



D7.4

The ENVRI-FAIR Knowledge Base V2

Work Package	WP7
Lead partner	UvA
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30 June 2023
Submission date	30 June 2023

Deliverable abstract

The overarching goal of ENVRI-FAIR is for all participating RIs to improve their FAIRness and prepare the connection of their data repositories and services to the European Open Science Cloud. With the development of FAIR implementations from the participating RIs and integrated services among the environmental subdomains, these data and services will be brought together at a higher level (for the entire cluster), providing more efficient services for researchers and policy makers.

This deliverable presents new developments to the knowledge base, initially presented in the ENVRI-FAIR Deliverable 7.3. The new developments focus on the knowledge base search engine. We introduce a new automated pipeline to index online resources: web pages, datasets, and computational notebooks. We also present a complete redesign of the search engine and interface. These new features were deployed in a cloud environment, making the service more performant and reliable. We also discuss the integration with the ENVRI community through the ENVRI-Hub and show a demonstration of the new search engine. By allowing users to find resources from ENVRI and its RIs, the knowledge base search engine improves interoperability.



DELIVERY SLIP

	Name	Partner Organisation	Date
Main Author	Gabriel Pelouze	UvA/LifeWatch ERIC	27 June 2023
Contributing Authors	Markus Stocker Barbara Magagna Zhiming Zhao	TIB EAA UvA	
Reviewer(s)	Keith Jeffery Peter Thijsse Christian Pichot	UKRI/EPOS MARIS/SeaDataNet INRA/ANAE	9 June 2023 26 June 2023 26 June 2023
Approver	Andreas Petzold	FZJ	30 June 2023

DELIVERY LOG

Issue	Date	Comment	Author
V 1.0	2 June 2023	First Draft	Gabriel Pelouze
V 2.0	9 June 2023	Second Draft	Gabriel Pelouze
	2 June 2023	Comments from Reviewer 1	
	26 June 2023	Comments from Reviewer 2	
	26 June 2023	Comments from Reviewer 3	
V 3.0	26 June 2023	Final version	Gabriel Pelouze

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the Project Manager at manager@envri-fair.eu.

GLOSSARY

A relevant project glossary is included in Appendix A. The latest version of the master list of the glossary is available at <http://doi.org/10.5281/zenodo.4471374>.

PROJECT SUMMARY

ENVRI-FAIR is the connection of the ESFRI Cluster of Environmental Research Infrastructures (ENVRI) to the European Open Science Cloud (EOSC). Participating research infrastructures (RI) of the environmental domain cover the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems and thus the Earth system in its full complexity.

The overarching goal is that at the end of the proposed project, all participating RIs have built a set of FAIR data services which enhances the efficiency and productivity of researchers, supports innovation, enables data- and knowledge-based decisions, and connects the ENVRI Cluster to the EOSC.

This goal is reached by: (1) well defined community policies and standards on all steps of the data life cycle, aligned with the wider European policies, as well as with international developments; (2) each participating RI will have sustainable, transparent, and auditable data services, for each step of data life cycle, compliant to the FAIR principles. (3) the focus of the proposed work is put on the implementation of prototypes for testing pre-production services at each RI; the catalogue of prepared services is defined for each RI independently, depending on the maturity of the involved RIs; (4) the complete set of thematic data services and tools provided by the ENVRI cluster is exposed under the EOSC catalogue of services.

TABLE OF CONTENTS

D7.4 – The ENVRI-FAIR Knowledge Base V2.....	4
1 Introduction.....	4
2 Development activities.....	4
2.1 Automated online content ingestion pipeline	4
2.1.1 Web pages.....	4
2.1.2 Datasets.....	5
2.1.3 Computational notebooks	6
2.2 Index storage and search engine.....	6
2.3 User interface redesign.....	7
3 Knowledge Base deployment.....	7
4 Integration with the ENVRI-Hub.....	9
5 Demonstration.....	10
5.1 Environmental indicators overview.....	10
5.2 Essential climate variables datasets.....	12
5.3 Science demonstrator notebooks	14
6 Summary	15
7 References.....	15
8 Appendix 1: Glossary.....	16

D7.4 – The ENVRI-FAIR Knowledge Base V2

1 Introduction

This document describes the progress made by WP7 on the new developments to the ENVRI Knowledge Base (KB) for RI service interoperation and competence. The goal of the knowledge base is to facilitate collaboration and information sharing among different users active within the ENVRI community, such as RI developers, data managers, and users. This is made possible by documenting technical practices, identifying common requirements, and allowing users to search and analyse existing solutions for interoperability challenges, as well as knowledge and resources across RIs.

The architecture of the knowledge base was chosen based on a requirements analysis and state-of-the art review presented in D7.3. In the prototype presented in D7.3, the WP7 team chose Ontowiki to manage the RDF triples, and open semantic search to develop the search engine for the knowledge base. A number of tools were developed for ingesting knowledge from specific sources. The detailed architecture of the knowledge base can be found in D7.3 and in [1].

In the current document, WP7 focuses on the new developments carried out since D7.3. These include new developments to the ENVRI Knowledge Base search engine (Section 2), a deployment in the cloud (Section 3), and integration efforts with the community (Section 4). A demonstration of the new features of the search engine is provided (Section 5).

2 Development activities

2.1 Automated online content ingestion pipeline

The ENVRI Knowledge Base Search Engine (KBSE) relies on the ingestion of resources, mainly from the RIs. Since D7.3, a new ingestion pipeline has been developed for these documents¹. In this section, we describe the new developments to the ingestion pipelines of web pages, datasets, and Jupyter notebooks.

The system uses the Extract-Transform-Load (ETL) process, where a resource is, step 1, gathered from an online source (e.g. a domain-specific data repository), then, step2, transformed into a metadata document following an internal schema, and, step3, loaded into Elasticsearch. During step 2, contextual information is extracted from the resource and stored into structured fields (e.g. topics, locations, or organisations). This step is crucial to create rich metadata that lead to more relevant search results.

2.1.1 Web pages

The ENVRI website and those of all RIs are crawled and indexed. Each page of the websites is retrieved and parsed. The following metadata are saved:

- page title,
- page URL,
- page textual content,
- metadata of documents and images in the page,
- research infrastructure (name, website, domain(s), etc.).

