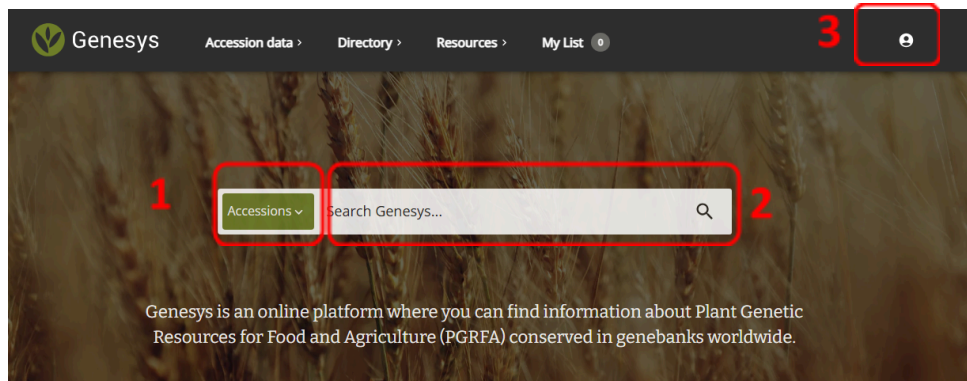


DOWNLOAD OCCURRENCE DATA

This section illustrates the process for downloading occurrence data from the available data sources for a specific crop or taxon.

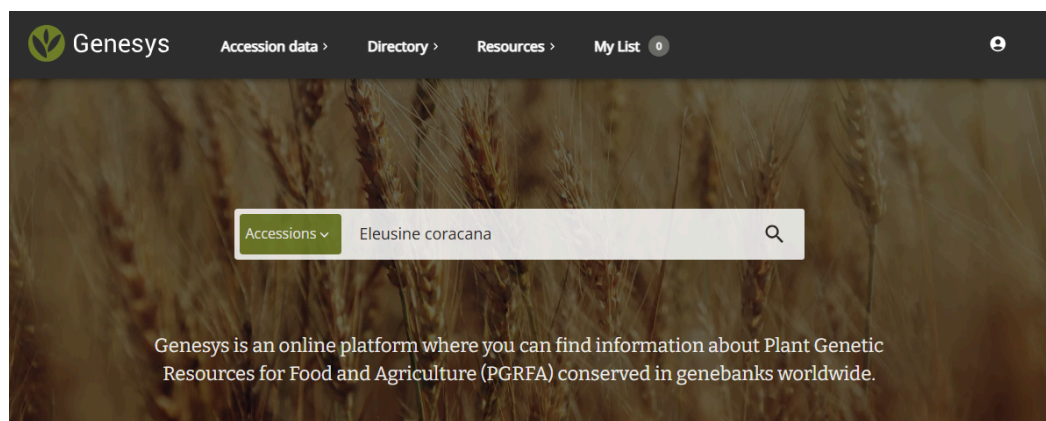
DOWNLOAD OCCURRENCES FROM GENESYS-PGR (<https://www.genesys-pgr.org>):

Genesys-PGR (Plant Genetic Resources) is an online platform that provides access to information about Plant Genetic Resources for Food and Agriculture (PGRFA) conserved in gene banks worldwide. Launched in 2011, Genesys aims to facilitate the discovery, access, and use of ex situ collections of plant germplasm. It serves researchers, plant breeders, and other users by offering a unified gateway to explore the diversity of crops stored in gene banks. Genesys integrates data from various global networks and gene banks, detailing millions of accessions, enabling efficient searches, and supporting the conservation and sustainable use of crop diversity.



1. Type of data for query: Make sure “Accessions” is selected.
2. Search bar: field to type a query, user can type any information related to accessions (accessionID, genebankID, genus, species, etc.).
3. Loggin button: it is recommended to create an account and log in (Google authentication is also available).

In this example we will query for finger millet (*Eleusine coracana*) accession data, we strongly recommend using the scientific name as query, then click on the magnified glass icon.



Once the genesys-pgr portal finish loading, we should see the next page.

