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.

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 [3]. The topic covers varied types of information, such as taxonomy [2], 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 created 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. Finally, **Conclusions and Future Work** highlights the outcomes and sets the direction for the upcoming phases of the project.

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) [9].

2.3 Pipeline

Utilizing the LXML [7] 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 [5] 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.

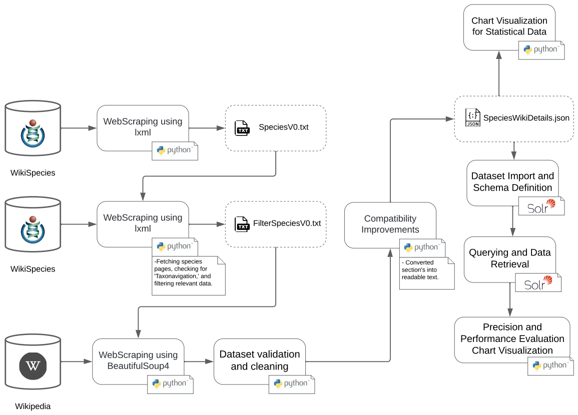
Chart Visualization for Statistical Data
python
WikiSpecies
WebScraping using lxml
python
SpeciesV0.txt
SpeciesWikiDetails.json
Dataset Import and Schema Definition
Solr
WikiSpecies
WebScraping using lxml
python
-Fetching species pages, checking for 'Taxonavigation,' and filtering relevant data.
FilterSpeciesV0.txt
Compatibility Improvements
python
- Converted section's into readable text.
Querying and Data Retrieval
Solr
Precision and Performance Evaluation Chart Visualization
python
Wikipedia
WebScraping using BeautifulSoup4
python
Dataset validation and cleaning
python

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.

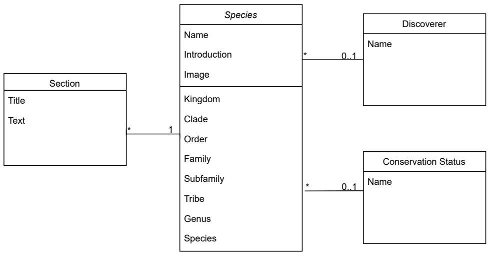

Figure 2: Conceptual Model.

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 [6] was utilized to analyze a CSV file, created with meaningful values from the JSON dataset, while Matplotlib [4] 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.

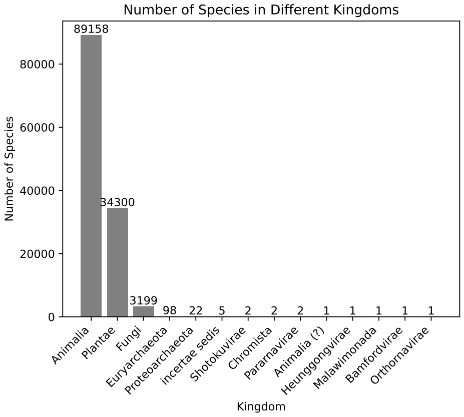

Figure 3: Species divided by kingdoms.

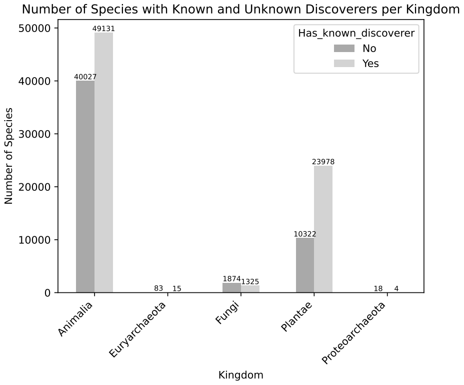

Figure 4: Known discoverer by species kingdom.

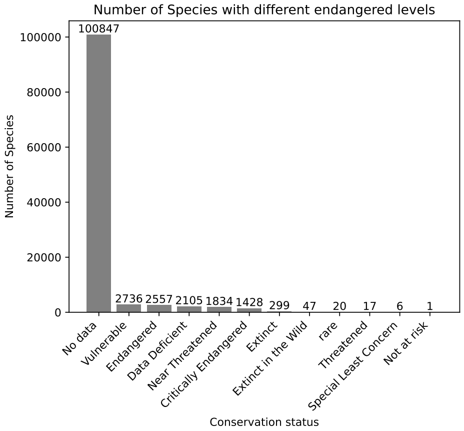

Figure 5: Species conservation status.

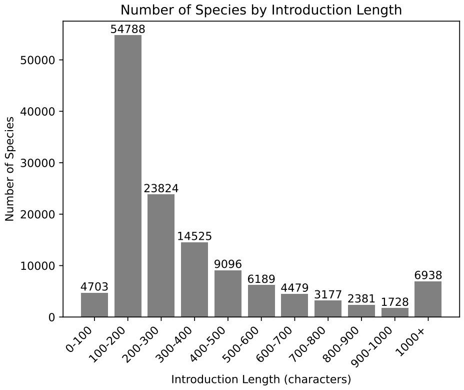

Figure 6: Species introduction length.

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 [8].

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 [1] 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 1.

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

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

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:
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.

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

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

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

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 Conclusions and Future Work

In conclusion, the completion of 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 this project phase, the challenge relied on the searching process of species when considering a specific information need. This was challenging due to having to deal with synonyms, and words in Latin that were not being considered when using the *PorterStemFilterFactory*. However, the Schema and Queries were completed successfully for this stage, as well as all the Documentation and evaluation done regarding these.