Furthermore, the following context metadata are extracted from the page using natural language processing:

- topics,
- dates,
- locations,
- organisations,
- people.

¹ <https://github.com/QCDIS/kb-indexer>

We used a pre-trained natural language processing pipeline from the spaCy library², to extract and label entities from the textual contents of the web pages. The raw and enriched metadata for each page are combined into a single index document and loaded into Elasticsearch.

2.1.2 Datasets

The ingestion of metadata of dataset records is more complex, because of the variety of sources. Despite standardisation efforts, different dataset repositories use different API formats to expose metadata records. In addition, metadata records themselves have different metadata schemas, which have to be mapped manually onto a common schema before ingestion into Elasticsearch that is used for indexing.

The ingestion pipeline for dataset metadata follows the approach developed by [2], allowing to enrich the metadata with contextual information. The pipeline for ingesting dataset metadata into the search engine follows the following steps:

1. Extraction
 - a. Crawling: a list of metadata records is retrieved from the dataset repository.
 - b. Metadata extraction: metadata records are retrieved for each metadata entry.

For this step, we rely as much as possible on machine-to-machine interfaces from the RIs metadata catalogues (e.g., SPARQL or REST APIs).
2. Mapping
 - a. Metadata mapping: metadata records are mapped onto an internal metadata schema. This requires manually creating a mapping between the original schema of the record, and the internal schema of the Knowledge Base. To simplify the creation of such mappings, we adopted an internal schema similar to the one used by EUDAT-B2FIND³, allowing to reuse the metadata mappings developed for B2FIND⁴. The internal dataset metadata schema used for the ENVRI KBSE can be found at: https://github.com/QCDIS/kb-indexer/blob/v0.6/indexers/dataset/data_sources/metadata_schema.json
The transformation of metadata for assets of the ENVRI RIs was established by TF1 of WP5, and uses EPOS-DCAT-AP as an intermediate schema. This transformation was still under development when the KBSE prototype was made, but will be adopted in the future.
 - b. The metadata is enriched with contextual information:
 - i. A list of potential topics is extracted from selected textual fields of the metadata record, using Latent Dirichlet Allocation (LDA) [2].
 - ii. Domain specific keywords are extracted from the list of potential topics. This is done by calculating the cosine similarity between the potential topics, and a list of essential variables from the sub-domain(s) of the RI from which the metadata record was retrieved.
3. Loading: the converted metadata records are loaded into Elasticsearch.

The dataset ingestion pipeline was refactored to follow a modular structure. This allows in particular to decouple the record extraction (specific on the protocols exposed by different dataset repositories) from the metadata mapping (specific to the metadata schema used in different repositories). This modular approach will make it easier to ingest records for more dataset repositories in the future.

A list of the currently implemented dataset repositories, and the number of records harvested into the ENVRI KBSE, are shown in Table 1. Currently, datasets are harvested from five repositories. This will be extended to all ENVRI RIs as part of an ongoing development effort, which could be part of ENVRI-Hub NEXT (currently a proposal).

² <https://spacy.io>

³ <https://b2find.eudat.eu>

⁴ <https://github.com/EUDAT-B2FIND/md-ingestion>

Table 1: Dataset repositories indexed in the ENVRI Knowledge Base

Dataset repository	API endpoint	Harvested records
DiSSCo	https://sandbox.dissco.tech/api/swagger-ui/index.html	418 k
ICOS	https://meta.icos-cp.eu/sparql/	124 k
SeaDataNet CDI	https://cdi.seadatanet.org/sparql/	240 k
SeaDataNet EDMED	https://edmed.seadatanet.org/sparql/	4.4 k
SIOS	https://sios.csw.met.no/csw/	28 k
Total		576 k

2.1.3 Computational notebooks

In addition to web pages and datasets, the KBSE allows users to search for computational notebooks. ENVRI science demonstrator notebooks are added manually to the index. In addition, notebooks relevant to the ENVRI topics are harvested from public repositories. The automatic ingestion pipeline uses the following steps:

1. Extraction:
 - a. Search public repositories (GitHub and Kaggle) for notebooks, using the repositories' search APIs. To get notebooks relevant to the ENVRI community, we use Essential Climate-, Ocean-, and Biodiversity variables as query terms. The list of query terms used to harvest notebooks can be found on GitHub⁵. The search is restricted to Jupyter notebooks using Python.
 - b. Retrieve metadata about the notebooks.
2. Mapping: the metadata are mapped onto the following fields:
 - a. Title
 - b. Description
 - c. Upvotes (or any other measure of community score, such as stars or forks)
 - d. URL
3. Loading: the converted metadata records are loaded into a dedicated index in Elasticsearch.

Currently, the pipeline has harvested metadata from about 52k notebooks (32k from GitHub and 20k from Kaggle).

Currently, the indexing and search of notebooks in the KBSE is restricted to the metadata. However, notebooks contain rich contextual information in the form of code and explanatory text. These can be used to significantly increase the relevance of search results. Indexing and search over the content of notebook is a topic of active research at UvA (see e.g. [3]). This research is highly relevant to the ENVRI Knowledge Base and will be used in the future to improve notebook search results.

2.2 Index storage and search engine

In the ENVRI Knowledge Base Search Engine (KBSE) prototype presented in D7.3, the indexed documents were managed by Open Semantic Search. In the current prototype, we switched to Elasticsearch⁶, which provides a few advantages: better scalability in a cloud ecosystem, better documentation, and community support. Furthermore, Elasticsearch provides results aggregation, allowing to implement faceted search in the KBSE.

⁵ https://github.com/QCDIS/kb-indexer/blob/main/indexers/notebook/data_sources/envri_queries.csv

⁶ <https://www.elastic.co/elasticsearch/>

2.3 User interface redesign

Along with switching to Elasticsearch, the user interface⁷ was entirely redesigned. The interface consists of a backend developed with Django, and a frontend developed with Bootstrap. A screenshot of the new interface is shown in Figure 1. This interface is separate from the ENVRI-Hub catalogue, in an effort to provide a free-text search experience targeted at human users, supporting segmented search, as well as more resource types than the catalogue. Integration with the ENVRI-Hub is achieved by displaying results from the KBSE on the ENVRI-Hub's starting page.

