Biofinder - A Species Information Retrieval System

Nuno França
up201807530@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

João Tomás Teixeira
up202108738@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Isabel Silva
up201904925@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Rodrigo Esteves
up202403070@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Abstract

This project focuses on biological species, where accurate and accessible information is essential for researchers and educators. The article explains the steps needed to develop a specialized search engine for species data, which is available on Wikipedia. The process involves collecting species-related data from Wikipedia, cleaning and preparing it for analysis, and conducting an in-depth review of the information. Following the data preparation, schemas, and queries were designed to index and retrieve relevant information. The accuracy and completeness of the retrieved results were evaluated by measuring the system's performance using precision and recall criteria. The ultimate goal is to create a robust system that enhances search capabilities, improves information retrieval, and ensures that users can easily access accurate data on a wide range of species, through a simple and responsive User Interface.

Keywords

Species, Information, Datasets, Scraping, Data Retrieval, Data Preparation, Data Analysis, Data Processing, Pipeline, Data Refinement, Data Cleaning

1 Introduction

The focus on species as a research topic stems from its relevance in biological and environmental sciences and the wide range of characteristics it presents. The study of species and their attributes is particularly significant in today's data-driven world, offering insights into biodiversity and ecosystems while showcasing biological data's complexity [6]. The topic covers varied types of information, such as taxonomy [4], distribution, and behavior, making structuring data more challenging and thus a suitable candidate for investigating advanced search and retrieval techniques. This theme aligns with the course's goal of exploring practical, real-world information retrieval systems.

In this project phase, we begin with **Dataset**, which introduces the sources of species data and assesses the data quality. Here we explain the selection, processing, and storage methodologies, ensuring a clear workflow and we evaluate and visualize the processed data, using different criteria to analyze relationships. After data preparation, information about the **Documents** used is presented, including the need to create some distinct files for the development process. Then, we made appropriate queries and implemented specialized indexing schemas to maximize information retrieval from the dataset in the section **Schema and Queries**. Precision and recall measurements were used to thoroughly assess these queries' performance, yielding significant details about how efficiently the retrieval system performed, in the **Evaluation** section. After these steps, and taking into account that some improvements were necessary to the project's results, we have a new section **Information Retrieval Improvements** where we explain the refinements done afterward. In the **Query Parameters Improvements** topic, for each of the six queries developed, the parameters improved are explained in more detail. For the **Schema Improvements** the process to improve the existing schema and the semantics of it are detailed. In the next section, **Semantic and Parameters Improvements**, both query parameters and schema improvements are joined to create an even better final retrieval system. The User Interface part is explained in the **UI Development** section. Finally, the **Conclusions** highlight the outcomes and results obtained throughout the complete project development.

2 Dataset

The dataset produced by the data extraction and enrichment method includes important contextual and visual information in addition to species names, providing a solid basis for further research and information retrieval tasks.

The subsequent subsections will provide further details on the procedures applied during the data collection and filtering processes.

2.1 Description

The Wikipedia Species Directory, an extensive and community-maintained database of biological and botanical species, provided the species information that assembled the dataset used for this research. Entries for several species from diverse biological and botanical groups are included, along with relevant data such as scientific names, introductory information, and other characteristics.

- **Species Name:** The scientific (usually Latin) and common name of each species is included. This improves the dataset's accessibility in a variety of research contexts while helping to distinguish species with similar common names.
- **Taxonomy:** This specific field can highlight, not only the historical context of species but also show the changes that the species have throughout time.
- **Introductory Information:** The introductory section, offers a brief description of the species, including general characteristics, habitat, and significant features.

- **Conservation Status (if available):** Knowing this information is crucial for determining whether a species is in danger of becoming extinct.
- **Discoverer Information (if available):** The dataset includes the name of the individual or team who first discovered the species.

2.2 Origin

The dataset was generated by extensively gathering all the items of data that could be found in the Wikipedia Species Directory (Wikispecies) [14].

2.3 Pipeline

Utilizing the LXML [11] Python module, HTML data from Wikispecies was scraped and converted into a text file with 1.500.000 lines, one line for each page name. As a result, the information was filtered using this same library, and only the data relevant to the scope of this project was stored. Subsequently, BeautifulSoup4 [8] was utilized in Wikipedia, to collect complete information on the species and save it across multiple JSON files.

It is possible to view the pipeline's methodology visually in figure 1.

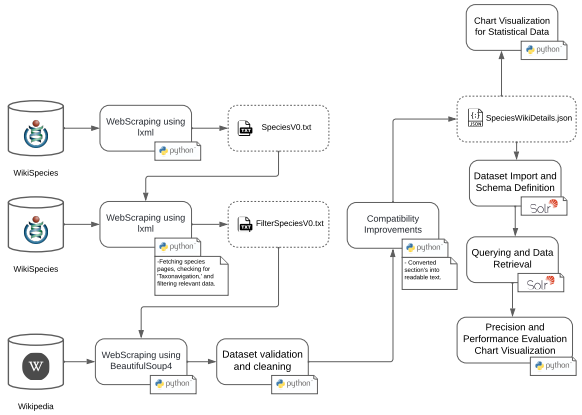

Figure 1: Pipeline.

2.4 Data Collection

The data extraction process began with scraping the Wikipedia species page to compile a comprehensive list of species. Although the page contains a wide variety of information, it also includes entries that are not specifically related to species. To ensure the dataset's relevance, filtering techniques were applied to retain only the species-related entries, resulting in a cleaner dataset.

After this initial filtering, the selected list of species was used to scrape Wikipedia for additional detailed information. For each species, several key pieces of information were gathered. The introductory section of the Wikipedia page, which offers a brief overview of the species, was extracted, along with other relevant sections detailing aspects such as habitat, behavior, and conservation status. Additionally, the primary image associated with each species was collected to provide a visual reference. The scientific classification of each species, covering the taxonomy from domain to species level, was also retrieved. Furthermore, details regarding the discoverer or the team responsible for first describing the species were obtained.

2.5 Data Cleaning

Throughout the data cleaning procedure, all entries were examined to ensure their relevance to species or whether the information gathered related to other data that was irrelevant to the project's scope.

Despite verifying that the species were included in Wikispecies, certain columns required removal due to their non-conformity to actual species.

To prepare the dataset for analysis, several duplicate entries were identified and subsequently removed to ensure data integrity and improve the accuracy of the results.

In addition, the species with a brief introduction (less than 150 words) and no sections were eliminated, as were the wiki sites lacking taxonavigation.

When the data was cleaned, it was discovered that, in certain instances, links about distinct species led to the same Wikipedia article. About 30,000 values in the dataset experienced this and were subsequently eliminated. Nonetheless, these statistics will be analyzed further in the document.

Another column was eliminated because not many species had a conservation status indicating whether or not they were in danger of becoming extinct.

2.6 Final Dataset

Following the conclusion of the data collection, and cleaning processes, the final dataset is organized and includes the relevant species information found in the Wikipedia Species Directory (Wikispecies).

Following the data cleansing procedure, the final dataset has about 120,000 records for distinct species.

There are still some limitations on the existing data in the final dataset. Some species don't have complete information, especially when it comes to the Discoverer and Conservation Status.

Because of the limited information in the source material (Wikipedia Species Directory - Wikispecies), some entries have limited descriptive details.

However, the final dataset, which is based on openly accessible data from the Wikipedia Species Directory (Wikispecies), provides a strong, organized basis for examining species diversity, classification, and associated information. Even though this corrected dataset is much smaller than the original extraction, it is nevertheless of excellent quality and relevance for further study or investigation.

2.7 Tables with the Main Fields

This project contains four distinct tables: Section, Species, Discoverer, and Conservation Status.

Information about the section title and the field containing the matching text can be found in the Section table.

The Conceptual Model discusses the various fields that constitute the species. Still, the primary fields are the image, the name of the species, and the introduction — which provides crucial details about the particular species.

The Discoverer table lists the name of the person who found the species, although this may not always exist.

Lastly, the name field is the only one in the Conservation Status table, and it may not be present for some species.

2.8 Conceptual Model

After the data preparation phase, the project's structure was built around a well-defined conceptual model in Figure 2.

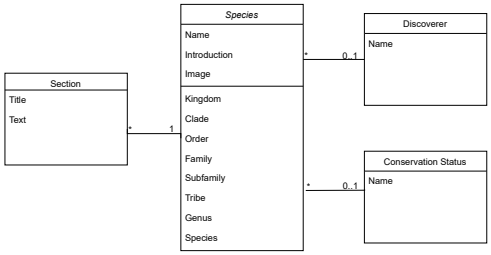

Figure 2: Conceptual Model.

