Faculty of Engineering, University of Porto
2024

BIOFINDER

Information Processing and Retrieval

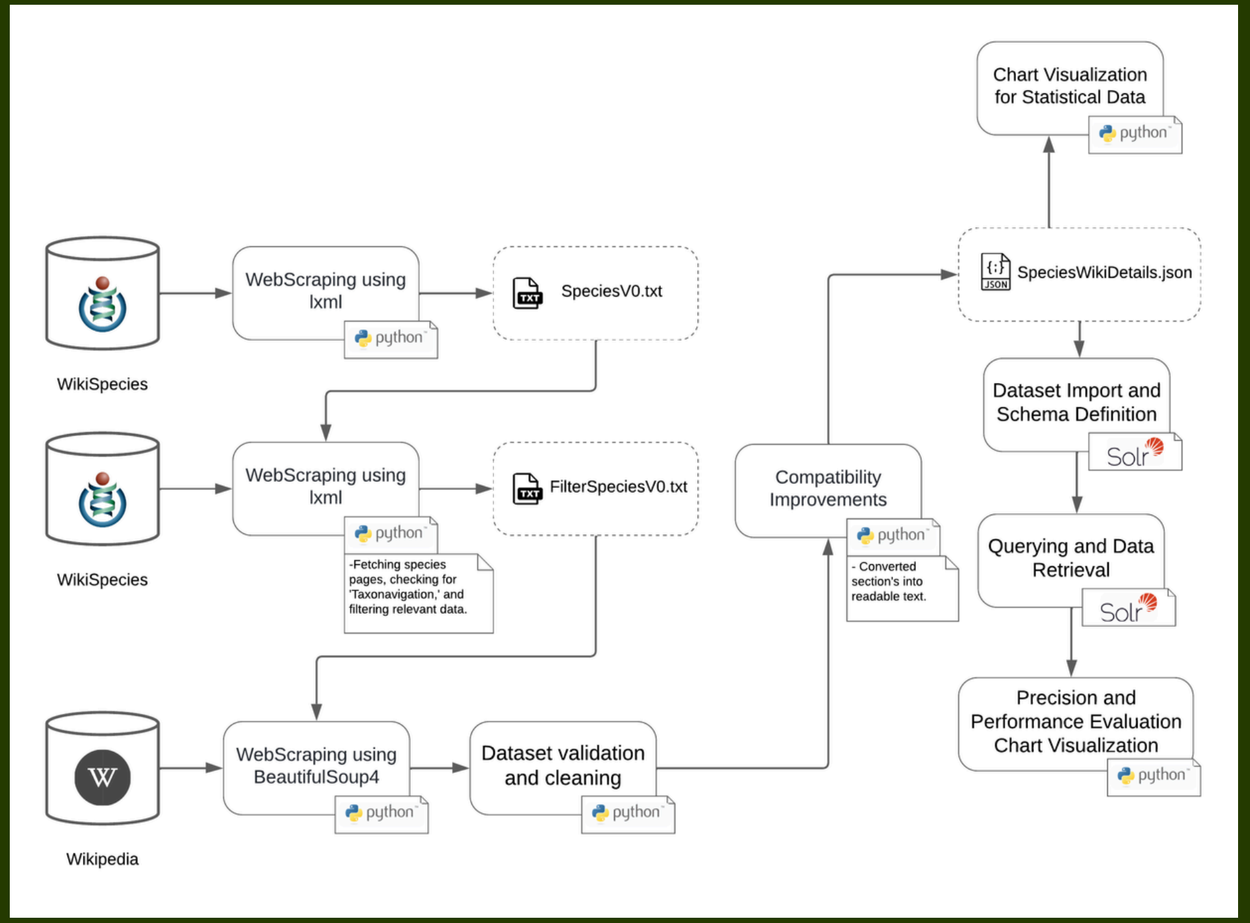
Nuno França; João Tomás Teixeira; Rodrigo Esteves; Isabel Silva

