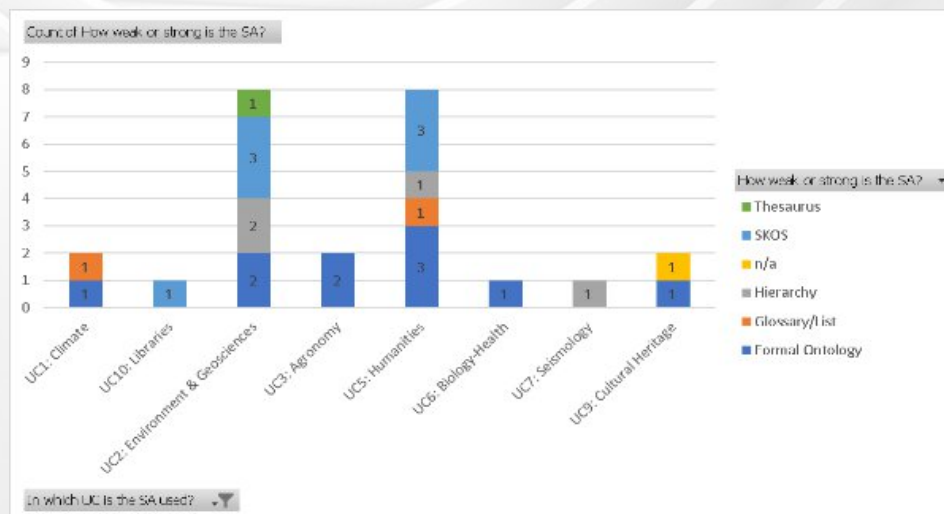
## Recommended SAs

- Recommend SAs to add to the search index

## Evaluate semantic artifacts

- Evaluate semantic artifacts based on defined criteria

UC1: Climate		
MatPortal*	<a href="https://matportal.org">https://matportal.org</a>	OntoPortal offers an actionable API to read directly from all the included ontologies
UC2: Environment & Geosciences		
The NERC (Natural Environment Research Council) Vocabulary Server (NVS)*	<a href="http://vocab.nerc.ac.uk/collection/">http://vocab.nerc.ac.uk/collection/</a>	While the NVS concepts are not available in plain rdf/ASCII the server exposes its resources with a dedicated API, both custom and SPARQL from where a harvesting might be possible
Semantic Sensor Network Ontology (SSN)	<a href="https://www.w3.org/TR/vocab-ssn/">https://www.w3.org/TR/vocab-ssn/</a>	RDF is available, not machine-actionable. Must first be stored in a machine-actionable environment to make the concepts accessible with an API
GEMET - General Multilingual Environmental Thesaurus	<a href="https://www.eionet.europa.eu/gemet/en/themes/">https://www.eionet.europa.eu/gemet/en/themes/</a>	RDF is available, not machine-actionable. Must first be stored in a machine-actionable environment to make the concepts accessible with an API
UC3: Agronomy		



# EOSC-Pillar OntoPortal

- OntoPortal: a repository service based on BioPortal
  - Publish, search and store semantic artefacts
- EOSC OntoPortal - a instance of OntoPortal
  - Populated with semantic artifacts that are not in a machine actionable environment
  - Status: > 60.000 concepts




