Optimizing Botanical Data Integrity: A Comparative Study of Text Similarity Methods

Luma G. R. Cerqueira[1], **Carina F. Dorneles**[1], **Simone S. Werner**[1]

[1]Department of Informatics and Statistics – Federal University of Santa Catarina
Postal Address 88040-370 – Florianópolis – SC – Brazil

lumariios@gmail.com, carina.dorneles@ufsc.br, simone.werner@usfc.br

Abstract. *In this study, we address the challenges of managing authorship nomenclature as dictated by the International Code of Nomenclature for algae, fungi, and plants (ICN), within the Begoniaceae and Bignoniaceae families databases. Our goal was to evaluate various text similarity algorithms for their effectiveness in deduplicating botanical data, ensuring accuracy in authorship and synonymy. Our results highlighted Smith-Waterman's superior balance in precision, recall, and F1 Score, suggesting its potential as a robust solution for improving database integrity. The study also demonstrates the importance of fine-tuning these algorithms to navigate the unique challenges of botanical data management, emphasizing the necessity for specialized approaches in this field.*

1. Introduction

Taxonomy in biology seeks to name and organize biological diversity, allowing universal communication by assigning scientific names. This practice, initiated by Carl Linnaeus in the 18th century, relies on an extensive community of researchers who update taxon names and describe new species. With the rapid growth in the number of new species described annually and the expansion of herbaria globally (Figure 1), effective data management strategies in botanical databases have become crucial. The International Code of Nomenclature for algae, fungi, and plants (ICN) sets precise rules for indicating authorship in botanical nomenclature to ensure clarity, consistency, and traceability. Basically, these rules define the differentiation between the original authors who first described a species and those who later may reclassify the same species into a different genus, and this nomenclature is then used in scientific papers that describe the botanical species.

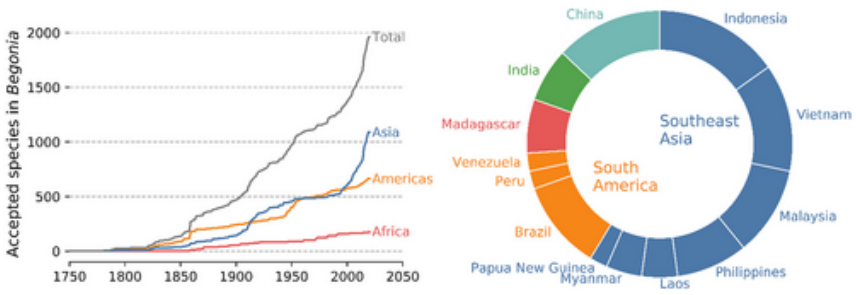

Figure 1. Accepted species in *Begonia* between 1750 and 2020 [Cheek et al. 2020]

The need to accurately represent authorship, which indicates who first described a species and any subsequent revisions or reclassifications, complicates data entry and retrieval in botanical databases. Issues of synonymy, where multiple names exist for a single

taxonomic entity due to historical revisions or taxonomic differences, further complicate database integrity. These challenges demand sophisticated similarity algorithms capable of recognizing and reconciling nuanced differences to ensure the consistency and reliability of botanical databases. Text similarity measurement algorithms play a fundamental role in identifying duplicate or erroneously cataloged records. Algorithms such as Levenshtein, Jaccard, Jaro-Winkler, Metaphone, N-grams, Smith-Waterman, and Fingerprinting have been widely explored for their effectiveness in detecting similarities between text strings, even with spelling errors or orthographic variations. Studies have demonstrated these algorithms' applicability in various biological contexts, emphasizing specific adaptations to increase their effectiveness in specialized databases. For example, in botanical collections of the *Begoniaceae* family, fine-tuning algorithms and similarity thresholds is necessary to accurately identify duplicates.

This study investigates the challenges posed by the "Authors" attribute in deduplicating data within botanical databases. Our research evaluates various similarity algorithms and thresholds to address these deduplication challenges, aiming to enhance the integrity and accuracy of biological databases. We conducted an empirical evaluation of different similarity functions to identify the most effective approach for handling potential duplicates, imprecise data, and misspellings. The results show improvements in data deduplication and standardization efforts, highlighting the effectiveness of tailored similarity algorithms in managing botanical information. However, the specific challenges presented by the databases in this study remain unsolved, indicating the need for further research and differentiating our work from related studies.