As we transition to the next stage, efforts will be focused on further improving the project. Therefore, 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, and to also experimenting with new tokenizers to improve the precision of our results. Another goal for the next phase is to develop a user interface application.

7 Annexes

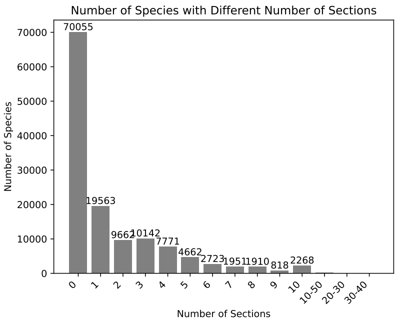

Figure 7: Species divided by number of sections.

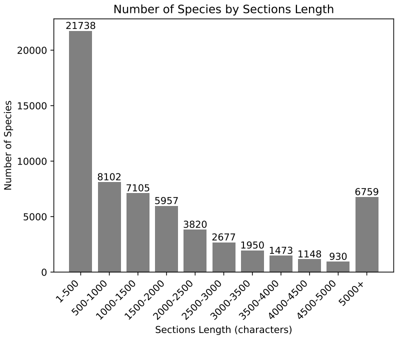

Figure 8: Species divided by total section length.

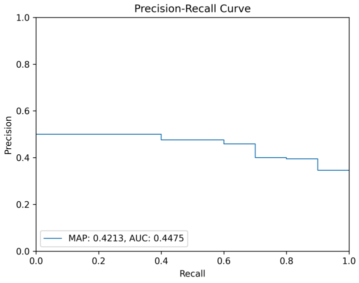

Figure 9: Precision recall for Query 1, System 1.

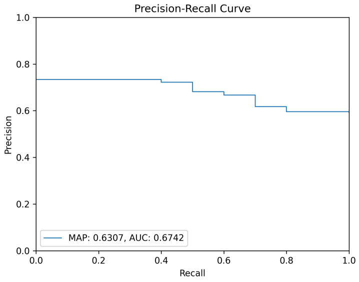

Figure 10: Precision recall for Query 1, System 2.

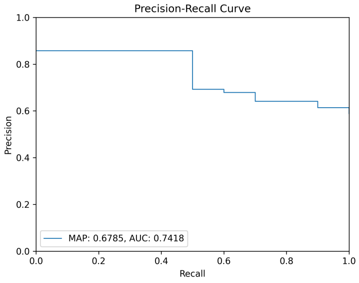

Figure 11: Precision recall for Query 2, System 1.

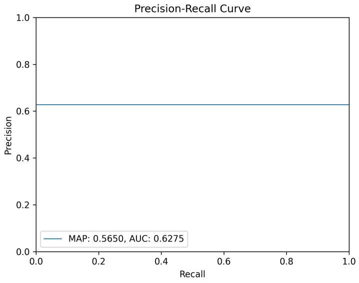

Figure 12: Precision recall for Query 2, System 2.

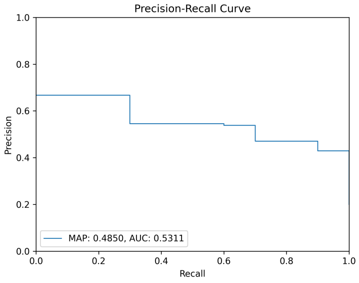

Figure 13: Precision recall for Query 3, System 1.

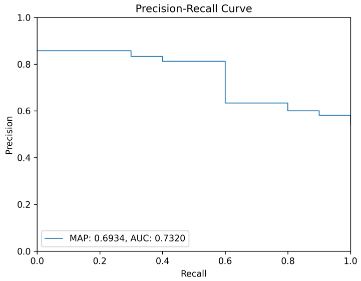

Figure 14: Precision recall for Query 3, System 2.

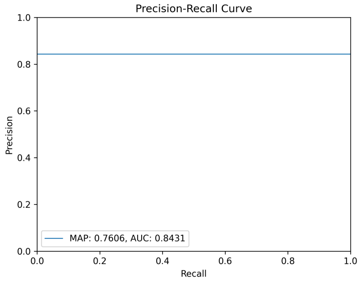

Figure 15: Precision recall for Query 4, System 1.

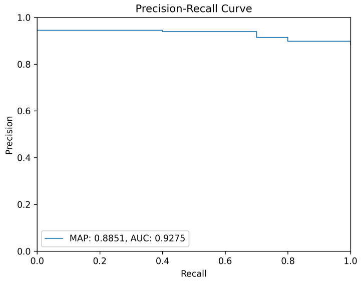

Figure 16: Precision recall for Query 4, System 2.

Listing 1: Schema developed

```
{
    "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
        }
    ]
}
```

References

[1] [n. d.]. *Makefile*. https://makefiletutorial.com/ Accessed: 2024-11-17.

[2] [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.

[3] [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.

[4] 2012 – 2024. *Matplotlib Library*. https://matplotlib.org/ Accessed: 2024-10-08.

[5] 2024. *Beautifulsoup4 Library*. https://pypi.org/project/beautifulsoup4/ Accessed: 2024-10-07.

[6] 2024. *Pandas Library*. https://pandas.pydata.org/ Accessed: 2024-10-10.

[7] 2024. *Pythonic XML Library*. https://pypi.org/project/lxml/ Accessed: 2024-10-11.

[8] 2024. *Solr 9.7.0*. https://solr.apache.org/ Accessed: 2024-11-20.

[9] April 202. WikiSpecies. https://species.wikimedia.org/wiki/Main_Page Accessed: 2023-10-03.