



Interoperable Information Retrieval System for Enhanced Knowledge Management

Enhancing insights through AI-driven workflows

Lorenzo GABRIELLI, Edoardo RAMALLI, Francesco CABRAS, Davide DE MARCHI, Alexander KOTSEV, Anders FRIIS-CHRISTENSEN

Lorenzo GABRIELLI

Our Work and Key Projects





Data Ecosystems and Enablers – ENABLED (2025 – 2027)

- **Why?**
- Sub-optimal data sharing in the EU
- Challenges: fragmented infrastructures, power imbalances, underused data, varying standards
- **How?**
- Experimental approaches
- JRC Data Sharing Innovation Lab
- **What?**
- Provide scientific evidence for data sharing practices
- Focus on technical and governance aspects
- **With Whom?**
- Partners in policy, academia, and data-sharing associations

Data analytics infrastructure and services - DAIS

- Centralized Platform for Scientific Data:
 - FAIR & AI-Ready Data
- Integrated Data & Computational Infrastructure:
 - Big Data Analytics Platform
- Support for Policy-Relevant Analytics:
 - Data Management, processing, analysis & Visualization

Background

-  **What**
Enhance the *findability* of content, making strategic information more accessible and actionable.
-  **Why**
Distill essential insights to support strategic analysis and decision-making across various domains, ensuring that relevant information could be easily retrieved and utilized.
-  **Problem: Lack of a Structured Corporate Solution**
 *No existing framework was in place to address our specific needs, requiring a tailored approach to enhance findability and knowledge extraction.*

In practice - Implementation



Semantic Text Analyser and More

- Enables ingestion and semantic search of text corpora
- Provides a range of APIs including a RAG API that returns relevant documents

voilà

GPT@JRC

API

GPT@JRC

Knowledge tool

SeTA ingestion:

- Data model mapping
- Chunking text in smaller units
- Calculate embeddings
- Store in Opensearch



DSSC Blueprint 1.5



European Commission

Joint Research Centre Data Catalogue



European Commission



MISSION LETTER

Brussels, 17 September 2024

Ursula von der Leyen
President of the European Commission



European Commission

Focus on Case Studies: RAG and LLM in Action

Data Spaces Prototype (DSSC)

- Data Source: Includes Blueprint 1.5 data manually downloaded due to lack of AI-ready approach from web pages.
- Additional Sources: Staff working documents and European Strategy for Data in PDF format, imported and processed using Python.
- Limitations: For now, PDFs are divided by pages, lacking optimized chunking, but necessary due to technical constraints.

JRC Data Catalogue

- 3700+ datasets
- DCAT-AP compliant metadata descriptions (semi-structured text)
- Varying levels of metadata quality
- APIs for efficient data extraction

Key learning from the Prototypes

- **Improved Data Retrieval:** Integrating RAG has significantly enhanced our ability to retrieve context-specific information, ensuring more accurate and relevant responses.
- **Efficiency in Summarization:** GPT@JRC models for summarization have streamlined the document review process, enabling quicker extraction of key insights without losing critical context.
- **Importance of AI-Ready Data:** Ensuring that data is AI-ready is crucial for improving interoperability and enabling more effective integration across systems.
- **Enhanced Usability:** The full-stack web application has simplified user interactions, making the data flow more intuitive and accessible, improving the overall user experience.

Future work and discussion points

Improving Data Quality and Integration

- **Enriched Metadata:** Improving metadata with richer descriptions and related resources (e.g., abstracts).
- **New Data Sources:** Adding and exposing new data sources in SeTA via RAG (e.g., CORDA data).
- **Collaboration with DIGIT (SEMIC Action):** Exploring LLMs for semantic interoperability and DCAT ingestion.

Enhancing Retrieval & Structured Data

- **Reducing Hallucinations:** Refining retrieval and response mechanisms.
- **Structured Data Integration:** Further experiments on schema-aware query generation for improved extraction.
- **Addressing Schema Semantics:** Improving RAG's understanding of complex queries.

Example

History

RAG

SeTA Collection Name
jrc-data-catalogue:jrc

Search limit 1 2 3 4 5 6 7 8 9 10

LLM Model
GPT@JRC - Nous Hermes 2 Mixtral 8x7B DPO

Temperature 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


System Prompt
You are a chatbot designed to assist scientists. Your primary goal is to help scientists navigate and select data on the JRC Data Catalogue and to guide them in addressing their specific requests. ✕

What data exist on air quality? i

There are several data sets available on air quality in the JRC Data Catalogue. The data sets provide measurements of various air pollutants such as NO, NO2, O3, CO, and SO2 at different locations and time periods. Here are a few examples:

1. "Ambient Air Quality measurements of NO, NO2, O3, CO and SO2 at the JRC Ispra site": This dataset provides daily unvalidated air quality data for the years 2022 and 2023 at the JRC Ispra site in Italy. The data includes measurements of nitrogen oxides (NO and NO2), ozone (O3), carbon monoxide (CO), and sulfur dioxide (SO2). The data can be accessed at the following URL: <http://data.europa.eu/89h/91cb1803-99b0-417b-99a5-6f1532497aef>
2. "Air pollution data from EMEP, LRTAP Convention": This dataset contains air pollution data from the EMEP (European Monitoring and Evaluation Programme) network, which is part of the United Nations Economic Commission for Europe's (UNECE) LRTAP (Long-Range Transboundary Air Pollution) Convention. The data covers a wide range of air pollutants, including particulate matter (PM10, PM2.5), sulfur dioxide (SO2), nitrogen oxides (NOx), ammonia (NH3), and more. The data is available for various European countries and time periods. You can access the dataset at the following URL: <http://data.europa.eu/89h/8d66b161-d7d2-475b-81e3-56c8f7d45223>
3. "Air Quality data from the European Environment Agency (EEA)" This dataset provides air quality data collected by the European Environment Agency (EEA) and its partners in Europe. The data includes measurements of various air pollutants, such as particulate matter (PM10, PM2.5), nitrogen oxides (NOx), sulfur dioxide (SO2), ozone (O3), and more. The data can be accessed through the EEA's Air Quality e-Reporting platform, which is available at the following URL: <https://www.eea.europa.eu/data-and-maps/data/aireporting-8>

These are just a few examples of available air quality data in the JRC Data Catalogue. To find more data sets related to air quality, you can search the catalogue using keywords such as "air quality," "air pollution," and "air pollutants."

 > SEND

Thank you



© European Union, 2024

Unless otherwise noted the reuse of this presentation is authorised under the [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.



EU Science Hub
joint-research-centre.ec.europa.eu