FloraNER: A new dataset for species and morphological terms named entity recognition in French botanical text

Ayoub Nainia, Régine Vignes-Lebbe, Eric Chenin, Maya Sahraoui, Hajar Mousannif, Jihad Zahir

Data in Brief 56 (2024) 110824

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

FloraNER: A new dataset for species and morphological terms named entity recognition in French botanical text

Ayoub Nainia[a,*], Régine Vignes-Lebbe[a], Eric Chenin[b], Maya Sahraoui[a,c], Hajar Mousannif[d], Jihad Zahir[b,d]

[a] Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, Muséum National d'Histoire Naturelle, CNRS, EPHE-PSL, Université des Antilles, F-75005, Paris, France
[b] UMMISCO, IRD France Nord, Bondy, France
[c] Institut des Systèmes Intelligents et de Robotique (ISIR), Sorbonne Université, CNRS, F-75005 Paris, France
[d] LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco

ARTICLE INFO

Article history:
Received 7 May 2024
Revised 17 July 2024
Accepted 5 August 2024
Available online 10 August 2024

Dataset link: FloraNER: a Named Entity Recognition Dataset for Botanical French Text (Original data)

Keywords:
NER Dataset
Biodiversity dataset
Species identification dataset
Plant morphology dataset

ABSTRACT

FloraNER is a distantly supervised named entity recognition dataset (NER). The dataset is built from botanical French literature extracted from the OCR-preprocessed flora of New Caledonia, provided by the National Museum of Natural History in France (MNHN), and distantly annotated with a botanical French corpus created by merging botanical lexicons available online. FloraNER comprises separate sub-datasets for the recognition of plant species names, as well as coarse-grained and fine-grained botanical morphological terms. The resulting datasets are in CSV format, displaying textual data, identified named entities, and their annotations, covering one named entity type "Species" (Espèce in French) for species name identification, two named entity types "Organ" and "Descriptor" for coarse-grained morphological term identification, and eight named entity types for fine-grained morphological term identification: Organ, Descriptor, Form, Color, Development, Structure, Surface, Position, Disposition, and Measure. This dataset can be utilized to train and evalu-

* Corresponding author.
E-mail address: Ayoub.Nainia@etu.sorbonne-universite.fr (A. Nainia).
Social media: @nainia_ayoub (A. Nainia)

https://doi.org/10.1016/j.dib.2024.110824

ate named entity recognition models for extracting information from botanical French literature.

Specification Table

Subject	Computational Biodiversity
Specific subject area	*Named Entity Recognition for identifying plant species names and morphological terms from French text data. Botanical information extraction from French text.*
Type of data	Tabular (csv) Text (Annotated dataset) Python Code (Jupyter Notebook)
Data collection	The data were programmatically extracted from the Flora of New Caledonia and augmented with information from Wikipedia through web scraping. ThePlantList was also utilized to automate the scraping process from Wikipedia. Additionally, online botanical glossaries were merged and used for distant annotation.
Data source location	Flora of New Caledonia, and the web.
Data accessibility	Public repository Repository name: Zenodo Data identification number: 10.5281/zenodo.10940913 Direct URL to data: https://zenodo.org/records/10940913

1. Value of the Data

- FloraNER is the first annotated dataset for botanical named entity recognition in French, encompassing the extraction of species names with coarse-grained and fine-grained morphological terms.
- This dataset is valuable for computer scientists, and botanical experts interested in automated information extraction.
- The dataset can serve as a tool for extracting structured information from the abundant French botanical literature on flora into knowledge bases.
- The dataset is provided as annotated CSV files containing text, identified named entities, their annotated start and end indices, and their pre-defined entity types, encompassing all the necessary information for conversion to any NER dataset format.

2. Background

The main objective behind building this dataset is to address the untapped botanical expertise documented in the literature, specifically from Flora of New Caledonia.[1] A flora is a book that lists and describes the plants found in a specific region. It often includes information about their characteristics, habitats, and distribution. The relevant textual information provided in floras needs to be identified, cleaned, extracted, and structured, and potentially leveraged to gain new insights into plant species distribution and characteristics. Additionally, it can enhance the development of knowledge bases on platforms like Xper3[2] [1] and Bioinspire-Explore[3] [2], by incorporating morphological knowledge extracted from relevant scientific literature using AI text analysis methods such as Named Entity Recognition. Hence, this serves as the main motivation behind compiling this dataset.

