A Transformer-based Nlp Pipeline for Enhanced Extraction of Botanical Information Using Camembert on French Literature

Ayoub Nainia[1], Régine Vignes-Lebbe[1], Eric Chenin[2], Maya Sahraoui[13], Hajar Mousannif[4], and Jihad Zahir[24]

[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Sorbonne Université, Muséum national d'Histoire naturelle, CNRS, EPHE-PSL, Université des Antilles, F-75005, Paris, France
[2] UMMISCO, IRD France Nord, Bondy
[3] Institut des Systèmes Intelligents et de Robotique (ISIR), Sorbonne Université, CNRS, F-75005 Paris, France
[4] LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco

Abstract. This research investigates the untapped wealth of centuries-old French botanical literature, particularly focused on floras, which are comprehensive guides detailing plant species in specific regions. Despite their significance, this literature remains largely unexplored in the context of AI integration. Our objective is to bridge this gap by constructing a specialized botanical French dataset sourced from the flora of New Caledonia. We propose a transformer-based Named Entity Recognition pipeline, leveraging distant supervision and CamemBERT, for the automated extraction and structuring of botanical information. The results demonstrate exceptional performance: for species names extraction, the NER model achieves precision (0.94), recall (0.98), and F1-score (0.96), while for fine-grained extraction of botanical morphological terms, the CamemBERT-based NER model attains precision (0.93), recall (0.96), and F1-score (0.94). This work contributes to the exploration of valuable botanical literature by underscoring the capability of AI models to automate information extraction from complex and diverse texts.

Keywords: Information Extraction, Natural Language Processing, Named Entity Recognition, Biodiversity Literature.

1 Introduction

In this paper, we explore the rich botanical diversity documented in French-written flora: *Flora of New Caledonia*[5]. France, characterized by its diverse geographical features, hosts a range of biogeographic zones including Atlantic, Continental, Mediterranean, and Alpine, making it a focal point in European botanical diversity, which reflect the extensive ecological variation across France's metropolitan and overseas territories. This is confirmed by the presence of about 6,000 species[6] of higher

[5] Flora of New Caledonia `https://bibliotheques.mnhn.fr`
[6] Biodiversity in France `https://inpn.mnhn.fr/informations/biodiversite/france`

David C. Wyld et al. (Eds): DMML, CSITEC, NLPI, BDBS - 2024
pp. 59-78, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140605

plants, a figure that underscores the country's significant contribution to Europe's floristic richness. The incorporation of data from French overseas regions further accentuates the global relevance of this botanical diversity.

Given this context, the automated extraction of botanical information from French texts is not only methodologically sound but also crucial for advancing botanical knowledge and conservation. This work aims to leverage these rich textual resources to offer new insights into plant species distribution and characteristics, contributing to the broader understanding of global biodiversity.

Furthermore, the extraction of morphological data from botanical floras holds significant implications for the advancement of knowledge bases on platforms dedicated to descriptive data management and interactive identification like Xper3[7] [1]. However, the current manual extraction process is both time-consuming and demands botanical expertise. Consequently, our proposed automated information extraction pipeline stands as a potential checkpoint in elevating the construction of knowledge bases on platforms such as Xper3.

This paper uniquely converges on the fundamental domains of botany: taxonomy and morphology. Our contribution to taxonomy involves the automatic extraction of plant species names, rooted in botanical nomenclature. This process relies on identifying and classifying segments of information referenced in text into predefined named entities [2] (known as Named Entity Recognition), facilitating the automatic labeling of extracted entities through distant supervision.

Traditional methods for acquiring this crucial information are acknowledged for their challenges in terms of time, cost, and complexity [3]. Acknowledging this, recent endeavors explore resources for trait information embedded in texts found in books, and online sources.

One notable contribution to this effort is found in the work proposed by *(Victor et.al (2023))* [4], introducing a natural language processing (NLP) pipeline designed to automatically extract trait information from unstructured textual descriptions of plant species. In the proposed pipeline, textual classification models are utilized for categorical attributes, and question answering models for numerical attributes, specifically targeting five categorical traits (growth form, life cycle, epiphytism, climbing habit, and life form) and three numerical traits (plant height, leaf length, and leaf width).

Another notable project [5] in the space of *Open Information Extraction (OIE)* [6] explores the extraction of information from plant morphological descriptions written in Spanish. This initiative utilizes information from the National Biodiversity Institute of Costa Rica (INBio), a pre-trained language models (PLM) and a language model specifically trained on Spanish plant morphological data.

However, as we survey the landscape of existing research, a conspicuous gap emerges: none have specifically addressed the morphological text of plants in the

[7] Xper3 `https://xper3.fr/en/`

French language. Notably, the focus of prior research has been primarily on plant morphological descriptions documented in Spanish and/or English. This linguistic specificity raises concerns regarding the broader applicability and transferability of models, hinting at potential challenges necessitating significant retraining or adaptation for other linguistic contexts. Importantly, the absence of a sequential extraction method for the organs associated with categorical traits in previous studies is a notable gap. Additionally, the oversight in addressing morphological relation extraction underscores the need for a more comprehensive exploration in our forthcoming research.

