Web visualisation of mappings between CoL and other checklists

Deliverable D10.4

29 April 2022

Olaf Bánki*, Joe Miller#, Donald Hobern*, Thomas Stjernegaard Jeppesen#, and Markus Döring#

** Catalogue of Life / Species 2000, Leiden, The Netherlands*
 # Global Biodiversity Information Facility, Copenhagen, Denmark

BiCIKL

BIODIVERSITY COMMUNITY INTEGRATED KNOWLEDGE LIBRARY

This project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Start of the project:	May 2021
Duration:	36 months
Project coordinator:	Prof. Lyubomir Penev Pensoft Publishers
Deliverable title:	Web visualisation of mappings between CoL with other Checklists
Deliverable n°:	D10.4
Nature of the deliverable:	Report + associated tools
Dissemination level:	Public
WP responsible:	WP10
Lead beneficiary:	Species 2000
Citation:	Bánki, O., Miller, J., Hobern, D., Stjernegaard Jeppesen T., & Döring, M. (2022). *Web visualisation of mappings between COL and other checklists*. Deliverable D10.4 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 12
Actual submission date:	29 April 2022

Deliverable status:

Version	Status	Date	Author(s)
1.0	Draft	8 April 2022	Bánki, O., Miller, J., Hobern, D., Stjernegaard Jeppesen, T., and Döring, M.
1.1	Review	18 April 2022	Agosti, D., Arvanatidis, C., and Penev, L.
2.0	Final report	28 April 2022	Bánki, O., Miller, J., Hobern, D., Stjernegaard Jeppesen, T., and Döring, M.

Table of contents

Preface

Taxonomy is a fundamental resource to align and interpret biodiversity data. Historically taxonomic data has been splintered by discipline and the global community has not been able to produce a unified taxonomic resource for interpreting other biodiversity data. This fault has severely limited the integration of biodiversity data and its use in research and policy to improve our global knowledge.

Catalogue of Life (and Species2000) has now built a community of more than 500 global experts that are responsible for vetted taxonomic resources that are used to interpret biodiversity data. It is recognized as the most comprehensive resource for taxonomy. A recent collaboration with GBIF has upgraded their joint infrastructure to now allow greater impact.

Biodiversity Community Integrated Knowledge Library (BiCIKL) is an EU funded project that recognized taxonomy as one of four pillars for work. The other pillars are DNA sequences, literature and specimens. The joint COL- GBIF infrastructure is a starting point to improve collaborative taxonomic functionality to hasten and improve biodiversity data integration.

BiCIKL Work Package 10 (WP10) focuses on FAIR data improvements to all stages within the process to harvest, aggregate and curate taxonomic information from publications and genomic research, in order to accelerate standardised access to streams of new or digitised Linnean or molecular classifications and treatments. These data are being mapped in Catalogue of Life to enable expert community curation of the derived products and encapsulate those in services and visualisations that support the infrastructures of all BiCIKL partners and other users.

IIn particular WP10 builds upon the COL-GBIF infrastructure by developing and implementing tools that will help to establish linkages among BiCIKL partner databases and their infrastructures. The present report outlines the advances performed especially under Task 10.4 Delivery and presentation of taxonomic information, but also displays some advances relevant to the Task 10.1, namely the automatic ingestion of newly published taxonomic information into the COL-GBIF developed ChecklistBank; Task 10.2 Data mapping for taxonomic information; and Task 10.3 Curation and validation of taxonomic information.

Summary

The deliverable 10.4 'web visualisation of mappings of the Catalogue of Life (COL) Checklist with other checklists' is aimed at delivering a means for users and editors of the COL Checklist to better understand differences in the use of taxonomic names between data sources. As part of the deliverable Catalogue of Life/Species 2000 and the Global Biodiversity Information Facility roll out three tools: a taxonomic name usage search, name match, and a difference viewer. These tools are particularly aimed at supporting the BiCKIL user community.

The taxonomic name usage search, name match, and difference viewer are embedded in ChecklistBank. ChecklistBank is an integral part of the COL and GBIF infrastructures, and is continuously improved as part of the COL and GBIF collaboration. ChecklistBank is aimed at providing a data repository for taxonomic and nomenclatural data sources, and provides a set of integrated tools for assembling the COL Checklist. In BiCIKL, improvements on ChecklistBank will be used to deliver taxonomic data services to the BiCIKL user community and the biodiversity data infrastructure partners.

List of abbreviations

BHL	Biodiversity Heritage Library
BiCIKL	Biodiversity Community Integrated Knowledge Library
COL	Catalogue of Life
ColDP	Catalogue of Life Data Package
DwC-A	Darwin Core Archive
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GBIF	Global Biodiversity Information Facility
ITIS	Integrated Taxonomic Information System
IPNI	International Plant Names Index
WCSP	World Checklist of Selected Plants
WCVP	World Checklist of Vascular Plants
WoRMS	World Register of Marine Species
WWW	World Wide Wattle

1. ChecklistBank

The following sections describe the ChecklistBank shared infrastructure that COL and GBIF have developed in collaboration. ChecklistBank will also take on a central role in BiCIKL where it concerns taxonomic names services to infrastructure partners and the BiCIKL user community.

1.1. An international collaboration

1.1.1. Catalogue of Life

The most comprehensive source on names and classification of species and higher rank taxa
Species 2000, Leiden, The Netherlands
https://www.catalogueoflife.org/

Catalogue of Life (COL) is an international collaboration bringing together the effort and contributions of taxonomists and informaticians from around the world. COL aims to address the needs of researchers, policy-makers, environmental managers and the wider public for a consistent and up-to-date listing of all the world's known species and their higher taxa. The COL Checklist is a consensus classification, based on the underlying taxonomic source databases, managed by a community of more than 500 experts (Costello et al. 2022). The higher taxa are partially based on a management hierarchy. COL also supports those who need to manage their own taxonomic information and species lists.

1.1.2. Global Biodiversity Information Facility

The world's most comprehensive source of primary biodiversity data
GBIF, Copenhagen, Denmark
https://www.gbif.org/

GBIF - the Global Biodiversity Information Facility - is an international network and data infrastructure funded by the world's governments and aimed at providing open access to data about all types of life on Earth to anyone and anywhere in the world.

Coordinated through its Secretariat in Copenhagen, the GBIF network of participating countries and organisations, working through participant nodes, provides data-holding institutions around the world with common standards and open-source tools that enable them to share information about where and when species have been recorded. This knowledge derives from many sources, including everything from, say, museum specimens collected in the 18th and 19th century to geotagged smartphone photos shared by amateur naturalists in recent days.

1.1.3. Alliance for biodiversity knowledge

GBIF convenes the *alliance for biodiversity knowledge* (the *alliance*), a lightweight umbrella framework mandated by major biodiversity-related data infrastructures to maximise impact of FAIR data in research and policy. The Alliance for biodiversity knowledge aligns efforts to

deliver current, accurate and comprehensive data, information and knowledge on the world's biodiversity.

This *alliance* is open to all institutions, agencies, organisations, researchers and communities working to measure and monitor biodiversity or dependent on accurate information on biodiversity. By joining forces, every stakeholder will benefit from free and open access to the best possible evidence to address questions at all scales.

The COL - GBIF infrastructure collaboration is an initial and exemplar collaboration under the umbrella of the *alliance*. As described below the collaboration has been successful in building a joint infrastructure used by both entities and is ready to be used by outside organisations with a goal of providing a global taxonomic resource. The collaboration has been successful in meeting the alliance's vision: support for science and evidence-based planning, support for open data and open science, support for highly-connected biodiversity data and support for international collaboration.

This vision is shared by the BiCIKL coalition as the project is centred around the FAIR principles. The COL - GBIF infrastructure collaboration is broadened to now include all BiCiKL partners as taxonomy is one of the four pillars of the BiCiKL project.

1.2. The COL Checklist

1.2.1. A global species checklist resource

The COL Checklist is assembled based on validated taxonomic data sources following a set of criteria for measuring progress (Hobern et al. 2021). The COL Checklist contains more than 2 million species, both living and extinct. The March 2022 release of the COL Checklist is based on 164 data sources underpinned by a network of more than 500 experts from around the world (Bánki et al. 2022). Most of these data sources are updated on a regular basis, and involve active editing communities.

Data sources need to be converted into one of the existing data standards (e.g. DwCA, ColDP) before these sources can be published in ChecklistBank. The COL consortium of partners assist in converting data into appropriate formats, and do supply in some cases automated feeds of data for the COL Checklist (e.g. ITIS, WoRMS). Editors of the Catalogue of Life populate taxonomic sectors in the COL Checklist based on the available data sources. During this process the editors apply editorial decisions on the data, e.g. like blocking names in the case of duplicates. The COL Checklist uses a management classification. The taxonomic communities that deliver data for a taxonomic sector in the COL Checklist determine the point of attachment in the classification. The COL contributors to a specific taxonomic sector approve their data in the COL Checklist before it becomes part of a formal release.

The COL Checklist faces a series of content challenges that should be addressed together with the taxonomic community and consortium partners in the coming years. For some taxonomic sectors no (active) taxonomic communities exist that can deliver or continuously update a global species checklist. Addressing these taxonomic gaps is one of the main purposes of the COL taxonomy group. In addition, in order for the COL Checklist to be useful for various biodiversity data infrastructures and initiatives, the pool of scientific names should be increased. A further challenge is to understand whether data sources are based on the same nomenclatural foundation. Barriers to update the COL Checklist with newly published

species from literature as well as barriers to update existing taxonomic data sources should be addressed as best as possible. Visualisation of mappings of the COL Checklist with other taxonomic checklists is one of the means that contribute to the overall usefulness of the COL Checklist. It can also assist in reviewing available taxonomic data sources with the data sources that are used to assemble the COL Checklist.

1.2.2. Persistent name identifiers

With the migration to the new Catalogue of Life infrastructure in December 2020, Catalogue of Life has also switched to a new algorithm to generate stable identifiers for name usages. Up until the 2019 annual COL Checklist a simple hashing of names has been applied to make sure the IDs between COL Checklist releases do not change. This resulted in name based identifiers that did change whenever a single character of the name or its authorship was altered. The new implementation is also a name based system, but tries to keep the identifiers stable if the authorship of a name has only been slightly modified. It forces a change in identifiers when an authorship is added. Status or parent/classification changes do not trigger any ID changes. So when names change status from an accepted name to a synonym or vice versa, there is no change in the ID. By combining a name identifier and the data set key the user would have a stable reference to a name in a particular release of the COL Checklist.

In case there have been multiple previous IDs for the same name we prefer the ID from the oldest release to keep stability. This can for example happen when we later realise a genus was included twice in the COL Checklist because it was added to the Checklist from different data sources at different positions in the hierarchy - something we internally refer to as a split genus.

