Faculty of Engineering, University of Porto
2024

BIOFINDER

Information Processing and Retrieval

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

PROJECT GOALS

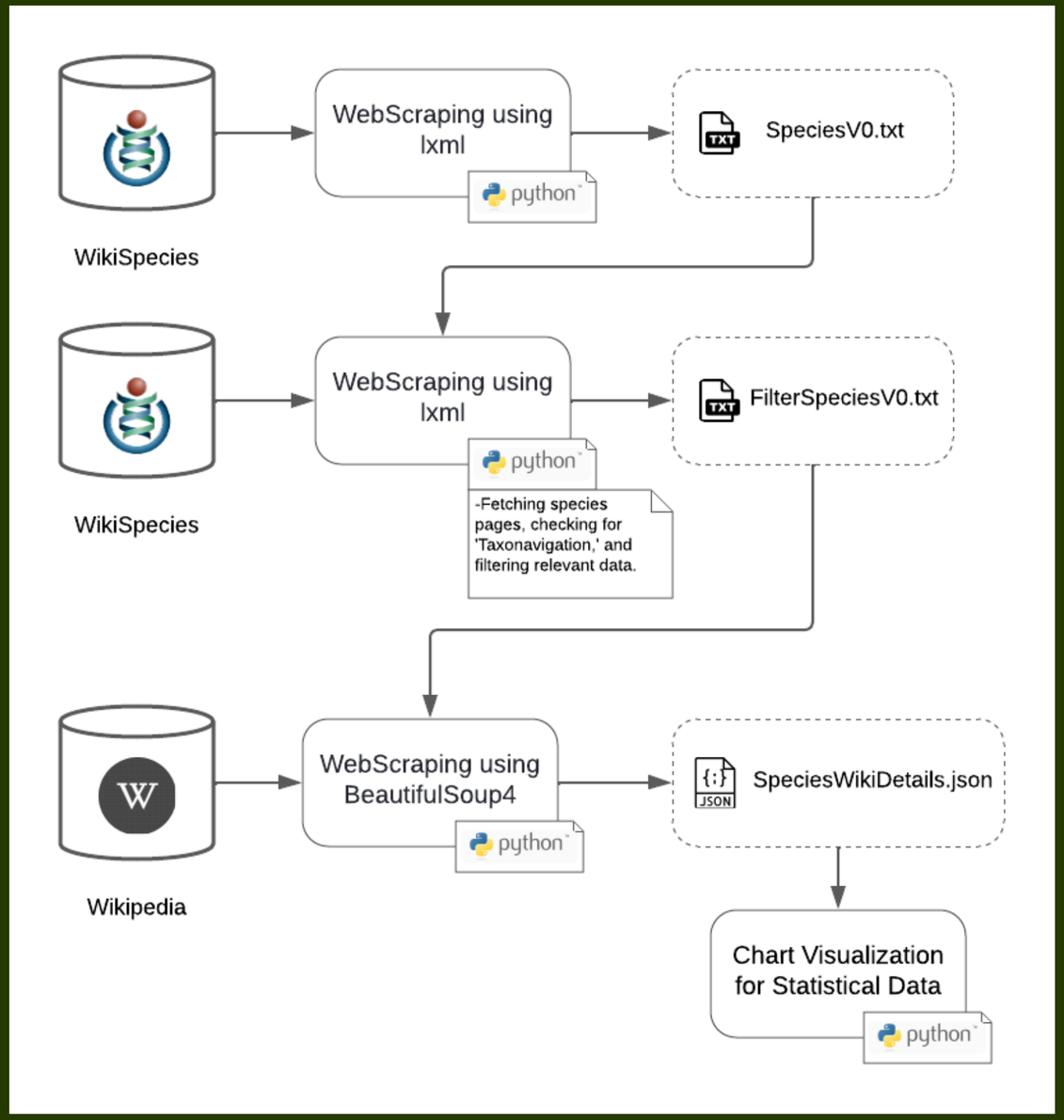
This project focuses on creating a search engine that can recover multiple types of information. The main goal is to develop a system that enhances the search and information recovery capabilities about many different species available in Wikipedia, providing a useful tool to study biodiversity.