Here, we present our distinctive contributions:

1. Firstly, we construct comprehensive morphological datasets for Named Entity Recognition of species names and morphological plant traits. We achieved this by leveraging the abundant yet untapped information present in the extensive published and online French literature, including the Flora of New Caledonia, glossary of morphological terms, and Wikipedia.
2. Furthermore, we develop a specialized named entity recognition model dedicated to extracting species names.
3. In addition, we introduce a second model for coarse-grained Named Entity Recognition of plant species' morphological terms. These terms are then categorized into two predefined named entity types: ORGAN and DESCRIPTOR.
4. Finally, we build upon this foundation of contributions by introducing a third model for fine-grained Named Entity Recognition. We explore specific descriptors associated with plant species' organs, extending beyond the predefined named entity types for fine-grained extraction. These descriptors include Descriptor, Form, Color, Development, Structure, Surface, Position, and Disposition.

2 Definition of Concepts

In this section, we present key concepts in botany. The definitions provided here serve as a primer for the specialized language employed in this research:

- ***Botany***: Botany is the branch of biology that focuses on the scientific study of plants, encompassing various aspects of plant life.
- ***Flora***: Flora denotes the entirety of plant species present within a specific geographical area or timeframe. It encompasses all plant species, their distribution, and interactions within an ecosystem.
- ***Organism***: An organism is a living individual capable of growth, reproduction, and response to its environment.
- ***Taxonomy***: Taxonomy is the field within biology that deals with organizing, naming, and distinguishing different organisms..

- ***Species***: A species constitutes the foundational element of biological classification. In botanical terms, a species is characterized by shared morphological, genetic, and ecological traits.
- ***Plant species name***: A plant species name, scientifically known as a binomial or botanical name, is a formalized two-part naming system used in botanical nomenclature. It consists of the capitalized genus name followed by the lowercase species epithet, creating a distinct label for a specific plant species.
- ***Identification keys***: Identification keys *(Figure 1 encompasses identification key section titled "clés des espèces" for the genus "Leptostylis")* in botany are systematic guides with a series of paired statements and species names used to identify plants based on their distinguishing characteristics.
- ***Organ***: Within botany, a plant organ refers to a discrete operational component of a plant, including entities such as roots, stems, leaves, flowers, or fruits.
- ***Descriptor***: In botany, descriptors are specific terms used to describe characteristics of plants, aiding in identification and classification, such as leaf shape or flower color.

3 Methodology

In this paper, our focus is to build Named Entity Recognition (NER) models, utilizing naturalist data sourced from the National Museum of Natural History. The initial phase involves extracting, normalizing, cleaning, pre-processing, and annotating the text data to facilitate the subsequent training of NER models.

This section delineates our overall pipeline for both contributions involving NER, as presented in *Figure 2*, beginning with the extraction pipeline of plant species names, followed by coarse-grained and fine-grained extraction of morphological terms from descriptions of plant species. We present the methodological steps as follows:

1. Data Collection: The process initiates with the collection and extraction of textual data from the floras.
2. Text Annotation: Rigorous annotation is applied to the collected data, a crucial step in preparing the ground truth for training the NER models.
3. Model Training: Transformer-based training using CamemBERT language model [7] and spacy-transformers[8] package as a wrapper.
4. Model Evaluation: A thorough evaluation of the trained models is conducted to gauge their efficacy and performance against predefined metrics.
5. Deployment: Upon successful evaluation, the models are deployed, making them ready for application to new morphological text.

[8] spacy-transformers `https://spacy.io/universe/project/spacy-transformers`

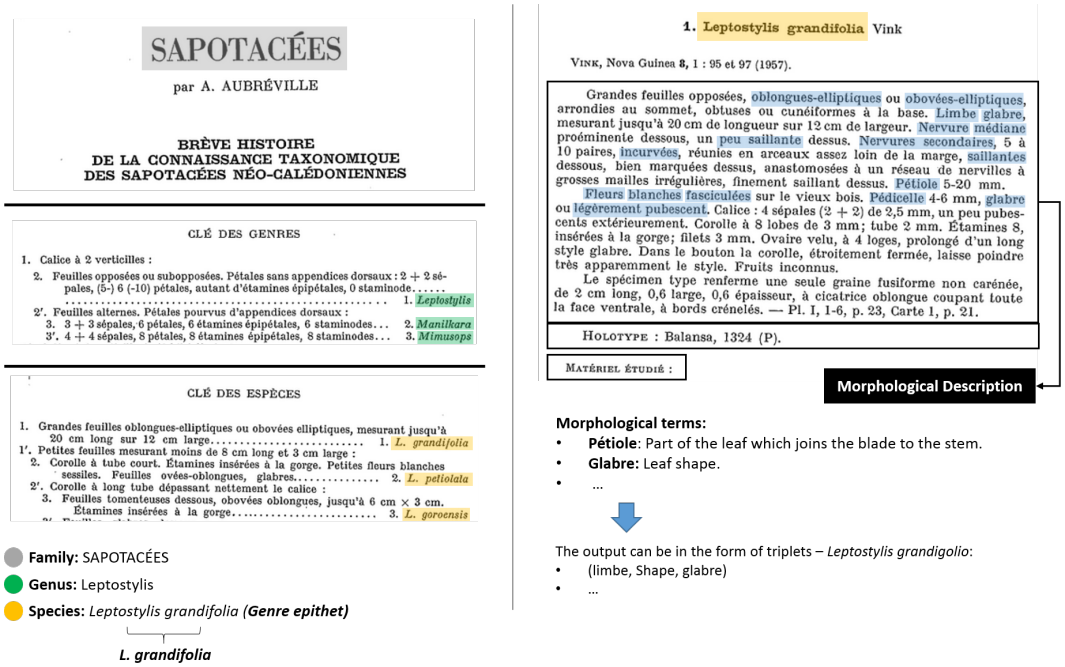


Fig. 1. Plant species names and their descriptive morphological terms from flora of New Caledonia.

