



D10.1

Semantic Search in ENVRI Catalogue and RI Catalogues

Status: Final

Dissemination Level: Public



Funded by
the European Union

Disclaimer: Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them

Abstract

Keywords	Semantic Search, Knowledge Graphs, SPARQL Endpoints, Model Context Protocol (MCP), Environmental Research Infrastructures, FAIR Principles
-----------------	--

This deliverable addresses data discovery challenges in European environmental research infrastructures by documenting semantic search capabilities within ENVRI-Hub NEXT. We assessed existing semantic infrastructures across the ENVRI community, identifying operational SPARQL endpoints in ICOS, Euro-Argo, SeaDataNet, and the NERC Vocabulary Server. Our architecture integrates Semantic Web technologies with Large Language Models through the Model Context Protocol, enabling natural language queries across federated knowledge graphs. We recommend adopting the SKG-IF standard for metadata harmonization and defining a 12-month roadmap with user stories and success criteria. This work aims to democratize environmental data access and advance FAIR principles implementation.

Revision History

Version	Date	Description	Author/Reviewer
V 0.1	15/12/2025	ToC and First Draft	Thierry Carval (Ifremer)
V 0.2	09/01/2026	WP10 Internal Review	Zhiming Zhao (UvA) Delphine Dobler (Euro-Argo ERIC) Erwan Bodéré (Ifremer)
V 0.3	14/01/2026	EHN Internal Review	Ulrich Bundke (Jülich), Alessandro Turco (EPOS ERIC)
V1.0	27/01/2026	Final Version	Thierry Carval (Ifremer)

Document Description			
D10.1 – Semantic Search in ENVRI Catalogue and RI Catalogues			
Work Package Number 10			
Document Type	Deliverable		
Document Status	Final	Version	1.0
Dissemination Level	Public		
Copyright Status	 This material by Parties of the ENVRI-Hub NEXT Consortium is licensed under a Creative Commons Attribution 4.0 International License .		
Lead partner	Ifremer		
Document Link	https://documents.egi.eu/document/4049		
DOI	https://zenodo.org/records/18433332		
Author(s)	<ul style="list-style-type: none"> Thierry Carval (Ifremer) Zhiming Zhao (UvA) Delphine Dobler (Euro-Argo ERIC) Erwan Bodéré (Ifremer) 		
Reviewers	<ul style="list-style-type: none"> Ulrich Bundke (Julich) Alessandro Turco (EPOS ERIC) 		
Moderated by:	Matteo Agati (EGI)		
Editing	Editing support with Claude by Anthropic		
Approved by:	Development Steering Board (DSB)		

Terminology / Acronyms

Term/Acronym	Definition

Useful Reference: [ENVRI Glossary](#)

Table of Contents

1. Introduction	8
1.1. Objectives of this Deliverable	8
1.2. Document Structure	8
2. General Principles of Semantic Search	9
2.1. Architecture Overview	9
2.2. SPARQL Endpoints	9
2.3. Knowledge Graphs	10
2.4. The SKG-IF Standard	10
2.5. Model Context Protocol (MCP)	11
3. State of the Art of Semantic Search in the ENVRI Community	12
3.1. Existing Semantic Resources	12
3.1.1. Central ENVRI-Hub Catalogue	12
3.1.2. NERC Vocabulary Server	12
3.1.3. ICOS (Integrated Carbon Observation System)	12
3.1.4. Euro-Argo	12
3.1.5. SeaDataNet CDI	13
3.1.6. Other Infrastructures	13
3.2. Examples of Semantic Queries	13
3.2.1. Parameter Mapping with NVS and I-ADOPT	14
3.2.2. Multi-Infrastructure Federated Queries	14
3.2.3. Resource Discovery through Semantic Relationships	14
4. Design of Search Algorithms	15
4.1. Overall Architecture	15
4.2. Conversational Agents	16
4.2.1. Understanding User Intent	16
4.2.2. Query Generation and Execution	16
4.2.3. Results Presentation	16
4.3. User Stories	16
4.3.1. Geographical Visualization of Observations	17
4.3.2. Query for Specific Data	17
4.3.3. Discovery of Related Resources	17
5. Roadmap and Next Steps	18
5.1. Phase 1: Consolidation of Existing Infrastructures (Months 23-30)	18
5.2. Phase 2: Development of Conversational Agents (Months 23-30)	18
5.3. Phase 3: Extension and Optimization (Months 30-36)	18
5.4. Associated Deliverables	18
5.5. Test Criteria, Benchmark	19
6. Conclusion	20
7. References	21

Table of Figures

- [Figure 1: MCP Bridge between Semantic Web Technologies](#)
- [Figure 2: The List of ICOS Stations](#)
- [Figure 3: The list of ENVRI-Hub Research Infrastructures](#)
- [Figure 4: ENVRI-Hub Semantic Search Architecture](#)

Executive Summary

This deliverable presents the state of the art and design of semantic search solutions within the ENVRI-Hub NEXT project. It describes existing infrastructures, deployed technologies, and the roadmap for the next 18 months.