MILESTONE #2

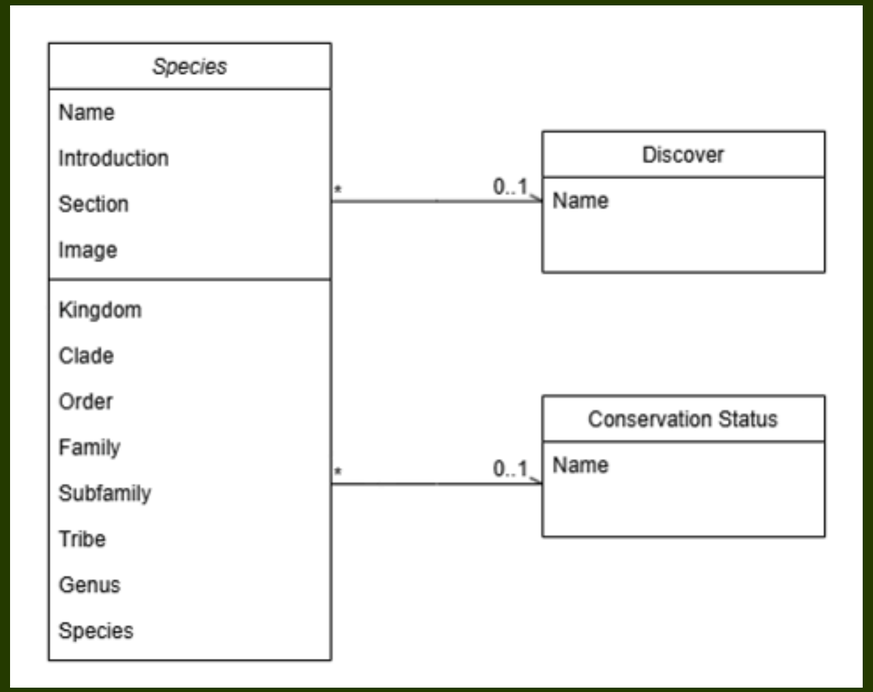
- Implement an information retrieval system for species data.
- Evaluate query performance using precision and recall metrics.
- Compare a simple system with an optimized version.
- Address specific queries related to species information.

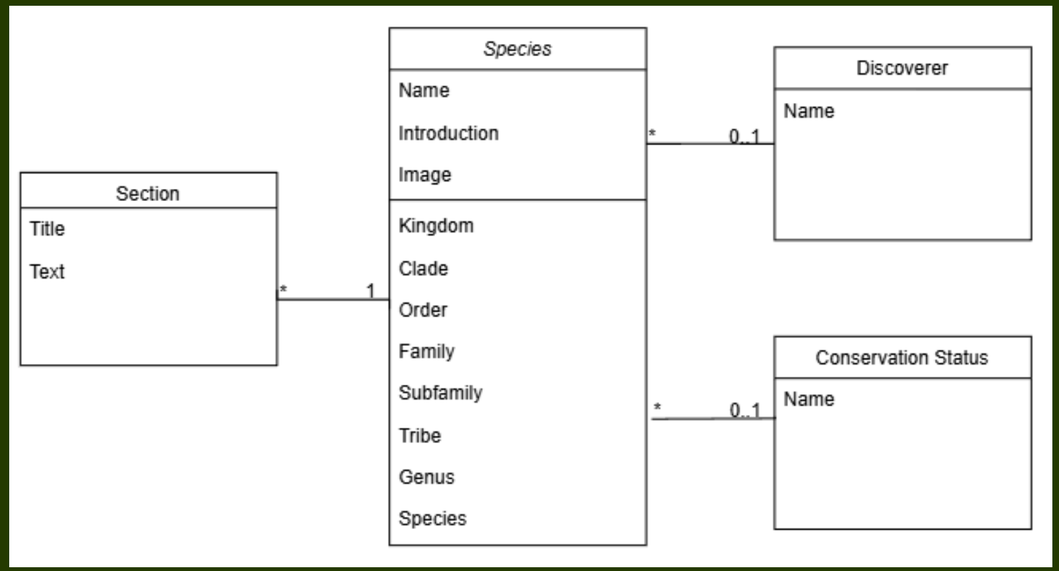
DATA PIPELINE

- Firstly, the JSON file with all the gathered information needed to be re-formatted for it to be correctly read by Solr.
- Secondly, the multiple subsections inside the section were merged into a big paragraph.
- After merging the subsections into one big section, there was no need to create subsets of data.
- To help automatize and simplify the Solr interface process, a Makefile was created.



CONCEPTUAL MODEL





SIMPLE SCHEMA VS ADVANCED 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.

When performing the queries, three distinct files were created that contained the most important synonyms for the information needs that we considered for this study.

For the synonyms, we also took into account the kingdom to which the species belonged.
Example: kingdom Animalia and in the search one of the keywords is animal then this species should appear as a match. However, since kingdom names are in Latin, stemming would be ineffective, so there was a need to force these synonyms in the created files mentioned at the beginning of the paragraph.

Advanced Schema
StandardTokenizerFactory
LowerCaseFilterFactory
PorterStemFilterFactory
SynonymGraphFilterFactory

QUERIES

Different queries were developed throughout this process when thinking of possible search inputs that could be inserted into our future application.

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)

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.

Query:
- q: (largest huge size) AND animal
- q.op: OR
- defType: edismax
- qf: introduction sections
- rows: 50

Complex query:
- q: (largest^2 huge^2 size)
- q.op: OR
- defType: edismax
- qf: introduction^4 sections kingdom^3
- pf: introduction^5
- ps: 4
- rows: 50

The query was evaluated manually by checking images to confirm the animals' size.

The Advanced System performed better, with higher MAP (0.68 vs. 0.54) and better precision at P@10 and P@50. This is primarily due to the advanced schema with stemming, and the inclusion of synonyms for words like *"large"* and *"huge"*.