 <https://ontoportal.eosc-pillar.eu/>

Welcome to EOSC Pillar Ontoportal, your ontology repository for your ontologies

---

Search for a class

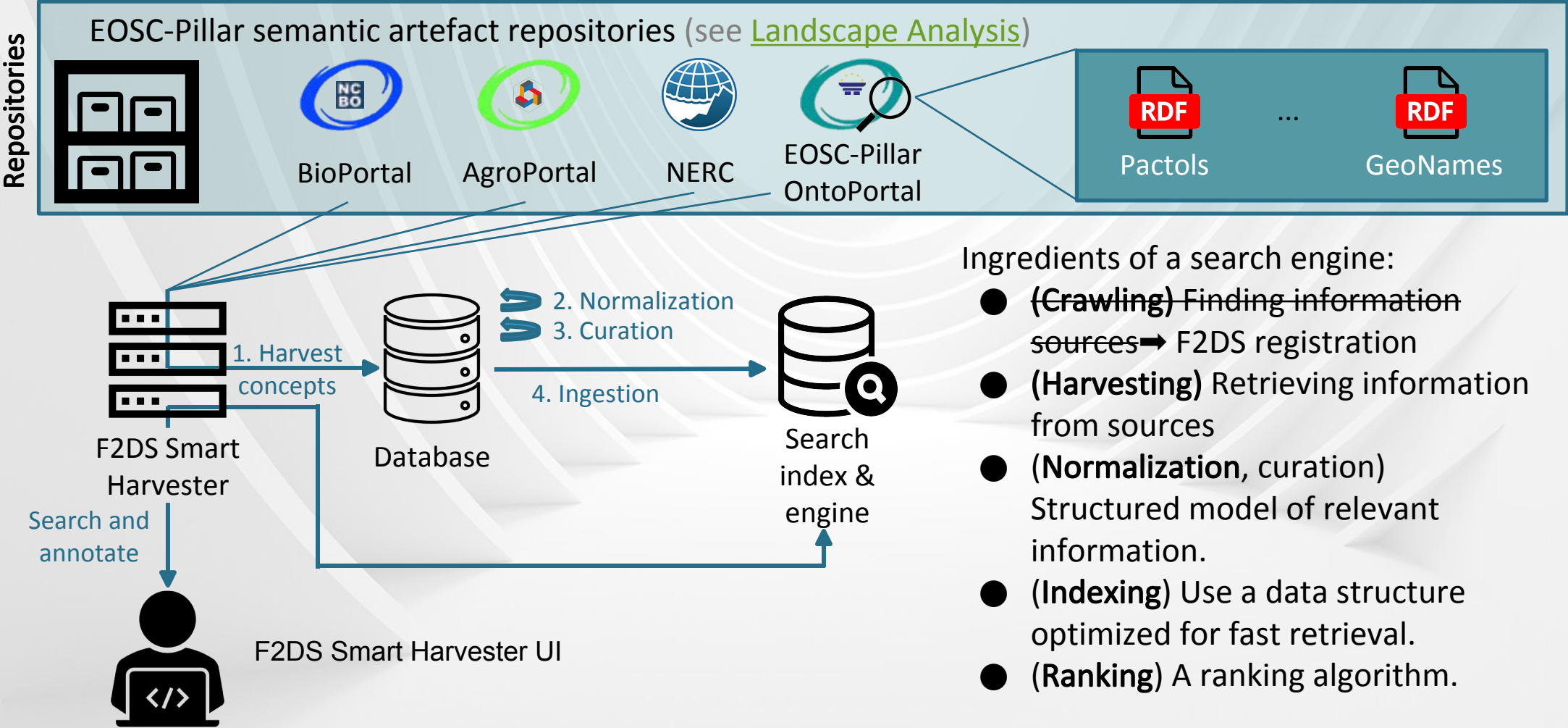
 

[Advanced Search](#)