The new identifiers try to be short and readable, avoiding characters that can easily be confused. Because they are based on a set of 29 alphanumeric characters we call the encoding LATIN29[1]. By preventing the use of vowels we also avoid most real words and potentially offensive meanings in various languages. For the COL Checklist with currently ~4.2 million name usages, the identifiers have a maximum length of 5 characters. We have reserved single character identifiers for kingdoms (except for viruses) and manually assigned these. For example the letter P is used for Plantae[2]. For a species identifier we use for example 4QHKG for Puma concolor[3].

The existence of stable name identifiers in the Catalogue of Life Checklist enable the visualisation of mappings of the COL Checklist with other taxonomic checklists on the basis of name usages and differences in names between taxonomic data sources.

1.3. The launch of ChecklistBank

In December 2020, prior to the start of the BiCIKL project, the new Catalogue of Life infrastructure in collaboration with GBIF was launched. This infrastructure consists of three parts. First is a public portal (https://catalogueoflife.org/) that facilitates access to the monthly updated COL Checklists, its underlying taxonomic databases, and general information on COL. The second component is ChecklistBank (https://www.checklistbank.org/), which is a

[1] https://github.com/CatalogueOfLife/backend/issues/491
[2] https://www.catalogueoflife.org/data/taxon/P
[3] https://www.catalogueoflife.org/data/taxon/4QHKG

data repository that facilitates access to original data sources underlying the COL Checklist, all COL Checklist releases, all GBIF taxonomic checklists, and assembly tooling for the COL Checklist. ChecklistBank tools will be publicly available for future users to build taxonomic backbones with resources publicly held within it. Thirdly, the infrastructure includes a set of APIs (https://api.catalogueoflife.org/) to render all COL Checklist data to ChecklistBank, the COL portal and users, provide persistent name and digital object identifiers, and support various data standards (e.g. DwC-A, Catalogue of Life Data Package).

1.3.1. A data repository

ChecklistBank is a high-functionality public repository and portal established to simplify FAIR data sharing for taxonomic and nomenclatural lists. 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 sections of the Catalogue of Life Checklist, and new releases of the COL Checklist are also published as ChecklistBank datasets.

All datasets can be downloaded in multiple formats and accessed via a consistent API. Aggregating taxonomic and nomenclatural lists through a common portal makes it possible for users to locate sources offering differing perspectives on nomenclature and taxonomy.

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.

1.3.2. Catalogue of Life Assembly Tooling

A primary function for ChecklistBank is to serve as a rich editing environment for the construction and management of complex taxonomic data products, including the COL Checklist. Integrated tools enable an editor to construct a project that combines components from multiple data sources and applies rule-based decisions (e.g. blocking of names, selection of taxonomic sectors from data sources) to construct a new dataset. This infrastructure now serves as the platform used each month to produce new releases of the COL Checklist. These tools are also made available for other projects to manage construction of taxonomic lists (part of BiCIKL deliverable 10.3). Embedding these functions in ChecklistBank will promote reuse of scientific name records and other data and allow new projects to benefit from the efforts of COL and others to organise data.

The tools developed for the BiCIKL user community and described in this document will benefit a wide range of users, but also contribute to the tool set available to COL editors and other projects using ChecklistBank as an assembly platform.

1.3.3. Technical implementation

ChecklistBank is an open-source project with multiple repositories hosted in GitHub[4]. The back-end[5] is implemented in Java as a Dropwizard application that drives the COL ChecklistBank API. The front-end[6] is a React user interface application that uses the ChecklistBank API and supports public exploration of all data in ChecklistBank. It also includes (for appropriately authorised users) the tools for assembling taxonomic checklists from multiple sources.

1.4. ChecklistBank implementation in BiCIKL

Species 2000 together with GBIF leads BiCIKL's work package 10 to deliver high-quality virtual access to the taxonomic framework for use by the research infrastructures involved in the BiCIKL project and strengthen linkages with the taxonomic community and taxonomic publishers to ensure the quality of this framework and its trustworthiness. The workflows developed in the Joint Research Activities will be implemented together with GBIF in the new Catalogue of Life (COL) research infrastructure, and especially through ChecklistBank. The BiCIKL project activities will be centred around the existing and ongoing developments by the COL-GBIF collaboration of ChecklistBank.

[4] https://github.com/CatalogueOfLife
[5] https://github.com/CatalogueOfLife/backend
[6] https://github.com/CatalogueOfLife/checklistbank

2. Tools of Deliverable 10.4

The existing and ongoing development of ChecklistBank by Catalogue of Life and GBIF, provide an ideal foundation for further tool development. For the BiCIKL user community tools have been developed that contribute to the web visualisation of mappings between the COL Checklist and other data sources. Three tools are presented as part of the deliverable 10.4.

2.1. Name usage search

2.1.1. Description of the tool

The 'taxonomic name usage search' gives an overview of the appearance of a particular scientific name in all data sources, taxonomic or nomenclatural, 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[7] that is used in a particular dataset or resource. The 'taxonomic name usage search' can be accessed in ChecklistBank[8]. The tool is in production and usable for the wider public, without login into ChecklistBank.

2.1.2. An example: *Quercus robur* L.

The functionality and usability of the 'taxonomic name usage search' can be best conveyed by providing a real example of a search on a scientific name. We have chosen a plant example to do this, but a zoological scientific name would work equally well. *Quercus robur* L. is an oak tree belonging to the family of the Fagaceae with an extensive distribution over Europe.

The scientific name of *Quercus robur* L. is typed into the search field of the 'taxonomic name usage search'. The search results show all datasets in ChecklistBank where the scientific name *Quercus robur* L. appears in (Figure 1). 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 through the advanced search.

Looking into one of the data sources, the 'GBIF Type Specimen Names' dataset, reveals a web page for an occurrence record for the scientific name *Quercus robur* L. (Figure 2) which provides details of the type specimen (a lectotype) for the species in GBIF.org (Figure 3). So by using the 'taxonomic name usage search' a user can make linkages to a biodiversity data infrastructure such as GBIF, to explore further evidence of specimen records of a scientific name as deposited in a natural history museum.

[7] see also the definition used by TDWG, http://rs.tdwg.org/tnu/terms/TaxonomicNameUsage).

[8] The taxonomic name usage search tool can be accessed in ChecklistBank through https://www.checklistbank.org/nameusage/search

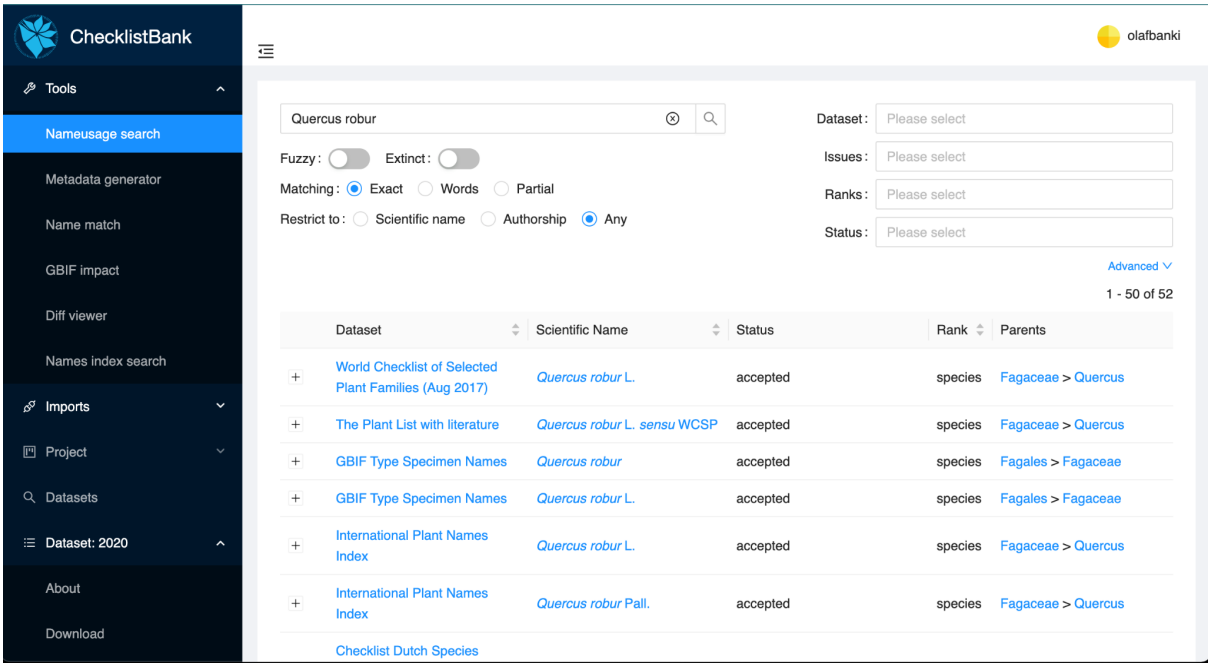


Fig. 1: Taxonomic Name Usage Search: Quercus robur L.

A screenshot of the search results of the 'taxonomic name usage search' on *Quercus robur* L.; note that this scientific name appears in 52 datasets in ChecklistBank (this includes all COL Checklist releases). The matching is set to be 'Exact'. This matching 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 also appear ; see https://www.checklistbank.org/dataset/2026/taxon/56093).