This article is organized as follows: Section 1 provides an introduction to the study, outlining the research problem and objectives. Section 2 discusses ICN rules for authorship data structure. Section 3 reviews related works. Section 4 describes the methodology, divided into Overview, Preprocessing, Similarity Function, and Threshold Choice. Section 5 covers the experimental evaluation, including Datasets, Evaluation Metrics, and Results. Section 6 discusses the results and their implications. Section 8 concludes the article, summarizing the main findings and suggesting avenues for future research. Section 9 acknowledges contributions, followed by the Bibliography.

2. ICN rules to Authorship Data Structure

The example of authorship shown in the Figure 2 illustrates some of the ICN authorship rules.

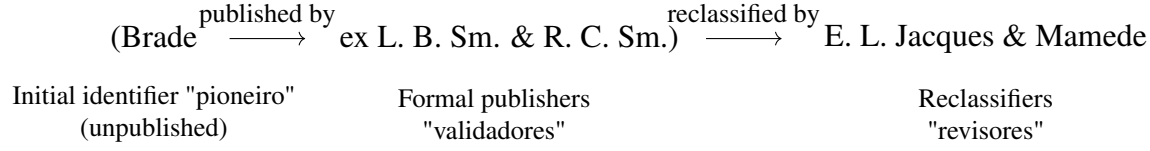

Figure 2. Authorship schema showing roles of different contributors as per ICN.

The ICN establishes intricate rules governing the attribution of scientific names. Authors listed outside the parentheses are those who originally published the name of the species, reflecting the original description and naming. When authors are mentioned within parentheses, it indicates that the species was originally described under a different

genus and later reclassified into a new genus, with the names inside the parentheses being the original describers and the name(s) outside the parentheses being those who performed the reclassification.

Central to these rules is the differentiation between the original authors who first described a species and those who later may reclassify the same species into a different genus. Authors whose names appear outside parentheses ("E.L. Jacques and Mamede" in Figure2) are credited with the initial description and naming of the species, they are labeled as "pioneiros" or "validadores". Conversely, when names are enclosed in parentheses, it signifies that the species was initially described under a different genus and has since been reclassified, with the parenthesized names belonging to the original describers and the names outside the parentheses to the reclassifiers, or "revisores". This nuanced approach not only maintains the historical integrity of species classification but also introduces a layer of complexity in the management and analysis of botanical data.

3. Related Works

This section reviews pivotal studies that leverage text similarity algorithms to address these challenges, demonstrating how they enhance data integrity and reliability in botanical databases.

Several studies focus on hybrid models and specific techniques for deduplication tasks. [Gyawali et al. 2020] present a hybrid model combining locality-sensitive hashing (LSH) and word embeddings to identify near and exact duplicates in scholarly documents, achieving a macro F1-score of 0.90. This method is beneficial for botanical databases where precise handling of minor variations in taxonomic descriptions and author names is required. Similarly, [Glick et al. 2020] explore various information-based similarity measures tailored for botanical databases, emphasizing the effectiveness of combining multiple similarity metrics to handle challenges such as synonymy and authorship variations. These studies highlight the importance of hybrid approaches in improving data accuracy and reliability in botanical data management.

Another group of studies provides comprehensive overviews and classifications of text similarity methods. [Gomaa and Fahmy 2013] offer a survey categorizing text similarity methods into string-based, corpus-based, and knowledge-based approaches, highlighting their strengths and weaknesses. This survey aids in selecting effective algorithms for botanical data management, with methods such as Jaccard, Levenshtein, Jaro-Winkler, and N-grams being directly applicable to improving data deduplication and standardization efforts. Complementing this, [Silva et al. 2019] discuss various text similarity measurement techniques and provide a classification system, offering insights into text distance and representation methods that refine algorithms used for deduplicating botanical data.