[1] https://bibliotheques.mnhn.fr.
[2] https://xper3.fr/en/.
[3] https://bioinspire-explore.mnhn.fr/.

3. Data Description

In a flora, various information including species names, morphological plant organs, and descriptors is found, highlighting the necessity of creating separate sub-datasets.

Our species names Named Entity Recognition (NER) dataset consists of 60,101 tokens. Among these, 8.4 % (5074) represent the species (SPECIES) entity type. The species named entity type is bound by the rules of botanical nomenclature, where the name of a species consists of a genus name and a specific epithet. In the flora of New Caledonia, our main data source, we encountered an abbreviated form of species names where the genus is written as an initial followed by a period. This abbreviated form is also covered in the dataset as a species named entity type. Furthermore, the species names NER dataset was enriched by introducing a new data source: Wikipedia. This extension resulted in a dataset of 2425 data rows, of which 57.7 % are from the flora of New Caledonia and 42.6 % are from Wikipedia.

The coarse-grained NER dataset for morphological plant organs and descriptors comprises 198,072 tokens. Among these, 22,977 tokens (11.6 %) are annotated as the Organ named entity type, and 19,333 tokens (9.7 %) are annotated as the Descriptor named entity type. Additionally, the average length of each text row in the coarse-grained NER dataset is 1264 characters, with an average of 19 annotated descriptors and 27 annotated organs in each data row.

The dataset also annotates the descriptor entity type from the coarse-grained NER dataset into the following fine-grained named entity types: Form, Measure, Surface, Color, Position, Disposition, Structure, and Development. The respective values of distribution throughout the dataset are as follows: 3728, 1019, 2306, 1058, 898, 705, 829, and 38.

The provided dataset repository [3] contains two CSV files representing all the listed sub-datasets:

1. Plant species names NER dataset
2. Coarse-grained and fine-grained NER dataset for botanical morphological terms. Both datasets are merged into one CSV file because they share the same morphological text and named entities, differing only in annotations.

The FloraNER repository also includes the code (under ‘./Code’) for extracting text and named entities, implementing the distant supervision annotation process, expanding the datasets using different data sources, and training and evaluating all NER models. Additionally, we provide the official train/test split of each sub-dataset under the folders ‘./Train’ and ‘./Test,’ respectively.

The number of instances of each type of named entities in the dataset is given in Table 1.

Table 1
Total instances of named entity types.

Named Entity Type	Total Instances
Species	2541
Organ (coarse-grained)	22,977
Descriptor (coarse-grained)	16,283
Form (shape)	3728
Measure	1019
Surface	2306
Color	1058
Position	898
Disposition	705
Structure	829
Development	38

Table 2
Summary of primary data sources.

Data	Flora of New Caledonia	Botanical Corpus
Size	25 documents	3530 Entries
Data Type	OCR preprocessed Text	Text (dataframe)
Data Source	National Museum of Natural History	Compiled from existing botanical glossaries online[4]

[4] https://atlasflore04.org/lexique.php.
http://www.pixiflore.com/pages/glossaire/glossaire.html
http://herbierfrance.free.fr/lexique.htm

4. Experimental Design, Materials and Methods

4.1. Data acquisition

The data acquisition process for the FloraNER dataset leverages three primary data sources: the flora of New Caledonia, a botanical corpus, and Wikipedia. Each data source is described in Table 2.

The unique feature of the FloraNER dataset is that it not only captures plant species names and botanical terms for Named Entity Recognition, but is also primarily constructed from botanical literature that details how botanists list and describe plant species.

4.2. Plant species names NER dataset

Plant species name is identified by a binomial name that consists of the scientific name of its genus and an epithet. This clarification is important since the flora of New Caledonia records plant species names in two written formats (Fig. 1 (section 1 and 2)):

- **Genus epithet**: The species name consists of the full genus name with the first letter capitalized, followed by a lowercase epithet.
- **G. epithet**: The species name consists of the initial letter of the genus name, followed by a lowercase epithet.

The plant species NER dataset requires a combination of text, named entities, and annotations. We extract the text from the flora by first identifying the occurrences of plant species names in the text, which we can find in two sections of the flora (Fig. 1 (section 1 and 2)):

- Species names of the format "G. epithet" are found in the plant species key sections.