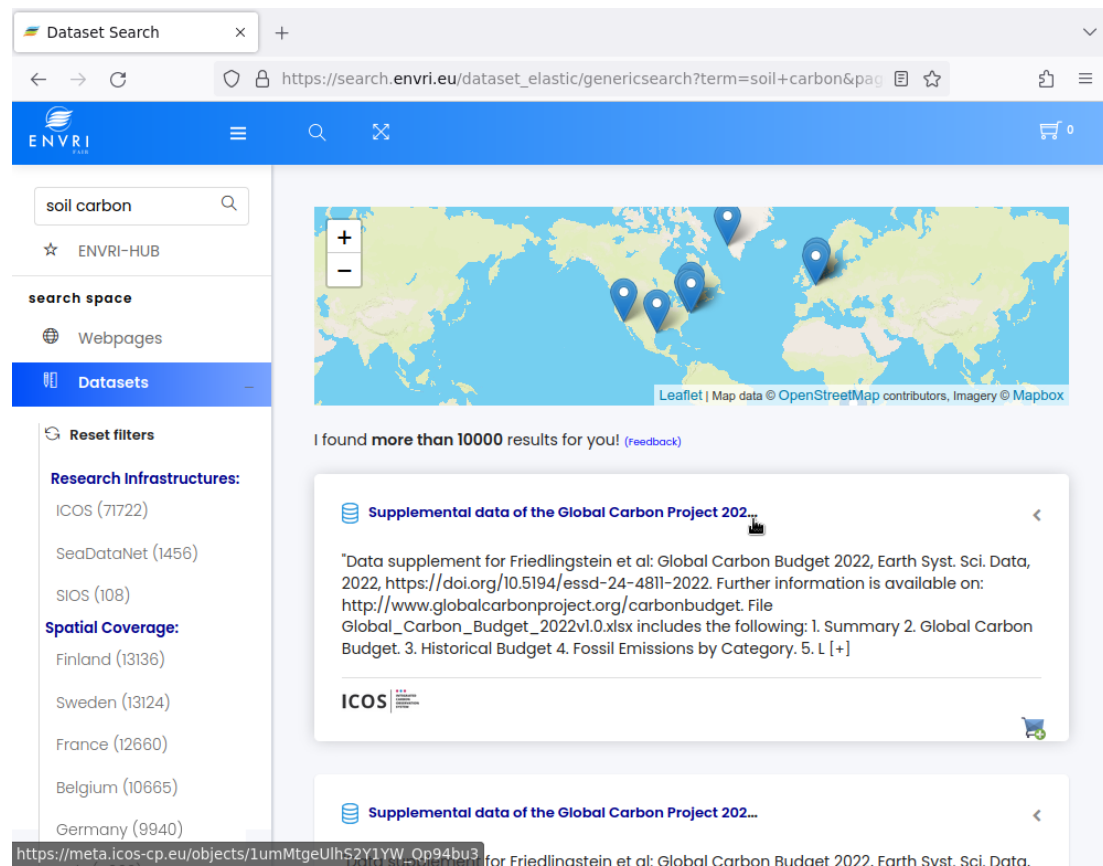


Figure 1: Screenshot of dataset search results in the new interface of the KBSE

3 Knowledge Base deployment

The KBSE is deployed in a cloud environment following modern DevOps practices. A deployment is shown in Figure 2. The search engine is deployed on two 4-node clusters managed by Kubernetes. The main cluster is provided by LifeWatch ERIC and hosted in Roubaix (France). The second cluster is hosted at UvA (The Netherlands). Incoming traffic goes through a dedicated VM running HAProxy. During normal operations, traffic is forwarded to the LifeWatch cluster. In case of failure, the load balancer can be reconfigured to point to the UvA cluster.

⁷ <https://github.com/QCDIS/KMS-generic>

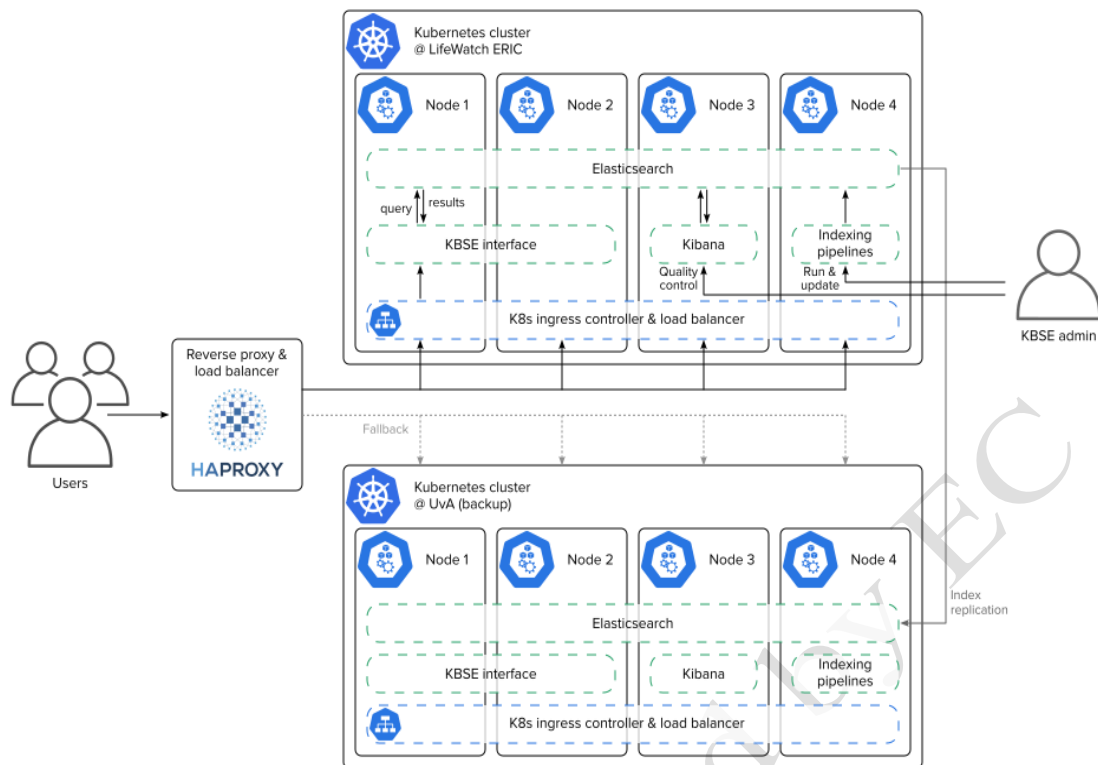


Figure 2: Deployment diagram of the KBSE

The services that make up the KBSE are containerised using Docker and deployed on both clusters:

- Elasticsearch stores indexed documents. Internally, it consists of four types of services: master, coordinating, indexing and data. We deploy two instances of each service, which are distributed across the cluster nodes by Kubernetes. To ensure data integrity, the indexes are replicated across both clusters.
- The KBSE interface⁸ serves the web pages, getting search results from Elasticsearch.
- The indexing pipelines⁹ are deployed alongside the database and search interface, and are operated by the KBSE administrator to index new resources, or update existing records.
- The Kibana¹⁰ dashboard is used for monitoring and quality control of the indexes stored into Elasticsearch.

