Essential Biodiversity Variables: extracting plant phenological data from specimen labels using machine learning

Maria Auxiliadora Mora-Cross‡, Adriana Morales-Carmiol‡, Te Chen-Huang‡, María José Barquero-Pérez‡

‡ School of Computer Engineering, Costa Rica Institute of Technology, Alajuela, Costa Rica

Corresponding author: Maria Auxiliadora Mora-Cross (mariamoracross@gmail.com)

Academic editor: Editorial Secretary

Abstract

Essential Biodiversity Variables (EBVs) make it possible to evaluate and monitor the state of biodiversity over time at different spatial scales. Its development is led by the Group on Earth Observations Biodiversity Observation Network (GEO BON) to harmonize, consolidate and standardize biodiversity data from varied biodiversity sources. This document presents a mechanism to obtain baseline data to feed the Species Traits Variable Phenology or other biodiversity indicators by extracting species characters and structure names from morphological descriptions of specimens and classifying such descriptions using machine learning (ML).

A workflow that performs Named Entity Recognition (NER) and Classification of morphological descriptions using ML algorithms was evaluated with excellent results. It was implemented using Python, Pytorch, Scikit-Learn, Pomegranate, Python-crfsuite, and other libraries applied to 106,804 herbarium records from the National Biodiversity Institute of Costa Rica (INBio). The text classification results were almost excellent (F1 score between 96% and 99%) using three traditional ML methods: Multinomial Naive Bayes (NB), Linear Support Vector Classification (SVC), and Logistic Regression (LR). Furthermore, results extracting names of species morphological structures (e.g., leaves, trichomes, flowers, petals, sepals) and character names (e.g., length, width, pigmentation patterns, and smell) using NER algorithms were competitive (F1 score between 95% and 98%) using Hidden Markov Models (HMM), Conditional Random Fields (CRFs), and Bidirectional Long Short Term Memory Networks with CRF (BI-LSTM-CRF).

Keywords

Essential Biodiversity Variables, plant phenology, Natural Language Processing, machine learning, Text Classification, Named Entity Recognition.

Introduction

Biological diversity is a fundamental pillar of life on Earth. Therefore, the governments of the world committed themselves through the United Nations Convention on Biological Diversity (CBD) to reduce the loss of biodiversity by intending to meet the Aichi Biodiversity Targets Convention on Biological Diversity (CBD) 2011. However, the ambitious Aichi Biodiversity Targets proposed in the 2011-2020 Strategic Plan for Biodiversity regarding this subject were not achieved. According to reports from different countries to the CBD, the causes of failure related to knowledge and technologies included the lack of biodiversity data for relevant taxa and locations and the lack of monitoring systems to support conservation actions Secretariat of the Convention on Biological Diversity 2020.

Essential Biodiversity Variables (EBVs) are recommended as a global biodiversity monitoring and reporting system to assess the state of biodiversity over time. They provide the basis for generating biodiversity indicators that allow repeated assessments of progress against national and global conservation goals (e.g., the Sustainable Development Goals and the Aichi Biodiversity Targets) Pereira et al. (2013), Kissling et al. (2018), Hardisty et al. (2019), Turak et al. (2017). The selected variables were proposed by a group of international ecologists led by the Group on Earth Observations Biodiversity Observation Network (GEO BON). The 22 EBV candidates were suggested in 2013 and organized into six classes (i.e., genetic composition, species populations, species traits, community composition, ecosystem structure, and ecosystem function) Pereira et al. (2013). Although EBVs were selected through a rigorous evaluation process of dozens of options considering criteria on scalability, temporal sensitivity, feasibility, and relevance, their practical implementation remains a challenge Kissling et al. (2018), Skidmore et al. (2015), Pettorelli et al. (2016), Brummitt et al. (2017), Turak et al. (2017), Kissling et al. (2018).

Species traits include any measurable morphological, phenological, physiological, reproductive, or behavioral characteristics of individual organisms; nevertheless, they can also be generalized at the taxa and population levels. Recently, increasing efforts to integrate species traits have resulted in a significant amount of data available Kissling et al. (2018), Schneider et al. (2019); however, most of these data are associated with taxa rather than with specimens. Aggregating species traits at the taxa level causes critical data for monitoring changes in individual organisms or populations in a particular geographic area (e.g., time and location) to be lost, Schneider et al. 2019.

Species traits have been suggested as indicator variables for monitoring the response of organisms to changes in the environment; for instance, phenological trait information related to changes in the timing of plant leafing, flowering, and fruiting can be used as an indicator of climate change impacts Kissling et al. 2018, Geijzendorffer et al. (2015), Kissling et al. (2018). Different authors suggest frameworks and ideas to feed the Phenology EBVs from specimen data Kissling et al. (2018), Pereira et al. (2016). Additionally, there are focused efforts to measure trends in particular species: for example,

the UK Spring Index that tracks the impact of temperature change on the annual mean observation date of four biological events. These events include the first flowering of hawthorn (*Crataegus monogyna*), the first flowering of horse chestnut (*Aesculus hippocastanum*), the first recorded flight of an orange-tip butterfly (*Anthocharis cardamines*), and the first sighting of a swallow (*Hirundo rustica*) Parliamentary Office of Science and Technology, UK Parliament (2021).

On the other hand, the transformation of texts from taxonomic literature into structured data remains a key challenge in Biodiversity Informatics Hobern and Miller (2019), Miralles et al. (2020). NLP tools and algorithms have been successfully applied in information extraction tasks in biodiversity texts; for example, to extract taxonomic names using rules based on syntax, fuzzy logic, and dictionaries Gerner et al. (2010), Leary (2014), Wei et al. (2010), Sautter et al. (2006), and, in some cases, probabilistic models Akella et al. (2012); to structure complete texts using rules, regular expressions, dictionaries, and heuristics based on text style Sautter et al. (2012), Cunningham et al. (2011), Curry and Connor 2016 ; and to extract species morphological characteristics using rules, dictionaries, and ontologies Mora and Araya (2018), Duan et al. (2013), Cui (2012), Cui (2013), Balhoff et al. (2014).

Additionally, some ML algorithms, such as NER and Classification have been successfully applied to bioinformatics and biomedicine, and, more recently, to BI. Text Classification and Named Entity Recognition (NER) are classic research topics in the NLP field. Text Classification is a fundamental technique in NLP to categorize unstructured text data into predefined labels or tags (widely used in sentiment analysis). The Allerdictor tool is an example of an application in bioinformatics that models sequences as text documents and uses Multinomial Naïve Bayes (NB) or Support Vector Machine (SVM) for allergen classification Dang and Lawrence (2014). In addition, Pan et al. in Pan et al. (2018) describe a method to predict bacteriophage virion proteins in ecology using a Multinomial Naïve Bayes classification model; Delizo et al. used Multinomial Naïve-Bayes to analyze users' tweets polarity concerning the COVID-19 with excellent results Delizo et al. (2020); and Demichelis et al. proposed a hierarchical Naïve Bayes Model to manage biological heterogeneity to improve classification accuracy using a prostate cancer tissue microarray dataset Demichelis et al. (2006).