Semantic search in ENVRI-Hub relies on three technological pillars: knowledge graphs to structure metadata according to standardized ontologies, SPARQL endpoints to query these graphs, and MCP (Model Context Protocol) connectors to interface these semantic resources with Large Language Models (LLMs).

Several research infrastructures in the ENVRI community already have operational SPARQL endpoints, including the central ENVRI-Hub catalogue, the NERC vocabulary server (NVS), ICOS, Euro-Argo, and SeaDataNet. For infrastructures that do not yet have a knowledge graph, we recommend adopting the SKG-IF (see [R1](#) Scientific Knowledge Graph Interoperability Framework) standard developed by the RDA.

The next steps will consist of exposing these semantic resources to LLMs via MCP connectors, thus enabling natural language querying of catalogues through conversational agents. This approach will transform the user experience by making data and services from environmental research infrastructures more accessible and queryable.

1. Introduction

The ENVRI-HUB-NEXT project aims to strengthen interoperability and accessibility of data and services from European environmental research infrastructures. In this context, semantic search constitutes a fundamental element to enable researchers to discover and access relevant resources in an intuitive and efficient manner.

This deliverable reviews the state of the art of semantic search within the ENVRI-Hub NEXT project, documenting existing deployed technologies and the design of search algorithms that will be implemented over the next 12 months. It is part of Work Package 10 dedicated to improving data discovery and access capabilities.

Semantic search differs from traditional keyword search by exploiting the semantics and relationships between concepts rather than relying on exact syntactic matching (e.g. Duhan et al., 2024). It relies on Semantic Web technologies (see [R10](#)) such as knowledge graphs, ontologies, triple stores and the SPARQL query language (see [R1](#), [R2](#), [R3](#), [R4](#), [R5](#)). The goal is to enable users to formulate queries in natural language and obtain relevant results that exploit the richness of structured metadata.

1.1. Objectives of this Deliverable

This document aims to:

- Establish an inventory of semantic resources currently deployed in the ENVRI community;
- Present the principles and technologies underlying semantic search in ENVRI-Hub;
- Design the architecture and semantic search algorithms;
- Define the roadmap for the development and implementation of advanced search functionalities.

1.2. Document Structure

The rest of this document is organized as follows:

- [Section 2](#) presents the general principles of semantic search and the key technologies used.
- [Section 3](#) provides a detailed inventory of existing semantic resources in the ENVRI community.
- [Section 4](#) describes the design of the search architecture and algorithms.
- [Section 5](#) presents the development roadmap for the coming months.
- Finally, [Section 6](#) concludes this deliverable and offers perspectives.

2. General Principles of Semantic Search

Semantic search in ENVRI-Hub relies on a set of principles and technologies that enable full exploitation of the richness of research infrastructure metadata. This section presents the conceptual and technical foundations of our approach.

2.1. Architecture Overview

ENVRI data and services are exposed in ENVRI catalogues and/or directly within individual research infrastructures. The ENVRI-Hub NEXT project – building upon the architectural framework of the original ENVRI-Hub (see [R12](#) Petzold et al. 2023) – dedicates significant efforts to creating catalogues compliant with FAIR principles (Findable, Accessible, Interoperable, Reusable) (see [R13](#) Wilkinson et al, 2016). These catalogues are designed to be queried semantically, thus enabling optimized discovery and access to scientific resources.

In ENVRI-Hub, research assets – including data, services, and documents – can be discovered via the catalogue or the ENVRI knowledge base search engine. This search engine indexes ENVRI resources using advanced information retrieval technologies, offering keyword-based search and similarity-based ranking. Additionally, the ENVRI knowledge base provides LLM-empowered dialogue agents to support personalized search experiences (see [R14](#) Zhao et al., *Advancing the ENVRI Knowledge Base with LLM Agents and Model Context Protocol. Abstract in SciDataCon 2025 / International Data Week*).

