Validation and Curation Workbench for Taxonomic Data

Deliverable D10.3

31 October 2022

Olaf Bánki*, Joe Miller#, Donald Hobern*, Peter Schalk*, 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:	Validation and Curation Workbench for Taxonomic Data
Deliverable n°:	D10.3
Nature of the deliverable:	Report + associated tools
Dissemination level:	Public
WP responsible:	WP10
Lead beneficiary:	Species 2000
Citation:	Bánki, O., Miller, J., Hobern, D., Schalk, P., Stjernegaard Jeppesen T., & Döring, M. (2022). *Validation and curaton workbench for taxonomic data*. Deliverable D10.3 EU Horizon 2020 BiCIKL Project, Grant Agreement No 101007492.
Due date of deliverable:	Month 18
Actual submission date:	31 October 2022

Deliverable status:

Version	**Status**	**Date**	**Author(s)**
1.0	Draft	25 October 2022	Bánki, O., Miller, J., Hobern, D., Schalk, P., Stjernegaard Jeppesen, T., and Döring, M. Species 2000 & GBIF
1.1	Review	26 October 2022	Güntsch, A. and Agosti, D.
2.0	Final	30 October 2022	Bánki, O., Miller, J., Hobern, D., Schalk, P., Stjernegaard Jeppesen, T., and Döring, M. Species 2000 & GBIF

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 example 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 Species 2000) 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 listing of the world's species. 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.

In 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 10.3 "Curation and validation 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 10.5 "Delivery and presentation of taxonomic information".

Summary

The BiCIKL deliverable D10.3 encompasses a 'Validation and curation workbench for taxonomic data'. ChecklistBank offers workbench tools that are used for the assembly of the COL Checklist. As part of D10.3 a more generic 'ChecklistBank project functionality' has been developed that enables the creation and curation of project datasets in ChecklistBank. The ChecklistBank workbench functionality is not intended to support the *de novo* creation of a new taxonomic dataset or to offer all the content features and functionality that platforms such as WoRMS, ITIS and TaxonWorks provide for managing nomenclatural and taxonomic data (including literature references, type specimens, distribution data, etc.). The BiCIKL deliverable D10.3 opens the way for wider use of ChecklistBank not only as a repository and access tool for species checklists, but also as an environment for production of useful checklist products based on the primary outputs of published taxonomic research.

List of abbreviations

BHL	Biodiversity Heritage Library
BiCIKL	Biodiversity Community Integrated Knowledge Library
COL	Catalogue of Life
ColDP	Catalogue of Life Data Package
DNA	Deoxyribo Nucleic Acid
DwC-A	Darwin Core Archive
ENA	European Nucleotide Archive
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
GBIF	Global Biodiversity Information Facility
ITIS	Integrated Taxonomic Information System
MOTU	Molecular Operational Taxonomic Unit
OTU	Operational Taxonomic Unit
UNITE	Unified system for rDNA sequences based identification of fungal species
WoRMS	World Register of Marine Species

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 (Bánki et al. 2022), 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[1]. COL, through ChecklistBank, also supports those who need to manage their own taxonomic data 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, for example, museum specimens collected in the 18th and 19th century to geotagged smartphone photos shared by amateur naturalists in recent days.

[1] https://www.catalogueoflife.org/about/glossary.html#classification

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 October 2022 release of the COL Checklist is based on 165 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. DwC-A, 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, such as blocking duplicate names. The COL Checklist uses a management classification - an agreed higher taxonomic classification of higher taxa that supports the organisation of all other sections of the checklist. 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, coverage of all scientific names should be increased, including those no longer considered the accepted name for a species. With an increasing proportion of new and already published taxonomic research published in digital formats, COL and its BiCIKL partners are working to remove or minimise barriers to immediate processing of data from these publications within the COL Checklist and for timely updates to other taxonomic databases, particularly COL's own data sources. Visualisation of mappings between the COL Checklist and other taxonomic checklists will contribute to the overall usefulness of the COL Checklist and will assist COL editors in assessing available taxonomic data sources as potential new data sources for assembling 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. Annual COL Checklist versions up to 2019 used a simple hashing of names to limit changes to record identifiers between releases. This resulted in name based identifiers that did change even when just a single character of a name or its authorship was corrected.

The new implementation aims to keep the identifiers stable when the authorship of a name has only been slightly modified, although it does force a change in identifiers when an authorship is added to a record that previously lacked one. Changes in status (accepted name or synonym) and parent/classification changes do not trigger any ID changes. So when name usages change status from an accepted name to a synonym or vice versa, there is no change in the ID. By combining a name usage identifier and the data set key the user has a stable reference to an immutable name usage in a particular release of the COL Checklist, no matter how the treatment of this name changes over time.

In some cases, multiple identifiers may previously have been issued for the same name, for example when the same genus has been added to the COL Checklist at different positions in the hierarchy based on classifications provided by different sources. In such cases, COL prefers the identifier issued in the oldest release to maximise stability.

The new identifiers aim 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[2]. 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[3]. For a species identifier we use for example 4QHKG for Puma concolor[4].

[2] https://github.com/CatalogueOfLife/backend/issues/491 - case-insensitive combinations of the characters 23456789BCDFGHJKLMNPQRSTVWXYZ

[3] https://www.catalogueoflife.org/data/taxon/P

[4] https://www.catalogueoflife.org/data/taxon/4QHKG

The existence of stable name usage 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. 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 data repository that facilitates access to original data sources underlying the COL Checklist, all COL Checklist releases, all GBIF taxonomic checklists, and workbench or 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.checklistbank.org/) to render all COL Checklist data to ChecklistBank, the COL portal and users, provide persistent name and digital object identifiers (DOIs), and support various data standards (e.g. DwC-A, ColDP).

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 interpreted into a standard data model and accessible through a standard API and reusable web browser components. In future, datasets in ChecklistBank will be cited using a ChecklistBank Digital Object Identifier (DOIs). At present, DOIs are only available for project releases and it's underlying data sources such as the Catalogue of Life (COL) Checklist. 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 COL 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. Current state of curation in ChecklistBank

ChecklistBank is an integrated data repository and workbench for taxonomic checklists. It replaces previously largely manual processes that COL editors have developed over many years to support the construction and versioning of the COL Checklist. By combining data management and versioning for component source datasets with rich tools for constructing, editing and versioning the integrated checklist, ChecklistBank not only increases the

robustness and FAIRness of the COL Checklist, but also offers a functional workbench for other checklist editing processes. The aggregation of thousands of source datasets in a single repository will speed development of other national, regional or thematic checklists. The versioning support for datasets in ChecklistBank will allow editors to refresh each constructed checklist with the latest versions of the source datasets.

The taxonomic data sources that are used to construct the COL Checklist have until now all been curated outside ChecklistBank using external editing platforms such as WoRMS, ITIS and TaxonWorks or bespoke tools, databases and spreadsheets. Taxonomists using these platforms then publish versions of their data sources to ChecklistBank. Data sources need to be converted into one of the existing data standards (e.g. DwCA, ColDP) before they can be published in ChecklistBank. ChecklistBank provides the contributors with tools to view their published dataset, review potential issues around data integrity, and update the dataset with new versions. ChecklistBank at present only maintains the latest version of a dataset. Contributors benefit further by ChecklistBank serving as a citable source for the latest version of their data, or the specific part of their dataset that has been integrated into the COL Checklist, and from standardised API access to their dataset.

