parseGBIF Manual

Pablo Hendrigo Alves de Melo[[1]](#footnote-1)

Nadia Bystriakova[[2]](#footnote-2)

Alexandre Monro[[3]](#footnote-3)

2023-06-27

# parseGBIF Manual

parseGBIF package is designed to convert [Global Biodiversity Information Facility - GBIF](https://www.gbif.org/) species occurrence data to a more comprehensible format to be used for further analysis, e.g. spatial. The package provides tools for verifying and standardizing species scientific names and for selecting the most informative species records when duplicates are available. The Manual provides a brief introduction to parseGBIF, with more information available from Help pages accessed via help(function\_name).

## Installation

You can install the development version of parseGBIF from [GitHub](https://github.com/pablopains/parseGBIF). To install parseGBIF, run

devtools::install\_github("pablopains/parseGBIF")

Please site parseGBIF as:

print(citation("parseGBIF"), bibtex = FALSE)

## To cite package 'parseGBIF' in publications use:  
##   
## Melo P, Bystriakova N, Monro A (2023). "ParsGBIF: An R package for  
## parsing species occurrence records." \_Methods in Ecology and  
## Evolution\_, \*1\*(11), 1-11. doi:doi..... <https://doi.org/doi.....>.

## Example

**parseGBIF makes it easy to get species occurrence records based on GBIF.**

### 1. GBIF data preparation

#### 1.1. Obtaining occurrence data of the species records from GBIF

1.1.1. Access a registered account in [GBIF](gbif.org)

1.1.2. Filter occurrences using available fields, for instance:

* Basis of record: *Preserved specimen*
* Occurrence status: *present*
* Scientific name: *Botanical family name* (e.g. Achatocarpaceae) or **filter by other fields**

1.1.3. Request to download information in **DARWIN CORE ARCHIVE FORMAT**

1.1.4. Download compressed file and unzip downloaded file

1.1.5. Use the **occurrence.txt** file as input to the prepare\_gbif\_occurrence\_data(gbif\_occurrece\_file = ‘occurrence.txt’) function

#### 1.2. Preparing occurrence data downloaded from GBIF

To prepare occurrence data downloaded from GBIF to be used by parseGBIF functions, run prepare\_gbif\_occurrence\_data.

library(parseGBIF)  
   
 occ\_file <- 'https://raw.githubusercontent.com/pablopains/parseGBIF/main/dataGBIF/Achatocarpaceae/occurrence.txt'  
   
 occ <- parseGBIF::prepare\_gbif\_occurrence\_data(gbif\_occurrece\_file = occ\_file, columns = 'standard')  
   
 head(occ)

## # A tibble: 6 × 54  
## Ctrl\_bibliographicCit…¹ Ctrl\_language Ctrl\_institutionCode Ctrl\_collectionCode  
## <chr> <chr> <chr> <chr>   
## 1 <NA> es Universidad de Anti… HUA   
## 2 <NA> es Universidad de Anti… HUA   
## 3 <NA> es Universidad de Anti… HUA   
## 4 <NA> es Universidad de Anti… HUA   
## 5 <NA> es Universidad de Anti… HUA   
## 6 <NA> es Universidad de Anti… HUA   
## # ℹ abbreviated name: ¹​Ctrl\_bibliographicCitation  
## # ℹ 50 more variables: Ctrl\_datasetName <chr>, Ctrl\_basisOfRecord <chr>,  
## # Ctrl\_informationWithheld <chr>, Ctrl\_dataGeneralizations <chr>,  
## # Ctrl\_occurrenceID <chr>, Ctrl\_catalogNumber <chr>, Ctrl\_recordNumber <chr>,  
## # Ctrl\_recordedBy <chr>, Ctrl\_georeferenceVerificationStatus <chr>,  
## # Ctrl\_occurrenceStatus <chr>, Ctrl\_eventDate <dttm>, Ctrl\_year <dbl>,  
## # Ctrl\_month <dbl>, Ctrl\_day <dbl>, Ctrl\_habitat <chr>, …

