Faculty of Engineering, University of Porto
2024

BIOFINDER

Information Processing and Retrieval

Nuno França; João Tomás Teixeira; Rodrigo Esteves; Isabel Silva

MILESTONE #3

- Two new queries were developed.
- New query field and semantic improvements.
- Simple and responsive UI developed.

CONCEPTUAL MODEL

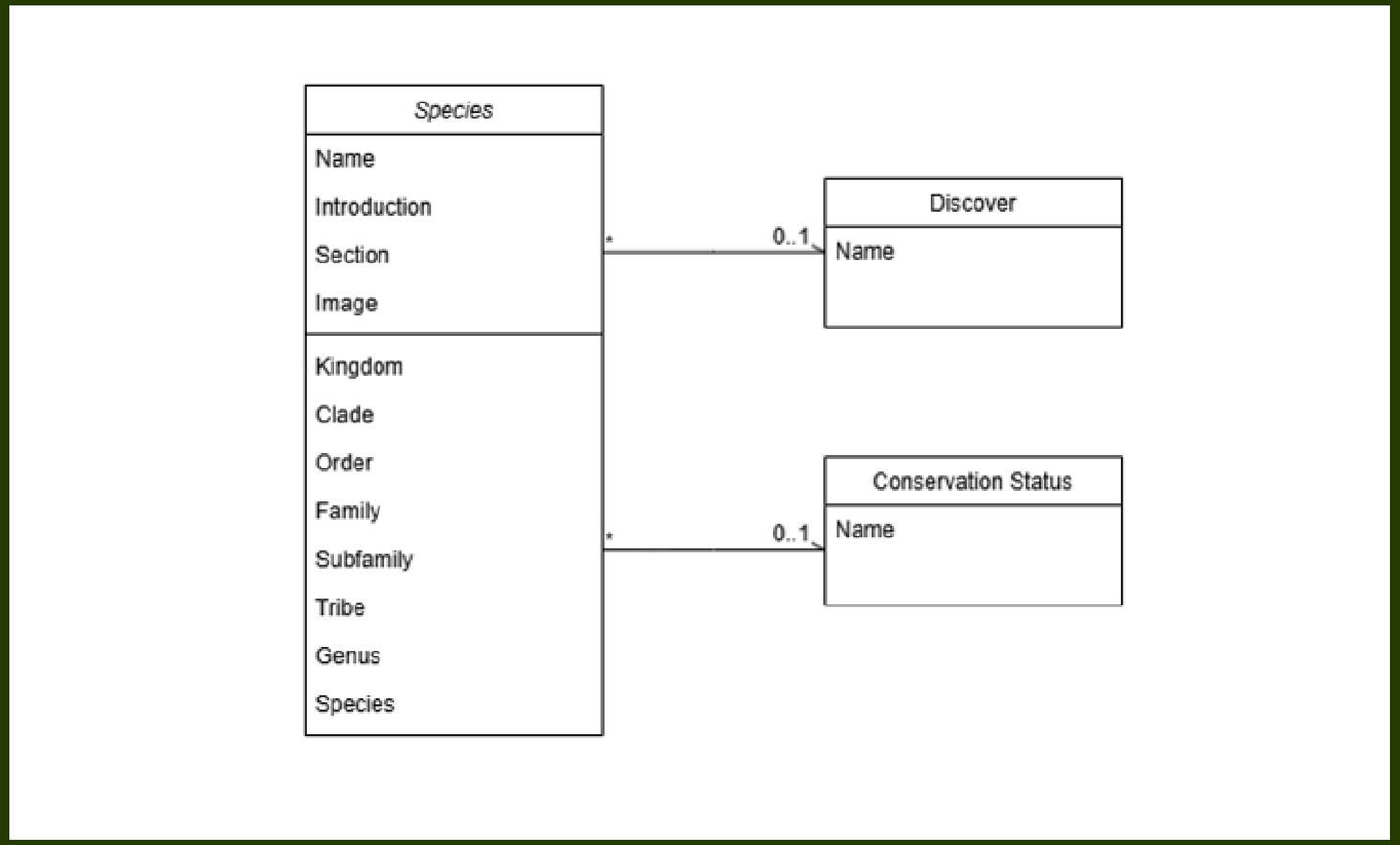


Figure 1: Conceptual model.

DATA PIPELINE

- The web-scrapping process was done again.
- Due to the 429 HTTP Error, the data extracted from Wikipedia was not complete.