Metric	Simple System	Complex System
MAP	0.54	0.68
P@10	0.6	0.8
P@50	0.36	0.58

Q1 - LARGEST ANIMALS ON EARTH

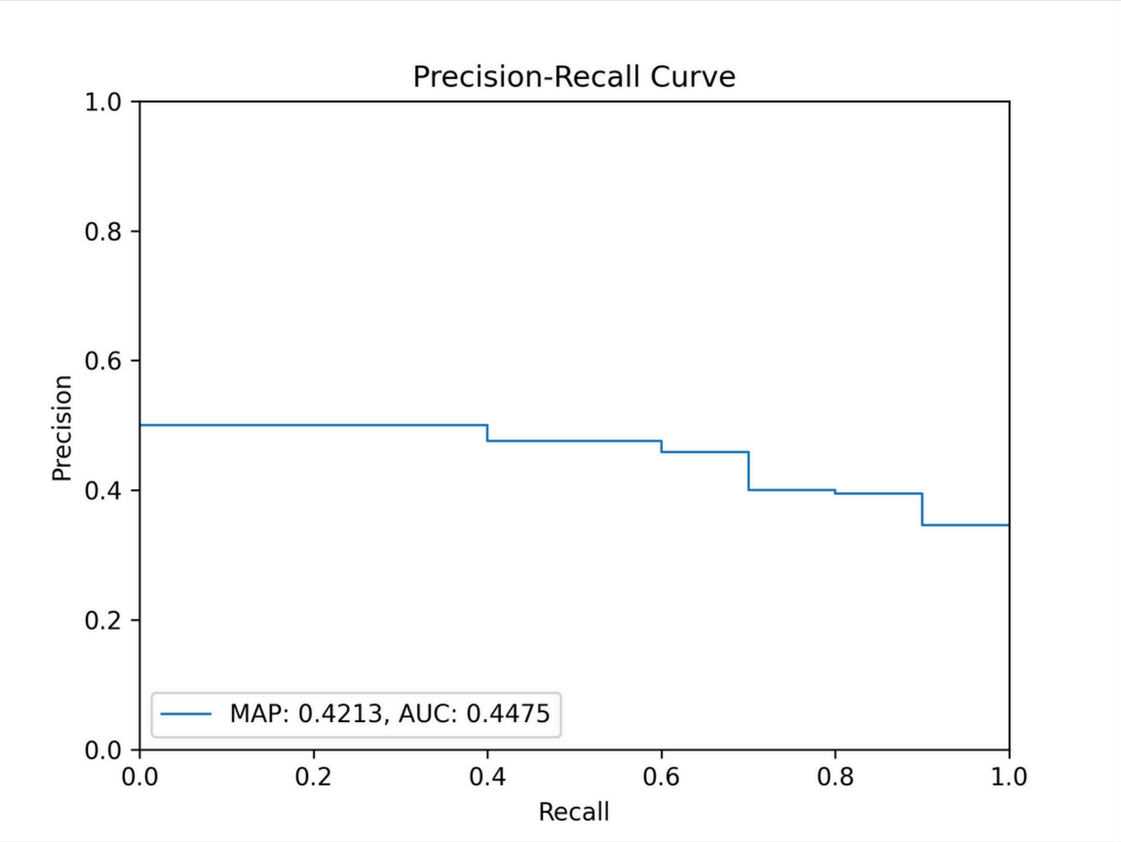


Figure 1: Precision-Recall Initial raw schema.

