ChecklistBank handleiding

Catalogue of Life

Versie 1, June 2023

GBIF | Global Biodiversity Information Facility

Inhoudsopgave

Beschrijving

ChecklistBank is a platform that supports the publication, analysis, and curation of checklists with emphasis in taxonomic and nomenclatural data. It is an integral part of the infrastructure developed and maintained by the Catalogue of Life (COL) and the Global Biodiversity Information Facility (GBIF).

ChecklistBank includes all checklists that form the Catalogue of Life, which are primarily global species lists provided by specific taxonomic communities. However, there are many other sources and uses for taxonomic checklists, such as national or local species lists, lists that share a common attribute such as red lists or invasive species lists, or lists that summarise taxonomic publications. Through ChecklistBank, we facilitate the discovery, use and citation of these valuable resources.

ChecklistBank is provided as a fundamental tool to ensure that basic data on species names and classifications can be shared and reused in support of the biological sciences and wider societal uses.

Currently, ChecklistBank provides a repository and a suite of tools to explore its content, and further examine checklists of particular interest. In this tutorial, we provide 4 exercises to explore these tools:

1. Explore the **ChecklistBank repository**: search, inspect and download checklists.
2. **Cross dataset search** tool: look up the appearance of a particular scientific name in all data sources available in ChecklistBank.
3. **Name match** tool: enables comparison of the COL Checklist with one or two other datasets in ChecklistBank in terms of taxon name matching.
4. **Dataset comparison** tool: allows for comparison of two taxonomic datasets in ChecklistBank on a scientific name by scientific name basis.

Audience

This tutorial is for anyone interested in discovering, curating and publishing checklists with emphasis in taxonomic and nomenclatural checklists.

ChecklistBank login

Use the following steps to login to ChecklistBank:

1. Go to https://www.checklistbank.org/
2. Click on Login (top right)
3. Login with your GBIF account

If you don't have a GBIF account yet, use the register at gbif.org now!
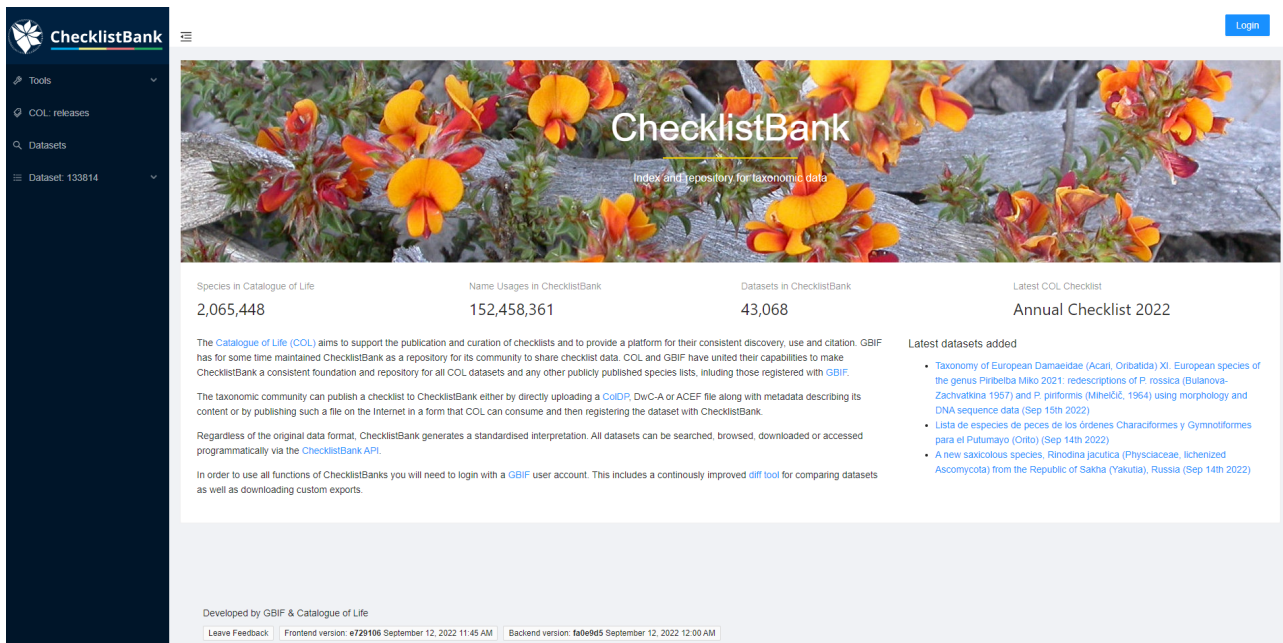
Explore the ChecklistBank repository