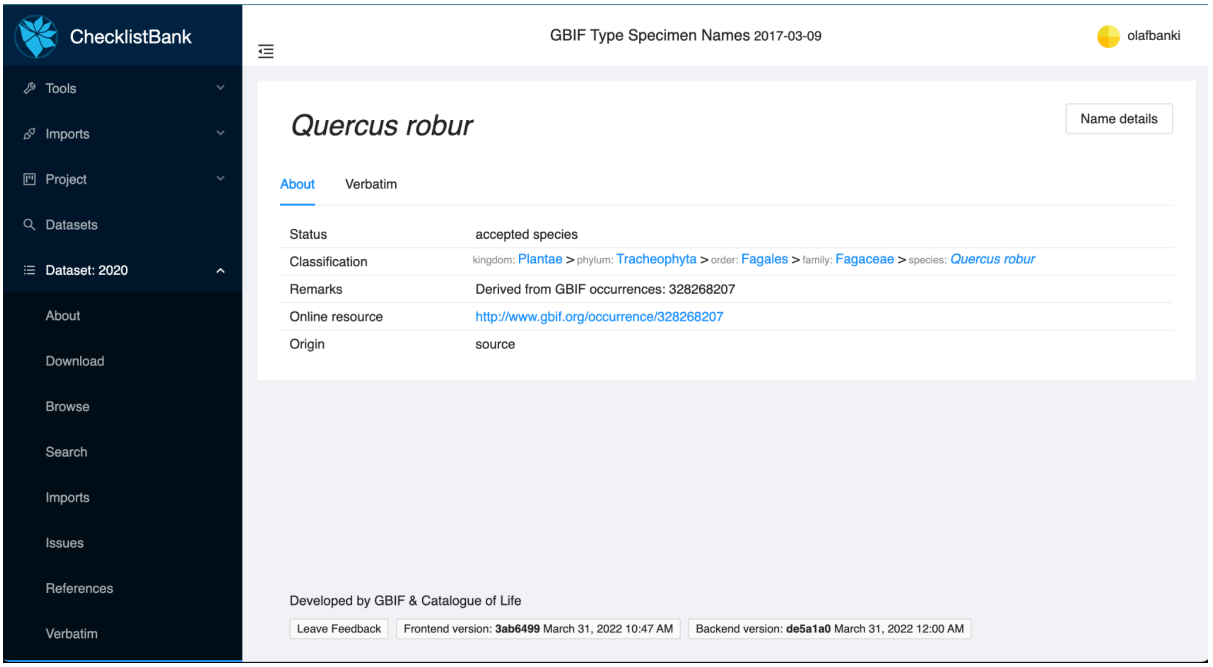


*Fig. 2: **Quercus robur** in GBIF Type Specimen Names*

A screenshot showing how *Quercus robur* L. occurs in the 'GBIF Type Specimen Names' dataset in ChecklistBank. A link is provided to the type specimen record in GBIF.org: http://www.gbif.org/occurrence/328268207

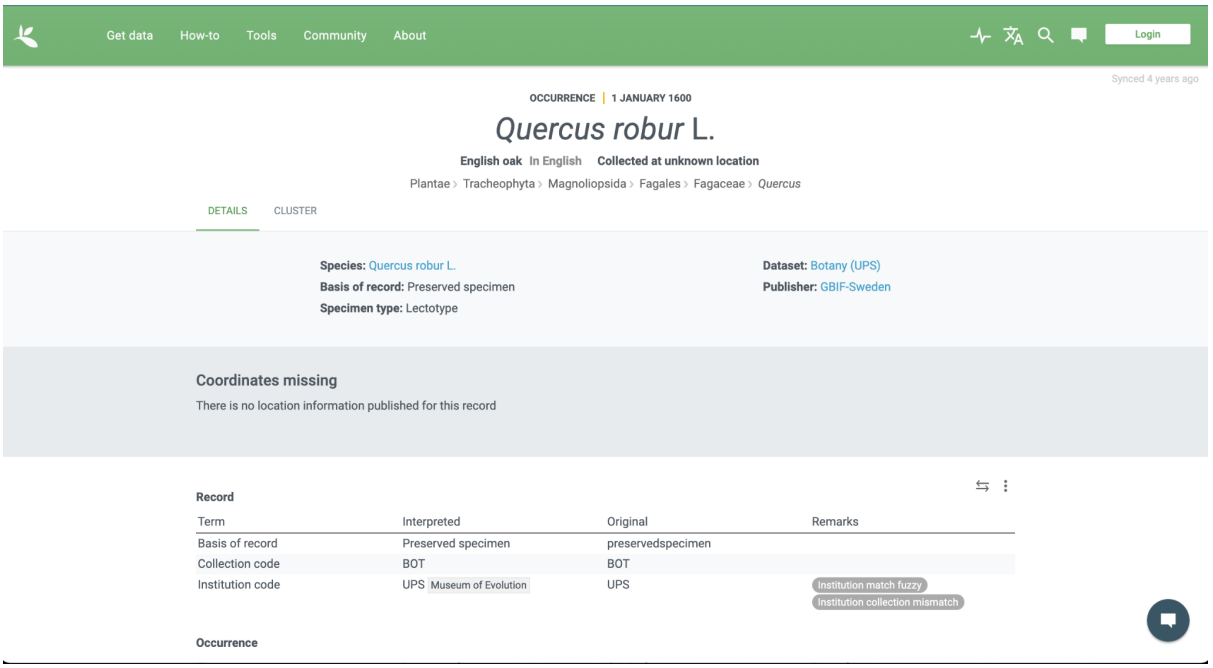


Fig. 3: Type specimen of Quercus robur L. in GBIF.org

A screenshot of GBIF.org that shows a lectotype of *Quercus robur* L. from a specimen record collected in 1600 in Passim and deposited in the Museum of Evolution in Uppsala (http://www.gbif.org/occurrence/328268207).

The search results on *Quercus robur* L. of the 'taxonomic name usage search' are at the species rank. Almost all of the accepted species instances refer to *Quercus robur* L.; except for a few. The scientific name of *Quercus robur* Pall. is only found in the International Plant Names Index (IPNI), a plant nomenclator dataset (Figure 4).

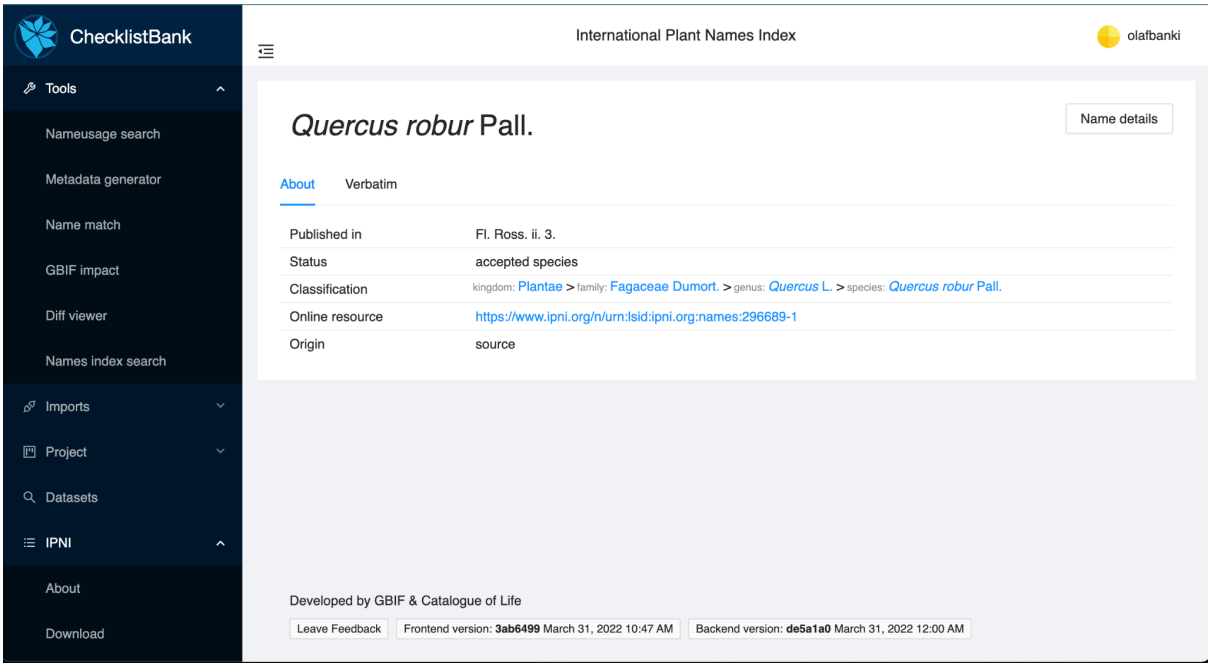


*Fig. 4: **Quercus robur** Pall. in IPNI*

A screenshot of the scientific name *Quercus robur* Pall. that is part of the IPNI dataset in ChecklistBank. Note that a link is provided that takes the user to IPNI: https://www.ipni.org/n/urn:lsid:ipni.org:names:296689-1

Exploration of the verbatim record of *Quercus robur* Pall. reveals the data is interpreted during the processing of the data source in ChecklistBank as 'nomenclature invalid' because its status could not be interpreted (Figure 5). Links to IPNI are made. In the case of *Quercus robur* L. digital access to the original publication in BHL can be found through IPNI. This example shows that also the nomenclature of a scientific name can be explored further.

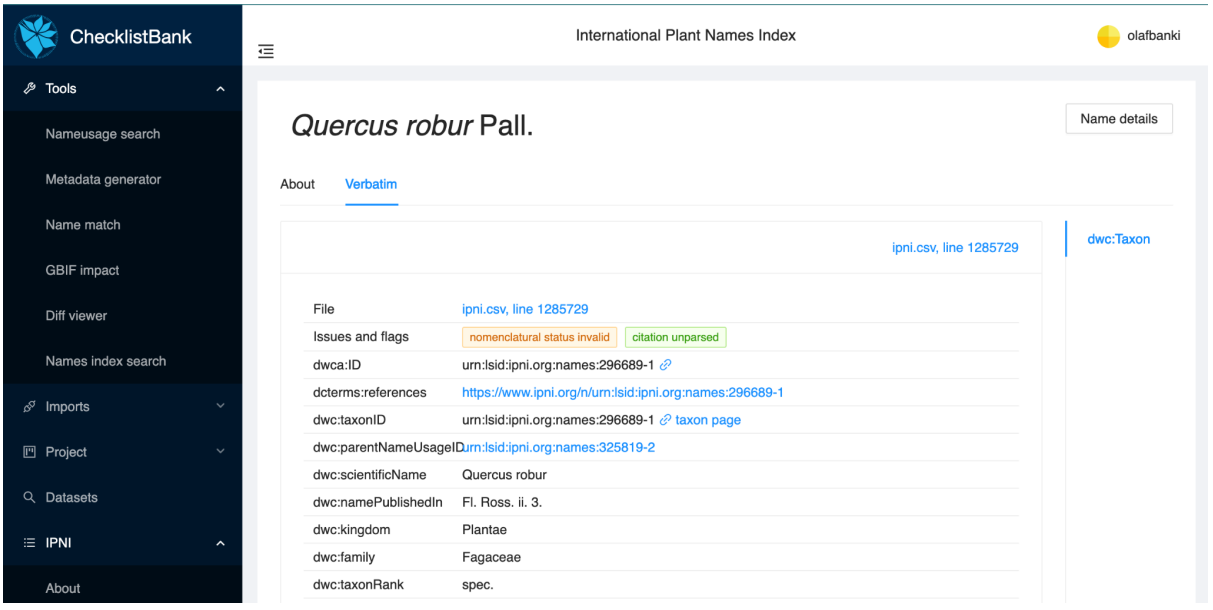


Fig. 5: Verbatim view of Quercus robur Pall. in IPNI

A screenshot of the verbatim of the scientific name of *Quercus robur* Pall. as provided by IPNI. The name record carries a flag of 'nomenclatural status invalid' because during the processing of the IPNI dataset in ChecklistBank the status could not be correctly interpreted.