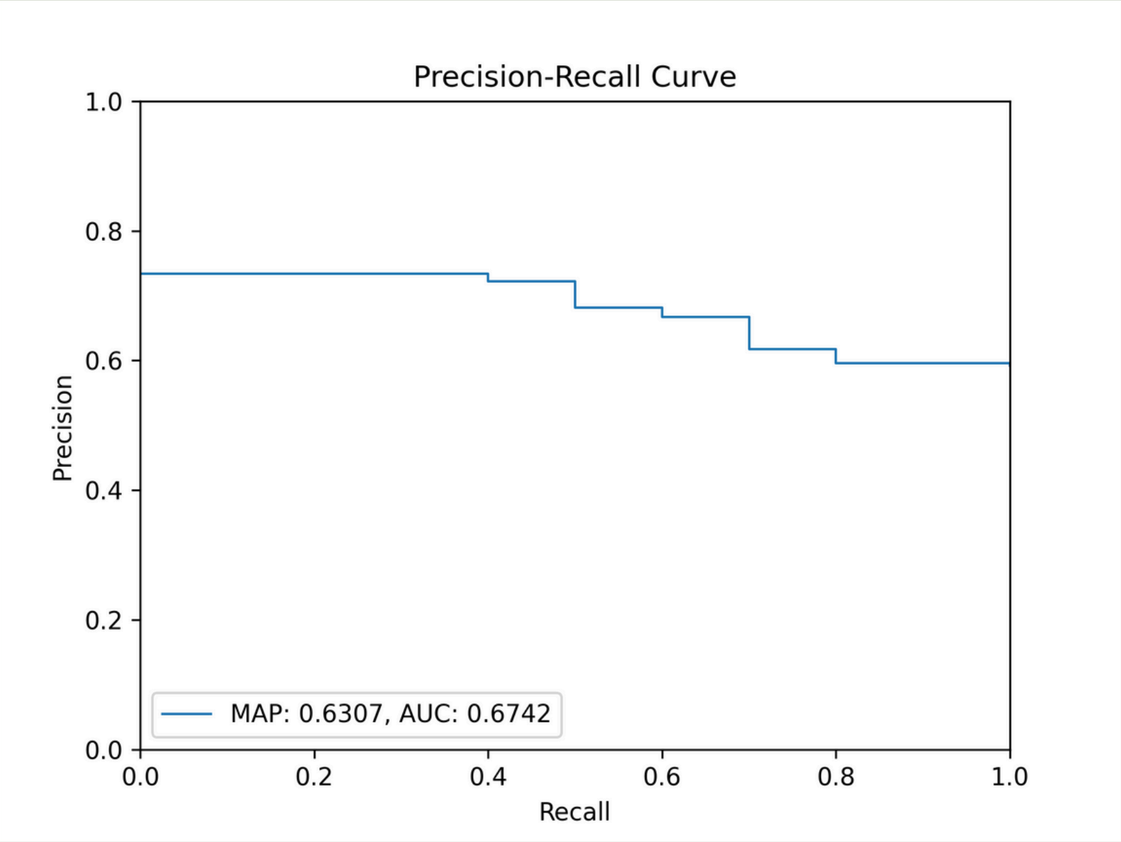


Figure 2: Precision-Recall Enhanced schema.

Q2 - LIFESPAN OF ATLANTIC OCEAN SPECIES

Information need:
What is the lifespan in years of the aquatic species that frequent the Atlantic Ocean?

Relevance:
The goal of this task is to search for the lifespan of all species that frequent the Atlantic Ocean.

Query:
- q: (lifespan years) AND (atlantic AND Aquatic)
- q.op: OR
- defType: edismax
- qf: introduction sections
- rows: 50

Complex query:
- q: (lifespan years) AND atlantic AND aquatic
- q.op: OR
- defType: edismax
- qf: introduction^4 sections^3
- pf: introduction^5 sections^3
- ps: 3
- rows: 50

The evaluation of this query was more methodical, since multiple parameters needed to be checked. This process involved reading the document in search for terms like “Atlantic Ocean” and those indicating lifespan.

TThe Simple System excels in MAP and P@10, achieving higher precision in top-ranked results, while the Complex System outperforms at P@50 showing improved performance for larger result sets. The flatter curve for System 2, is due to its use of synonyms that created multiple false positives spreading across the results.