Find an ontology

# Search engine: architecture



## Ingredients of a search engine:

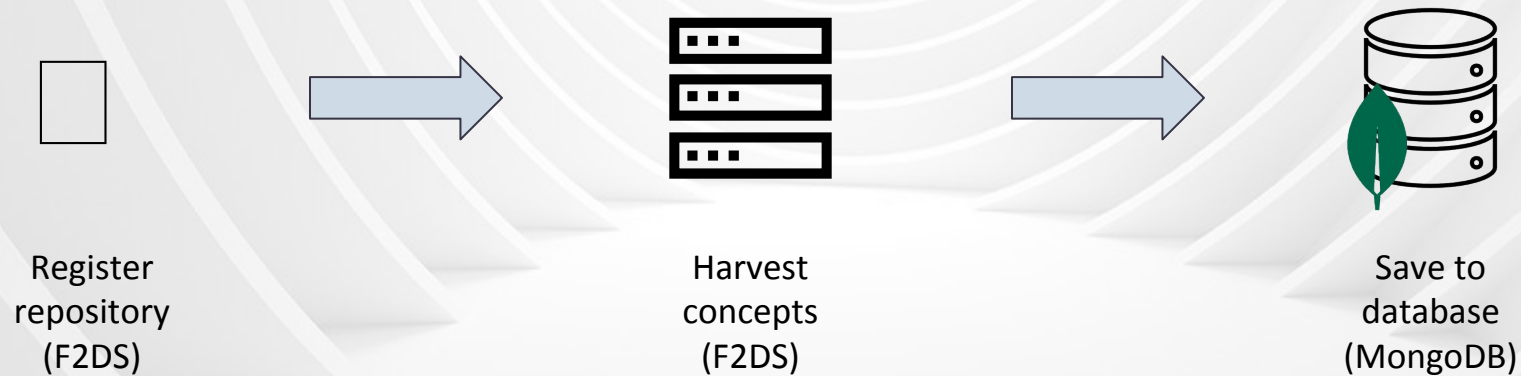
- **(Crawling)** Finding information sources → F2DS registration
- **(Harvesting)** Retrieving information from sources
- **(Normalization, curation)** Structured model of relevant information.
- **(Indexing)** Use a data structure optimized for fast retrieval.
- **(Ranking)** A ranking algorithm.



# Search engine: harvesting concepts

The F2DS Smart Harvester provides a graphical user interface to register concept repositories and retrieve their information.

(aimed at **administrators**)



# Search engine: harvesting concepts

## OpenAPI specification

```
{
  "openapi": "3.0.0",
  "info": {"title": "nerc", ..., "x-catalog-id": "1b0930f3-b0c5-415a-b5cb-2d612a70e873" },
  "servers": [{"url": "http://vocab.nerc.ac.uk"}],
  "paths": {"/scheme":
    {"get": {"tags": ["dataset"], "summary": "",
      "parameters": [
        {"in": "query",
          "name": "_profile",
          "schema": {"type": "string", "default": "nvs"},
          "required": false,
          "description": ""},
        {...}
      ]
    },
    "responses": {"200": {...}, ...}
  }
},
  [...]
}
```