When parsing data, the user can choose between “standard” and “all” columns to be selected. The “standard” format has 54 data fields (columns), and the “all” format, 257 data fields (columns).

col\_standard <- parseGBIF::select\_gbif\_fields(columns = 'standard')  
   
 str(col\_standard)

## chr [1:54] "bibliographicCitation" "language" "institutionCode" ...

col\_all <- parseGBIF::select\_gbif\_fields(columns = 'all')  
  
 str(col\_all)

## chr [1:257] "gbifID" "abstract" "accessRights" "accrualMethod" ...

#### 1.3. Extracting GBIF issue

## EnumOccurrenceIssue

## Enumeration GBIF issue An enumeration of validation rules for single occurrence records.

### Description

There are many things that can go wrong and we continously encounter unexpected data. In order to help us and publishers improve the data, we flag records with various issues that we have encountered. This is also very useful for data consumers as you can include these issues as filters in occurrence searches. Not all issues indicate bad data. Some are merley flagging the fact that GBIF has altered values during processing. On the details page of any occurrence record you will see the list of issues in the notice at the bottom.

### Usage

data(EnumOccurrenceIssue)

### Format

A data frame with 69 rows and 9 columns

### Details

constant

GBIF issue constant

description

GBIF issue description

definition

Our definition for classifying geographic issues

type

Type issue

priority

Impact of the issue for the use of geospatial information

score

Impact, in number, of the issue for the use of geospatial information

selection\_score

Value used to calculate the quality of the geospatial information according to the classification of the issue

reasoning

Reasoning of the impact of the theme for the use of geospatial information

notes

Notes

### Source

* [GBIF Infrastructure: Data processing](https://www.gbif.org/article/5i3CQEZ6DuWiycgMaaakCo/gbif-infrastructure-data-processing)
* [An enumeration of validation rules for single occurrence records](https://gbif.github.io/gbif-api/apidocs/org/gbif/api/vocabulary/OccurrenceIssue.html)

data(EnumOccurrenceIssue)  
  
 colnames(EnumOccurrenceIssue)

## [1] "constant" "description" "definition" "type"   
## [5] "priority" "score" "selection\_score" "reasoning"   
## [9] "notes"

occ\_gbif\_issue <- parseGBIF::extract\_gbif\_issue(occ = occ)  
  
 names(occ\_gbif\_issue)

## [1] "occ\_gbif\_issue" "summary"

head(occ\_gbif\_issue$summary)

## issue n\_occ  
## 1 INSTITUTION\_MATCH\_FUZZY 1919  
## 2 GEODETIC\_DATUM\_ASSUMED\_WGS84 1883  
## 3 OCCURRENCE\_STATUS\_INFERRED\_FROM\_INDIVIDUAL\_COUNT 1336  
## 4 CONTINENT\_DERIVED\_FROM\_COORDINATES 1038  
## 5 COORDINATE\_ROUNDED 978  
## 6 TYPE\_STATUS\_INVALID 858

### 2. Check species names against WCVP database

The World Checklist of Vascular Plants (WCVP) database is available from the (Royal Botanic Gardens, Kew) [<https://powo.science.kew.org/about-wcvp>]. It can be downloaded to a folder of the user’s choice or into memory using get\_wcvp function. The output has 33 columns.

data(wcvp\_names\_Achatocarpaceae)  
 wcvp\_names <- wcvp\_names\_Achatocarpaceae  
   
 # wcvp\_names <- wcvp\_get\_data(read\_only\_to\_memory = TRUE)$wcvp\_names  
   
 colnames(wcvp\_names)

## [1] "plant\_name\_id" "ipni\_id"   
## [3] "taxon\_rank" "taxon\_status"   
## [5] "family" "genus\_hybrid"   
## [7] "genus" "species\_hybrid"   
## [9] "species" "infraspecific\_rank"   
## [11] "infraspecies" "parenthetical\_author"   
## [13] "primary\_author" "publication\_author"   
## [15] "place\_of\_publication" "volume\_and\_page"   
## [17] "first\_published" "nomenclatural\_remarks"   
## [19] "geographic\_area" "lifeform\_description"   
## [21] "climate\_description" "taxon\_name"   
## [23] "taxon\_authors" "accepted\_plant\_name\_id"   
## [25] "basionym\_plant\_name\_id" "replaced\_synonym\_author"  
## [27] "homotypic\_synonym" "parent\_plant\_name\_id"   
## [29] "powo\_id" "hybrid\_formula"   
## [31] "reviewed" "TAXON\_NAME\_U"   
## [33] "TAXON\_AUTHORS\_U"