A species consists of the following attributes: a name, an introduction, an image (which may not exist), and a scientific classification. Both the discoverer and the conservation status have corresponding names and can be linked to multiple species. Each Wikipedia section for a species is treated as a unique entity. Sections with the same title but different content are considered distinct and are associated with only one species.

2.9 Data Characterization

To evaluate the collected entries' quality and determine whether they are suitable for further analysis, data characterization attempts to reveal the underlying characteristics, structure, and significant elements of the entries.

2.9.1 Statistics

With the final dataset ready, the group created some graphs to visualize better and analyze the data that will be further explored in the following parts of the project.

To complete the mentioned task, Pandas [10] was utilized to analyze a CSV file, created with meaningful values from the JSON dataset, while Matplotlib [7] helped with the graphics creation.

Figure 3, 4, 5 and 6, are the most relevant statistics, where it can be seen distributions of size and the most relevant values for the future work of full-text search.

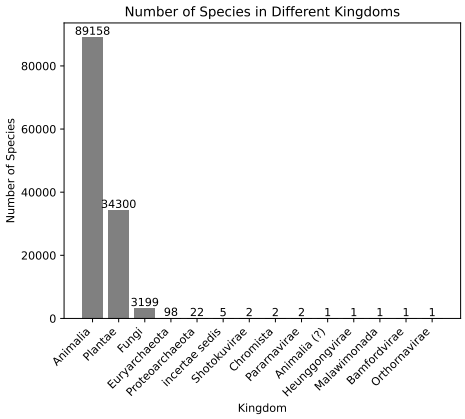

Figure 3: Species divided by kingdoms.

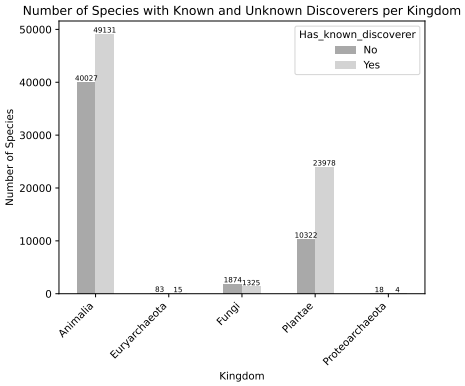

Figure 4: Known discoverer by species kingdom.

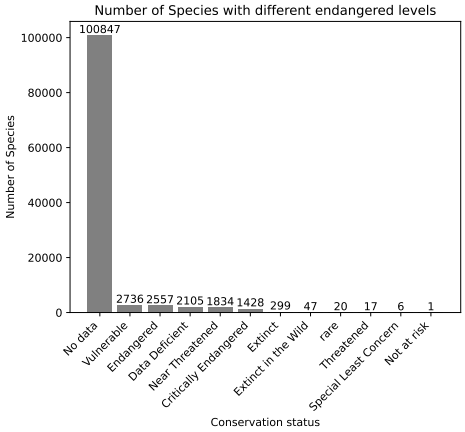

Figure 5: Species conservation status.

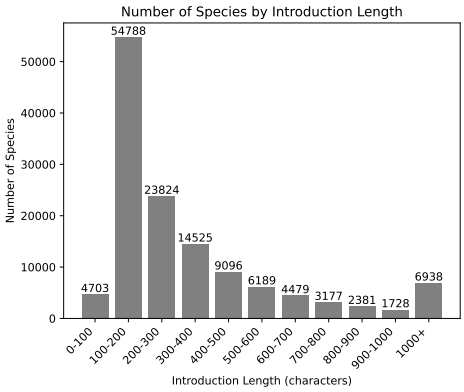

Figure 6: Species introduction length.

2.9.2 Fields/Values Comparison

Figure 3 illustrates the distribution of species across the different kingdoms. The most common kingdoms are Animalia, Plantae, and Fungi. The discrepancy in size is so pronounced that, for this project, the other kingdoms could be ignored.

Figure 4 shows the distribution of the number of species of a given kingdom that have a known discoverer. The number of known discoverers of species of the kingdoms Euryarchaeota and Proteoarchaeota is so low, compared with the number of known discoverers of the kingdoms Animalia, Fungi, and Plantae, that it is only possible to visualize its columns in the graphic because of the values written above said columns. When comparing the distinct columns Animalia, Plantae, and Fungi, the dispersion of the values is high. The Plantae kingdom has approximately twenty times the number of known discoverers of the Fungi kingdom, and therefore the Animalia kingdom has around twice the number of known discoverers of the Plantae. This means that the Animalia kingdom has almost fifty times more known discoverers than the Fungi kingdom, which proves the value difference amongst columns.

Figure 5 shows the big difference between species that have no data regarding the conservation status and the species that have relevant information on this field. It is also important to notice that there is a column labeled Data Deficient, which denotes that there is no known information about the conservation status.

2.9.3 Textual analysis

The dataset has two main fields with rich textual data, the introduction field and the sections field.

The statistics related to the introduction field can be seen in Figure 6. The figure shows the distribution of species introductions relative to the number of characters in their description. As seen in the graphic, the most common length for the introduction is between 100 and 200 characters. Beyond this range, there is an exponential decrease in the number of species as the length of introductions increases.

The statistics related to the size of the section field are presented in the annexes in Figure 8. This figure shows the distribution of the total length of species sections, measured by the number of characters. It's important to note that the content themes within these sections are arbitrary and can vary deeply from one species to another. Just like the graphic mentioned above, the most common length for the total of sections sits between 1 and 500 characters. However, unlike the introduction length graphic, the number of species decreases more gradually, as the section length increases.

2.10 Research Scenarios

During our data analysis, several key themes emerged, providing valuable insights into the study of species. By examining the most commonly referenced elements within the data, we gained a deeper understanding of the critical factors impacting biodiversity research and conservation efforts.

One of the most important aspects identified was Conservation Status. Terms such as "endangered," "vulnerable," and "extinct" frequently appeared, showcasing the importance of tracking species' population trends and the urgency of conservation efforts. Research inquiries like "species at risk of extinction" or "species with declining populations" can offer crucial information for prioritizing protection strategies.

The theme of Discovery was also identified, focusing on the individuals who have contributed to identifying new species. This highlights the historical and scientific value of species discovery and classification, encouraging research into who first described certain species. For example, exploring "species discovered by [scientist]" allows for a deeper look into the historical context of taxonomic research.

The dataset also frequently references scientific classification, such as genus, family, and order, indicating the importance of understanding a species' biological categorization. Questions like "related species within the same genus" or "species classification hierarchy" can help comprehend the evolutionary relationships between different organisms.

Some examples of research scenarios are "What is the lifespan or years of the aquatic species that frequent the Atlantic Ocean?", "I want to know the endangered species that currently habit in Portugal.", or "I'm visiting Australia, I want to know the dangerous species to avoid."

Finally, detailed Species Descriptions (Sections), including characteristics like size, color, and behavior, were often noted, emphasizing the role of specific traits in identifying and studying species. With these extra sections, the dataset can also answer more broad questions, such as, "venomous animal located in Southern Europe", or "species of fish with more than 2 meters, that frequents the Atlantic Ocean".

3 Documents

During this phase of the development process, the initial dataset was changed in two distinct ways.

Firstly, the JSON file with all the gathered information needed to be re-formatted for it to be correctly read by Solr [13].

Secondly, the multiple subsections inside the section were merged into a big paragraph. This was done since not every species had every subsection, and those sub-divisions were making it more complicated to search for species in the dataset.

Furthermore, after merging the subsections into one big section, there was no need to create subsets of data. All information gathered and inserted into the Solr core was treated as a whole.

To help automatize and simplify the Solr interface process, a Makefile [2] was created to be used with the different schemas.

When performing the queries that will be further explained in more detail in the next section, three distinct files were created that contained the most important synonyms for the information needs that we considered for this study. If the application was used widely, then it would be necessary to support every possible query, and not only the ones taken into account here. For the synonyms, we also took into account the kingdom to which the species belonged. For example, if a species is from the kingdom *Animalia* and in the search one of the keywords is *animal* then this species should appear as a match. However, since kingdom names like *Animalia*, *Plantae*, or *Fungi* 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.

4 Schema and Queries

The schema and queries developed were based on the Research Scenarios presented previously. Taking into account the enormous number of queries possible to implement, the group decided to choose some that seemed like they could be searched given our data.

4.1 Schema

The raw schema only checks the lowercase and all attributes were indexed with a simple *tokenizer*.

