








**PREPRINT**

*Author-formatted, not peer-reviewed document posted on 30/03/2026*

DOI: <https://doi.org/10.3897/arphapreprints.e193155>

# **A Browser-Based Curation Tool for Expert Review of DNA Barcode Records from BOLD Systems**

**Stephan Kühbandner, Fabian Deister,  Axel Hausmann,  Michael Raupach,  Ben Price,  Torbjørn Ekrem, Elisabeth Stur, Brent Emerson, Peter Hollingsworth,  Rutger Vos, Leonardo Dapporto,  Adele Bordoni,  Claudia Bruschini,  Sónia Ferreira**

# A Browser-Based Curation Tool for Expert Review of DNA Barcode Records from BOLD Systems

Stephan Kühbandner<sup>‡</sup>, Fabian Deister<sup>‡</sup>, Axel Hausmann<sup>‡</sup>, Michael J. Raupach<sup>‡</sup>, Ben Price<sup>§</sup>, Torbjørn Ekrem<sup>||</sup>, Elisabeth Sturl<sup>||</sup>, Brent Emerson<sup>||</sup>, Peter M. Hollingsworth<sup>#</sup>, Rutger Aldo Vos<sup>□</sup>, Leonardo Dapporto<sup>«</sup>, Adele Bordini<sup>«</sup>, Claudia Bruschini<sup>«</sup>, Sónia Ferreira<sup>»</sup>,<sup>^</sup>

<sup>‡</sup> Bavarian State Collection of Zoology, Munich, Germany

<sup>§</sup> Natural History Museum, London, United Kingdom

<sup>||</sup> Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

<sup>||</sup> Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), La Laguna, Tenerife, Islas Canarias, Spain

<sup>#</sup> Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

<sup>□</sup> Naturalis Biodiversity Center, Leiden, Netherlands

<sup>«</sup> Dipartimento di Biologia, Università degli Studi di Firenze, Florence, Italy

<sup>»</sup> CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661 Vairão, Vila do Conde, Portugal

<sup>^</sup> BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, 4485-661 Vairão, Vila do Conde, Portugal

Corresponding author: Stephan Kühbandner ([kuehbandner@snsb.de](mailto:kuehbandner@snsb.de)), Torbjørn Ekrem ([torbjorn.ekrem@ntnu.no](mailto:torbjorn.ekrem@ntnu.no))

Reviewable

v 1

## Abstract

## Background

We present a browser-based curation tool developed to support expert validation of taxonomic records derived from the Barcode of Life Data System (BOLD). This tool forms a critical component of a two-step approach designed within the EU Horizon Europe project Biodiversity Genomics Europe (BGE) to build a high-quality, curated DNA barcode reference library for European species. The upstream component—a bioinformatics pipeline described in a companion publication—automatically filters, cleans, and ranks BOLD records based on metadata completeness, sequence quality, and taxonomic consistency. However, certain complex cases, such as misidentifications, nomenclatorial problems (e.g. synonymy), BIN-sharing (multiple species sharing one BIN) or BIN-splitting (a single species associated with multiple BINs), cannot be fully resolved by automated methods and require expert judgment.

## New information

Our curation tool enables taxonomic experts to interactively inspect, validate, or exclude individual records, update species names, assign curation statuses, and provide curator

notes. The tool supports real-time statistics for BIN conflicts and dynamically updates curation metrics as the expert interacts with the data. Its user interface is designed to simplify the review of large datasets while ensuring consistency, traceability, and minimal risk of structural errors common in spreadsheet-based curation workflows.

The curated output from this tool, combined with the automated pipeline, forms the foundation of a reference library suitable for accurate DNA-based species identification in biodiversity monitoring and ecological studies. By integrating expert knowledge into a standardized and scalable interface, the tool supports the creation of FAIR, version-controlled reference libraries essential in the era of accelerating biodiversity loss and declining taxonomic capacity.

## Keywords

Reference library curation, BOLD, Taxonomic records, BIN, DNA barcoding

## Introduction

### Background

DNA barcoding (Hebert et al. 2003, DeSalle and Goldstein 2019, Rani et al. 2026) has become a widely used approach for species identification and biodiversity assessment across a broad range of taxa. The Barcode of Life Data System (BOLD, Ratnasingham and Hebert 2007) is the central repository for DNA barcode data, hosting over 17,8 million specimen records globally from 1.3 million Barcode Index Numbers (BINs, which are proxies for species (Ratnasingham and Hebert 2013, Ratnasingham 2024a), including over 1.5 million records for European species alone. Despite the enormous value of this resource, many of the records — particularly those from early barcoding efforts or mirrored from GenBank — often lack critical metadata, contain outdated taxonomic names, or do not meet current quality standards (Baena-Bejarano et al. 2023, Meiklejohn et al. 2019). As a result, accurate species-level identification using BOLD data often depends on extensive post-processing and expert validation.