Some studies emphasize the importance of context and semantic understanding in text similarity measures. [Prakoso et al. 2021] focus on methods for measuring similarity in short texts, crucial for managing concise botanical descriptions and author names. Their review classifies these methods into string-based, corpus-based, knowledge-based, and hybrid-based categories, supporting the refinement of text similarity algorithms for more accurate deduplication of botanical data.

Additionally, tools designed for data quality and validation play a significant role

in improving botanical database management. [Silva et al. 2021] present a tool for validating and importing data into herbarium databases, addressing similar data quality issues as this study. The tool's implementation of filters and validations to check taxonomic and geographic data accuracy aligns with our goal of improving database integrity through rigorous data preprocessing and similarity checks.

These studies provide a robust foundation for applying text similarity and matching algorithms in botanical databases, highlighting the broader implications for ensuring data accuracy and reliability. By improving data deduplication and standardization, these methodologies significantly enhance the integrity and utility of botanical information systems. Our research situates itself within this context, empirically evaluating various similarity functions to identify the most effective approaches for handling potential duplicates, imprecise data, and misspellings, thereby facilitating accurate botanical research. Nonetheless, the challenge posed by the specific databases used in this study remains unresolved, indicating an area where further work is needed and differentiating our study within the field of biological database management.

4. Author's Name Similarity Method

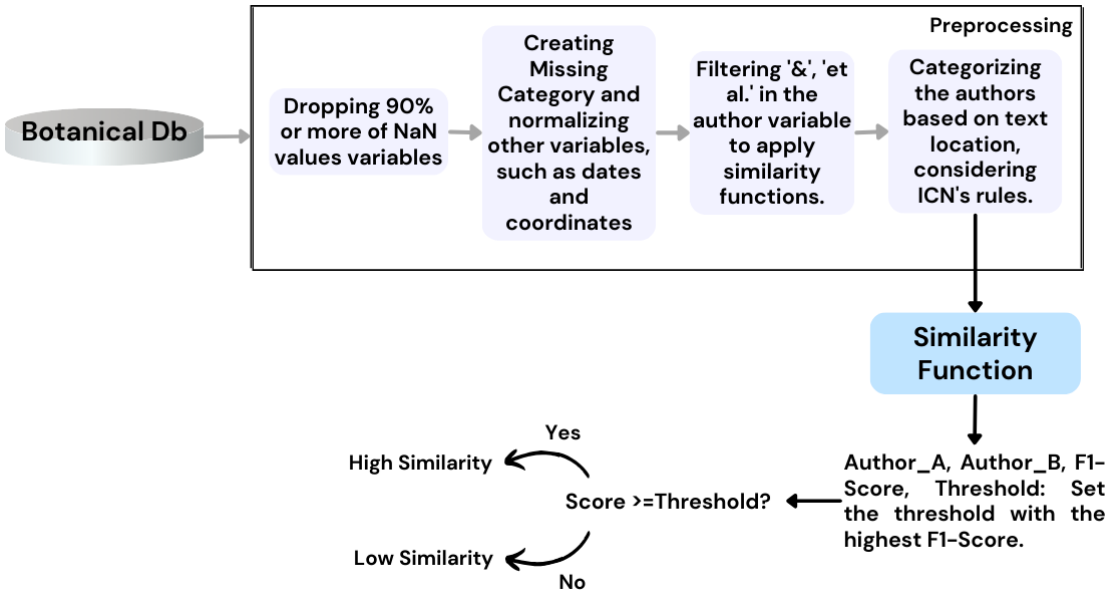

Figure 3. Methodology Overview

In the outlined methodology, as shown in the Figure 3, an initial data cleansing step is executed by discarding variables from the dataset that are comprised of 90% or more missing values. This is followed by the standardization of data types, with date columns converted to integers and all other columns to strings, which facilitates uniform data handling. A "Missing Category" is then introduced to systematically categorize any data that remains unaccounted for, ensuring comprehensive dataset integrity.