	Complex system	New system
Total entries	183 171	328 223
Duplicate entries	36 439	84 248
Final entries	131 828	215 437

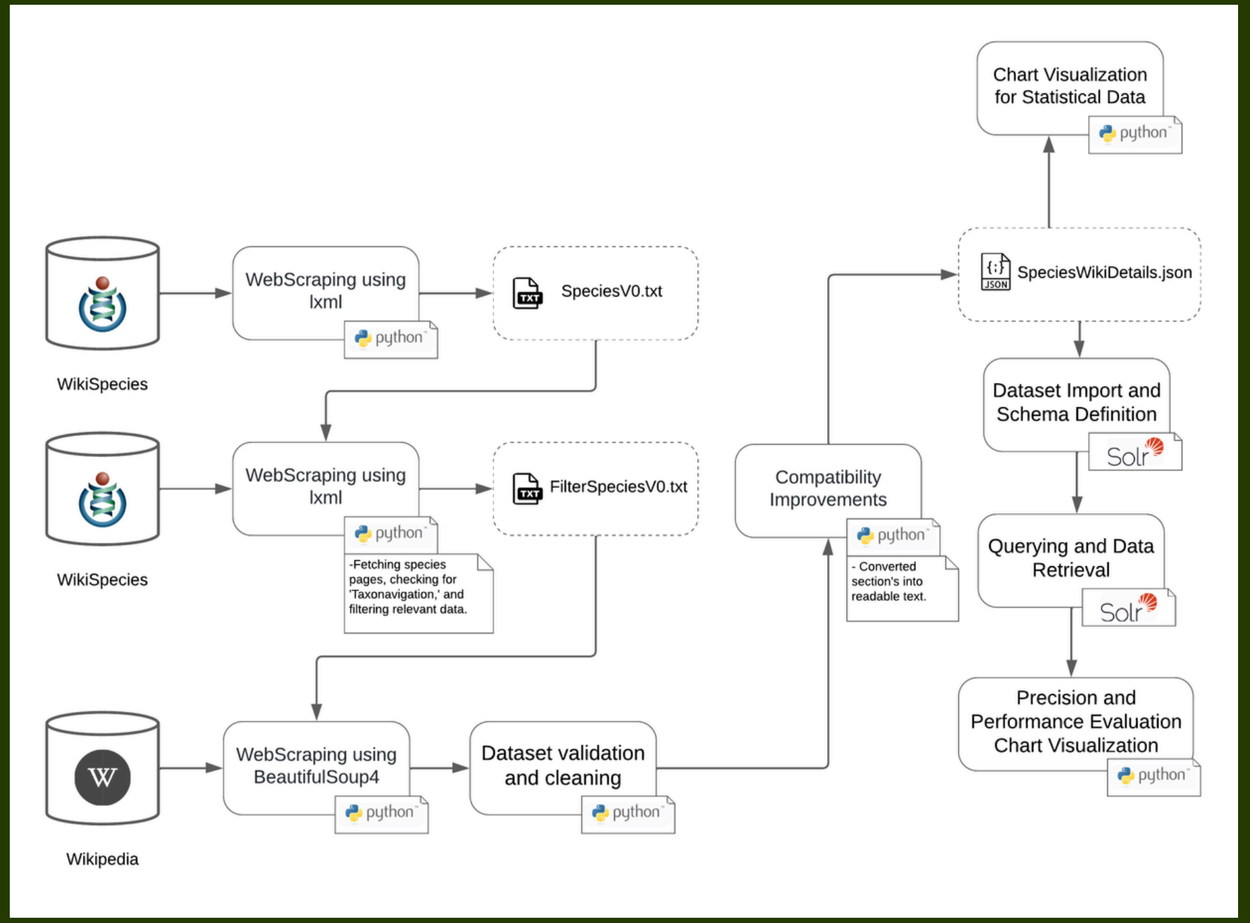


Figure 2: Data Pipeline.

QUERIES

Information Need Query Developed	Information Need Query Developed
What are the largest and huge animals on the planet Earth?	(largest huge size animal)
What is the lifespan in years of the aquatic species that frequent the Atlantic Ocean?	(lifespan years) AND (Atlantic AND Aquatic)
I want to know the endangered species that currently inhabit Portugal.	(endangered species Portugal)
I’m visiting Australia, I want to know the dangerous species to avoid.	(venomous dangerous Australia)
I want to know the evasive species in the planet that have a negative impact in the ecosystem	(species negative impact) AND invasive
I want to know species that have a night activity or nocturnal behavior	(activity behavior) AND (night nocturnal)

SUBSET

- The complete dataset with the semantic improvements had a size of around 4 GigaBytes of information
- The subset created has 300 distinct values.
- The creation of the dataset could not be randomized, due to the diversity of values that are complete opposites in our queries.
  - Instead, it was decided to use the previously developed system and, for each query, we chose to:
    - select 30 true positive results
    - select 20 false positives
- No duplicate results
- The pool of true positive results for each query does not overlap with the true positive results for any other query.
- This method of evaluation was improved by the development of a script, enabling fully automated evaluation.
- A re-evaluation of the subset with the Complex System was also performed.

