EIFFEL Cognitive Search Engine in GEOSS Platform

Yannis Kopsinis, Dimitris Bliziotis

LIBRA MLI, I-SENSE Group/ICCS

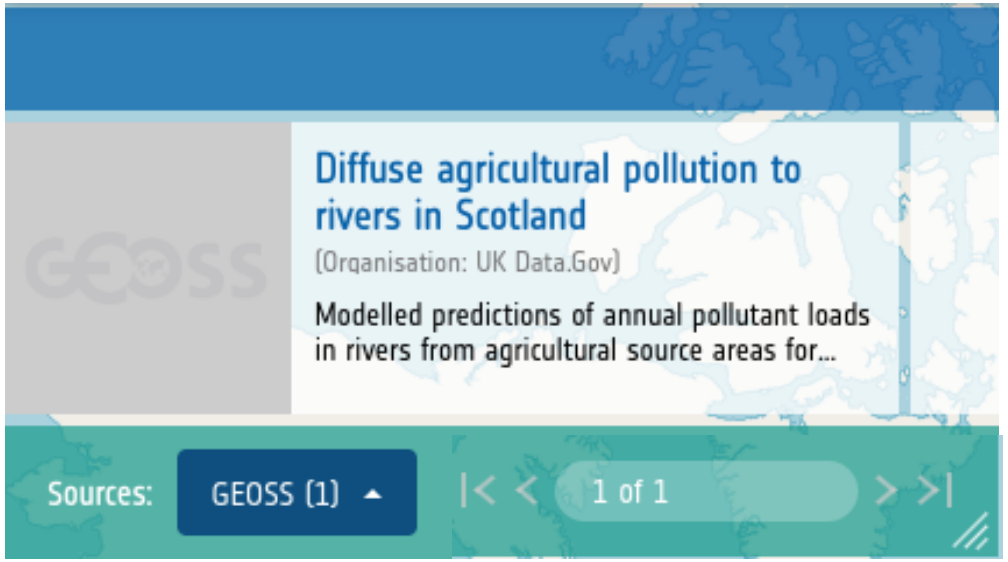
Learn more here:

Data Discovery in Data Spaces/Platforms

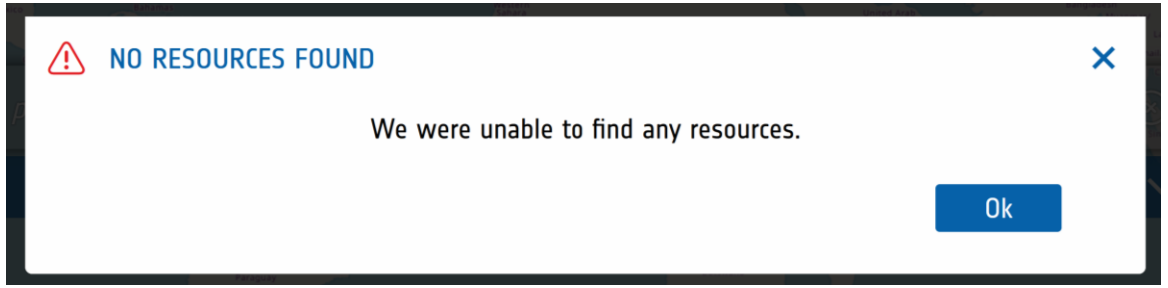
- Data discovery is the ultimate goal of data platforms
- Search queries can be ambiguous in nature
- Search queries are performed on data object metadata
- Is conventional keyword-based search good enough?
- Ontology-based search might bring much overhead to the data publisher
- NLP-based search is a valuable alternative