NER is the first step in many NLP tasks. It seeks to locate and classify entities' names in free text into categories. The traditional NER task has expanded beyond identifying people, location and organization to identify dates, email addresses, book titles, protein names, numbers, amongst other applications. Additionally, there has been a strong interest in using NER for extracting product attributes from online data due to the rapid growth of E-Commerce Zheng et al. (2018), for assessing people skill sets in Skill Analysis Fareria et al. (2021) and in information retrieval, to extract the main elements of a user query to better identify what the user is looking for Putra et al. (2020). In E-Commerce, NER is used to autofill in attribute specifications, to improve search and to build product graphs. Some examples in Biodiversity Informatics include: the Specialized Information Service Biodiversity Research (BIOfid), which facilitates automatic extraction of regular categories (e.g., person, location, organization) and taxon names from printed literature about plants,

birds, moths and butterflies hidden in German libraries for over the past 250 years Akella et al. (2012), Rössler (2004). The National Commission for Knowledge and Understanding of Biodiversity (CONABIO) in Mexico has trained models for NER to extract species names from text written in Spanish, Barrios et al. (2015). The "TaxonGrab" method is a web-based project that allows users to upload information and then displays the list of the candidates' taxonomic names mentioned in the text Koning et al. (2005), NetiNeti (Name Extraction from Textual Information-Name Extraction for Taxonomic Indexing) and TaxoNERD to recognize scientific names in biodiversity publications using NER Akella et al. (2012), Le Guillarme and Thuiller (2021). However, at this point, no applied research results have been published to extract phenological data from morphological descriptions of specimens using ML algorithms.

The main objective of this project was to obtain baseline data to feed the Species Traits Variable Phenology and other biodiversity indicators by extracting species characters and structure names from morphological descriptions of specimens and classifying the descriptions using machine learning (ML). To achieve this goal, an ML workflow was tested to classify specimen descriptions to determine if the plant had flowers and/or fruits at the time of collection and to extract species characters and structure names mentioned in the descriptions. A database with 106,804 records from the Herbarium of the National Biodiversity Institute of Costa Rica (INBio) was used to illustrate the proposed approach, Vargas (2016).

The remainder of the paper is structured as follows: Section "Materials and methods" provides the detailed workflow of the proposed material and methods, section "Results" presents the evaluation metrics and results, and section "Discussion" analyze the results. Finally, conclusions and future work are discussed in "Conclusions".

Materials and methods

This research work presents an effort to extract species morphological characters and structure names using NER algorithms and classify specimen morphological descriptions to determine if a given plant had flowers or/and fruits at the time of collection.

Successfully applying ML algorithms to NLP problems requires defining a workflow that includes phases like data selection and pre-processing, model training and test and model deployment. Fig. 1 shows the general workflow used in this research.

A. Data Selection and Processing Phase

A.1. Atta Dataset: Atta is an information system developed by INBio to manage data of specimens of different biological groups, such as plants, arthropods, fungi, and nematodes.

The database contains 350,007 records from the kingdom Plantae. Data related to taxonomy (i.e., scientific name and higher taxonomy); plant specimens (i.e., morphological description, date collected, locality, collectors, and sampling protocol, amongst other data);

and geospatial data (i.e., locality and geographic coordinates) were obtained from Atta. All the selected specimens were collected in Costa Rica.

Fig. 2 shows an example of a specimen collected by INBio; Fig. 3 presents the collection sites for specimens available at the INBio's Herbarium, which represent 354 plant families and span 124 years from 1892 to 2016, with a higher concentration of records in the period from 1990-2006; and Fig. 4 displays a histogram of records by year of collection.

A.2. Cleaning and Random Selection of Data: In this project, 106,804 records from Atta were used. Atta contains 350,007 records from the kingdom Plantae. Herbarium rules and regulations state to send duplicate specimens to the National Museum of Costa Rica and the Missouri Botanical Garden, so from this figure, 64% are duplicate records. After removing duplicate records, records without morphological description, discarded specimens, and descriptions written in English, about 93% of the remaining records (i.e., 106,804 records) were tagged (i.e., they were assigned to one of the classification target classes: has_flowers and has_fruits).

A.3. Tagging Data for Multi-label Classification: The texts used in the experiments correspond to the morphological description of 106,804 specimens. Morphological descriptions contain statements that detail morphological aspects (i.e., shape and structure) of specimens, which are useful to study and identify them. Statements may describe structures, substructures, characters, states, and relationships between structures (e.g., leaves, apex, flowers, flower buds, or fruits). The characters are, for instance, length, width, pigmentation patterns, smell, or architecture. An example of a description is the statement *"Arbolito de 7-9 m x 10 cm dap. Corteza lisa, amarillo-cafezuzco, exfoliante. Brotes vegetativos verde-tenue con pubescencia blanca, conspicua, caulifloro. Frutos inmaduros, esferoides, verde-tenue*". (Small tree 7-9 m x 10 cm DBH. Smooth, yellow-brown, exfoliating bark. Faint-green vegetative shoots with white, conspicuous, cauliflorous pubescence. Immature, spheroid, faint-green fruits).

Morphological descriptions of plant specimens use a semi-structured language characterised by Mora and Araya (2018):

- They use many abbreviations and omit functional words and verbs, making sentences become telegraph phrases to save space in scientific publications;
- Texts are written in a very technical language because the formal terminology is based on Latin;
- They contain primarily names, adjectives, numbers (measures) and adverbs to a lesser extent. Verbs are seldom used;
- The vocabulary used is repetitive;
- They are short because they are included on the specimen label and sometimes the text is shortened to fit on the label. Fig. 5 shows the distribution of the descriptions length of specimens from the INBio Herbarium;
- They use highly standardised syntax even though they are written in natural language.

Supervised machine-learning algorithms were used to classify descriptions. Training supervised models involves adjusting their parameters using examples that allow models to map an input to the desired output, in this case, the target classes. Examples were built from the specimens' morphological descriptions by manually assigning each description to one of the classes (i.e., has_flowers and has_fruits). For example, the morphological description *"Creciendo en tronco seco. Flores naranjas. Muestra conservada en alcohol"* ("Growing on the dry trunk. Orange flowers. Sample preserved in alcohol", in English) was assigned to the has_flowers class, and the description *"Arbusto de 35 m. en el sotobosque. Frutos de color verde y rojo a púrpura oscuro cuando están maduros. Escaso"* ("35 m shrub in the understorey. Green and red fruits to dark purple when ripe. Scarce", in English) was assigned to the has_fruits class. Descriptions were standardised by changing their contents to lowercase, removing special characters, and tokenising each description (i.e., breaking descriptions into words, symbols, or other elements called tokens).

Two classes were used to classify specimen morphological descriptions and determine if a plant had flowers or/and fruits at the time of collection: has_flowers and has_fruits, accordingly. The 106,804 records from INBio's database (i.e., Atta) were tagged. Fig. 6 shows the number of records with zero, one, or two classes assigned in the selected samples. Records were tagged manually using SQL statements in a PostgreSQL database. Descriptions such as *"sin flores"* (no flowers), *"sin frutos"* (no fruits), *"sin flores ni frutos"* (no flowers or fruit), amongst others, were not included in the experiments because very few descriptions presented that pattern.

A.4. Tagging Data for NER: A small part of the clean records used in the classification process was randomly selected for extracting species characters and structure names using supervised ML algorithms. Eight thousand specimen records were chosen for this purpose.