The schema developed for this project enables Solr to index species data, improving search and retrieval efficiently. There are three types of custom fields in it: **advanced_search**, **convert_kingdom** and **convert_conservation_status** as shown in Table 1.

Field	Type	Indexed
introduction	advanced_search	Yes
sections	advanced_search	Yes
kingdom	convert_kingdom	Yes
conservation status	convert_conservation_status	Yes

Table 1: Schema field types.

Using various filters, the **advanced_search** field type improves text analysis. To achieve this, each entered word is converted to lowercase, a synonym graph filter is used to look for synonyms as well as exact matches, and a stemming filter *PorterStemFilterFactory* is applied to reduce words to their base form, increasing search flexibility.

A similar setup, specifically designed for the "kingdom" field, was applied on the **convert_kingdom** field type. This may include special synonym expansions linked to taxonomic categories. For example when a user searches for "animal", the kingdom "Animalia" should also be a match.

For the **convert_conservation_status** field a related approach was used as well. The data was transformed into lowercase and the synonym graph filter is used to look for synonyms and exact matches.

The schema developed defines four main fields: **introduction**, **sections**, **kingdom** and **conservation_status**.

Species introductions are stored in the **introduction**, which is indexed with synonym handling. For most query scenarios, the searched words have a higher weight when searched in the introduction compared to the sections attribute.

Text data from different parts of species entries are included in the **sections**, which is also indexed with synonyms.

The "kingdom" column is handled by the **kingdom**, which uses synonym-based conversion to uniformly format the entries.

Lastly, the "**conservation_status**" column stores the information about the species regarding their conservation status and was changed to the **convert_conservation_status** field to make the information uniform and accurate to handle queries.

The schema developed can be observed in further detail 2.

4.2 Queries

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

The information needs will be presented together with the queries developed to search in the Solr application in Table 2.

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)

Table 2: Information Needs and Corresponding Queries

5 Evaluation

To evaluate these changes, precision and recall values were extracted using the scripts provided.

For each of the information needs presented, it is possible to see the initial query, and then an enhancement done to that initial query, for better results.

With the values of precision and recall obtained, some graphics were done to better visualize the correlation of both variables, by comparing the first schema (the one with no enhancements) to the one where the previously mentioned enhancements were applied.

5.1 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

Boosts:

For this query, the boosts used were query fields, phrase fields, and phrase slops.

Evaluation:

The evaluation for this query was done manually, primarily by examining images from the results and judging accordingly, since images are a straightforward way to check if an animal is in the larger size of the spectrum.

Results:

Table 3: Q1 results

Rank	Syst. Simple	Syst. Complex
MAP	0.54	0.68
P@10	0.6	0.8
P@50	0.36	0.58

The systems differ in performance, with the Complex System consistently outperforming the Simple System in MAP, P@10, and P@50 in retrieving the desired information, this is primarily due to the advanced schema with stemming, and synonyms for words like **large** and **huge**. Advanced weighting, proximity boosts, and field prioritization were also useful. All these techniques were able to help the complex system create a higher Precision-Recall AUC and flatter curve when compared to the simple system. The precision-recall curve for Simple System can be observed in Image 9 and for the Complex System, in Image 10.

5.2 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

Boosts:

For this query, the boosts used were query fields, phrase fields, and phrase slops.

Evaluation:

In contrast to the previous query, the evaluation was more methodical, since multiple parameters needed to be checked. This process involved reading the documents from the retrieved species. Any document that contained information about a species with an aquatic lifestyle in the Atlantic Ocean and showed their lifespan in years was marked as a successful retrieval.

Results:

Table 4: Q2 results

Rank	Syst. Simple	Syst. Complex
MAP	0.737	0.61
P@10	0.9	0.6
P@50	0.6	0.64

The systems display contrasting performance, with the Simple System excelling in MAP and P@10, reflecting superior precision in the top-ranked results. However, the Complex System surpasses P@50, showing improved performance for larger result sets. The flatter curve for System 2, which can be seen in image 12, is due to its use of weighted fields and proximity boosting, using synonyms for the complex system also created multiple false positives spreading across the results, which distributes more evenly the success in the results. This approach was still able to prioritize broader contextual matches, maintaining precision as recall increases, however at the cost of top-ranked precision. As for system 1, the precision-recall curve had a drastic impact on precision, once the recall of 0.5 was reached, such can be seen in image 11.

5.3 Species endangered in Portugal

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

Relevance: It is important to know the species 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

Boosts:

For this query, the only boosts used were query field boosts.

Evaluation:

As for this query, the evaluation of retrieved documents was done by checking if the retrieved species currently have habitats in Portugal and checking their current **conservation_status**. It is important to note, that most species don't have a conservation_status attribute, just like it can be seen in image 5, however information about their current status could still appear in the sections or introduction attributes.

Results:

Table 5: Q3 results

Rank	Syst. Simple	Syst. Complex
MAP	0.6	0.785
P@10	0.6	0.9
P@50	0.2	0.52

The Complex System outperforms the Simple System, achieving higher MAP, P@10, and P@50, indicating better precision across both top-ranked and larger result sets. The improved performance of the Complex System is attributed to the weighted emphasis on critical fields like conservation status and introduction, as well as stemming and synonyms for "endangered" since this status can be defined by multiple different words. These enhancements allow the system to prioritize documents that better address the query intent. The precision-recall curve for the Simple and Complex System can be seen, respectively, in Image 13 and 14.

5.4 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

Boosts:

For this query, the only boosts used were query field boosts.

Evaluation:

For this query, any species retrieved that had Australia as their current habitat and posed a potential danger to humans was classified as a successful retrieval.

Results:

Table 6: Q4 results

Rank	Syst. Simple	Syst. Complex
MAP	0.8	0.958
P@10	0.8	1.0
P@50	0.86	0.9

The results show that both systems perform well, as indicated by their relatively flat Precision-Recall curves, but the Complex System slightly outperforms the Simple System. It achieves a higher MAP, reflecting greater overall precision across results. The Complex System also reaches perfect precision at P@10 and demonstrates slightly better performance in P@50. This improvement is due to the optimized query's weighted emphasis on critical fields like Introduction and applying the advanced schema, since having a broader range of similar words to **venomous** and **dangerous**, enables the retrieval of more relevant documents. Both precision-recall curves for the Simple and Complex System can be seen, in Images 15 and 16 respectively.

6 Information Retrieval Improvements

After the developments explained until this point, some improvements were required to increase the precision and recall of our search system.

First of all, two new queries were created, to analyze other aspects of our dataset. (1) *I want to know the evasive species on the planet that have a negative impact on the ecosystem* and (2) *I want to know species that have a night activity or nocturnal behavior.*

To meet the expected requirements for this phase of the project, three main areas were tackled.

The first subject explored was a recollection of the documents. At the beginning of this project phase, it became evident that issues occurred while extracting data from Wikipedia. While these problems didn't impact the results, it was decided to perform again the web-scrapping process to ensure the dataset was as complete as possible. The process and the results obtained will be explained further in the next section, along with the challenges posed by a large dataset. These challenges led to the decision to focus on a smaller, handpicked subset of documents, which changed how the evaluation of the systems was performed when compared to the previous iteration.

The second subject explored was how to improve the information retrieval system. This was divided in three parts. Enhancements to query parameters, improvements to the schema with a focus on semantic exploration, and the combination of the updated schema with the new parameters.

Finally, to provide better usage for the user, a simple, yet responsive user interface (UI) was also developed for this stage of the project.

6.1 Web-Scrapping

The web-scrapping process was re-done, and we concluded that due to the 429 HTTP Error [1], the data extracted from Wikipedia was not complete, so it was decided to perform this once again.

The results when compared to the previous web-scrapping process were very distinct.

Table 7: Web-Scrapping Results

	First data extraction	New data extraction
Total entries	183.171	328.223
Duplicate entries	36.439	84.248
Final entries	131.828	215.437

As shown in Table 7, approximately 145.000 new entries were added to the complete dataset. This increase in the number of entries also led to an increase in the number of duplicate documents. The count of duplicate entries grew significantly, rising from approximately 35,000 to 84,000.

Despite the increase in duplicate entries, a total of 83.609 new entries were added to our dataset, bringing the overall count to 215,437 documents and finalizing this phase of the project.

6.2 Subset

After the developments mentioned and explained in the previous section, problems related to the quantity and size of data available emerged. This enhanced the need to have a smaller subset, so that we could compare the results of the field boosts improvements, with the results of the semantic enhancements.