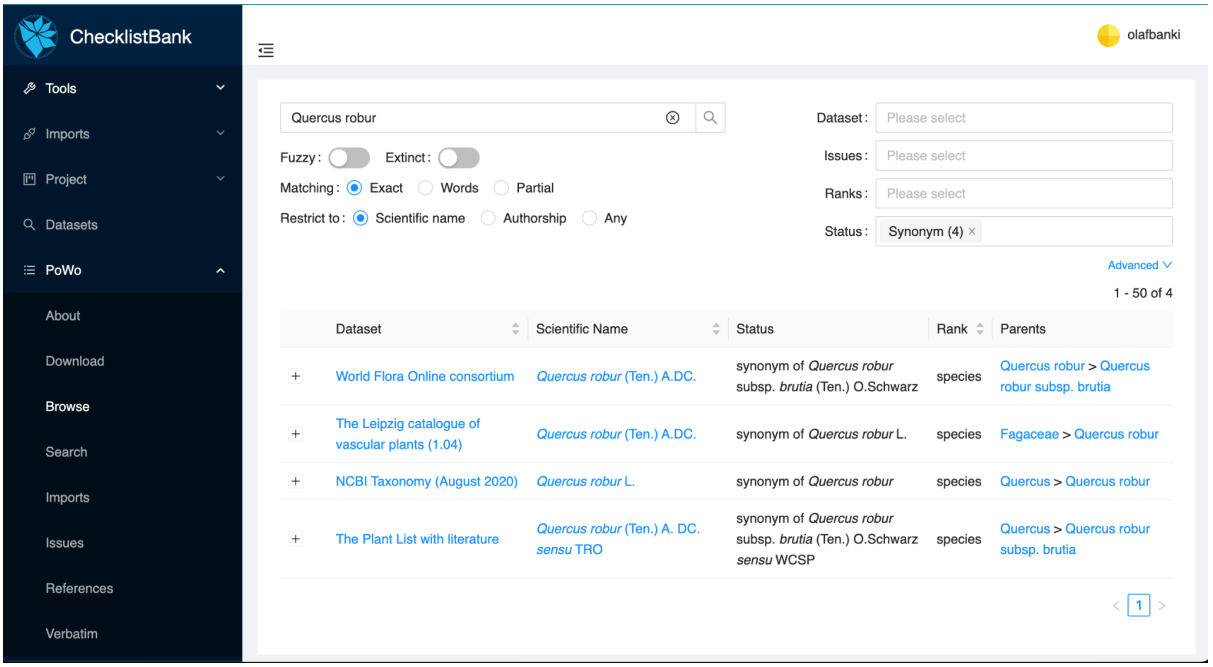


Fig. 6: Synonym of Quercus robur L.: Quercus robur (Ten.) A. DC.

A screenshot showing all synonyms at a species rank in datasets in ChecklistBank found to be associated with *Quercus robur* L.. Note that the NCBI Taxonomy places *Quercus robur* L. as a synonym of *Quercus robur*. The synonym mostly found is *Quercus robur* (Ten.) A. DC.

Apart from the accepted species, the search results also contain four instances of a synonym (Figure 6). Looking at one of the synonyms, *Quercus robur* (Ten.) A. DC., in the Leipzig Catalogue of Vascular Plants reveals that it is a synonym of a *Quercus robur* L. subspecies *Quercus robur* subsp. *brutia* (Ten.) O.Schwarz. Three subspecies are accepted in this dataset in ChecklistBank (Figure 7). Clicking on one of the subspecies reveals all associated synonyms provided by the Leipzig Catalogue of Vascular Plants dataset in ChecklistBank.

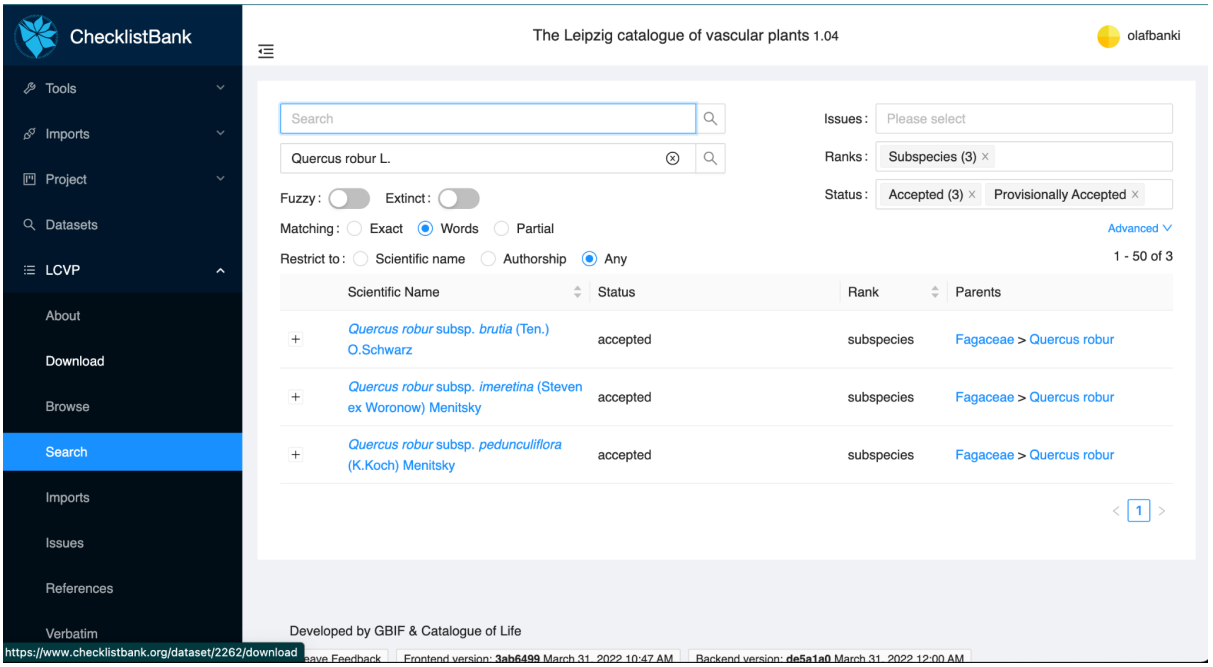


Fig. 7: Subspecies of Quercus robur L. in the Leipzig catalogue of vascular plants

A screenshot showing the subspecies of *Quercus robur* L. that are recognised by the Leipzig catalogue of vascular plants 1.04.

From the search results of the 'taxonomic name usage search' for *Quercus robur* L., the COL Checklist can itself also be explored. In ChecklistBank monthly releases of the COL Checklist can be found for the last year. Monthly releases of the COL Checklist older than a year are likely to be removed from the database in ChecklistBank, and only available as data dumps[9]. The annual releases of the COL Checklist will receive long term support, and remain available as datasets in ChecklistBank. For this example we use the March 2022 release of the COL Checklist (Bánki et al. 2022). Exploration of *Quercus robur* L. in this COL Checklist release reveals the data is provided by the 'World Checklist of Selected Plants' dataset (Govaerts 2017, Figure 8). Several scientific names are shown that are not validly published. The page also provides access to 4 subspecies (Figure 9). One of these synonyms is *Quercus robur* subsp. *brutia* (Ten.) O.Schwarz just as in the example given in the Leipzig Catalogue of Vascular Plants (Figure 6 & 7). Also in the COL Checklist all associated synonyms to the subspecies are provided including *Quercus robur* subsp. *robur*.

Apart from web visualisations of mappings between the COL Checklist and other data sources in ChecklistBank, different releases of the COL Checklist itself can also be compared on the basis of the usage of a scientific name.

Fig. 8: Quercus robur L. in the COL Checklist
A screenshot showing that in the March 2022 release of the COL Checklist the 'World Checklist of Selected Plants' dataset (Govaerts 2017) provides the data related to *Quercus robur* L.

[9] Annual and monthly versions of the COL Checklist can be downloaded from here: https://www.catalogueoflife.org/data/download

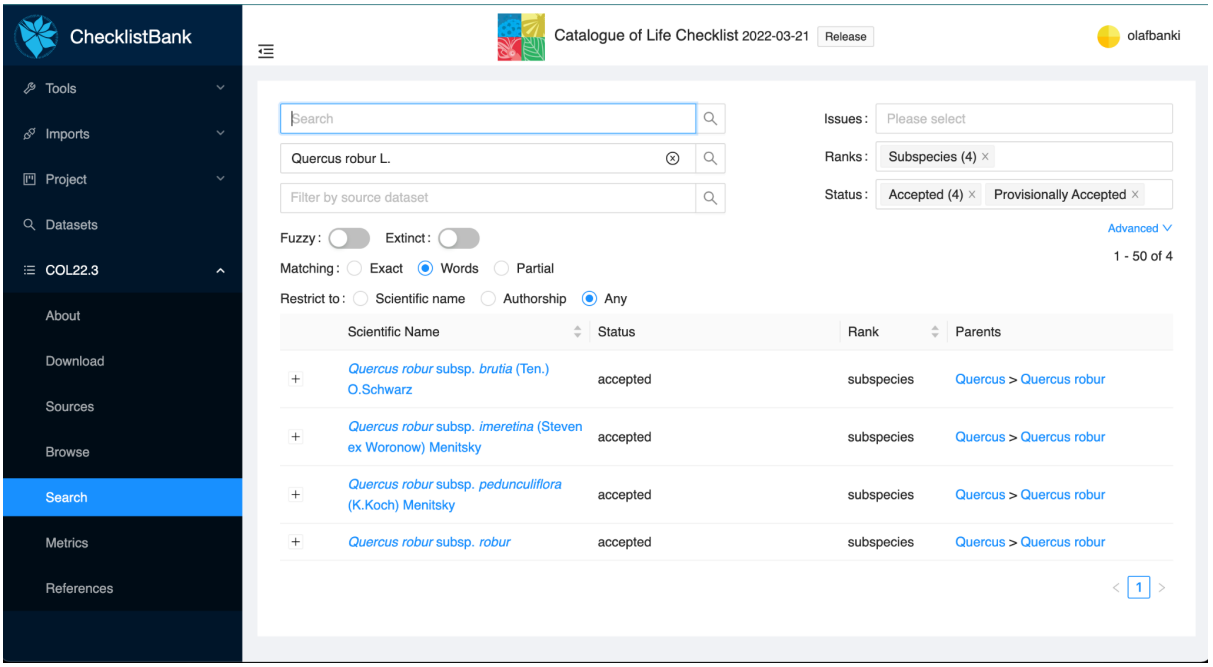


Fig. 9: Subspecies of Quercus robur L. in the COL Checklist

Screen capture showing the subspecies associated with *Quercus robur* L. in the March 2022 release of the COL Checklist originating from the 'World Checklist of Selected Plants' dataset (Govaerts 2017). By clicking on one of the subspecies all associated synonyms are revealed.