To prepare examples, different standard approaches to sequence tagging Goyal et al. (2018) were evaluated, such as IO (Inside, Outside), BIO (Begin, Inside, Outside) Ramshaw and Marcus (1999), and BIOE (Begin, Inside, Outside, End) Sang and Veenstra (1999). Due to the characteristics of the morphological descriptions mentioned above, the BIO standard was selected. BIO assigns a tag or class to each token within the text of the descriptions; it captures the named entity type, the entity boundary, and tokens outside. For example, the description *"palma solitaria de 3 m. tallos de hasta 0.50 m. inflorescencias interfoliares, botones florales crema. comun."* (solitary palm of 3 m. stems up to 0.50 m. interfoliar inflorescences, cream flower buds. common.) was tagged as *"palma solitaria de 3 m. tallos[B] de hasta 0.50 m. inflorescencias[B] interfoliares, botones[B] florales[I] crema. comun."*, where, [B] represents the beginning of an entity, [I] represents the intermediate tokens in multi-word entities (e.g., *"botones florales"*, flower buds), and O any other token including punctuation marks (not marked in the example). Very few multi-word entities were found in the specimen morphological descriptions. Fig. 7 shows the number of words in the records assigned with each label (i.e., B, I, O).

The following activities were carried out for the tagging process:

- In addition to the has_flowers and has_fruits classes, the 106,804 specimens were associated with other classes such as has_leaves and has_stems (has_root was not used because very few descriptions mentioned roots). These classes were used to randomly select two thousand records of each to balance the presence of structures belonging to all classes. In total, eight thousand records were selected, including records for classes has_flowers, has_fruits, has_leaves, and has_stems.
- FreeLing v.4.2 morphological analyzers and taggers Padró and Stanilovsky (2012) were used for tokenizing, lemmatizing and POS-tagging (part-of-speech tagging) the morphological descriptions. POS-tagging was used to semi-automatically assign a class (e.g., noun, adjective, verb, adverb, article) to each token. Most plant structures and characters correspond to nouns in sentences.
- Using the POS tags generated by FreeLing, each token was assigned a B, I, or O tag, depending on its role in the sentence.
- Two thousand records randomly selected from the eight thousand were assigned to each team member to manually review the labels (four team members).

B) Models Training and Evaluation Phase

B.1. Classification: Train Models using NB, SVC, and LR: The classification of morphological description involved 106,804 specimen records used for training and test models. The experiments were carried out using Python version 3 Python Software Foundation (2021), Scikit Learn version 0.24.2 Pedregosa (2011), and the Natural Language Toolkit (NLTK) version 3.5 Elhadad (2010).

The classification objective was to determine if each of the morphological descriptions of the specimens mentioned or not the presence of flowers or fruits, that is, to assign each description to the has_flowers and/or has_fruits classes. Each sample could be assigned to zero, one, or both classes; therefore, the classification problem corresponds to a multi-label classification task. The algorithms Multinomial Naive Bayes (NB) Klampanos (2009), Linear Support Vector Classification (SVC) Chang et al. (2008), and Logistic Regression (LR) Bishop (2006) were used for the experiments.

The input to the models was a one-dimensional vector (x1, x2, ..., xn) with the morphological descriptions. Features were extracted from this 1D vector that was converted to a matrix of values using TF-IDF (Term Frequency-Inverse Document-Frequency) or the frequency of words occurring in the descriptions with a lower and upper boundary of the range of (1,3) for different n-grams to be extracted.

To estimate the skill of the models on new data, ten-fold cross-validation was used with the function cross_val_score (Scikit Learn) in combination with the NB, SVC, and LR algorithms Pedregosa (2011). The One-vs.-Rest (OvR) strategy was applied to solve the problem of multi-label Classification. The parameters used with each of the algorithms were as follows: NB (with learn class prior probabilities equal to true and priors adjusted according to the data), SVC (with hinge loss function, tolerance equal to 1e-4, strength

regularization inversely proportional to 1.0, calculate the intercept equal to true, multi-class strategy one-vs.-res, and 1000 maximum number of iterations), and LR (with tolerance equal to 1e-4, L2 the norm of the penalty, strength regularization inversely proportional to 1.0, and L-BFGS solver for optimization Liu and Nocedal (1989).

B.2. NER: Train Models using HMM, CRFs, and BI-LSTM-CRF: Out of the 106,804 specimen records, 8,000 were randomly selected, where 80% of the records were used for training, while the remaining 20% were for testing the models. The training and testing of the models were done using Python version 3 Python Software Foundation (2021), Pomegranate 0.14.7 Schreiber (2018) for HMM, Pytorch 1.8.1 Paszke et al. (2019) for BI-LSTM-CRF, and Pycrfsuite Wijffels and Okazaki (2007) for CRFs.

The aim of applying NER tagging to the data was to extract characters and structure names from morphological descriptions (e.g., flowers, trunk, color, height) where every token of a description was assigned a B, I or O tag. With this purpose in mind, the algorithms CRFs Lafferty et al. (2001), BI-LSTM-CRF Huang et al. (2015), and HMM Baum and Petrie (1966) were used for NER tagging. The information considered relevant to train the models of CRFs and BI-LSTM-CRF was the token, its POS tag, and the label assigned; for the case of HMM, only the token and its label were considered.

In order to train the HMM model, bigram, sequence starting, and sequence ending counts were used to estimate the probability distribution and generate every state and transition that the model would use for its predictions.

The way the data were handled to train the CRFs model was to convert each token in the training data into a feature that would later be fed to the model. The characteristics considered for every word were the word itself, its last three letters, if it was a punctuation mark or if it was a digit, its POS tag, and the first two letters of the POS tag. Each feature was processed using its own characteristics combined with the next and previous words in the sentence (if applicable). Afterwards, the model was trained with the hyperparameters established in Table 1.

To train the BI-LSTM-CRF model, every word in the dataset was put into a dictionary that was later passed to the model; this had to be done with all records. The model worked with every sentence not as a string of words, but as a tensor of their respective indexes in the word dictionary. After obtaining the ready data, the model ran a forward pass with the negative log-likelihood cost function, then computed the loss and gradients, and updated the model parameters. This process was done for every sentence in the training set for every epoch. The model was trained with the hyperparameters established in Table 2.

B.3. Models Evaluation (Accuracy, Precision, Recall, and F1 score): The metrics generally used in classification and NER problems to evaluate the results are precision and recall Dandapat (2011). They measure the percentage of correct classification and the completeness of the method, respectively. In addition, the accuracy and the F1-score (the harmonic mean between precision and recall) were computed.

Results

This section presents a report of the experimental results for both classification and NER tests.

Classification of morphological descriptions of specimens. Performance of the NB, SVC and LR algorithms: Fig. 1 presents the general workflow of the project. Table 3 gives examples of morphological descriptions obtained from the Atta database. After a cleaning process that involved removing duplicate records, specimens without morphological description, discarded specimens, descriptions written in English, and records with phrases such as *"sin flores"* (no flowers), *"sin futos"* (no fruits), *"sin flores ni frutos"* (no flowers or no fruit), amongst other issues in the data, 106,804 records were tagged for the experiments. Table 4 presents the amount of specimen morphological descriptions distributed by class, the average length in characters of the descriptions, and the standard deviation. The objective of the experiment was to train models that could automatically associate the non-exclusive classes has_flowers and has_fruits to the morphological descriptions. As each description can be assigned to more than one class, the One-vs.-Rest (OvR) strategy was used with three traditional ML algorithms: NB, SVC, and LR.

Models' skills were estimated using ten-fold cross-validation to prevent overfitting and reduce bias. After executing the ten training sequences and tests of different models, metrics such as accuracy, precision, recall, and F1 score by algorithm and class were computed, and the average of the results was calculated. Table 5 presents the results of the metrics used to estimate the model's skills.