3.1 Overview of Data Source

The creation of the datasets used in this research involved strategic utilization of theses primary data sources:

Table 1. Dataset Summary

Data	Flora of New Caledonia	Botanical Lexicon
Size	25 documents	3530 entries
Data Type	Ocerized Text	Text (Dataframe)
Data Source	National Museum of Natural History	Compiled from various glossaries online

– **Flora of New Caledonia**: This comprehensive botanical work comprises twenty-five floras, each averaging 201 pages *(Table 1)*. Each flora encompasses botanical families, genera, and species, with each species accompanied by a detailed morphological description. The Flora of New Caledonia offers a rich source of botanical information, covering a spectrum of plant families, genera, and species. To extract meaningful insights, a nuanced comprehension of each flora's structure is essential *(Figure 1)*.

– **Botanical Lexicon**: We utilized a botanical lexicon comprising more than 2500 terms, compiled from various glossaries accessible online[9] (Table 1). The part-of-speech tagging (POS), coupled with the definition of each term, in this lexicon, serves as a pivotal resource for the identification of morphological terms.

3.2 Extracting Plant Species Names from Textual Data

The concept of species holds paramount significance in systematics, particularly within the domain of taxonomy. Identifying species names provides a crucial context for understanding and extracting more complex information such as morphological descriptions [8]. Species names act as identifiers and help establish a clear link between the morphological information provided and the specific organism it refers to. Without knowing the species name, the morphological details lack a specific reference point. Therefore, by first extracting and establishing the species names, we create a foundation for structuring the subsequent morphological data.

In this section, we demystify our approach to identifying and classifying named entities within textual descriptions. Specifically, we aim to categorize these entities into the pre-defined type "SPECIES" (translated as "ESPECE" in french).

3.2.1 Data Understanding

Initiating the NLP pipeline for our named entity recognition model requires a profound understanding of the textual data to be extracted from flora of New Caledonia. This flora lists and describes the diverse plant species found in New Caledonia. The fundamental task is to identify and extract the names of these species for constructing the training dataset pivotal to our model.

Within the realm of botany, each species is designated by a binomial name [9], consisting of the scientific name of its genus and a descriptive-specific epithet *(as portrayed in Figure 1)*. Notably, upon scrutinizing the numerized version of the flora, we observed that species names can manifest in three distinct ways:

1. **Genus epithet**: The species designation consists of the genus name, capitalized at the beginning, followed by a lowercase epithet
2. **G. epithet**: Alternatively, the species name may present with the initial letter of the genus in uppercase, followed by the epithet in lowercase.
3. **Genus epi-thet or G. epi-thet**: In the OCR-processed version of the floras, the epithet may break at the end of a line, introducing a third form that necessitates consideration.

[9] Atlas floristique: https://atlasflore04.org/lexique.php
Pixiflore: `http://www.pixiflore.com/pages/glossaire/glossaire.html`
Herbierfrance: `http://herbierfrance.free.fr/lexique.htm`

The overarching objective is to autonomously extract species names, regardless of their form, while retaining the contextual information. This process is crucial for providing accurate annotations to train the named entity recognition model for the entity type "SPECIES."

Species names following the pattern *G. epithet* are located within the species keys sections ("clés des espèces") of each flora (Figure 1). The primary objective of this stage is to extract the text containing these species names for subsequent annotation.

Considering the potential errors introduced during the OCR process of each Flora, such as non-normalized text and inconsistent titles for the identification keys section *(e.g., "CLÉS DES ESPÈCES" similarly to Figure 1, "clés des espèce," "clés des especes," "cles des espèces," "cles des especes," etc.)*, normalization becomes imperative. The normalized title for this section is achieved by converting it to "cles des espece" in lowercase, irrespective of its initial form. The text within the section is then segmented using a line break.

The normalization and preparation of OCR-processed floras result in a list of character strings extracted from documents where the title "cles des especes" appears. For each character string, we employ a regular expression to identify occurrences of species names of the form G. epithet.

The extracted data is formatted into a dataframe, comprising the text sequence and its named entities of the type "SPECIES".

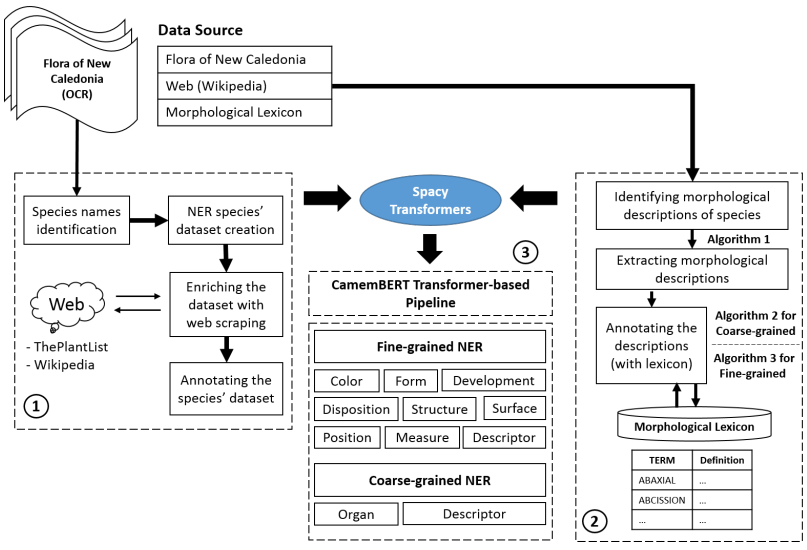


Fig. 2. NER Pipeline for Botanical Information Extraction from Naturalist Data - A Case Study with Floa of New Caledonia.