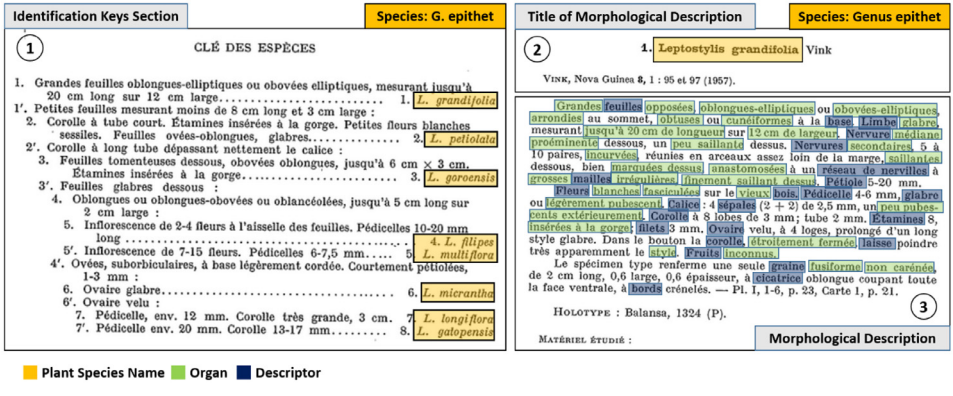

Fig. 1. Data acquisition from Flora of New Caledonia.

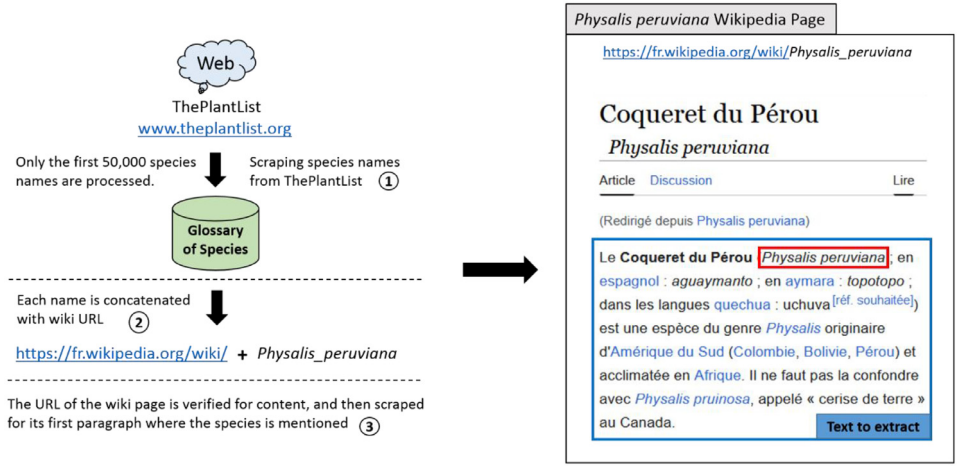

Fig. 2. Scraping Wikipedia species articles using ThePlantList.

- Species names of the format "Genus epithet" are found as titles to the morphological descriptions of plant species.

The extraction of plant species names from the flora begins with locating the species identification keys section, specifically by searching for the phrase ***"clé des espèces"*** (Fig. 1 **(section 1)**), which can be written in either lowercase or uppercase. Therefore, a lowercase normalization step is required.

Next, we split the identification keys section by line breaks. For each line, we identify occurrences of plant species names in the form "G. epithet" using the regular expression *"([A-Z][.][][a-z]+[-][a-z]+|[A-Z][.][][a-z]+)"*.

We proceed similarly for species of the form "Genus epithet" (Fig. 1 **(section 2)**). We search for occurrences of plant species names as titles using the regular expression *"(?<!\S)\d[.][][A-Z][a-z]+[][a-z]+[][A-Z][a-z]+"*, from which we extract the full plant species name using the regular expression *"[A-Z][a-z]+[][a-z]+"* to represent the named entity.

The flora-derived text for plant species names captures identification keys and title-based species, but it falls short in including plant species names within text paragraphs. To address this limitation, we leverage Wikipedia, which contains numerous articles about plant species where we can find occurrences of plant species names within text.