1 Manual filtering

2 FILTER ACCESSIONS

3 Accession data > Directory > Resources > My List 0

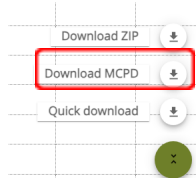
4 Accession browser

Approximately 15,373 accessions Full-text: Eleusine coracana Excluding Historical

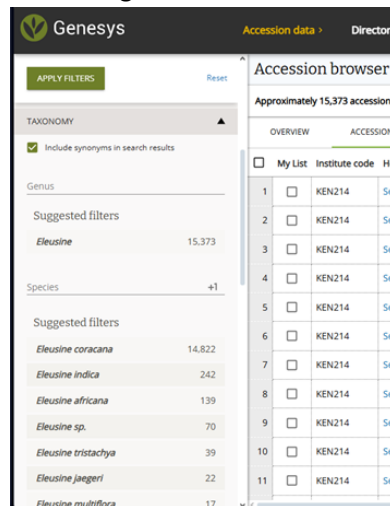
OVERVIEW	ACCESSIONS	MAP	IMAGES	SUBSETTING TOOL	ANALOGUES TOOL	UNITS		
My List	Institute code	Holding Instit...	Accession number	Accession name	Taxonomy	Crop name	Crop	Biologic
1	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN1000	Eleusine coracana			
2	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0127	Eleusine coracana			
3	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0112	Eleusine coracana			
4	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0134	Eleusine coracana			
5	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0140	Eleusine coracana			
6	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0142	Eleusine coracana			
7	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0145	Eleusine coracana			
8	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0118	Eleusine coracana			
9	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0912	Eleusine coracana			
10	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0919	Eleusine coracana			
11	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0865	Eleusine coracana			

5 Download Button

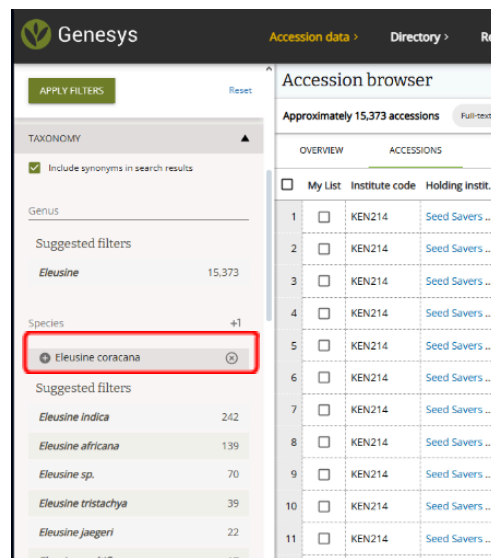
1. Make sure “Manual filtering” tab is selected.
2. Filter bar: shows the available filters – after a filter is selected/modified make sure of click on “APPLY FILTERS” button to apply the filters.
3. Section with information of the filters that are currently applied and the number of records retrieved.
4. Passport data database visualization.
5. Download Button: Currently data can be downloaded using three different options. We recommend downloading the MCPD as illustrated below.



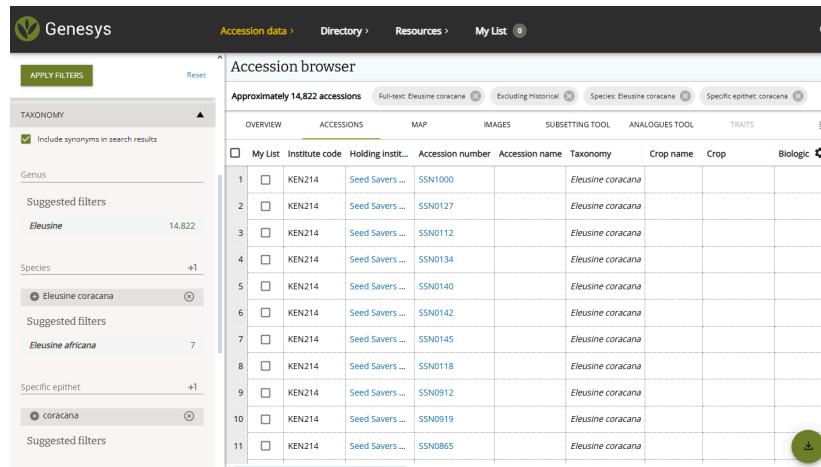
To refine the query further, we will select only accessions of *Eleusine coracana*. As shown, the data is currently filtered to include only the *Eleusine* genus.