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.

At present there are no ways to alter data sources directly in ChecklistBank. A long desired feature is as well to provide a form of a feedback mechanism to data custodians, so adjustments to the original data can be made after processing in ChecklistBank has revealed issues with the data. A 'lightweight' editing option in ChecklistBank could be an interesting asset for some taxonomic communities and may be further explored as part of Deliverable D10.5 CoL services for direct encapsulated use within search and discovery services of BiCIKL partner infrastructures.

1.3.3. Technical implementation

ChecklistBank is an open-source project with multiple repositories hosted in GitHub[5]. The back-end[6] is implemented in Java as a Dropwizard application that drives the COL ChecklistBank API. The front-end[7] 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.

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

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 by Sp2000/Catalogue of Life and GBIF in ChecklistBank. The BiCIKL project activities will be centred around the existing and ongoing developments by the COL-GBIF collaboration of ChecklistBank.

2. Functionality of Deliverable 10.3

The focus of BiCIKL Deliverable D10.3 is to make the workbench tooling[8], used for the assembly of the Catalogue of Life Checklist, also available in general for the assembly of other checklists. The workbench tools for assembling the COL Checklist are made available through the 'ChecklistBank project functionality'. Making use of this functionality enables a user to curate and manage a project dataset in ChecklistBank.

2.1. Access to Project Functionality in ChecklistBank

The following section describes the services included in the 'ChecklistBank project functionality'. The Figures 1 to 11 subsequently provide a workflow example of how a project is set up, how a project dataset is managed and curated, and lastly how a project dataset is published as a ChecklistBank dataset.

The access to the 'ChecklistBank project functionality' provides the following services:

- Users of ChecklistBank with the appropriate editing rights will be able to create a new "project" checklist dataset (see Figure 1 and Figure 2).
- Each project will include a ChecklistBank project view accessible to the user in the ChecklistBank menu (Figure 3 and see Figure 4 for the entire project menu in ChecklistBank).
- Within the project view, the metadata of the project dataset can be created or adjusted, for example by uploading a metadata file or editing the metadata directly within the project (Figure 4).
- Users may grant access for other registered users to have edit access to the project they manage (Figure 5).
- Managed projects will initially be empty (Figure 6).
- Users may create records for one or more root taxa or craft a higher classification to attach sectors (the highest-ranked taxa to be included in the project; Figure 6 and Figure 7).
- Users may select taxon records from any public ChecklistBank dataset and attach these as sectors within the project dataset (Figure 7 and 8).
- Each sector can be synchronised with the source dataset and will then contain a replica of the taxonomic hierarchy below the source taxon record (Figure 9).

[8] The workbench tooling is also referred to as assembly tooling.

- The user may request any sector to be refreshed to reflect new versions of the source dataset (Figure 9).
- The user may attach sectors at any level within the project dataset (Figure 7 and 9).
- The user may delete any branch of the project taxonomic tree - this will exclude that branch if the contributing sector is refreshed (Figure 6).
- The user may define decisions that control the inclusion of taxon records from a source dataset (Figure 10).
- If the dataset is ready for release this can be triggered and the dataset will be made available for public use (unless it is a private dataset) in ChecklistBank (Figure 11).

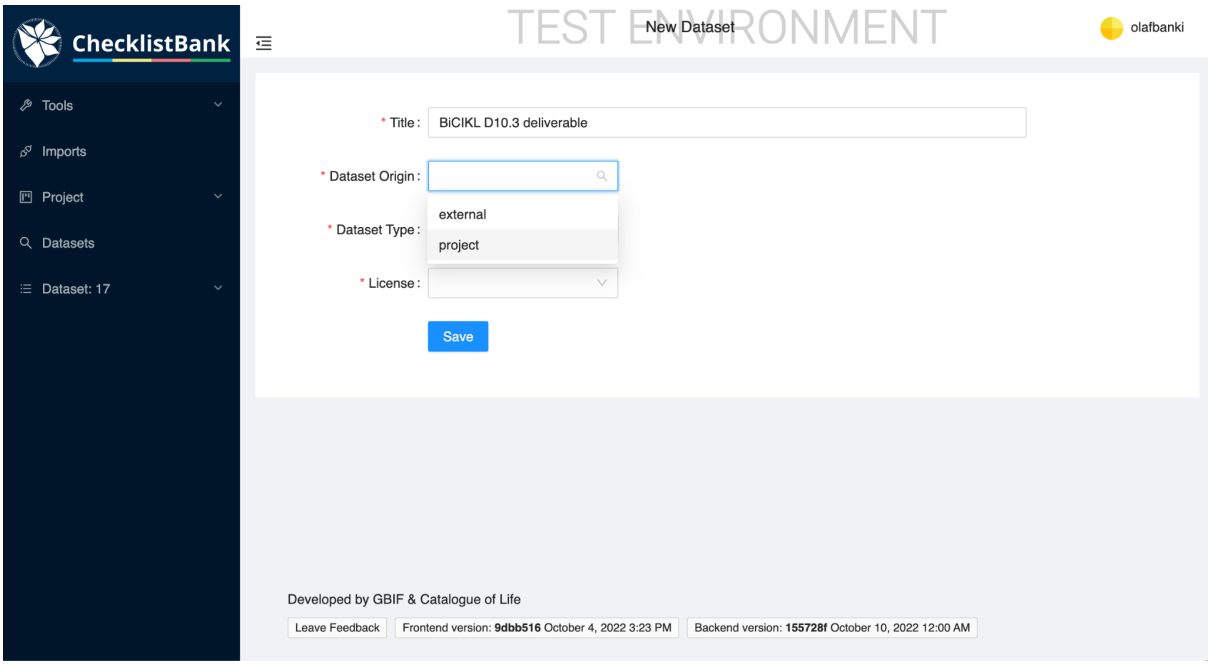

Figure 1: *Setting up a project dataset.*

A project dataset is set up with the title BiCIKL D10.3 deliverable. The dataset origin needs to be set to 'project' and not 'external'. There is an option to choose the dataset type: nomenclatural, taxonomic, thematic, legal, etc. In addition, one of the supported Creative Commons open data license types must be selected.