3.2.2 Species Names of the Form "Genus epithet / epi-thet"

This section is dedicated to extracting a distinctive category of species names, characterized as "Genus epithet". These names are situated following the species keys section of each genus, appearing as titles. The primary objective here is to capture the text encompassing this specific form of species name, contributing to the expansion of our training corpus.

The extraction process begins with capturing the entire title, followed by the application of a regular expression to precisely extract the species name. The results of this extraction are thoughtfully consolidated into a structured dataframe, detailing the text sequence alongside its named entities of type "SPECIES." This organized format optimizes the utility of the collected data for subsequent analytical phases.

Recognizing the importance of enhancing model generalization for more accurate classification of named entities, we intentionally introduce another data source, from the web, for species descriptions *(Figure 2 Section 1)*. Beyond the Flora of New Caledonia, this additional data source enriches the training dataset with diverse text sequences containing species names, thus guaranteeing the model's ability to generalize across varied textual structures.

3.2.3 Enriching Species Text Data with a New Data Source

Recognizing the limitations of extracting species data solely from floras—confined to titles— we introduce a supplementary data source: *Wikipedia*. While the flora-derived data was valuable for title-based species identification, it fell short of encompassing species within paragraphs. To bridge this gap, we leverage Wikipedia, a rich repository of descriptive articles on various plant species.

Our methodology *(Figure 3)* involves utilizing The Plant List[10] as a foundational resource to scrap a diverse set of species names. We strategically concatenate these names with the Wikipedia URL structure *(https://fr.wikipedia.org/wiki)*, creating unique URLs for potential articles. Subsequently, we employ a scraping mechanism to retrieve textual descriptions from these Wikipedia articles, ensuring a broad and varied dataset for robust model training.

To address the potential dataset imbalance resulting from Wikipedia's dominance, we selectively chose 50,000 species names out of over 500,000 scraped from The Plant List. This process resulted in the creation of a set of 50,000 unverified Wikipedia URLs (articles).

The practical implementation involves filtering the articles with content first, and then extracting the first paragraphs from the selected Wikipedia articles, totaling 1,051 lines of data. These lines constitute existing articles about plant species, serving as the enriching addition to our dataset.

[10] The Plant List `http://www.theplantlist.org/`

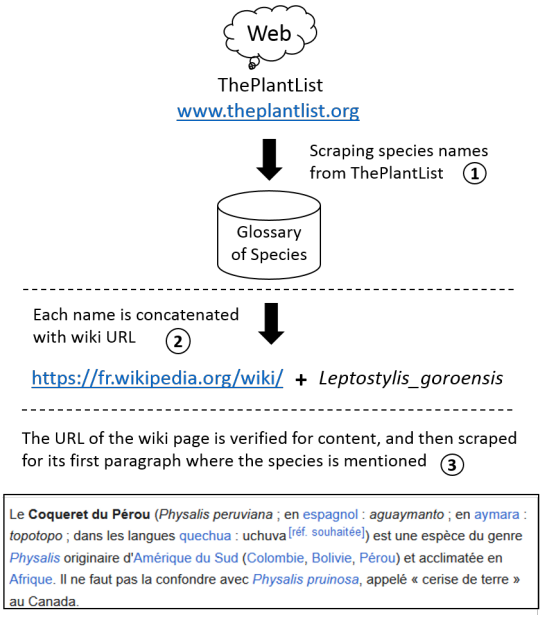


Fig. 3. The framework for enriching plant species text data by scraping ThePlantList and Wikipedia.

3.2.4 Text Annotation for Species Names Extraction

In this section, we focus on annotating the extracted text of plant species to allow the training of a Named Entity Recognition (NER) model for plant species names. Named Entity Recognition entails generating a list of tuples *(Is, Ie, t)*, where "Is" represents the start index, "Ie" represents the end index, and "t" signifies the predefined type of the named entity—in our case, "SPECIES". [10].

Our custom NER model requires a curated dataset comprising text with plant species, paired with annotation tuples for effective training. This training dataset *(Table 2)* encapsulates both the text and the corresponding named entities marked for annotation. The significance of this annotation process cannot be overstated, as the efficacy of our NER model is intricately linked to the quality and precision of the annotated dataset (Figure 2, Section 1). The process involves generating annotation tuples by aligning named entities with the texts to which they pertain, thereby extracting the start and end indexes.

Table 2. Species Dataset Summary Statistics

Total Dataset Rows	Average Text Length	Average Token Count	Average Species Count
2425	125.1357	19.6462	1.0478

3.2.5 Training Custom NER model for Plant Species Names

In the preceding section, we annotated the data to meet the formatting requirements for spaCy[11], an open-source library renowned for advanced natural language processing.

Our approach leverages spaCy's sophisticated Named Entity Recognition system that integrates a state-of-the-art word embedding strategy, incorporating subword features and Bloom embeddings. Additionally, it deploys a deep convolutional neural network employing residual connections, combined with a transition-based strategy for named entity parsing. This amalgamation forms the foundation of our custom NER model.

We initiate the training process by acquiring the dataset crafted in the preceding section, where "SPECIES" (ESPECE) serves as the sole annotated named entity. This dataset is then converted into JSON format tailored for spaCy training.