For the KBSE interface and the indexing pipelines, new images are built from source and published to Docker Hub using GitHub actions. All resources are described and deployed on Kubernetes using Helm Charts¹¹.

This cloud deployment provides better flexibility in handling variable loads. For instance, the Elasticsearch database and search interface can be scaled dynamically to respond to peaks in demand. It also improves data resilience through duplication across two geographic locations, and service availability. In case of failure, traffic can be redirected to the backup cluster through the load balancer. Because data and services on the backup cluster are kept ready, downtime is limited to a few minutes. Finally, the whole service infrastructure and configuration can easily be recreated in a new infrastructure by deploying the Helm Charts.

Compared to the previous service deployment on a single VM, we were able to increase the number of requests per seconds by a factor of 5.8 (going from 4 to 23 Req / s).

⁸ <https://github.com/QCDIS/KMS-generic>

⁹ <https://github.com/QCDIS/kb-indexer>

¹⁰ <https://www.elastic.co/kibana/>

¹¹ <https://github.com/QCDIS/KMS-generic-helm-charts>

4 Integration with the ENVRI-Hub

ENVRI-Hub is the community effort to provide a single access point. To improve interoperability between the different ENVRI services, the KBSE has been integrated with the ENVRI-Hub. Both systems are maintained and deployed independently. The integration is achieved through hyperlinks on different pages of the ENVRI-Hub and the KBSE, as well as a search API from the KBSE¹². The ENVRI-Hub home page¹³ displays a search box which allows users to find resources from the KBSE (Figure 3). Thanks to the search API, results are displayed directly on the ENVRI-Hub page. To find more results or refine their search, users can follow a link to the KBSE interface. On this interface, a link allows them to navigate back to the ENVRI-Hub.

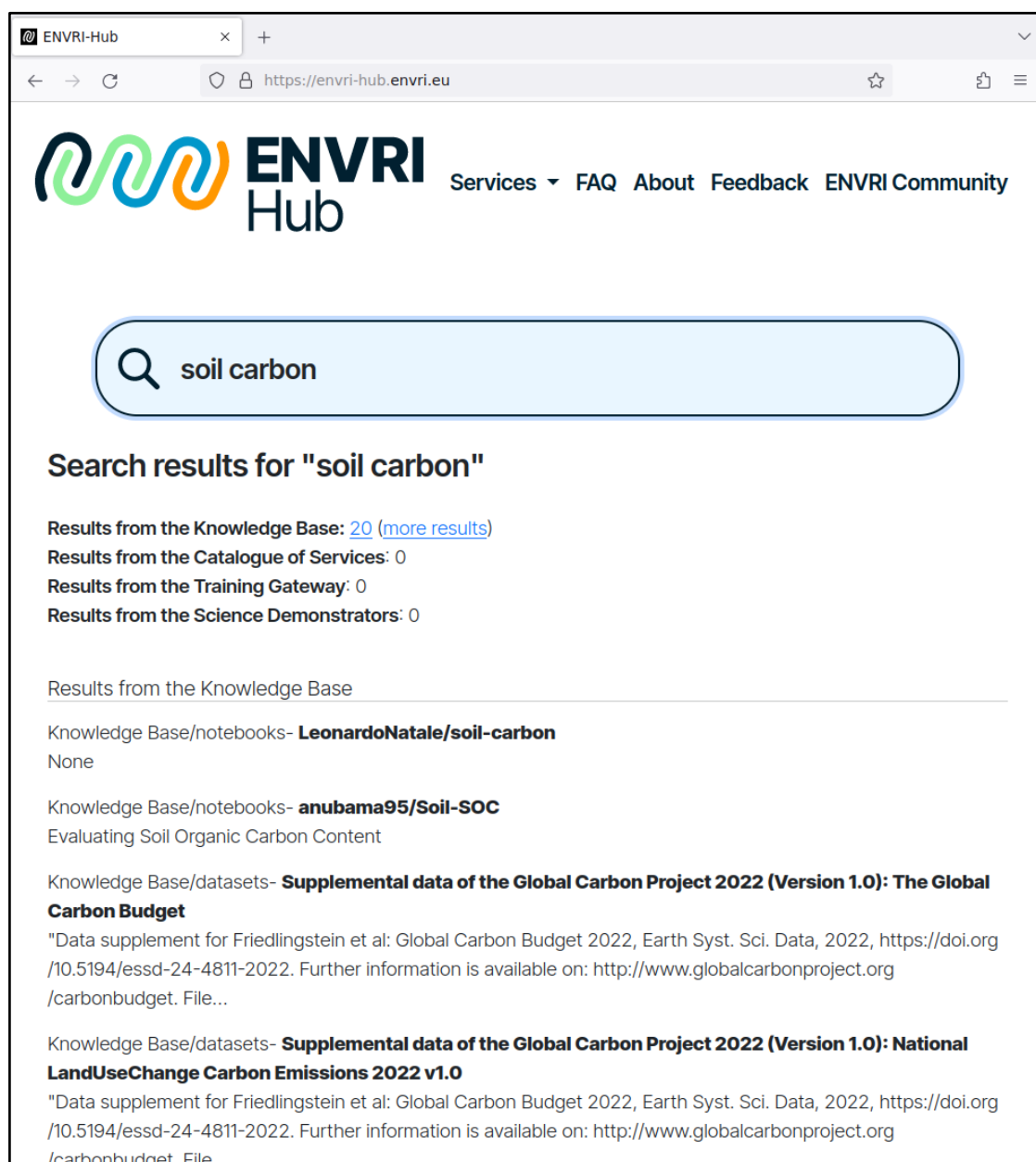


Figure 3: Search results of the ENVRI Knowledge Base search engine are shown on the ENVRI-Hub