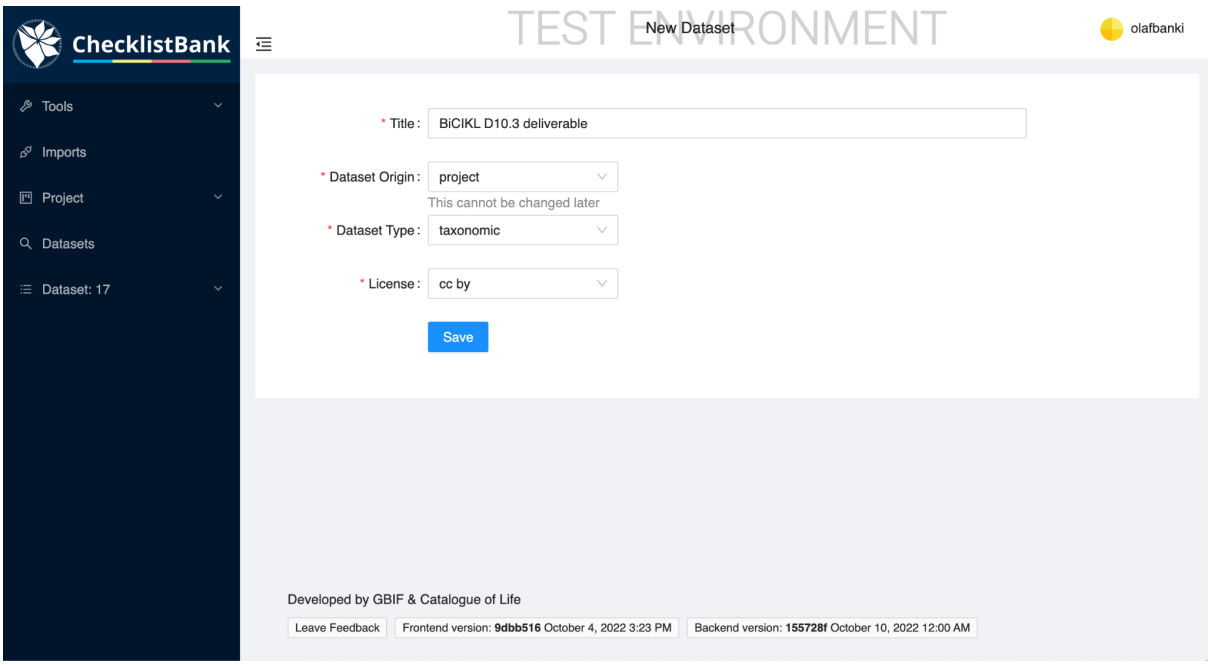

Figure 2: *Defining the type for a project dataset.*

The project dataset with the title BiCIKL D10.3 deliverable is a taxonomic dataset with Creative Commons open data license (CC-BY). In choosing the license, the user should be aware that the license selected for the dataset must be compatible with the licenses for the included sources. For example, the project dataset cannot be licensed CC0 if it contains sections from datasets licensed CC-BY.

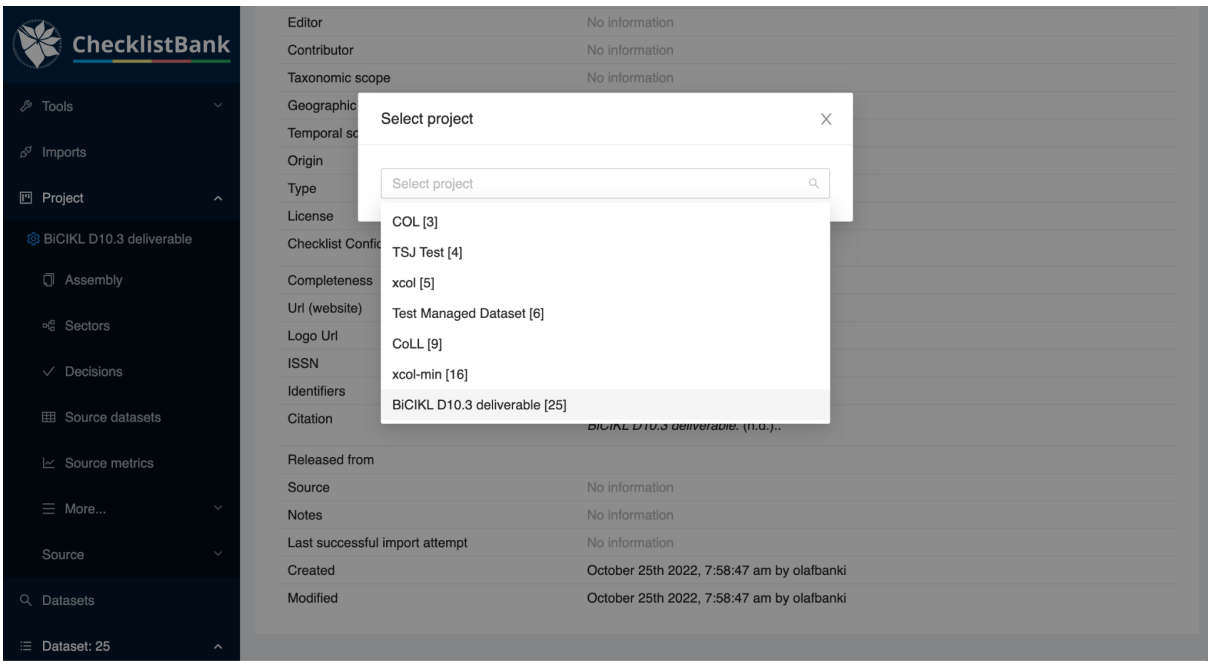

Figure 3: *The project view of Dataset BiCIKL D10.3 deliverable.*

On the left side of the ChecklistBank menu you can open the project, right click on the following icon  next to the project dataset title, and select the specific project dataset you would like to open. In Figure 3, the BiCIKL D10.3 deliverable is chosen.