2.1.3. Contribution to web visualisation of mappings

The 'taxonomic name usage search' provides web visualisation of mappings between the COL Checklist, and the various releases thereof, with other taxonomic and nomenclature data sources published in ChecklistBank. Section 2.1.2 just gives an example of the possible avenues that could be explored by using the tool. The first results on the search of a scientific name can be the entry point for exploration of all datasets in ChecklistBank from a scientific name perspective. Explorations can venture into taxonomic data sources or data sources that are mostly concerned with names (nomenclatural data sources). Differences between the COL Checklist and other data sources or between data sources in ChecklistBank can be explored in detail. It is possible to investigate the verbatim form of a scientific name, meaning getting an insight into the scientific name that is provided by the original data source and how ChecklistBank has interpreted the information. ChecklistBank provides avenues for linkages with other data infrastructures or initiatives.

Apart from the actual web pages where scientific names and their usage can be explored, the existing ChecklistBank API can also be used by developers or knowledgeable users. It is to be expected that the web visualisation of mappings and the use of the API both will be useful for the BiCIKL community of users as these may serve different purposes. It is also to be expected that extensive scientific usage may lead to improvements of the tool in the near future.

2.2. Name match

2.2.1. Description of the tool

The 'name match' tool in ChecklistBank enables comparison of the COL Checklist with one or two other datasets in ChecklistBank in terms of taxon name matching. In the case matching is performed against the COL Checklist, by default this is always performed against the latest release of the COL Checklist. It is also possible to compare two datasets without any matching against the COL Checklist. A user is also able to upload a custom CSV file for comparison. The tool is a front-end application, and there is a limit to matching a maximum of 6000 taxon names in one request. The results of the matchings can be explored in web pages in ChecklistBank or downloaded as a CSV file. The 'taxon name match' can be accessed in ChecklistBank[10]. The tool is in production and usable for the wider public, without login into ChecklistBank.

2.2.2. An example: *Acacia* Mill.

The functionality and usability of the 'taxon name match' can be best conveyed by providing a real example of a comparison between datasets in ChecklistBank. We have chosen a comparison of the plant family Fabaceae, and specifically the genus *Acacia* Mill..

The intent is to compare the latest version of the COL Checklist with the World Checklist of Vascular Plants (WCVP) Fabaceae dataset (Legume Phylogeny Working Group 2021). We select the latter in the subject dataset field. As root taxon we choose the genus *Acacia* Mill. (Figure 10). By clicking the match button, 3,289 scientific names from the genus Acacia Mill. of the WCVP-Fabaceae are matched with the latest release of the COL Checklist (Bánki et al. 2022).

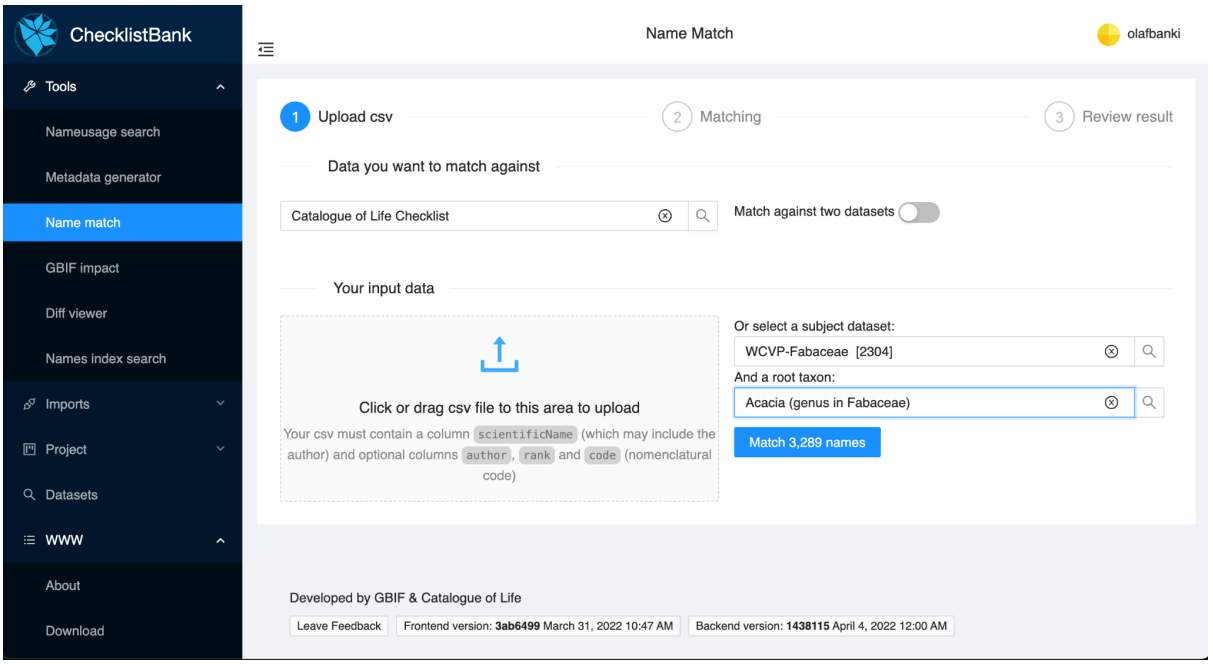
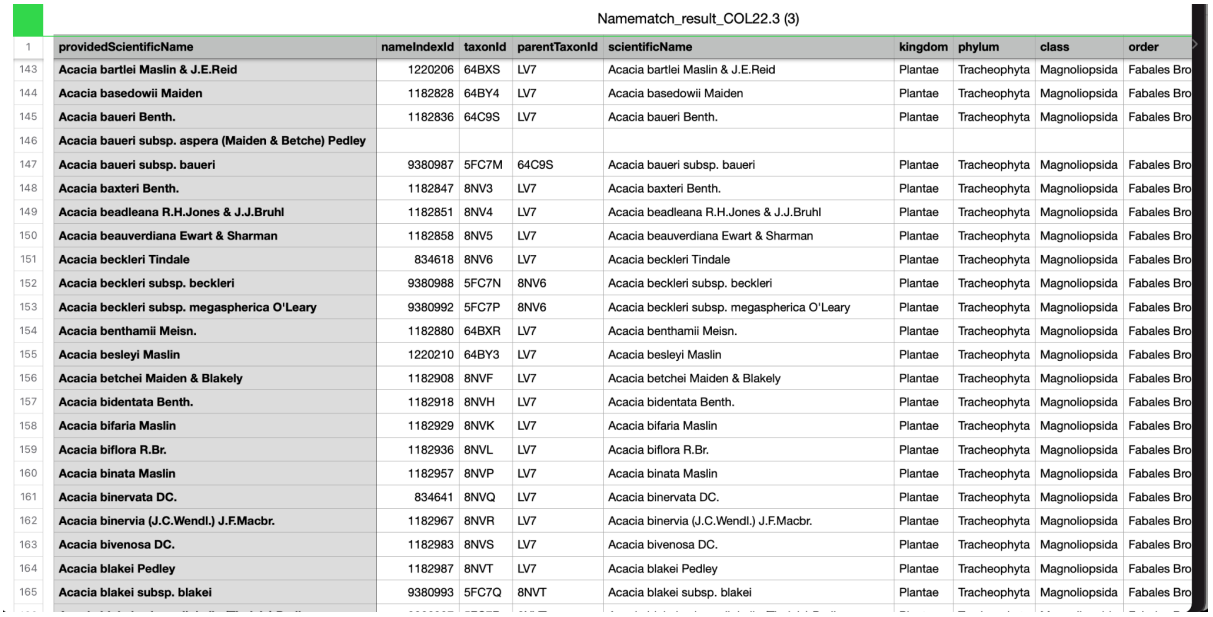


Fig. 10: Name match COL Checklist vs WCVP-Fabaceae

[10] The 'taxon name match' tool is accessible in ChecklistBank through the following URL: https://www.checklistbank.org/tools/name-match

A screenshot of the 'taxon name match' tool console with the WCVP-Fabaceae as subject dataset (Legume Phylogeny Working Group 2021) with the genus *Acacia* Mill. selected as root taxon for matching with the latest release of the COL Checklist (Bánki et al. 2022).

The results of the matching can either be explored in a web page or downloaded as a CSV file (Figure 11). In the case of the taxon name *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley the cells are empty of the COL Checklist and there seems to be no match.



Namematch_result_COL22.3 (3)

providedScientificName	nameIndexId	taxonId	parentTaxonId	scientificName	kingdom	phylum	class	order
Acacia bartlei Maslin & J.E.Reid	1220206	64BXS	LV7	Acacia bartlei Maslin & J.E.Reid	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia basedowii Maiden	1182828	64BY4	LV7	Acacia basedowii Maiden	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia baueri Benth.	1182836	64C9S	LV7	Acacia baueri Benth.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia baueri subsp. aspera (Maiden & Betche) Pedley								
Acacia baueri subsp. baueri	9380987	5FC7M	64C9S	Acacia baueri subsp. baueri	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia baxteri Benth.	1182847	8NV3	LV7	Acacia baxteri Benth.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia beadleana R.H.Jones & J.J.Bruhl	1182851	8NV4	LV7	Acacia beadleana R.H.Jones & J.J.Bruhl	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia beauverdiana Ewart & Sharman	1182858	8NV5	LV7	Acacia beauverdiana Ewart & Sharman	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia beckleri Tindale	834618	8NV6	LV7	Acacia beckleri Tindale	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia beckleri subsp. beckleri	9380988	5FC7N	8NV6	Acacia beckleri subsp. beckleri	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia beckleri subsp. megaspherica O'Leary	9380992	5FC7P	8NV6	Acacia beckleri subsp. megaspherica O'Leary	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia benthamii Meisn.	1182880	64BXR	LV7	Acacia benthamii Meisn.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia besleyi Maslin	1220210	64BY3	LV7	Acacia besleyi Maslin	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia betchei Maiden & Blakely	1182908	8NVF	LV7	Acacia betchei Maiden & Blakely	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia bidentata Benth.	1182918	8NVH	LV7	Acacia bidentata Benth.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia bifaria Maslin	1182929	8NVK	LV7	Acacia bifaria Maslin	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia biflora R.Br.	1182936	8NVL	LV7	Acacia biflora R.Br.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia binata Maslin	1182957	8NVP	LV7	Acacia binata Maslin	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia binervata DC.	834641	8NVQ	LV7	Acacia binervata DC.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia binervia (J.C.Wendl.) J.F.Macbr.	1182967	8NVR	LV7	Acacia binervia (J.C.Wendl.) J.F.Macbr.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia bivenosa DC.	1182983	8NVS	LV7	Acacia bivenosa DC.	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia blakei Pedley	1182987	8NVT	LV7	Acacia blakei Pedley	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro
Acacia blakei subsp. blakei	9380993	5FC7Q	8NVT	Acacia blakei subsp. blakei	Plantae	Tracheophyta	Magnoliopsida	Fabales Bro

Fig. 11: CSV file of the taxon name matching

A screenshot of the downloaded CSV file of the comparison of the genus Acacia Mill. between the WCVP-Fabaceae dataset (Legume Phylogeny Working Group 2021) and the latest release of the COL Checklist (Bánki et al. 2022). Where there is an empty cell the taxon names cannot be matched.

