The training dataset of the LifeCLEF2021 Plant Identification challenge is based on the same visual data used during the previous LifeCLEF 2020 Plant Identification challenge but also introduces new data related to 5 functional traits covering exhaustively all the 1000 species of the challenge.

1. Visual training dataset

The training dataset can be downloaded here:
https://zenodo.org/record/3658343#.Xj2k1eEo-V4

This package is organized into three subfolders:

- "herbarium" subdirectory contains the vast majority of the data: it is a collection of about 327k herbarium scans relating to a selection of 1000 species of Amazonian plants mainly centered on French Guiana. The herbarium sheets are coming from two sources: the Herbier IRD de Guyane" digitized in the context of the e-ReColNat project, and iDigBio, a large international platform aggregating and giving access millions of images of herbarium specimens hosted by various National Museum of Natural History and botanical institutes around the world. Pictures and theirs related metadata xml files are organized into subfolders, one for each species. The name of the subfolders are directly the content of ClassId field that can be found in the xml content. The xml file contains various information (when available) like longitude, latitude, place, date, taxonomy, and some tags on the pictures. Some herbarium sheets are related to a same plant observation or "specimen" and can be found through the ObservationId field. All the pictures were resized to a maximum height of 1024 pixels, but the field OriginalUrl can be used to get pictures with a higher resolution.

- the "herbarium_photo_associations" subdirectory contains more than 3 hundreds specimens related to about 250 species where we are supposed to have for each individual plant identified by the (ObservationId field) some pictures in the field and one or more herbarium sheets. The PhotoType field in the xml can take the value of "herbarium" or "Photo" in order to identify if the content is related to an herbarium sheet of a picture in the field. The field "HerbariumPhotoAssociation" explicitly indicates if there is an association or not between pictures in the field and herbarium sheets related to a same specimen (but it's possible that sometimes there are missing photos...). As the previous "herbarium" directory, pictures and their related metadata xml files are organized into subfolders, one for each species identified by a ClassId.

- Finally, the "photo" subfolder contains few pictures in the field that was provided by the IdigBio API when the training species were requested.

Pictures in the field contained into the "herbarium_photo_associations" and "photos" subdirectories could be used classically as an extra training dataset for fine tuning directly a ConvNet model for species classification. In the same vein, it would be possible also to use

pictures in the field related to Amazonian plants like the PlantCLEF2019 training dataset. But we really encourage the participants to act as if no data were available other than herbarium sheets in the world (which is actually the case for many species in the training set and the test set). Photos in the “herbarium_photo_associations”, and eventually “photos”, subdirectory/ies are essentially provided to allow learning a mapping between the herbarium sheets domain and the field pictures domain.

2. Functional traits

The file PlantCLEF2021_all_ClassId_Species_to_five_traits.csv contains additional metadata at the species level expressing functional traits, a very valuable information that can potentially help improve prediction models.
Indeed, it can be assumed that species which share the same functional traits also share to some extent common visual appearances. This information can then potentially be used to guide the learning of a model through auxiliary loss functions.
These data were collected through the Encyclopedia of Life API. The 5 most exhaustive traits were verified and completed by experts of the Guyanese flora, so that each of the 1000 dataset species all have a value for each trait. The first two columns of the file PlantCLEF2021_all_ClassId_Species_to_five_traits.csv contain basically the ClassId and the species name used in the visual training dataset, while the remaining columns contain the values of the 5 traits.

Below we list the names of the 5 traits as well as the possible values associated with these traits.

- **plant growth form**

This trait describes which plant growth form(s) can take a species among these 4 possibilities:

climber
herb
shrub
tree

it is important to note that a species can sometimes be associated with several forms of growth. For example, a young plant of the species Justicia betonica L. can be considered as an herb while in adulthood it would be described as a shrub.

- **habitat**

This trait is actually a non standardised free tag(s) describing the typical habitat of a given species.
Examples of most used words: tropical, moist, broadleaf, forest, flooded, grassland, rocky, non-wetland, savanna, shrubland, coastal, etc.

- **plant lifeform**

This trait often refers to the physical support of development used by a species. Below a list of possible values and some definitions

- aquatic plant
- epiphyte: an organism that grows on the surface of a plant and derives its moisture and nutrients from the air, rain, water (in marine environments) or from debris accumulating around it
- geophyte: species that develop organs for storing energy (water or carbohydrates)
- helophyte: a plant that grows in or near water and is either emergent, submergent, or floating
- hemiepiphyte: a plant that spends part of its life cycle as an epiphyte
- hydrophyte: close to helophyte
- lithophyte: plants that grow in or on rocks
- pleustophyte: it is a plant living in the thin surface layer existing at the air-water interface of a body of water which serves as their habitat
- succulent plant: is a plant with parts that are thickened, fleshy, and engorged, usually to retain water in arid climates or soil conditions (close to geophyte)
- terrestrial plant

- **trophic guild**

This trait can be common to any group of species that exploit the same resources, or that exploit different resources in related ways.

- carnivorous plant
- hemiparasite: partially parasite (see below)
- parasite: is a plant that derives some or all of its nutritional requirement from another living plant
- photoautotroph: these plants are capable of synthesizing their own food from inorganic substances using light as an energy source
- saprotrophic: plants which secrete digestive juices in dead and decaying matter and convert it into a solution and absorb it.

- **woodiness**

Woodiness basically expresses whether the species is capable of producing "lignin", in other words wood. The values are then:

herb
woody