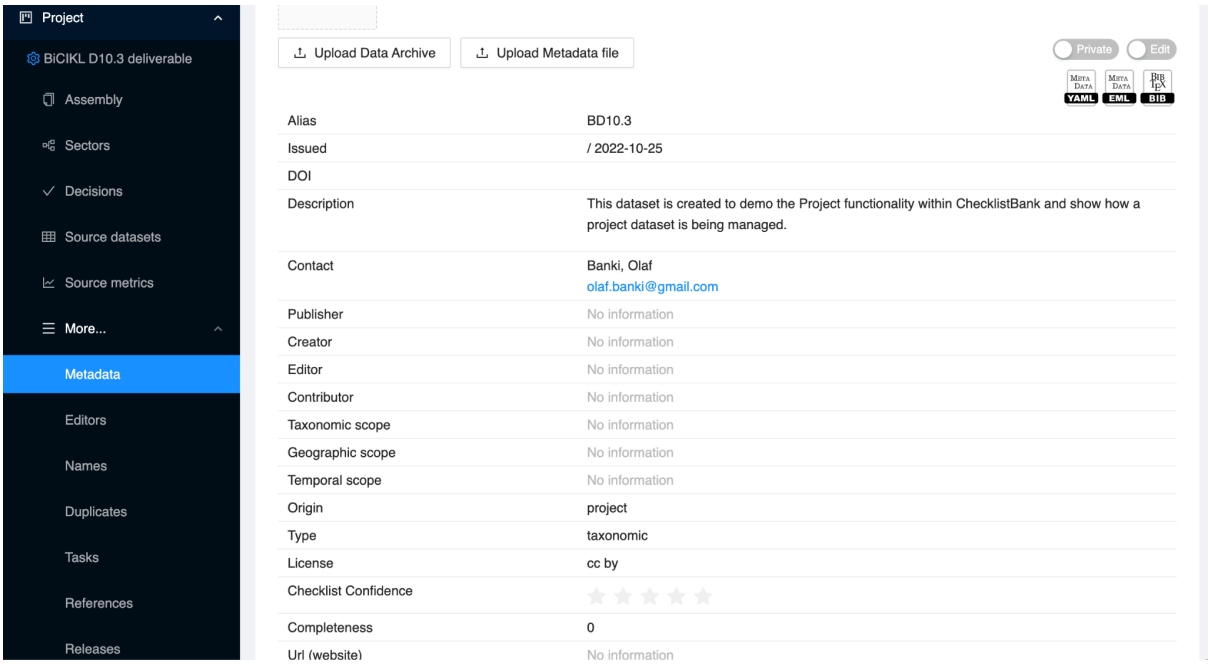

Figure 4: *Editing the metadata of the dataset BiCIKL D10.3 deliverable.*

The metadata of the dataset BiCIKL D10.3 deliverable can be edited directly in the project. An alias, BD10.3, is given to the dataset as well as a small description and contact information. There is also an option to upload a metadata file (in YAML or EML formats), logo, or data archive to bootstrap the project content. The dataset can also be made available as a private dataset. This means the dataset will not be publicly visible in ChecklistBank.

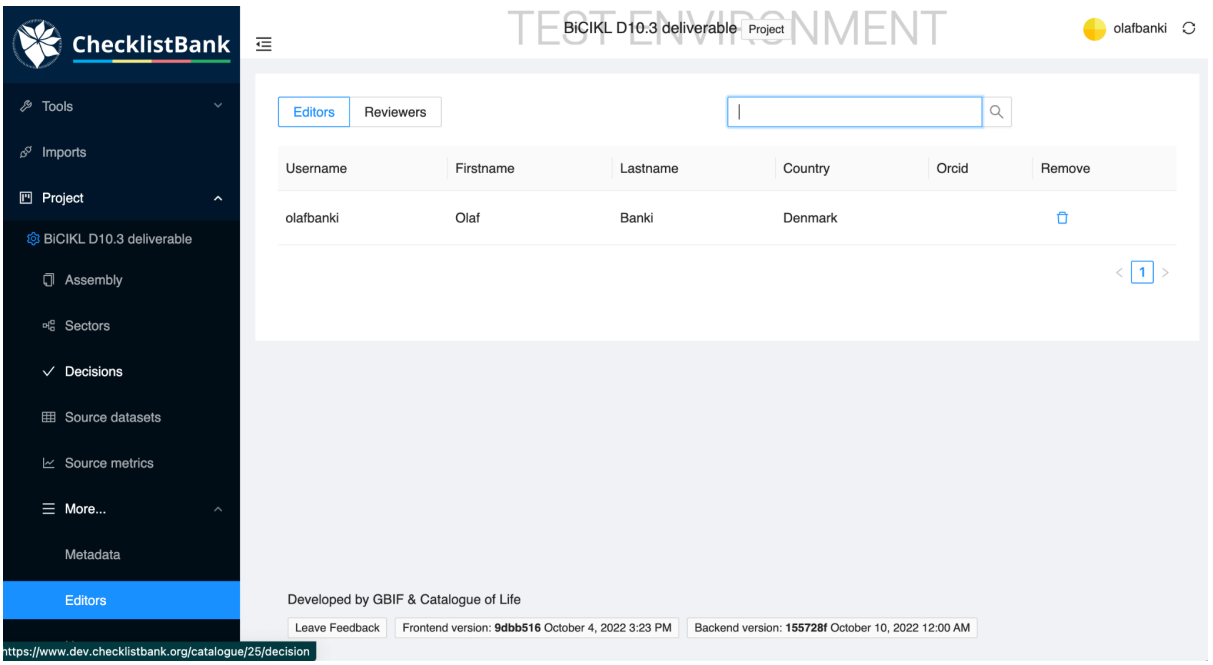

Figure 5: *Assigning editors to the project dataset.*

Multiple editors could be assigned to the dataset in case this is needed.

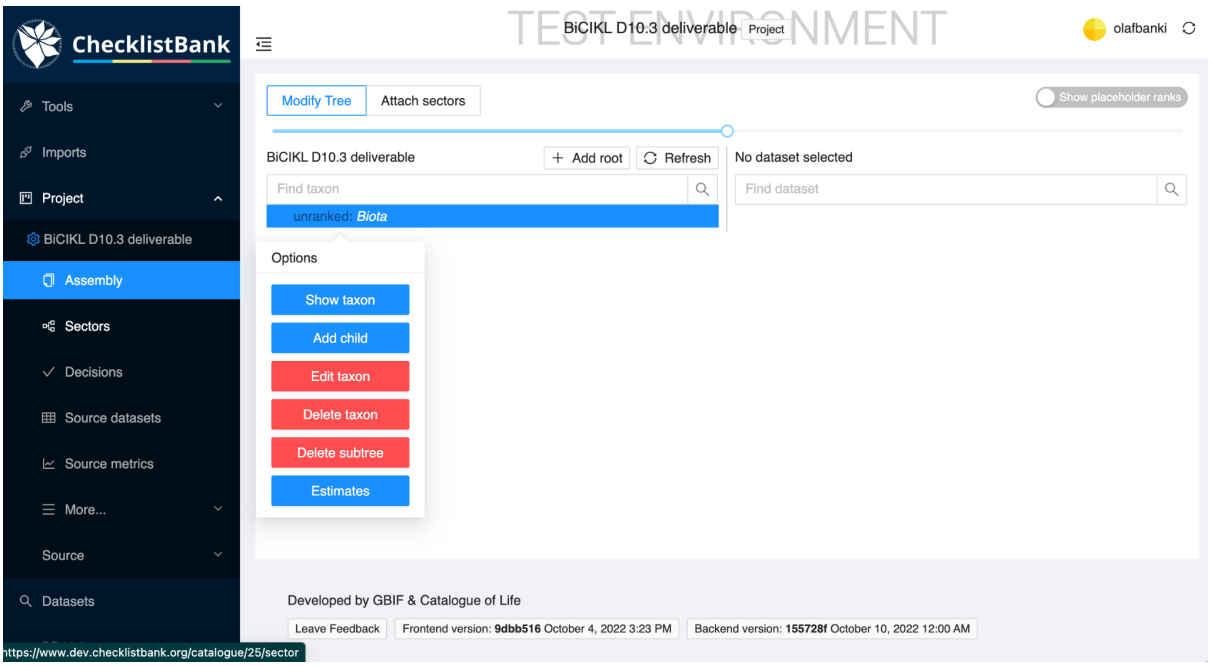

Figure 6: *Assembly of the tree.*

The tree of the project dataset is empty at first. The tree can be modified at any time. For example by starting to add higher ranks such as Biota to the tree. Organising higher ranks can be done, with a right click on Biota. By doing this several options are provided, to show the taxon, add a child taxon, edit the taxon, delete the taxon and subtree, etc.