Metric	Simple System	Complex System
MAP	0.737	0.61
P@10	0.9	0.6
P@50	0.6	0.64

Q2 - LIFESPAN OF ATLANTIC OCEAN SPECIES

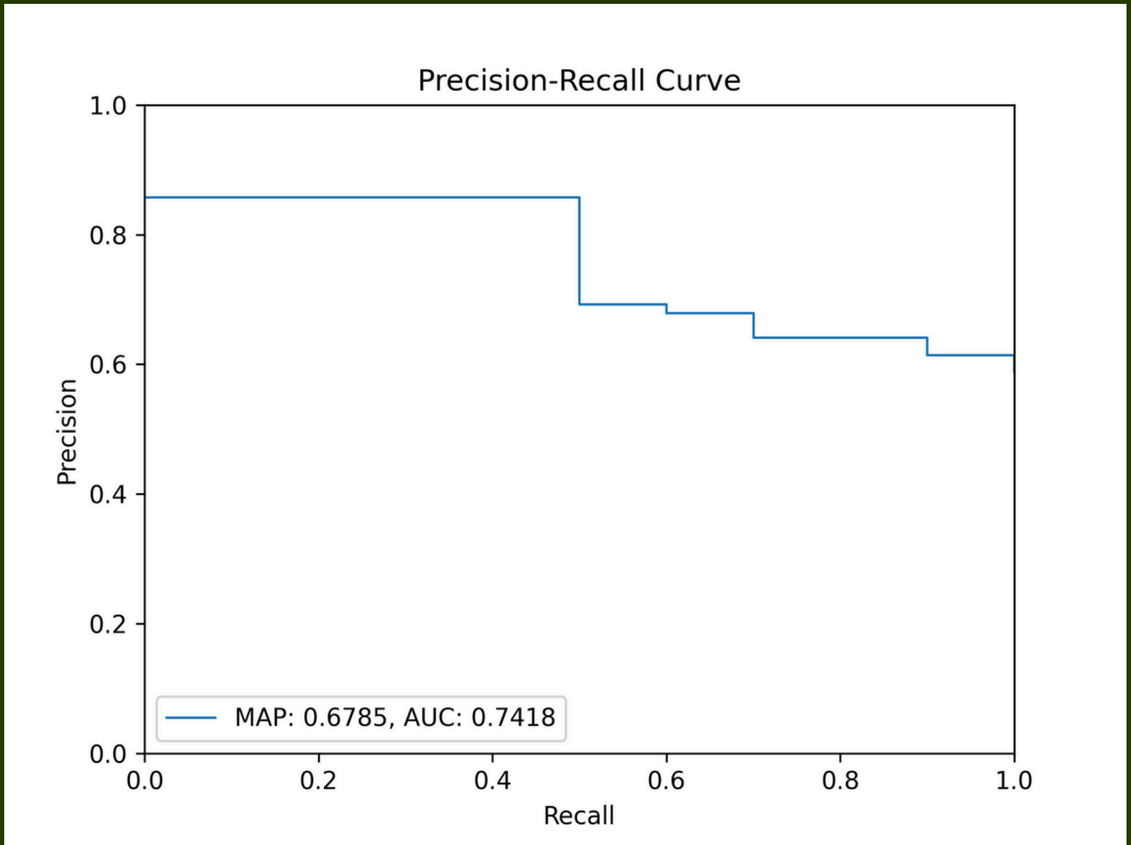


Figure 3: Precision-Recall Simple schema.

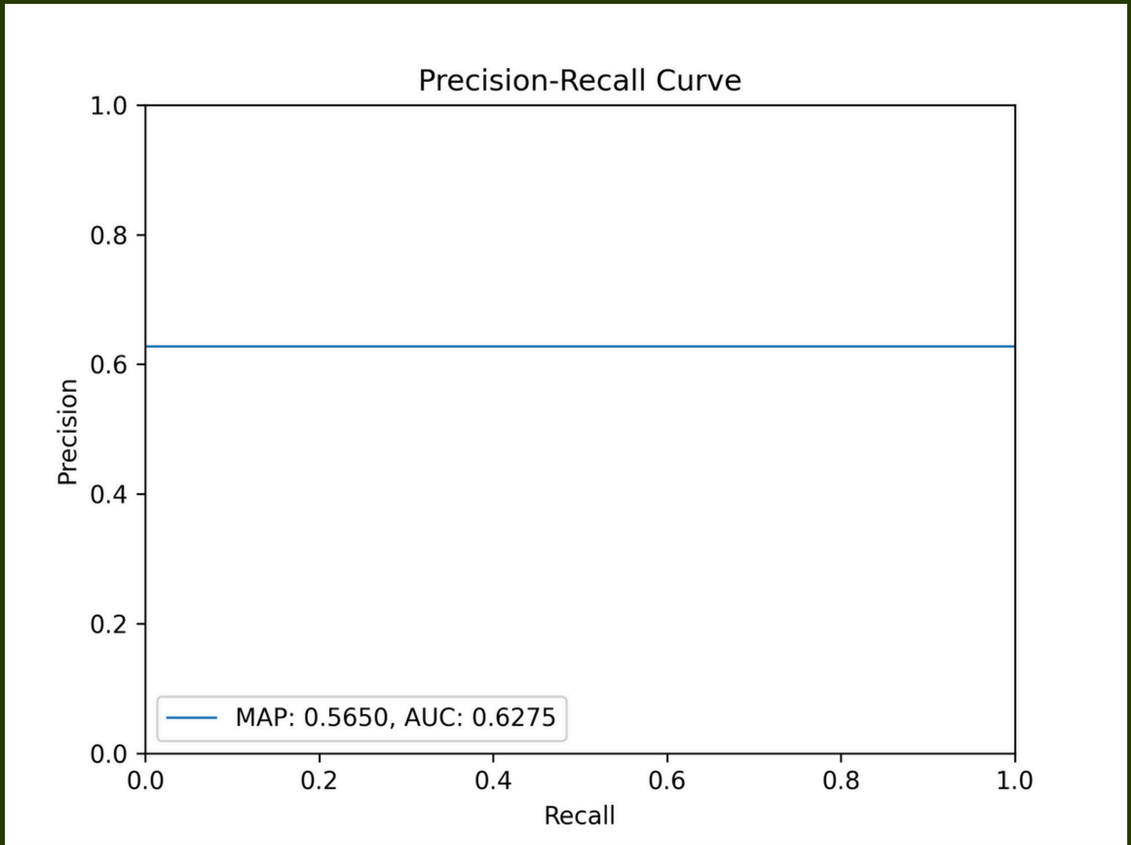


Figure 4: Precision-Recall Advanced schema.

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.