To measure the impact of different collector's writing on the result, in a second experiment, training and test data were partitioned using the number of specimens gathered per collector. The test was carried out to verify if the resulting models were just trained to parse the writing of the prolific collectors. Specimen descriptions written by collectors with different amounts of gatherings were selected for testing models, the rest of the samples were used to train the models. Fig. 8 shows the results of applying the algorithms to text written by collectors with one collected sample up to 500 samples. In all tests the model results remained above 98% (macro-average F1-score) for the algorithms SVC and LR. Only in the case of NB, the result fell to 94% (macro-average F1-score) for collectors with less than 10 samples.

NER tagging of morphological descriptions. Performance of the CRFs, BI-LSTM-CRF and HMM algorithms: Records, such as those shown in Table 6, were used to test the models and data cleaning was similar to the one used in the classification experiments. Records that were duplicated, discarded, lacked a morphological description, or contained descriptions in English were not used in the research.

As seen in the examples, the aim was to tag the entities that appeared in the specimen's description. With this purpose in mind, CRFs, HMM, and BI-LSTM-CRF were used.

The Sklearn Pedregosa (2011) library was used to obtain the metrics to evaluate every model's performance with the test data, including the accuracy, precision, recall, and F1-score. Table 7 shows the results obtained by each model used in the NER experiment.

Discussion

A successful workflow was tested with the current project to extract phenological data from morphological descriptions of botanical specimens. Some elements of the project to highlight are:

- The results achieved in the classification experiments showed that was feasible and generalisable to other biological groups to use the specimen morphological descriptions to automatically obtain phenological data, which most of the time, is only available in text format. The SVC models surpassed NB and LR models with an average F1 score higher than 0.995 (Table 5 compares the performance of SVC with two other ML methods). For more complex texts, more robust algorithms, such as Recurrent Neural Networks - LSTM and Transformers, can be applied.
- The NER experiments results showed that the HMM and CRFs model's performance had better results than the BI-LSTM-CRF model as shown in Table 7. The most significant difference between HMM and CRFs can be observed with the [I] tag results where CRFs outperform HMM in precision and F1-score while HMM surpassed CRFs in accuracy and recall. Otherwise, the models showed very similar results in the other two tags.
- Certain words in the Spanish vocabulary had mistaken POS tags, where FreeLing would often confuse nouns with similar-sounding verbs, for example, words like *"morado / morar"* (purple / dwell) or *"sépalo / sepa-lo"* (sepal / to know). This resulted in many words being mistagged and others being separated into two different words.
- The NER models had problems differentiating when an entity was composed by the name of an entity and an adjective (i.e., “frutos[B] maduros[I] rojos” - ”red ripe fruit”) and when that same adjective was used to describe the entity (i.e., “frutos[B] maduros” - ”ripe fruits”).
- The characteristics of descriptions could have influenced that FreeLing tools were not as effective in tagging nouns that are key elements to perform NER. This result made the manual review of the tagging text more time-consuming.
- Although classes were highly unbalanced in all experiments and the description length ranges from 4 to 952 characters, the model's performance was not affected. This was mainly due to the large amount of data used during the training phase and the characteristics of the descriptions.
- The data used were collected by INBio throughout the country, over a long time and by more than 400 botanists and technicians, which gives an idea of how variable the descriptions were. Figures 4 and 5 present these data in detail.
- Most of the time, data of morphological descriptions of specimens are not shared in global networks that integrate biodiversity data, such as the Global Biodiversity

Information Facility (GBIF), which could make it easier to carry out experiments integrating multiple sources and multiple languages.

Conclusions

Phenological traits data, such as the timing of plant leafing, flowering, and fruiting, have been suggested as indicators to measure how organisms respond to disturbances and changes in environmental conditions. This document has proposed a workflow that uses ML and NLP algorithms to integrate phenological data extracted from morphological descriptions in text format with other structured data available in specimen records (such as geographic coordinates, taxonomy and collection date). The integrated data, combined with abiotic records (e.g., temperature, precipitation, and humidity), could enable users (e.g., decision-makers, researchers, biodiversity institutes) to answer questions related to the possible effects of environmental changes that occur in time and space on particular species.

As far as we know, this work is the first to apply ML algorithms to specimen morphological descriptions to extract phenological data on flowering and fruiting. Results showed that it is possible to classify specimen morphological descriptions with more than 99% success (F1-score) using a multi-label approach (with classes like has_flowers and has_fruits) and to extract the characters and structure names from descriptions with more than 98% success (F1-score) using NER.

Although models, like the one proposed in this project, achieve excellent results, it is crucial to consider that, even though there are records of the planet's biodiversity that have been systematically collected over hundreds of years, the available data are strongly unbalanced regarding taxa, locality, time, and the number of individuals.

The results of this project can be used to generate baseline data to feed the Phenology EBV from morphological descriptions of specimens written in any language, amongst other applications. Although data about the event duration as proposed by the USA-National Ecological Observatory Network (NEON) Jones et al. (2014) cannot be obtained from specimens, different authors present frameworks and ideas to feed the Phenology EBVs from specimen data Kissling et al. (2018), Pereira et al. (2016).

The proposed workflow can be applied to the morphological descriptions of specimens of different biological groups,, and there are no restrictions on the language used. For biodiversity networks that integrate data from multiple sources using different languages, it is also vital to evaluate cross-lingual algorithms to alleviate the need to manually tag descriptions in a target language by leveraging tagged descriptions from other languages. For more complex texts, more robust algorithms, such as Recurrent Neural Networks - LSTM and Transformers, can be applied.

Data and Code

Data from the National Biodiversity Institute of Costa Rica is used in this paper. The full dataset and documentation can be downloaded from https://www.gbif.org/dataset/3717f916-d983-4a81-bb13-5f91200871a6. Code for data cleaning and analysis is provided as part of the replication package. It is available at https://github.com/colibri-itcr.

References

- Akella LM, Norton CN, Miller H (2012) NetiNeti: discovery of scientific names from text using machine learning methods. BMC Bioinformatics 13 (1). https://doi.org/10.1186/1471-2105-13-211
- Balhoff JP, Dahdul WM, Dececchi T, Lapp H, Mabee PM, Vision TJ (2014) Annotation of phenotypic diversity: decoupling data curation and ontology curation using Phenex. Journal of Biomedical Semantics 5 (1). https://doi.org/10.1186/2041-1480-5-45
- Barrios J, Molina A, Sierra-Alcocer R, Enrique D, Zenteno J (2015) A Text Mining Library for Biodiversity Literature in Spanish. International Journal of Computational Linguistics and Applications 6 (2).
- Baum L, Petrie T (1966) Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics 37 (6): 1554-1563. https://doi.org/10.1214/aoms/1177699147
- Bishop C (2006) Pattern Recognition and Machine Learning. . Information Science and Statistics. Springer
- Brummitt N, Regan E, Weatherdon L, Martin C, Geijzendorffer I, Rocchini D, Gavish Y, Haase P, Marsh C, Schmeller D (2017) Taking stock of nature: Essential biodiversity variables explained. Biological Conservation 213: 252-255. https://doi.org/10.1016/j.biocon.2016.09.006
- Chang K, Hsieh C, Wang X, Lin C (2008) LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9.
- Convention on Biological Diversity (CBD) (2011) Strategic Plan for Biodiversity 2011-2020, Including Aichi Biodiversity Targets. https://www.cbd.int/sp/. Accessed on: 2021-9-01.
- Cui H (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of the American Society for Information Science and Technology 63 (4): 738-754. https://doi.org/10.1002/asi.22618
- Cui H (2013) MARTT: Automatic Markup of Taxonomic Descriptions with XML. Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI https://doi.org/10.29173/cais277
- Cunningham H, Maynard D, Bontcheva K, Tablan V (2011) GATE: an architecture for development of robust HLT applications. Associate Computer Linguistics.
- Curry G, Connor R (2016) Automated Extraction of Biodiversity Data from Taxonomic Descriptions. Biodiversity Databases63-81. https://doi.org/10.1201/9781439832547-6

