The training dataset of the LifeCLEF 2020 Plant Identification challenge can be found by following this link:
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 content various information (when available) like longitude, latitude, place, date, taxonomy, 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 where 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 inf the field. The field "HerbariumPhotoAssociation" explicitly indicated 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 theirs 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.