SCHEMA

The initial schema, provided basic functionality by indexing all attributes using a simple tokenizer and ensuring lowercase processing. However, to improve search flexibility and retrieval accuracy, a more advanced schema was developed.

For this milestone, the schema was enhanced in three key aspects:

- Incorporation of semantic embeddings
- Improved synonym dictionary
- Addition of a new filter: StopFilterFactory.

To incorporate Semantic embeddings in the schema, the Python Library *Sentence_Transformers* was used to change the dataset with the respective embeddings.

Advanced Schema
StandardTokenizerFactory
LowerCaseFilterFactory
PorterStemFilterFactory
SynonymGraphFilterFactory

A synonym dictionary was created with help from the NLTK Python Library, where all unique words in the dataset were analyzed to generate an extensive list of synonyms for each word. Additionally, a stopwords dictionary was developed with NLTK, enabling the removal of unnecessary words to further refine the document's search.

NEW PARAMETERS

- Similarly to the improvements done for the complex system developed previously, additional parameter improvements were implemented to increase the precision and recall of the data.

- These parameters were curated and adjusted specifically for each query.

- While a universal parameter configuration for every query was not created, the results gave a better understanding of how each parameter works and the adjustments required to develop a generalized configuration.

- qf (Query Fields): Specifies the fields to search and their relative importance.
- bq (Boost Query): Applies additional boosts to specific terms or conditions.
- pf(Phrase Fields): Boosts results where query terms appear together as phrases in specific fields.
- ps (Phrase Slop): Controls the flexibility for matching query terms in phrases.
- qs (Query Slop): Defines the flexibility for word order in phrase matching.
- tie (Tie Breaker): Balances the scoring contributions from multiple fields.
- mm (Minimum Match): Specifies how many query terms must match in a document.

FIELDS + SEMANTIC IMPROVEMENTS

- To further analyze the results obtained, it was decided to join, both query parameters improvements and the schema improvements, to understand if there were differences in the final results of precision and recall.

Q1 - LARGEST ANIMALS ON EARTH

Information need:
What are the largest and huge animals on the planet Earth?

Relevance:
The objective of this query is to know which are the largest animals that live on our planet.

Metric	Complex System	Param. Improv.	Schema Improv.	Schema + Param.
MAP	0.8502	0.9113	0.9071	0.9388
P@10	0.9000	1.0000	0.9000	1.0000
P@30	0.7667	0.7667	0.8000	0.8333

Q1 - LARGEST ANIMALS ON EARTH

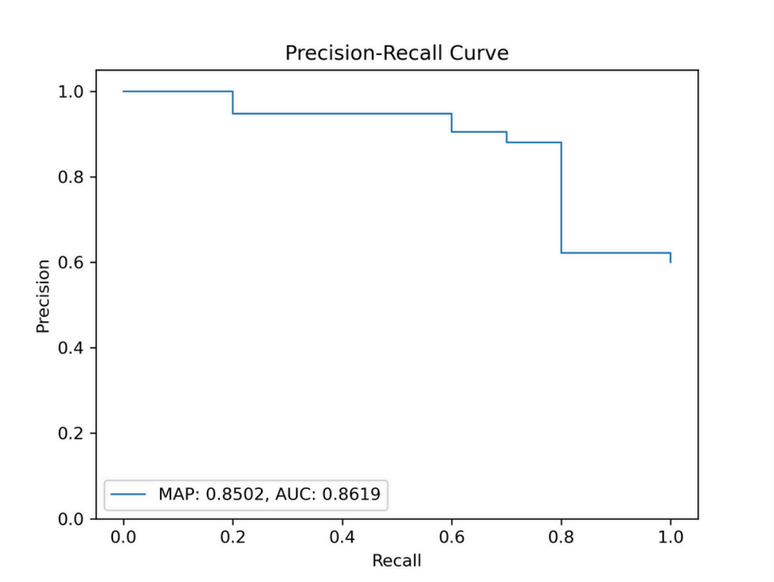


Figure 3: Precision-Recall original schema.

