



## EIFFEL Cognitive Search Engine in GEOS Platform

Yannis Kopsinis, Dimitris Bliziotis

LIBRA MLI, I-SENSE Group/ICCS



Learn more here:



ATHENS 7-9 DECEMBER 2022



## 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

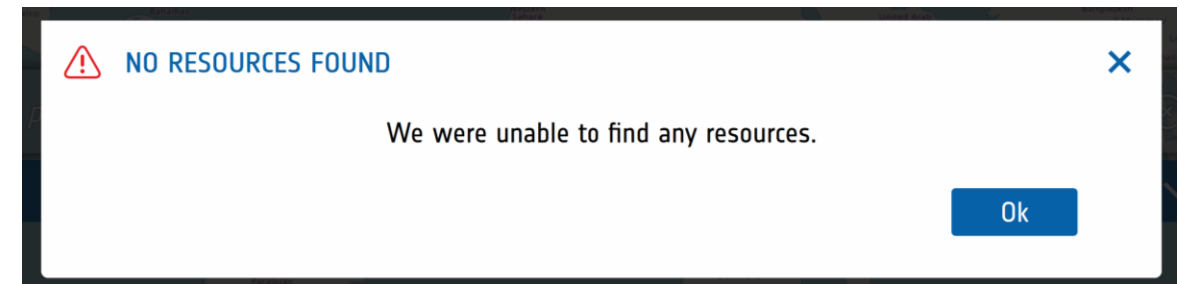


## GEOSS Platform data discovery experience

*Agricultural pollution to rivers*



*pollution to rivers due to agriculture*



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