The choice of the Species as a theme is interesting, since there is a growing amount of data available with a complexity that aligns with the project's central theme.

DATA PIPELINE

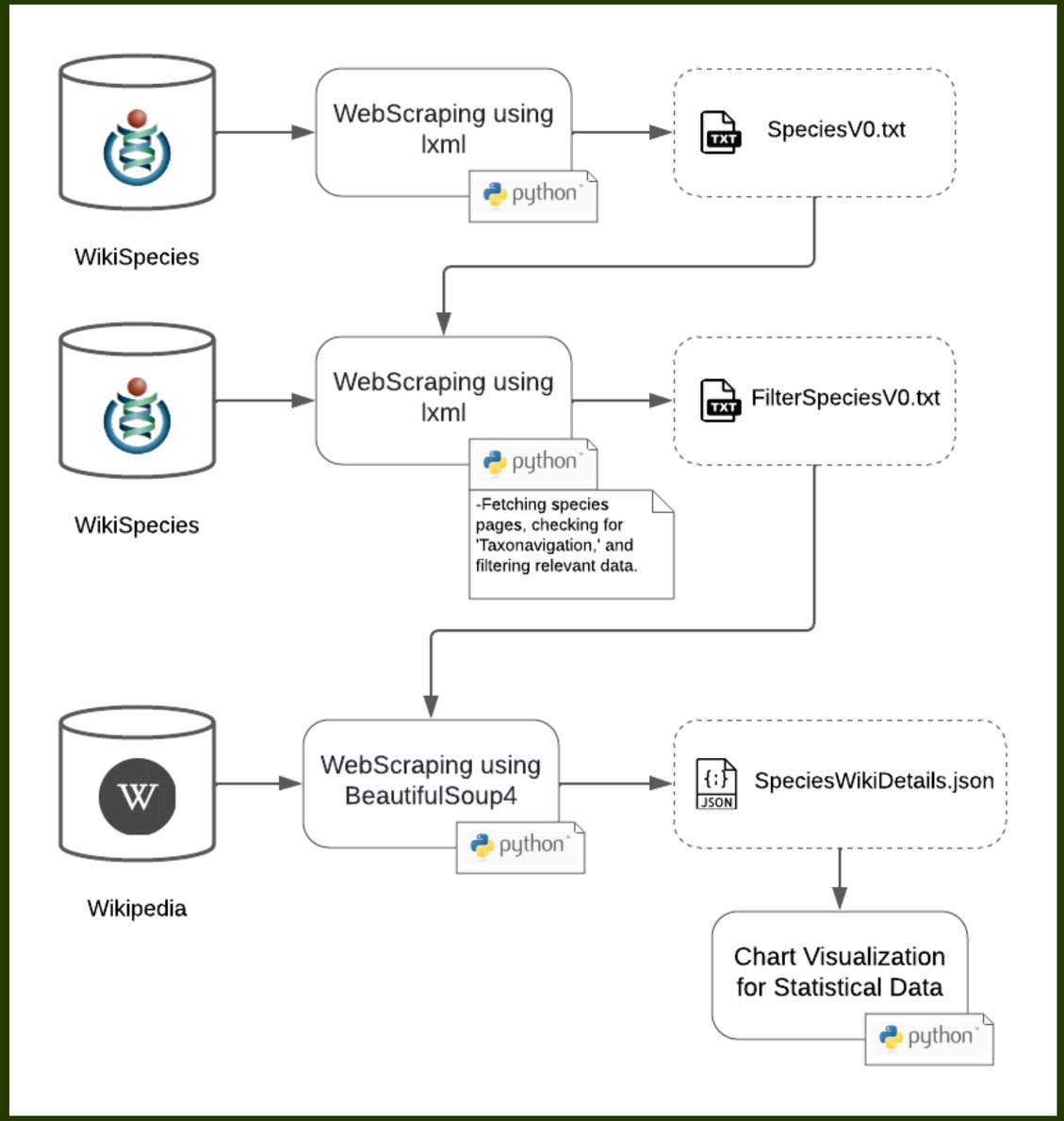
The data for our search engine was extracted from Wikipedia, based on the species names available at WikiSpecies. The dataset includes scientific names, descriptions, characteristics, and the species conservation status.