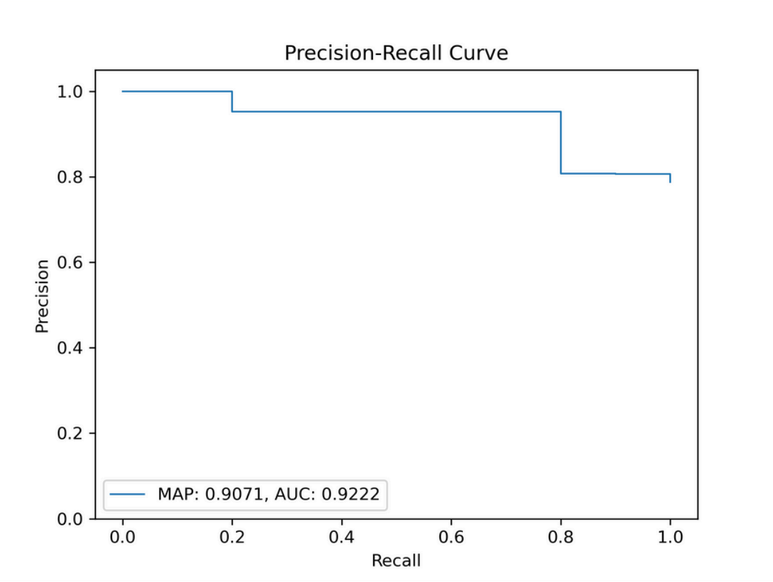


Figure 4: Precision-Recall improved semantics schema.

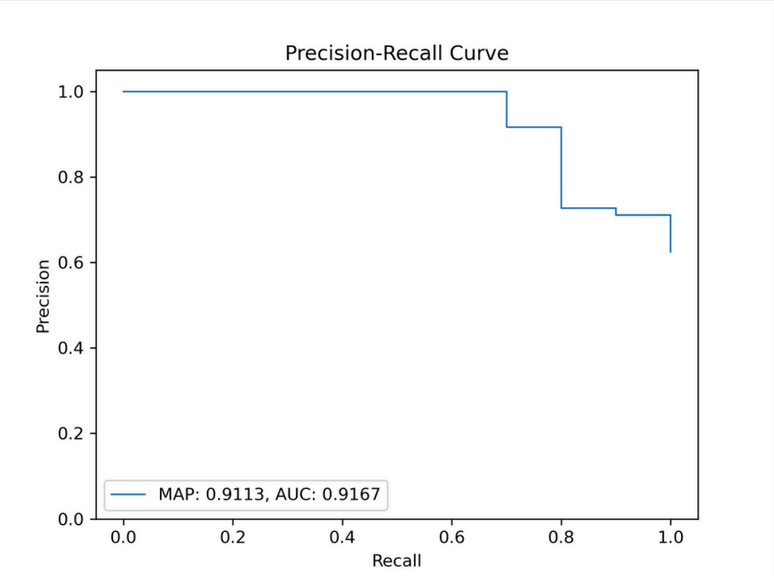


Figure 5: Precision-Recall query improvements.

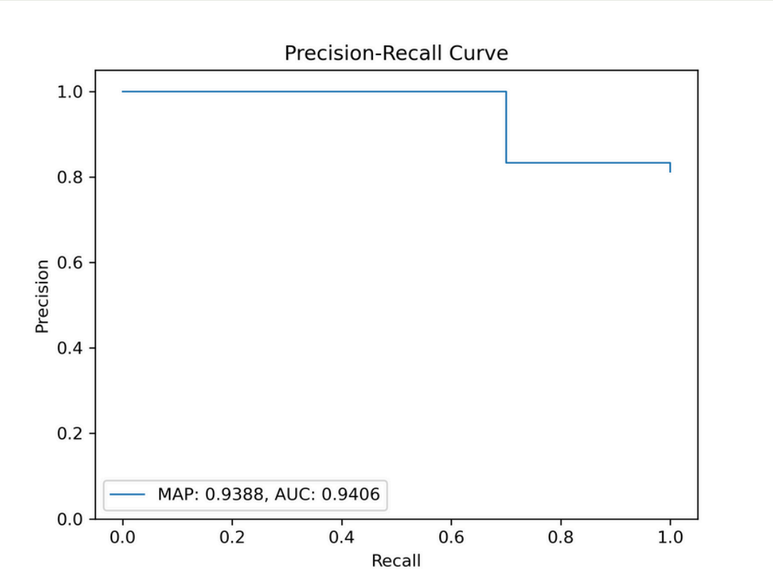


Figure 6: Precision-Recall query and schema improvements.

Q3 - ENDANGERED SPECIES IN PORTUGAL

Information need:
I want to know the endangered species that currently inhabit Portugal.

Relevance:
It is important to know the species that are in danger in Portugal, so that measures can be taken to safeguard the individuals that still exist.