In the **Taxonomy** filter, we can search for the specific species to select (can select multiple species), in this case, we will choose “*Eleusine coracana*” from the **Species** filter



Once it's selected, click on “APPLY FILTERS” button to apply.



Genesys Accession browser

Approximately 14,822 accessions

Full-text: Eleusine coracana Excluding historical Species: Eleusine coracana Specific epithet: coracana

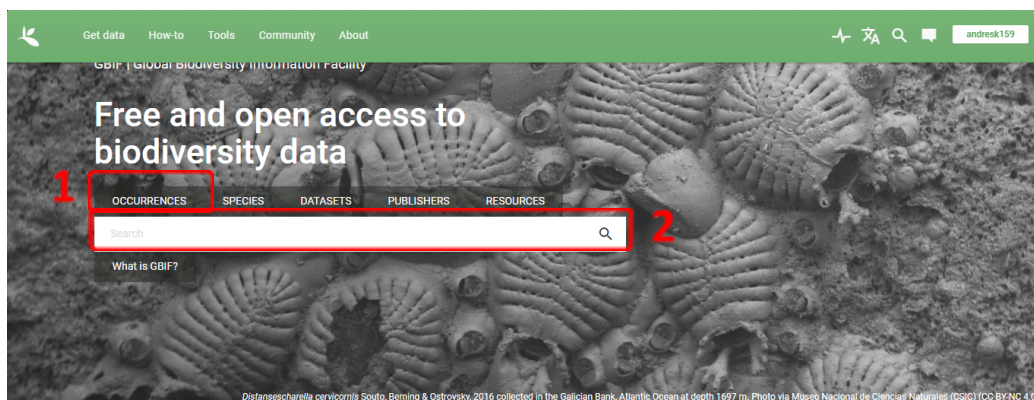
OVERVIEW ACCESIONS MAP IMAGES SUBSETTING TOOL ANALOGUES TOOL TRAITS

	My List	Institute code	Holding instit...	Accession number	Accession name	Taxonomy	Crop name	Crop	Biologic
1	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN1000		Eleusine coracana			
2	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0127		Eleusine coracana			
3	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0112		Eleusine coracana			
4	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0134		Eleusine coracana			
5	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0140		Eleusine coracana			
6	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0142		Eleusine coracana			
7	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0145		Eleusine coracana			
8	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0118		Eleusine coracana			
9	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0912		Eleusine coracana			
10	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0919		Eleusine coracana			
11	<input type="checkbox"/>	KEN214	Seed Savers ...	SSN0865		Eleusine coracana			

Now, the results are restricted to 14.822 accessions. Now we can download the database.

DOWNLOAD OCCURRENCES FROM GBIF(<https://www.gbif.org>):

GBIF stands for the Global Biodiversity Information Facility. It is an international network and data infrastructure funded by governments and organizations around the world. GBIF provides free and open access to biodiversity data from a wide range of sources, including scientific institutions, museums, and research projects.



1. Navigate to the occurrence search tool.
2. Search bar for text queries.

By default, when any text is entered into the search bar, the GBIF portal automatically initiates a search for all the data types. We suggest clicking on OCCURRENCES to restrict the search. Then it will initiate the **Occurrences** search, shown below.

The screenshot shows the GBIF Occurrences search interface. The top navigation bar includes links for 'Get data', 'How-to', 'Tools', 'Community', and 'About'. On the right, there are icons for a heart, a star, a magnifying glass, and a user profile labeled 'andresk159' (annotated with a red '5'). Below the navigation bar, the left sidebar is titled 'Occurrences' and contains a search bar (annotated with a red '1') and a filter menu (annotated with a red '2'). The filter menu includes options for 'Occurrence status' (with 'Present' selected), 'Licence', 'Scientific name', 'Basis of record', 'Year', 'Month', 'Location', 'Administrative areas (gadm.org)', 'Country or area', 'Continent', 'Dataset', 'Publisher', and 'IUCN Global Red List Category'. The main content area is titled 'SEARCH OCCURRENCES | 3,035,911,374 RESULTS' and features a 'DOWNLOAD' button (annotated with a red '4'). Below the button is a table of search results (annotated with a red '3') with columns for 'Scientific name', 'Country or area', 'Coordinates', 'Event date', 'Occurrence status', and 'Basis of record'. The table lists various species such as *Acacia implexa* Benth., *Usnea intermedia* (A.Massal.) Jatta, *Tiliqua scincoides* (White, 1790), *Mareca strepera* (Linnaeus, 1758), *Pseudocheirus peregrinus* (Boddaert, 1785), *Ondatra zibethicus* (Linnaeus, 1766), *Sitta europaea* Linnaeus, 1758, *Anas superciliosa* J.F.Gmelin, 1789, *Prunella modularis* (Linnaeus, 1758), *Callidemum* Blanchard, 1853, *Spherothores vieilloti* Vigors & Horsfield, 18..., and *Tiliqua scincoides* (White, 1790).