ChecklistBank is a public repository including over 40,000 checklists, established to simplify FAIR data sharing. It allows contributors to publish lists using a variety of typical data formats. Each list is then accessible through a standard API and reusable web browser components and can be cited using a ChecklistBank Digital Object Identifier.

Data publishers benefit both by making their datasets accessible for reuse and attribution and also through ChecklistBank tools for data review and detection of possible issues. Some of the datasets in ChecklistBank serve as authoritative sources for the Catalogue of Life Checklist, and new releases of the COL Checklist are also published as ChecklistBank datasets.

You can search through the ChecklistBank repository and explore the information and content of each checklist. All checklists can also be downloaded in multiple formats.

Before you start, please make sure you are logged in to ChecklistBank. If you don't know how to do that, follow the steps here.

Search for a checklist

Click on Datasets in the menu on the left. This will bring you to the dataset search page.
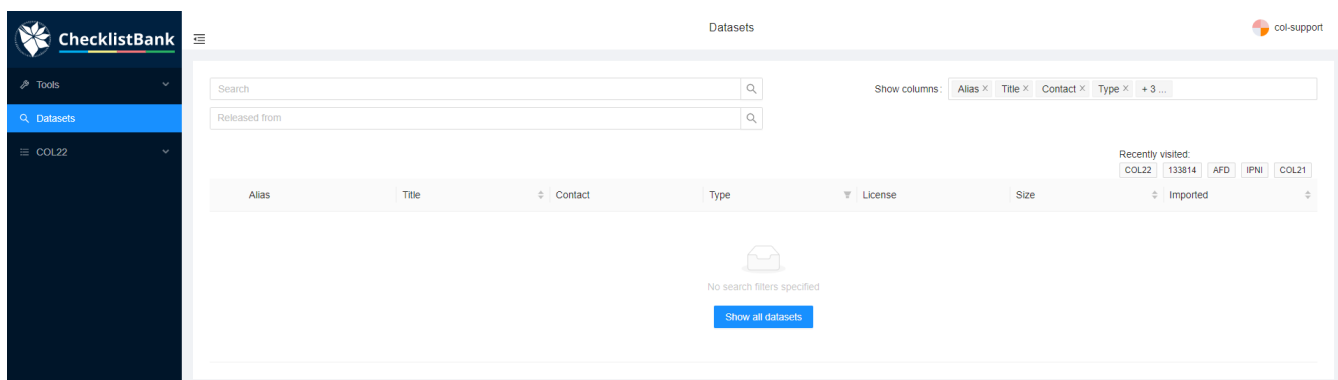
You can use the Search field to type any keyword of the checklist you are searching for. Here are a few suggestions that you can use for this tutorial:

- search 'COL22' : this will only give 1 result, the Catalogue of Life Checklist 2022.
- search 'Carabcat': the top result shows the Carabcat checklist, a global database of ground beetles, but additionally it will also show COL Checklists in which this checklist was included (since 2021).
- search 'Foraminifera': the top result shows the WoRMS Foraminifera checklist, below that the COL Checklists in which this checklist was included, and also a range of additional checklists that have been published in GBIF.