The subset created for the application 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, for the creation of the dataset, it was decided to use the previously developed systems to select 30 true positive results and 20 false positives for each of the six queries.

Another important aspect of the subset is that there are no duplicate results. Additionally, the pool of true positive results for each query does not overlap with the true positive results for any other query. These two features simplify the evaluation method since there is only a need to compare the true positive values from a given query with the corresponding results. This method of evaluation was further improved by the development of a script, enabling fully automated evaluation.

Finally, with the newly created subset, in order to compare the results from the systems developed in this iteration with the previous one, a re-evaluation of the subset with the Complex System was also performed.

7 Query Parameters Improvements

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. Each query will be analyzed and explained throughout the next subsections.

7.1 Largest animals on Earth

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

Query:

- q: (largest^2 huge^2 size) AND animal
- q.op: OR
- defType: edismax
- qf: introduction^5 sections^6 kingdom^2
- pf: introduction^4 sections^3
- bq: kingdom:Animalia^2
- ps: 6
- qs: 2
- tie: 0.5
- rows: 300

Query Parameters Explained: The query parameters introduced include the updated "qf" parameter, which applies new boosts to prioritize fields based on their importance. Fields like sections and introductions are weighted more heavily than kingdoms because they are expected to contain more critical information relevant to the query. Similarly, the "pf" parameter was updated to enhance fields, such as introduction and sections, by boosting documents where query terms appear together as a phrase. To further prioritize specific results, the "bq" parameter is used to boost documents that contain "Animalia" in the kingdom field. Flexibility in matching is introduced through a larger "ps" (phrase slop) and the introduction of "qs" (query slop) parameters. These allow slight variations in the word order or proximity of terms, ensuring that relevant results are not excluded due to minor discrepancies in phrasing. The "tie"

parameter plays an important role in balancing how scores from different fields contribute to the final ranking. In this query, the tie value of 0.5 ensures that while the highest-scoring field remains a dominant factor, contributions from other matching fields are also considered.

Results:

Table 8: Q1 results

Rank	Syst. Complex	New System
P@10	0.90	1.00
P@30	0.77	0.77

The results demonstrate that both systems perform well, but the New System achieves superior performance at lower recall values. Attaining perfect precision at P@10.

As we can see from image 17, the New System maintains perfect precision for a longer range of recall and achieves significantly higher overall metrics, with a MAP of 0.9113 and AUC of 0.9167, compared to the Complex System's MAP of 0.6307 and AUC of 0.6742. This demonstrates the New System's ability to retrieve more relevant results earlier and maintain higher precision as recall increases.

7.2 Lifespan of Atlantic Ocean species

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

Query:

- q: (lifespan years) AND (Atlantic AND Aquatic)
- q.op: OR
- defType: edismax
- qf: introduction^{5} sections^{6}
- pf: introduction^{6} sections^{4}
- bq: sections:(lifespan^{5} years^{4} aquatic^{3} Atlantic^{2})
- ps: 3
- qs: 2
- mm: 2<80%,
- tie: 0.4
- rows: 300

Query Parameters Explained: Higher boosts for key fields like introduction and sections prioritize documents with the most informative content. The addition of the "qs" parameter introduces flexibility for minor variations in term order, while the "mm" parameter ensures a minimum match threshold, in this case, 2 query terms, preventing irrelevant results. The "bq" parameter boosts documents containing specific high-priority terms, such as "lifespan" and "aquatic," and adjusting the tie parameter to 0.4 ensures a balanced contribution from all fields.

Results:

From the rank-based results, the Complex System performs better at both P@10 and P@30, when compared to the New System. As

Table 9: Q2 results

Rank	Syst. Complex	New System
P@10	0.80	0.70
P@30	0.57	0.53

we can see from the image 18, the New System achieves a slightly higher MAP, along with a marginally better AUC. The New System maintains perfect precision at the early stages of recall for a longer range.

7.3 Species endangered in Portugal

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

Query:

- q: (endangered^{2} species) AND Portugal
- q.op: OR
- defType: edismax
- qf: introduction^{4} sections^{3} conservation_status^{1}
- pf: introduction^{4} sections^{3} conservation_status^{8}
- bq: conservation_status:endangered^{1}0 sections:(Portugal^{5} introduction^{1})
- qs: 3
- tie: 0.3
- mm: 2<80%
- rows: 300

Query Parameters Explained: The new "qf" parameter boosts now assigns a significantly higher boost to the conservation status field (up to 10), ensuring documents mentioning "endangered" in this field are prioritized. Similarly, the "pf" parameter boosts phrases in key fields, with conservation status receiving the highest weight, reflecting its critical importance for the query focus. The "bq" also prioritizes documents with "endangered" in conservation status and "Portugal" in sections, ensuring the most relevant results rise to the top. The "qs" of 3 allows slight variations in word order, while the tie parameter (0.3) balances contributions from all fields. The "mm" condition of 2 ensures query focus by requiring sufficient term coverage.

Results:

Table 10: Q3 results

Rank	Syst. Complex	New System
P@10	0.90	0.90
P@30	0.63	0.70

The results show that both systems achieve equal precision at P@10. However, the New System outperforms the Complex System at P@30. This improvement suggests that the New System retrieves more relevant documents as the result set expands, likely due to better weighting of fields and boosted terms.

As we can see from the image 19, the precision-recall curve for the New System remains more stable, indicating that it consistently retrieves relevant documents across a broader range of the subset. In contrast, the Complex System's curve shows a sharper drop in precision after an initial high value, reflecting weaker performance as more documents are retrieved.

7.4 Dangerous species in Australia

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

Query:

- q: (venomous dangerous) AND Australia
- q.op: OR
- defType: edismax
- qf: introduction^{4} sections^{3}
- pf: introduction^{5} sections^{4}
- bq: sections:(venomous^{6} dangerous^{5} australia^{4}) introduction:australia^{5}
- tie: 0.4
- mm: 2<75%
- rows: 300

Query Parameters Explained: The "qf" has been refined with higher boosts to prioritize matches in the introduction and sections that contain critical information to the query need. The "bq" applies strong boosts to terms like "venomous" and "dangerous" within both sections and the introduction, ensuring that documents explicitly mentioning these terms are ranked higher. The "pf" further emphasizes the importance of phrases containing these terms, boosting their influence in scoring. Additionally, the tie parameter, set to 0.4, balances contributions from all matching fields, ensuring more comprehensive scoring without overemphasizing a single field. The minimum match (mm) condition of 3 ensures that all three query terms ("venomous," "dangerous," and "Australia") must appear in the document, guaranteeing precision.

Results:

Table 11: Q4 results

Rank	Syst. Complex	New System
P@10	1.00	1.00
P@30	0.97	0.73

The System Complex still outperforms the New System in overall precision and recall performance.

As we can see from image 20, the precision-recall curves further emphasize this difference. The System Complex maintains near-perfect precision across almost the entire recall range. However, the New System shows a gradual decline in precision as recall increases. This indicates that the New System was overly strict and optimized for precision at the cost of recall, leading to struggles in maintaining relevance as the number of retrieved documents grows.

7.5 Evasive species with a negative impact on the ecosystem

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

Query:

- q: (species negative impact) AND invasive^{2}
- q.op: OR
- defType: edismax
- qf: introduction^{5} sections^{6}
- pf: introduction^{3} sections^{2}
- bq: sections:invasive^{2} introduction:invasive^{4}
- ps: 5
- qs: 3
- tie: 0.4
- rows: 300

Query Parameters Explained: The "qf" now assigns higher boosts to introductions and sections, prioritizing content in these fields where critical information is more likely to be found. The "pf" boosts phrases in the introduction and sections, emphasizing results where query terms appear together, preserving contextual meaning. The "bq" applies additional weight to the term "invasive" within sections and introduction, ensuring documents explicitly mentioning "invasive" are prioritized. The "ps" set to 5 and "qs" of 3 allow slight variations in term order, improving recall by matching phrases with minor word rearrangements. Finally, the tie parameter, set to 0.4, balances contributions from all fields, ensuring that while the most relevant field dominates, secondary matches still influence the final ranking.

Results:

Table 12: Q5 results

Rank	Syst. Complex	New System
P@10	0.70	0.90
P@30	0.80	0.80

From the rank-based results, the New System achieves a higher precision at P@10 compared to the System Complex. At P@30, both systems deliver equal performance with a precision of 0.80, indicating that the New System retrieves more relevant results earlier, while both systems are comparable as more results are retrieved.