1. Text search bar
2. Filter Menu
3. Data retrieved visualization
4. Download button
5. Login button

In this example we will query for finger millet (*Eleusine coracana*) accession data, we strongly recommend using the scientific name as query, then click on the magnified glass icon.

Always download the “**DARWIN CORE ARCHIVE**” file. This will download a ZIP file with all the metadata related to the query. Within this zipped file you will find the next files:

dataset	File folder		
citations.txt	Text Document	3 KB	No
meta.xml	Microsoft Edge HTML Do...	4 KB	No
metadata.xml	Microsoft Edge HTML Do...	2 KB	No
multimedia.txt	Text Document	5 KB	No
occurrence.txt	Text Document	27 KB	No
rights.txt	Text Document	2 KB	No
verbatim.txt	Text Document	20 KB	No

The **occurrence.txt** file contains all the passport data for the downloaded accessions. Additional files provide supplementary information, such as **citations.txt**, which include details on how to cite the data for a scientific article.

DOWNLOAD OCCURRENCES FROM USDA-GRIN

NOTE: This database contains occurrences for taxa available in the United States Department of Agriculture (This database only contains Germplasm bank accessions!).

In your browser type: <https://npgsweb.ars-grin.gov/gringlobal/search>. The following scheme will appear in your browser. The website possesses four tabs 1) Simple search (Search using PI number), 2) List search (This option serves to search a list of accessions (e.g. PI 651794,PI 651649,PI 651650) 3) Advanced search (This option is the only option that includes the search with species name 4) Results (This tab returns the search results). Also on the top of the website the scroll menu allows you to increase the number of accessions required by the user.

Select the tab for the type of search. Each tab has everything you need to do to perform that type of search.

Return up to 500 ▾

(Results of more than 500 will not return images.)

Simple Search	List Search	Advanced Search	Results
<hr/>			
<input type="text" value="e.g., PI 651650"/>		<input type="button" value="Search"/>	

This search will show only accessions that have material that may be requested today.

Please use the Advanced search and change the Return up to 10000 as follows to obtain an accurate number of accessions. Advanced search allows you to search by species name and limit the accession to countries if it is necessary. The following graphic shows how to search for records of ***Amaranthus caudatus***. The search also will be limited to searching active accessions by activating the option Accessions that can be requested (including out of season). Finally, to search click on the search button.

Select the tab for the type of search. Each tab has everything you need to do to perform that type of search.

Return up to

10000

Update Limit

(Results of more than 500 will not return images.)

Simple Search

List Search

Advanced Search

Results

The more information you provide, the better the search will be.

Q

e.g., PI 651649

Q Search

✕ Clear All

Scientific name (any part, no hybrid symbols)

Amaranthus caudatus

Plant name

e.g., Rufa

Genebank

Country of Origin

Afghanistan

Albania

Algeria

Angola

...

Reset Countries

Other search criteria:

Select one

Search for:

☒ Accessions that can be requested (including out of season)

☐ All accessions - Including historic (not in the NPGS collections, information only)

Limit accessions displayed:

☐ With genomic data

☐ With NCBI link

☐ With images

☐ Only non-Genetically Engineered

Q Search

The results will appear in the Results tab as shown in the following graphics. Please click on the View Accession Details to add the information attached to each germplasm passport and show/hide columns to show the new columns.