The training process unfolds by initially creating an empty (blank) model, specifying the desired language model (French). Subsequently, we add the entity recognizer to the pipeline while deactivating all other pipeline components. This approach ensures that during the training phase, only the entity recognizer undergoes refinement.

3.3 NER for coarse-grained and fine-grained morphological terms

Plant morphology is concerned with describing the external features of plants. This includes both aerial and underground organs, encompassing size, shape, stems, leaves, flowers, roots, bulbs, and tubers.

The morphological description of a species is presented in the form of semi-structured text. These are not literary texts with sentences containing verbs *(Figure 1)*, but rather a succession of assertions that review the properties of various parts of the plant, including the general appearance, leaves, veins, flowers, fruits, and seeds. Each assertion is, therefore, the combination, whether implicit or explicit, of an organ name, a property, and a value. For example, in the phrase *"White flowers"* the organ is the flower, specifically the petals of the flower, the property is the color (implicitly stated), and the value is white.

This section aims to clarify and explain the process of training a Named Entity Recognition (NER) model with the objective of extracting morphological terms, specifically of the types "ORGAN" and "DESCRIPTOR," from morphological plant species' descriptions. We build a distantly annotated species description dataset for NER by leveraging a curated list of common morphological terms and combining it with the flora of New Caledonia

[11] SpaCy `https://spacy.io/`

3.3.1 Extracting Morphological Descriptions from Flora of New Caledonia

Within the Flora of New Caledonia, morphological descriptions immediately follow the species names as titles in the format (Genus epithete) *(Figure 1)*. The beginning of a morphological description is consistently marked by the species title, with relevant botanist information interposed between the title and the description.

Algorithm 1 Extracting Morphological Descriptions

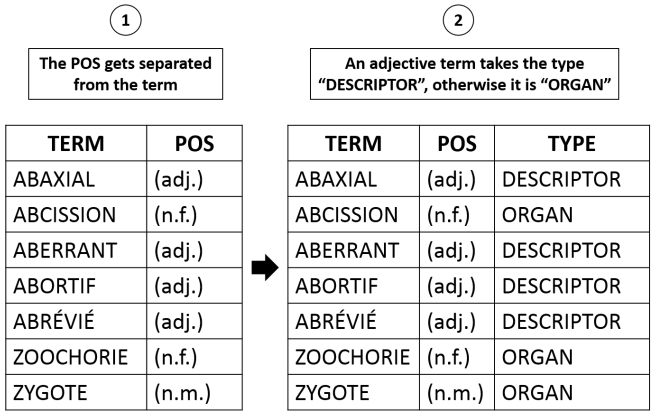
```
procedure IdentifySpeciesTitles(floraPages)
  for each page in floraPages do
    Identify species titles with RegEx
    Criteria:
        - Titles start with a two-digit number.
        - The species name is in the form (Genus epithet)
        - The botanist's name is presented as:
            * A single uppercase name ("Linnaeus")
            * Enclosed in parenthesis ("(Linnaeus) Carl")
            * Two names separated by "&."
    procedure TextRetrieval(speciesTitle)
    if speciesTitle matches regular expression then
      Retrieve all text found after the species title
    end if
    end procedure
  end for
end procedure
```

By applying *Algorithm1* for extracting morphological descriptions, organized by species titles, we create a dataset comprising 944 rows and two columns: *Title* and *Text (as morphological description)*.

To refine the extracted morphological descriptions, we eliminate preceding botanist information and irrelevant text at the beginning and end of morphological descriptions. A meticulous inspection reveals that the sequence **'/x97'** denotes the beginning of irrelevant text, and its counterpart signifies the end. The removal of these segments yields a set of refined morphological descriptions.

3.3.2 Extracting Morphological Terms from Descriptions for NER Model Training

To train our Named Entity Recognition (NER) model, which classifies morphological terms into ORGAN and DESCRIPTOR categories, our initial step involves extracting these terms from the retrieved morphological descriptions obtained from the Flora of New Caledonia. A crucial aspect of this extraction process is the utilization of a specialized lexicon for morphological terms *(Figure 2 Section 2)*.



TERM	POS
ABAXIAL	(adj.)
ABCISSION	(n.f.)
ABERRANT	(adj.)
ABORTIF	(adj.)
ABRÉVIÉ	(adj.)
ZOOCHORIE	(n.f.)
ZYGOTE	(n.m.)

TERM	POS	TYPE
ABAXIAL	(adj.)	DESCRIPTOR
ABCISSION	(n.f.)	ORGAN
ABERRANT	(adj.)	DESCRIPTOR
ABORTIF	(adj.)	DESCRIPTOR
ABRÉVIÉ	(adj.)	DESCRIPTOR
ZOOCHORIE	(n.f.)	ORGAN
ZYGOTE	(n.m.)	ORGAN

Fig. 4. Conceptualizing the morphological lexicon.

This morphological lexicon *(Figure 4)* encompasses a comprehensive set of morphological terms, each accompanied by its definition and grammatical functions. Notably, the inclusion of Part-of-Speech Tagging (PoS) allows us to discern between terms functioning as descriptors and those representing organs. Specifically, we consider the terms categorized as adjectives to be descriptors, while the remaining terms are classified as organs.

While the lexicon contains a broader range of terms beyond organs and descriptors, the morphological descriptions exclusively contain terms falling into the ORGAN and DESCRIPTOR types. To harmonize the lexicon with the extraction requirements, we add a "TYPE" column within the lexicon, enabling the categorization of terms into ORGAN and DESCRIPTOR classes.