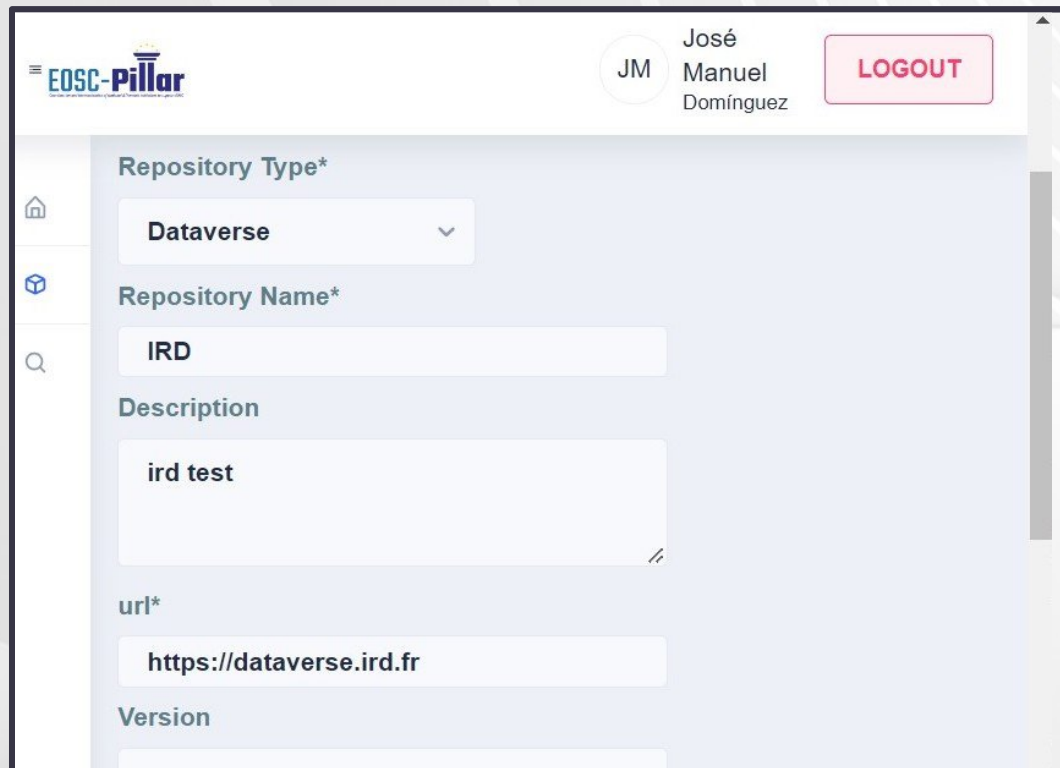


Human and computer-readable specification  
for HTTP APIs.

# Search engine: harvesting concepts

## F2DS Registration

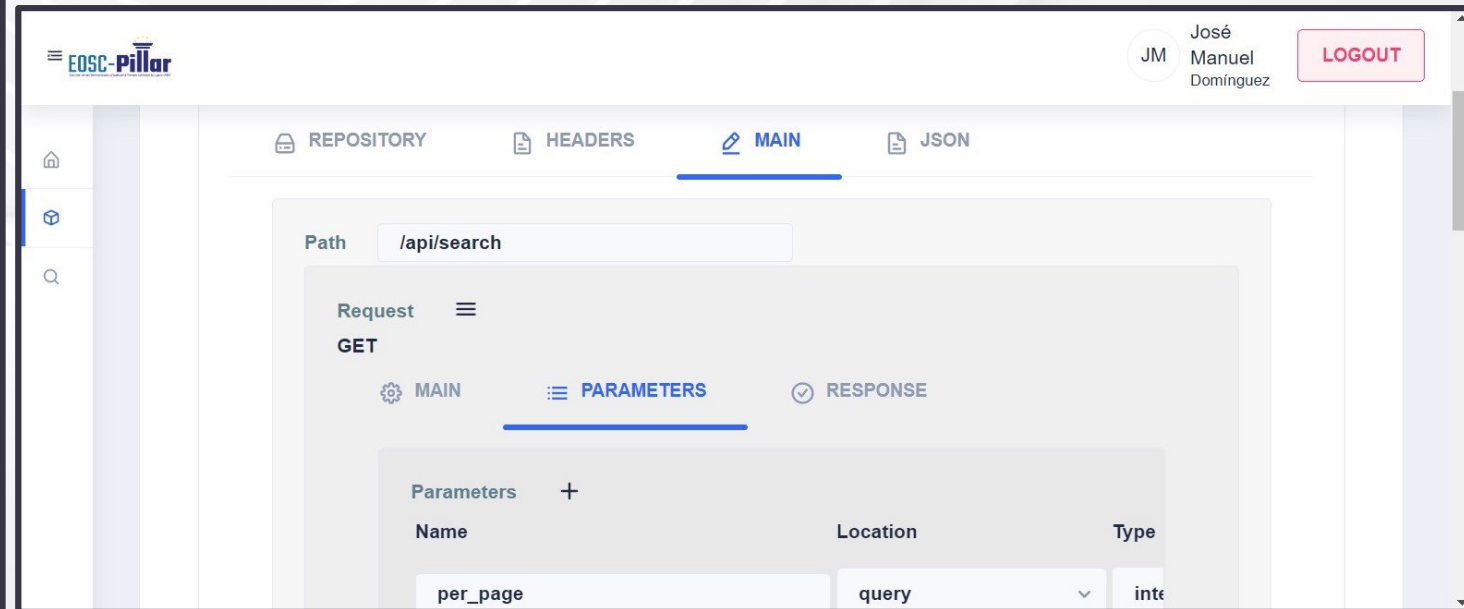
Graphical user interface for creating the API description



The screenshot shows the registration form for a repository. The user is logged in as José Manuel Domínguez. The form includes the following fields:

- Repository Type\***: A dropdown menu with "Dataverse" selected.
- Repository Name\***: A text input field containing "IRD".
- Description**: A text area containing "ird test".
- url\***: A text input field containing "https://dataverse.ird.fr".
- Version**: A text input field (empty).