Within ChecklistBank the datasets can be explored further to understand why there is no match related to *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley. The World Wide Wattle is the taxonomic data source that forms the basis of the genus *Acacia* Mill. in the COL Checklist (Maslin 2018). It appears that the taxon name *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley does exist in the World Wide Wattle dataset, but without any synonyms (Figure 12).

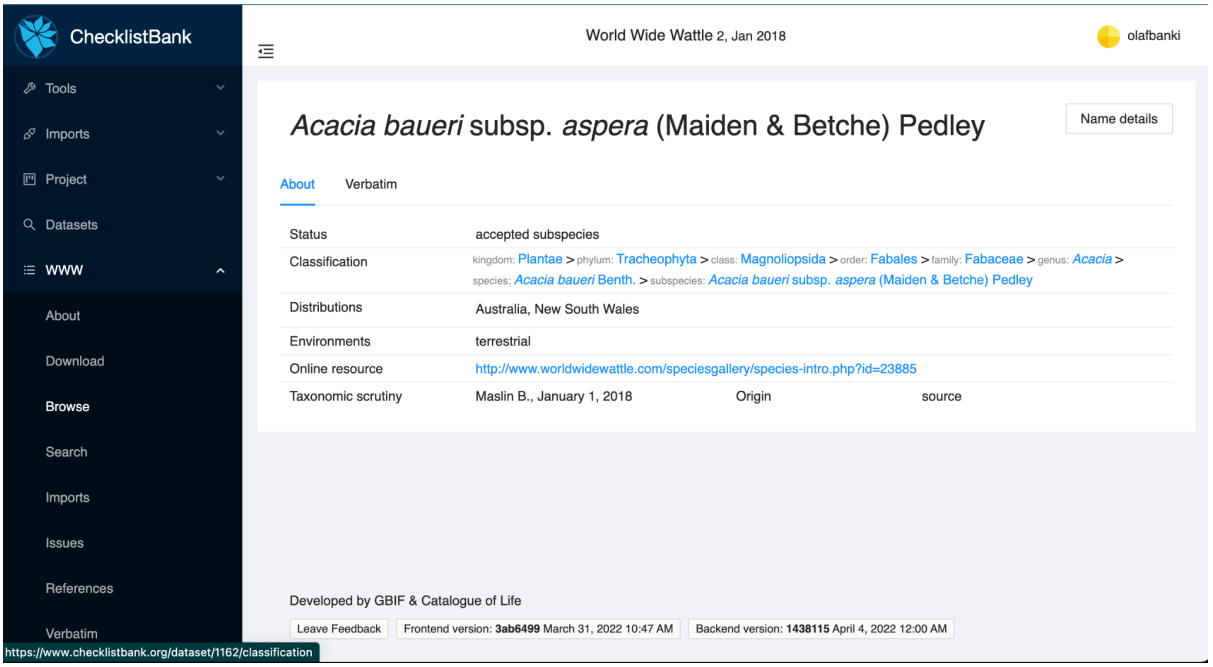


Fig. 12: Acacia baueri subsp. aspera (Maiden & Betche) Pedley in World Wide Wattle

World Wide Wattle is the source dataset used for *Acacia* Mill. in COL Checklist (Maslin 2018). *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley does exist as an accepted name but does not contain any synonyms.

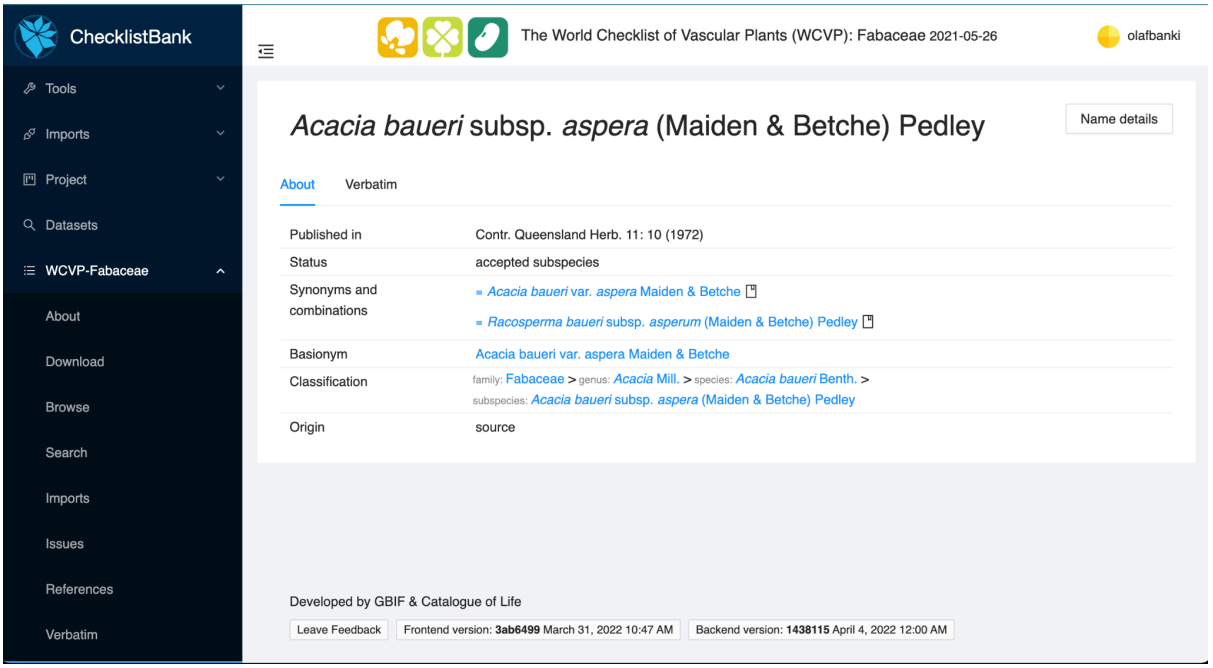


Fig. 13: Acacia baueri subsp. aspera (Maiden & Betche) Pedley in WCVP - Fabaceae

In WCVP-Fabaceae (Legume Phylogeny Working Group 2021) *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley is associated with two synonyms.

Closer inspection of the WCVP-Fabaceae dataset (Legume Phylogeny Working Group 2021) reveals that *Acacia baueri subsp. aspera* (Maiden & Betche) Pedley contains two synonyms (Figure 13). These synonyms may have caused the taxon names not to match up between WCPVP-Fabaceae and the COL Checklist (Maslin 2018).

2.2.3. Contribution to web visualisation of mappings

The 'taxon name match' tool provides the functionality to compare taxonomic datasets in ChecklistBank with each other. The default is to match taxon names with the latest COL Checklist release. An attractive functionality is that a user can also upload a custom CSV file and match this with the COL Checklist. The matching results can be downloaded.

Even though the 'taxon name match' tool is in production and can be used directly, there are still some improvements that could be made in facilitating the interpretation of the results of the matching by users. At a certain point in time it may also be desired to lift the 6,000 names limit, and enable more matching against the back-end database instead of a front-end application that is limited by the number of names to be matched.

2.3. Difference viewer

2.3.1. Description of the tool

The 'difference viewer' allows for comparison of two taxonomic 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 and authorship, if requested, between datasets. The set of names to be compared can be filtered by various parameters: One or more root taxa can be selected to include only certain groups. Synonyms and names below a certain rank can be excluded on demand, i.e. you can just compare accepted genera or families. The results of the comparison can be explored in web pages in ChecklistBank or can be downloaded as a text file. The 'difference viewer' can be accessed in ChecklistBank[11]. The tool is in production and usable, but a user has to login into ChecklistBank. The login is required, because using the tool can take up a lot of capacity of the available servers.

2.3.2. An example: *Acacia* Mill.

The start screen of the 'difference viewer' allows the user to choose two datasets to make the comparison on the basis of scientific names. The latest release of the COL Checklist (Bánki et al. 2022) is selected and the comparing dataset is the World Checklist for Vascular Plants - Fabaceae (Legume Phylogeny Working Group 2021). It is possible to choose a root taxon to make comparisons upon; in this case it is the genus *Acacia* Mill.. The user is able to select whether the scientific name comparisons include authorship and synonyms, and whether the parent should be shown in the results (Figure 14).

The Legume Phylogeny Working Group is a group of Fabaceae specialists that have worked on a comprehensive global taxonomic list for Fabaceae. This list is compared to the World Wide Wattle (Maslin 2018), the data source of the latest release of the COL Checklist for the genus *Acacia* Mill.. In a web page the results of the comparisons can be explored (Figure 15). In the web page two columns are visible, one for each dataset. Scientific name differences are shown in each column where a '-' or '+' sign indicates in which datasets there is a difference. The results of the comparisons can also be downloaded in a CSV file for further investigation.