Metric	Complex System	Param. Improv.	Schema Improv.	Schema + Param.
MAP	0.7637	0.8347	0.8113	0.8223
P@10	0.9000	0.9000	1.0000	0.9000
P@30	0.6333	0.7000	0.6667	0.7000

Q3 - ENDANGERED SPECIES IN PORTUGAL

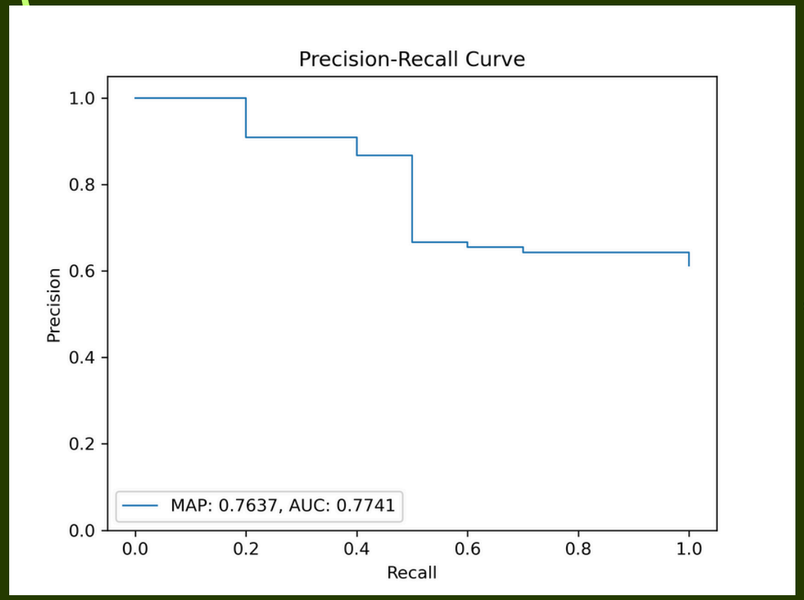


Figure 7: Precision-Recall original schema.

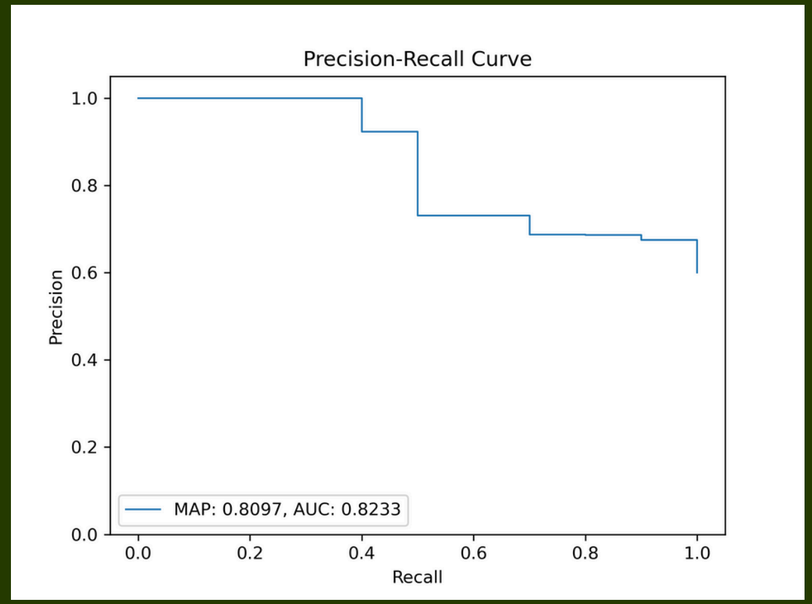


Figure 8: Precision-Recall improved semantics schema.

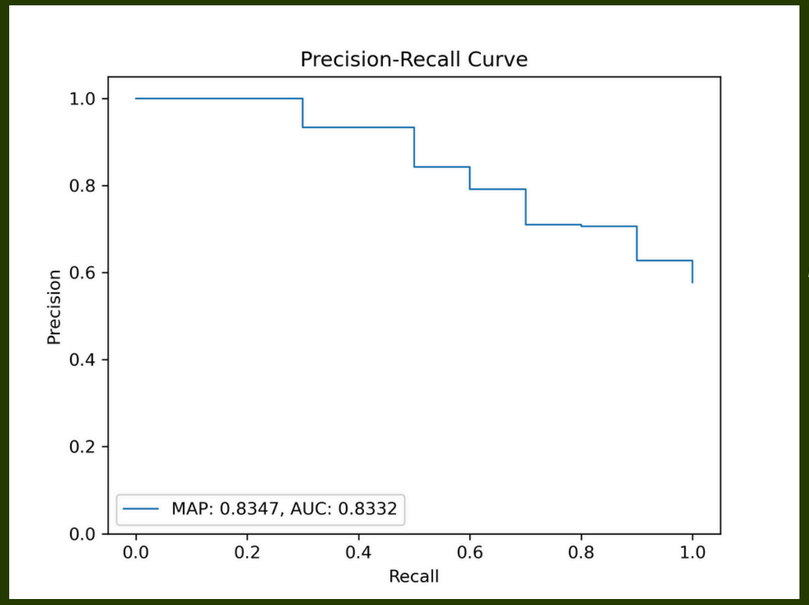


Figure 9: Precision-Recall query improvements.