Our approach for scraping relevant text from Wikipedia[5] (illustrated in Fig. 2) begins with gathering a large set of plant species names from ThePlantList[6] through web scraping. These names are then concatenated with corresponding Wikipedia URLs. For example, for the species name ***"Calvoa grandifolia"***, and the given Wikipedia URL (https://fr.wikipedia.org/wiki/) the corresponding Wikipedia URL would be the result of concatenating the Wikipedia base URL the species name as follows:

•Plant species Wikipedia article: https://fr.wikipedia.org/wiki/Calvoa_grandifolia.

We verify the availability of potential articles for each plant species. If an article exists, we extract the first paragraph from the Wikipedia article, using the plant species name as the identified entity. To ensure diversity in our dataset and avoid dominance by Wikipedia text, we

[5] https://fr.wikipedia.org/wiki/.
[6] http://www.theplantlist.org/.

index	Text	Species	Annotations
0	Kibatalia stenopetala est une espèce de plantes de la famille des Apocynaceae.	Kibatalia stenopetala	[(0, 21, 'ESPECE')]
1	La ciboulette ou civette (Allium schoenoprasum L.) est une plante aromatique de la famille des Amaryllidacées (anciennement Liliacées ou Alliacées), cultivée pour ses feuilles souvent utilisées comme condiment.	Allium schoenoprasum	[(26, 46, 'ESPECE')]
2	3. Cryptocarya velutinosa Kostermans, sp. nov.	Cryptocarya velutinosa	[(3, 25, 'ESPECE')]
3	Pittosporum leroyanum est une espèce de plante à fleurs de la famille des Pittosporaceae.	Pittosporum leroyanum	[(0, 21, 'ESPECE')]
4	1. Feuilles charnues. Silicules de longueur > 4,5 mm ; graines non ailées. 1. L. bidentatum Y. Feuilles membraneuses. Silicules de longueur < 4 mm ; graines ailées.	L. bidentatum	[(78, 91, 'ESPECE')]
5	2. Symplocos baptica Brongniart & Gris	Symplocos baptica	[(3, 20, 'ESPECE')]
6	Seseli libanotis, le Libanotis en français, est une espèce de plantes herbacées de la famille des Apiaceae (Ombellifères), sous-famille des Apoideae, tribu des Apieae.	Seseli libanotis	[(0, 16, 'ESPECE')]
7	courte, robuste, de 1-1,5 cm .....5. C. parvifolia	C. parvifolia	[(58, 71, 'ESPECE')]

Fig. 3. Annotated data samples from the species NER dataset.

limit our search to the first 50,000 scraped species, resulting in the discovery of 1051 relevant Wikipedia articles. The details of this scraping process, utilizing the Plant List, are outlined in the accompanying code in the folder *'/Code'*:

1. *'./Code/1)_FloraNER_Extracting_text_with_species_names.ipynb'*
2. *'/Code/2)_FloraNER_Scraping_plant_species_text_from_Wikipedia.ipynb'*

As a result, we obtain a species names dataset of 2425 data rows for one named entity type: 'SPECIES' (Fig. 3).

4.3. Coarse-grained and fine-grained NER dataset for morphological terms

We extracted descriptions capturing plant morphological terms (including plant organs and descriptors) exclusively from the flora of New Caledonia (Fig. 1 **(section 3)**). While considering Wikipedia as an additional data source for this FloraNER subset might appear beneficial, it is not the case since Wikipedia does not provide literature-level morphological descriptions for plant species.

In extracting descriptions from the flora of New Caledonia, we observed that these descriptions can vary widely in length, ranging from multiple pages to concise paragraphs. They typically follow titles containing species names and often conclude with the specific phrase *"Matériel Étudié"* (as illustrated in Fig. 1 **(section 3)**).

Our extraction process begins by identifying species name titles using the regular expression *"((?<!\S)\d{1,2}[.][][A-Z][a-z]+[][a-z]+[][A-Z][a-z]+|(?<!\S)\d{1,2}[.][][A-Z][a-z]+[][a-z]+[][A-Z][.][][A-Z][a-z]+|(?<!\S)\d{1,2}[.][][A-Z][a-z]+[][a-z]+[][(][A-Z][a-z]+[)][][A-Z][a-z]+|(?<!\S)\d{1,2}[.][][A-Z][a-z]+[][a-z]+[][(][A-Z][a-z]+[][&][][A-Z][a-z]+[)][][A-Z][a-z]+)"*.