Subsequently, the approach involves filtering '&' and 'et al.' in the author variable to apply similarity functions. Authors are categorized based on text location, considering ICN's rules. Nine different similarity functions are then applied to quantify the likeness between author names, and a customized threshold is established to determine the criteria

for author name grouping by setting the threshold with the highest F1-Score. The final step employs evaluation metrics to assess the efficacy of the data processing techniques.

4.1. Preprocessing

The preprocessing step aims to ensure the integrity and usability of the data for further analysis. In this work, preprocessing involves cleaning the dataset and using a recursive function to flag the authors. Variables exhibiting more than 90% NaN (Not a Number) values were systematically removed from the dataset, as their high levels of missing data would likely contribute little to meaningful analysis and could introduce bias or inaccuracies in the study's findings.

A "missing value category" was additionally created to normalize other variables, such as dates and coordinates. This approach streamlined the dataset by focusing on variables that offer unique insights and contribute significantly to the biological data's diversity and specificity, ensuring a cleaner, more manageable dataset optimized for subsequent data analysis stages. Authors were categorized based on text location, as shown in Fig. 2, considering ICN's rules, and specific author name, as "Collector Unspecified" or "Taxononomia de grupo 2017/grupo verde" patterns, which don't specify the authors, were filtered and standardized to ensure consistency for applying similarity functions.

4.2. Threshold choice

The selection of thresholds for similarity algorithms is a critical step in text analysis and information retrieval applications. Thresholds determine whether two entities are considered similar or not, affecting the sensitivity and specificity of the algorithm in identifying matches. The optimal threshold often varies depending on the specific task, the nature of the data, and the desired balance between false positives and false negatives. [Manning 2008] discuss the importance of empirical validation in information retrieval settings. Adjustments based on observations allow for fine-tuning the algorithm's performance to the peculiarities of the specific dataset or application domain.

The threshold values were chosen based on empirical analysis, where researchers observed the algorithm's performance across a range of values and select the value that shows more matches made correctly. Building upon the critical importance of selecting thresholds for similarity algorithms in text analysis and information retrieval, as highlighted by [Manning 2008], it is also essential to standardize this selection process to facilitate comparative analysis across different algorithms and applications. To this end and to make a threshold observations and comparison, a structured approach involving the definition of a customized threshold approach were created. All thresholds were calculated, and the one with the highest F1 score was selected. In the event of a tie in the F1 score with more than one threshold, the lowest threshold value was chosen.

5. Experimental evaluation

The main objective of the experiments is to evaluate different similarity functions to identify the most effective approach for handling potential duplicates for indicating authorship in botanical nomenclature as dictated by the ICN. In this section, we present the dataset we have used, the ground truth we have built, the similarity functions we have chosen to test, the evaluation metrics, and the results.

5.1. Setup

The experimental setup was conducted using a Lenovo Ideapad 310-15ISK notebook, equipped with an Intel Core i5-6200U processor, 8GB of RAM, a 1TB HDD, running Windows 10. The computational tasks were executed on Google Colab, an online platform that provides access to cloud-based computing resources, allowing for the execution of Python code in a Jupyter Notebook environment. The Python libraries utilized in the experiment included ‘pandas‘ for data manipulation, ‘os‘ for operating system interactions, ‘google.colab.drive‘ for managing Google Drive connections, ‘random‘ for generating random numbers, ‘re‘ for regular expression operations, ‘seaborn‘ and ‘matplotlib.pyplot‘ for data visualization, and ‘itertools.combinations‘ for generating combinations of data elements. Notably, only two similarity functions relied on external libraries: Metaphone, utilizing the ‘metaphone‘ function, and N-grams, using the NLTK library. All other similarity functions were implemented from scratch.

5.2. The Datasets

SPLINK[1] is a digital platform that consolidates botanical data from various herbaria and collections across Brazil, enhancing research and conservation of Brazilian flora. It offers access to detailed records, including images, taxonomic classifications, and geographic distributions. By integrating data from multiple sources, SPLINK facilitates scientific study and promotes the visibility of Brazil's botanical diversity to a global audience. It integrates data from various herbaria within Brazil, offering access to a wealth of information including specimen images, taxonomic classifications, geographic locations of collections, collector details, and collection dates. From SPLINK, two botanical families were used: *Begoniaceae* and *Bignoniaceae*. The data for *Begoniaceae* comprises approximately 16,900 collections, while the data for *Bignoniaceae* comprises approximately 34,900 collections. In the *Begoniaceae* dataset from SPLINK, 25% of the dataset contained variables with 90% or more NaN values, while in the *Bignoniaceae* dataset, 47.70% of the data had 90% or more NaN values.