Explore a checklist

Let's have a closer look at the information you can find for a particular dataset. In this example, we use the Alucitoidea dataset by Donald Hobern: https://www.checklistbank.org/dataset/2207/about.

The menu on the left shows various options to explore this dataset:

About: here you can find all the metadata about this dataset including the author and contact person, the version of the dataset and date of when it was issued, licensing information and how to cite this source, link to the website of the dataset (if applicable), taxonomic scope, etc.

Download: on this page you can download this dataset in various formats:

- Darwin Core Archive (DwC-A): commonly used standard for biodiversity data.
- Annual Checklist Exchange Format (acef): data format previously used by Catalogue of Life.
- TextTree: a simple format to represent taxonomic trees using indented, plain text.
- Catalogue of Life Data Package (ColDP): this is a tabular text format with a standard set of files and columns recommended for exchanging checklists with the Catalogue of Life.
- Newick: a way of representing graph-theoretical trees with edge lengths using parentheses and commas, often used with phylogenetic data.
- Graphviz dot: a simple widely used format for representing graphs as nodes and edges. ChecklistBank exports will include synonym and basionym relations in the final graph that can be rendered with many software tools.

For your download you have the following options to select a subset of the checklist to be included in the download. This can be useful if you only want to download a selection of a large checklist.

- Choose root taxon: you can select a taxon within the checklist that will be the top level for the data included in your download. This means only taxonomic levels from this root level and further down are selected.
  For example, if we choose 'Cenoloba' here as root taxon, only this genus and the species within this genus would be included in the download.
- Exclude ranks below: select the highest taxonomic level you want to include in the download.
  For example, if we choose 'genus' all levels below will be excluded, thus species will not be listed.

The download consists of a zip file containing several tsv files of which the file 'NameUsage.tsv' includes the names in the checklist.

Browse: you can use this page to browse through the taxonomic tree for this particular checklist. This is useful to get a quick overview of which taxonomic levels and names are included in the checklist.

This is the view if you unfold the taxonomic tree for the Alucitoidea checklist:

Search: here you can search for specific taxonomic names. This functionality is most useful if you are looking for a particular name.

Imports: this shows the history and metrics of the import of this checklist into ChecklistBank including the different versions of the checklist. The timeline compares counts across imports, making it possible to see the growth in taxon counts, new datatypes added, etc.
Note: for the Catalogue of Life monthly and annual releases, the Imports is not available, because each release is a new checklist. For these checklists, there is just a Metrics page.

Issues: this lists the issues found by automatic checks. When you hover your mouse over a particular issue it will show the definition of that issue. A full list of all issues can be found here.
When you click on the count number it will take you to the search page displaying the names with that particular issue.

References: these are the references used to generate this checklist.

Verbatim: this shows the exact data as it was supplied without any alterations done by the importer to best fit the ChecklistBank data model. You can perform a full text search and browse the csv files. This page can be useful for data providers in debugging why ChecklistBank has interpreted things in a particular way, but also for others to search the original data and browse it.

Cross dataset search

The 'cross dataset search' tool gives an overview of the appearance of a particular scientific name in all data sources, taxonomic or nomenclatural, that are available in ChecklistBank.

A 'Taxonomic Name Usage' is a use of a particular scientific name in a particular resource. The resource can either be a dataset itself or it is any other source (e.g. human expert, literature reference) that is provided by a data custodian in the 'accordingTo' data standard term. The tool is helpful in understanding how a scientific name refers to different synonyms and hence a taxonomic concept that is used in a particular dataset or resource.

Please make sure you are logged in to ChecklistBank. If you don't know how to do that, follow the steps here.

You can find the tool in the menu on the left. Expand the Tools menu by clicking on the arrow. Then click 'Cross dataset search'.