¹² <https://search.envri.eu/api/v1/doc/>

¹³ <https://envri-hub.envri.eu>

Such integration has been successfully demonstrated at different events, such as EGU23 General Assembly and the RDA's 20th plenary in 2023. During the EGU23 demonstration, most visitors were early career scientists, as well as a few seniors. They were mainly interested in the data provided by ENVRI. A few other visitors were data engineers and developers, who were in the ENVRI services, and design of the ENVRI-Hub. The KBSE was used to look for information answering most questions, as it was the only service that could provide results starting from free text queries. Visitors also found the notebook search functionality interesting. It was noted that some of the dataset results linked to XML metadata records, which were not useful to the end users. This unintended behaviour will be fixed in the future, as we make sure that all search results link to landing pages, designed for humans. In the future, we plan to provide separate links to machine- and human-readable records in search results. Another suggestion was to make the KBSE and the ENVRI training catalogue interoperable. Currently, these two tools are only linked through the ENVRI-Hub, making it tricky for users to discover the tutorials or courses from the training catalogue while using the KBSE.

5 Demonstration

In this section, we demonstrate the new developments to the search engine through three user stories:

1. As a researcher, I would like to find resources that provide an overview of environmental indicators.
2. As a researcher, I would like to find datasets on several specific essential climate variables.
3. As a researcher, I would like to find science demonstrators developed by the ENVRI community.

These three stories begin on the homepage of the ENVRI Knowledge Base search engine, accessible at <https://search.envri.eu>, and shown on Figure 4.

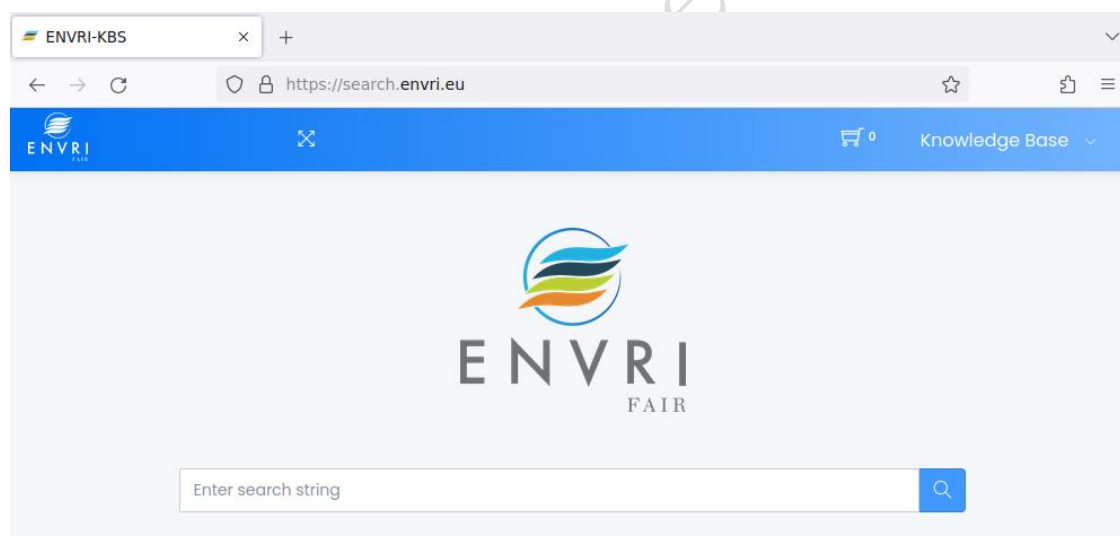


Figure 4: Homepage of the ENVRI Knowledge Base search engine

5.1 Environmental indicators overview

Starting from the home page of the KBSE, the user can search for “environmental indicators” and press search. The results page shows web pages from the ENVRI network (Figure 5). At the top of the results list, the user can find the Dashboard for the State of the Environment¹⁴, developed by ENVRI and integrated with EOSC Future.