- Dandapat S (2011) Nitin Indurkhya and Fred J. Damerau (eds): Handbook of Natural Language Processing (second edition). Machine Translation 25 (4): 377-381. https://doi.org/10.1007/s10590-011-9117-6
- Dang H, Lawrence C (2014) Allerdictor: fast allergen prediction using text classification techniques. Bioinformatics 30 (8): 1120-1128. https://doi.org/10.1093/bioinformatics/btu004
- Delizo J, Abisado M, De Los Trinos M (2020) Philippine Twitter Sentiments during Covid-19 Pandemic using Multinomial Naïve-Bayes. International Journal of Advanced Trends in Computer Science and Engineering 9 (1.3): 408-412. https://doi.org/10.30534/ijatcse/2020/6491.32020
- Demichelis F, Magni P, Piergiorgi P, Rubin MA, Bellazzi R (2006) A hierarchical Naïve Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays. BMC Bioinformatics 7 (1). https://doi.org/10.1186/1471-2105-7-514
- Duan H, Hei Y, Cui Z (2013) Heuristics based semantics annotation of biodiversity documents in Chinese. Chinese Library Information Science
- Elhadad M (2010) Natural Language Processing with Python Steven Bird, Ewan Klein, and Edward Loper (University of Melbourne, University of Edinburgh, and BBN Technologies) Sebastopol, CA: O'Reilly Media, 2009, xx+482 pp; paperbound, ISBN 978-0-596-51649-9, $44.99; on-line free of charge at nltk.org/book. Computational Linguistics 36 (4): 767-771. https://doi.org/10.1162/coli_r_00022
- Fareria S, Melluso N, Chiarello F, Fantoni G (2021) SkillNER: Mining and mapping soft skills from any text. Expert Systems with Applications Volume 184 (1). URL: https://arxiv.org/pdf/2101.11431.pdf
- Geijzendorffer I, Regan E, Pereira H, Brotons L, Brummitt N, Gavish Y, Haase P, Martin C, Mihoub J, Secades C, Schmeller D, Stoll S, Wetzel F, Walters M (2015) Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. Journal of Applied Ecology 53 (5): 1341-1350. https://doi.org/10.1111/1365-2664.12417
- Gerner M, Nenadic G, Bergman CM (2010) LINNAEUS: A species name identification system for biomedical literature. BMC Bioinformatics 11 (1). https://doi.org/10.1186/1471-2105-11-85
- Goyal A, Gupta V, Kumar M (2018) Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review 29: 21-43. https://doi.org/10.1016/j.cosrev.2018.06.001
- Hardisty A, Michener W, Agosti D, Alonso García E, Bastin L, Belbin L, Bowser A, Buttigieg PL, Canhos DL, Egloff W, De Giovanni R, Figueira R, Groom Q, Guralnick R, Hobern D, Hugo W, Koureas D, Ji L, Los W, Manuel J, Manset D, Poelen J, Saarenmaa H, Schigel D, Uhlir P, Kissling WD (2019) The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecological Informatics 49: 22-31. https://doi.org/10.1016/j.ecoinf.2018.11.003
- Hobern D, Miller J (2019) An alliance for biodiversity knowledge: Rethinking international collaboration in biodiversity informatics. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37324
- Huang Z, Xu W, Yu K (2015) Bidirectional LSTM-CRF Models for Sequence Tagging . arXiv. Cornell University URL: https://arxiv.org/abs/1508.01991v1

- Jones K, Elmendorf S, Enquist C, Rosemartin A, Thorpe A, Weltzin J, Brown J, Powers L, Wee B (2014) Using Essential Biodiversity Variables (EBVs) as a framework for coordination across research and monitoring networks: A case study with phenology. . In: National Ecological Observatory Network (Ed.) Conference: 99th ESA Annual Convention 2014, 2014. 99th ESA Annual Convention.
- Kissling WD, Walls R, Bowser A, Jones M, Kattge J, Agosti D, Amengual J, Basset A, van Bodegom P, Cornelissen JC, Denny E, Deudero S, Egloff W, Elmendorf S, Alonso García E, Jones K, Jones O, Lavorel S, Lear D, Navarro L, Pawar S, Pirzl R, Rüger N, Sal S, Salguero-Gómez R, Schigel D, Schulz K, Skidmore A, Guralnick R (2018) Towards global data products of Essential Biodiversity Variables on species traits. Nature Ecology & Evolution 2 (10): 1531-1540. https://doi.org/10.1038/s41559-018-0667-3
- Klampanos I (2009) Manning Christopher, Prabhakar Raghavan, Hinrich Schütze: Introduction to information retrieval. Information Retrieval 12 (5): 609-612. https://doi.org/10.1007/s10791-009-9096-x
- Koning D, Sarkar IN, Moritz T (2005) TaxonGrab: Extracting Taxonomic Names From Text. Biodiversity Informatics 2 https://doi.org/10.17161/bi.v2i0.17
- Lafferty J, McCallum A, Pereira FN (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Scholarly Commons, University of Pennsylvania.
- Leary P (2014) TaxonFinder. Available: http://taxonfinder.org/about. Accessed on: 2020-10-12.
- Le Guillarme N, Thuiller W (2021) TaxoNERD: deep neural models for the recognition of taxonomic entities in theecological and evolutionary literature. bioRxiv https://doi.org/10.1101/2021.06.08.444426
- Liu D, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Mathematical Programming 45: 503-528. https://doi.org/10.1007/bf01589116
- Miralles A, Bruy T, Wolcott K, Scherz MD, Begerow D, Beszteri B, Bonkowski M, Felden J, Gemeinholzer B, Glaw F, Glöckner FO, Hawlitschek O, Kostadinov I, Nattkemper TW, Printzen C, Renz J, Rybalka N, Stadler M, Weibulat T, Wilke T, Renner SS, Vences M (2020) Repositories for Taxonomic Data: Where We Are and What is Missing. Systematic Biology 69 (6): 1231-1253. https://doi.org/10.1093/sysbio/syaa026
- Morales J (2005) Studies in the neotropical apocynaceae XIX: The family apocynaceae (Rauvolfioideae, Apocynoideae) in Costa Rica. Darwiniana 43: 90-191.
- Mora M, Araya J (2018) Semi-automatic Extraction of Plants Morphological Characters from Taxonomic Descriptions Written in Spanish. Biodiversity Data Journal 6 https://doi.org/10.3897/bdj.6.e21282
- National Commission for Biodiversity Management (CONAGEBIO), Ministry of Environment and Energy (MINAE) Costa Rica. (2018) BiodataCR. National Biodiversity Infrastructure for the Management of Knowledge and Information on Biodiversity. https://www.chmcostarica.go.cr/node/256. Accessed on: 2021-9-10.
- Padró L, Stanilovsky E (2012) FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA, Istanbul, Turkey.