You will find several options to improve your dataset search:

- Fuzzy: is an approximate string matching that matches two or more words even if there are typos or misspellings.
- Extinct: if selected match only within extinct taxa names.
- Matching:
    - Exact: compares the entire name string (without considering authors).
    - Words: matches word parts of the name, regardless of how they are combined.
    - Partial: a prefix search that matches names that begin with those letters.
- Restrict to:
    - Scientific name: search within the scientific names without authorship
    - Authorship: search within the scientific name
    - Any: search in all both scientific name and authorship

Example: *Quercus robur* L.
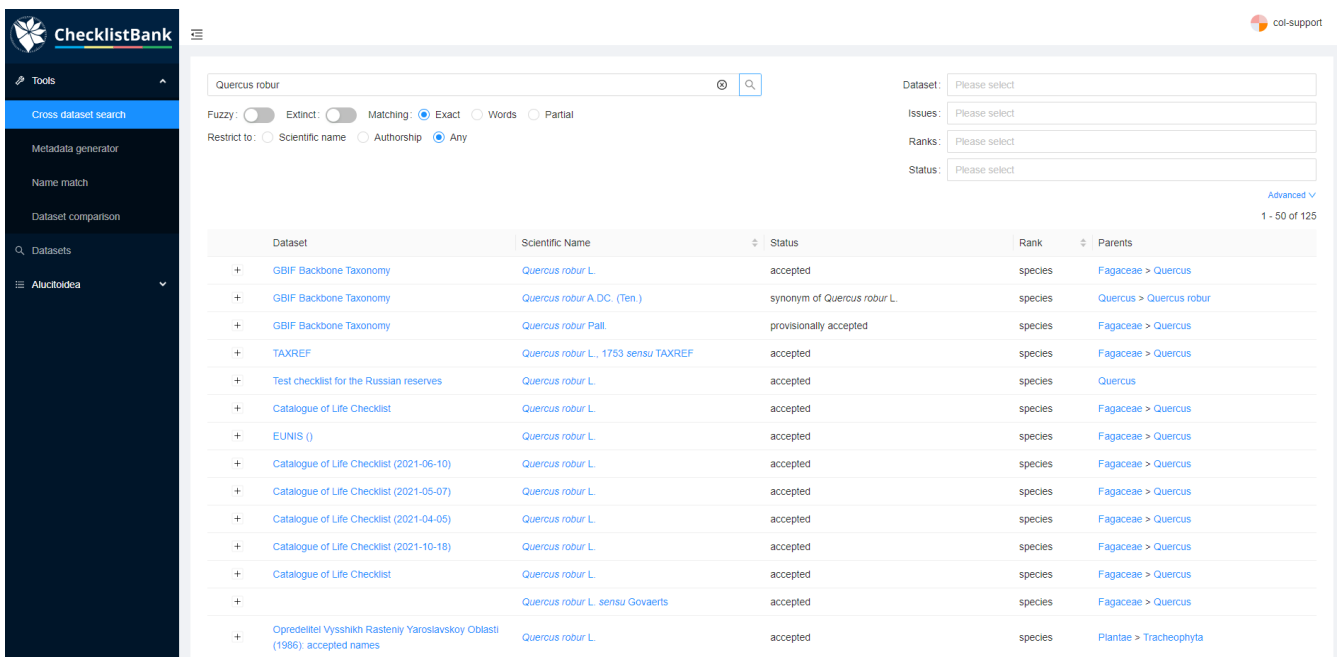
We showcase the functionality and usability of the 'cross dataset search' tool by using a real example of a search on a scientific name. *Quercus robur* L. is an oak tree belonging to the family of the Fagaceae with an extensive distribution across Europe.

Step 1: Type Quercus robur in the search bar.

Step 2: Select Exact in the options.

Setting the matching to 'Exact' ensures that all results relate to tree taxa in the family of the Fagaceae plant family. If the matching on 'Words' is used, zoological taxa (for example *Stomaphis quercus* (Linnaeus, 1758) from the Fauna Europaea dataset would also appear.

The search results show all datasets in ChecklistBank where the scientific name *Quercus robur* L. appears in.