GEOSS Platform data discovery experience

Agricultural pollution to rivers



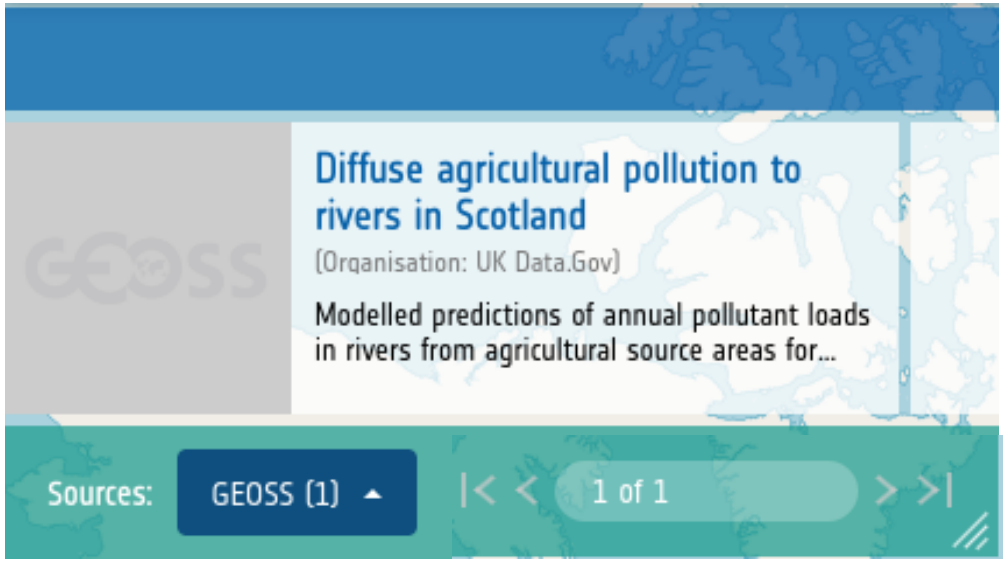
pollution to rivers due to agriculture



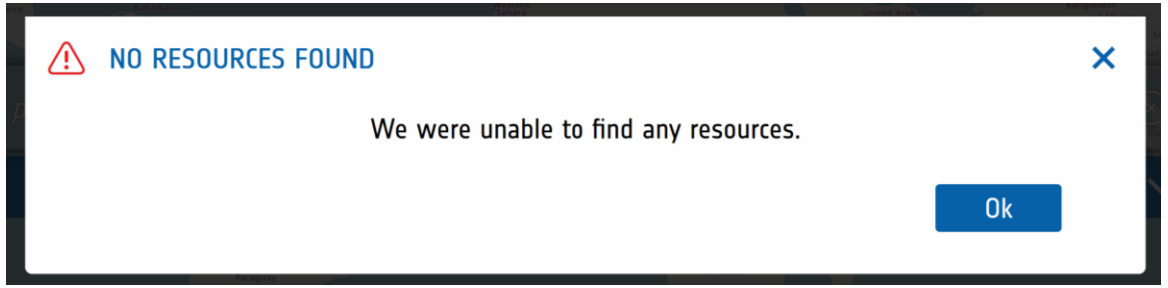
- *The search engine needs to be resilient to rephrased queries*

GEOSS Platform data discovery experience

Pollution AND rivers AND agriculture



Pollution AND inland waters AND agriculture



- *The search engine needs to tackle Semantically similar words and concepts consistently*

HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

The Cognitive search data discovery experience

Agricultural pollution to rivers

Or

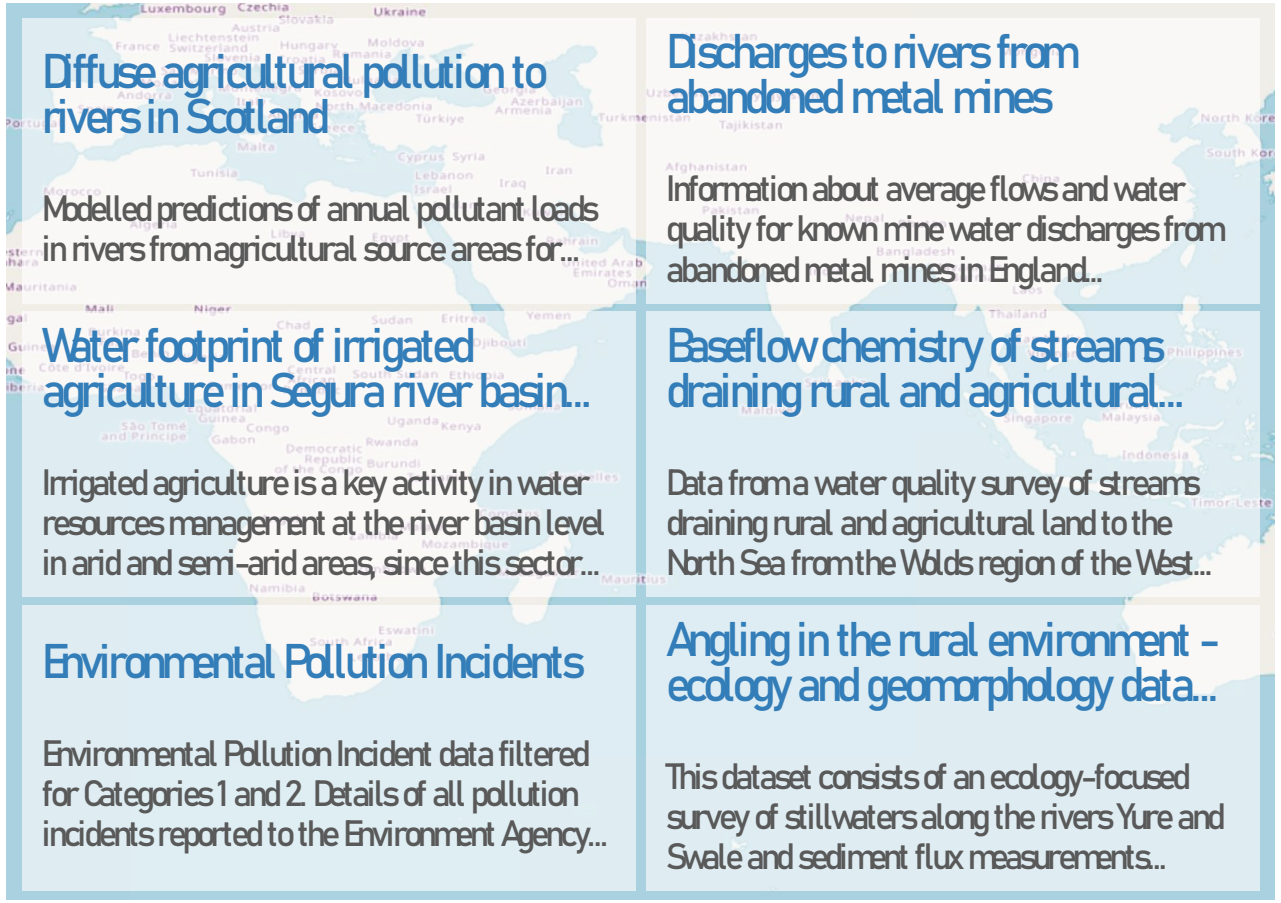
pollution to rivers due to agriculture

Or

Pollution, rivers, agriculture

Or

Pollution, inland waters, agriculture

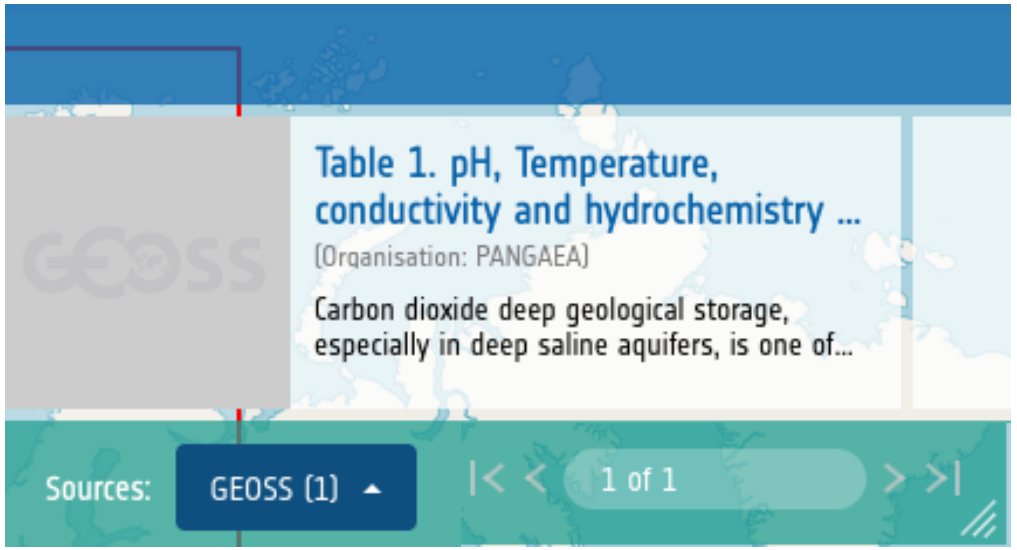


HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS | GSRI GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION

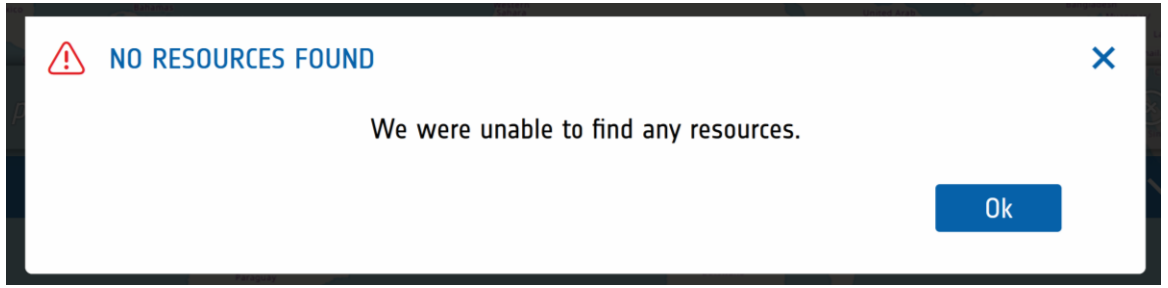
European Commission

GEOSS Platform data discovery experience

greenhouse gases emissions



greenhouse gasses emissions



- *The search engine needs to be resilient to misspelling*

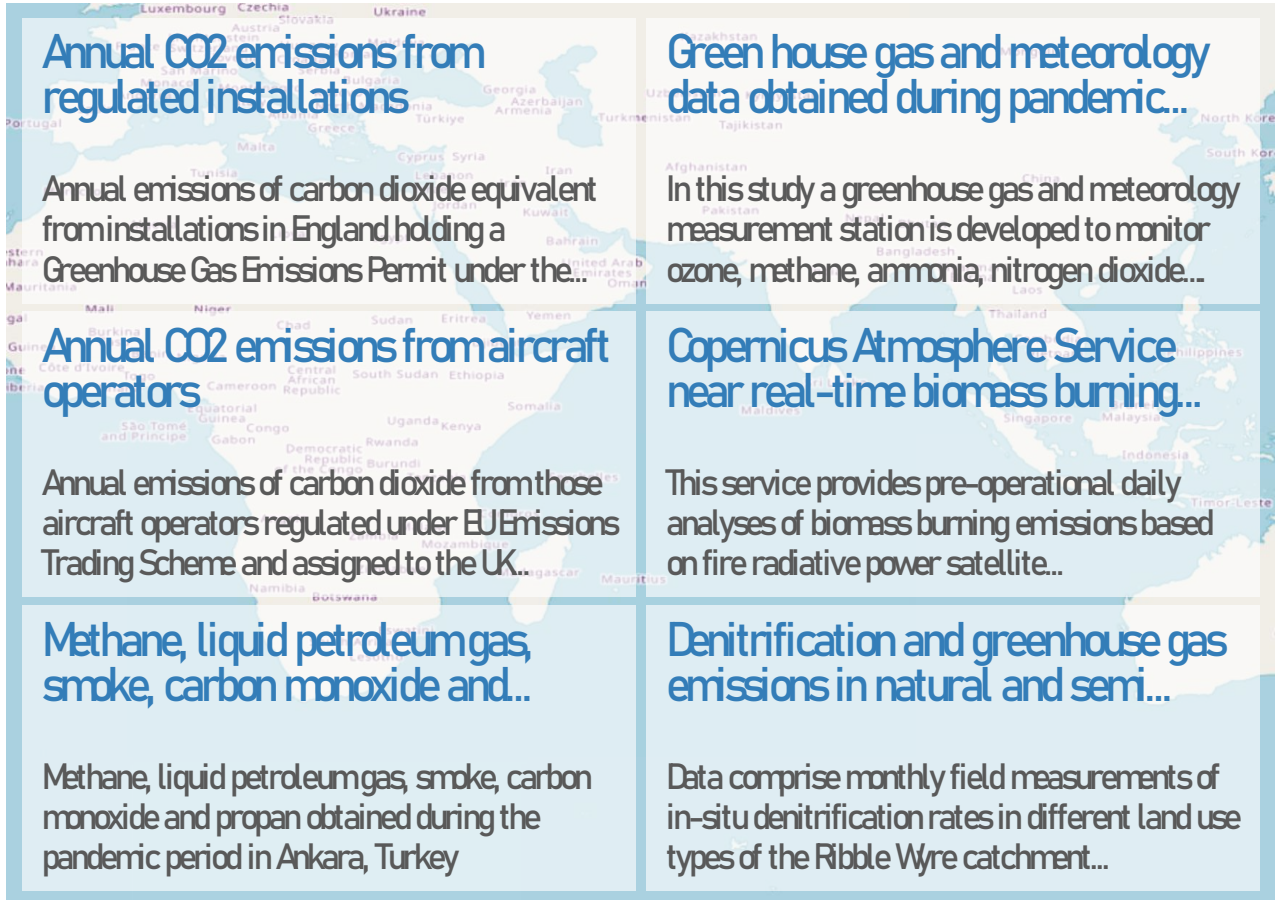
HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

