**Syracuse University**

# Synoptic Key of Life II

## Final Project Report

David Caspers

Christopher Murpy

La Monte Yarroll

IST 718, Big Data Analytics

25 March 2025

## Table of Contents

# Project Overview

In this project, we aim to contribute to the broader field of computational taxonomy by automating the parsing and categorization of text in biological journals to support the automatic construction of synoptic keys. A synoptic key is a flexible identification tool that allows users to classify organisms based on any observable characteristic rather than following a rigid, ordered decision tree, as in traditional binary keys. Automating the construction of synoptic keys is crucial for researchers because it significantly reduces the manual effort required to extract taxonomically relevant information from vast amounts of unstructured scientific literature. This, in turn, enhances species identification, accelerates biodiversity research, and improves taxonomic consistency across large datasets.

Our project addresses a critical challenge in text-based taxonomic analysis: mycological journal articles contain a mixture of structured nomenclatural information, detailed species descriptions, and general exposition. Extracting and classifying these components manually is labor-intensive, inconsistent, and difficult to scale. By implementing a PySpark-based classification pipeline, we seek to replace prior ad-hoc methods, enabling scalable and efficient parsing of mycological journals. This framework automates the identification and labeling of paragraphs into three categories—Nomenclature (species names), Description (detailed organism characteristics), and Miscellaneous Exposition (general text). Beyond streamlining text classification, this approach lays the foundation for feature extraction and the automated generation of synoptic keys, allowing researchers to rapidly identify fungal species based on descriptive text. Ultimately, our project enhances the accessibility and usability of mycological data, contributing to more efficient species identification and a deeper understanding of fungal biodiversity.

# Exploratory Analysis

## About the Data

Our dataset consists of biological literature from three major mycological journals: Mycologia, Mycotaxon, and Persoonia (Hennebert, Nauta, Murrill, Fungi). It is divided into two corpora: a labeled corpus containing 190 journal issues with manually annotated paragraphs and an unlabeled corpus of 1,021 journal issues derived primarily from OCR-processed PDFs. The labeled corpus includes 60,754 paragraphs, categorized into 6,072 nomenclature paragraphs containing species names, 6,192 description paragraphs detailing species traits, and 48,564 miscellaneous exposition paragraphs consisting of general text, introductions, and discussions. In total, this corpus contains 301,874 lines and 2,123,285 words, making it a valuable resource for training classification models. The unlabeled corpus, significantly larger, contains 4,272,880 lines and 25,552,432 words, presenting a wealth of raw textual data that requires structured classification and annotation before further analysis.

A major challenge with this dataset is in the diverse formatting styles found in scientific literature, compounded by inconsistencies introduced during OCR processing. The text structure varies significantly across journals and time periods, which introduces significant noise when

trying to parse meaningful features for classification. Our approach will leverage a combination of heuristics to parse the dataset into reasonable chunks and normalization techniques to better capture differentiating features for categorizing the texts to surpass prior ad-hoc methods.

## Visualizations:

- Paragraph Distance Analysis

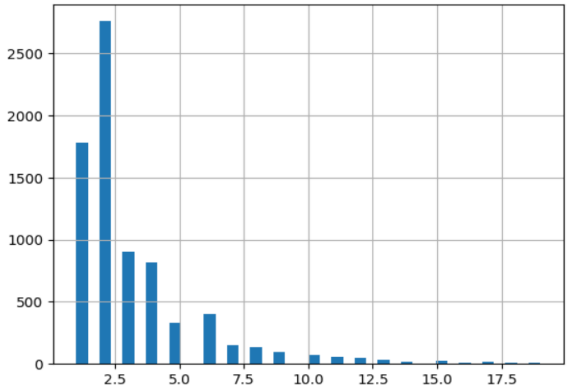


*Figure 1 - Distance Between Nomenclature Paragraphs and Corresponding Description Paragraphs*

An important aspect of our analysis is understanding the distance between nomenclature paragraphs (species names) and their corresponding descriptions. Ideally, species descriptions should follow immediately after nomenclature entries, but as can be seen in the chart above, formatting inconsistencies, OCR errors, and variations in journal styles introduce wide variations in paragraph separation. To assess whether these gaps reflect meaningful structural differences or dataset artifacts, we sampled a subset of the labeled paragraphs, which suggested that the heuristic methods used to capture paragraphs at times generated short “paragraphs”, consisting of only a few words. These artificially inflate measured distances and often consist of metadata, footnotes, or fragmented text from OCR processing rather than true interruptions in descriptive content. This finding underscores the need for our models to handle significant noise and variability in paragraph structure.