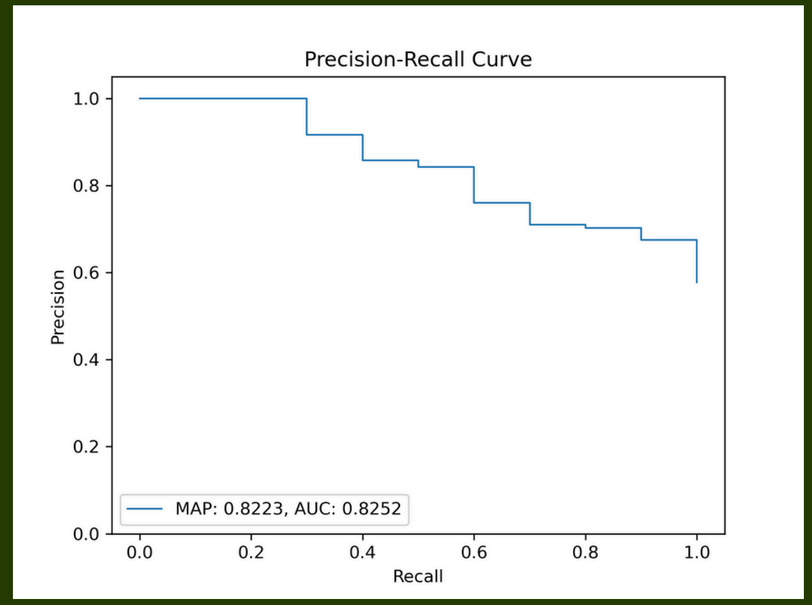


Figure 10: Precision-Recall query and schema improvements.

Q5 - EVASIVE SPECIES WITH A NEGATIVE IMPACT

Information need:
I want to know the evasive species in the planet that have a negative impact in the ecosystem.

Relevance:
While doing a study to verify the invasive species on the planet, I also want to check the impact they have in the ecosystem.

Metric	Complex System	Param. Improv.	Schema Improv.	Schema + Param.
MAP	0.7000	0.9006	0.9484	0.9521
P@10	0.7000	0.9000	1.0000	1.0000
P@30	0.8000	0.8000	0.7000	0.7333

Q5 - EVASIVE SPECIES WITH A NEGATIVE IMPACT

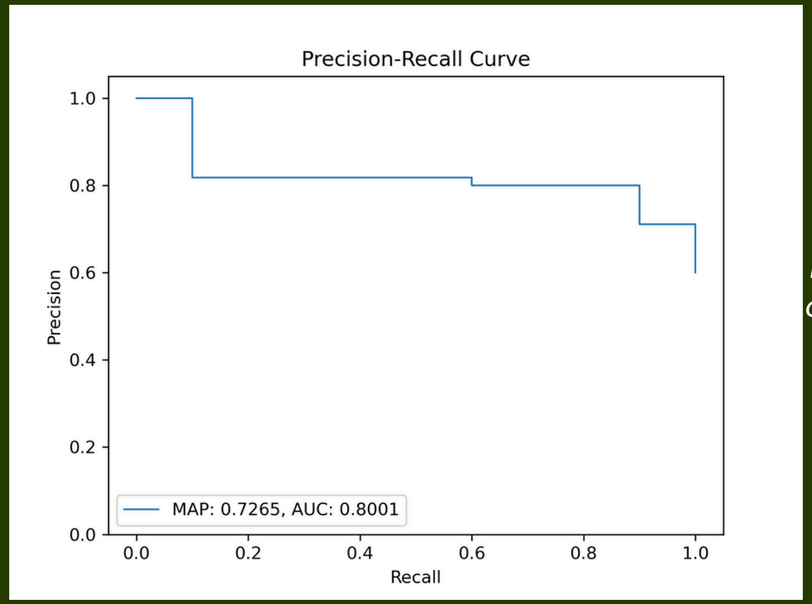


Figure 11: Precision-Recall original schema.