¹⁴ <https://env-dashboard.eoscfuture.eu/dashboard>

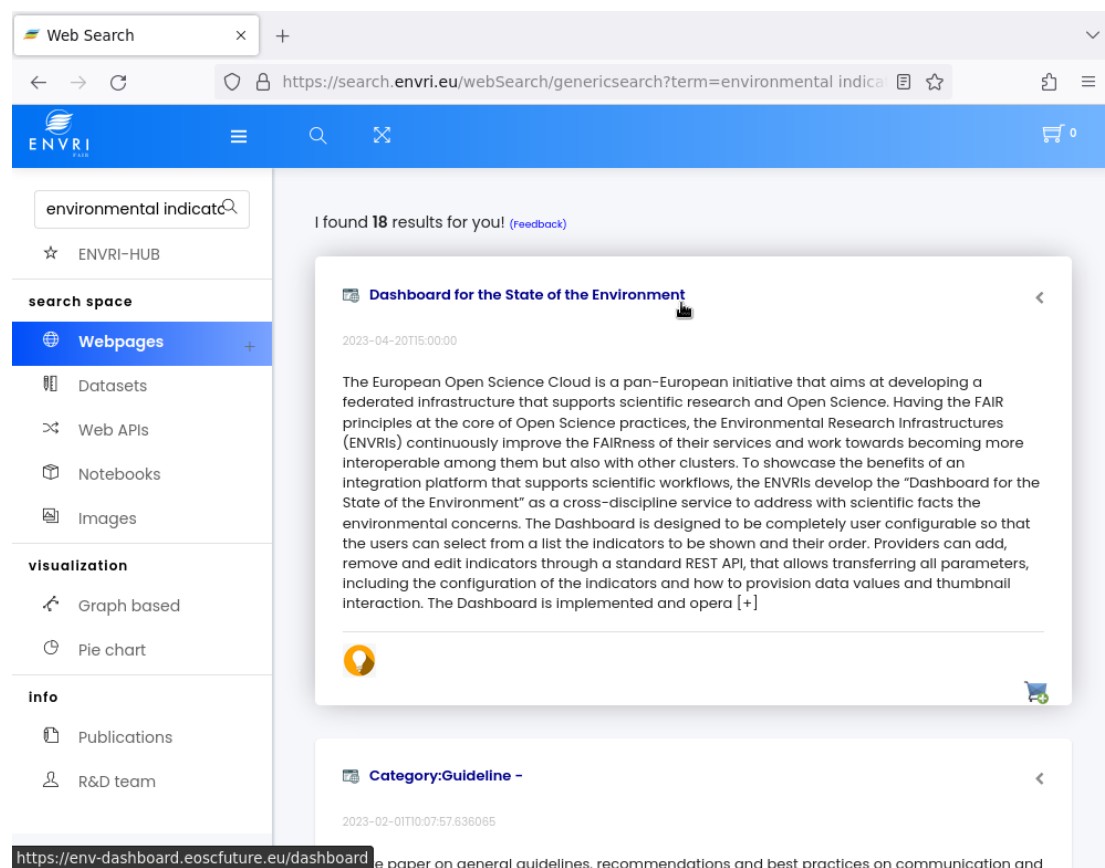


Figure 5: Search results showing web pages for “environmental indicators”

Switching to the web APIs search space (Figure 6), users can find the web API behind the EOSC Future Dashboard for the State of the Environment¹⁵.

¹⁵ <https://env-dashboard.eoscfuture.eu:4000/docs>

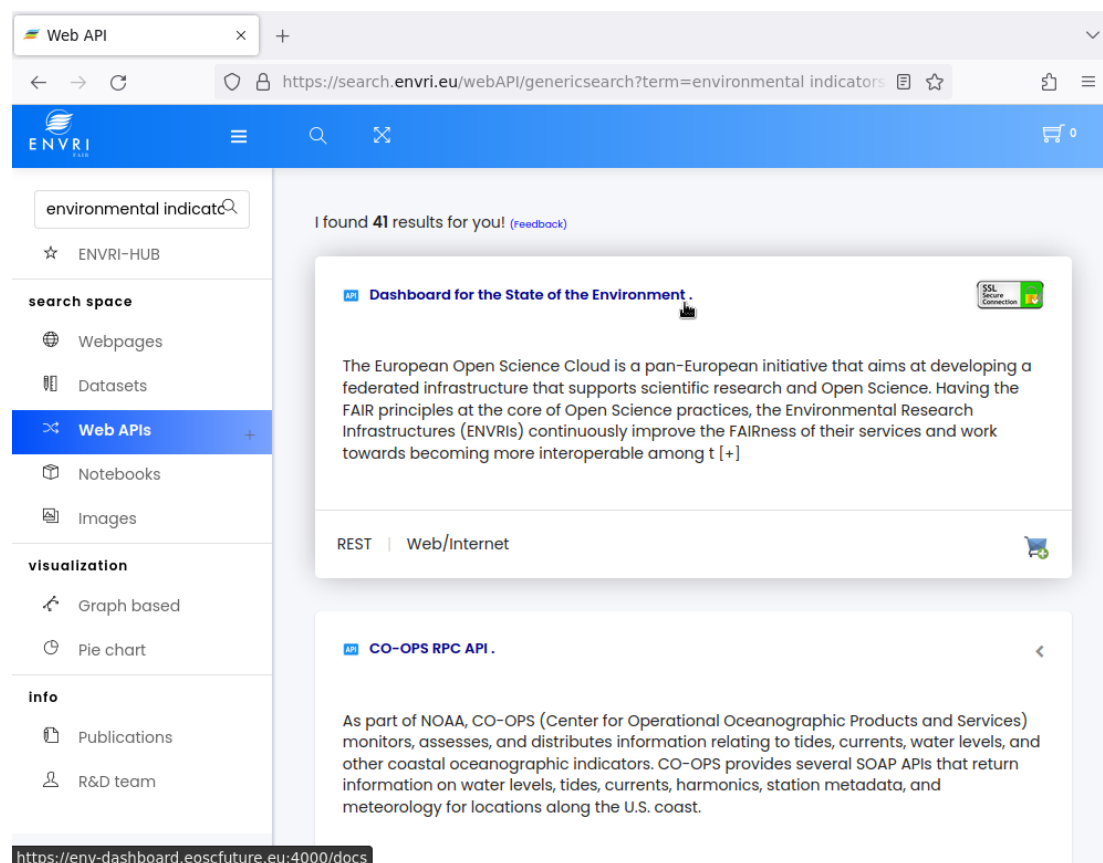


Figure 6: Search results showing web APIs for “environmental indicators”

5.2 Essential climate variables datasets

Users can find datasets relevant to specific essential climate, ocean, or biodiversity variables using the KBSE. For instance, searching for “soil carbon” in the datasets search space returns a list of datasets from different RIs¹⁶, relevant to this environmental indicator (Figure 7). Users can use the abstract displayed on the search results page and visit the datasets landing pages on the RIs catalogues to determine the relevance of the datasets.

¹⁶ https://search.envri.eu/dataset_elastic/genericsearch?term=soil%20carbon&page=1