To address these limitations, the EU Horizon Europe project Biodiversity Genomics Europe (BGE, Naturalis Biodiversity Center 2025) is developing a curated DNA barcode reference library for European species, with a primary focus on pollinators, freshwater, and marine taxa. The reference library curation is performed in two steps (Fig. 1). In the first phase, an automated bioinformatics pipeline (Vos 2024, Price 2025) filters and ranks public BOLD records (Ratnasingham 2024b) based on metadata completeness, sequence quality, taxonomic consistency, and known issues such as synonymy and typographical errors. This process results in a pre-curated dataset suitable for further review.

However, automated filtering alone is insufficient to resolve certain biologically complex or taxonomically ambiguous cases. For example, BIN-sharing events (multiple species share a single Barcode Index Number) or BIN-splitting events (a single species is

assigned to multiple BINs) require expert taxonomic knowledge to interpret and resolve (Fontes et al. 2020, Hausmann et al. 2013). In addition, validating taxon names, updating nomenclature, and reviewing metadata in light of ongoing taxonomic revisions are tasks that necessitate manual curation.

Several approaches have been proposed to support the systematic curation of BOLD records. One such approach is the Barcode Audit and Grade System (BAGS; Fontes et al. 2020), which assigns grades (A–E) to species based on the number of records per BIN and the occurrence of BIN-sharing or BIN-splitting (Table 1). Within the BGE project, this concept has been extended through the definition of *country representatives* — pre-selected records with the best metadata quality for each combination of species, OTU and country. Additionally, we have developed a metadata quality rating system (Price 2025, Vos et al. 2026) that ranks records on a scale from 1 to 6, incorporating factors such as sequence length, presence of voucher information, and completeness of collection data (Table 2). These frameworks provide the quantitative foundation for prioritizing records during manual review.

While it is technically possible to conduct manual curation using common spreadsheets, this approach becomes impractical and error-prone for large and metadata-rich datasets (Broman and Woo 2018). Spreadsheet-based workflows are prone to formatting inconsistencies, accidental overwriting of fields, unstandardized status entries, and loss of data integrity when files are transferred between curators (Broman and Woo 2018). They also lack the ability to dynamically update key indicators such as BIN-sharing, BIN-splitting, or BAGS scores in real time. In contrast, a dedicated curation tool can enforce consistent data structures, provide immediate feedback on changes, and reduce the volume of data that needs to be exchanged with experts (Broman and Woo 2018).

To support this critical second phase, we developed a dedicated, browser-based curation tool tailored to the needs of taxonomic experts.

Here we introduce the design and functionality of the curation tool, and highlight its potential role in producing high-quality barcode reference data for DNA-based species identification. In doing so, we aim to provide a scalable, transparent, and expert-driven solution for curating large and complex barcode datasets, particularly in the context of biodiversity research and monitoring initiatives.

## Project description

**Title:** BGE Library Curation Tool

### **Design description:**

The BGE Library Curation Tool allows curators to review, validate, and annotate BOLD-derived records using a structured, user-friendly interface (cf. Fig. 2). It provides real-time statistics on species coverage, BIN conflicts, and curation progress, and ensures data integrity through constrained input options and exportable, version-ready outputs.

The manual curation process using this tool generally follows these steps:

1. **Load dataset:** Select the .db file containing the data for the target taxonomic group generated by the pre-curation pipeline.
2. **Filter and search:** Narrow down the dataset by species, BIN, or other metadata fields.
3. **Review records:** Examine species names, BIN assignments, metadata quality, and potential conflicts.
4. **Assign status:** Mark each record as valid, invalid, or excluded; correct species names where needed.
5. **Add notes:** Document curation decisions with curator comments.
6. **Monitor statistics:** Use dynamic counters to track BIN-sharing/splitting events and curation completeness.
7. **Export results:** Save curated data as .csv along with a changes.log file for audit purposes.

**Funding: Biodiversity Genomics Europe (Grant no.101059492)** is funded by Horizon Europe (European Commission 2021) under the Biodiversity, Circular Economy and Environment call (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract numbers 22.00173 and 24.00054; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

## Web location (URIs)

Homepage: <https://doi.org/10.5281/zenodo.18877943>

## Technical specification

**Platform:** Browser (Edge, Chrome, Firefox, etc.)

**Programming language:** Java Script, HTML, CSS

**Operational system:** Windows, Linux, Mac OS

Interface language: English

## Repository

Type: Git

Browse URI: <https://doi.org/10.5281/zenodo.18877909>

## Usage licence

Usage licence: Creative Commons Public Domain Waiver (CC-Zero)

## Implementation

### Implements specification

The BOLD Curation Tool (cf. Fig. 3) is a browser-based application designed to assist taxonomic experts in the manual validation of DNA barcode records, particularly those derived from the Barcode of Life Data System (BOLD). It serves as the expert-driven interface in a two-phase curation workflow. The upstream component is a semi-automated bioinformatics pipeline that filters, ranks, and enriches raw BOLD data. This tool builds on that output by providing an intuitive environment for manual review and expert decision-making. The steps of manual curation carried out by taxonomic experts with this tool comprise inspection of records (view pre-selected records, filter data, view BAGS grade, BIN-sharing and -splitting events and metadata) and take action (validate or invalidate records, exclude (reinclude) species, changes species name, choose reason for name corrections and add curator notes as freetext).