Our extraction approach involves matching tokens from the set of morphological descriptions with the terms in the lexicon. This process ensures the identification of named entities categorized as ORGAN and DESCRIPTOR types.

Total Data rows	Average Description Length	Average Entities per Description	Average Descriptor Entities	Average Organ Entities
838	1264.31	41.43	16.41	25.02

Table 3. Morphological Terms Dataset Summary Statistics

3.3.3 Coarse-grained Data Annotation for Custom spaCy NER Model Training

To align with the requirements of our custom spaCy model, the training data must adhere to the same structured JSON format of the previous species names NER model, constituting a data dictionary inclusive of morphological descriptions and annotations for named entities. These annotations encompass the start and end indices of the identified entities within the text, along with their designated types (ORGAN and DESCRIPTOR).

In this section, we present the approach to annotating data with two distinct named entity types. The extraction of morphological terms from descriptions introduces a potential challenge—overlap among terms. Overlapping entities can share start or end indices or be nested within the index range *([Is, Ie])* of other entities, presenting a significant annotation concern.

The annotation process revolves around matching entities from the specialized botanical lexicon with those extracted from morphological descriptions. Crucially, this process is executed independently for each entity type (ORGAN and DESCRIPTOR), with a sequential resolution of overlapping entities within the same type and subsequently across all entities. Our proposed annotation approach unfolds as follows:

Algorithm 2 Annotating Morphological Descriptions for NER

```
procedure ANNOTATION(morphologicalDescriptions, namedEntities)
  - Initialize empty lists for ORGAN and DESCRIPTOR
  for each description in MorphologicalDescriptions do
    - Lowercase the morphological description.
    - Retrieve associated set of entities for the description.
    for each entity in set of entities do
      - Determine its start and end index in the description.
      if entity is the first in the description then
        - Add annotation (Is, Ie, TYPE) to the entity lists.
      else
        Annotation is only added if the named entity:
        1- Does not share its indices with other entities.
        2- Its lenght >than that of the overlapping entity.
      end if
      In case of lenght inequality, annotate the largest.
    end for
  end for
  Consolidate the ORGAN and DESCRIPTOR lists into a unified list for further processing.
end procedure
```

3.3.4 Fine-Grained Data Annotation for Custom spaCy Model Training

The extraction of fine-grained named entities within the descriptor entity type demands specialized botanical knowledge, necessitating the precise identification and classification of these entities. To address this, we present a methodical lexicon-based automated annotation approach, ensuring the accurate annotation of morphological descriptions with predefined entity types: ***surface, color, development, structure, shape, position, disposition (arrangement), and measure***.

Our lexicon of morphological terms lacks explicit distinctions for the specified entity types. However, leveraging the associated definitions with each morphological term enables a nuanced classification of the entity types referenced.

The process of assigning entity types to morphological terms unfolds as follows:

Algorithm 3 Assigning fine-grained descriptor entity types to morphological terms

for each fine-grained entity type **do**
  - Specify a set of keywords/expressions for exploration within the respective definitions of entities in the specialized lexicon.
end for
for each term in the lexicon **do**
  Search for the occurrence of keywords.
  for each keyword do
    if keyword is in the definition of the term **then**
      - Assign the corresponding entity type associated with the keyword to the morphological term.
    end if
  end for
end for

The outcome is a fine-grained lexicon, categorized by the fine-grained descriptor entity type.

3.3.5 Training NER models with spacy-transformers and CamemBERT

Spacy-Transformers[12] emerges as a pivotal component in our research methodology as it seamlessly integrates with HuggingFace's[13] transformers, affording us access to state-of-the-art transformer architectures, notably exemplified by BERT [11]. This integration represents a strategic fusion of advanced language representation capabilities with the robust information extraction functionalities and production-ready features inherent in spaCy.

[12] spacy-transformers `https://spacy.io/api/transformer`
[13] HuggingFace `https://huggingface.co/`

The choice of spaCy as our framework was particularly based on spaCy's innate extensibility by incorporating custom components and attributes, a crucial feature for fine-tuning NER models to the intricacies of morphological term recognition. Furthermore, spaCy's commendable efficiency in handling voluminous datasets, including entire web dumps, due to being written from the ground up in the memory-managed Cython [12], significantly contributes to the scalability of our models.

The coarse-grained morphological terms extraction model *(Figure 2 Section 3)* was trained following the same approach used for training the species extraction model, with a slight modification in the definition of named entities. Instead of having only one entity type, "ESPECE" (SPECIES), we introduced both entities named ORGAN and DESCRIPTOR into the pipeline.

Similarly, we trained the fine-grained model *(Figure 2 Section 3)* with the same approach with a different set of named entites: ***Organ, Descriptor, Form, Color, Development, Structure, Surface, Position, and Disposition***. The inclusion of the ***"Descriptor"*** entity type in this fine-grained extraction proves particularly valuable, ensuring comprehensive coverage of morphological traits not specified within the predefined entity types.

For a second fine-grained NER model, we fine-tuned CamemBERT [7], a language model for French, based on the RoBERTa model [13], via spaCy's standard *nlp.update* training API on the french morphological corpus we created.

4 Evaluation and Assessment

We assess the performance of the species extraction NER model by initially excluding test set data rows in which the named entity type 'SPECIES' appeared in the training set. This ensures that the model is evaluated solely on unseen data and unseen named entities.