Select the tab for the type of search. Each tab has everything you need to do to perform that type of search.

Return up to

(Results of more than 500 will not return images.)

If your results aren't what you expected, try using the Advanced Search tab and filling in more information.

Your query included: **All accessions**

Scientific name (any part, no hybrid symbols): *Amaranthus caudatus*

☐ **View Observation Data**

Selected item(s) below:

Basic Info

Source Info

Show all columns

Show/hide columns

Show 10 rows

Excel

Search:

Showing 1 to 10 of 763 entries

Previous

1

2

3

4

5

...

77

Next

☐ ▲

ACCESSION

PLANT NAME

TAXONOMY

ORIGIN

GENEBANK

AVAILABILITY

IMPROVEMENT LEVEL

Search ACCESSION

Search PLANT NAME

Search TAXONOMY

☐

PI 654390

DB 2005652

Amaranthus caudatus L.

Iowa, United States

NC7

Genetic material

☐

PI 632249

DB 2001949

Amaranthus caudatus L.

Iowa, United States

NC7

Genetic material

☐

PI 628794

Ames 23921

Amaranthus caudatus L.

Lima, Peru

NC7

Cultivar

☐

PI 669853

UC122

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

☐

PI 669855

UC124

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

☐

PI 669883

UC153

Amaranthus caudatus L.

Himachal Pradesh, India

NC7

Landrace

☐

PI 669885

UC158

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

☐

PI 669888

UC176

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

☐

PI 669889

UC177

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

☐

PI 669891

UC185

Amaranthus caudatus L.

Uttar Pradesh, India

NC7

Landrace

Showing 1 to 10 of 763 entries

Previous

1

2

3

4

5

...

77

Next


The search had 763 entries. The columns possible to evaluate show accession number, taxonomy, origin, genebank, etc...Please download by clicking on the button Excel

Table 1. Technical description of the fields available in USDA download.

field	description	example
ACCESSION	Accession name	PI 634914
PLANT NAME	variety name	Ghanair
TAXONOMY	Species or taxon name	Amaranthus caudatus L.
ORIGIN	Locality information	Azad Jammu and Kashmir, Pakistan
GENEBANK	FAO WIEWS genebank id	NC7
AVAILABILITY	Availability in USDA	
RECEIVED	Received data	1989
SOURCE TYPE	Source	COLLECTED
SOURCE DATE	Date	10/25/1988
COLLECTION SITE	Collection Site data	Below road at Bandi Bakhalan, 5 km north of Chikar.
COORDINATES	decimal coordinates	34.15000000, 73.68333333
ELEVATION	Elevation data	1540
HABITAT	Habitat	Terraced farm. On a 5 degree slope facing northwest.
IMPROVEMENT LEVEL	type of record: CWR, landrace, or plant breeding	Landrace
NARRATIVE	Accession description	A grain type, with black seeds, red foliage and a drooping, red seed head. Used as a vegetable. As reported by the collectors. The seeds are black, flowers dark pink, leaves green. The RRC class type is: South American. In the greenhouse the leaves turned reddish as the plant matured. Observations from the Rodale Research Center, 1988 Rodale Amaranth Germplasm Catalog. Emmaus, PA.
	Accession unique id in USDA system	1051840

DOWNLOAD OCCURRENCES FROM FAO-WIEWS ([link](#))

WIEWS is the information system used by FAO for the preparation of periodic, country-driven global assessments of the status of conservation and use of PGRFA. WIEWS also monitors, on the basis of country reports, the implementation of the Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture, adopted in 2011.


**Food and Agriculture Organization
of the United Nations**

ENHANCED BY Google

العربية
 中文
 English
 Français
 Русский
 Español

WIEWS - World Information and Early Warning System on Plant Genetic Resources for Food and Agriculture

Home
 Background
 Data
Resources
 Glossary

Ex situ search

Accession-level information of plant genetic resources secured in genebanks (*ex situ*) under medium and long term storage can be retrieved through the search below.