Query:
- q: (endangered species) AND Portugal
- q.op: OR
- defType: edismax
- qf: introduction sections conservation_status
- rows: 50

Complex query:
- q: (endangered^2 species) AND Portugal
- q.op: OR
- defType: edismax
- qf: introduction^3 sections conservation_status^4
- rows: 50

The evalution focused on identifying species that inhabit in Portugal and are classified as endagered. Although many species lacked a direct *“conservation_status”,* revelant information was often found in the introduction and sections.
The *Complex System* outperformed the Simple System with a higher MAP and P@10 , due to its effective use of weighted fields as well as stemming and synonyms for "endangered"

Metric	Simple System	Complex System
MAP	0.6	0.785
P@10	0.6	0.9
P@50	0.2	0.52

Q3 - ENDANGERED SPECIES IN PORTUGAL

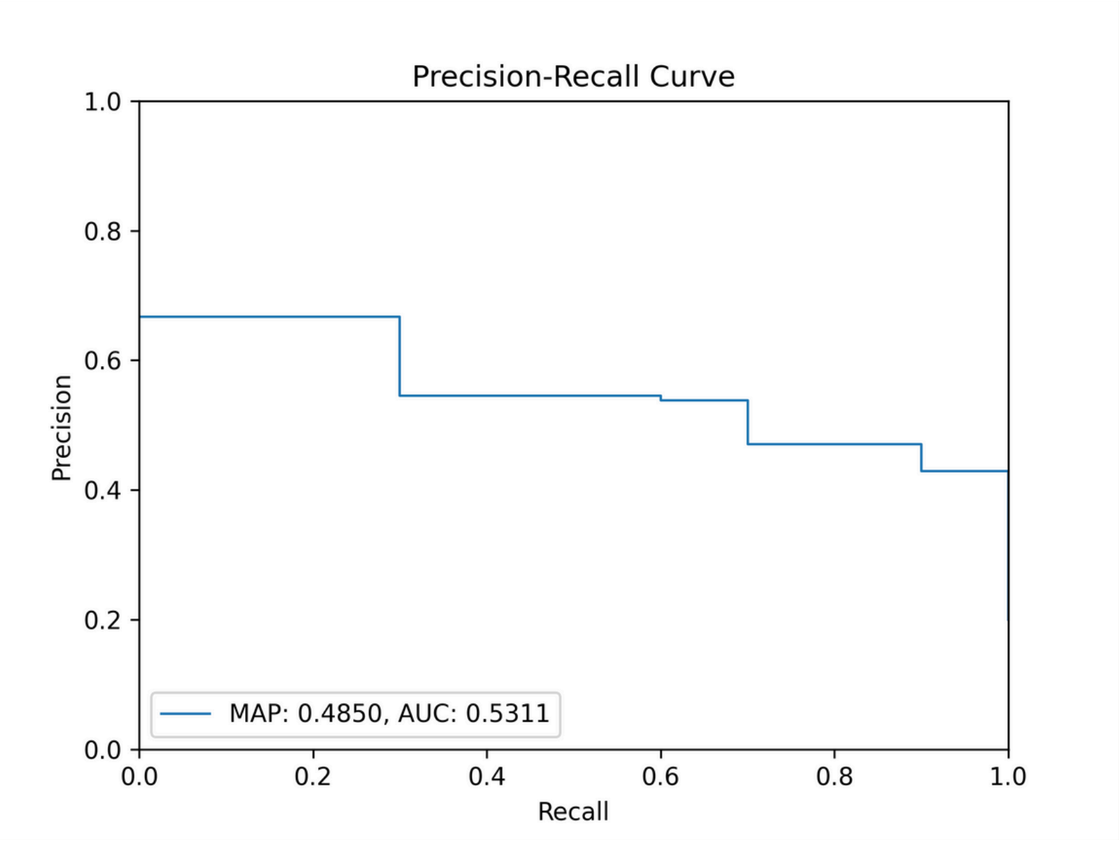


Figure 5: Precision-Recall Simple schema.