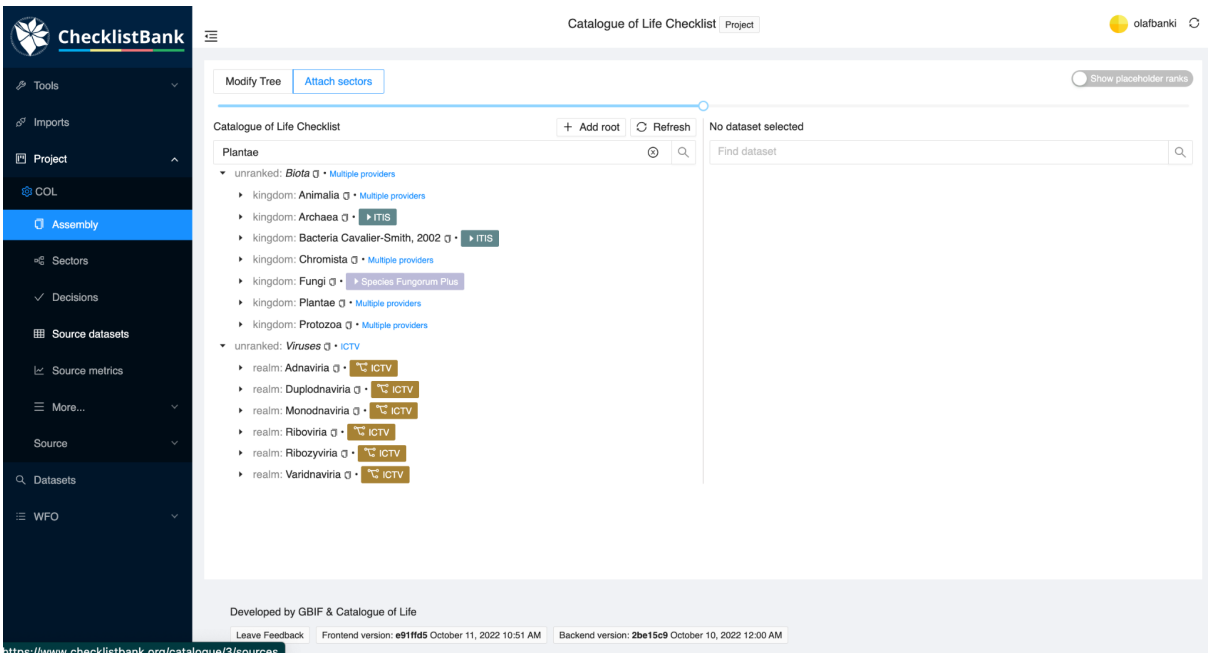

Figure 7: *Attaching sectors to the tree.*

Sectors can be attached to the tree. The point of attachment into the hierarchy of the tree is determined by the matching taxon in the source dataset selected from ChecklistBank. In the above case, the working draft of the COL Checklist (also a project dataset) is shown. The colored blocks beside a taxon (in this case kingdoms from ITIS and realms from ICTV) show the underlying data sources the checklist is built from.

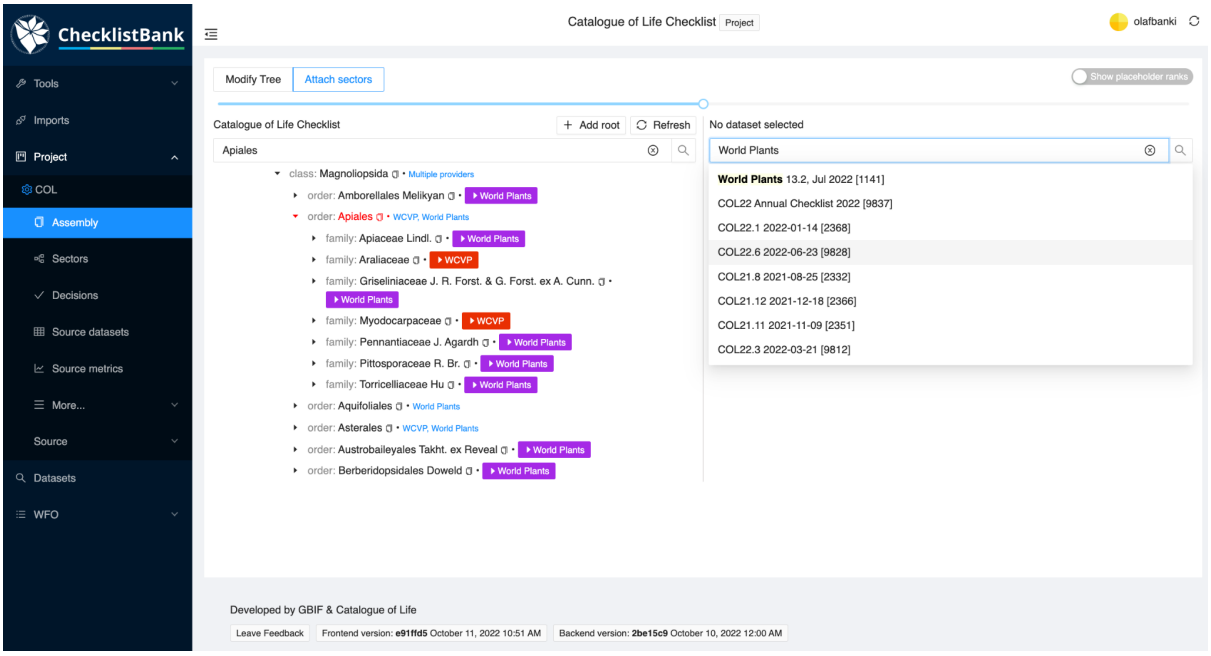

Figure 8: *Selecting datasets from ChecklistBank.*

In the right hand panel, a search interface allows source datasets to be found and selected for inclusion. Selected datasets are shown as updates to the working draft of the project dataset shown in the left panel.