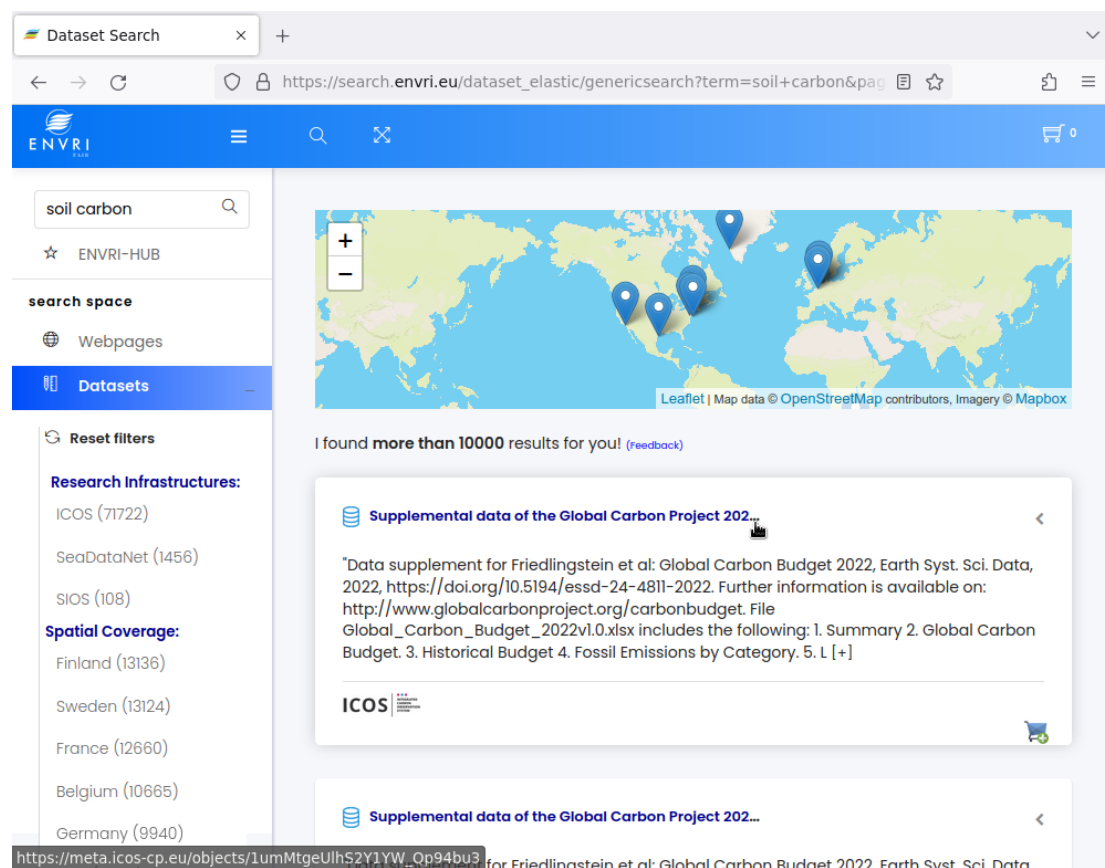


Figure 7: Search results showing datasets for the “soil carbon” environmental indicator

Switching to the “graph based” visualisation¹⁷ described in Deliverable 7.3 (Figure 8), users can get an overview of the search results for this query, across all search spaces. Results are categorised by source RI, and resource type.

Other examples search terms for this use case include:

- [methane](#)
- [anthropogenic water use](#)
- [surface radiation budget](#)
- [sensible heat flux](#)
- [latent heat flux](#)

¹⁷ https://search.envri.eu/dataset_elastic/genericsearch?term=soil carbon

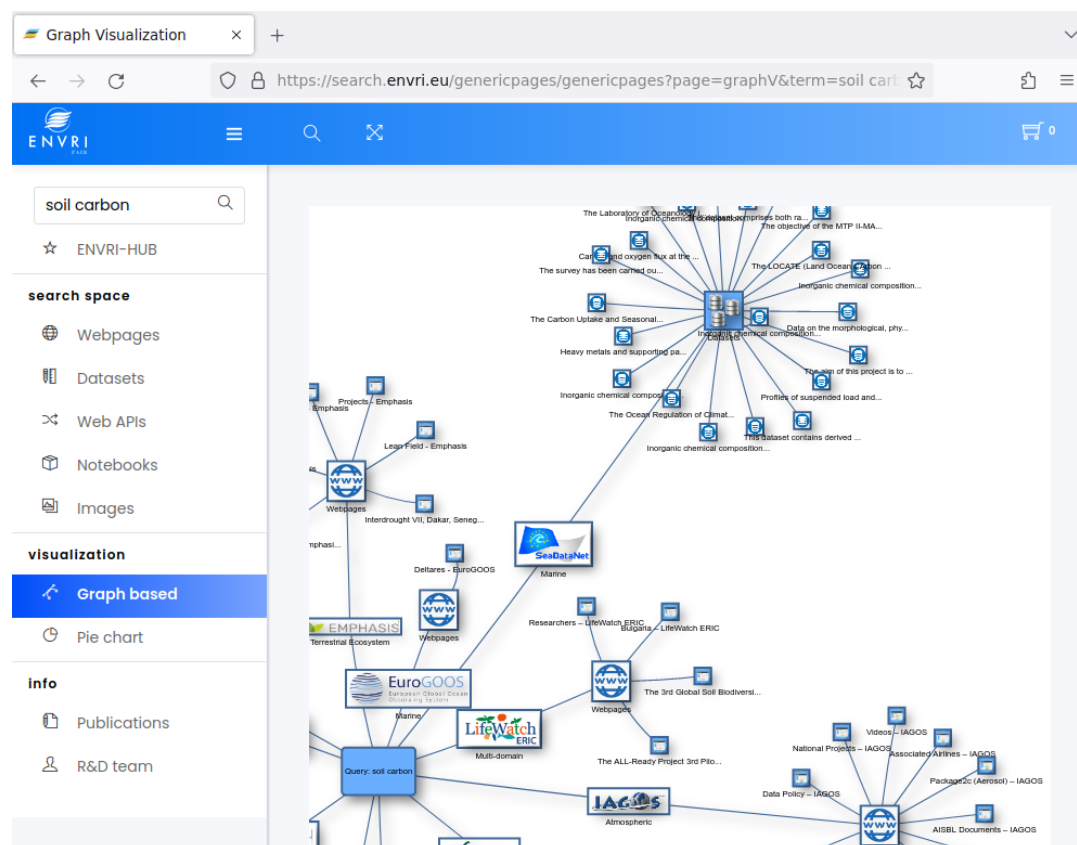


Figure 8: Graph visualisation of all search results for the “soil carbon” query

5.3 Science demonstrator notebooks

In order to find science demonstrator notebooks developed by the ENVRI community, users can search for “ENVRI demonstrators” and select the notebooks search space¹⁸. The first results show the four computational notebooks associated with ENVRI science demonstrators (Figure 9).