Through these results, it is possible to explore individual datasets in ChecklistBank which include *Quercus robur* L. Several options for filtering the search results are also available.

Step 3: Select a Dataset.

In the menu 'Dataset' on the top right, you can easily select the dataset which you want to explore further. Click on 'International Plant Names Index (IPNI)', and then click on the Scientific Name (Quercus robur L.) in the results.

This brings you to the record of this name, which includes its original resource.

Step 4: Navigate external resources.

Click on the link of the Online resource.

This brings you to the information page about this name of the original source, in this case the International Plant Names Index (IPNI).

This example shows that by using the 'cross dataset search' tool, you can make linkages to other infrastructures to explore further evidence of a scientific name.

Exploring synonyms

You can also check whether the name that you have searched is used as a synonym for other names.

Go back to the search page in the 'cross dataset search' tool, and again search for Quercus robur, select Exact, and under Status, select 'Synonym'.

You can now see how may synonyms of *Quercus robur* are amongst the resources in ChecklistBank.

Name match tool

The 'name match' tool in ChecklistBank enables comparison of any checklist against one or two other datasets in ChecklistBank in terms of taxon name matching. There is a limit to matching a maximum of 6000 taxon names in one request.

Please make sure you are logged in to ChecklistBank. If you don't know how to do that, follow the steps here.

You can find the tool in the menu on the left. Expand the Tools menu by clicking on the arrow. Then click 'Name match'.
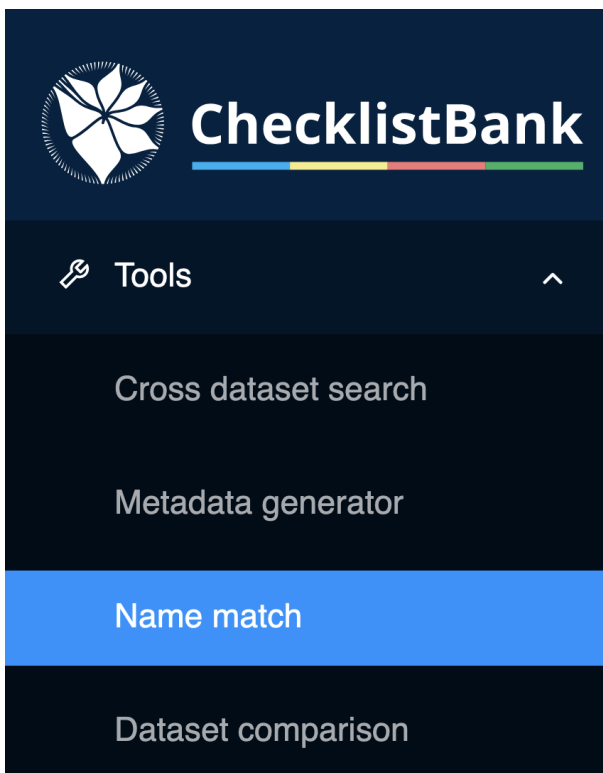
The user interface guides you through the steps to match your lists of interest.
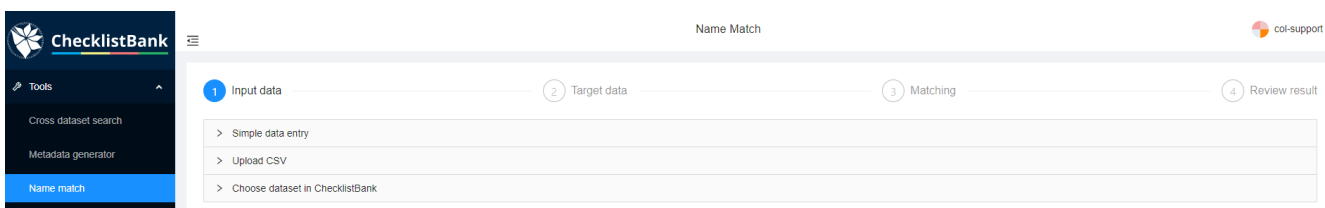
Step 1: Select your input data

There are three different options for your input data:

1. Simply paste a list of names in the text box. Use one name per row, and names can include author string.

2. Upload a csv file with a list of names. This csv file must contain at least a column with the header 'scientificName' (which may include the author) and optional columns 'author', 'rank', 'code'(nomenclatural code).