These titles include the numerical designation of the species name from the identification keys, along with the species name and the scientist who described it. Some titles do not contain all this information, which highlights the complexity of our regular expression to cover all possible title variations.

Subsequently, we retrieve all text following the title until we encounter the phrase *"Matériel Étudié"*, which marks the end of each morphological description.

This extraction resulted in 838 data rows of morphological descriptions (Fig. 4).

The full extraction process is detailed in the notebook under the *'./Code'* folder *('./Code/3)_FloraNER_Extracting_morphological_descriptions_from_Flora_of_New_Caledonia')*

4.4. Data annotation

The FloraNER sub-datasets were annotated using distant supervision rather than manual annotation. For plant species names, we extracted named entities using regular expressions, as mentioned earlier. These extracted named entities were then used to define their character-level start and end indices in the text. Each annotated named entity is represented as a tuple con-

Index	Text	Organ Entities	Descriptor Entities	Coarse-grained Annotation	Fine-grained Annotation
0	grandes feuilles opposées, oblongues-elliptiques ou obovées-elliptiques, arrondies au sommet, obtuses ou cunéiformes à la base limbe glabre, mesurant jusqu'à 20 cm de longueur sur 12 cm de largeur nervure médiane proéminente dessous, un peu saillante dessus nervures secondaires, 5 à 10 paires, incurvées, réunies en arceaux assez loin de la marge, saillantes dessous, bien marquées dessus, anastomosées à un réseau de nervilles à grosses mailles irrégulières, finement saillant dessus pétiole 5-20 mm fleurs blanches fasciculées sur le vieux bois pédicelle 4-6 mm, glabre ou légèrement pubescent galice : 4 sépales (2 + 2) de 2,5 mm, un peu pubes-cents extérieurement corolle à 8 lobes de 3 mm; tube 2 mm étamines 8, insérées à la gorge; filets 3 mm ovaire velu, à 4 loges, prolongé d'un long style glabre dans le bouton la corolle, étroitement fermée, laisse poindre très apparemment le style fruits inconnus le spécimen type renferme une seule graine fusiforme non carénée, de 2 cm long, 0,6 large, 0,6 épaisseur, à cicatrice oblongue coupant toute la face ventrale, à bords crénelés	['bouton', 'pédicelle', 'corolle', 'tube', 'feuilles', 'marge', 'filets', 'loges', 'nervure', 'sépales', 'nervilles', 'fleurs', 'style', 'lobes', 'nervure', 'laisse', 'réseau', 'bois', 'corolle', 'base', 'limbe', 'cicatrice', 'fruits', 'pétiole', 'ovaire', 'style', 'graine', 'bords']	['fermée', 'pubes-cents', 'cunéiformes', 'ventrale', '20 cm de longueur', 'obovées-elliptiques', '12 cm de largeur', 'oblongues-elliptiques', 'incurvees', 'insérées', 'velu', 'blanches', 'fusiforme', 'pubescent', 'secondaires', 'proéminente', 'anastomosées', 'opposées', 'fasciculées', 