The semantic search capability is often structured around three main components: knowledge graphs that structure metadata according to standardized ontologies, SPARQL endpoints that enable querying of these graphs, and MCP (Model Context Protocol) connectors that bridge knowledge graphs and Large Language Models (LLMs). This modular architecture ensures both flexibility and extensibility of the system.

2.2. SPARQL Endpoints

The consensus within the ENVRI community is to manage service metadata in a knowledge graph, generally hosted in a triple store (graph-oriented database). A triple store stores information as subject-predicate-object triples, in accordance with RDF (Resource Description Framework) Semantic Web standards.

For each knowledge graph, a SPARQL endpoint enables semantic queries. SPARQL (SPARQL Protocol and RDF Query Language) is the standard query language for RDF data, comparable to SQL for relational databases. It allows expression of complex queries exploiting relationships between entities and metadata properties.

A central SPARQL endpoint can perform federated queries, simultaneously interrogating multiple local SPARQL endpoints hosted by different research infrastructures. This federated approach enables a unified view of distributed resources while respecting the autonomy of each infrastructure.

2.3. Knowledge Graphs

A knowledge graph represents the semantic content exposed by a research infrastructure, structured according to an ontology. An ontology formally defines the concepts of a domain and the relationships between these concepts, thus enabling standardized and interoperable representation of knowledge.

Some research infrastructures in the ENVRI community already have knowledge graphs describing their data and services, exposed via SPARQL endpoints. These graphs may use different ontologies according to the specific needs of each infrastructure and its scientific domain.

For research infrastructures that do not yet have a knowledge graph, or wish to improve the interoperability of their existing metadata, we recommend adopting the SKG-IF (Scientific Knowledge Graph Interoperability Framework) standard. This standard, described in the following section, offers a common data model specifically designed for scientific research infrastructures.

2.4. The SKG-IF Standard

SKG-IF (Scientific Knowledge Graph Interoperability Framework) (see [R15](#) Mannocci et al. 2025) is a common data model developed by the RDA (see [R1](#) Research Data Alliance) to enable interoperability between scientific knowledge graphs. It is a community effort aimed at establishing standards for representing research information across different disciplines and infrastructures.

SKG-IF provides a standardized ontology and schema for representing research outputs, datasets, organizations, researchers, projects, and their relationships across different research infrastructures. The model notably covers:

- Publications and other research outputs,
- Datasets and their metadata,
- People and their affiliations,
- Organizations and institutions,
- Research projects and their funding,
- Relationships between these different entities.

Adopting SKG-IF presents several advantages for the ENVRI community. It facilitates harmonization of metadata between infrastructures, simplifies federated queries by relying on a common vocabulary, and reduces development and maintenance costs by pooling efforts around a shared standard. Moreover, SKG-IF is designed to be extensible, allowing infrastructures to add domain-specific concepts while maintaining compatibility with the base model.

The complete documentation of SKG-IF is available at:

- <https://skg-if.github.io/data-model>

For infrastructures that already have knowledge graphs, we will study migration or compatibility possibilities towards SKG-IF. This approach can take different forms depending on the case: complete migration to SKG-IF, maintaining a dual model with mappings between the existing ontology and SKG-IF, or progressive extension of the existing ontology to include SKG-IF concepts.

2.5. Model Context Protocol (MCP)

The Model Context Protocol (MCP) (see [R16 Anthropic, 2024](#)) constitutes the interface layer between knowledge graphs and Large Language Models (LLMs). An MCP connector exposes the semantic content of the ENVRI catalogue and research infrastructure catalogues to LLMs, thus enabling exploitation of these resources in a conversational artificial intelligence context.

MCP acts as a bridge between Semantic Web technologies (knowledge graphs, SPARQL) and generative AI technologies (LLMs). It translates natural language queries formulated by users into structured SPARQL queries, executes these queries on appropriate endpoints, and then reformats results so they can be presented intelligibly by the LLM.

A chatbot interfaced with the LLM enables natural language querying of research infrastructure resources. Users can thus ask complex questions without needing to know SPARQL or the structure of knowledge graphs. The system understands user intent, identifies relevant resources, formulates appropriate queries, and presents results conversationally.

This conversational approach represents a major evolution in the accessibility of scientific data, considerably reducing the technical barrier for researchers who wish to discover and access resources from environmental research infrastructures.