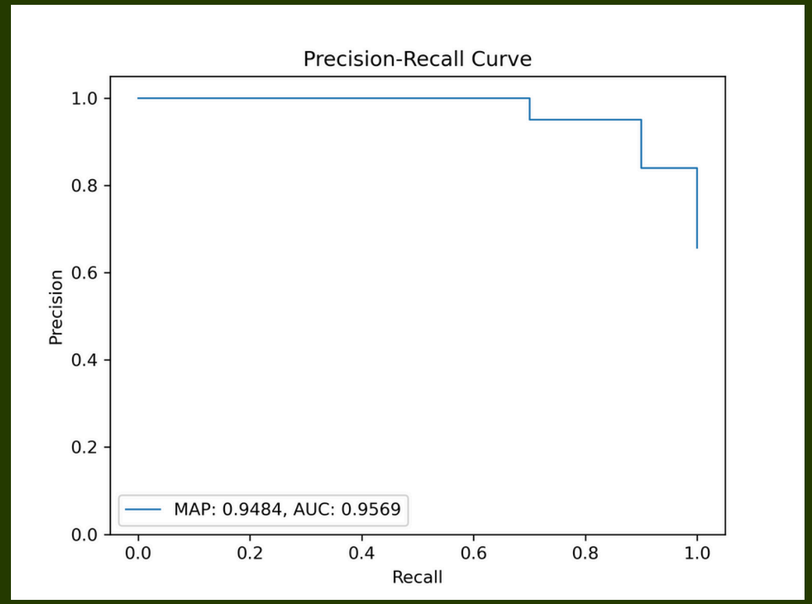


Figure 12: Precision-Recall improved semantics schema.

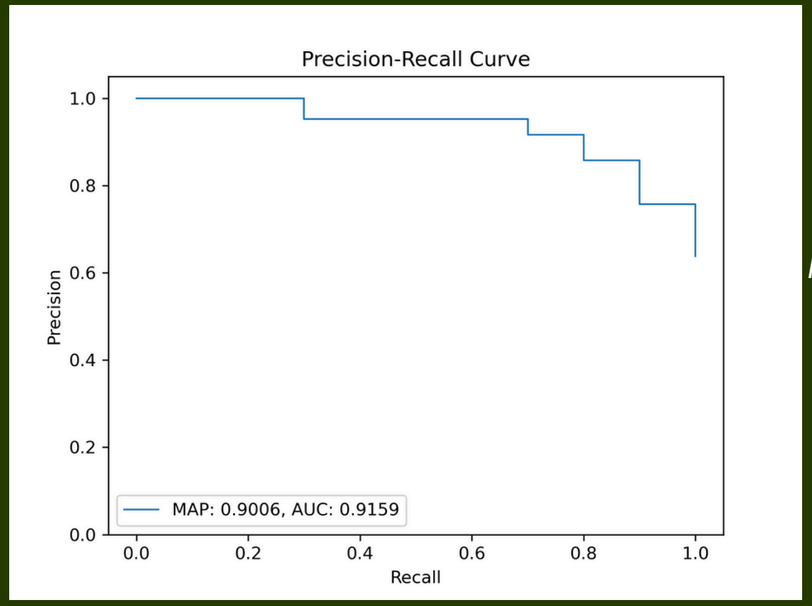


Figure 13: Precision-Recall query improvements.