¹⁸ <https://search.envri.eu/notebookSearch/genericsearch?term=ENVRI+demonstrators&page=1>

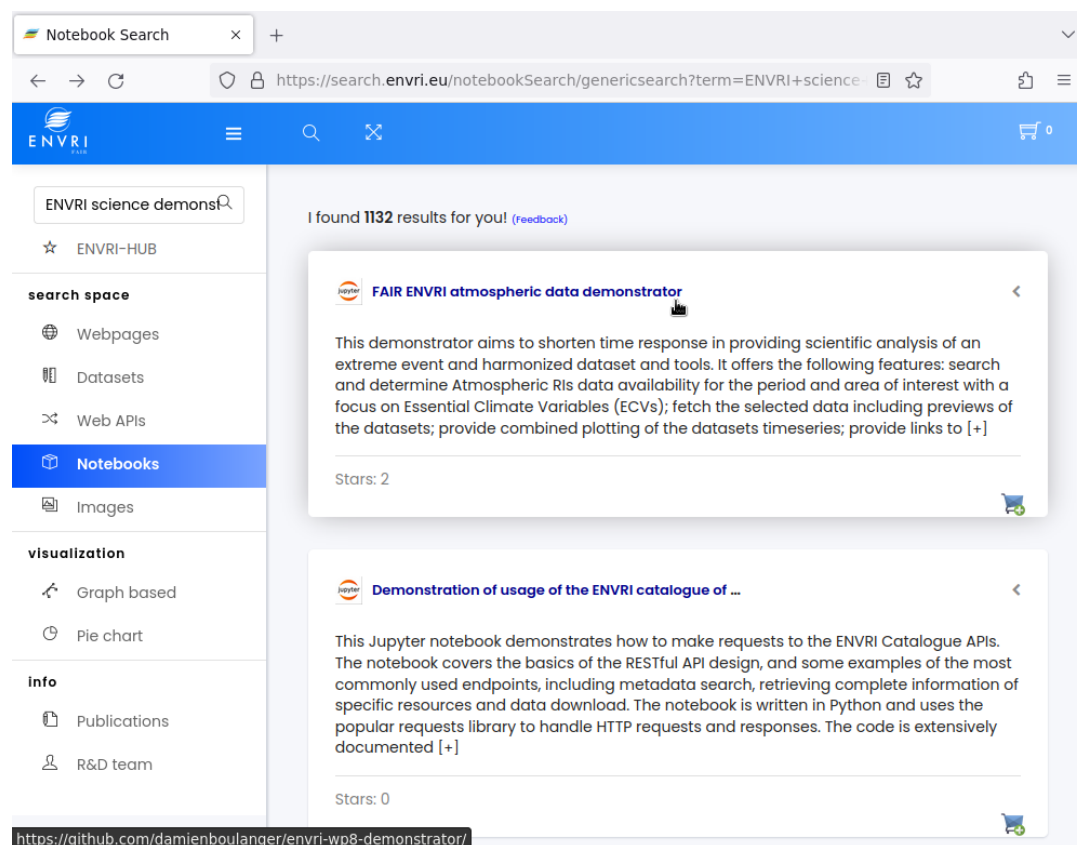


Figure 9: Search results showing computational notebooks for “ENVRI science demonstrators”

6 Summary

Since D7.3, new developments to the ENVRI Knowledge Base have focused on the search engine part. These include the development of an automated pipeline to index online resources, such as web pages, datasets, API descriptions, and computational notebooks from the ENVRI RIs. While web pages are ingested from all the ENVRI RIs, datasets metadata are currently harvested from five repositories, with the goal to expand it to all repositories within ENVRI. The list of resources to index could be updated in the future, as part of a new project such as ENVRI-Hub next. In addition to the indexing pipeline development, the search engine backend has been changed, and the user interface entirely redesigned. The ENVRI Knowledge Base is now deployed in a cloud environment, following modern DevOps practices. This allows us to better respond to peaks in traffic, as well as increase data resilience and service availability. Finally, better integration with the community was achieved by coordinating with the ENVRI-Hub, and user feedback was gathered during a live demo of the knowledge base search engine.

7 References

- [1] S. Farshidi *et al.*, ‘Knowledge sharing and discovery across heterogeneous research infrastructures [version 2; peer review: 1 approved, 1 approved with reservations, 1 not approved]’, *Open Res. Eur.*, vol. 1, no. 68, 2021, doi: 10.12688/openreseurope.13677.2.
- [2] S. Farshidi and Z. Zhao, ‘An Adaptable Indexing Pipeline for Enriching Meta Information of Datasets from Heterogeneous Repositories’, in *Advances in Knowledge Discovery and Data Mining*, J. Gama, T. Li, Y. Yu, E. Chen, Y. Zheng, and F. Teng, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, pp. 472–484. doi: 10.1007/978-3-031-05936-0_37.
- [3] N. Li, S. Farshidi, R. Bianchi, S. Koulouzis, and Z. Zhao, ‘Context-Aware Notebook Search in a Jupyter-Based Virtual Research Environment’, in *2022 IEEE 18th International Conference on e-Science (e-Science)*, Oct. 2022, pp. 393–394. doi: 10.1109/eScience55777.2022.00054.

8 Appendix 1: Glossary

ENVRI	(1) The ENVRI Community of Environmental Research Infrastructures. (2) FP7 project on Implementation of common solutions for a cluster of ESFRI infrastructures in the field of Environmental Sciences.
ENVRI-FAIR	An EU-funded project which stands for ENVironmental Research Infrastructures building Fair services Accessible for society, Innovation and Research.
FAIR	Findability, Accessibility, Interoperability, and Reusability of digital assets
Elasticsearch	Elasticsearch is a search engine based on the Lucene library.
EOSC	European Open Science Cloud
Knowledge Base (KB)	(1) A store of information or data that is available to draw on. (2) The underlying set of facts, assumptions, and rules which a computer system has available to solve a problem.
LifeWatch ERIC	European e-Science infrastructure for biodiversity and ecosystem research
Metadata	Data that describes other data. Metadata summarises basic information about data, which can make finding and working with particular instances of data easier.
Ontowiki	A free and open-source semantic wiki application, meant to serve as an ontology editor and a knowledge acquisition system.
Open Semantic Search	A free software for building own Search Engine, an explorer for discovery of large document collections, media monitoring, text analytics, document analysis & text mining platform based on Apache Solr or Elasticsearch.
RDA	Research Data Alliance
RDF	Resource Description Framework
RI	Research Infrastructure
SPARQL	SPARQL is an RDF query language—that is, a semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework format.
Triple	A triple is a data entity composed of subject-predicate-object
Triplestores	A triplestore or RDF store is a purpose-built database for the storage and retrieval of triples through semantic queries.
WP	Work Package