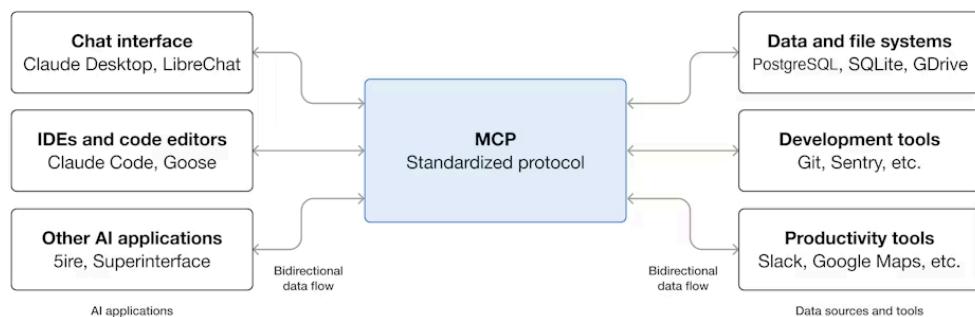


Figure 1: MCP Bridge between Semantic Web Technologies

3. State of the Art of Semantic Search in the ENVRI Community

The ENVRI community has already established several operational semantic infrastructures. This section presents an inventory of existing resources and concrete examples of semantic search usage.

3.1. Existing Semantic Resources

A series of SPARQL endpoints is already deployed and operational within the ENVRI community. These endpoints expose metadata from different research infrastructures and enable advanced semantic queries. Here is an inventory of the main resources:

3.1.1. Central ENVRI-Hub Catalogue

The ENVRI-Hub Catalogue Of Services (COS) has a SPARQL endpoint that aggregates metadata from different member research infrastructures. This central catalogue plays a federating role and enables cross-cutting searches across the entire ENVRI ecosystem.

- <https://envri-hub.staging.envri.eu/sparql/rdflib>

3.1.2. NERC Vocabulary Server

The NVS (NERC Vocabulary Server) constitutes an essential resource for terminological harmonization. It hosts a vast collection of controlled vocabularies for marine and environmental sciences. The NVS SPARQL endpoint is accessible at:

- <https://vocab.nerc.ac.uk/sparql>

NVS notably exposes resources from the I-ADOPT framework (Interoperability of Data and Procedures in Ocean sciences), which provides a conceptual model for standardized description of observed variables and measured parameters in the oceanographic and environmental domain.

3.1.3. ICOS (Integrated Carbon Observation System)

ICOS (see [R6](#)), the European research infrastructure for carbon observation, has a SPARQL endpoint for its metadata. This catalogue exposes information on measurement stations, datasets, and observed variables related to the carbon cycle. The ICOS SPARQL client is accessible at:

- <https://meta.icos-cp.eu/sparqlclient>

3.1.4. Euro-Argo

Euro-Argo, the European infrastructure for oceanographic profiling floats, has developed a knowledge graph based on Linked Data. The Euro-Argo SPARQL endpoint (see [R7](#)) enables querying of float, mission, and observation metadata. It is accessible at:

- <https://co.ifremer.fr/co/argo-linked-data/html/Argo-HTML-SPARQL>

3.1.5. SeaDataNet CDI

SeaDataNet, the pan-European infrastructure for oceanographic and marine data, exposes its metadata via the CDI (Common Data Index, see [R8](#)). The CDI SPARQL endpoint enables querying of a rich catalogue of marine datasets from numerous European institutions. It is accessible at:

- <https://cdi.seadatanet.org/sparql>

3.1.6. Other Infrastructures

Other research infrastructures in the ENVRI community are developing their semantic capabilities, notably ANAEE (Analysis and Experimentation on Ecosystems) and several FAIR Data Point (FDP) implementations. These initiatives will be integrated into the semantic search ecosystem as they mature.