- Pan Y, Gao H, Lin H, Liu Z, Tang L, Li S (2018) Identification of Bacteriophage Virion Proteins Using Multinomial Naïve Bayes with g-Gap Feature Tree. International Journal of Molecular Sciences 19 (6). https://doi.org/10.3390/ijms19061779
- Parliamentary Office of Science and Technology, UK Parliament (2021) Biodiversity indicators. UK Parliament
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32.8024-8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
- Pedregosa (2011) Machine Learning in Python. JMLR 12: 2825-2830.
- Pereira H, Belnap J, Böhm M, Brummitt N, Garcia-Moreno J, Gregory R, Martin L, Peng C, Proença V, Schmeller D, van Swaay C (2016) Monitoring Essential Biodiversity Variables at the Species Level. The GEO Handbook on Biodiversity Observation Networks79-105. https://doi.org/10.1007/978-3-319-27288-7_4
- Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Höft R, Hurtt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M, Wegmann M, et al. (2013) Essential Biodiversity Variables. Science 339 (6117): 277-278. https://doi.org/10.1126/science.1229931
- Pettorelli N, Wegmann M, Skidmore A, Mücher S, Dawson T, Fernandez M, Lucas R, Schaepman M, Wang T, O'Connor B, Jongman RG, Kempeneers P, Sonnenschein R, Leidner A, Böhm M, He K, Nagendra H, Dubois G, Fatoyinbo T, Hansen M, Paganini M, de Klerk H, Asner G, Kerr J, Estes A, Schmeller D, Heiden U, Rocchini D, Pereira H, Turak E, Fernandez N, Lausch A, Cho M, Alcaraz-Segura D, McGeoch M, Turner W, Mueller A, St-Louis V, Penner J, Vihervaara P, Belward A, Reyers B, Geller G (2016) Framing the concept of satellite remote sensing essential biodiversity variables: challenges and future directions. Remote Sensing in Ecology and Conservation 2 (3): 122-131. https://doi.org/10.1002/rse2.15
- Putra HS, Priatmadji FS, Mahendra R (2020) Semi-supervised Named-Entity Recognition for Product Attribute Extraction in Book Domain. Digital Libraries at Times of Massive Societal Transition43-51. https://doi.org/10.1007/978-3-030-64452-9_4
- Python Software Foundation (2021) Python Language Reference, version 3. Available at . http://www.python.org. Accessed on: 2020-3-01.
- Ramshaw LA, Marcus MP (1999) Text Chunking Using Transformation-Based Learning. Text, Speech and Language Technology157-176. https://doi.org/10.1007/978-94-017-2390-9_10
- Rössler M (2004) Adapting an NER-system for German to the biomedical domain. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications - JNLPBA '04 https://doi.org/10.3115/1567594.1567615
- Sang E, Veenstra J (1999) Representing text chunks. Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics - https://doi.org/10.3115/977035.977059

- Sautter G, Böhm K, Agosti D (2006) A combining approach to find all taxon names (FAT). Biodiversity Informatics 3 https://doi.org/10.17161/bi.v3i0.34
- Sautter G, Böhm K, Agosti D (2012) Semi-automated xml markup of biosystematic legacy literature with the Goldengate editor. Biocomputing 2007 https://doi.org/10.1142/9789812772435_0037
- Schneider F, Fichtmueller D, Gossner M, Güntsch A, Jochum M, König-Ries B, Le Provost G, Manning P, Ostrowski A, Penone C, Simons N (2019) Towards an ecological trait-data standard. Methods in Ecology and Evolution 10 (12): 2006-2019. https://doi.org/10.1111/2041-210x.13288
- Schreiber J (2018) Pomegranate: fast and flexible probabilistic modeling in python. Journal of Machine Learning Research 18 (164): 1-6.
- Secretariat of the Convention on Biological Diversity (2020) Global Biodiversity Outlook 5. URL: https://www.cbd.int/gbo5
- Skidmore A, Pettorelli N, Coops N, Geller G, Hansen M, Lucas R, Mücher C, O'Connor B, Paganini M, Pereira HM, Schaepman M, Turner W, Wang T, Wegmann M (2015) Environmental science: Agree on biodiversity metrics to track from space. Nature 523 (7561): 403-405. https://doi.org/10.1038/523403a
- Turak E, Brazill-Boast J, Cooney T, Drielsma M, DelaCruz J, Dunkerley G, Fernandez M, Ferrier S, Gill M, Jones H, Koen T, Leys J, McGeoch M, Mihoub J, Scanes P, Schmeller D, Williams K (2017) Using the essential biodiversity variables framework to measure biodiversity change at national scale. Biological Conservation 213: 264-271. https://doi.org/10.1016/j.biocon.2016.08.019
- Vargas M (2016) Plantae of Costa Rica (INBio). Occurrence dataset. Version 1.14. Instituto Nacional de Biodiversidad (INBio), Costa Rica. . URL: https://doi.org/10.15468/tgno8a
- Wei Q, Heidorn PB, Freeland C (2010) Name Matters: Taxonomic Name Recognition in Biodiversity Heritage Library. Methods.
- Wijffels J, Okazaki N (2007) crfsuite: Conditional Random Fields for Labelling Sequential Data in Natural Language Processing based on CRFsuite: a fast implementation of Conditional Random Fields (CRFs). R package version 0.1. URL: https://github.com/bnosac/crfsuite
- Zheng G, Mukherjee S, Dong XL, Li F (2018) OpenTag. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining https://doi.org/10.1145/3219819.3219839

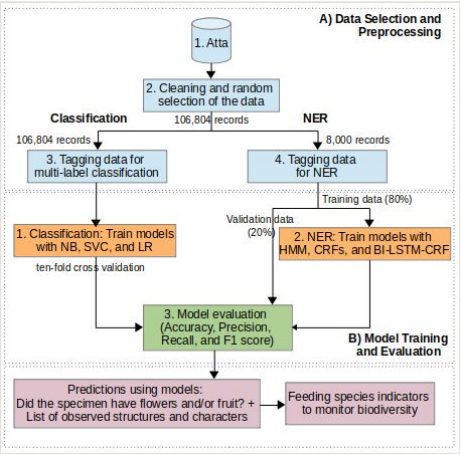

Figure 1.

The proposed general workflow includes two phases: A) Data Selection and Preprocessing using the Atta database (INBio). First, the data were cleaned by removing duplicate records, records written in English and null morphological descriptions, amongst other processes. Then, two datasets were selected for the next phase, one for Classification and one for NER. Those datasets were used for training and test activities. B) During the Models Training and Test phase, models were generated using algorithms such as: Multinomial Naive Bayes (NB), Linear Support Vector Classification (SVC) and Logistic Regression (LR) for Classification and Hidden Markov Model (HMM), Conditional Random Fields (CRF), and Bidirectional Long Short Term Memory Networks with CRF (BI-LSTM-CRF) for NER. Metrics like accuracy, precision, recall, and F1 score were used to test them.

Figure 2.

Specimen from INBio's collection shows the morphological description of a holotype of Stemmadenia abbreviata J. F. Morales, Novon 9(2): 236. 1999. TYPE. Costa Rica. Heredia: La Selva, OTS Field Station on the Río Peje, April 1982, B. Hammel 11680 (holotype, INB) Morales (2005).