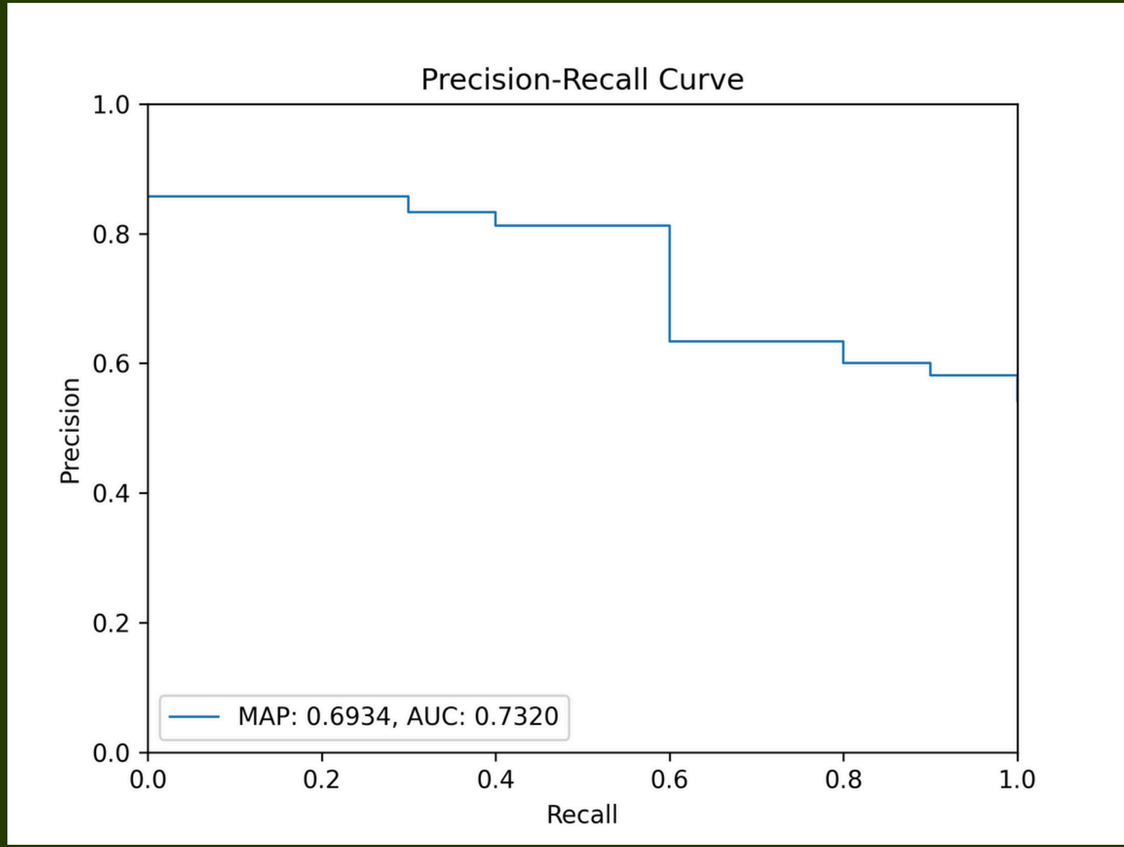


Figure 6: Precision-Recall Advanced schema.

Q4 - DANGEROUS SPECIES IN AUSTRALIA

Information need:
I'm visiting Australia, and I want to know the dangerous species to avoid.

Relevance:
When traveling to another country, namely Australia which is known for the distinct dangerous species that live there, I might want to know in better detail the most dangerous and the ones that should be avoided at all costs.

Query:
- q: (venomous dangerous) AND Australia
- q.op: OR
- defType: edismax
- qf: introduction sections
- rows: 50

Complex query:
- q: (venomous dangerous) AND Australia
- q.op: OR
- defType: edismax
- qf: introduction^3 sections
- rows: 50

The evaluation focused on identifying if a species habitats in Australia and is dangerous to humans

Complex System slightly outperforms the Simple System. It achieves a higher MAP, reflecting greater overall precision across results. This improvement is due to applying the advanced schema, since having a broader range of similar words to "*venomous*" and "*dangerous*", enables the retrieval of more relevant documents.