Species’ names can be checked against WCVP database one by one or in a batch mode. To verify individual names, the function wcvp\_check\_name is used.

name.checked <- parseGBIF::wcvp\_check\_name(searchedName = 'Achatocarpus mollis H.Walter',  
 wcvp\_names = wcvp\_names,  
 if\_author\_fails\_try\_without\_combinations = TRUE)  
 name.checked[,c(3:5,22,23,40)]

## wcvp\_taxon\_rank wcvp\_taxon\_status wcvp\_family wcvp\_taxon\_name  
## 9 Species Accepted Achatocarpaceae Achatocarpus pubescens  
## wcvp\_taxon\_authors wcvp\_searchNotes  
## 9 C.H.Wright Updated

To check names in a batch mode, there is wcvp\_check\_name\_batch function. It uses the occurrence data (occ) and WCVP names list (wcvp\_names) generated in the previous steps.

names.checked <- parseGBIF::wcvp\_check\_name\_batch(occ = occ,  
 wcvp\_names = wcvp\_names,  
 if\_author\_fails\_try\_without\_combinations = TRUE,  
 wcvp\_selected\_fields = 'standard')

names(names.checked)

## [1] "occ\_wcvp\_check\_name" "summary"

head(names.checked$summary)

## wcvp\_plant\_name\_id wcvp\_taxon\_rank wcvp\_taxon\_status wcvp\_family  
## 4 500156 Species Accepted Achatocarpaceae  
## 22 500161 Form Accepted Achatocarpaceae  
## 15 500146 Species Accepted Achatocarpaceae  
## 9 500163 Species Accepted Achatocarpaceae  
## 2 500150 Species Accepted Achatocarpaceae  
## 16 500149 Species Accepted Achatocarpaceae  
## wcvp\_taxon\_name wcvp\_taxon\_authors  
## 4 Achatocarpus nigricans Triana  
## 22 Achatocarpus praecox f. obovatus (Schinz & Autran) Hauman  
## 15 Achatocarpus balansae Schinz & Autran  
## 9 Achatocarpus pubescens C.H.Wright  
## 2 Achatocarpus gracilis H.Walter  
## 16 Achatocarpus brevipedicellatus H.Walter  
## wcvp\_accepted\_plant\_name\_id wcvp\_reviewed  
## 4 500156 Y  
## 22 500161 Y  
## 15 500146 Y  
## 9 500163 Y  
## 2 500150 Y  
## 16 500149 Y  
## wcvp\_searchedName wcvp\_taxon\_status\_of\_searchedName  
## 4 Achatocarpus nigricans Triana <NA>  
## 22 Achatocarpus obovatus Schinz & Autran Synonym  
## 15 Achatocarpus balansae Schinz & Autran <NA>  
## 9 Achatocarpus pubescens C.H.Wright <NA>  
## 2 Achatocarpus gracilis H.Walter <NA>  
## 16 Achatocarpus brevipedicellatus H.Walter <NA>  
## wcvp\_plant\_name\_id\_of\_searchedName wcvp\_taxon\_authors\_of\_searchedName  
## 4 NA <NA>  
## 22 500158 Schinz & Autran  
## 15 NA <NA>  
## 9 NA <NA>  
## 2 NA <NA>  
## 16 NA <NA>  
## wcvp\_verified\_author wcvp\_verified\_speciesName wcvp\_searchNotes  
## 4 100 100 Accepted  
## 22 100 100 Updated  
## 15 100 100 Accepted  
## 9 100 100 Accepted  
## 2 100 100 Accepted  
## 16 100 100 Accepted

To bring species’ names into line with the format used by WCVP, the function standardize\_scientificName inserts space between the hybrid separator (x) and specific epithet, and also standardizes abbreviations of infrataxa (variety, subspecies, form).