3.2. Examples of Semantic Queries

Existing SPARQL endpoints already enable sophisticated semantic queries exploiting relationships between concepts and metadata properties. Here are some usage examples:

ENVRI-Hub SPARQL Client

Endpoint: <https://envri-hub.staging.envri.eu/sparql/>

Pre-build queries: [ICOS Stations](#)

Query +

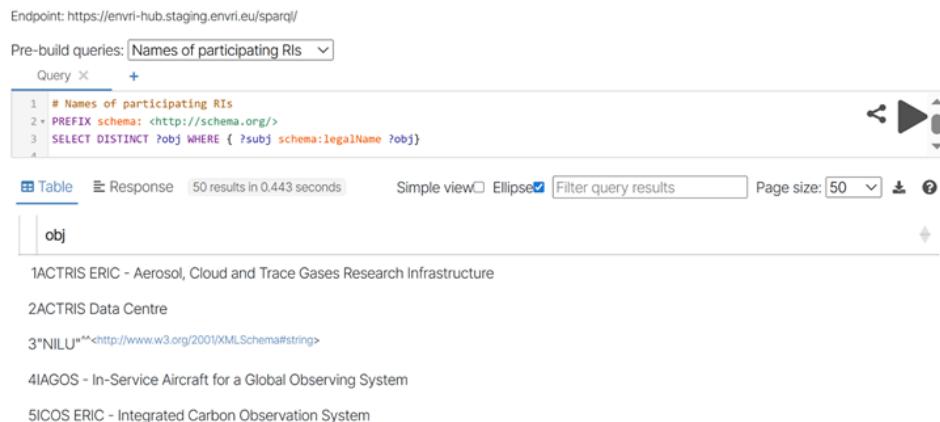
1 # ICOS Stations
2 PREFIX cpmeta: <http://meta.icos-cp.eu/ontologies/cpmeta/>
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 PREFIX cpst: <http://meta.icos-cp.eu/ontologies/stationentry/>
5 SELECT
6 ?tcName
7 ?hoName
8 ?Theme
9 ?tcClass
10 ?hoClass
11 ?tcCountry
12 ?hoCountry
13 FROM <http://meta.icos-cp.eu/resources/stationentry/>
14 FROM <http://meta.icos-cp.eu/ontologies/stationentry/>

Table Response 50 results in 34.799 seconds Simple view Ellipse Filter query results Page size: 50

tcName	hoName	Theme	tcClass	hoClass	tcCountry	hoCountry
1Lampedusa	Lampedusa	Atmospheric Station	2		IT	IT
2Monte Cimone	Monte Cimone	Atmospheric Station	2		IT	IT
3Hytemossa	Hytemossa	Atmospheric Station	1		SE	SE
4Jungfraujoch	Jungfraujoch	Atmospheric Station	1		CH	CH
5Puy de Dôme	Puy de Dome	Atmospheric Station	2		FR	FR
6Plateau Rosa	Plateau Rosa	Atmospheric Station	2		IT	IT

Figure 2: The List of ICOS Stations

ENVRI-Hub SPARQL Client



The screenshot shows the ENVRI-Hub SPARQL Client interface. At the top, it displays the endpoint: <https://envri-hub.staging.envri.eu/sparql/>. Below this is a dropdown menu for 'Pre-build queries' set to 'Names of participating RIs'. The main area shows a SPARQL query in the editor:

```

1 # Names of participating RIs
2+ PREFIX schema: <http://schema.org/>
3 SELECT DISTINCT ?obj WHERE { ?subj schema:legalName ?obj }
4

```

Below the query editor, the results are displayed in a table. The table has two columns: 'obj' and 'label'. The results are:

obj	label
1	ACTRIS ERIC - Aerosol, Cloud and Trace Gases Research Infrastructure
2	ACTRIS Data Centre
3	"NILU"^^<http://www.w3.org/2001/XMLSchema#string>
4	IAGOS - In-Service Aircraft for a Global Observing System
5	ICOS ERIC - Integrated Carbon Observation System

At the bottom of the interface, there are buttons for 'Table' (selected), 'Response' (50 results in 0.443 seconds), 'Simple view' (unchecked), 'Ellipsis' (checked), 'Filter query results' (unchecked), and 'Page size' (set to 50). There are also icons for back, forward, and search.

Figure 3: The list of ENVRI-Hub Research Infrastructures

3.2.1. Parameter Mapping with NVS and I-ADOPT

An important use of semantic search consists of mapping parameters measured by different research infrastructures with NVS standardized vocabularies and the I-ADOPT framework. This allows identification of equivalent or similar variables across different infrastructures, even when they use different terminologies.

For example, a SPARQL query can identify all infrastructures that measure sea surface temperature, relying on the formal definition of this parameter in controlled vocabularies. This approach greatly facilitates comparative studies and multi-source analyses.