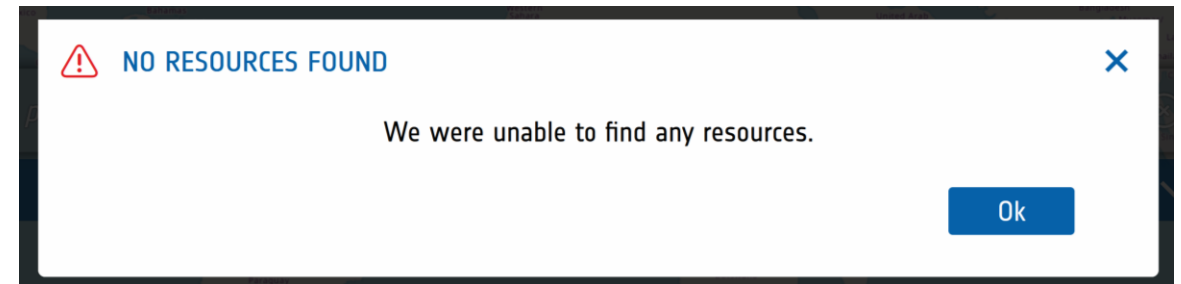


## GEOSS Platform data discovery experience

*Pollution AND rivers AND agriculture*



*Pollution AND inland waters AND agriculture*



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



## The Cognitive search data discovery experience

*Agricultural pollution to rivers*

*Or*

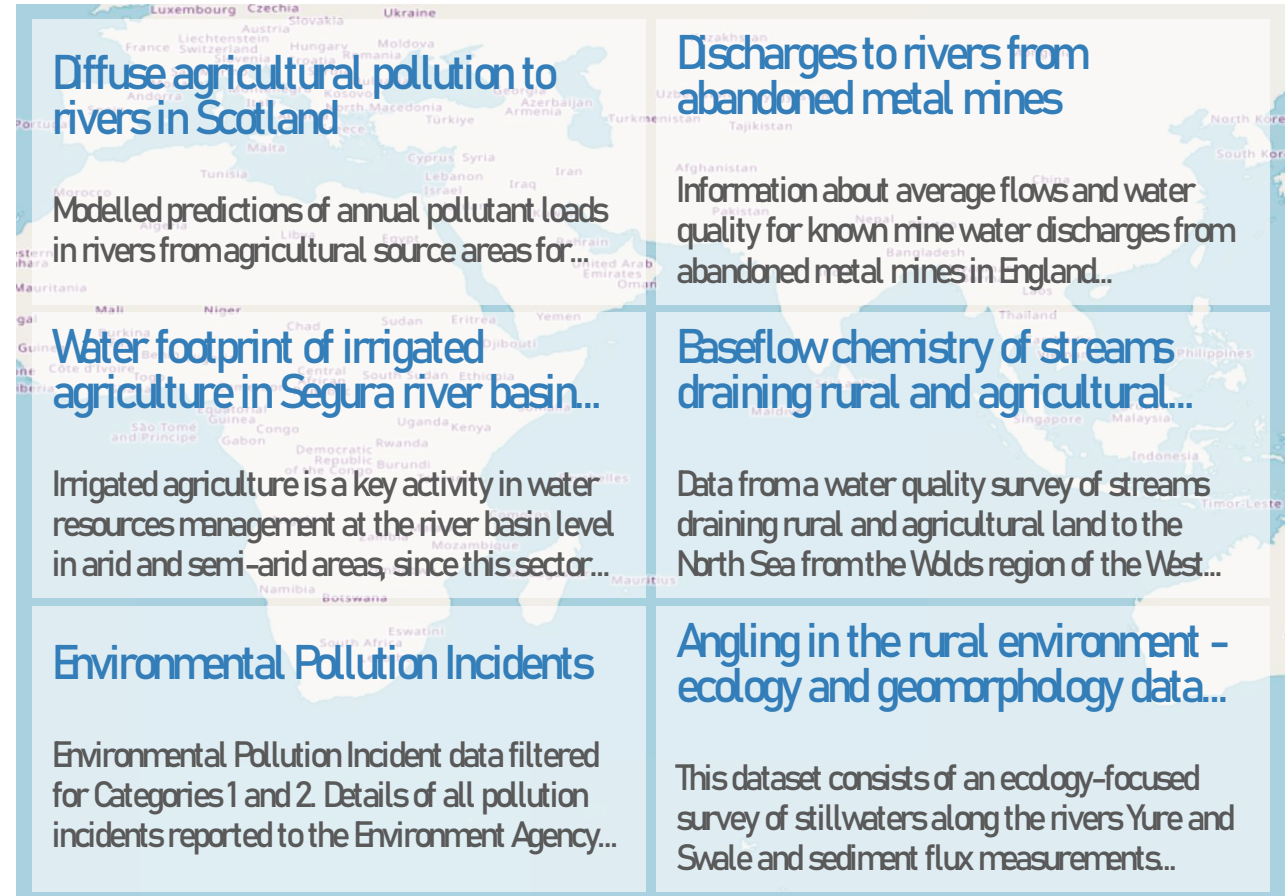
*pollution to rivers due to agriculture*