The module BeautifulSoup4 was used to extract information directly from Wikipedia's HTML code. Only the relevant data is filtered, collected, and stored in multiple JSON files, creating an organized structure for future analysis.

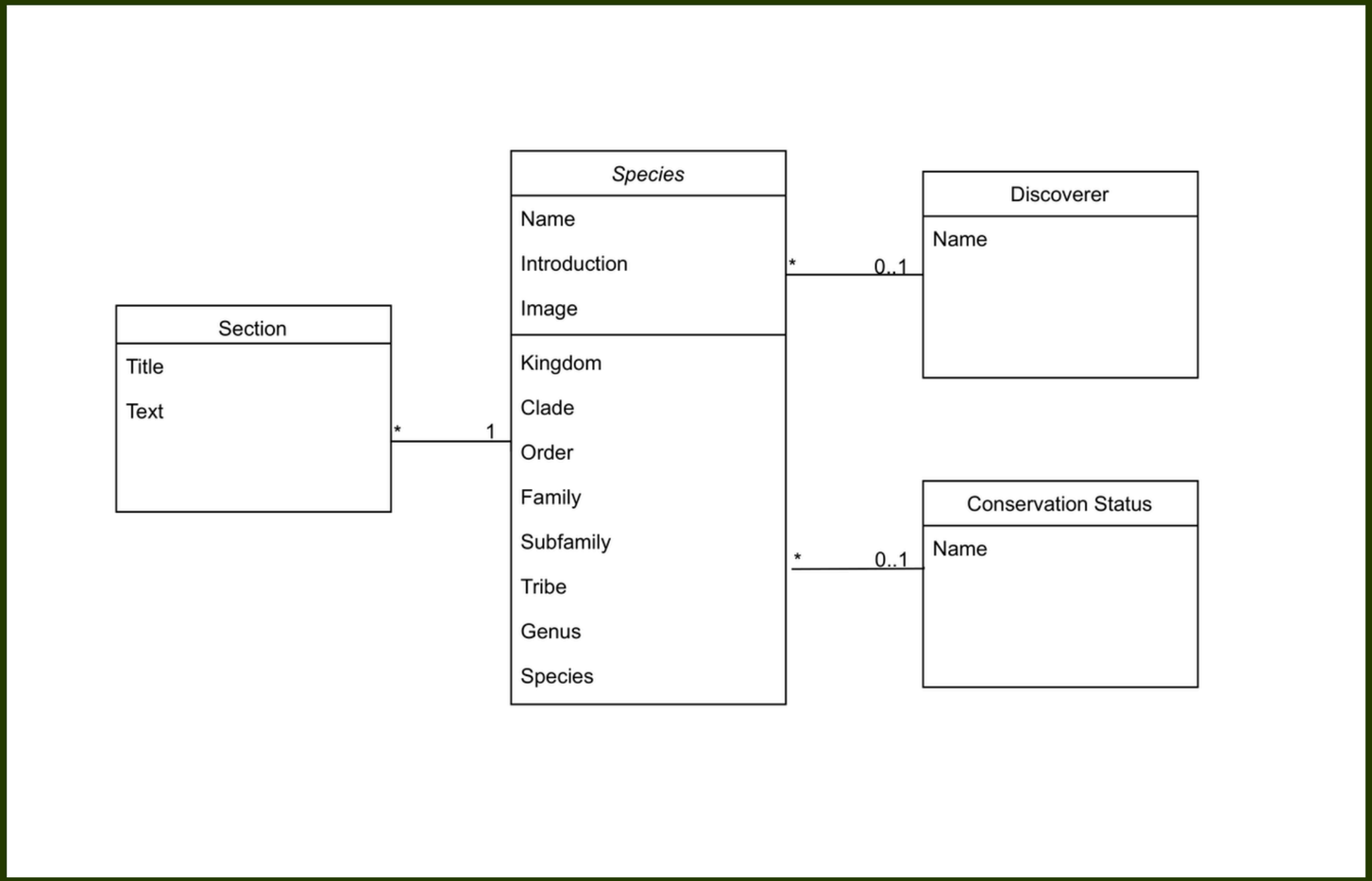


DATA CLEANING

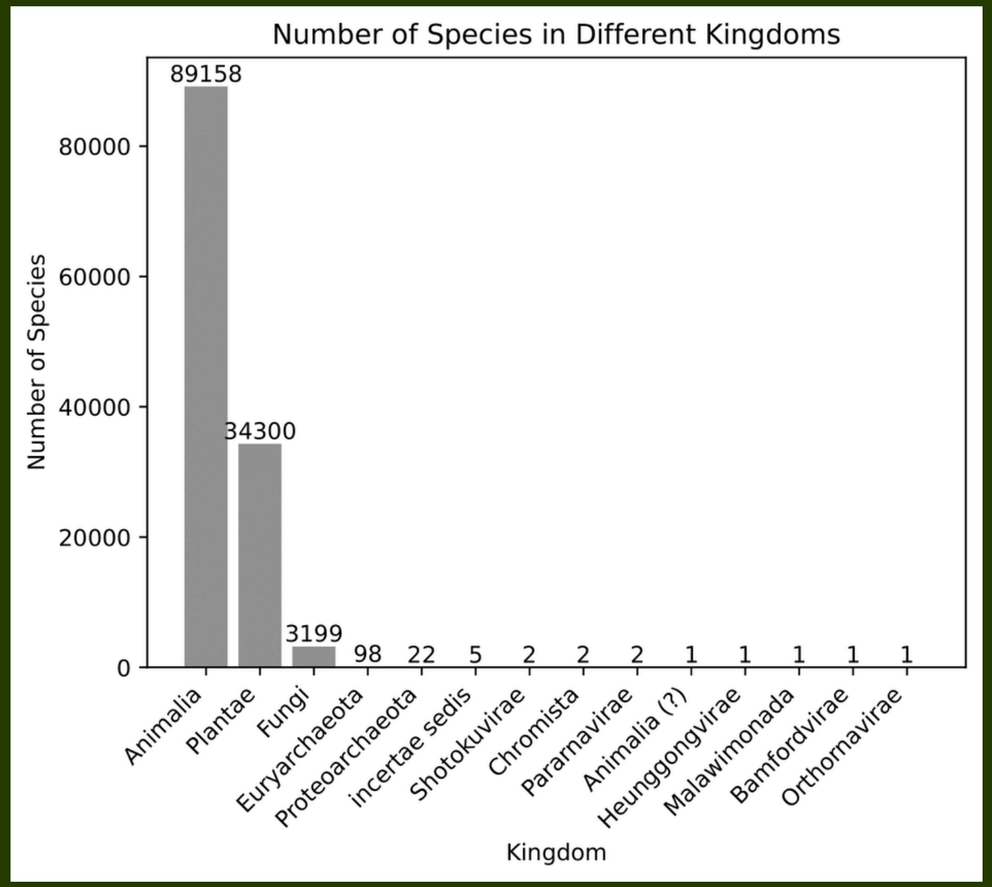
During the data cleaning phase, from the 760.000 entries at WikiSpecies, only around 180.000 had an actual Wikipedia page. With this collection of data gathered, we proceeded to eliminate duplicate entries, irrelevant information, and entries with insufficient descriptions. This resulted in a final dataset with around 130.000 unique entries.