- Word Clouds


*Figure 2 - Misc Exposition Word Cloud*


*Figure 3 - Nomenclature Word Cloud*

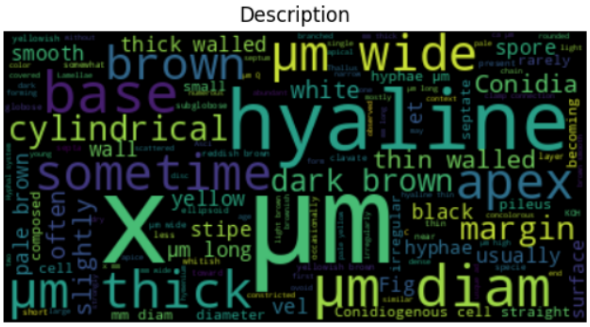


*Figure 4 - Description Word Cloud*

To gain insight into the most frequent terms associated with each classification category, we generated word clouds for Nomenclature, Description, and Miscellaneous Exposition paragraphs. These visualizations highlight distinguishing terms that classifiers can use to differentiate between categories. The Nomenclature word cloud prominently features species names and taxonomic terminology, which is expected, while the Description word cloud emphasizes morphological traits and scientific descriptors. The Miscellaneous Exposition cloud contains much more noise and includes general language related to introductions, discussions, and scientific context. This is expected as the Misc-exposition is a catch-all for other categories.

## Data Preparation and Transformations

Processing the mycological journal dataset required key transformations to structure the text for classification. The labeled corpus, provided in YEDDA annotation format, was parsed to extract paragraph labels—Nomenclature, Description, and Miscellaneous Exposition—and align them

with their corresponding text. This ensured that each paragraph retained its intended classification while preserving the dataset's structure.

For the OCR-processed unlabeled corpus, paragraph segmentation was more complex. We applied heuristic rules using line breaks, indentation, common acronyms, and formatting markers to define paragraphs. However, many were artificially short due to OCR artifacts, inflating paragraph distances and introducing noise. To mitigate this, we refined segmentation by merging fragmented text to improve classification accuracy and dropping empty lines.

To reduce noise, we used TF-IDF weighting, which assigns importance to words based on frequency while downweighting common, uninformative terms like author names and formatting artifacts. This ensures classifiers focus on taxonomy-specific language. Additionally, we extracted suffix-based features to enhance classification, particularly for species descriptions, by extracting the last 2, 3, and 4 letters of each word. Many mycological terms have characteristic suffixes (-aceae, -mycetes, -phore, -spore,  -ous, -ate, µm), which help distinguish nomenclature from descriptions. By extracting and analyzing these suffixes, we aim to improve the model's ability to distinguish between these paragraphs.

# Statistical Analysis

## Classification Models

To classify paragraphs into Nomenclature, Description, and Miscellaneous Exposition, we experimented with Logistic Regression and Random Forest classifiers using various feature engineering strategies. TF-IDF weighting was applied to emphasize key taxonomic terms while reducing noise from frequent but uninformative words. This also helped normalize sentences, which had varying lengths. Additionally, suffix-based features were extracted as detailed above to capture domain-specific morphological patterns commonly found in species descriptions. We evaluated the impact of these features by testing models with TF-IDF alone, suffixes alone, and a combination of both.

## Results

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **Logistic Regression, TFIDF** | 0.9418 | 0.9603 | 0.9666 | 0.9414 |
| **Random Forest, TFIDF** | 0.7878 | 0.7878 | 1.000 | 0.6943 |
| **Logistic Regression w/** | 0.9396 | 0.9593 | 0.9649 | 0.9392 |

| | | | | |
|---|---|---|---|---|
| **Suffixes Only** | | | | |
| **Logistic Regression w/ TFIDF + Suffixes** | 0.9421 | 0.9640 | 0.9630 | 0.9419 |

The best-performing model was Logistic Regression with both TF-IDF and suffix-based features, achieving an accuracy of 94.21% and an F1-score of 94.19%. This demonstrates that suffix-based features contribute additional discriminatory power, particularly for distinguishing description paragraphs from general exposition. The second-best model, Logistic Regression with only TF-IDF, achieved an accuracy of 94.18%, showing that TF-IDF alone is already a strong baseline for classification.

Interestingly, the suffix-only Logistic Regression model performed almost as well as the TF-IDF model, with an accuracy of 93.96% and an F1-score of 93.92%, indicating that morphological patterns alone provide substantial classification power. However, the Random Forest model using TF-IDF performed significantly worse, with an accuracy of only 78.78%. While its recall was perfect (1.00), its precision and F1-score were much lower, suggesting that it over-classifies certain categories, likely due to high variance and overfitting to noisy patterns in the text.

The results highlight that Logistic Regression is a more effective approach for this text classification task, particularly when TF-IDF and suffix-based features are combined. The strong performance of suffixes alone suggests that future models could further explore linguistic patterns commonly found in taxonomy, such as Latinized species names and morphological trait descriptors, could serve as a key feature in classification models. Additionally, Random Forest's weaker performance indicates that tree-based models may require further tuning or additional feature selection to handle high-dimensional textual data effectively.