3.2.2. Multi-Infrastructure Federated Queries

Federated queries enable simultaneous interrogation of multiple SPARQL endpoints to obtain an overview of available resources. For example, a query can search for all datasets available for a given geographical area (such as the Bay of Biscay) by simultaneously querying the catalogues of ICOS, Euro-Argo, SeaDataNet, and other infrastructures.

3.2.3. Resource Discovery through Semantic Relationships

The richness of knowledge graphs enables resource discovery through navigation of semantic relationships. For example, from a research project, one can discover produced datasets, involved researchers, associated publications, and used infrastructures. This exploratory approach is particularly useful for data discovery and linking related resources.

4. Design of Search Algorithms

This section describes the technical architecture and design of semantic search algorithms that will be developed and deployed within the ENVRI-Hub NEXT project. The goal is to create an integrated system enabling intuitive querying of research infrastructure resources.

4.1. Overall Architecture

The ENVRI-Hub semantic search architecture is structured around several interconnected technological layers. [Figure 4](#) illustrates the general organization of the ENVRI HUB Next system:

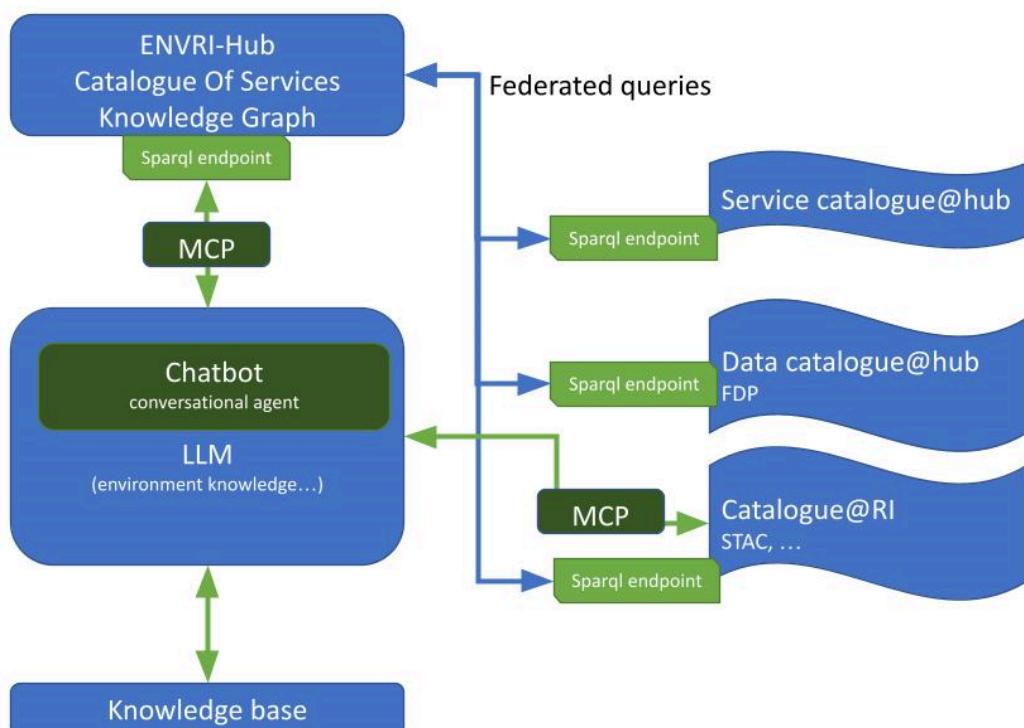


Figure 4: ENVRI-Hub Semantic Search Architecture

Data Layer:

At the base of the system are triple stores hosting knowledge graphs from different research infrastructures (see Catalogue@RI, Data catalogue@hub, Service Catalogue@hub in [Figure 4](#)). These graphs are structured according to standardized ontologies, notably SKG-IF for infrastructures adopting this standard. This layer ensures the persistence and integrity of semantic metadata.

Data Access Layer:

SPARQL endpoints (light green in [Figure 2](#)) expose knowledge graphs and enable their querying. A central endpoint (top left) coordinates federated queries to local endpoints of different

infrastructures. This distributed architecture preserves the autonomy of each infrastructure while allowing a unified view of resources.

Artificial Intelligence Layer:

MCP connectors bridge SPARQL endpoints and Large Language Models (LLMs). They translate natural language queries into structured SPARQL queries, manage execution of these queries on appropriate endpoints, and reformat results for presentation by the LLM. This layer also integrates reasoning and query optimization logic.