Metric	Simple System	Complex System
MAP	0.737	0.61
P@10	0.9	0.6
P@50	0.6	0.64

Q4 - DANGEROUS SPECIES IN AUSTRALIA

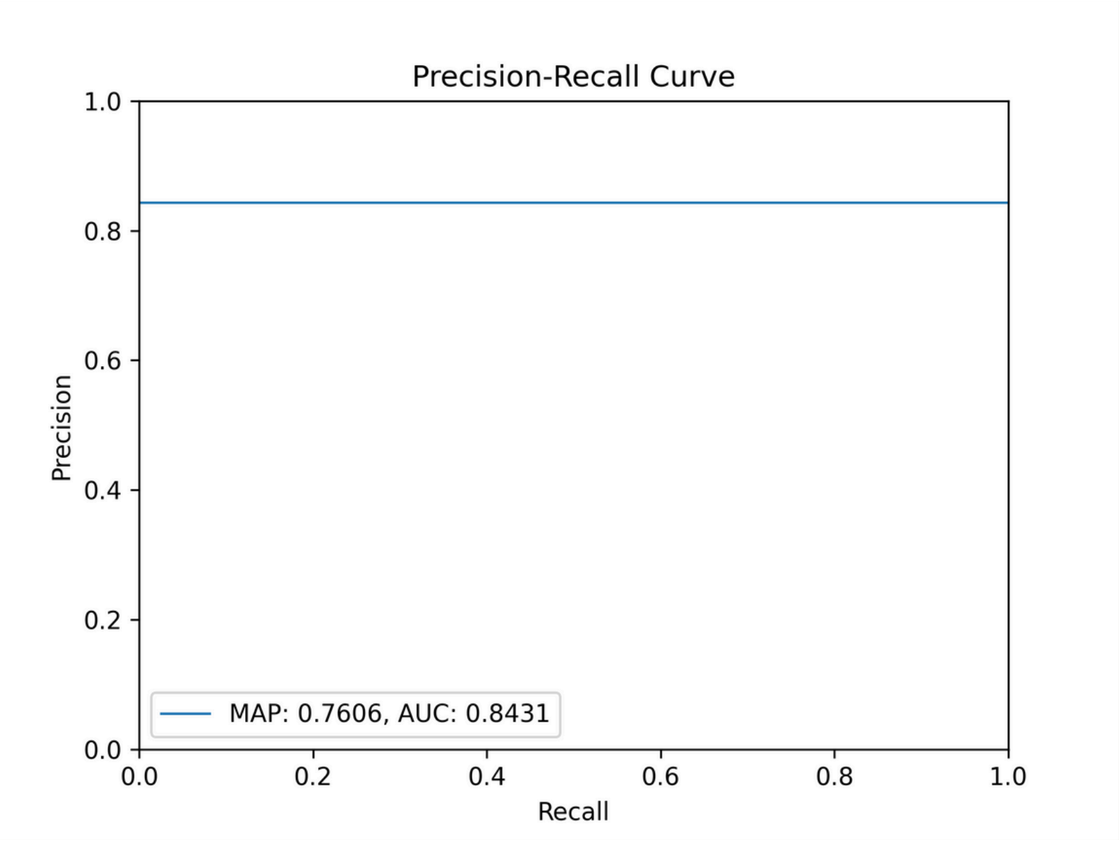


Figure 7: Precision-Recall Simple schema.

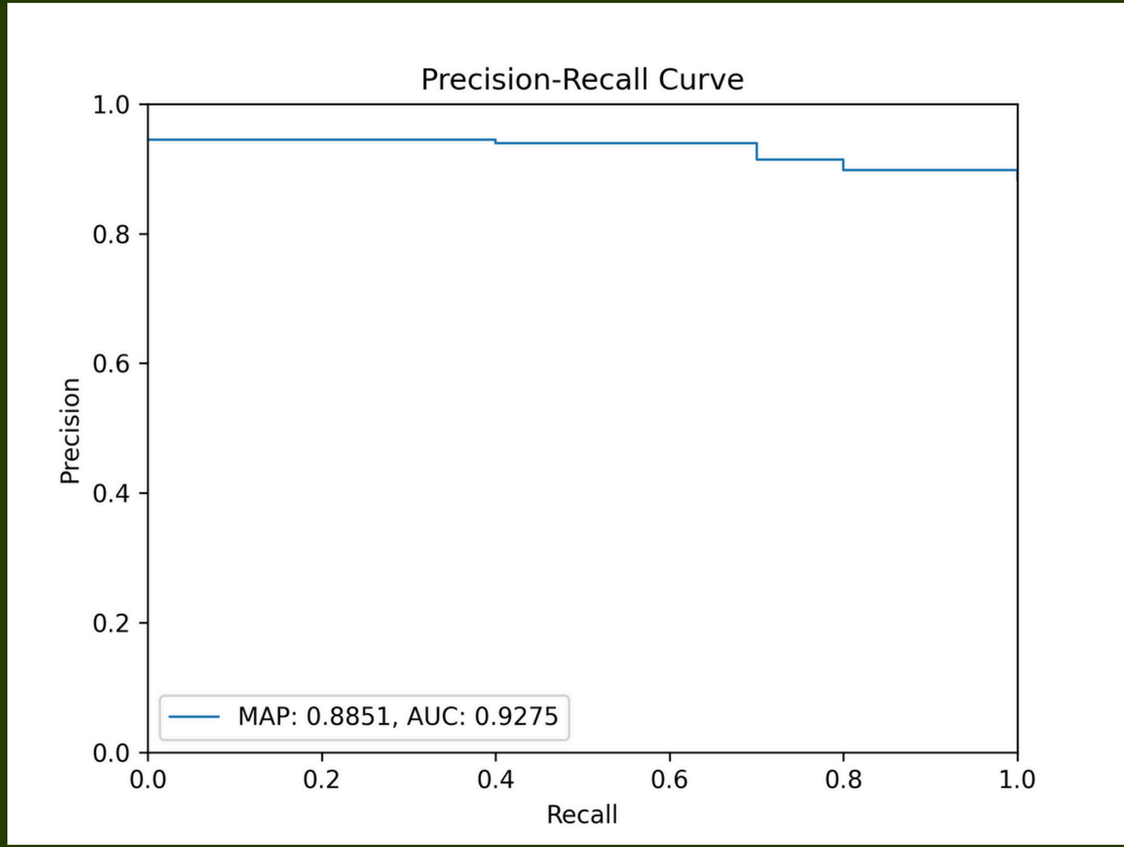


Figure 8: Precision-Recall Advanced schema.

FUTURE WORK

As we transition to the next stage, efforts will be on further improving the project.

The focus of the next phase will consist of:

- Improving the schema developed to this point
- Increasing the list size of the synonyms to improve the search engine
- Try new tokenizers to improve the precision of our results.
- Development of an UI application.

Q&A

Any questions?