The Cognitive search data discovery experience

greenhouse gases emissions

Or

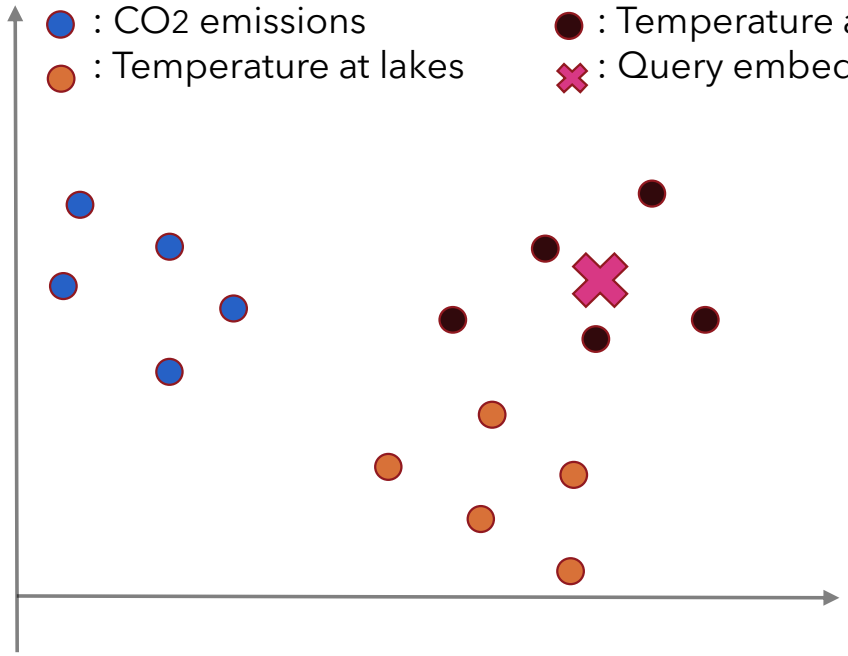
greenhouse gasses emissions



HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GSRI GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Cognitive search engine facts

- Uses an ML-powered **language model** that performs NLP in unstructured text
- The language model inherently performs semantic analysis
- The language model converts text into mathematical vectors
- It can be domain-aware: In Eiffel we aim for CC domain specificity
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top

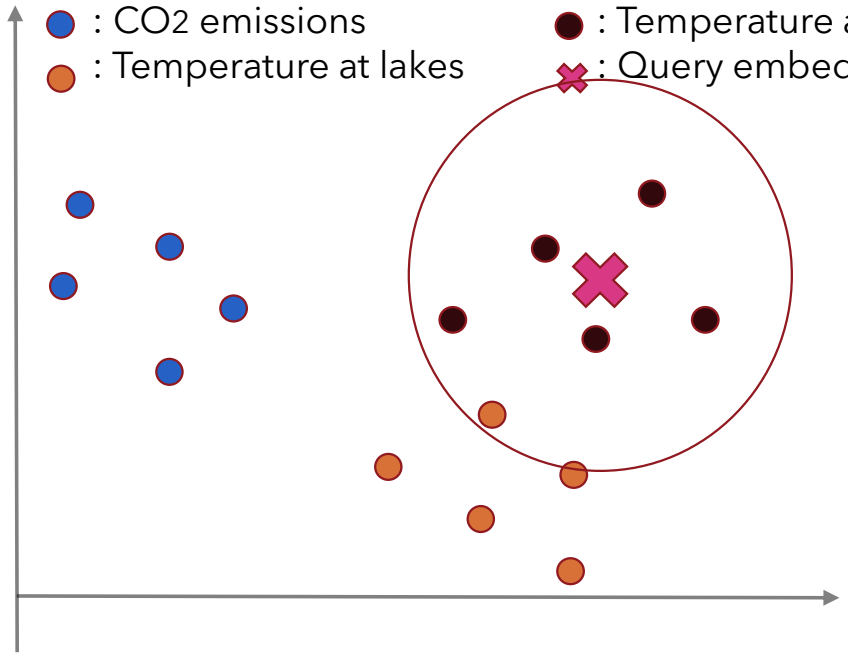


Text embedding encodes words and sentences as numeric vectors

HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GSRI GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Cognitive search engine facts