*Or*

*Pollution, rivers, agriculture*

*Or*

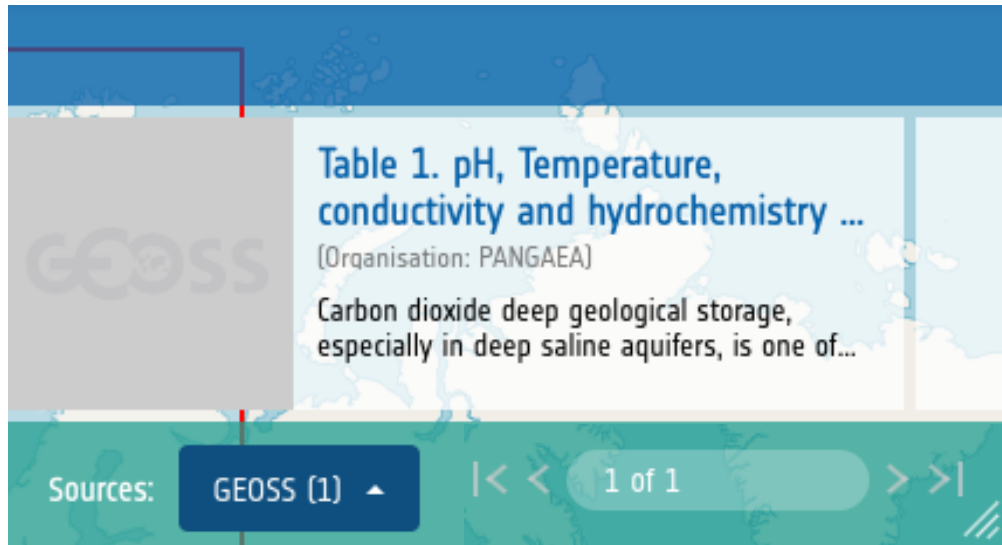
*Pollution, inland waters, agriculture*



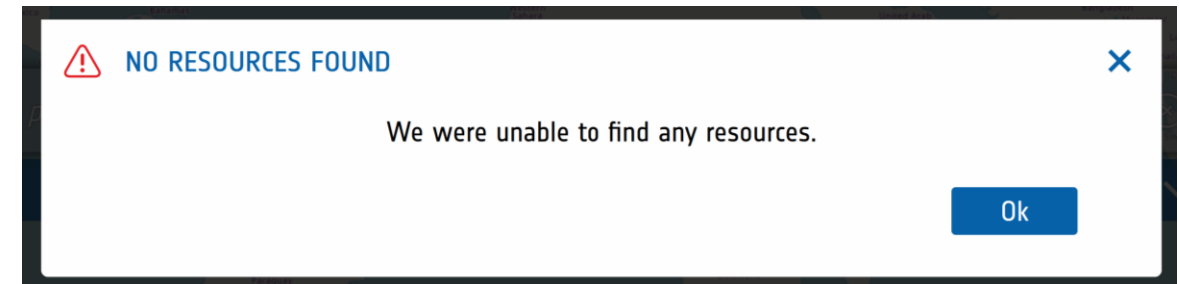


## GEOSS Platform data discovery experience

*greenhouse gases emissions*



*greenhouse gasses emissions*



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





## The Cognitive search data discovery experience

*greenhouse gases emissions*

*Or*

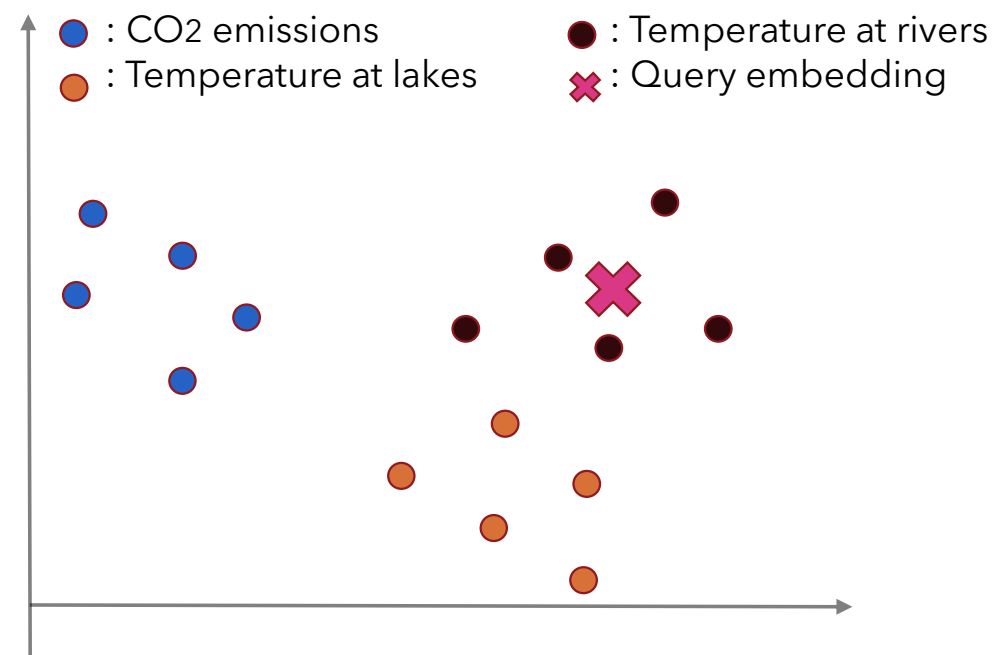
*greenhouse gasses emissions*

<p><b>Annual CO2 emissions from regulated installations</b></p> <p>Annual emissions of carbon dioxide equivalent from installations in England holding a Greenhouse Gas Emissions Permit under the...</p>	<p><b>Green house gas and meteorology data obtained during pandemic...</b></p> <p>In this study a greenhouse gas and meteorology measurement station is developed to monitor ozone, methane, ammonia, nitrogen dioxide...</p>