A "LOGOUT" button is visible in the top right corner.



The screenshot shows the API description editor for the path "/api/search". The user is logged in as José Manuel Domínguez. The interface includes the following elements:

- Navigation tabs**: REPOSITORY, HEADERS, MAIN (selected), JSON.
- Path**: A text input field containing "/api/search".
- Request**: A dropdown menu with "GET" selected.
- Configuration tabs**: MAIN (selected), PARAMETERS, RESPONSE.
- Parameters table**: A table with columns "Name", "Location", and "Type".

Name	Location	Type
per_page	query	inte

A "LOGOUT" button is visible in the top right corner.

# Search engine: harvesting concepts

## Harvested concept

```
{
  "repositoryId": "1b0930f3-b0c5-415a-b5cb-2d612a70e873",
  "url":
  "http://vocab.nerc.ac.uk/collection/GS8/current?_profile=nvs&_mediatype=application/ld+json",
  "document": {
    "identifier": "SDN:GS8::PASR",
    "note": {"@value":"accepted", "@language":"en"},
    "@type":"skos:Concept",
    ...,
    "prefLabel":{"@value":"Passive seismic refraction", "@language":"en"},
    ...,
    "definition":{"
      "@value":"The elucidation of geological structure by quantifying the refraction of waves
from naturally occurring low frequency earth movements (e.g. earthquakes) by sub-surface layers .",
      "@language":"en"},
      "@id":"http://vocab.nerc.ac.uk/collection/GS8/current/PASR/",
      "dc:identifier": "SDN:GS8::PASR"
    },
    "_class":"com.smartharvester.model.Ontologies"
  }
}
```

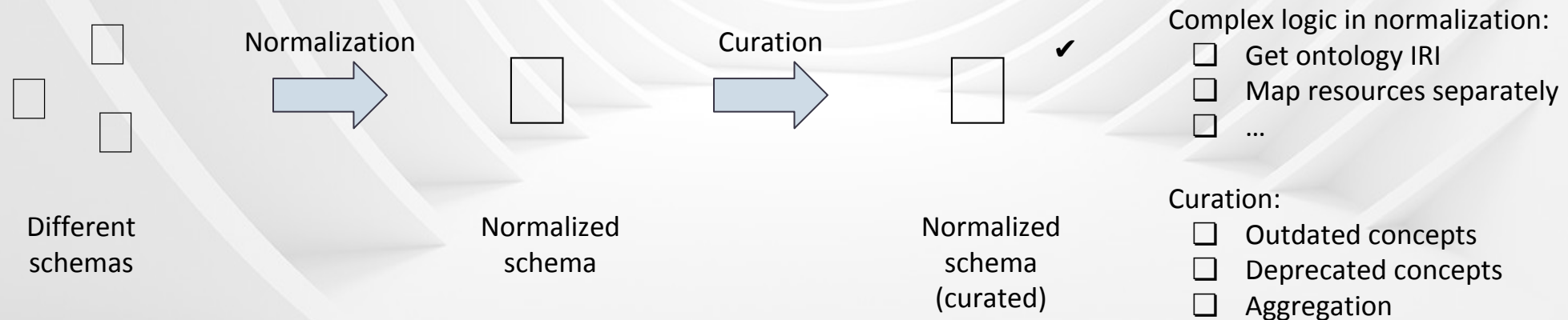
⚠ Different schema for each concept repository!