[11] The 'difference viewer' is accessible in ChecklistBank through the following URL: https://www.checklistbank.org/tools/diff-viewer

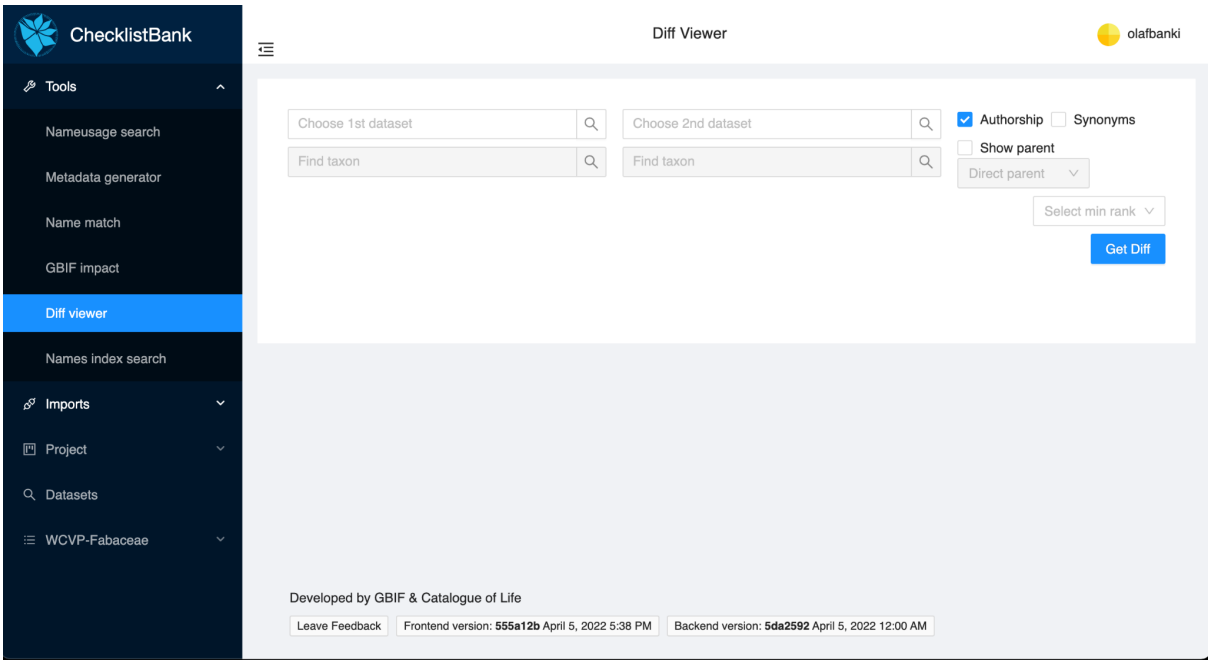


Fig. 14: Start screen of the difference viewer in ChecklistBank

Two datasets in ChecklistBank can be selected for comparison on a scientific name to name basis. A root taxon can be selected, as well whether comparisons should include authorship, synonyms and whether the parent should be indicated.

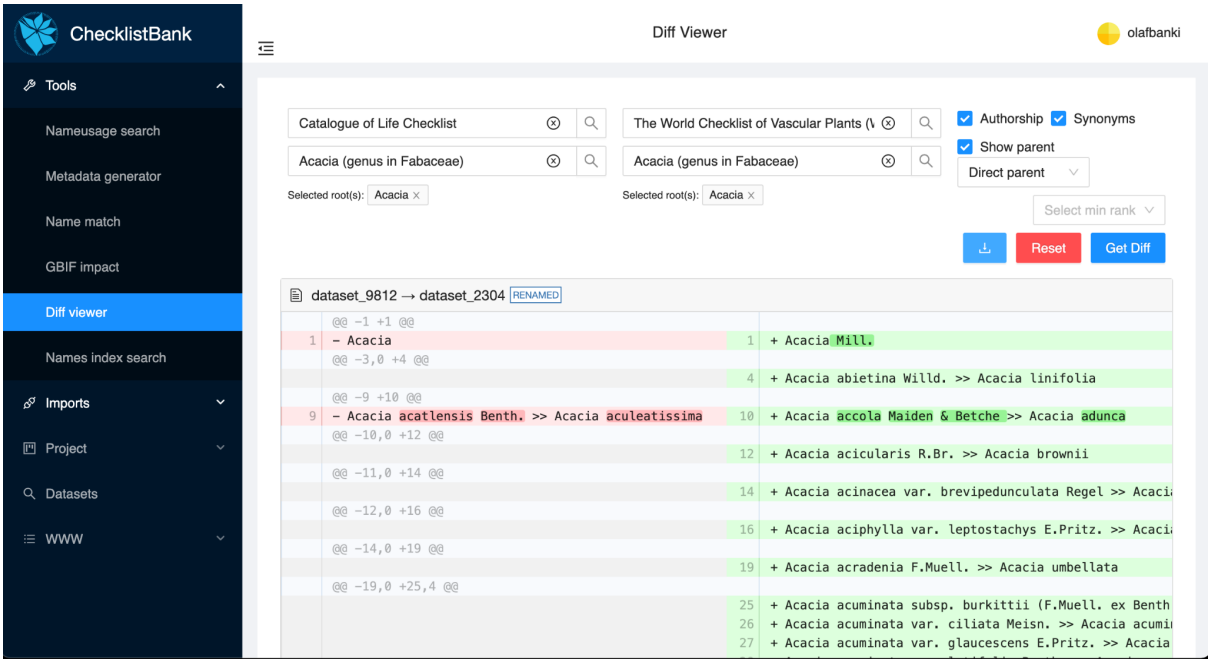


*Fig. 15: Difference between the COL Checklist and WCVP - Fabaceae: **Acacia Mill.***

A screenshot showing the results of the 'difference viewer' on the genus *Acacia* Mill. between the COL Checklist (Maslin 2018) and the WVCP-Fabaceae (Legume Phylogeny Working Group 2021). In two columns the differences in scientific names are shown per dataset.

The results of the 'difference viewer' (Figure 15) show that in the WCVP-Fabaceae dataset (Legume Phylogeny Working Group 2021) there is a scientific name difference for *Acacia*

accola Maiden & Betche to *Acacia adunca*. In ChecklistBank under datasets it is possible to search for the latest release of the COL Checklist (in this case the March 2022 release). In the search it is possible to look for the scientific name of *Acacia adunca* A.Cunn. ex G.Don in the COL Checklist. This scientific name appears to be an accepted species in the World Wide Wattle dataset in ChecklistBank (Figure 16). The species therefore does also appear in the COL Checklist, however the species page shows the species is not associated with any synonyms.

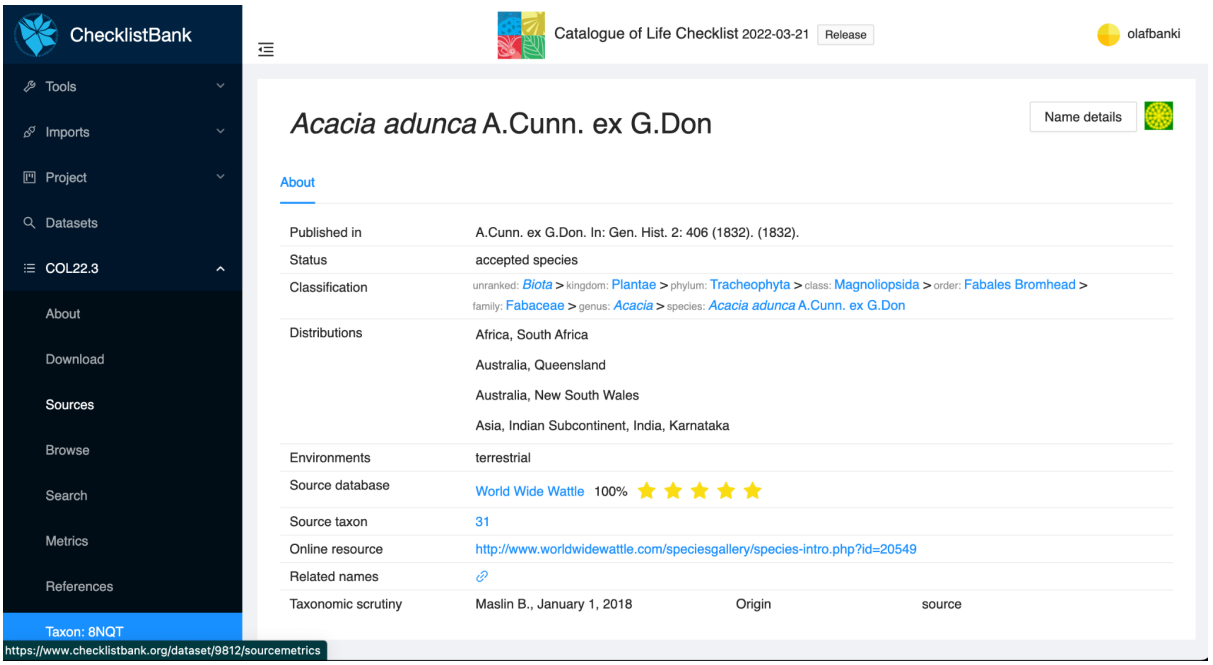


Fig. 16: Acacia adunca A. Cunn. ex G. Don in World Wide Wattle (COL Checklist)
A screenshot showing the accepted species *Acacia adunca* A.Cunn. ex G.Don in the World Wide Wattle dataset (Maslin 2018). The species is not associated with any synonyms.

In ChecklistBank also the WCVP-Fabaceae dataset (Legume Phylogeny Working Group 2021) can be explored further. *Acacia adunca* A.Cunn. ex G.Don is also an accepted species in this dataset. The species page of *Acacia adunca* A.Cunn. ex G.Don in WCVP-Fabaceae however does reveal three associated synonyms. One of these synonyms is *Acacia accola* Maiden & Betche, which is present in WCVP-Fabaceae but not in the COL Checklist (Figure 17).

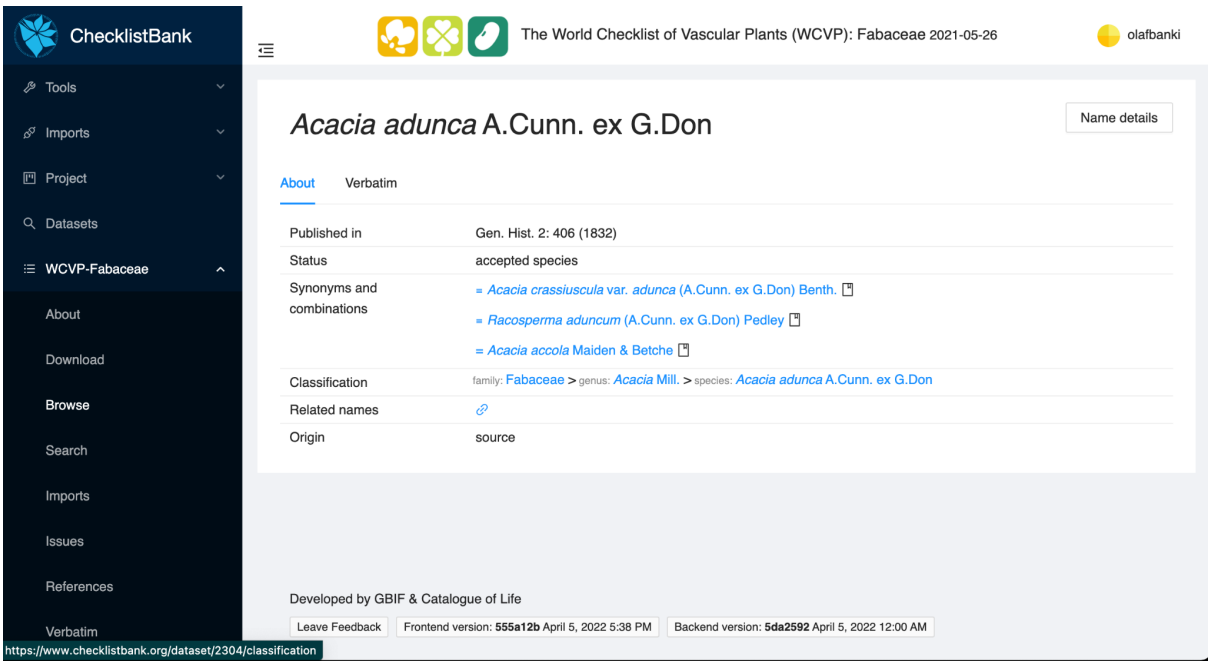


Fig. 17: Acacia adunca A. Cunn. ex G. Don in World Checklist of Vascular Plants - Fabaceae
A screenshot showing the species page of the accepted species *Acacia adunca* A.Cunn. ex G.Don in the WCVP-Fabaceae dataset (Legume Phylogeny Working Group 2021). Three synonyms are associated with this accepted species, including *Acacia accola* Maiden & Betche. The difference in terms of scientific names with the World Wide Wattle (Maslin 2018) are these three synonyms.