<p><b>Annual CO2 emissions from aircraft operators</b></p> <p>Annual emissions of carbon dioxide from those aircraft operators regulated under EU Emissions Trading Scheme and assigned to the UK...</p>	<p><b>Copernicus Atmosphere Service near real-time biomass burning...</b></p> <p>This service provides pre-operational daily analyses of biomass burning emissions based on fire radiative power satellite...</p>
<p><b>Methane, liquid petroleum gas, smoke, carbon monoxide and...</b></p> <p>Methane, liquid petroleum gas, smoke, carbon monoxide and propan obtained during the pandemic period in Ankara, Turkey</p>	<p><b>Denitrification and greenhouse gas emissions in natural and semi...</b></p> <p>Data comprise monthly field measurements of in-situ denitrification rates in different land use types of the Ribble Wyre catchment...</p>



## 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



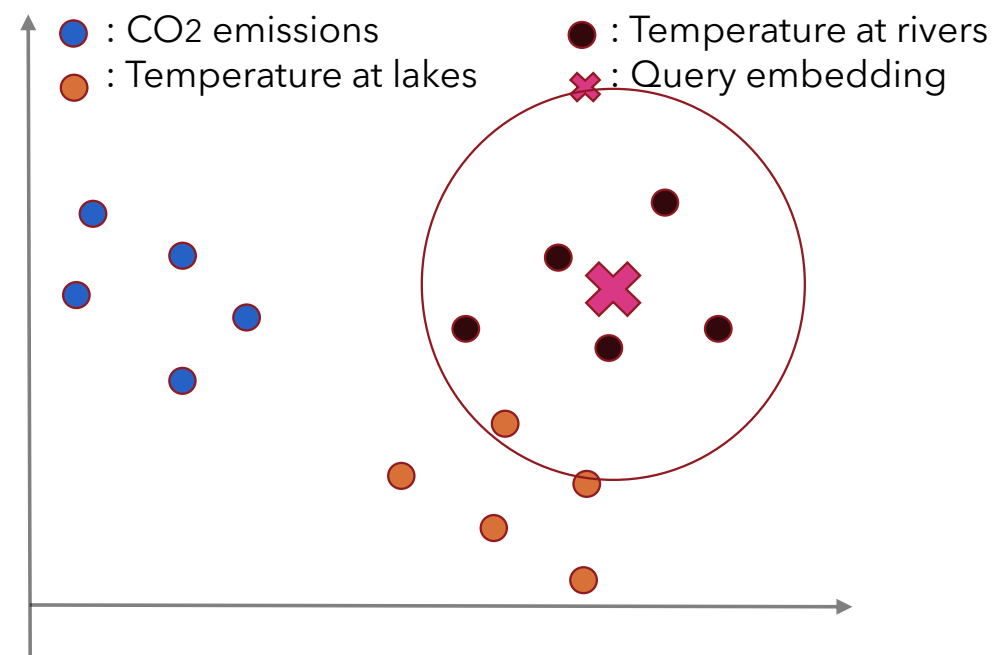
Text embedding encodes words and sentences as numeric vectors