# hybrid separator  
 parseGBIF::standardize\_scientificName('Leucanthemum ×superbum (Bergmans ex J.W.Ingram) D.H.Kent')

## $searchedName  
## [1] "Leucanthemum ×superbum (Bergmans ex J.W.Ingram) D.H.Kent"  
##   
## $standardizeName  
## [1] "Leucanthemum × superbum"  
##   
## $taxonAuthors  
## [1] "(Bergmans ex J.W.Ingram) D.H.Kent"  
##   
## $taxonAuthors\_last  
## [1] "D.H.Kent"

# variety   
 parseGBIF::standardize\_scientificName('Platymiscium pubescens subsp. fragrans (Rusby) Klitg.')

## $searchedName  
## [1] "Platymiscium pubescens subsp. fragrans (Rusby) Klitg."  
##   
## $standardizeName  
## [1] "Platymiscium pubescens subsp. fragrans"  
##   
## $taxonAuthors  
## [1] "(Rusby) Klitg."  
##   
## $taxonAuthors\_last  
## [1] "Klitg."

# subspecies

The funtion get\_lastNameRecordedB returns the last name of the main collector in recordedBy field.

# library(parseGBIF)  
 parseGBIF::collectors\_get\_name('Melo, P.H.A, Bystriakova, N. & Monro, A.')

## [1] "Melo"

parseGBIF::collectors\_get\_name('Monro, A.; Bystriakova, N. & Melo, P.H.A')

## [1] "Monro"

parseGBIF::collectors\_get\_name('Bystriakova, N., Monro, A.,Melo, P.H.A')

## [1] "Bystriakova"

### 3. Collectors Dictionary

To extract the last name of the main collector based on the recordedBy field and assemble a list relating the last name of the main collector and the raw data from the recordedBy, use the collectors\_prepare\_dictionary function. It uses the occurrence data (occ) generated in the previous step.

#### 3.1 Prepare dictionary collectors

collectorsDictionary.dataset <- parseGBIF::collectors\_prepare\_dictionary(occ = occ)

head(collectorsDictionary.dataset)

## Ctrl\_nameRecordedBy\_Standard Ctrl\_recordedBy  
## 1 ACEVEDO R. ACEVEDO R.; C. REYES  
## 2 ACOSTA R. ACOSTA P.; F. VÁZQUEZ B.  
## 3 ACOSTA R. ACOSTA P.; N. ACOSTA B.  
## 4 ADARVE J. ADARVE  
## 5 AGUILAR AGUILAR, R.M.  
## 6 AGUILAR JOSÉ AGUILAR CANO;SANDRA MEDINA;JHONATAN GUZMÁN  
## Ctrl\_notes collectorDictionary Ctrl\_update collectorName Ctrl\_fullName  
## 1 <NA> <NA> <NA> <NA>  
## 2 <NA> <NA> <NA> <NA>  
## 3 <NA> <NA> <NA> <NA>  
## 4 <NA> <NA> <NA> <NA>  
## 5 <NA> <NA> <NA> <NA>  
## 6 <NA> <NA> <NA> <NA>  
## Ctrl\_fullNameII CVStarrVirtualHerbarium\_PersonDetails  
## 1 <NA> <NA>  
## 2 <NA> <NA>  
## 3 <NA> <NA>  
## 4 <NA> <NA>  
## 5 <NA> <NA>  
## 6 <NA> <NA>

#### 3.2 Check the main collector’s last name

It is recommended to check the main collector’s last name in the nameRecordedBy\_Standard field. The goal is to standardize the main collector’s last name, which was automatically extracted from the recordedBy field, standardized in uppercase and with non-ascii characters replaced, so that a botanical collector is always recognized by the same last name.

If the searched recordedBy exists in the collector’s dictionary, the function retrieves the last name of the main collector referring to the recordedBy (in this case the CollectorDictionary field will be indicated with ‘checked’), otherwise, it returns the last name of the main collector, extracted automatically from the recordedBy field .