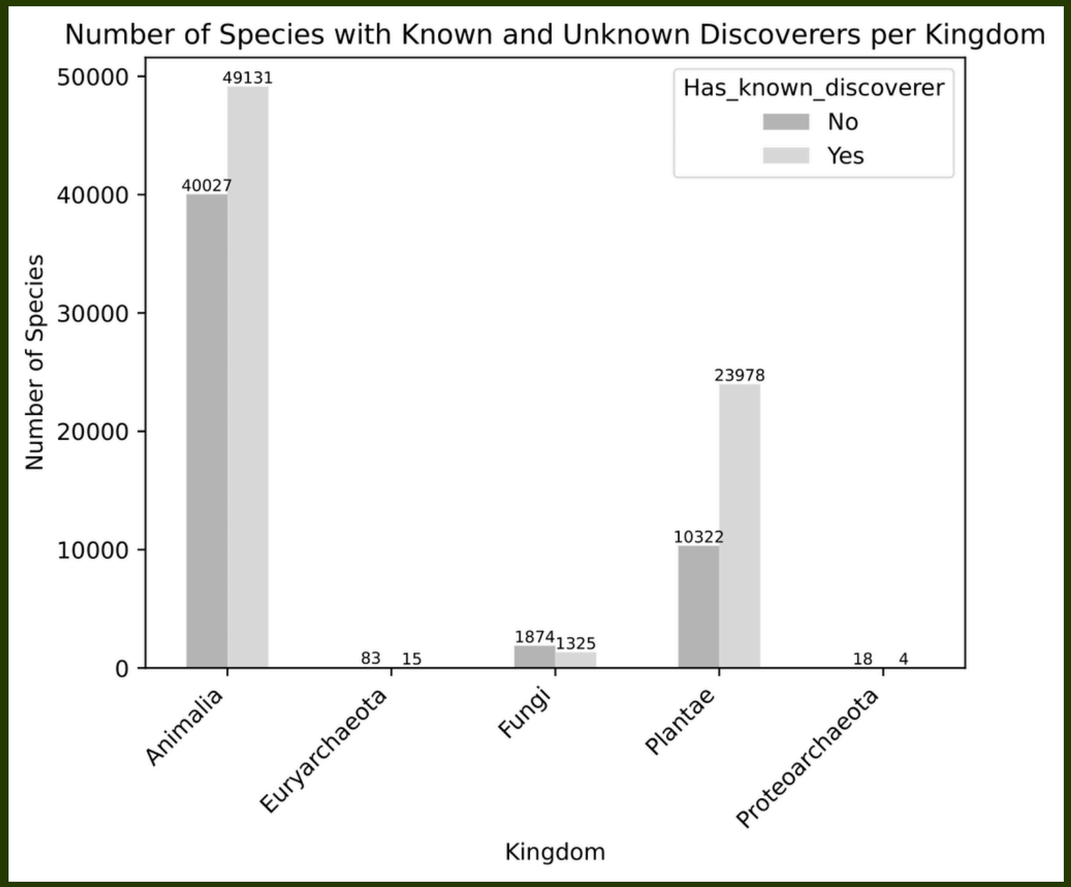


CONCEPTUAL MODEL



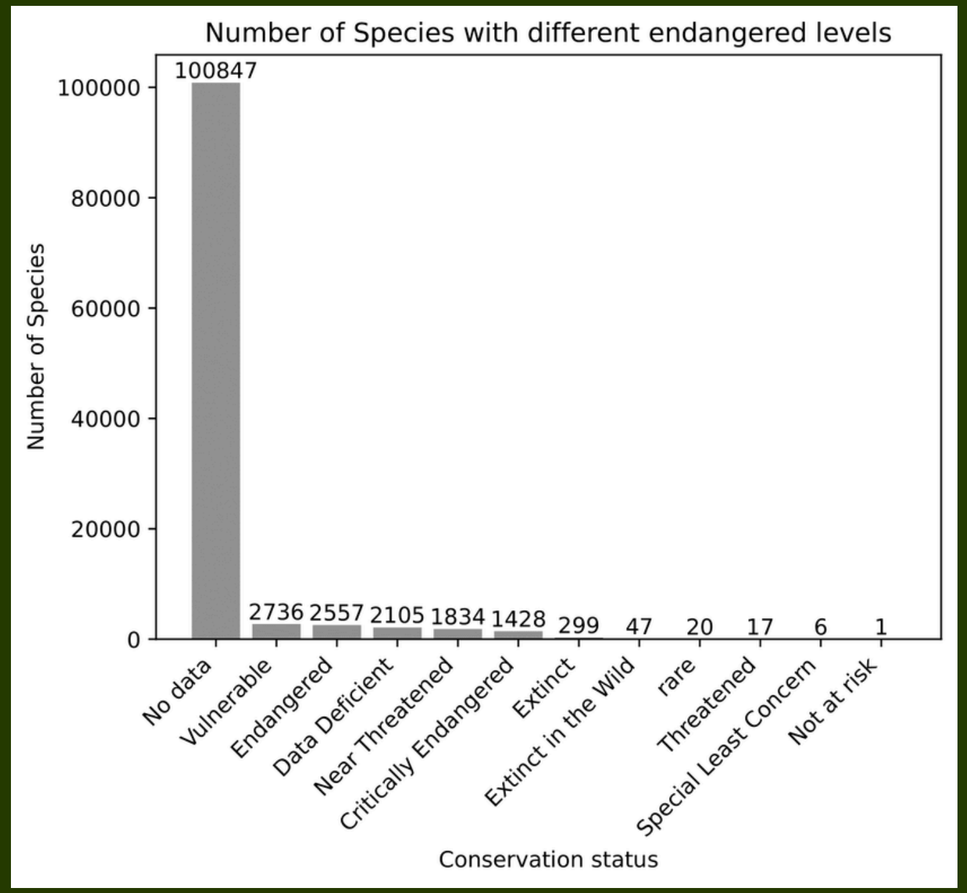
DATA CHARACTERIZATION

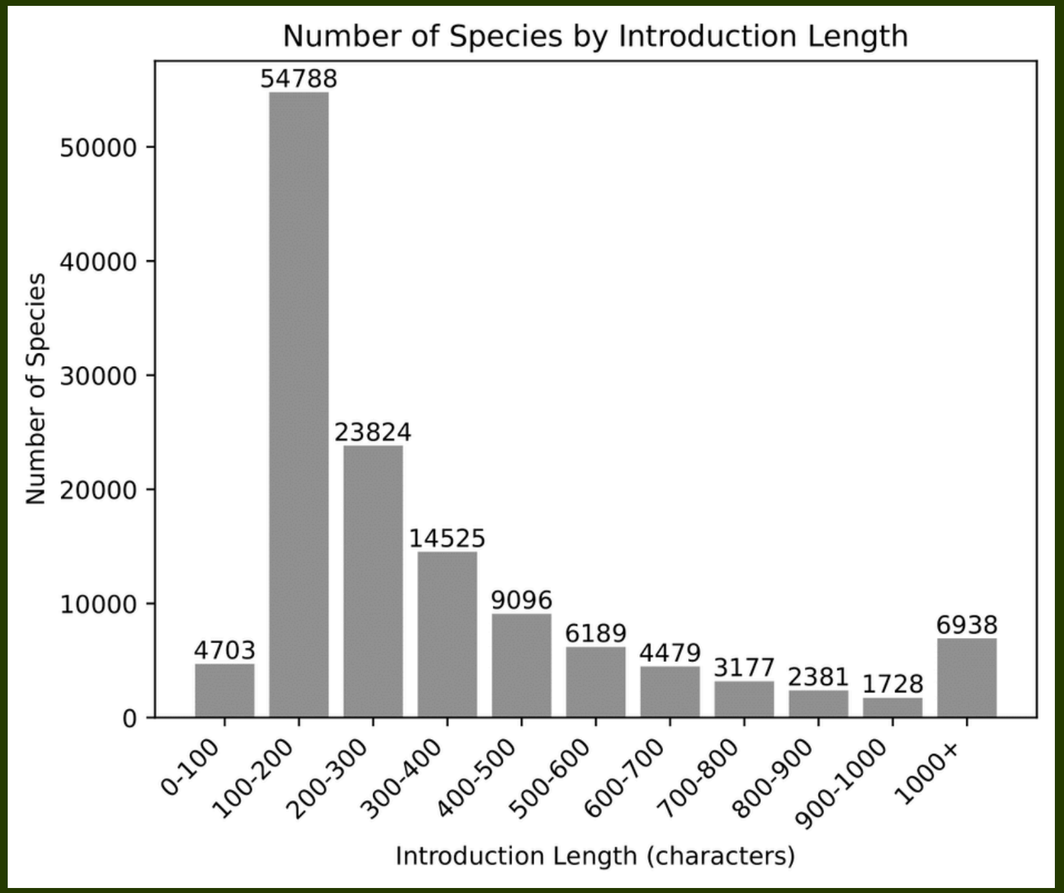