'carénée', 'crénelés', 'saillante', 'obtuses', 'glabre']	[(650, 661, 'DESCRIPTEUR'), (968, 977, 'DESCRIPTEUR'), (616, 623, 'ORGANE'), (1086, 1091, 'ORGANE'), (1033, 1042, 'ORGANE'), (299, 308, 'DESCRIPTEUR'), (1092, 1100, 'DESCRIPTEUR'), (95, 102, 'DESCRIPTEUR'), (53, 72, 'DESCRIPTEUR'), (160, 177, 'DESCRIPTEUR'), (901, 906, 'ORGANE'), (508, 514, 'ORGANE'), (827, 833, 'ORGANE'), (908, 914, 'ORGANE'), (18, 26, 'DESCRIPTEUR'), (858, 864, 'DESCRIPTEUR'), (9, 17, 'ORGANE'), (749, 755, 'ORGANE'), (549, 553, 'ORGANE'), (346, 351, 'ORGANE'), (200, 207, 'ORGANE'), (982, 989, 'DESCRIPTEUR'), (690, 695, 'ORGANE'), (129, 134, 'ORGANE'), (811, 817, 'DESCRIPTEUR'), (135, 141, 'DESCRIPTEUR'), (395, 407, 'DESCRIPTEUR'), (123, 127, 'ORGANE'), (594, 603, 'DESCRIPTEUR'), (555, 564, 'ORGANE'), (216, 227, 'DESCRIPTEUR'), (762, 768, 'ORGANE'), (182, 198, 'DESCRIPTEUR'), (769, 773, 'DESCRIPTEUR'), (106, 117, 'DESCRIPTEUR'), (837, 844, 'ORGANE'), (244, 253, 'DESCRIPTEUR'), (961, 967, 'ORGANE'), (28, 49, 'DESCRIPTEUR'), (524, 535, 'DESCRIPTEUR'), (423, 432, 'ORGANE'), (1074, 1082, 'DESCRIPTEUR'), (705, 709, 'ORGANE'), (779, 784, 'ORGANE'), (805, 810, 'ORGANE'), (413, 419, 'ORGANE'), (678, 685, 'ORGANE'), (728, 736, 'DESCRIPTEUR'), (271, 282, 'DESCRIPTEUR'), (491, 498, 'ORGANE'), (573, 579, 'DESCRIPTEUR'), (866, 872, 'ORGANE'), (515, 523, 'DESCRIPTEUR')]	[(650, 661, 'DESCRIPTEUR'), (395, 407, 'DISPOSITION'), (616, 623, 'ORGANE'), (1086, 1091, 'ORGANE'), (1033, 1042, 'ORGANE'), (299, 308, 'DESCRIPTEUR'), (271, 282, 'POSITION'), (53, 72, 'DESCRIPTEUR'), (515, 523, 'COULEUR'), (811, 817, 'SURFACE'), (135, 141, 'SURFACE'), (901, 906, 'ORGANE'), (508, 514, 'ORGANE'), (827, 833, 'ORGANE'), (908, 914, 'ORGANE'), (182, 198, 'MESURE'), (858, 864, 'DESCRIPTEUR'), (9, 17, 'ORGANE'), (749, 755, 'ORGANE'), (594, 603, 'SURFACE'), (549, 553, 'ORGANE'), (346, 351, 'ORGANE'), (200, 207, 'ORGANE'), (524, 535, 'DISPOSITION'), (690, 695, 'ORGANE'), (129, 134, 'ORGANE'), (982, 989, 'DESCRIPTEUR'), (769, 773, 'SURFACE'), (1092, 1100, 'STRUCTURE'), (123, 127, 'ORGANE'), (244, 253, 'SURFACE'), (555, 564, 'ORGANE'), (216, 227, 'DESCRIPTEUR'), (762, 768, 'ORGANE'), (106, 117, 'FORME'), (837, 844, 'ORGANE'), (961, 967, 'ORGANE'), (573, 579, 'SURFACE'), (160, 177, 'MESURE'), (28, 49, 'DESCRIPTEUR'), (423, 432, 'ORGANE'), (1074, 1082, 'DESCRIPTEUR'), (705, 709, 'ORGANE'), (18, 26, 'DISPOSITION'), (779, 784, 'ORGANE'), (805, 810, 'ORGANE'), (413, 419, 'ORGANE'), (678, 685, 'ORGANE'), (728, 736, 'DESCRIPTEUR'), (491, 498, 'ORGANE'), (95, 102, 'FORME'), (968, 977, 'FORME'), (866, 872, 'ORGANE')]

Fig. 4. Coarse-grained and fine-grained annotated data sample.

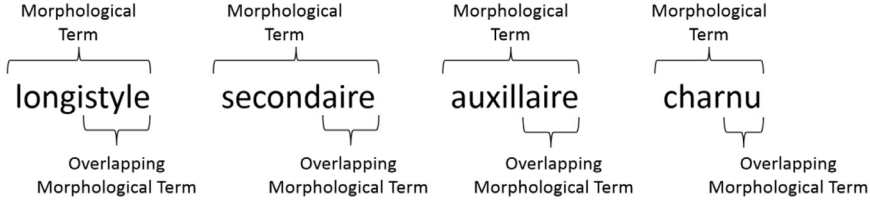

Fig. 5. Examples of overlapping morphological terms.

taining its start index, end index, and type ***(start_index, end_index, named-entity type)***, in this case, 'SPECIES'.