- Uses an ML-powered **language model** that performs NLP in unstructured text
- The language model inherently performs semantic analysis
- The language model converts text into mathematical vectors
- It can be domain-aware: In Eiffel we aim for CC domain specificity
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top

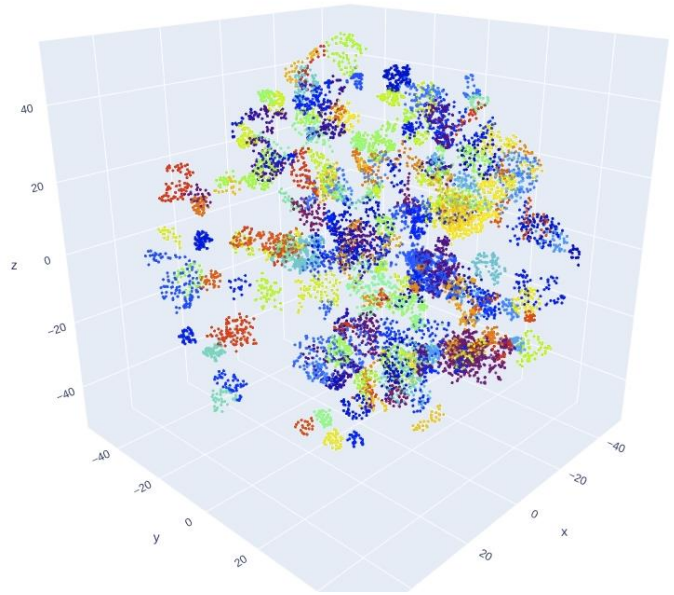


Text embedding encodes words and sentences as numeric vectors

HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Cognitive search engine facts

- Uses an ML-powered **language model** that performs NLP in unstructured text
- The language model inherently performs semantic analysis
- The language model converts text into mathematical vectors
- It can be domain-aware: In Eiffel we aim for CC domain specificity
- Semantic search adds language understanding to search results, promoting the most semantically relevant results to the top