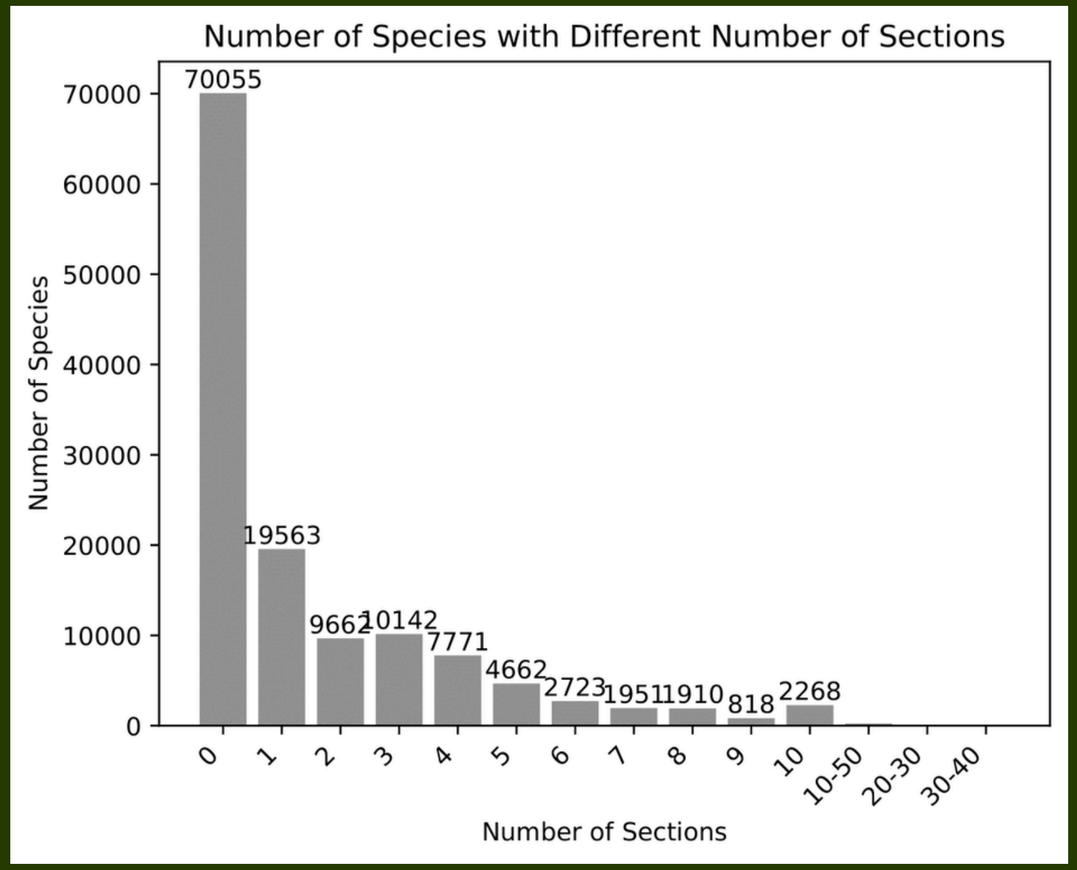
The final dataset was used to create graphics that facilitated the visualization of the distribution of species by kingdoms. Most species are from the kingdoms Animalia, Plantae, and Fungi, providing a clear vision of the huge biological diversity available at the data collection.