Once verified, the collector’s dictionary can be reused in the future.

file.collectorsDictionary.dataset <- 'file\_collectorsDictionary\_dataset.csv'  
  
 write.csv(collectorsDictionary.dataset,  
 file.collectorsDictionary.dataset,   
 row.names = FALSE,   
 fileEncoding = "UTF-8",   
 na = "")

#### 3.3 Update dictionary collectors

Create a key to group duplicates of a sample.

It also returns new collectors to be added to the collector dictionary that can be reused in the future.

occ\_collectorsDictionary <- parseGBIF::collectors\_update\_dictionary(occ=occ,  
 collectorDictionary\_checked\_file = file.collectorsDictionary.dataset)

names(occ\_collectorsDictionary)

## [1] "occ\_collectorsDictionary" "summary"   
## [3] "collectorsDictionary\_add"

head(occ\_collectorsDictionary$occ\_collectorsDictionary[,c(1,3)])

## # A tibble: 6 × 2  
## Ctrl\_nameRecordedBy\_Standard Ctrl\_key\_family\_recordedBy\_recordNumber  
## <chr> <chr>   
## 1 FONNEGRA ACHATOCARPACEAE\_FONNEGRA\_1629   
## 2 ROLDAN ACHATOCARPACEAE\_ROLDAN\_957   
## 3 FONNEGRA ACHATOCARPACEAE\_FONNEGRA\_1657   
## 4 BUNCH ACHATOCARPACEAE\_BUNCH\_   
## 5 DUQUE ACHATOCARPACEAE\_DUQUE\_4101   
## 6 TRUJILLO ACHATOCARPACEAE\_TRUJILLO\_5470

### 4. Select digital voucher

To group duplicates and choose the digital voucher:

1. If the key for grouping duplicates is complete with collector information and collection number, sample duplicates can be grouped. In this case, the voucher with the highest score is selected among the duplicates in the sample.
2. If the key to group duplicates is incomplete, sample duplicates cannot be grouped due to missing collector information and/or collection number. In this case, each record is considered a sample, without duplicates, and a voucher is selected for each sample.

#### How is the information score calculated?

moreInformativeRecord = sum of **textual quality** + **quality of geospatial information**.

#### How is the **quality of textual** information calculated?

##### The **Text quality** is the sum of the number of flags with text quality equal to TRUE.

Is there information about the collector?  
Is there information about the collection number?  
Is there information about the year of collection?  
Is there information about the institution code?  
Is there information about the catalog number?  
Is there information about the collection site?  
Is there information about the municipality of collection?  
Is there information about the state/province of collection?  
Is there information about the bibliographic citation?

#### How is the **quality of geospatial information** calculated?

##### The **quality of geospatial information** is based on geographic issues made available by GBIF.

GIBF issues on the quality of geospatial information were classified into three levels.

* Not applicable, with selection\_score equal to 0
* Does not affect coordinating accuracy, with selection\_score equal to -1
* Potentially affect coordinate accuracy, with selection\_score equal to -3
* Records to be excluded from spatial analysis, with selection\_score equal to -9

#### How is the taxonomic identification of the sample chosen?

1. When the key to group the duplicates is complete:

The accepted TAXON\_NAME identified at or below the specified level and the most frequent among the duplicates is chosen.

In case of a tie in frequency, in alphabetical order, the first accepted TAXON\_NAME identified up to or below the specific level is chosen.

If there is no identification, equal to or less than the specific level, for the sample, the sample is indicated as unidentified.