Both the coarse-grained and fine-grained models undergo evaluation using a train-test split of 0.8 and 0.2. Unlike the previous model, named entities encountered during training are not removed from the test set. Given that each morphological description contains, on average, 41 named entities, it implies that each entity is more likely to be repeated across different descriptions. Eliminating these descriptions would potentially reduce the size of the training set to an insignificant level. Therefore, our coarse-grained and fine-grained models are assessed on unseen data (descriptions) but not necessarily on unseen named entities.

We also compare the performances of both regular spaCy and CamemBERT based pipelines for fine-grained NER models (Table 6). The CamemBERT-based model shows a slight improvement in precision (0.93) compared to SpaCy (0.92), indicating a higher percentage of correctly identified entities among those predicted. Notably, its recall (0.96) surpasses SpaCy's NER (0.94), suggesting a better ability to capture actual entities present in the data. Additionally, the F1 score for

CamemBERT (0.94) is slightly higher than SpaCy's (0.93), reflecting a balanced performance between precision and recall.

Table 4. Performance of Species Names NER Model

Species NER - Pipeline: fr_core_news_lg			
Entity Type	**P**	**R**	**F1**
SPECIES	0.94	0.98	0.96

Table 5. Performance of fine-grained and coarse-grained NER *(*SpaCy Pipeline)

Coarse-grained NER - *SpaCy*			
Entity Type	**P**	**R**	**F1**
Organ	0.97	0.987	0.98
Descriptor	0.91	0.96	0.94
Fine-grained NER - *SpaCy*			
Entity Type	**P**	**R**	**F1**
Organ	0.97	0.98	0.978
Descriptor	0.88	0.929	0.907
Surface	0.76	0.69	0.72
Color	0.77	0.73	0.75
Development	0.88	0.66	0.76
Structure	0.88	1	0.93
Form	0.82	0.88	0.85
Position	0.93	0.98	0.96
Disposition	0.95	0.93	0.94
Measure	0.909	0.829	0.867

Table 6. Comparing CamemBERT NER to SpaCy's NER

Comparing Global Metrics for Fine-grained NER			
Pipeline	**P**	**R**	**F1**
SpaCy	0.92	0.94	0.93
CamemBERT	0.93	0.96	0.94

5 Discussion

We present the performance metrics for all NER models. The species NER model (Table 4) demonstrates high precision, recall, and F1 score for identifying plant species names. These results indicate that the model performs well in extracting species names from unseen text data and unseen named entities.

While the performance metrics of the species NER model accurately reflect its capabilities, given its training on just one named entity type and evaluation on unseen data from various sources (Flora of New Caledonia and Wikipedia), assessing the performance of the coarse-grained and fine-grained NER models can be more challenging even when the performance metrics are high (Table 5).

Our lexicon-based distant supervision annotation process (Figure 4) makes the assumption that terms with a POS of *'Adj'* are Descriptors, while other terms are categorized as Organs. This assumption would be accurate if our lexicon included only organs and descriptors; however, this is not the case. Nevertheless, the accuracy of the assumption also relies on the fact that morphological terms in the extracted descriptions fall into one of two types: Descriptor or Organ, which holds true. Therefore, the accuracy of the annotation process is still maintained.

We also observe high precision, and recall (Table 5) for most named entity types, suggesting that the model accurately captures a significant portion of actual instances of the named entity in the text. Recall specifically measures the model's ability to correctly identify and include all relevant instances of a specific entity class. Although this doesn't guarantee the complete accuracy of the annotation process, it remains a positive indicator that the process does not suffer from the noise labeling problem.

While comparing the performance of both pipelines (regular spaCy NER and spaCy's CamemBERT-based NER) (Table6), we observe that although the improvement in precision, recall, and F1 score for CamemBERT is not substantial, it still indicates enhancement in named entity recognition performance. Furthermore, the improvement is consistent across all metrics (precision, recall, and F1 score), and is not limited to just one.

To assess the real-world applicability and performance of our NER model, we operationalized the NER models into a proof of concept web application *(Figures 5, 6, and 7)*. This deployment enabled engagement with domain experts, specifically botanists, who not only validated the model but also highlighted its potential effectiveness when utilized by end-users in relevant fields.

6 Conclusion

In this paper, we proposed a transformer-based Named Entity Recognition (NER) pipeline to extract botanical information from French-written flora. Our research addresses the scarcity of available datasets and the lack of information extraction

models for botanical French literature. We have detailed the proposed pipeline from data collection to the evaluation of trained models for each of the following contributions: 1) NER for plant species. 2) Coarse-grained NER to extract Organs and Descriptors from morphological descriptions. 3) Fine-grained NER for extracting the following entity types from morphological descriptions: *Organ, Descriptor, surface, color, development, structure, shape, position, disposition (arrangement), and measure.*

Through our lexicon-based distant supervision annotation approach, we achieved high recall and precision across most entity types. Specifically, for species names extraction, the NER model achieved precision (0.94), recall (0.98), and F1-score (0.96). For coarse-grained extraction, the NER model achieved precision (0.97), recall (0.98), and F1-score (0.98) for the *"Organ"* named entity type, and precision (0.91), recall (0.96), and F1-score (0.94) for the *"Descriptor"* named entity type. Furthermore, the CamemBERT-based NER model for fine-grained extraction of botanical morphological terms attained precision (0.93), recall (0.96), and F1-score (0.94).