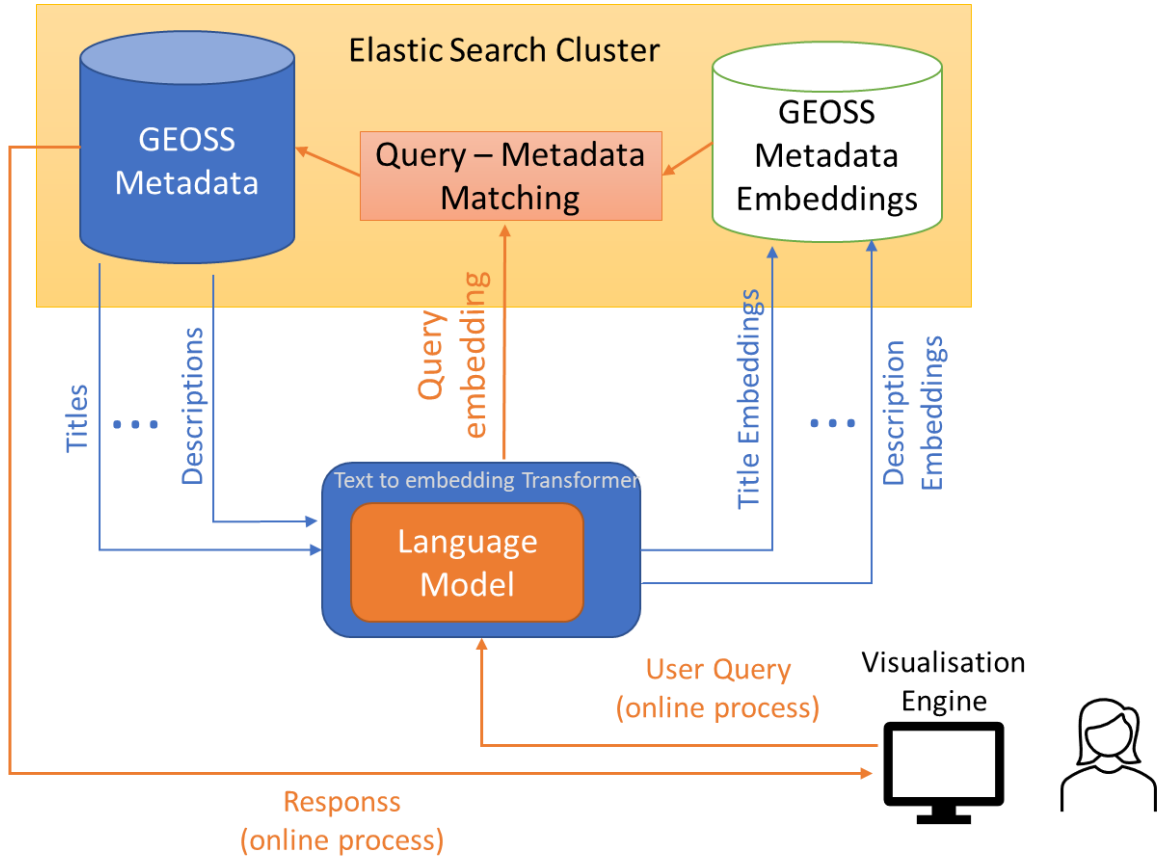


Text embedding encodes words and sentences as numeric vectors

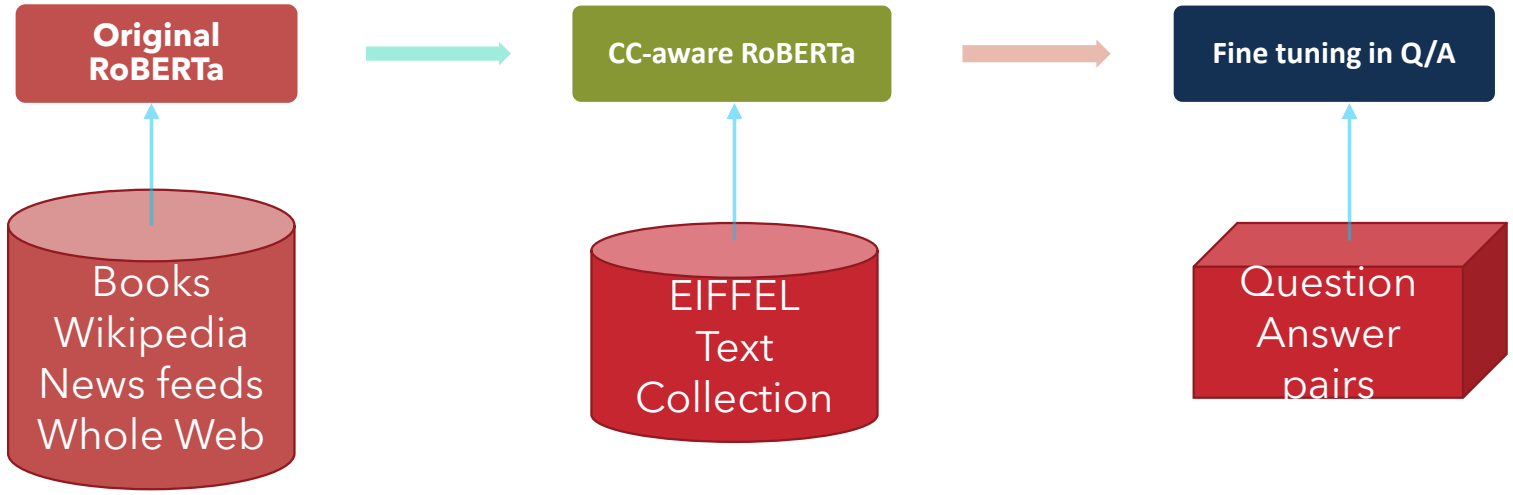
HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Cognitive search pipeline

- The metadata (e.g., titles, descriptions, keywords) pass through the language model to produce metadata embeddings (offline process)
- The user query passes through the language model to produce the query embedding (online process)
- The semantically similar data objects are returned in ranked order
- Elasticsearch stores embeddings and calculates vector similarity fast