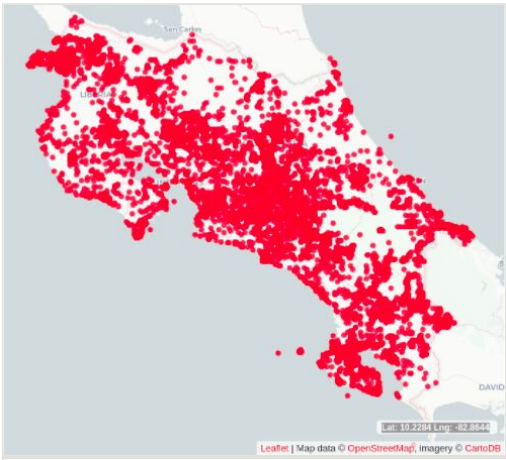

Figure 3.

Collection sites of INBio's erbarium specimens currently available at the data portal of the National Commission for Biodiversity Management (CONAGEBIO), Ministry of Environment and Energy (MINAE) Costa Rica. (2018).

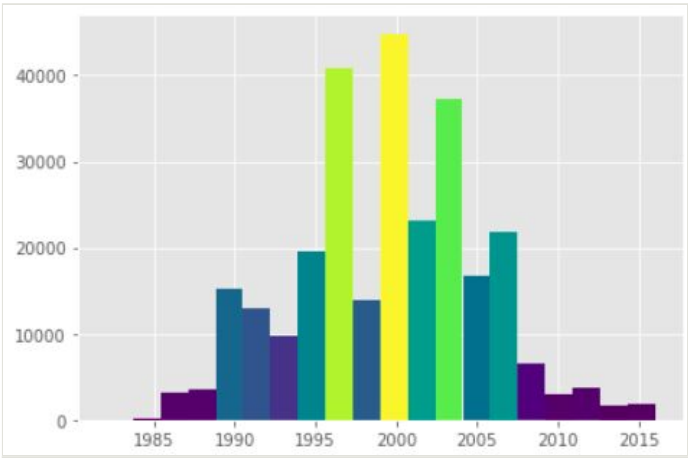

Figure 4.

Histogram of records by year of collection. Years with few records, from 1892 to 1981, were excluded in the graph (i.e., 110 specimen records were not taken into consideration).

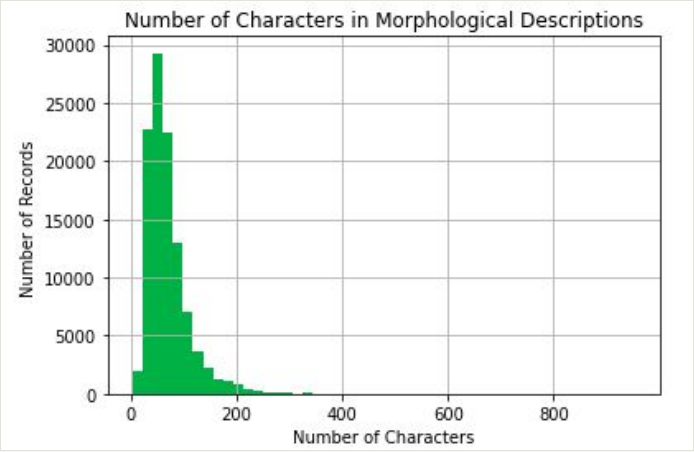

Figure 5.

Histogram of the number of characters, including blanks, in specimen morphological descriptions from the INBio Herbarium.

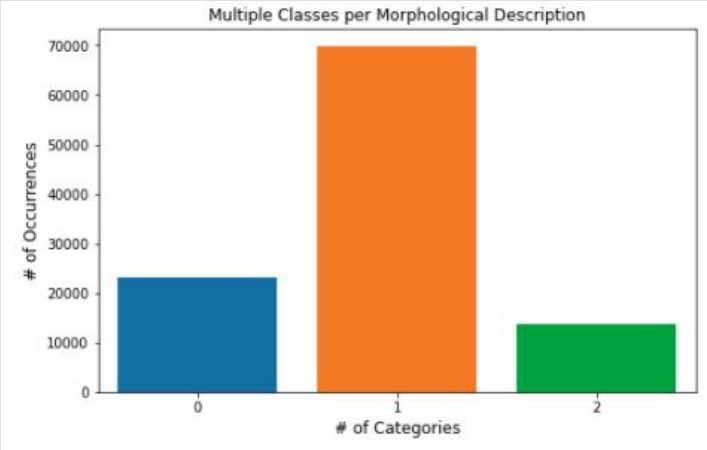

Figure 6.

The number of morphological descriptions assigned to zero, one, or two classes (i.e., has_flowers and has_fruits).

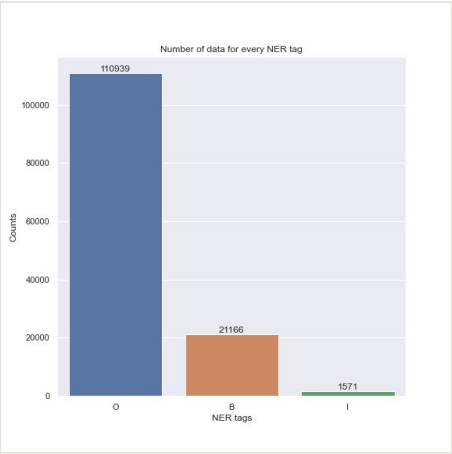

Figure 7.

Number of words in the specimen morphological descriptions with the B, I, O labels assigned in the selected samples.

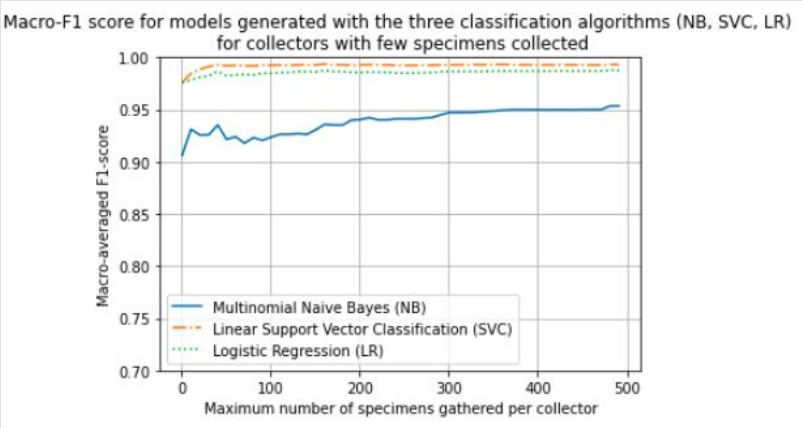

Figure 8.

Results of applying the algorithms to text written by collectors with one collected sample up to 500 samples. The test was carried out to measure the impact of different collector's writing on the result and to verify if the resulting models were just trained to parse the writing of the prolific collectors. Training and test data were partitioned using the number of specimens gathered per collector. Specimen descriptions written by collectors with different amounts of gathering were selected for testing models, the rest of the samples were used to train the models.

Table 1.

Hyperparameters used to train the CRFs model.

Hyperparameters	Values
Coefficient for L1 penalty	0.1
Coefficient for L2 penalty	0.1
Maximum Iterations	40

Table 2.

Hyperparameters used to train the BI-LSTM-CRF model.

Hyperparameters	**Values**
Hidden dimension	4
Embedding dimension	5
Learning rate	0.01
Weight decay	1e-4
Epochs	20

Table 3.

Examples of types of morphological descriptions used in these experiments.