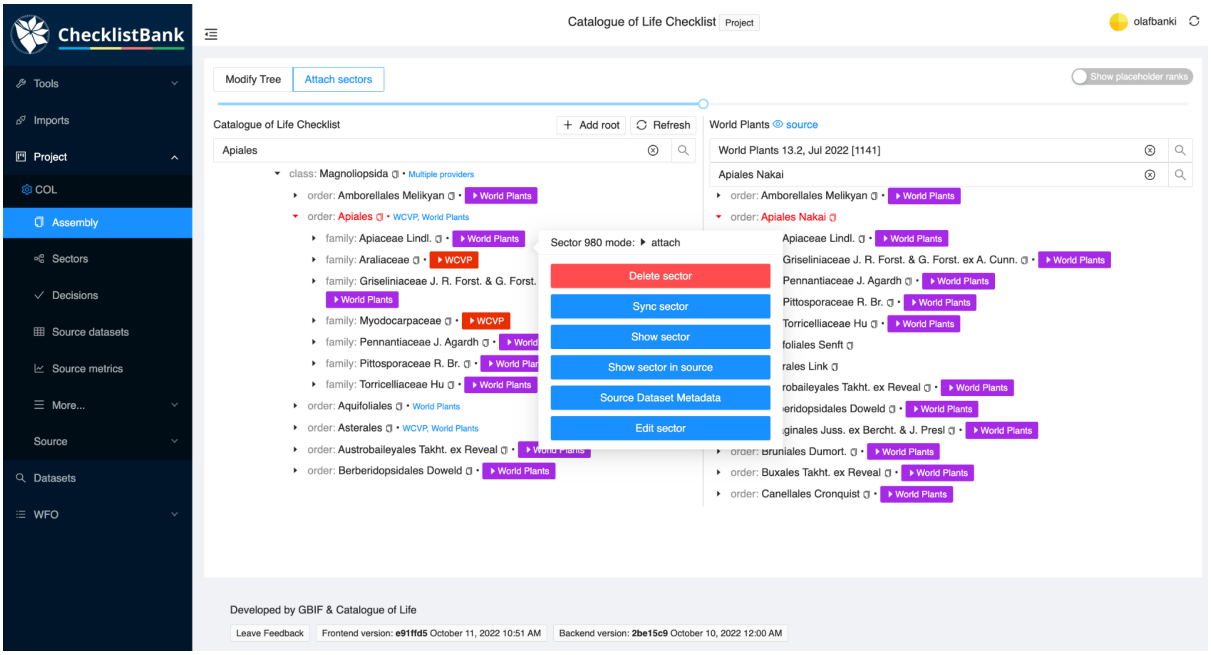

Figure 9: *Selected sectors can be synchronised or replaced.*

Selected sectors in the tree of the project dataset can be synchronised with a new version of the same sector when the underlying datasource has been updated. The taxonomic sector can also be deleted or various edits could be applied to the sector.

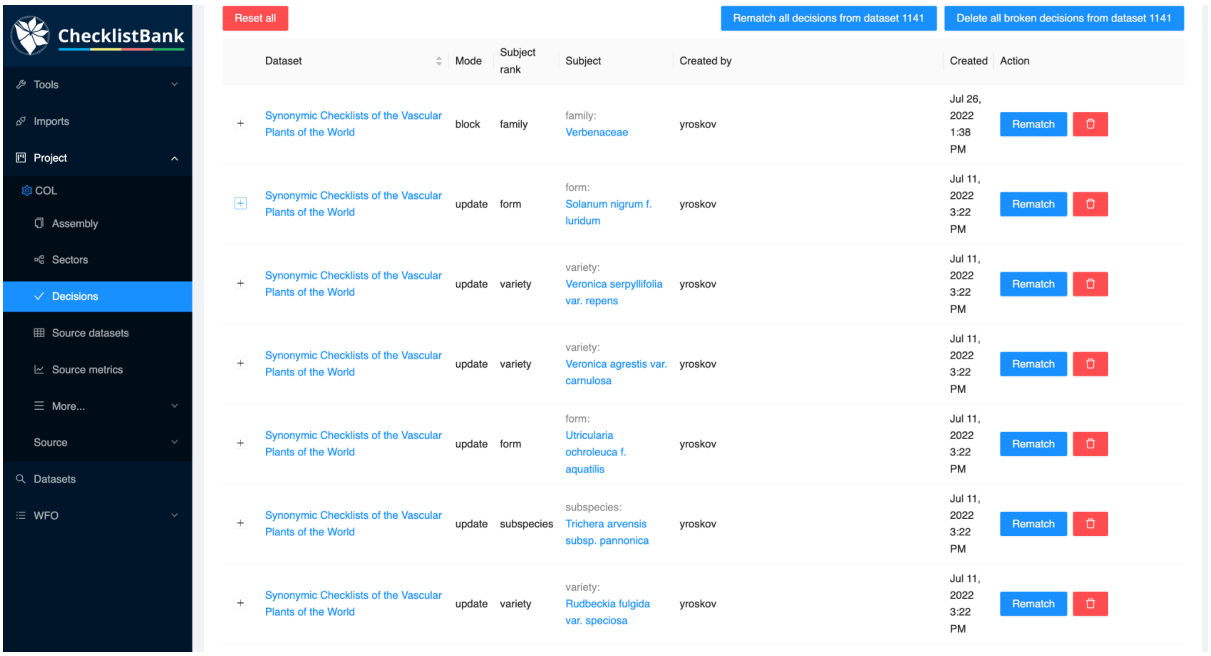

Figure 10: *Editorial decisions.*

Editorial decisions can be applied to modify taxonomic sectors that are attached to the tree of a project dataset. For example, certain taxonomic ranks, such as a specific family, could be excluded (by blocking) from the project dataset. Statuses (accepted/synonym) can also be modified when taxa are attached from the sector. These editorial decisions are persistent and can be reapplied every time that a sector is synched with the working draft of the project

dataset. The editorial decisions are just applied on top of the dataset and no changes are made to the source dataset in ChecklistBank.

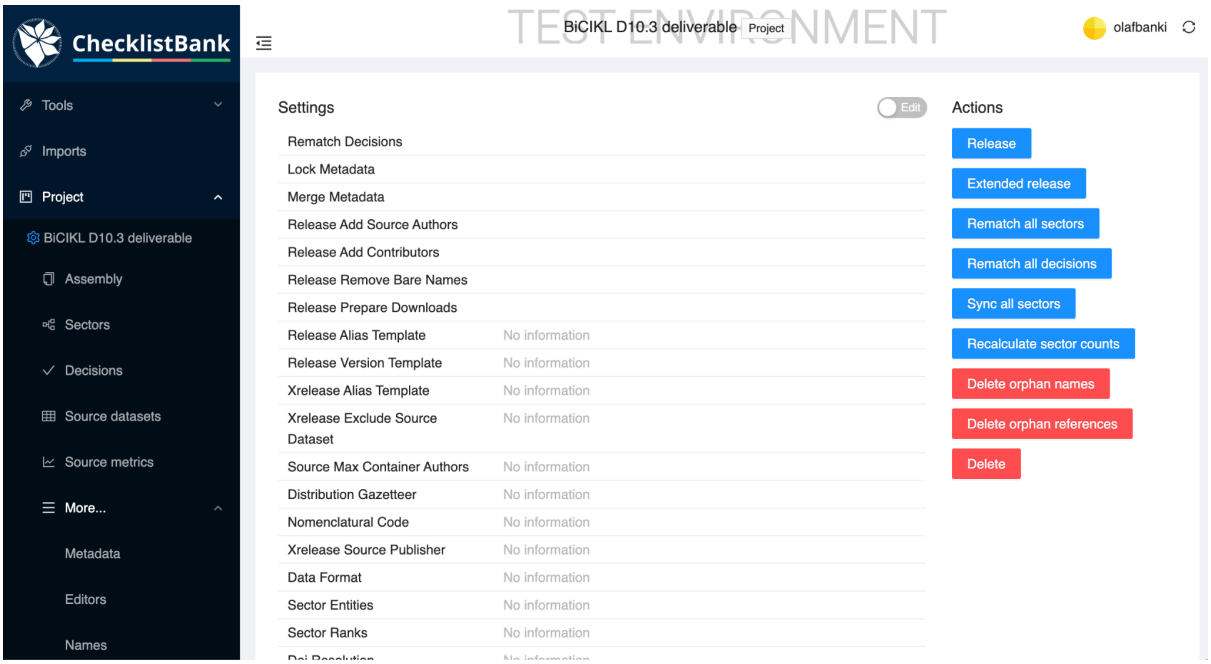

Figure 11: *Additional Project options.*

There are several additional project options available for a project dataset in ChecklistBank. If the working draft of a project is ready for release, the user can trigger this release. Releasing a dataset means that it is available in ChecklistBank for use and that versioning is established for the dataset. Other project options include, amongst others, rematching of all sectors or decisions, syncing of sectors, or various options for deleting components.