The distant annotation process for both coarse-grained and fine-grained approaches involve utilizing a corpus of botanical French terms. This corpus was created by aggregating available botanical glossaries online, which capture a comprehensive set of botanical terms. Each term in the corpus is accompanied by its definition and part-of-speech tagging.

For coarse-grained Named Entity Recognition (NER), we distinguish between two types of named entities: "Organ" and "Descriptor", categorized based on their part of speech as nouns and adjectives, respectively.

In our annotation corpus, we label the noun botanical terms as "Organ", while the adjectives as "Descriptor" (Fig. 6). Subsequently, we align these labeled morphological terms from the annotation corpus with their occurrences in the morphological text to annotate them as named entities Morphological text in Fig. 4.

Our annotation approach involves matching tokens from the morphological descriptions with terms in the corpus. This process ensures that the identification of named entities is categorized either as "Organ" or "Descriptor" within the coarse-grained subset. The annotation tuple includes character-level start and end indices along with the named entity type. After annotating the named entities in terms of their star and end index, and their entity type, we implement a step to eliminate any potential overlapping annotations (Fig. 5). This means that the annotation of a given description should not have the same start or end index as that would mean that the annotated named-entities are overlapping. We solve this by eliminating the annotation that has the least number of characters. Meaning, we eliminate the contained annotation and not the containing one.

The code for extracting and annotating the coarse-grained named entities is available in:

1. '*./Code/4)_FloraNER_Extracting_and_annotating_coarse_grained_named_entities.ipynb*'
2. '*./Code/5)_FloraNER_Extarcting_remaining_(more)_Named_Entities.ipynb*'

Morphological Corpus

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

An adjective term is assigned the type "**Descriptor**"; otherwise, it is "**Organ**".

Labeled Morphological Corpus

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. 6. Assigning the coarse-grained named entity types in the botanical corpus.

Type	Assigned Phrases / Key words
Form (Shape)	• Shaped like • Having the shape of • In the shape of • That resembles, which resembles, ...
Color	• Colored, Having the color • Presenting a mix of different colors • Of the color, Of the same color, ...
DEVELOPMENT	• Development, ...
STRUCTURE	• Structure, ...
SURFACE	• Surface, ...
POSITION	• Position, in position, ...
DISPOSITION	• Disposition, disposition

Searching the occurrence of assigned phrases in definitions to assign the appropriate type to the terms.

Botanical Corpus

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

Fig. 7. Assigning fine-grained named entity in the botanical corpus based on the definition of terms.

For fine-grained annotation, we utilize the definitions of morphological terms from the botanical corpus to determine the specific types (surface, color, development, structure, form (shape), position, and disposition) to which the identified named entity belongs. For example, if the definition of a specific botanical term in the corpus includes the expression "in the shape of", we assign the fine-grained named entity type "Form" to this term. This means we prepare a set of expressions for each fine-grained named entity type, and each expression will be searched within the definitions of terms in the corpus (Fig. 7).

Regarding measurements in the morphological descriptions, we specify the fine-grained entity type as "Measure". Plant species measurements such as diameter, width, height, and length are extracted using the following regular expressions:

- **Diameter:** *"((de)?(([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)-([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)|([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)) (cm|mm) de diamètre)"*
- **Height:** *"(([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+) m de (hauteur|haut))"*
- **Width:** *"((de)?(([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)-([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)|([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)) (cm|mm) de largeur)"*
- **Length:** *"((de)?(([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)-([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)|([\d]+[,.\d]+|[\d]*[.,\d]+|[\d]+)) (cm|mm) de (longueur|long))"*

Morphological descriptions of plant species in the flora of New Caledonia include specific descriptor named entities that are not covered in the botanical corpus. These named entities generally consist of two descriptor entities joined by a hyphen. We extract these compound

Table 3
Evaluation of reference model on FloraNER dataset.