## 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

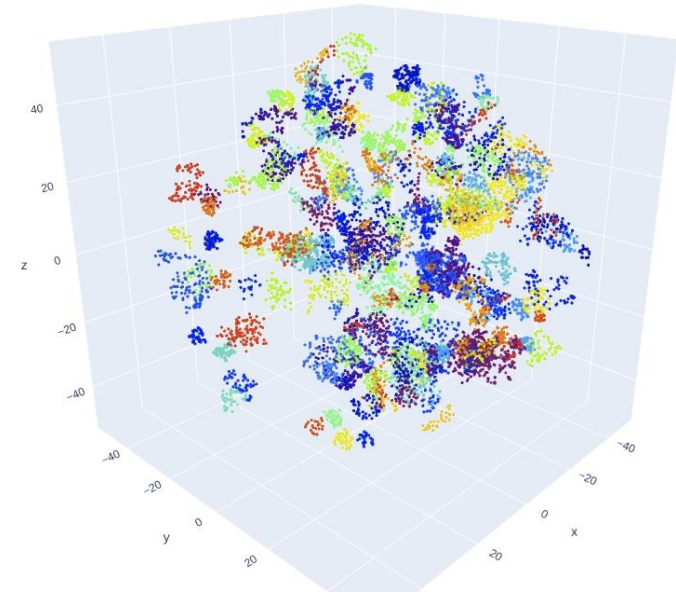


Text embedding encodes words and sentences as numeric vectors



## 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

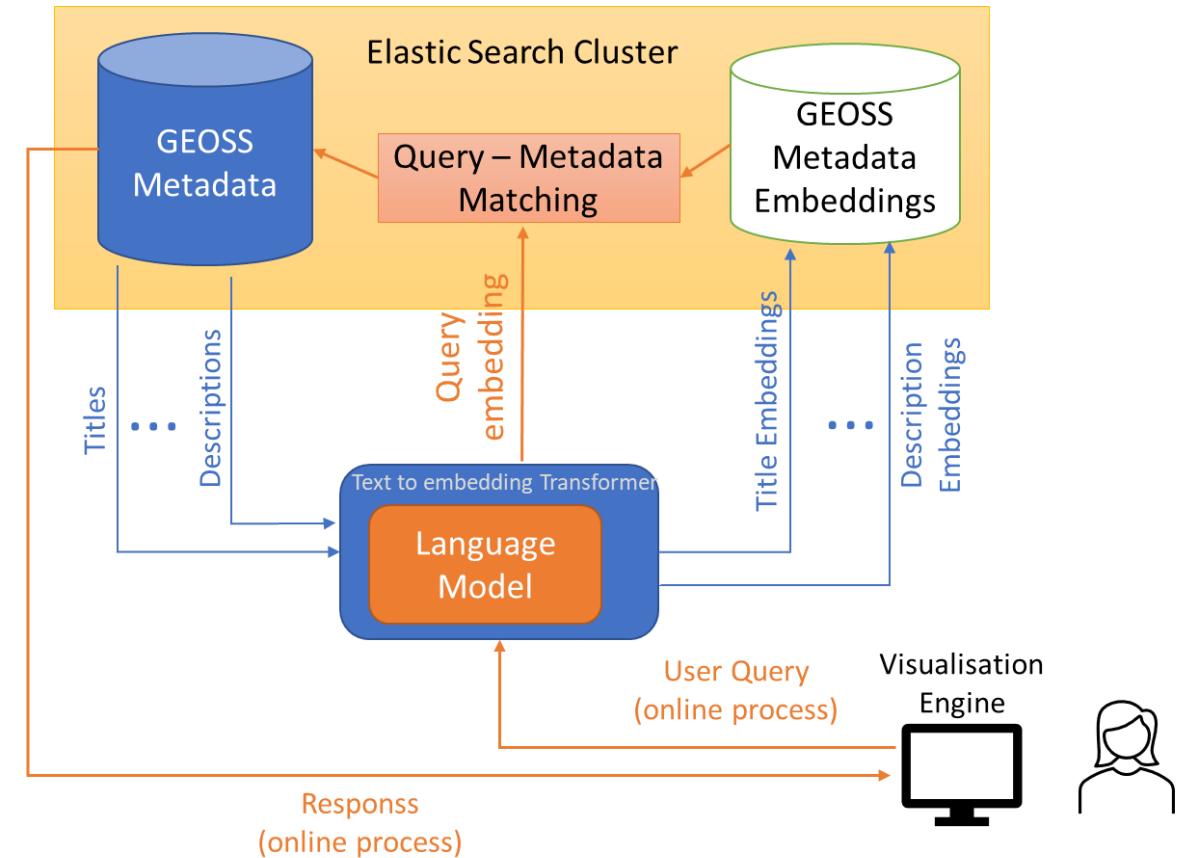


Text embedding encodes words and sentences as numeric vectors