This outcome serves as a positive indicator that the annotation process does not suffer from the noise labeling problem. Additionally, we have addressed the lack of training data by creating new NER datasets for both taxonomic and morphological use cases. Lastly, we productionized the trained models as proof-of-concept applications and gathered the impressions and judgments of domain experts regarding the accuracy of the provided output.

Fig. 5. Proof of concept for fine-grained NER of morphological terms.

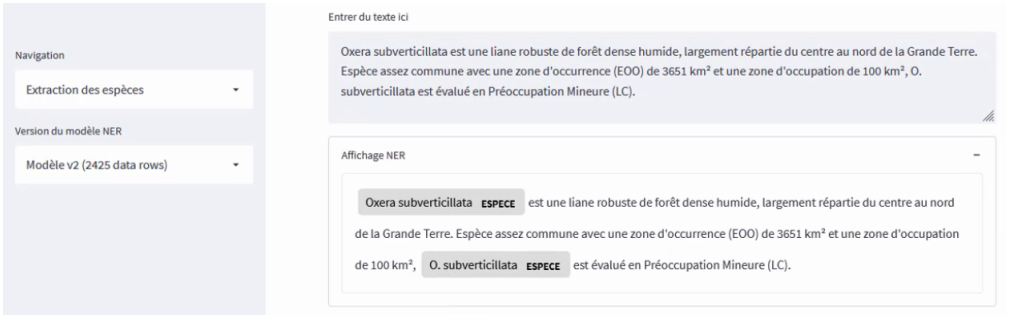
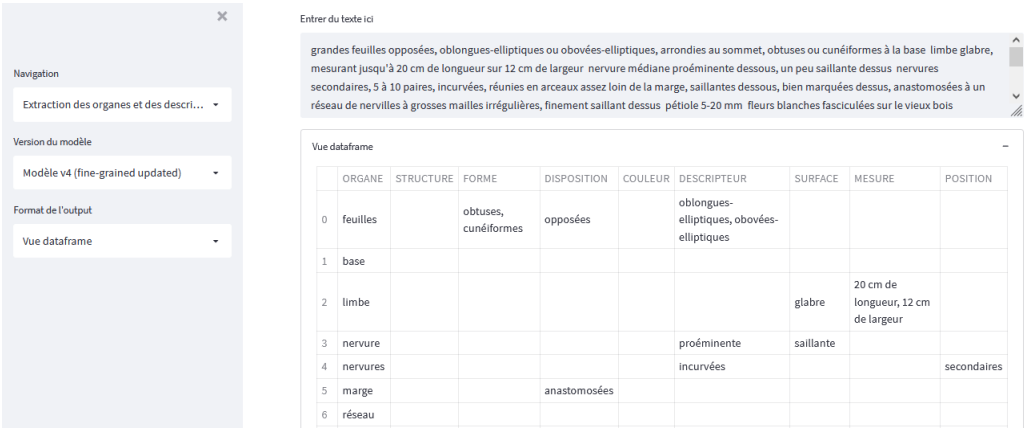


Fig. 6. Proof of concept for plants species names NER.



	ORGANE	STRUCTURE	FORME	DISPOSITION	COULEUR	DESCRIPTEUR	SURFACE	MESURE	POSITION
0	feuilles		obtuses, cunéiformes	opposées		oblongues-elliptiques, obovées-elliptiques			
1	base								
2	limbe						glabre	20 cm de longueur, 12 cm de largeur	
3	nervure					proéminente	saillante		
4	nervures					incurvées			secondaires
5	marge			anastomosées					
6	réseau								

Fig. 7. Proof of concept for fine-grained NER with dataframe format showing the relationship between organs and fine-grained descriptors.

Acknowledgments

The authors express gratitude for the support received from the French National Research Institute for Sustainable Development (IRD) within the framework of the e-COL+ project.

References

1. Adeline Kerner, Sylvain Bouquin, Rémy Portier, and Régine Vignes Lebbe. The 8 years of existence of xper3: State of the art and future developments of the platform. *Biodiversity Information Science and Standards*, 5:e74250, 2021.
2. Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017, 2023.
3. J. Mason Heberling. Herbaria as big data sources of plant traits. *International Journal of Plant Sciences*, 183(2):87–118, 2022.

4. Viktor Domazetoski, Holger Kreft, Helena Bestova, Philipp Wieder, Radoslav Koynov, Alireza Zarei, and Patrick Weigelt. Using natural language processing to extract plant functional traits from unstructured text. *bioRxiv*, 2023.
5. Maria Auxiliadora Mora-Cross, William Ulate, Brandon Sthuar Retana Chacón, María Fernanda Biarreta Portillo, Josué David Castro Ramírez, and Jose Alejandro Chavarria Madriz. Structuring information from plant morphological descriptions using open information extraction. *Biodiversity Information Science and Standards*, 7:e113055, 2023.
6. Shaowen Zhou, Bowen Yu, Aixin Sun, Cheng Long, Jingyang Li, Haiyang Yu, Jian Sun, and Yongbin Li. A survey on neural open information extraction: Current status and future directions, 2022.
7. Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.
8. Lakshmi Manohar Akella, Catherine N. Norton, and Holly Miller. Netineti: Discovery of scientific names from text using machine learning methods. *BMC Bioinformatics*, 13(1):211, 2012.
9. Bradley C. Bennett and Michael J. Balick. Does the name really matter? the importance of botanical nomenclature and plant taxonomy in biomedical research. *Journal of Ethnopharmacology*, 152(3):387–392, 2014.
10. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022.
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
12. Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The best of both worlds. *Computing in Science Engineering*, 13(2):31–39, 2011.
13. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.