3. Choose a dataset in ChecklistBank. You can select any checklist from the ChecklistBank repository. If this is a large checklist, it might be helpful to select a root taxon. The matching is then done for all names below the root taxon, ignoring higher taxonomic levels.

Step 2: Select the target data

The default selection for the target data is the latest release of the COL Checklist. It is also possible to select a different checklist from ChecklistBank, or even match against two datasets.

Step 3: Matching

Once you click next in step 2, the matching will start. You can follow the progress in the progress bar. It is best to wait with exploring the results until the matching is complete.

Step 4: Review result

The results of the matchings can be explored in web pages in ChecklistBank or downloaded as a CSV file.

Example: Acacia Mill.

In this tutorial, we showcase the functionality of the name match tool with a comparison of the genus *Acacia* Mill. in the plant family Fabaceae.

Step 1. Select your input data

- Go to 'Choose a dataset in ChecklistBank'.
- In the 'Select a subject dataset' box, start typing and select 'WCVP-Fabaceae [2304]'.
  This is a dataset from the Legume Phylogeny Working Group 2021, which is a group of Fabaceae specialists that have worked on a comprehensive global taxonomic list for this taxon.
- Under 'And a root taxon', start typing and select 'Acacia (genus in Fabaceae)'
- Click on 'Fetch 3,297 names'.

Step 2. Select your target data

For this tutorial select the COL Checklist 2021 (COL21). - Type 'COL21' in the textbox and select 'COL21 Annual Checklist 2021 [2328]'.
- Click Next.

Step 3. Matching

The matching will start, please wait until it is complete.

Step 4. Review result

The first column lists the scientific name from the provided dataset (in this case the Fabaceae dataset).
The 'Scientific Name' column has a filter in which you can select whether you want to view the names that are also listed in the COL Checklist 2021 (usage in COL21), or the names that do not have a match (NO usage in COL21).

Dataset comparison

The 'dataset comparison' allows for comparison of two datasets in ChecklistBank on a scientific name by scientific name basis.

All datasets in ChecklistBank can be compared including the COL Checklist releases. The tool provides an overview of the differences in scientific names between datasets. The set of names to be compared can be filtered by various parameters, and a root taxon can also be selected to include only a certain group. Synonyms and names below a certain rank can be excluded on demand, i.e. you can just compare accepted genera or families.

Please make sure you are logged in to ChecklistBank. If you don't know how to do that, follow the steps here.

You can find the tool in the menu on the left. Expand the Tools menu by clicking on the arrow. Then click 'Dataset comparison'.
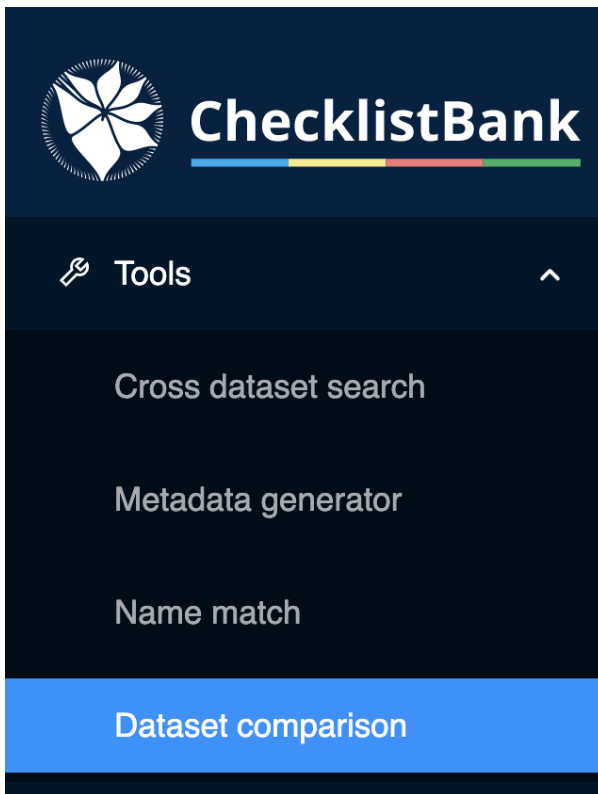
Example: *Lysiloma* (genus in Fabaceae)