Domain-specific language model training

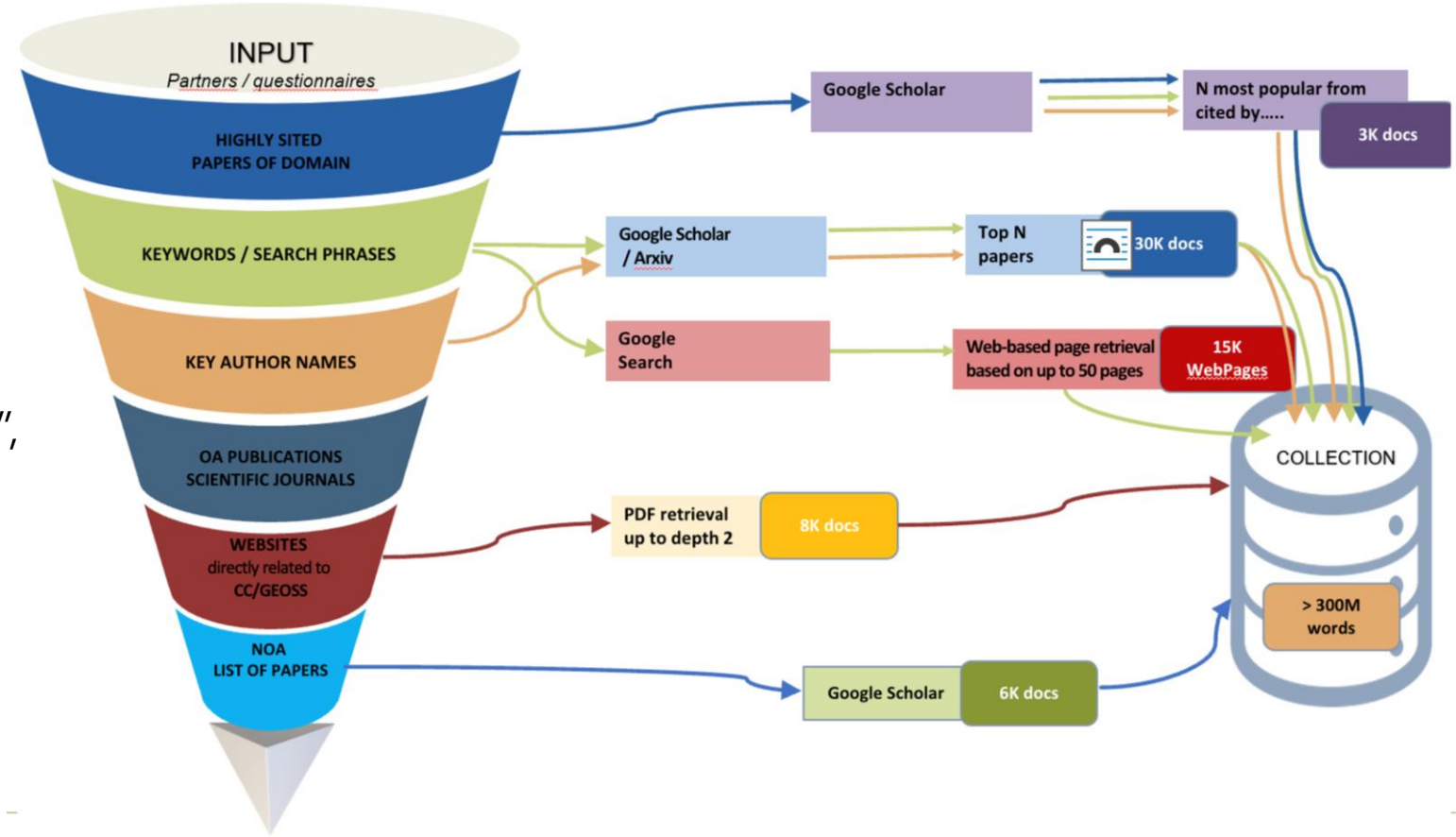


- Chose a baseline language model (Distilled RoBERTa)
- Collect extensive domain-relevant text
- Further train the language model with the domain-relevant text
- Fine tune in Q/A task (optional)

Domain-specific corpus collection (13M sentences)

Newly included terms in the language model

"basin", "distribution", "parameters", "factors", "regions", "environmental", "variables", "emissions", "simulation", "atmospheric", "correlation", "modelling, "measurement", "estimation", "greenhouse", "radiation", "percentage", "climatic", "cooling", "rainfall", "regression", "gases", "pollution", "meteorological", "dioxide", "flux", "anthropogenic", "indicator", "humidity", "ocean", "baseline", "ecosystems", "renewable", "hydrological", "sustainable", "socioeconomic", "CO2"



HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS
GSRI GENERAL SECRETARIAT FOR RESEARCH AND INNOVATION
European Commission

Further advance AI-enabled Cognitive Search

- Metadata curation and augmentation
- Fuse with filtering (e.g., geospatial)
- Perform advanced reranking
- Exploit domain-specific ontologies (Eiffel ontology)

Thank you!

The EIFFEL Consortium

Dimitris Bliziotis, dimitris.bliziotis@iccs.gr, ICCS

Yannis Kopsinis, yannis.kopsinis@libramli.ai, LIBRA