Brazil's botanical data ecosystem is further enriched by the REFLORA database. REFLORA[2] is a robust initiative aimed at digitizing and disseminating historical and contemporary botanical data pertinent to Brazil's flora. It hosts a comprehensive repository of specimens gathered from both national and international herbaria. This project plays a critical role in the recovery and digital archiving of Brazilian plant specimens, originally housed overseas. The database offers access to high-quality digital images of specimens, enhanced metadata, and vital taxonomic information. REFLORA's platform is instrumental in supporting research by providing a centralized resource that aids in the identification and study of plant species, promoting the conservation of Brazil's unique botanical heritage. From REFLORA, only the botanical family *Begoniaceae* was used. The data comprises approximately 1,900 collections, of which 25% of the dataset contained variables with 90% or more NaN values. Additionally, the ICN prescribes the use of specific designations such as "ex" (Figure 2 and Figure 4, item 14) and "&" to further detail the contributions of various authors to the taxonomic history of a species.

The "ex" notation is used when an author, "validadores", formally publishes a species name that was originally proposed by another, "pioneiros", often unpublished,

[1] `http://www.splink.org.br/`

[2] `http://www.reflora.jbrj.gov.br/`

Scientificnameauthor Frequency:

	name	frequency_scientificnameauthor
0	Willd.	1334
1	A.DC.	1085
2	Missing	1058
3	Raddi	1035
4	Schrank	930
5	(Klotzsch) A.DC.	857
6	Brade	854
7	Vell.	795
8	Irmsch.	653
9	Dryand.	355
10	Link	339
11	A. DC.	309
12	Thunb.	250
13	Schott	174
14	Schott ex A.DC.	167

Figure 4. The 15 most frequent author names in our database and differents applications of ICN's rules

author, thereby recognizing the contribution of both parties. The ampersand ("&") is employed to link multiple authors who jointly published the name of a species. These rules can be applied either jointly or individually. The presence of parentheses, "ex", "&", or any other delimiter are not conditioned upon each other. This makes the process of identifying duplicates, or even text similarity, unique for this variable in this type of database. These conventions, while facilitating a more precise attribution of authorship, pose significant challenges in data deduplication efforts within botanical databases, particularly when aligning records from diverse sources.

5.3. The Ground Truth

The ground truth was established based on the 114 unique author values identified within the *Begoniaceae* speciesLink's dataset, 48 unique author values in *Begoniaceae* REFLORA's dataset, and 151 unique values in *Bignoniaceae* speciesLink's dataset. This involved a manual process where, for each group of similar names, a single correct value was chosen to represent all variants. For instance, variations such as 'A. DC' (1 dot), 'A. DC.'(2 dots), 'A.D.C.'(3 dots), and 'A.DC.' (no space between letters) were consolidated under a singular, correct equivalent, 'A. DC'. This decision was predicated on the understanding that the aforementioned variants were not distinct entities but rather the result of typographical inconsistencies. By selecting one correct value for each set of similar names, the ground truth effectively rectifies these errors, serving as a critical reference for data cleaning and normalization efforts. This approach ensures that the dataset is both accurate and reliable, facilitating more precise analyses and interpretations.

Additionally, the data analysis executed during the ground truth creation revealed other challenges in the authors' list of unique values. Names such as 'Aitch.', 'Downs', 'Klotz.', 'Meisn.', and 'Moric.' do not appear as authors, reviewers, or validators in the publications describing the species. They are referenced to the International Plant Names Index (IPNI) website (https://www.ipni.org/), probably referencing plant names and indicating potential recording errors in the datasets. The value 'Hort. Berol.' was found as the author name, but is actually a botanical garden and museum in Berlin, not an author. The authors 'G.' and 'L. B.' appear in the database after processing but were absent before preprocessing, suggesting that the authors' names were separated during database preprocessing, highlighting the challenge of finding preprocessing solutions that do not

result in such issues.