2.2. Project datasets and avenues for curation

2.2.1. Project datasets

In addition to current options for external datasets (those maintained in an external editing platform), users will have the option to create project datasets; constructed by using the 'ChecklistBank project functionality'.

Project datasets have the following benefits for users:

- A project can be initialised via a data import into ChecklistBank in the same way as for an external dataset.
- Users may treat the ChecklistBank copy of any project dataset as the primary version for editing.
- For users of ChecklistBank, a project dataset will support all uses and behaviours offered for an external dataset and potentially all functionality associated with a project in ChecklistBank.

- The contents of any public project will be accessible as a ChecklistBank dataset with all associated behaviour and functionality (metadata, API access, download functions, version history, availability as a source for other projects, etc.).

2.2.2. Avenues for curation of project datasets

The 'ChecklistBank project functionality' is not intended to support the *de novo* creation of a new taxonomic dataset or to offer all the content features and functionality that platforms such as WoRMS, ITIS and TaxonWorks provide for managing nomenclatural and taxonomic data (including literature references, type specimens, distribution data, etc.). At present there are no ways to alter data sources directly in ChecklistBank. Offering some annotation functions may be explored as part of Deliverable D10.5 'CoL services for direct encapsulated use within search and discovery services of BiCIKL partner infrastructures'.

Extensive taxonomic editing and curation of a project datasets needs to be done outside ChecklistBank. Datasets in ChecklistBank may be downloaded in various formats (e.g. DwCA, ColDP, texttree, newick, dot; Figure 12). The user may edit the download externally using any appropriate tools (which may be developed to work with the download formats). The COL consortium partners offer various taxonomic editing tools, such as Aphia, TaxonWorks, Rhakhis, the ITIS taxonomic workbench or the EDIT platform.

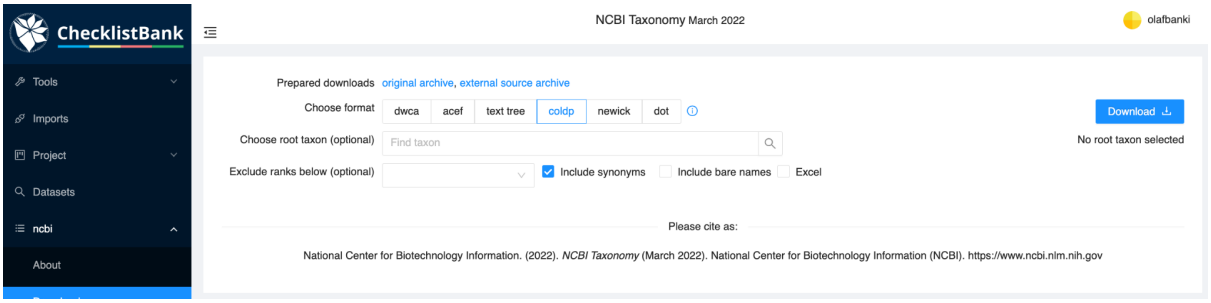

Figure 12: *Download options of ChecklistBank datasets.*

Datasets in ChecklistBank can be downloaded through various formats (e.g. DwCA, ColDP, texttree, newick, dot, excel). The original archive or external source archives can be downloaded as zip files. Instead of downloading the entire dataset, there are also options to start a download from a particular root taxon and/or exclude certain ranks from or include synonyms in a downloaded dataset.

ChecklistBank also offers a metadata generator to its users. Through this generator, metadata in YAML or EML formats can be generated (Figure 13). The generated metadata file can be uploaded to a project dataset, and serve as the prime source of metadata for this project dataset.

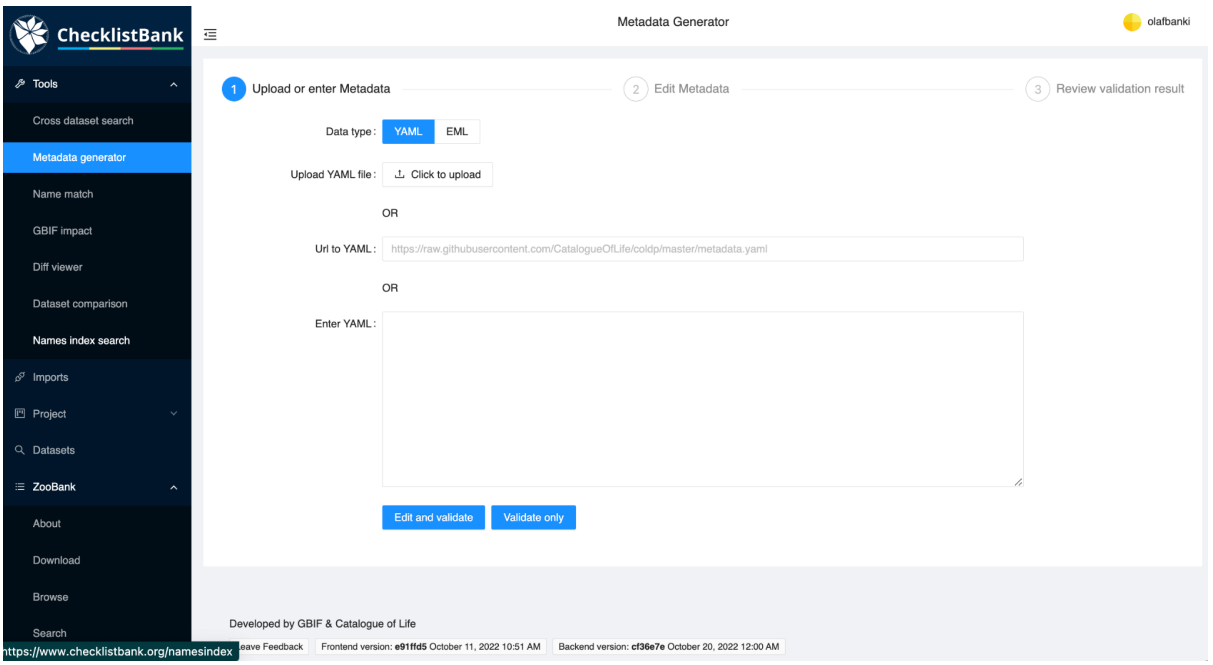

Figure 13: *Metadata generator.*

As part of its suite of tools that are offered to users, ChecklistBank also offers a metadata generator. This generator enables a user to create a metadata file in YAML or EML.