Year
 2022

Country
 Country

Holding institute
 Holding institute

Current selected institute(s)
 Selection currently empty

Crop
 Crop

Crop Wild Relatives
 Excluded

Genus
 Genus

Species
 Species

Current selected element(s)
 Selection currently empty

☐ Among reported genera and species
 ☐ Among accepted genera and species

Accession number
 Accession number

DOI
 DOI

Country of origin
 Country of origin

Status under the Multilateral System
 Status under the Multilateral System

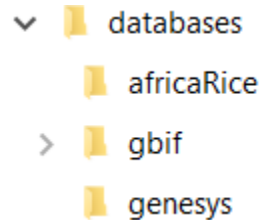
Biological status of accession
 Biological status of accession

Type of germplasm storage
 Type of germplasm storage

1. Genus's filter: type genus, automatically web portal will display the available genus (if nothing is displayed that means no data available for it).
2. Species filter: type one or more species, available species will be displayed (if nothing is displayed that means no data available).
3. Click to add genus and specie to the query
4. Displays the current selection of specie (usar can remove those species that do not match the query)
5. Select option **"300) Traditional cultivar/Landrace"**
6. Click on search
7. Download resulting table

PROTOCOL FOR DATABASES CONSOLIDATION - GAP ANALYSIS

1. Download each database and save them in separated folders, we encourage to name folders with the same name as the data source.



2. Apply the next filters for each one of the passport data we have downloaded (we encourage users to save this new filtered data table in a new file, preserving always the raw file):

a. GBIF:

- i. Make sure that in the species column only appears the specie we wanted.
- ii. Remove rows with missing decimalLatitude or decimalLongitude.
- iii. Keep rows that matches "PRESERVED_SPECIMEN" and "LIVING_SPECIMEN" from **basisOfRecord** column. Remove any row that does not match the criteria.
- iv. Keep rows that match "ACCEPTED" and "SYNONYM" from **taxonomicStatus** column. Remove any row that does not match the criteria.
- v. Create a new column named **"status"** from the **basisOfRecord** following the next rules: set **G** if **basisOfRecord** == **"LIVING_SPECIMEN"**, set **H** in any other case.
- vi. Create a new column named **"ISO3"**. From column **countryCode**, convert the existing ISO2 country code to ISO3 format.
- vii. Create a new column named **"COLLSITE"**. Concatenate columns **locality**, **municipality**, **country** and **stateProvince** into one single character string.
- viii. Create a new column named **"source_db"**. This column should have the name of the data source where it was downloaded.
- ix. Create a new column named **"database_id"**. This column should have a consecutive sequence of numbers that identifies each row from the data table.
- x. Keep the next columns: **ISO3**, **elevation**, **institutionCode**, **decimalLatitude**, **decimalLongitude**, **catalogNumber**, **COLLSITE**, **status**, **source_db**, **database_id**.
- xi. Rename columns as follows:
ORIGCTY = **ISO3**,
ELEVATION = **elevation**,
INSTCODE = **institutionCode**,
DECLATITUDE = **decimalLatitude**,
DECLONGITUDE = **decimalLongitude**,
ACCENUMB = **catalogNumber**

b. Gensys:

- i. Create a new column named **status**, fill this column with a fixed lab **"G"**.
- ii. Create a new column named **source_db**, fill this column with a fixed lab **"GENESYS"**.
- iii. Create a new column named **database_id**, This column should have a consecutive sequence of numbers that identifies each row from the data table.

- iv. Remove rows with missing **DECLATITUDE** or **DECLONGITUDE**.
 - v. Select rows that match “300” from **SAMPSTAT**, remove rows that does not match.
 - vi. Keep the next columns: “**SAMPSTAT**”, “**ORIGCTY**”, “**ELEVATION**”, “**DECLATITUDE**”, “**DECLONGITUDE**”, “**ACCENUMB**” and “**INSTCODE**”.
3. Once you have the databases, please check that they only have records of the target race, gender or class. In case of GBIF, for example, the database usually has records of the same crop but these are not always landraces or have the same class that you are looking for:

gbifid	species	infraspeci	taxonrank	countryco	locality	decimalla	decimallo	elevation	basisofrec	institution	address	lon.geoco	lat.geoco
910479696	Oryza sativa	NA	SPECIES	ES	Guadiana	38.95932	-6.63416	NA	PRESERVE	HSS	NA	NA	NA
910479701	Oryza sativa	NA	SPECIES	ES	Guadiana	38.95932	-6.63416	NA	PRESERVE	HSS	NA	NA	NA
912100508	Oryza sativa	NA	SPECIES	IN	Raigarh	NA	NA	NA	PRESERVE	K	Raigarh, IN	73.43333	18.25
912295775	Oryza glaberrima	NA	SPECIES	ML	French Su	NA	NA	NA	PRESERVE	K	French Su	-3.25756	14.77709
912295802	Oryza glaberrima	NA	SPECIES	SD	Casamanc	NA	NA	NA	PRESERVE	K	Casamanc	-16.7592	12.55611