As we can see from the image 21, the New System maintains higher precision across the recall range, achieving a MAP of 0.9006 and an AUC of 0.9159, which are significantly better than the System Complex's MAP of 0.7265 and AUC of 0.8001. The New System starts with perfect precision and maintains a more gradual decline.

7.6 Species with nocturnal behavior

Information need: I want to know species that have a night activity or nocturnal behavior

Query:

- q: (activity behavior) AND (night nocturnal)^2
- q.op: OR
- defType: edismax
- qf: introduction^5 sections^4
- pf: introduction^4 sections^3
- bq: sections:nocturnal^3 introduction:night^2
- ps: 5
- qs: 3
- tie: 0.5
- rows: 300

Query Parameters Explained: The higher boosts in "qf" for introduction and sections ensure that key terms like "night" and "nocturnal" in critical fields receive greater importance, improving precision. The "pf" prioritizes documents where query terms appear together, making results more contextually relevant. To further refine relevance, the "bq" specifically amplifies "nocturnal" in sections and "night" in the introduction, prioritizing documents where these terms are prominent. The reduced phrase slop (ps = 5) allows slight flexibility in word order while keeping phrases cohesive, and the query slop (qs = 3) introduces additional tolerance for term rearrangement, increasing recall without sacrificing relevance. Lastly, the tie parameter of 0.5 ensures a balanced contribution from all matching fields, preventing any single field from dominating the ranking.

Results:

Table 13: Q6 results

Rank	Syst. Complex	New System
P@10	0.80	0.70
P@30	0.57	0.77

From the precision perspective, the System Complex performs better at P@10, compared to the New System. However, at P@30, the New System significantly outperforms the System Complex. This suggests that the New System retrieves more relevant documents as the result set expands, maintaining consistency at broader ranks.

As we can see from image 22, the precision-recall curves reinforce this observation. The New System achieves a higher MAP and AUC compared to the System Complex. The New System maintains higher precision over a larger portion of the recall range, while the Complex System shows a sharper drop in precision as recall increases.

8 Schema Improvements

The other approach used to obtain better results and to better analyze the dataset was improving the system's schema. This phase started with the complex schema from the previous delivery. The schema was enhanced in three key aspects. The incorporation of semantic embeddings, an improved synonym dictionary, and the addition of a new filter, *StopFilterFactory*.

To incorporate Semantic embeddings in the schema, the Python Library Sentence_Transformers [5] was used to change the dataset with the respective embeddings. A synonym dictionary also was created with help from the NLTK [3] Python Library, where all unique words in the dataset were analyzed to generate an extensive list of synonyms for each word. Additionally, a stopword dictionary was developed with NLTK, enabling the removal of unnecessary words to further refine the document's search.

The results for each query will be analyzed and explained throughout the next subsections.

8.1 Largest animals on Earth

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

Query:

- q: (largest^2 huge^2 size)
- fq: {!knn f=vector topK=60} (Large and Huge sized animal)
- q.op: OR
- defType: edismax
- qf: introduction^4 sections kingdom^3
- pf: introduction^5
- ps: 4
- rows: 300

Results:

Table 14: Q1 results

Rank	Syst. Complex	New System
P@10	0.90	0.90
P@30	0.76	0.80

For the first query, the precision values at P@10 and P@30 do not show great differences between the systems. However, the precision-recall curves (Image 35 for the Complex System and Image 23 for the new system) reveal that the new schema consistently achieves better precision at lower recall values. This improvement can be attributed to the semantic embeddings, which group similar terms such as "large," "huge," and "size" in the vector space.

8.2 Lifespan of Atlantic Ocean species

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

Query:

- q: (lifespan years) AND (Atlantic AND Aquatic)
- fq: {!knn f=vector topK=80} (Lifespan in years of atlantic species)
- q.op: OR
- defType: edismax
- qf: introduction^4 sections^3
- pf: introduction^5 sections^3
- ps: 3

- rows: 300

Results:

Table 15: Q2 results

Rank	Syst. Complex	New System
P@10	0.80	0.70
P@30	0.57	0.53

This system does not show an improvement given the query. The primary reason is that the embedding function can't find good similarities between "lifespan" and the "Aquatic/Atlantic". In our documents, the habitat and the lifespan can be separated by a high number of words, which reduces the effectiveness of semantic search. Such can also be seen in the precision-recall curves for both systems in image 24 and 36.

8.3 Species endangered in Portugal

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

Query:

- q: (endangered^2 species) AND Portugal
- fq: {!knn f=vector topK=300} (Endangered species in Portugal)
- q.op: OR
- defType: edismax
- qf: introduction^3 sections conservation_status^4
- rows: 300

Results:

Table 16: Q3 results

Rank	Syst. Complex	New System
P@10	0.90	1.00
P@30	0.63	0.67

For this query, the new system for lower recall values has better precision. However, as the recall increases the precision becomes similar to that of the complex system. Similar to the previous query, while some documents have a strong correlation between "endangered" and "Portugal" with those words even appearing in the same sentence others documents do not. This inconsistency leads to a precision-recall curve characterized by a great decline at higher recall values. These curves can be seen for both systems in images 25 and 19.

8.4 Dangerous species in Australia

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

Query:

- q: (venomous dangerous) AND Australia
- fq: {!knn f=vector topK=80} (Venomous or Dangerous species in Australia)
- q.op: OR
- defType: edismax
- qf: introduction^3 sections
- rows: 300

Results:

Table 17: Q4 results

Rank	Syst. Complex	New System
P@10	1.00	1.00
P@30	0.97	0.93

For this query, while the theme is interesting, there is almost no difference in the results between the two systems. This is likely because most species in Australia can, in fact, be categorized as dangerous. This makes the query results almost perfect and consequently difficult to improve with the new system. Both precision-recall curves for the systems can be seen in images 26 and 38.

8.5 Evasive species with a negative impact on the ecosystem

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

Query:

- q: (species negative impact) AND invasive
- fq: {!knn f=vector topK=100} (Invasive species with negative impact)
- q.op: OR
- defType: edismax
- qf: introduction^3 sections
- rows: 300

Results:

Table 18: Q5 results

Rank	Syst. Complex	New System
P@10	0.70	1.00
P@30	0.80	0.70

In this query, similar to the previous ones, the new schema that takes advantage of semantic search, was able to achieve great results at lower and medium recall levels. This can be attributed, again, to the fact that the key terms in the query are often grouped together in most documents, showing a strong correlation. Precision-recall curves for both systems can be seen in images 27 and 39.

8.6 Species with nocturnal behavior

Information need: I want to know species that have a night activity or nocturnal behavior

Query:

- q: (activity behavior) AND (night nocturnal)
- fq: {!knn f=vector topK=100} (Species with night behavior or nocturnal behavior)
- q.op: OR
- defType: edismax
- qf: introduction[3] sections
- ps: 6
- rows: 300

Results:

Table 19: Q6 results

Rank	Syst. Complex	New System
P@10	0.80	0.70
P@30	0.57	0.63

The results for both systems are very similar across different recall levels, primarily because both systems retrieved a consistent high number of false positives. The new schema always found terms like "night" and "activity", where those appeared in the documents but they were not consistently relevant to each other. Precision-recall curves for both systems can be seen in images 28 and 40.

9 Semantic and Parameters 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.

Table 20: P@30 Results for Semantic + Extra Parameters

Queries	Syst. Complex	Semantic	Parameters	Sem.+Param.
Q1	0.77	0.80	0.77	0.83
Q2	0.57	0.53	0.53	0.5
Q3	0.63	0.67	0.70	0.7
Q4	0.97	0.93	0.73	0.73
Q5	0.80	0.70	0.80	0.73
Q6	0.57	0.63	0.77	0.7

As shown in the previous tables, we can observe that combining both developed systems led to better performance. This is further illustrated in the images from 29 to 34, which show how the new system with Extra Parameters and the new developed Schema has an higher MAP across most queries.

10 UI Development

In the previous section, it was shown that incorporating semantic embeddings into the schema and tailoring the parameters for each specific query significantly improved the results for the retrieved

Table 21: MAP Results for Semantic + Extra Parameters

Queries	Syst. Complex	Semantic	Parameters	Sem.+Param.
Q1	0.85	0.90	0.91	0.94
Q2	0.64	0.63	0.64	0.69
Q3	0.76	0.81	0.83	0.82
Q4	0.98	0.99	0.91	0.88
Q5	0.73	0.95	0.90	0.95
Q6	0.73	0.75	0.77	0.74

documents. However, since the user application was designed to accommodate broader use cases, as such, all 215.000 documents were indexed using a simpler schema. This schema has as most important filters: *PorterStemFilterFactory*, *SynonymGraphFilterFactory* and *StopFilterFactory*.