User Interface Layer:

At the top of the architecture is the conversational interface (chatbot) that allows users to interact with the system in natural language. This interface masks underlying technical complexity and offers an intuitive and accessible user experience.

4.2. Conversational Agents

Conversational agents constitute the privileged interface between users and ENVRI-Hub semantic resources. These agents exploit the capabilities of Large Language Models to understand natural language queries and orchestrate access to different data sources.

4.2.1. Understanding User Intent

The first step in processing a query consists of analyzing user intent. The LLM identifies key concepts mentioned (parameters, geographical areas, time periods, infrastructures), determines the desired search type (data discovery, metadata navigation, comparative analysis), and extracts constraints and filtering criteria.

4.2.2. Query Generation and Execution

Based on intent analysis, the system generates one or more SPARQL queries. The MCP connector translates the conceptual representation of the query into formal SPARQL syntax, selects relevant endpoints, and orchestrates query execution. In case of federated queries, the system coordinates calls to different endpoints and aggregates results.

4.2.3. Results Presentation

Raw results from SPARQL queries are transformed by the LLM into a conversational presentation adapted to the context of the initial question. The system can synthesize results, organize them by relevance, suggest complementary searches, and provide links to complete resources. The goal is to present information clearly and actionable for the user.

4.3. User Stories

To illustrate the expected capabilities of the semantic search system, we present some user stories representative of researcher needs, giving guidelines and motivation for the needed development steps presented in the roadmap (see [Section 5](#)):

4.3.1. Geographical Visualization of Observations

User Query:

"I want to see ENVRI observations on a map"

Processing:

The system queries different catalogues to identify all geolocated datasets, extracts their geographical coordinates and associated metadata, and then generates an interactive map visualization. The user can then filter by observation type, infrastructure, time period, etc.

4.3.2. Query for Specific Data

User Query:

"I want to know the average temperature of the Bay of Biscay this October"

Processing:

The system identifies the geographical area (Bay of Biscay), the searched parameter (surface temperature), and the time period (October of the previous year). It queries relevant catalogues (Euro-Argo, SeaDataNet, ICOS, etc.) to find corresponding datasets, accesses the data itself if available, calculates the requested average, and presents the result with metadata of sources used.

4.3.3. Discovery of Related Resources

User Query:

"What data are available to study the impact of climate change on marine ecosystems in the Mediterranean?"

Processing:

This complex query requires an understanding of semantic relationships between concepts. The system identifies relevant data types (temperature, salinity, biodiversity, water chemistry), the geographical area (Mediterranean), and searches for associated datasets, publications, and projects. It presents results organized by theme and suggests complementary resources based on relationships in the knowledge graph.

5. Roadmap and Next Steps

This section presents the roadmap for the development and deployment of semantic search functionalities over the next 12 months of the ENVRI-Hub NEXT project.

5.1. Phase 1: Consolidation of Existing Infrastructures (Months 23-30)

Objectives:

- Document existing SPARQL endpoints,
- Evaluate the compatibility of existing knowledge graphs with SKG-IF,
- Develop first versions of MCP connectors for priority infrastructures,
- Establish a registry of vocabularies and ontologies used in the ENVRI community.

5.2. Phase 2: Development of Conversational Agents (Months 23-30)

Objectives:

- Implement a conversational agent prototype with a limited set of use cases,
- Develop algorithms for natural language → SPARQL query translation,
- Integrate MCP connectors with selected LLMs,
- Conduct user testing with a pilot group of researchers.

5.3. Phase 3: Extension and Optimization (Months 30-36)

Objectives:

- Extend the system to all infrastructures in the ENVRI community,
- Optimize federated query performance,
- Improve intent understanding and query generation algorithms,
- Develop advanced features: proactive suggestions, faceted navigation, etc.,
- Prepare technical documentation and user guides.

5.4. Associated Deliverables

Implementation of this roadmap will result in the following deliverables:

- D10.2: Conversational agent prototype and evaluation (Month 36),
- Scientific publications on the architecture and algorithms developed,

5.5. Test Criteria, Benchmark

The following criteria will serve as benchmarks for the project:

- Number of integrated infrastructures: at least 8 infrastructures with operational SPARQL endpoints,
- Query performance: average response time < 5 seconds for simple queries,
- Understanding accuracy: success rate > 85% on a set of test queries,
- User satisfaction: satisfaction score > 4/5 in user testing,
- Adoption: at least 100 active users within 6 months following deployment.