# Challenges and Areas for Future Study

## Challenges

Several challenges arose during this study, particularly related to OCR artifacts, computational constraints, and the limitations of a single-node Spark environment. OCR-related challenges were a significant hurdle, as many scanned documents contained artifacts such as misaligned text, excessive hyphenation, and irregular spacing, which were difficult to clean without extensive manual intervention. This introduced additional noise into paragraph segmentation and classification.

Additionally, handling state in Spark presented difficulties. Since much of the NLP pipeline required processing individual lines before aggregating them into structured paragraphs, ensuring that order was maintained was challenging. Spark’s distributed nature inherently makes ordered

text processing more complex, as it does not natively preserve document structure. Addressing this issue required extra steps in grouping, reordering, and aligning text fragments, adding complexity to our pipeline.

Computational constraints also played a role in limiting certain techniques. TF-IDF generated over 20,000 dimensions. We initially aimed to apply Principal Component Analysis (PCA) for dimensionality reduction, but it proved computationally infeasible given the high number of features in our dataset. The lack of efficient dimensionality reduction meant that our models had to process high-dimensional TF-IDF feature spaces, which increased processing time and complexity.

Another limitation stemmed from Spark's single-node environment in Google Colab. While Spark offers distributed processing capabilities, our implementation was constrained to a single node, which negated many of its performance advantages beyond lazy loading. A truly distributed Spark environment would have significantly improved processing speed and likely made PCA feasible.

## Future Studies

The results suggest several promising directions for future research. One of the most significant findings was that suffix-based features were highly discriminative, particularly for distinguishing Nomenclature and Description paragraphs. Future studies could explore further refinements to suffix-based classification, possibly incorporating a taxonomy-specific morphological dictionary to enhance precision.

Additionally, refining paragraph segmentation is a key area for improvement. Many noisy paragraphs resulted from arbitrary breaks in text due to OCR artifacts, rather than natural linguistic or structural breaks. Future research could focus on developing better heuristics or deep learning-based approaches to detect true paragraph boundaries, reducing fragmentation and improving classification quality. However, this would require manipulating the labeled corpus to ensure consistent paragraph structuring, adding an extra layer of preprocessing complexity.

# Conclusion

This study demonstrates that automated classification of mycological journal text can be achieved with a high degree of accuracy using relatively simple models. Logistic Regression with TF-IDF and suffix-based features produced an accuracy of 94.21%, showing that structured linguistic patterns and term frequency information are highly effective in distinguishing between Nomenclature, Description, and Miscellaneous Exposition paragraphs. Even suffix-based classification alone proved to be highly discriminative, underscoring the potential of morphological patterns as key indicators in taxonomic literature.

A key strength of this approach is its scalability, enabled by Apache Spark. Spark's distributed computing framework allows for efficient processing of large text corpora, making it well-suited

for handling the growing volume of digitized biological literature. Although this study was conducted in a single-node environment, deploying the system on a fully distributed Spark cluster would significantly improve performance, allowing for faster model training, real-time classification, and more efficient feature extraction. This scalability makes it possible to expand automated taxonomic analysis beyond mycology to other biological domains, supporting broader efforts in computational taxonomy and biodiversity informatics.

The success of these approaches suggests significant opportunities for future work. By leveraging automated classification pipelines, researchers can streamline feature extraction for synoptic key construction, reducing the reliance on manual curation of species descriptions. Furthermore, with additional refinements—such as transformer-based models for contextual analysis and improved paragraph segmentation techniques—this approach can be expanded to enhance biodiversity research and improve the accessibility of mycological taxonomic data.

Ultimately, this work lays the foundation for scalable, automated synoptic key generation, providing a valuable tool for taxonomists, ecologists, and biodiversity researchers in organizing and extracting taxonomic insights from vast bodies of scientific literature.

# Works Cited

Balasi, Jen, Chris Murphy, Shintaro Osuga, La Monte Yarroll: “Synoptic Key of Life” PowerPoint presentation for IST 664, 2024, GitHub Repository, retrieved Feb 4, 2025.

Hennebert, G.L., Richard P. Korf eds., Mycotaxon: A New Journal on Taxonomy and Nomenclature of Fungi and Lichens, Ithaca, NY., 1974-2010.

Nauta, M.M., M.E. Noordeloos, eds., Persoonia: A Mycological Journal, Riksherbarium, Leiden, The Netherlands, 1959-1998.

Murrill, William A., Mycologia, New York Botanical Garden, Mycological Society of America, 1909-1961.

“Fungi – Periodicals.” *The Online Books Page*, University of Pennsylvania, https://onlinebooks.library.upenn.edu/webbin/book/browse?type=lcsubc&key=Fungi%20%2D%2D%20Periodicals&c=x.