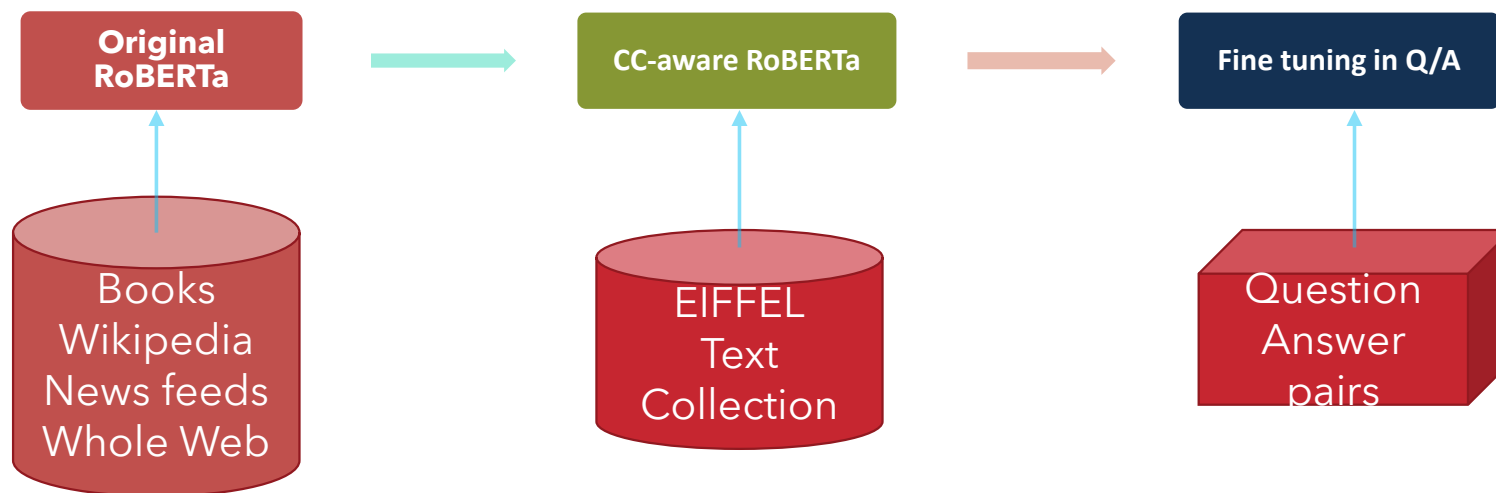
## Cognitive search pipeline

- 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



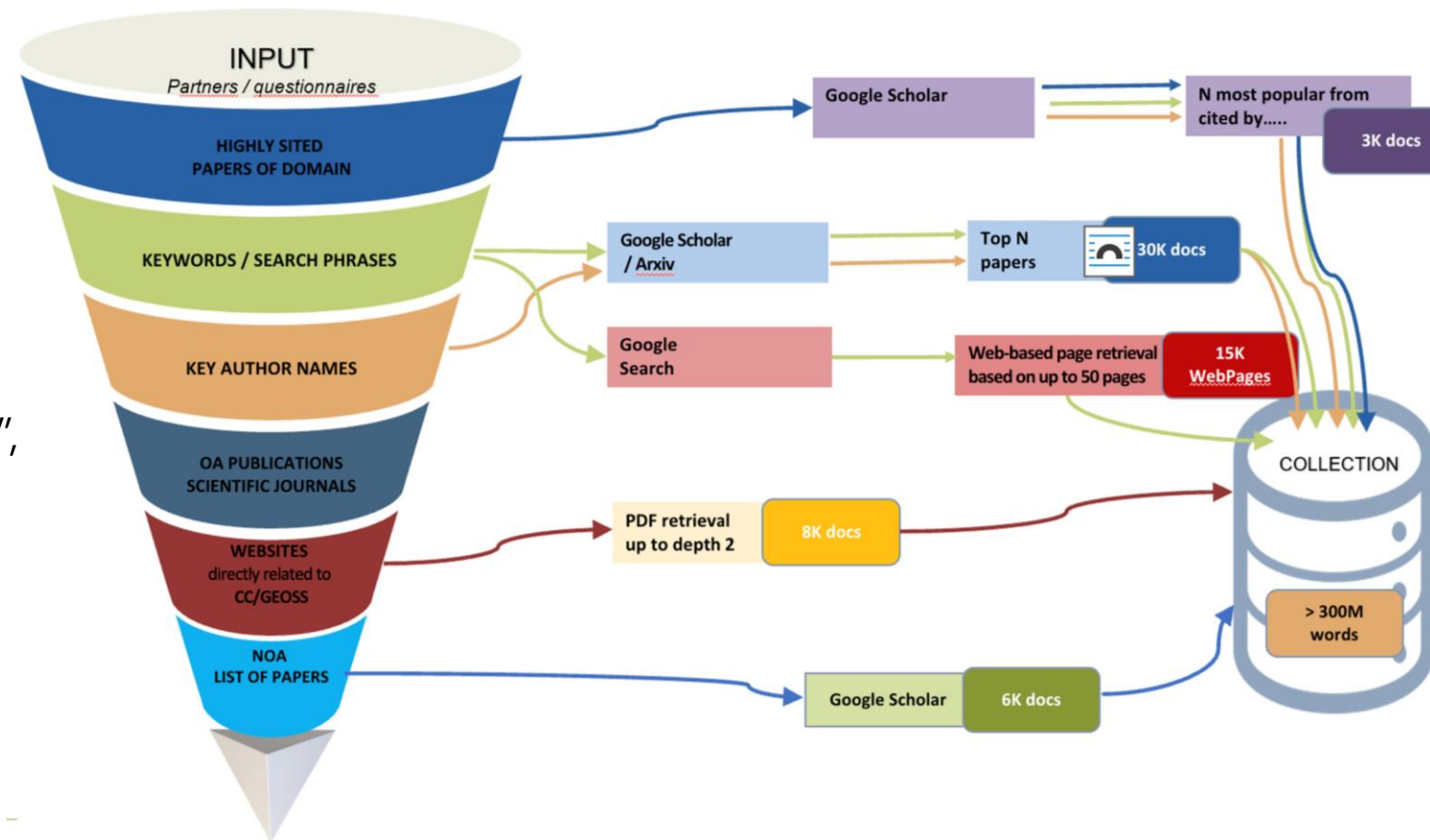
- Chose a baseline language model (Distilled RoBERTa)
- Collect extensive domain-relevant text
- Further train the language model with the domain-relevant text
- Fine tune in Q/A task (optional)



## 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"





## 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)





## The EIFFEL Consortium



*Thank  
you!*



Dimitris Bliziotis, [dimitris.bliziotis@iccs.gr](mailto:dimitris.bliziotis@iccs.gr), ICCS

Yannis Kopsinis, [yannis.kopsinis@libramli.ai](mailto:yannis.kopsinis@libramli.ai), LIBRA