3. Impact and way forward

3.1. Impact of deliverable D10.3

3.1.1. Impact of the functionality delivered as part of deliverable D10.3

The 'ChecklistBank project functionality' enables the following use cases:

1. Construction and versioning of the COL Checklist by the COL editors, including the collaborative work by COL and GBIF to expand the COL Checklist to provide tentative unreviewed placement for names and molecular OTU (MOTU) identifiers that are not present in the primary source datasets included in the COL Checklist (see below under 3.2.1).
2. Construction, validation and versioning of checklist datasets combining multiple sources for use in the COL Checklist. For example, information in COL for the Lepidoptera (butterflies and moths) is derived from multiple datasets with different scope. The ChecklistBank workbench tools will allow taxonomic experts to manage the combination and validation of these components as a single expert-reviewed Lepidoptera resource which can then be included within the COL Checklist.
3. Construction, validation and versioning of national, regional or thematic checklists. There has been a long-standing requirement for national or regional authorities and intergovernmental bodies to have good tools to manage species lists for regulatory

and other purposes. By using the ChecklistBank workbench, these authorities will benefit from access to all source datasets published in ChecklistBank, the ability to upload and manage other component datasets, the ability to select sectors of the COL Checklist itself as component datasets, and the suite of tools for rules-based construction and validation of a constructed checklist.

3.1.1. Relation to other deliverables in BiCIKL

Within BiCIKL work package 10, the deliverable D10.3 plays an important role in enabling the assembly of taxonomic checklists that could be geared towards the needs and requirements of BiCIKL partner infrastructures. D10.3 provides functionality that could be built upon in especially the deliverables D10.1 New taxonomic data products automatically appear in COL and D10.5 CoL services for direct encapsulated use within search and discovery services of BiCIKL partner infrastructures. D10.3 is also likely to strengthen deliverables D6.2 and D8.3 in providing alternative avenues for the assembly of project checklist datasets.

3.2. Wider biodiversity data landscape

3.2.1. The development of a semi-automated part of the COL Checklist

GBIF and COL are developing a semi-automated part of the COL Checklist. The addition of a semi-automated part will make the COL Checklist more comprehensive by including extended information and enriched data for example coming from Plazi mediated species information and OTU's coming from ENA and UNITE. It will improve taxonomic coverage and usefulness of the COL Checklist also in delivering taxonomic services for GBIF-mediated occurrences (Figure 14). The semi-automated part of the assembly of a checklist may in the future also become available as a generic function in the 'ChecklistBank project functionality'.

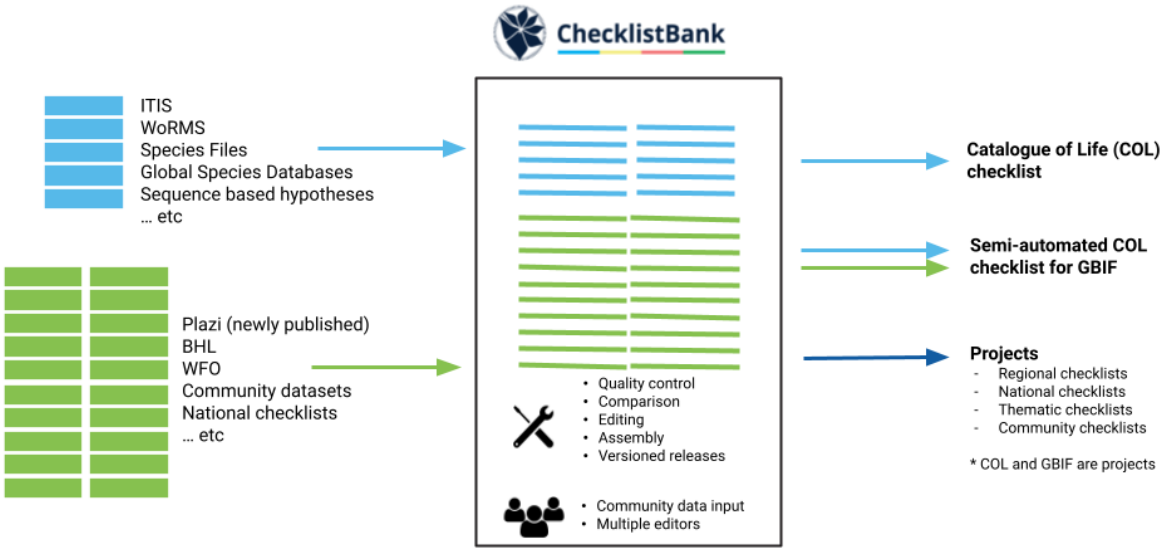

Figure 14: *Schematic figure of ChecklistBank datasets and the products coming out of it.*

At present datasets underpinning the Catalogue of Life Checklist (blue) and the GBIF Backbone Taxonomy (green) come into ChecklistBank. The COL Checklist is constructed from datasets in ChecklistBank. COL and GBIF now construct a semi-automated part of the COL Checklist that would serve as a candidate for the replacement of the GBIF Backbone Taxonomy. Data products, checklists of different scope, can also be generated through the 'ChecklistBank project functionality'.

3.2.2. Alignment of biodiversity data infrastructures

Currently ChecklistBank contains taxonomic checklists used by and underpinning the COL Checklist, taxonomic checklists that are part of the current GBIF Backbone Taxonomy, and more than 43K datasets mediated through Plazi. 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 a 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 now automatic harvesting and ingestion of taxonomic data into ChecklistBank from the literature (e.g. Plazi, Pensoft and European Journal of Taxonomy) is partially in place (with a contribution to the BiCIKL project deliverables D10.1, D10.2, as well as D6.3), any potentially new and already published taxonomic names can be automatically compared to checklists, such as the COL Checklist and the GBIF Backbone Taxonomy, to determine whether they need to be added to these checklists. This will greatly decrease the time lag from taxonomic publication to use by community researchers and taxonomic infrastructures. Given that ChecklistBank is making 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 'project functionality' together with the workbench tooling to assemble a taxonomic checklist, and project datasets in ChecklistBank will be made available to the BiCIKL user community and partners. Considering these tools will allow the creation of datasets in ChecklistBank, editing rights will only be assigned to specific users upon request. It is likely that COL and GBIF will run in 2023 a series of pilots to test the workbench tools with a variety of different requirements for the assembly of species checklists (e.g. taxonomic lists, national lists, and thematic or policy relevant species checklists). Catalogue of Life and the Global Biodiversity Information Facility will promote ChecklistBank and its associated tools to their respective user communities. It is expected that training materials will be developed for the workbench tooling. COL and the results in BiCIKL will also play an important role in the EU Commission project 'Transforming European Taxonomy through Training, Research, and Innovations' (TETTRIS) for the mapping and linking of local taxon lists with European and international checklists.

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., Alexander, S., et al. (2022). Catalogue of Life Checklist (Version 2022-10-20). Catalogue of Life. https://doi.org/10.48580/dfqf

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

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

Plazi, https://plazi.org/

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