To better understand the capabilities of the Frontend System, an evaluation was performed by comparing the top 30 documents retrieved from the subset using the best-developed system with the top 30 documents retrieved from the entire dataset using the Frontend System, for all created queries.

Table 22: MAP Results for Frontend System

Queries	Semantic + 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

So, in order to improve user's interaction with our information retrieval system, a basic user interface (UI) was created. This built front-end application allows users to query our system and visualize the results in a user-friendly manner, as well as view species information and images when they exist.

The user interface was developed using React [12] and taking full advantage of the Material-UI library [9] to ensure a modern and responsive design. For handling the user queries the application sends requests to the Solr API. In this query, a field boost of 2 is applied to the introduction field, query operator is "AND".

Images for the UI can be found in the annexes, as Image 41,42 and 43 .

11 Conclusions

In conclusion, completing all tasks in the data preparation phase marks a significant step forward in this project. This phase completion certifies that a well-organized and structured dataset is in place, which will serve as the basis for more detailed analysis.

A major challenge encountered previously was correctly processing species attributes from their respective Wikipedia page, such as handling irrelevant introductions and missing values, like the absence of conservation status on some pages.

In the last project phase, the challenge relied on the enormous size of the complete file with synonyms to be inserted into Solr, which led to the need to create a subset of data. Besides that, the

analysis of the results with semantic improvements and query parameter boosts was also a bit challenging when trying to make comparisons.

Overall, the main goal of the project was successful, and a precise and responsive system was developed, as well as a user interface to help interact with the system.

Additionally, we understand now the fact that when developing a fully comprehensive retrieval system certain compromises need to be made.

12 Annexes

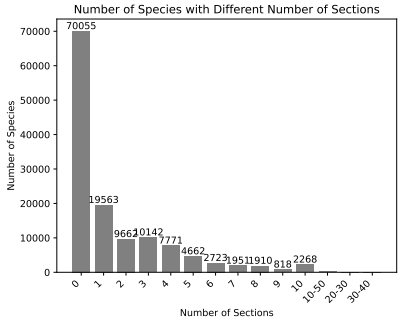

Figure 7: Species divided by number of sections.

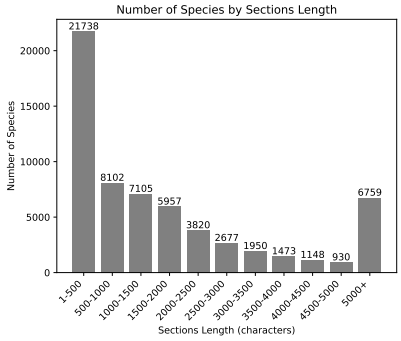

Figure 8: Species divided by total section length.

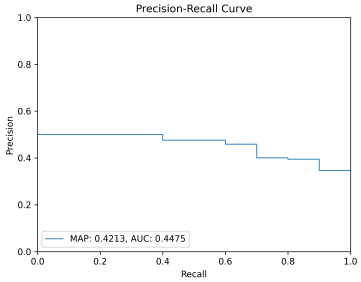

Figure 9: Precision recall for Query 1, System 1.

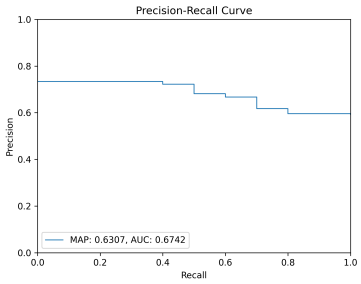

Figure 10: Precision recall for Query 1, System 2.

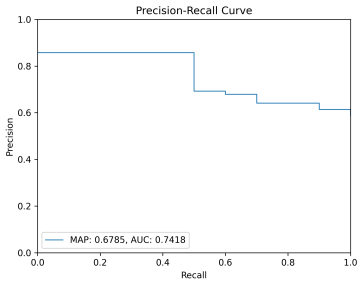

Figure 11: Precision recall for Query 2, System 1.

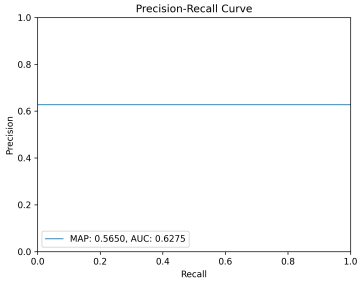

Figure 12: Precision recall for Query 2, System 2.

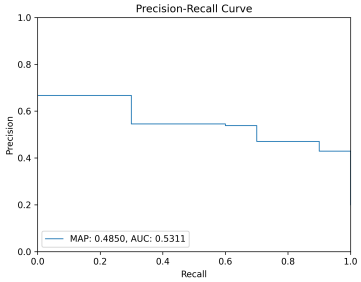

Figure 13: Precision recall for Query 3, System 1.

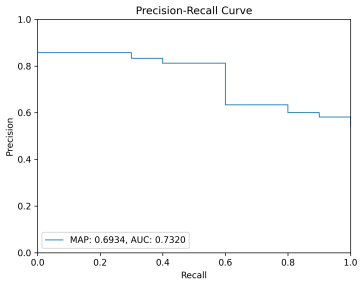

Figure 14: Precision recall for Query 3, System 2.

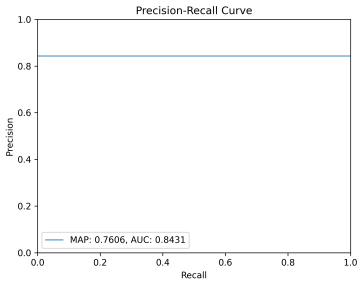

Figure 15: Precision recall for Query 4, System 1.

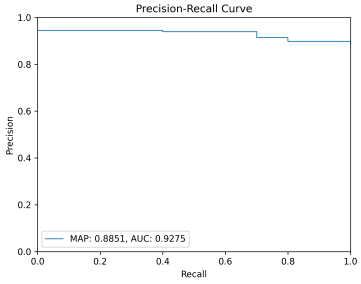

Figure 16: Precision recall for Query 4, System 2.

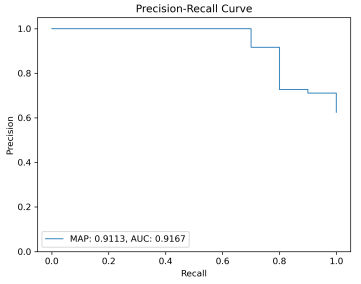

Figure 17: Precision recall for Query 1, Field Improvements.

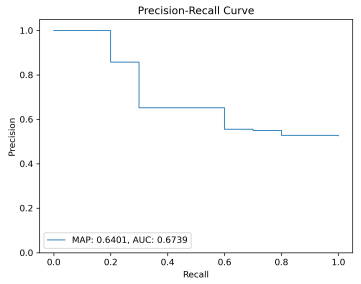

Figure 18: Precision recall for Query 2, Field Improvements.

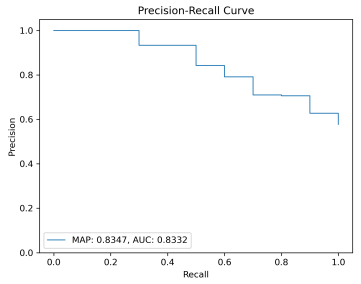

Figure 19: Precision recall for Query 3, Field Improvements.

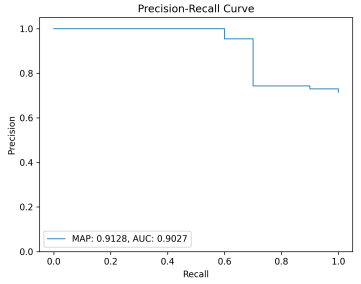

Figure 20: Precision recall for Query 4, Field Improvements.

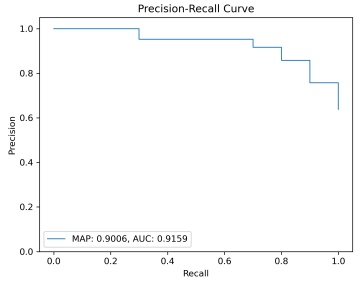

Figure 21: Precision recall for Query 5, Field Improvements.

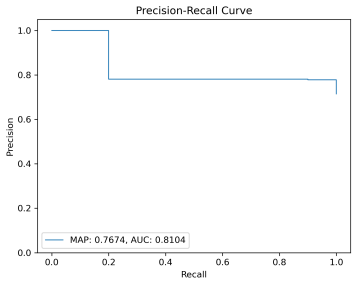

Figure 22: Precision recall for Query 6, Field Improvements.