Lastly, two authors were mistakenly recorded under the same abbreviation 'Gomes da Silva': Ary Gomes da Silva, whom the abbreviation is 'Gomes da Silva', and is the "pioneiro" for the species *Begonia mamedeana*, and Sandra Jules Gomes da Silva, whom the abbreviation is 'S. J. Gomes da Silva', and is the "pioneiro" for species such as *Begonia salesopolensis* and *Begonia jureiensis*. However, some records incorrectly use the same abbreviation for both, an error since even authors sharing a surname should have distinct abbreviations to accurately reflect their individual contributions. After these steps, to create the ground truth, we applied the similarity functions to identify the possibilities of two values representing the same real object among the unique values for authors names.

5.4. Similarity Functions

The following algorithms were incorporated into the methodology:

1. Jaccard Similarity: The Jaccard similarity measure assesses similarity and diversity between sets by comparing the intersection of items to the union of items. It finds application in scenarios where the presence or absence of features is more pertinent than their frequency, particularly in evaluating similarity between botanical species in databases.
2. Levenshtein Distance: The Levenshtein distance, or edit distance, quantifies dissimilarity between two strings by calculating the minimum number of operations needed for transformation. It encompasses insertions, deletions, or substitutions of single characters, proving valuable for rectifying typographical errors and accommodating minor variations in names.
3. Jaro-Winkler Similarity for Names: The Jaro-Winkler similarity algorithm specializes in comparing strings, particularly names, highlighting common prefixes. It assigns higher similarity scores to strings with similar beginnings, thus improving matching and correction of name variations.
4. Metaphone or Double Metaphone: Metaphone or Double Metaphone are phonetic algorithms encoding names based on pronunciation, facilitating comparison of names with similar sounds. These algorithms handle cases where variations in spelling result in similar or identical pronunciations.
5. N-grams: N-grams involve breaking names into sequences of contiguous letters or sounds of length 'n', capturing similarities in names by considering overlapping subsequences. It enhances the matching process by identifying structural similarities in names.
6. Smith-Waterman Similarity: The Smith-Waterman similarity algorithm, a local sequence alignment method prevalent in bioinformatics, identifies and corrects local similarities in names, accounting for sub-sequence variations.
7. Fingerprinting Algorithm: The Fingerprinting algorithm generates unique fingerprints for names, aiding in efficient comparison and identification of similarities. It employs binary vectors, known as molecular fingerprints, quantifying similarity using the Tanimoto coefficient, which evaluates overlap between binary vectors.

The use of these diverse algorithms aimed to comprehensively address the intricacies of variations in the representation of author's names in the biological databases.

5.5. Evaluation Metrics

To measure the effectiveness of our approach, we have used the classical metrics of evaluation: accuracy, precision, recall, and F1-measure [Baeza-Yates and Ribeiro-Neto 2008]. These evaluation metrics play vital roles in assessing the performance of author name deduplication algorithms. Ensuring accuracy and reliability in deduplication within botanical databases, especially concerning author names following ICN rules, is crucial for maintaining scientific record integrity.

Precision measures the proportion of accurately identified duplicates among all instances classified as duplicates, indicating the true positives rate. High precision is imperative to minimize false positives, particularly in botanical databases where merging distinct author names could result in significant information loss. Recall evaluates the algorithm's capability to detect all true duplicates within the dataset. In botanical databases, high recall ensures the correct identification of all variations of an author's name, conforming to ICN rules, belonging to the same individual, notwithstanding challenges posed by diverse name formats and potential typographical errors. The F1 score offers a harmonic mean of precision and recall, presenting a single metric that balances both the accuracy and completeness of the deduplication process. Given the equal importance of minimizing false positives (to prevent incorrect merges) and false negatives (to ensure comprehensive deduplication), the F1 score serves as a critical indicator of overall algorithm efficacy.