Specimen Morphological Description	English Translation	Data
"Epífita colgante. Brácteas y cáliz morado. Corola morado y blanco, estilo y estambres verde-morado, pedicelo blanco-morado. Orillas del sendero."	Hanging epiphyte. Bracts and calyx purple. Purple and white corolla, purple-green style and stamens, purple-white pedicel. Path shores.	**Scientific name** : *Cavendishia atroviolacea* **Classes**: has_flowers = Yes has_fruits = No
"Arbusto juvenil, 1.2 m; Hojas nuevas rojizas, las viejas coriáceas. Común en barrancos al lado de la carretera. Voucher para estudio filogenético/ adn- k. sytsma."	Juvenile shrub, 1.2 m; New leaves reddish, old leathery. Common in ravines next to the road. Voucher for phylogenetic/k-dna study. sytsma.	**Scientific name:** Alzatea verticillata **Classes:** has_flowers = No has_fruits = No
"Hierba de 4-5 m. Pecíolos ca. 1.5-2.5 m, lámina foliar de 2-4 m. Inflorescencia péndulas,bracteas circinadas disticas, rojas, 1/4 basal rojo-amarillo. Flores amarillas, escondidas entre las brácteas. Frutos violeta, inmadura. Bajo dosel, escaso."	Grass of 4-5 m. Petioles ca. 1.5-2.5 m, leaf blade 2-4 m. Pendulous inflorescence, circinate distichous bracts, red, basal 1/4 red-yellow. Yellow flowers, hidden amongst the bracts. Fruits violet, immature. low canopy, scarce.	**Scientific name:** *Heliconia pogonantha* **Classes:** has_flowers = Yes has_fruits = Yes
"Arbolito de 6 m x 8 cm dap. Follaje de haz verde-intenso y envés verde-tenue. Brotes vegetativos y ramitas café-tenue. Frutos anaranjado-tenue con semillas blanco-verdoso, recubierta de arillo rojo-intenso, brillante."	Small tree of 6 m x 8 cm dbh. Foliage with an intense-green upper surface and a faint green underside. Vegetative buds and twigs light brown. Faint-orange fruits with greenish-white seeds, covered with bright, intense red arils.	**Scientific name:** *Trichilia quadrijuga* **Classes:** has_flowers = No has_fruits = Yes

Table 4.

Amount of specimen morphological descriptions distributed by class, average length in characters, and standard deviation.

has_flowers	has_fruits	Amount of records	Min-Max Length (number of characters)	Average Length	Standard Deviation of Length
No	No	23,254	4-575	59.88	40.18
No	Yes	26,900	7-952	66.15	34.64
Yes	No	42,949	11-708	69.55	38.44
Yes	Yes	13,701	27-895	93.88	51.13

Table 5.

Average precision (P), recall (R), accuracy, and F1- score (F1) computed using ten-fold cross-validation for each algorithm and class.

Algorithm	**Class**	**Accuracy**	**Precision**	**Recall**	**F1- score**
Multinomial Naive Bayes (NB)	has_flowers	0.9626	0.9462	0.9855	0.9655
	has_fruits	0.9759	0.9851	0.9510	0.9677
	Average	0.9693	0.9657	0.9682	0.9666
Logistic Regression (LR)	has_flowers	0.9888	0.9979	0.9810	0.9894
	has_fruits	0.9904	0.9998	0.9749	0.9872
	Average	0.9896	0.9989	0.9780	0.9883
Linear Support Vector Classification (SVC)	has_flowers	**0.9946**	**0.9996**	**0.9903**	**0.9949**
	has_fruits	**0.9958**	**0.9999**	**0.9891**	**0.9944**
	Average	**0.9952**	**0.9997**	**0.9897**	**0.9947**

Table 6.

Examples of types of morphological descriptions used in NER experiments.

Specimen Morphological Description	English Translation	Tagged Data
"Epifita. Flores con corola rojo rosado de bordes blancos, tubo floral externo rojo rosado con pubescencia blanca, filamentos blancos, anteras y caliz verde tenue."	Epiphyte. Flowers with corolla pink red with white borders, external floral tube pink red with white pubescence, white filaments, dim green anthers and corolla.	Epifita. Flores**[B]** con corola**[B]** rojo rosado de bordes blancos, tubo**[B]** floral **[I]** externo**[I]** rojo rosado con pubescencia**[B]** blanca, filamentos**[B]** blancos, anteras**[B]** y caliz**[B]** verde tenue.
"Liana trepadora, colgante. Brotes vegetativos cafe-rojizo. Caliz verde, corola blanca. Frutos inmaduros verdes, maduros rosado brillante."	Hanging climbing liana. Vegetative buds reddish-brown. Green calyx, white corolla. Immature fruits green, mature bright pink.	Liana trepadora, colgante. Brotes**[B]** vegetativos**[I]** cafe-rojizo. Caliz**[B]** verde, corola**[B]** blanca. Frutos**[B]** inmaduros**[I]** verdes, maduros rosado brillante.
"Arbol 15 m x 25 m dap; nervios secundarios casi invisibles; vena principal hundida en el haz; hojas muy suaves; el peciolo carece de savia lechosa. Nombre comun: ninguno."	Tree 15 m x 25 m dbh; secondary nerves almost invisible; main vein sunken in the adaxis; very smooth leaves; the petiole lacks milky sap. Common name: none.	Arbol 15 m x 25 m dap**[B]**; nervios**[B]** secundarios**[I]** casi invisibles; vena**[B]** principal**[I]** hundida en el haz**[B]**; hojas **[B]** muy suaves; el peciolo**[B]** carece de savia**[B]** lechosa. Nombre comun: ninguno.
"Arbol de 13 m x 25 cm dap. Flores blancas con un exquisito olor a dulce de caramelo. Floracion abundante. Tronco derecho, corteza escamosa pardo clara. Hojas lustrosas en ambas caras."	Tree of 13 m x 25 cm dbh. White flowers with an exquisite smell of sweet caramel. Abundant flowering. Straight trunk, light brown scaly bark. Glossy leaves on both sides.	Arbol de 13 m x 25 cm dap**[B]**. Flores**[B]** blancas con un exquisito olor**[B]** a dulce de caramelo. Floracion**[B]** abundante. Tronco**[B]** derecho, corteza**[B]** escamosa pardo clara. Hojas**[B]** lustrosas en ambas caras**[B]**.

Table 7.

Average precision (P), recall (R), accuracy, and F1- score (F1).

Algorithm	Class	Accuracy	Precision	Recall	F1-score
Conditional Random Fields (CRFs)	B	0.9739	**0.9799**	0.9739	0.9769
	I	0.8908	**0.9480**	0.8908	**0.9185**
	O	**0.9953**	0.9933	**0.9954**	0.9943
	Average	0.9533	**0.9737**	0.9534	**0.9633**
	Weighted Average	0.9905	0.9906	0.9906	0.9906
BIi-LSTM Conditional Random Field (BI-LSTM-CRF)	B	0.9781	0.9573	0.9782	0.9676
	I	0.8821	0.8037	0.8822	0.8411
	O	0.9887	0.9944	0.9887	0.9916
	Average	0.9494	0.9495	0.9536	0.9515
	Weighted Average	0.9856	0.9880	0.9880	0.9880
Hidden Markov Model (HMM)	B	**0.9823**	0.9776	**0.9824**	**0.9800**
	I	**0.9712**	0.8346	**0.9713**	0.8977
	O	0.9927	**0.9962**	0.9927	**0.9945**
	Average	**0.9820**	0.9361	**0.9821**	0.9574
	Weighted Average	**0.9908**	**0.9912**	**0.9908**	**0.9909**