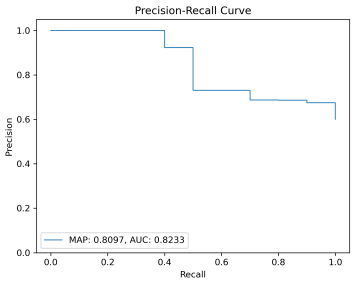

Figure 25: Precision recall for Query 3, Semantic Improvements.

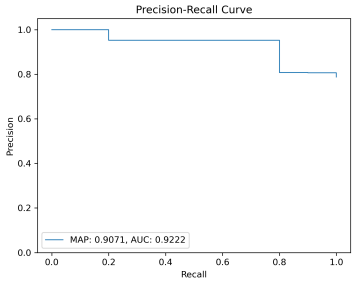

Figure 23: Precision recall for Query 1, Semantic Improvements.

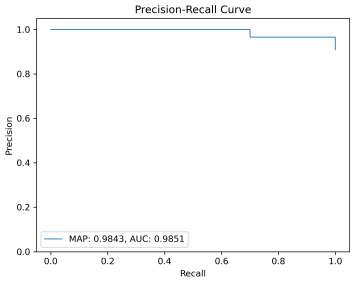

Figure 26: Precision recall for Query 4, Semantic Improvements.

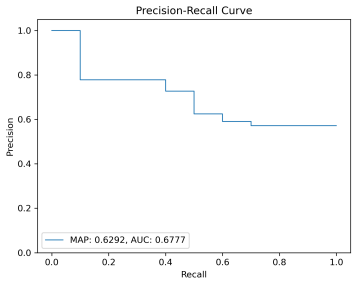

Figure 24: Precision recall for Query 2, Semantic Improvements.

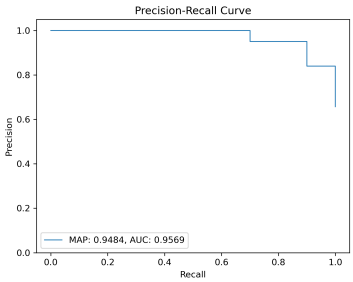

Figure 27: Precision recall for Query 5, Semantic Improvements.

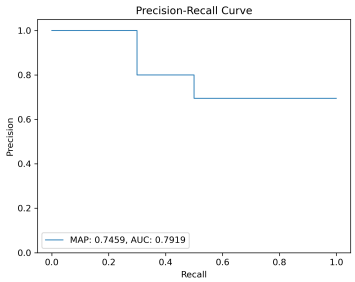

Figure 28: Precision recall for Query 6, Semantic Improvements.

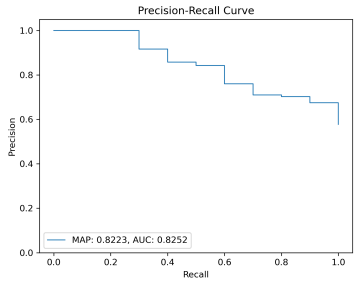

Figure 31: Precision recall for Query 3, Semantic Improvements.

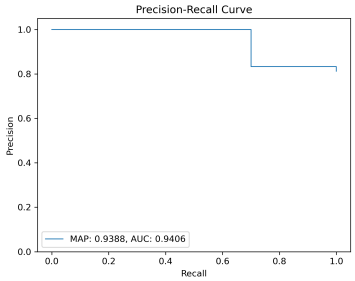

Figure 29: Precision recall for Query 1, Semantic Improvements.

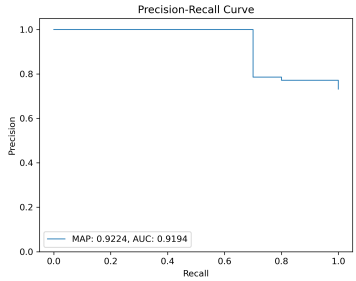

Figure 32: Precision recall for Query 4, Semantic Improvements.

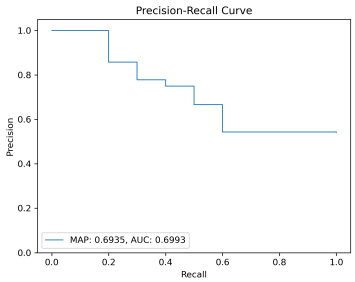

Figure 30: Precision recall for Query 2, Semantic Improvements.

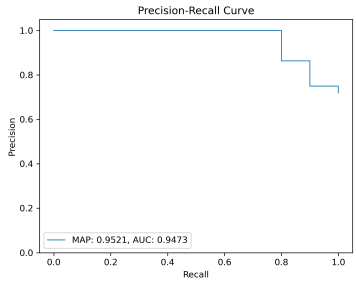

Figure 33: Precision recall for Query 5, Semantic Improvements.

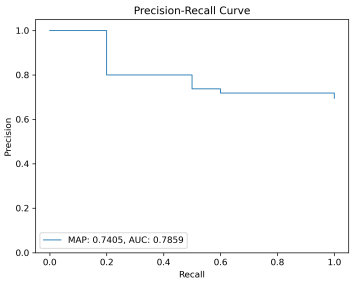

Figure 34: Precision recall for Query 6, Semantic Improvements.

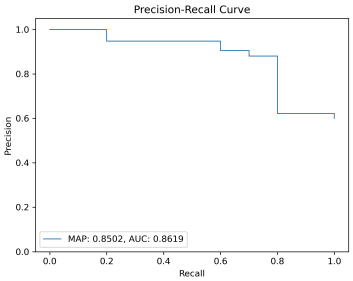

Figure 35: Precision recall for Query 1, original system in the new subset.

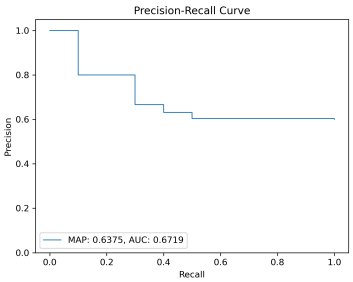

Figure 36: Precision recall for Query 2, original system in the new subset.

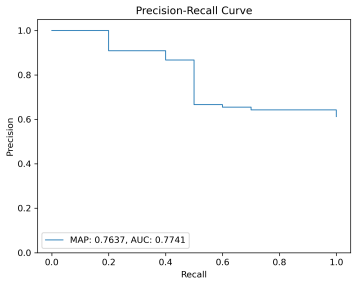

Figure 37: Precision recall for Query 3, original system in the new subset.

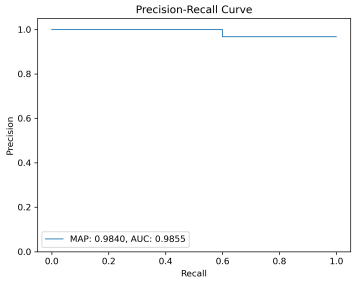

Figure 38: Precision recall for Query 4, original system in the new subset.

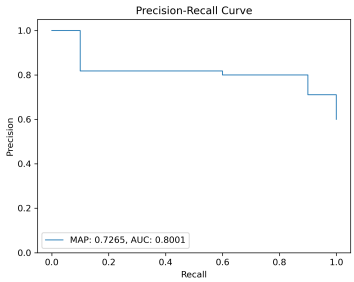

Figure 39: Precision recall for Query 5, original system in the new subset.

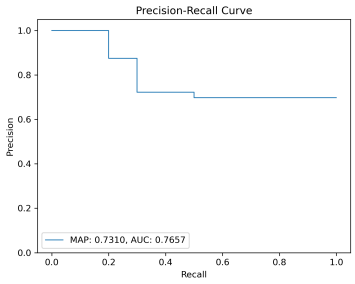

Figure 40: Precision recall for Query 6, original system in the new subset.