6. Conclusion

This deliverable has presented the state of the art and design of semantic search solutions within the ENVRI-Hub NEXT project. The ENVRI community already has a solid foundation with several operational SPARQL endpoints and rich knowledge graphs exposing metadata from environmental research infrastructures.

The proposed architecture relies on three complementary technological pillars: knowledge graphs structured according to standardized ontologies (notably SKG-IF), SPARQL endpoints enabling querying of these graphs, and MCP connectors interfacing these semantic resources with Large Language Models. This approach combines the advantages of the Semantic Web and conversational artificial intelligence to offer an intuitive and powerful user experience.

The next 12 months will be dedicated to progressive development and deployment of semantic search functionalities. The established roadmap provides for an iterative approach in three phases: consolidation of existing infrastructures, development of conversational agents, and then extension and optimization of the system. This approach will enable validation of technical choices and integration of user feedback throughout development.

The final goal is to transform the experience of scientific data discovery and access by enabling researchers to query resources from environmental research infrastructures in natural language. This democratization of data access will contribute to accelerating environmental research and fostering interdisciplinary collaborations within the ENVRI community.

The next deliverable (D10.2) will demonstrate developments carried out and present an evaluation of the operational system. There will also be an opportunity to share lessons learned and best practices identified during implementation.

7. References

Reference	
No	Description/Link
R1	Research Data Alliance (RDA). Scientific Knowledge Graph Interoperability Framework (SKG-IF). Available at: https://skg-if.github.io/data-model
R2	W3C. SPARQL 1.1 Query Language. W3C Recommendation, 21 March 2013. Available at: https://www.w3.org/TR/sparql11-query
R3	[3] W3C. RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation, 25 February 2014. Available at: https://www.w3.org/TR/rdf11-concepts
R4	NERC Vocabulary Server (NVS). Available at: https://vocab.nerc.ac.uk
R5	I-ADOPt Working Group. InteroperAbility of Data and Procedures in Ocean sciences for inTeroperable data exchange. Available at: https://www.i-adopt.org
R6	ICOS Carbon Portal. ICOS Metadata Service. Available at: https://meta.icos-cp.eu
R7	Euro-Argo. Argo Linked Data. Available at: https://co.ifremer.fr/co/argo-linked-data
R8	SeaDataNet. Common Data Index (CDI). Available at: https://www.seadatanet.org
R9	Wilkinson, M. D., et al. The FAIR Guiding Principles for scientific data management and stewardship. <i>Scientific Data</i> , 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
R10	Berners-Lee, T., Hendler, J., & Lassila, O. The Semantic Web. <i>Scientific American</i> , 284(5), 34-43 (2001).
R11	ENVRI-HUB-NEXT project website. Available at: https://envri.eu/envri-hub-next
R12	Petzold, A., Gomes, A. R., Bundke, U., Schleiermacher, C., Adamaki, A., Vermeulen, A., Zhao, Z., Stocker, M., Myhre, C. L., Boulanger, D., Hienola, A., & Bailo, D. (2023). ENVRI-Hub Design and Architecture White Paper. Environmental Research Infrastructures Community ENVRI, Version 1. Zenodo: https://zenodo.org/records/8046894
R13	Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Tolia, A., & van der Velde, M. (2016). The FAIR Guiding Principles for scientific data management and stewardship. <i>Scientific Data</i> , 3, 160018. doi:10.1038/sdata.2016.18

	M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. <i>Scientific Data</i> , 3, 160018. Zenodo Lund University https://doi.org/10.1038/sdata.2016.18
R14	Zhao et al., Advancing the ENVRI Knowledge Base with LLM Agents and Model Context Protocol. Abstract in SciDataCon 2025 / International Data Week
R15	Mannocci, A., Manghi, P., & SKG-IF WG. (2025). Advancing Integration and Reuse Across Diverse Scholarly Data Sources. Introducing the SKG-IF: the Interoperability Framework for Scientific Knowledge Graphs. ResearchGate In I. Heibi, C. Di Giambattista, & S. Peroni (Eds.), <i>Proceedings of the Workshop on Open Citations and Open Scholarly Metadata (WOOC 2025)</i> , Bologna, Italy, 28-29 May 2025. Zenodo. https://doi.org/10.5281/zenodo.16365716
R16	Anthropic. (2024). Introducing the Model Context Protocol. Fedcloud Anthropic. https://www.anthropic.com/news/model-context-protocol