1. When the key to group the duplicates is incomplete:

If so, the accepted TAXON\_NAME identified at or below the specified level is used. If there is no identification, equal to or less than the specific level, the sample is indicated as unidentified.

occ\_digital\_voucher <- parseGBIF::select\_digital\_voucher\_and\_sample\_identification(occ = occ,  
 occ\_gbif\_issue = occ\_gbif\_issue$occ\_gbif\_issue,  
 occ\_wcvp\_check\_name = names.checked$occ\_wcvp\_check\_name ,  
 occ\_collectorsDictionary = occ\_collectorsDictionary$occ\_collectorsDictionary)

names(occ\_digital\_voucher)

## [1] "occ\_digital\_voucher\_and\_sample\_identification"  
## [2] "occ\_join\_results"

colnames(occ\_digital\_voucher$occ\_digital\_voucher\_and\_sample\_identification)

## [1] "Ctrl\_geospatial\_quality" "Ctrl\_verbatim\_quality"   
## [3] "Ctrl\_moreInformativeRecord" "Ctrl\_selectedMoreInformativeRecord"  
## [5] "Ctrl\_thereAreDuplicates" "Ctrl\_unmatched"   
## [7] "Ctrl\_unidentifiedSample" "Ctrl\_sampleTaxonName"   
## [9] "Ctrl\_matchStatusDuplicates" "Ctrl\_sampleIdentificationStatus"   
## [11] "Ctrl\_numberTaxonNamesSample" "Ctrl\_useful\_spatial\_analysis"

head(occ\_digital\_voucher$occ\_digital\_voucher\_and\_sample\_identification[,c(8,9,11,12)])

## Ctrl\_sampleTaxonName Ctrl\_matchStatusDuplicates  
## 1 Achatocarpus nigricans matched  
## 2 Achatocarpus nigricans matched  
## 3 Achatocarpus nigricans matched  
## 4 Achatocarpus nigricans unmatched: no recordNumber   
## 5 Achatocarpus nigricans matched  
## 6 Achatocarpus nigricans matched  
## Ctrl\_numberTaxonNamesSample Ctrl\_useful\_spatial\_analysis  
## 1 1 TRUE  
## 2 1 TRUE  
## 3 1 TRUE  
## 4 1 TRUE  
## 5 1 TRUE  
## 6 1 TRUE

file.occ\_digital\_voucher <- 'occ\_digital\_voucher.csv'  
  
 write.csv(occ\_digital\_voucher$occ\_join\_results,  
 file.occ\_digital\_voucher,   
 row.names = FALSE,   
 fileEncoding = "UTF-8",   
 na = "")

### 5. Export of results

Separate records into three data frames  
Export of results:

* Useful data for spatial and taxonomic analysis
* Data in need of revision of spatial information or without identification
* Duplicates of the previous two datasets

occ\_resuts <- parseGBIF::export\_data(occ\_digital\_voucher\_file = file.occ\_digital\_voucher )

names(occ\_resuts)

## [1] "occ\_in" "occ\_out\_to\_recover" "occ\_dup"

NROW(occ\_resuts$occ\_in)

## [1] 2472

NROW(occ\_resuts$occ\_out\_to\_recover)

## [1] 305

NROW(occ\_resuts$occ\_dup)

## [1] 849

file.in <- 'parseGBIF\_occ\_in.csv'  
 write.csv(occ\_resuts$occ\_in,  
 file.in,   
 row.names = FALSE,   
 fileEncoding = "UTF-8",   
 na = "")  
  
 file.occ\_out\_to\_recover <- 'parseGBIF\_occ\_out\_to\_recover.csv'  
 write.csv(occ\_resuts$occ\_out\_to\_recover,  
 file.occ\_out\_to\_recover,   
 row.names = FALSE,   
 fileEncoding = "UTF-8",   
 na = "")  
  
  
 file.dup <- 'parseGBIF\_occ\_dup.csv'  
 write.csv(occ\_resuts$occ\_dup,  
 file.dup,   
 row.names = FALSE,   
 fileEncoding = "UTF-8",   
 na = "")

1. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais, [pablopains@yahoo.com.br](mailto:pablopains@yahoo.com.br) [↑](#footnote-ref-1)
2. Natural History Museum, London, [n\_bystriakova@yahoo.com](mailto:n_bystriakova@yahoo.com) [↑](#footnote-ref-2)
3. Royal Botanic Gardens, Kew, [a.monro@kew.org](mailto:a.monro@kew.org) [↑](#footnote-ref-3)