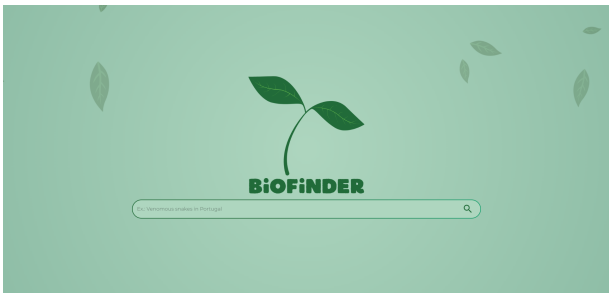

Figure 41: Home Page for UI

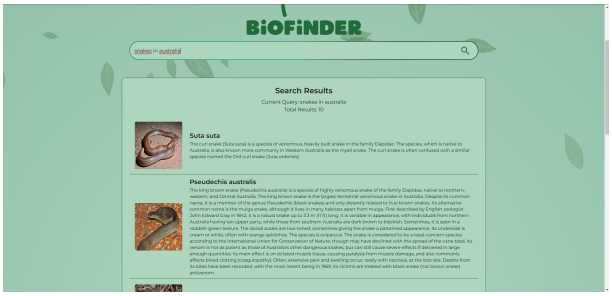

Figure 42: Query Results

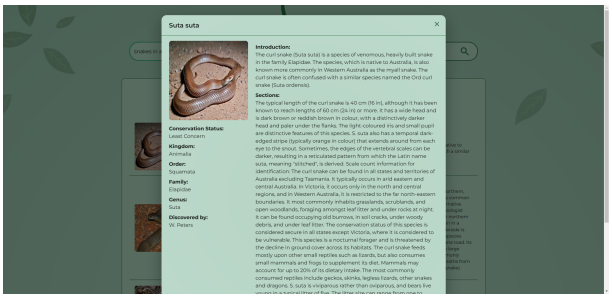

Figure 43: Specific Result

Listing 1: Complex Schema

```
{
    "add-field-type": [
        {
            "name": "advanced_search",
            "class": "solr.TextField",
            "indexAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        StandardTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "
                            intro_section_synonyms.
                            txt",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    },
                    {
                        "class": "solr.
                            PorterStemFilterFactory"
                    }
                ]
            },
            "queryAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        StandardTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "
                            intro_section_synonyms.
                            txt",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    },
                    {
                        "class": "solr.
                            PorterStemFilterFactory"
                    }
                ]
            }
        },
        {
            "name": "convert_kingdom",
            "class": "solr.TextField",
            "indexAnalyzer": {
                "tokenizer": {
```

```
                    "class": "solr.
                        KeywordTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "
                            kingdom_synonyms.txt",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    }
                ]
            },
            "queryAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        KeywordTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "
                            kingdom_synonyms.txt",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    }
                ]
            }
        },
        {
            "name": "convert_conservation_status",
            "class": "solr.TextField",
            "indexAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        KeywordTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "cs_synonyms.txt
                            ",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    }
                ]
            },
            "queryAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        KeywordTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "cs_synonyms.txt
                            ",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    }
                ]
            }
        }
    ],
    "add-field": [
        {
            "name": "introduction",
            "type": "advanced_search",
            "indexed": true,
            "stored": true
        },
        {
            "name": "sections",
            "type": "advanced_search",
            "indexed": true,
            "stored": true
        },
        {
            "name": "kingdom",
            "type": "convert_kingdom",
            "indexed": true,
            "stored": true
        },
        {
            "name": "conservation_status",
            "type": "convert_conservation_status",
            "indexed": true,
            "stored": true
        }
    ]
}
```

Listing 2: Schema Semantic

```
{
    "add-field-type": [
        {
            "name": "advanced_search",
            "class": "solr.TextField",
            "indexAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        StandardTokenizerFactory"
                },
```

```
            "filters": [
                {
                    "class": "solr.
                        SynonymGraphFilterFactory
                        ",
                    "synonyms": "
                        intro_section_synonyms.
                        txt",
                    "expand": true
                },
                {
                    "class": "solr.
                        LowerCaseFilterFactory"
                },
                {
                    "class": "solr.
                        StopFilterFactory", "
                        words":"stopwords.txt"
                },
                {
                    "class": "solr.
                        PorterStemFilterFactory"
                }
            ]
        },
        "queryAnalyzer": {
            "tokenizer": {
                "class": "solr.
                    StandardTokenizerFactory"
            },
            "filters": [
                {
                    "class": "solr.
                        SynonymGraphFilterFactory
                        ",
                    "synonyms": "
                        intro_section_synonyms.
                        txt",
                    "expand": true
                },
                {
                    "class": "solr.
                        LowerCaseFilterFactory"
                },
                {
                    "class": "solr.
                        StopFilterFactory", "
                        words":"stopwords.txt"
                },
                {
                    "class": "solr.
                        PorterStemFilterFactory"
                }
            ]
        }
    },
    {
        "name": "convert_kingdom",
        "class": "solr.TextField",
        "indexAnalyzer": {
            "tokenizer": {
                "class": "solr.
                    KeywordTokenizerFactory"
            },
            "filters": [
                {
                    "class": "solr.
                        SynonymGraphFilterFactory
                        ",
                    "synonyms": "
                        kingdom_synonyms.txt",
                    "expand": true
                },
                {
                    "class": "solr.
                        LowerCaseFilterFactory"
                }
            ]
        },
        "queryAnalyzer": {
            "tokenizer": {
                "class": "solr.
                    KeywordTokenizerFactory"
            },
            "filters": [
                {
                    "class": "solr.
                        SynonymGraphFilterFactory
                        ",
                    "synonyms": "
                        kingdom_synonyms.txt",
                    "expand": true
                },
                {
                    "class": "solr.
                        LowerCaseFilterFactory"
                }
            ]
        }
    },
    {
        "name": "convert_conservation_status",
        "class": "solr.TextField",
        "indexAnalyzer": {
            "tokenizer": {
                "class": "solr.
                    KeywordTokenizerFactory"
            },
            "filters": [
                {
                    "class": "solr.
                        SynonymGraphFilterFactory
                        ",
                    "synonyms": "cs_synonyms.txt
                        ",
                    "expand": true
                },
                {
                    "class": "solr.
                        LowerCaseFilterFactory"
                }
            ]
        },
```

```
            "queryAnalyzer": {
                "tokenizer": {
                    "class": "solr.
                        KeywordTokenizerFactory"
                },
                "filters": [
                    {
                        "class": "solr.
                            SynonymGraphFilterFactory
                            ",
                        "synonyms": "cs_synonyms.txt
                            ",
                        "expand": true
                    },
                    {
                        "class": "solr.
                            LowerCaseFilterFactory"
                    }
                ]
            }
        },
        {
            "name": "speciesVector",
            "class": "solr.DenseVectorField",
            "vectorDimension": 384,
            "similarityFunction": "cosine",
            "knnAlgorithm": "hnsw"
        }
    ],
    "add-field": [
        {
            "name": "introduction",
            "type": "advanced_search",
            "indexed": true,
            "stored": true
        },
        {
            "name": "sections",
            "type": "advanced_search",
            "indexed": true,
            "stored": true
        },
        {
            "name": "kingdom",
            "type": "convert_kingdom",
            "indexed": true,
            "stored": true
        },
        {
            "name": "conservation_status",
            "type": "convert_conservation_status",
            "indexed": true,
            "stored": true
        },
        {
            "name": "vector",
            "type": "speciesVector",
            "indexed": true,
            "stored": true
        }
    ]
}
```

References

[1] [n. d.]. *HTTP 429 Error Code.* https://learn.microsoft.com/pt-br/azure/logic-apps/handle-throttling-problems-429-errors?tabs=consumption Accessed: 2024-12-10.
[2] [n. d.]. *Makefile.* https://makefiletutorial.com/ Accessed: 2024-11-17.
[3] [n. d.]. *NLTK.* https://www.nltk.org/ Accessed: 2024-12-13.
[4] [n. d.]. Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives. https://link.springer.com/protocol/10.1007/978-1-0716-0997-2_1 Accessed: 2024-11-14.
[5] [n. d.]. *Sentence Transformers.* https://sbert.net/ Accessed: 2024-12-13.
[6] [n. d.]. User centered and ontology based information retrieval system for life sciences. https://link.springer.com/article/10.1186/1471-2105-13-S1-S4 Accessed: 2024-11-14.
[7] 2012 – 2024. *Matplotlib Library.* https://matplotlib.org/ Accessed: 2024-10-08.
[8] 2024. *Beautifulsoup4 Library.* https://pypi.org/project/beautifulsoup4/ Accessed: 2024-10-07.
[9] 2024. *Material-UI.* https://mui.com/ Accessed: 2024-12-14.
[10] 2024. *Pandas Library.* https://pandas.pydata.org/ Accessed: 2024-10-10.
[11] 2024. *Pythonic XML Library.* https://pypi.org/project/lxml/ Accessed: 2024-10-11.
[12] 2024. *React.* https://react.dev/ Accessed: 2024-12-14.
[13] 2024. *Solr 9.7.0.* https://solr.apache.org/ Accessed: 2024-11-20.
[14] April 202. WikiSpecies. https://species.wikimedia.org/wiki/Main_Page Accessed: 2023-10-03.