# Search engine: normalization and curation



Goals:

- Filter and structure information
- Enforce logical constraints (e.g. no deprecated concepts)



# Search engine: normalization and curation

## Normalized schema

Adapted from previous work<sup>1</sup>

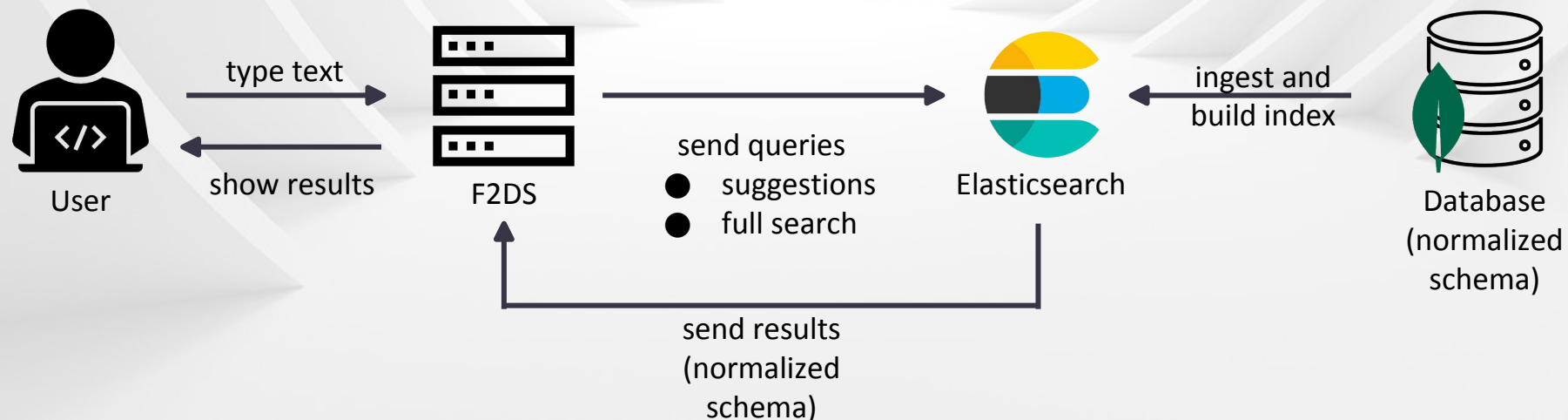
```
{  
  "iri": "http://edamontology.org/data\_0872",  
  "label": "Phylogenetic tree",  
  "description": "A phylogenetic tree is usually constructed from a set of sequences from which an alignment (or data matrix) is calculated. See also 'Phylogenetic tree image'.",  
  "synonyms": ["Phylogeny"],  
  "domains": [],  
  "resource iri": "https://identifiers.org/ito:ontology",  
  "harvested from": "https://data.bioontology.org/ontologies/ITO",  
  "ui link":  
  "http://bioportal.bioontology.org/ontologies/ITO?p=classes&conceptid=http%3A%2F%2Fedamontology.org%2Fdata\_0872",  
  "resource name": "Intelligence Task Ontology",  
  "resource acronym": "ITO",  
  "resource date": "2022-01-13T00:00:00.000Z",  
  "resource version": "1.01 (PWC export dated 2021-06-16)",  
  "resources reusing": ["http://edamontology.org"],  
  "resources reusing acronyms": ["EDAM"]  
}
```

<sup>1</sup> Goldfarb, Doron & Le Franc, Yann. (2017). Enhancing the Discoverability and Interoperability of Multi-disciplinary Semantic Repositories.

# Search index and engine

Built using  elasticsearch

- Data structures for **very fast** retrieval (index)
- Distributed architecture (several machines)
- Query language (select fields to be searched)
- Built-in scoring algorithms (give importance to each field)



# Search engine: integration into F2DS smart harvester

- Reuses the **keywords** of datasets to **automatically search and assign concepts from semantic artefacts**.
  - Fine-tuning: suggestions can be discarded or a different result chosen.
- Annotate with **additional concepts** if desired.

Demo

next presentation



# EOSC-Pillar

Coordination and Harmonisation of National & Thematic Initiatives to support EOSC

# Thank you!

# Get in touch with us!



[www.eosc-pillar.eu](http://www.eosc-pillar.eu)



[@EoscPillar](https://twitter.com/EoscPillar)



[/company/eosc-pillar](https://www.linkedin.com/company/eosc-pillar)



EOSC-Pillar has received funding from the European Union's Horizon 2020 research and innovation Programme under Grant Agreement No. 857650.  
This material by the EOSC-Pillar Consortium is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