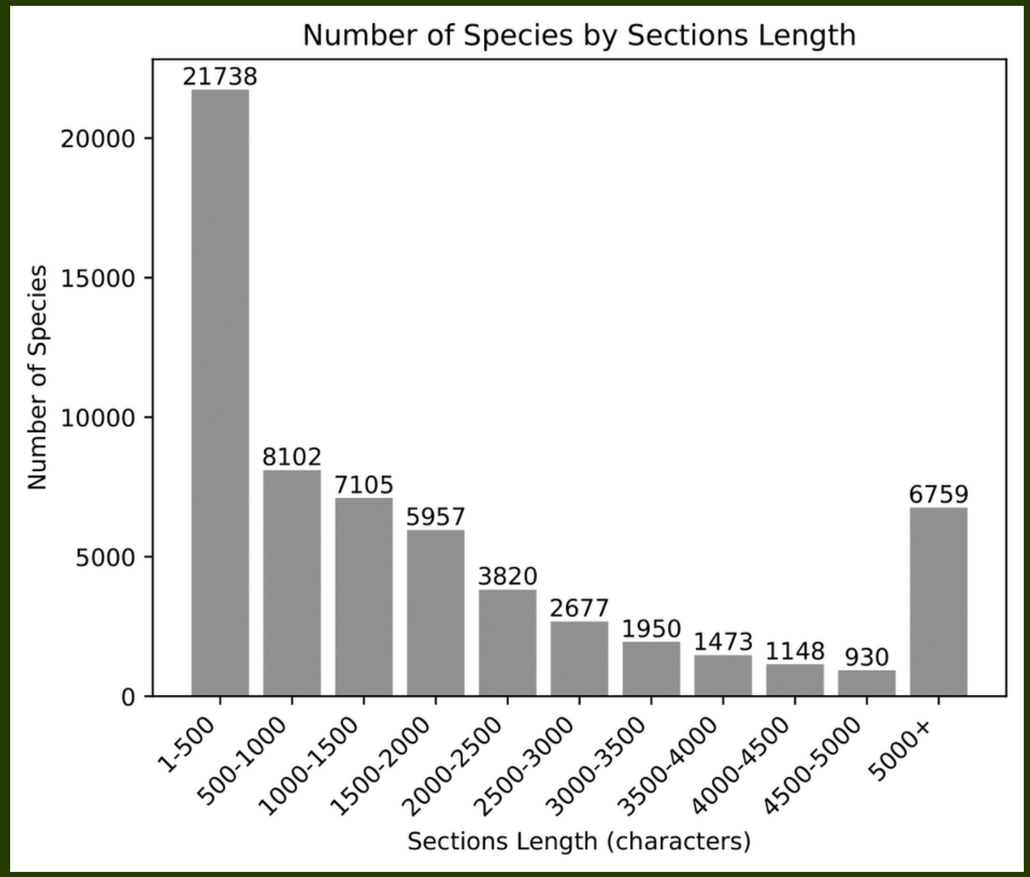
DATA CHARACTERIZATION





DATA CHARACTERIZATION





RESEARCH SCENARIOS

The data analysis about species revealed four distinct main categories, that offer valuable insights:

1. Conservation status: Terms like "Endangered", "Vulnerable", and "Extinct" highlight the importance of monitoring populational tendencies to protect the threatened biodiversity.
2. Species discovery: The exploration of the historical context of the taxonomy by investigating "species found by [Discoverer]" revealed the impact of these said findings in modern biology.
3. Scientific classification: References to kingdom, family, and order show the need to better understand the taxonomic hierarchy, and to study the evolutionary relationships between organisms.
4. Species detailed descriptions: Characteristics like size, color, and behavior help in the identification of the species and the comprehension of the vast existing biodiversity.

FUTURE WORK

The conclusion of the tasks of this phase of the project represents a significant advancement, which results in a well-organized data collection for detailed analysis.

An important challenge was to correctly process the species attributes from their corresponding Wikipedia species, including the need to deal with irrelevant introductions and missing values, the absence of conservation status, and the failures in information retrieval and processing.

The next phase's focus will be the development of a search engine that will allow the exploration of all species attributes identified so far.

Q&A

Any questions?