We use the Fabaceae checklist, which was also used in the 'name match tool' tutorial. We compare this list to the 2021 COL Checklist for which the data on the genus *Lysiloma* are provided by the ILDIS World Database of Legumes checklist (Roskov et al. 2014).

When you open the tool, the start screen of the dataset comparison allows you to choose two datasets to make the comparison on the basis of scientific names.

Step 1. Select dataset 1

- Start typing WCVP-Fabaceae and this dataset will appear in the menu. Select it.
- In the box 'Find taxon', type 'Lysiloma' and select it.

Step 2. Select dataset 2

- Type COL 2021, and select the COL21 Annual Checklist 2021.
- Click on 'Go to Lysiloma' to select the same genus as in dataset 1.
- Click on 'Show diff'

Step 3. Check you selection

The next view shows you which datasets and roots you have selected.
You can choose whether the scientific name comparisons should include authorship and synonyms, for this example select both options.
Additionally, you can select whether to show the parent, and at which level. For this example, also select the box 'Show parent'.

Step 4. Get results*

Once you have made your selection, click 'Get diff'.

The results of the comparison can be explored in the webpage.

You can see two columns, one for each dataset. Scientific name differences are shown in each column where a '-' or '+' sign indicates in which dataset the difference occurs. The results of the comparisons can also be downloaded in a CSV file for further investigation.

Glossary

API

Application Programming Interface. A set of clearly defined methods of communication between various software components.

CC Licences

Creative Commons. These are a series of licenses set up by the Creative Commons organization that enable sharing and reuse of creativity and knowledge through the provision of free legal tools.

COL

Catalogue of Life. A global collaboration bringing together the effort and contributions of taxonomists and informaticians from around the world.

Controlled Vocabulary

This is a restricted set of terms that are used as possible values for a given field. One can think of it as a lookup list or dropdown for a particular field. For example the DwC field basisOfRecord should only contain one of these values: "PreservedSpecimen", "FossilSpecimen", "LivingSpecimen", "HumanObservation", "MachineObservation". We would say that list of values is a controlled vocabulary for that field.

DOI

Digital Object Identifier. A persistent identifier or handle used to uniquely identify objects. DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports and data sets, and official publications.

DwC

Darwin Core is a biodiversity data standard, maintained by TDWG & widely used within the GBIF community and partners. It is a set of standardized terms (or field names) and their definitions, which are used to share biodiversity information.

DwC-A

Darwin Core Archive. A compressed (zipped) file containing all the information needed to share

with GBIF, for a particular resource. Each zip contains three types of files:

1. the actual data, in one or more text files: occurrence.txt/event.txt/measurmentoffact.txt etc

2. a mapping file: rtf.xml

3. a metadata (EML) file: eml.xml When you publish using the IPT, it creates a Darwin Core Archive, which is shared with GBIF. Also, when you download data from the GBIF website you can choose a DwC-A format as well.

GUID

Globally Unique IDentifier

Colophon

Suggested citation

Authors

Chantal Huijbers, Olaf Bánki, Markus Döring, Thomas Stjernegaard Jeppesen, Diana Hernández, Camila Plata and Joe Miller.

Contributors

Licence

Persistent URI

Document control

First edition, June 2023