6. Results

Figures 5 presents the results of comparing the text similarity methods using the three botanical datasets: SPLINK for *Bignoniaceae*, SPLINK for *Begoniaceae*, and REFLORA for *Begoniaceae*. It shows results for precision, recall, and F1-score for each method across all three datasets: a) *Bignoniaceae* - SPLINK, b) *Begoniaceae* - SPLINK, and c) *Begoniaceae* - REFLORA.

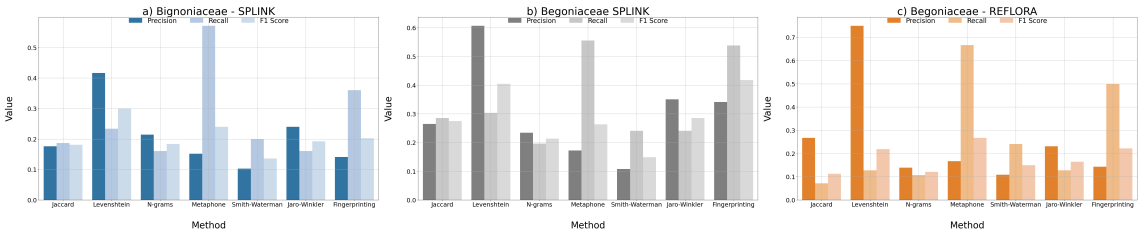

Figure 5. Comparison of text similarity methods applied to three botanical datasets (*Bignoniaceae* - SPLINK, *Begoniaceae* SPLINK, and *Begoniaceae* - REFLORA). The three graphs of subplots (a, b and c) presents bar charts showing precision, recall, and F1 score for each method across the three datasets.

The Levenshtein method achieved the highest precision among all methods, particularly in the *Begoniaceae* REFLORA database (Figure 5, c). However, its lower recall indicated it missed certain genuine similarities. Conversely, the Smith-Waterman method demonstrated a balanced performance with high scores in precision, recall, and F1 score across all databases (Figure 5, a, b, c), indicating robust capability in identifying true

similarities while maintaining low false positives and negatives. While the Metaphone method displayed high recall across all databases, especially in *Begoniaceae* REFLORA and SPLINK (Figure 5, b, c), but its lower precision suggested a higher incidence of false positives. Similarly, the Jaccard index showed consistent performance with moderate precision and recall, leading to balanced F1 scores (Figure 5, a, b, c), while the Jaro-Winkler method had high recall but lower precision, indicating more false positives (Figure 5, a, b, c).

For *Begoniaceae* in the SPLINK database, the Fingerprinting method achieved high precision and recall but performed lower in the REFLORA database, suggesting variability based on dataset (Figure 5, b, c). The N-grams method demonstrated moderate precision and recall, resulting in moderate F1 scores (Figure 5, a). Both Jaccard and Smith-Waterman showed high precision for *Bignoniaceae* in SPLINK, with Smith-Waterman achieving better balance (Figure 5, a). The Metaphone method's highest recall suggests it effectively identifies relevant similarities without false negatives, despite its lower precision (Figure 5, b, c). On the other hand, the Levenshtein method provided balanced performance, excelling in both precision and recall, and achieving a good F1 score (Figure 5, a, b, c). The Jaccard index was robust in precision and recall for *Bignoniaceae* in SPLINK (Figure 5, a), whereas Jaro-Winkler's high precision but lower recall indicated more false negatives (Figure 5, a, b, c).

A particular observation for the N-grams method in REFLORA (Figure 5, c) was its moderate precision and recall, highlighting the importance of dataset size and threshold settings. These results underscore the necessity of selecting appropriate text similarity methods based on the specific requirements of precision, recall, and F1 score. The Metaphone and Smith-Waterman methods demonstrated superior performance in this comparative analysis.

7. Discussion

The comparative analysis of text similarity methods underscores the criticality of selecting appropriate techniques tailored to specific application needs, particularly regarding precision, recall, and F1 Score. The Metaphone method exhibited exceptional recall, achieving perfect scores in both *Begoniaceae* and *Bignoniaceae* on the SPLINK database, as well as in the REFLORA database for *Begoniaceae*. It's high recall indicates its effectiveness in capturing all potential matches, making it particularly suitable for applications where minimizing false negatives is crucial. However, its lower precision suggests a higher incidence of false positives, necessitating its combination with other methods to ensure accurate duplicate identification.