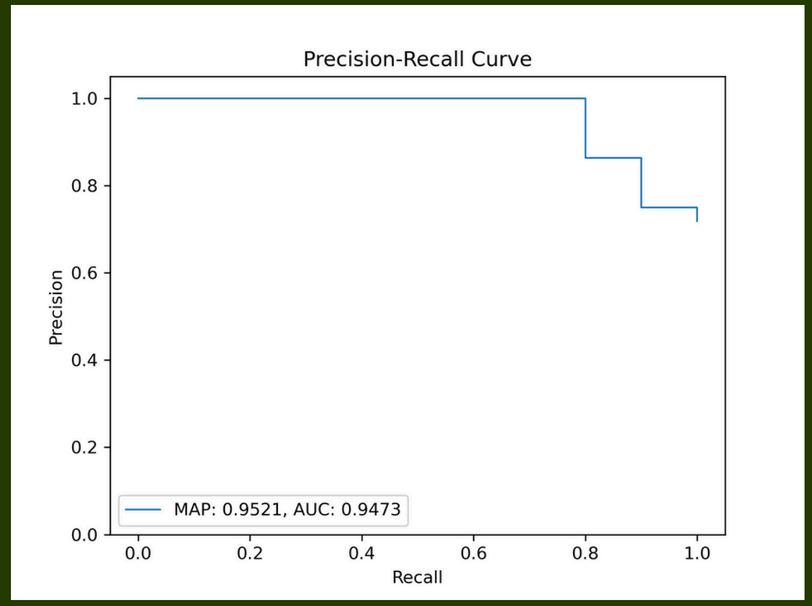
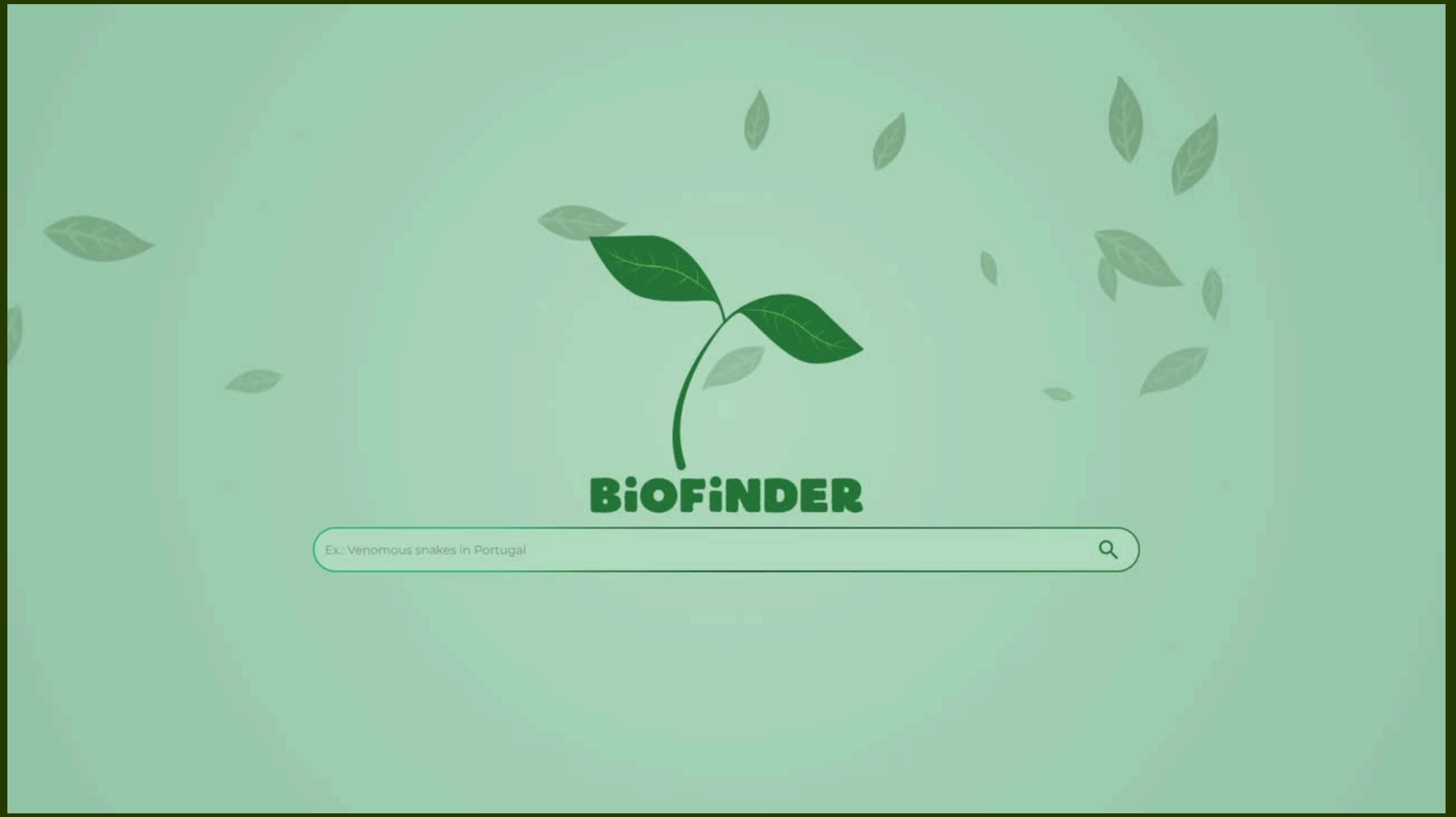


Figure 14: Precision-Recall query and schema improvements.

DEMO

Queries	**Semantic + Extra Parameters**	**Frontend System**
Q1	0.90	0.66
Q2	0.63	0.48
Q3	0.81	0.71
Q4	0.99	0.71
Q5	0.95	0.60
Q6	0.75	0.81

DEMO



Q&A

Any questions?