In this case, although all the *Oryza* records were downloaded, we want to analyze just **glaberrima**, that is why every record different to *Oryza glaberrima* must be removed. The filtered database must be saved as a new file called “<db_name>_cleaned.csv” and the original database **MUST BE KEPT WITHOUT MODIFICATIONS OR FILTERS**.

4. **[IMPORTANT]** Add two new columns in each database:
 - **source_db**: Data source name, i.e. Genesys, gbif, CIAT, etc.
 - **database_id**: Identifier of the records in each database. It has an abbreviator of the data source name, then an underscore “_”, and a unique number for each record. The unique number can be the “accession number”.

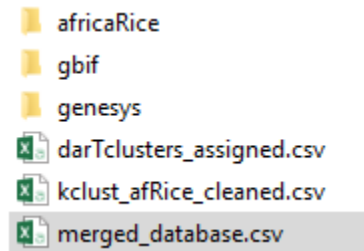
ecology	dartsec	dartsec	declat	declon	MAIN	Total.n	Numbe	Plant.h	Panicle	Leaf.bl	Leaf.bl	status	source_db	database_id
NA	NA	NA	14.77709	-3.25756	NA	NA	NA	NA	NA	NA	NA	H	gbif	gbif_1
NA	NA	NA	12.55611	-16.7592	NA	NA	NA	NA	NA	NA	NA	H	gbif	gbif_2
NA	NA	NA	10.9552	-2.6	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_3
NA	NA	NA	12.5	4.333333	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_4
NA	NA	NA	12.32528	-16.5478	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_5
NA	NA	NA	11.38333	5.5	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_6
NA	NA	NA	8.133333	-10.8833	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_7
NA	NA	NA	11.16667	4.666667	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_8
NA	NA	NA	10.4	14.85	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_9
NA	NA	NA	8.766667	-10.5	NA	NA	NA	NA	NA	NA	NA	G	gbif	gbif_10

5. Do not remove the columns for each downloaded database. As you know, GBIF database has a huge amount of variables that are usually unnecessary, however, do not remove them.

gbifID	datasetKey	occurrenceID	kingdom	phylum	class	order	family	genus	species
1840510370	b740eaa0-0679	http://www.botani	Plantae	Tracheop	Liliopsida	Poales	Poaceae	Oryza	Oryza sativ
1840516273	b740eaa0-0679	http://www.botani	Plantae	Tracheop	Liliopsida	Poales	Poaceae	Oryza	Oryza sativ
1840516392	b740eaa0-0679	http://www.botani	Plantae	Tracheop	Liliopsida	Poales	Poaceae	Oryza	Oryza sativ

belong to GBIF. However, in order to determine if the accession is **G** or **H**, please follow the next:

- a. Filter by **basisOfRecord** the accession different from "MATERIAL_SAMPLE" and "FOSSIL_SPECIMEN".
 - b. Add a new field/variable/column called **status** to the database by Filtering **basisOfRecord**; if this variable is equal to "LIVING_SPECIMEN" for a specific record, then the **status** of that accession must be **G**, otherwise **H**.
 - c. Make sure to preserve the column catalogNumber.
7. Merge the databases in one file, which must be saved in:
<root_path>/input_data/by_crop/rice_african/databases". Please include the word "**merged**" in the file name, in order to identify the final database.



In the above example, the file **merged_database.csv** is the final database created by the union of three different data sources: AfricaRice, GBIF and Genesys. In addition, there are two databases created from the final one: darTclusters_assigned and kclust_afRice_cleaned in which different African rice genetic structures are proved, hence, it is easy to identify which is the database to do the analysis from which other databases are created.