Conversely, the Levenshtein method demonstrated a balanced performance, with notable scores in precision, recall, and F1 Score, suggesting a robust capability to identify true similarities while maintaining a lower rate of false positives and negatives. Levenshtein's high precision is possibly due to its sensitivity to small differences between strings, making it particularly effective at identifying exact or near-exact matches. However, this same sensitivity can lead to lower recall, as the method may fail to capture legitimate variations in strings that represent the same entity, especially when dealing with spelling errors or alternative abbreviations.

The Jaccard index presented consistent performance across different datasets,

with moderate precision and recall scores leading to balanced F1 Scores. The Jaro-Winkler method also showed relatively high precision scores in most datasets, indicating its strength in identifying accurate matches. However, its recall was lower, which may point to a higher incidence of false negatives. In the assessment of *Begoniaceae* species within the SPLINK database, the N-grams method demonstrated moderate performance with lower precision and recall scores, leading to a lower F1 score. The Fingerprinting method, while achieving high precision in both the SPLINK and REFLORA databases, had a significantly lower recall, highlighting a tendency to miss genuine similarities. It's high recall is due to Fingerprinting's ability to recognize and match a broad spectrum of similar patterns across different names. However, like Metaphone, its broad approach can result in lower precision because it may overgeneralize, grouping distinct names together based on shared features that do not necessarily indicate identical entities.

Analyzing the *Bignoniaceae* species within the SPLINK database, the Smith-Waterman method showed high precision and recall scores across all databases, achieving a harmonious balance as evidenced by its F1 score. This balance indicates its robust capability in identifying true duplicates while minimizing false positives and negatives, making it a versatile choice for various datasets and applications.

These results underscore the importance of selecting appropriate text similarity methods based on the specific requirements of precision, recall, and F1 score. The superior performance of the Metaphone and Smith-Waterman methods in our comparative analysis highlights their potential as robust solutions for improving database integrity.

8. Conclusion and Future Works

This analysis of text similarity methods reveals performance variations based on precision, recall, and F1 Score metrics, underscoring the importance of selecting methods that align with specific application needs. The Smith-Waterman method emerged as a balanced and versatile choice, performing reliably across all metrics and highlighting its applicability to diverse text similarity tasks, ensuring data integrity across botanical datasets. Overall, these findings emphasize the necessity for tailored approaches in text similarity assessments. Future work will focus on developing a technique to address these challenges in the specific context of botanical databases using Large Language Model - LMM.

9. Acknowledgements

The authors thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the financial support to out research with following process number: 131227/2023-8.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (2008). *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition.

Cheek, M., Nic Lughadha, E., Kirk, P., Lindon, H., Carretero, J., Looney, B., Douglas, B., Haelewaters, D., Gaya, E., Llewellyn, T., Ainsworth, A. M., Gafforov, Y., Hyde, K., Crous, P., Hughes, M., Walker, B. E., Campostrini Forzza, R., Wong, K. M., and

Niskanen, T. (2020). New scientific discoveries: Plants and fungi. *PLANTS, PEOPLE, PLANET*, 2(5):371–388.

Glick, J. et al. (2020). Information-based similarity measures for botanical data. *Journal of Data Science and Botanical Information*, 8(2):101–119.

Gomaa, W. H. and Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13):13–18.

Gyawali, B., Anastasiou, L., and Knoth, P. (2020). Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 901–910, Marseille, France. European Language Resources Association (ELRA).

Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing,.

Prakoso, D. et al. (2021). Short text similarity measurement methods: A review. *Journal of Big Data and Analytics in Practice*, 3(1):33–44.

Silva, C. et al. (2019). Measurement of text similarity: A survey. *Information*, 11(421):1–25.

Silva, J. et al. (2021). Tool for validation and import in herbarium database. In *Proceedings of the Botanical Data Conference*, pages 123–130. Botanical Society.