2.3.3. Contribution to web visualisation of mappings

The 'difference viewer' provides web visualisation of mappings between the COL Checklist, and the various releases thereof, with other taxonomic data sources published in ChecklistBank. Section 2.3.2 just gives an example of the possible avenues that could be explored by using the tool. The 'difference viewer' provides a comparison on a scientific name by scientific name basis. The results of the 'difference viewer' can be further explored in ChecklistBank. At times the results may be dependent on the specifics of a dataset. Some datasets for example do not contain genera as a separate field, so a root taxon can only be chosen at a family level.

The 'difference viewer' can aid users and COL editors to understand better what the difference in terms of scientific names is between data sources that are used to assemble the Catalogue of Life Checklist and other taxonomic data sources in ChecklistBank. The results of the viewer can assist with a review if a taxonomic data source is better suited to become the source for a particular taxonomic sector in the Catalogue of Life Checklist. For end-users we expect the 'difference viewer' will provide more transparency as to why certain data sources are selected over others in becoming the source used in the COL Checklist.

The 'difference viewer' is in production, and could be used instantly in ChecklistBank. However, with more usage it is expected that the tool might need some more improvements. For example, at present the tool only performs well if comparisons of datasets with a couple of thousands of names are made. This can be achieved by using a root taxon at the family rank, for example. The tool is not equipped to show the differences immediately of hundreds of thousands of scientific names at present.

The 'difference viewer' provides web visualisations of mappings between the COL Checklist and other taxonomic lists. This truly is a major step forward in the usage and transparency of the COL Checklist. The tool is expected to be useful to the BiCIKL user community as well.

3. Impact and way forward

3.1. Impact of deliverable 10.4

3.1.1. Impact of the tools of deliverable 10.4

The tools of Deliverable 10.4 are built within a new robust publicly available resource – ChecklistBank. The connection of the tools in this database is expected to quickly become a global resource due to its novelty and rich functions. While name matching tools exist they are not integrated in a comprehensive taxonomic checklist and nomenclatural data resource like ChecklistBank. It is the integration that is thought to provide immediate benefits to users.

The three tools of the deliverable 10.4 provide for the first time in the more than 20 years of existence of Catalogue of Life a means to much better explore and understand the differences between the COL Checklist and other taxonomic data sources, a function which can show the evolution of the COL Checklist consensus classification in time. In addition, taxonomic and nomenclatural data sets in ChecklistBank can be explored in comparison to each other. ChecklistBank already provides data and web linkages to other biodiversity data infrastructures and databases. It is expected that also a BiCIKL user community will benefit from the 'taxonomic name usage search', the 'taxon name match', and the 'difference viewer' tools.

3.1.2. Relation to other deliverables in BiCIKL

Within BiCIKL work package 10, the deliverable 10.4 plays an important role in promoting the usage of ChecklistBank as a whole. For each of the deliverables within work package 10, the tools of deliverable 10.4 will contribute in understanding the relations between taxonomic checklists (D10.1), the connection to literature references (D10.2), the validation and curation of taxonomic and nomenclatural data resources (D10.3), and by understanding the taxonomic backbone services of other biodiversity data infrastructure partners within BiCIKL (D10.5).

Within the rest of BiCIKL, the deliverable 10.4 probably is most useful for the projects under work package 4 that will be submitted by the BiCIKL user community.

3.2. Wider biodiversity data landscape

3.2.1. A complimentary GBIF impact tool

As part of the ongoing collaboration between the Catalogue of Life and GBIF, a tool has been developed complementary to the deliverable 10.4. The tool is called the 'GBIF impact' tool.

This tool compares taxonomic interpretation of GBIF occurrence records between the current GBIF taxonomic backbone and the Catalogue of Life Checklist. The aim of the tool is to understand the differences between the current GBIF taxonomic backbone and the Catalogue of Life Checklist and the effect these differences have on the correct display of the occurrence records within GBIF.org. It is a tool to help understand where the COL Checklist could be improved with more up to date taxonomic information, additional scientific names from taxonomic and nomenclatural data sets or from published literature.

The 'GBIF impact' tool is currently only available for editors of the Catalogue of Life Checklist, and cannot be accessed by external users.

3.2.2. Alignment of biodiversity data infrastructures

Currently ChecklistBank contains taxonomic checklists used by COL and/or GBIF. ChecklistBank is open for publication of other taxonomic datasets, and community publishing to ChecklistBank will allow quick comparison to well-used and established taxonomic checklists. In time, all data sets in ChecklistBank will be made available with an unique DOI.

A current common but duplicative task for a researcher is to download taxonomic checklists from multiple sources, combine them and then compare and assess the results in a separate software. This is time consuming and does not allow researchers to take advantage of previous knowledge. The new tools embedded in the ChecklistBank infrastructure will alleviate this problem.

An exciting implication of this work is that once automatic harvesting and ingestion of taxonomic data into ChecklistBank from the literature (Plazi, BHL) is in place (to be partially addressed in the BiCIKL project), any potentially new taxonomic names can be automatically compared to checklist, such as COL and the GBIF taxonomic backbone, to determine whether they need to be added. This will greatly decrease the time lag from taxonomic publication to use by community researchers and taxonomic infrastructures. Given that ChecklistBank will make taxonomic and nomenclatural data available through a standard API where data can be found, made accessible, data could be integrated and (re-)used this epitomises the FAIR principles on which BiCIKL is based.

3.3. Way forward

The 'taxonomic name usage search', the 'taxon name match', and the 'difference viewer' are available to users and the BiCIKL user community. It is expected that with more usage of the tools, improvements will be made in the future to better serve the use cases brought forward by these users. The Catalogue of Life and GBIF will promote ChecklistBank and its associated tools to their respective user communities. Within BiCIKL the tools are expected to play a role in the projects that will be submitted by the BiCIKL user community.

4. Acknowledgements

The Catalogue of Life and the Global Biodiversity Information Facility especially would like to thank Thomas Stjernegaard Jeppesen and Markus Döring for developing ChecklistBank and their associated tools.

5. References

Alliance for Biodiversity Knowledge, https://www.allianceforbio.org/

Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Adlard, R., Adriaenssens, E. M., Aedo, C., Aescht, E., Akkari, N., Alfenas-Zerbini, P., et al. (2022). Catalogue of Life Checklist (Version 2022-03-21). Catalogue of Life. https://doi.org/10.48580/dfpd

Biodiversity Heritage Library (BHL), https://www.biodiversitylibrary.org/

Biodiversity Community Integrated Knowledge Library (BiCIKL), https://bicikl-project.eu/

Catalogue of Life (COL), https://catalogueoflife.org

ChecklistBank, https://www.checklistbank.org/

Costello, Mark John, DeWalt, R. E., Orrell, Thomas M., and Banki, Olaf. 2022. "Two million species catalogued by 500 experts." Nature, 601, (7892), 191–191. , https://doi.org/10.1038/d41586-022-00010-z.

Global Biodiversity Information Facility (GBIF), https://gbif.org

Govaerts, R. (2017). World Checklist of Selected Plant Families. In O. Bánki, Y. Roskov, M. Döring, G. Ower, L. Vandepitte, D. Hobern, D. Remsen, P. Schalk, R. E. DeWalt, M. Keping, J. Miller, T. Orrell, R. Aalbu, R. Adlard, E. M. Adriaenssens, C. Aedo, E. Aescht, N. Akkari, P. Alfenas-Zerbini, et al., Catalogue of Life Checklist (Aug 2017). https://doi.org/10.48580/dfp3-38c

Hobern, D, Barik, S.K., Christidis, L., Garnett, S.T., Kirk, P., Orrell, T.M., Pape, T., Pyle, R.L., Thiele, K.R., Zachos, F.E., & Bánki, O. Towards a global list of accepted species VI: The Catalogue of Life checklist. Org Divers Evol 21, 677–690 (2021). https://doi.org/10.1007/s13127-021-00516-w

Legume Phylogeny Working Group (LPWG), Andrella, G. C., Atahuachi Burgos, M., Bagnatori Sartori, Â. L., Balan, A., Bandyopadhyay, S., Barbosa Pinto, R., Barrett, R. L., Boatwright, J. S., Broich, S. L., Brullo, S., Bruneau, A., Cardinal-McTeague, W., Cardoso, D., Castro Silva, I. C., Cervantes, A., Choo, L. M., Cobra e Monteiro, T., Compton, J., et al. (2021). The World Checklist of Vascular Plants (WCVP): Fabaceae (R. Govaerts, Ed.). The Royal Botanic Gardens, Kew,. http://sftp.kew.org/pub/data_collaborations/Fabaceae/DwCA/

Maslin, B. (2018). World Wide Wattle. In O. Bánki, Y. Roskov, M. Döring, G. Ower, L. Vandepitte, D. Hobern, D. Remsen, P. Schalk, R. E. DeWalt, M. Keping, J. Miller, T. Orrell, R. Aalbu, R. Adlard, E. M. Adriaenssens, C. Aedo, E. Aescht, N. Akkari, P. Alfenas-Zerbini, et al., Catalogue of Life Checklist (2, Jan 2018). https://doi.org/10.48580/dfp3-3f4

Plazi, https://plazi.org/

Species 2000, https://www.sp2000.org