Named Entity Type	Precision	Recall	F1-score
Species	94.31 %	98.36 %	96.29 %
Organ (Coarse-grained)	94.46 %	97.48 %	95.95 %
Descriptor (Coarse-grained)	90.88 %	94.54 %	92.68 %
Organ (Fine-grained)	95.21 %	98.82 %	96.98 %
Descriptor (Fine-grained)	94.96 %	90.63 %	92.74 %
Surface	71.91 %	96.34 %	82.35 %
Color	83.27 %	100 %	90.87 %
Development	100 %	100 %	100 %
Structure	90.62 %	97.31 %	93.85 %
Form (Shape)	86.56 %	98.02 %	91.94 %
Position	98.34 %	99.44 %	98.88 %
Disposition	93.91 %	96.52 %	95.20 %
Measure	89.07 %	100 %	94.22 %

descriptors using a regular expression that searches for a hyphenated word where at least one part is listed in the botanical corpus. This ensures that it is a morphological term and not just a random compound word.

The code for coarse-grained and fine-grained distant annotation is available in the FloraNER repository *(./Code/6)_FloraNER_Fine_grained_extraction_and_annotation.ipynb)*

5. Reference Results

To assess the effectiveness of our data for training botanical NER systems, we fine-tuned the pre-trained SpaCy[7] French language model to recognize named entities specific to the FloraNER sub-datasets. The results show good performance in terms of precision, recall, and F1-score, as given in Table 3.

For the experimental setup for training NER models with the FloraNER dataset, we used an 80 % train and 20 % test split for each dataset. To prepare the data for training, we converted the CSV files to JSON format, which is compatible with spaCy's training requirements. Each JSON data point consists of the text and its appropriate annotation.

Since the text provided in the FloraNER dataset is in French, we based our blank spaCy model on the French (*fr*) language model and then added the NER pipeline. We evaluated the models using the Scorer class to provide the evaluation metrics. The full code for training and evaluation is provided under './Code' *('./Code/Training_FloraNER_models.ipynb')*.

Limitations

While our distant annotation process has enabled the creation of a valuable dataset for named entity recognition (NER), it is important to acknowledge certain limitations inherent in this approach for the coarse-grained and fine-grained datasets:

- **Incomplete named entities**: The distant annotation process does not extract and annotate all named entities present in the text. Due to its automated nature, it may miss certain entities, resulting in incomplete annotations within the dataset.
- **Partial Coverage**: As a consequence of the limited scope of the distant annotation process, the dataset may not encompass the full range of named entities relevant to the domain. Some fine-grained named entities may be overlooked or underrepresented in the dataset.

[7] https://spacy.io/.

Ethics Statement

The data used to compile the FloraNER dataset do not pose any ethical concerns as they were collected from the available Flora of New Caledonia and Wikipedia and not a social media platform or other sensitive data sources. We did not need permission to use data from Wikipedia and Flora of New Caledonia. We did not conduct human or animal studies in our work.

Data Availability

FloraNER: a Named Entity Recognition Dataset for Botanical French Text (Original data) (ZENODO).

CRediT Author Statement

Ayoub Nainia: Methodology, Software, Data curation, Writing – original draft, Visualization, Validation; **Régine Vignes-Lebbe:** Supervision, Conceptualization, Validation, Project administration; **Eric Chenin:** Resources, Supervision, Project administration; **Maya Sahraoui:** Methodology, Validation; **Hajar Mousannif:** Supervision, Project administration; **Jihad Zahir:** Supervision, Conceptualization, Methodology, Validation, Project administration.

Acknowledgements

This project is co-funded by the European Union's Horizon Europe research and innovation program Cofund SOUND.AI under the Marie Sklodowska-Curie Grant Agreement No 101081674. This research is also part of the e-COL+ project (ANR-21-ESRE-0053).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Kerner, S. Bouquin, R. Portier, R. Vignes Lebbe, The 8 years of existence of Xper3: state of the art and future developments of the platform, Biodivers. Inf. Sci. Stand. 5 (2021) e74250, doi:10.3897/biss.5.74250.
[2] A. Saint-Sardos, A. Aish, N. Tchakarov, T. Bourgoin, L-M Petit, J.-S. Sun, R. Vignes-Lebbe, Bioinspire-explore: taxonomy-driven exploration of biodiversity data for bioinspired innovation, Biomimetics 9 (2) (2024) 63, doi:10.3390/biomimetics9020063.
[3] A. Nainia, R. Vignes Lebbe, E. Chenin, M. Sahraoui, H. Mousannif, J. Zahir, FloraNER: a named entity recognition dataset for botanical French Text [Data set], Zenodo (2024), doi:10.5281/zenodo.10940913.