All actions performed in the library curation tool are documented in log files including timestamps, the executed action, and all information provided by the experts. These log files are analysed using a dedicated script ([https://github.com/FabianDeister/BGE\\_library\\_curation\\_tool\\_log\\_processing](https://github.com/FabianDeister/BGE_library_curation_tool_log_processing)) together with the output of the automated pipeline, whereby timestamps ensure that only the most recent version of each change is retained. In this way, both automatically pre-validated and manually reviewed records are merged to form the curated reference library.

The application is implemented using standard web technologies and can be run entirely on a local computer, without internet access (aside from optional loading of remote CSS assets). It includes the following components:

- **Backend:** The backend logic is handled by a lightweight Node.js (Open JS Foundation 2024) server (server.js), which manages HTTP requests and interacts with the input dataset—a structured SQLite-compatible .db file produced by the upstream curation pipeline. This file contains metadata-rich sequence records,

including taxonomic information, BIN URIs, quality scores, and precomputed BAGS (Barcode, Assignment, Grade, Status) values.

- **Frontend:** The user interface is built in index.html using HTML, CSS, and JavaScript, and runs in a modern web browser (e.g. Chrome, Firefox). It utilizes the [DataTables](#) library to provide interactive, searchable, and paginated tables. Custom JavaScript code supports advanced functionalities such as row coloring based on status, BIN visualization, in-table dropdowns for status selection, curator note entry, and per-record submission.
- **Database Input:** The tool operates on local .db database files. These files are placed in the data/ subdirectory and loaded dynamically through a dataset selector. Each file corresponds to a taxonomic group and contains hundreds to thousands of records to be curated.
- **Execution Environment:** The application is platform-independent and distributed as a self-contained folder Fig. 4. On Windows, users simply double-click start\_tool.bat to launch the server and automatically open the tool in a browser via <http://localhost:3000>. On Linux and macOS, the tool can be launched manually from the command line using Node.js. A detailed installation manual for Linux and macOS can be found here: [https://github.com/bge-barcoding/BGE\\_library\\_curation\\_tool](https://github.com/bge-barcoding/BGE_library_curation_tool).
- **Export and Audit:** All curation actions (status changes, species name updates, notes) are logged in a changes.log file, ensuring transparency and reproducibility. Curators can export their results (for own purposes) in .csv format and submit the log file as a standardized feedback mechanism. The tool prevents structural errors common in spreadsheet-based curation by enforcing consistent fields and controlled input types.
- **Dynamic Scoring:** The tool includes dynamic logic for recalculating BAGS scores and BIN statistics in real time. This allows experts to see how their actions (e.g. excluding a species or marking a record invalid) influence BIN-sharing, BIN-splitting, and representative selection.

In summary, the BOLD Curation Tool is a locally hosted, browser-accessible interface purpose-built for scalable expert curation of DNA barcode data. It bridges the gap between automated pipeline output and final expert-reviewed reference libraries, facilitating the creation of high-quality, FAIR-compliant resources for molecular biodiversity research.

There are several levels of support for the user. First, there is a user manual within the main folder of the curation tool. Second, contextual help is provided the user interface by red question marks - clicking on them opens a menu with additional information. Third, a video tutorial and FAQ section are available on the project website (<https://bge-barcoding.github.io/manual-curation/>). An overview of the associated github repositories is presented in Table 3.

## Audience

Taxonomic experts will curate records from BOLD Systems that have been pre-curated using this pipeline: <https://github.com/bge-barcoding/bold-library-curation>.

## References

- Baena-Bejarano N, Reina C, Martínez-Revelo DE, Medina C, Tovar E, Uribe-Soto S, Neita-Moreno JC, Gonzalez M (2023) Taxonomic identification accuracy from BOLD and GenBank databases using over a thousand insect DNA barcodes from Colombia. Barcode of Life Data Systems <https://doi.org/10.5883/ds-cbih2020>
- Broman K, Woo K (2018) Data Organization in Spreadsheets. The American Statistician 72 (1): 2-10. <https://doi.org/10.1080/00031305.2017.1375989>
- DeSalle R, Goldstein P (2019) Review and Interpretation of Trends in DNA Barcoding. Frontiers in Ecology and Evolution 7 <https://doi.org/10.3389/fevo.2019.00302>
- European Commission (2021) HORIZON-CL6-2021-BIODIV-01-01. European Commission.
- Fontes J, Vieira P, Ekrem T, Soares P, Costa F (2020) BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21 (2): 573-583. <https://doi.org/10.1111/1755-0998.13262>
- Hausmann A, Godfray HCJ, Huemer P, Mutanen M, Rougerie R, van Nieukerken EJ, Ratnasingham S, Hebert PDN (2013) Genetic patterns in European geometrid moths revealed by the Barcode Index Number (BIN) system. PLoS one 8 (12): e84518. <https://doi.org/10.1371/journal.pone.0084518>
- Hebert PDN, Ratnasingham S, de Waard J (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London. Series B: Biological Sciences 270 <https://doi.org/10.1098/rsbl.2003.0025>
- Meiklejohn K, Damaso N, Robertson J (2019) Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PLOS ONE 14 (6). <https://doi.org/10.1371/journal.pone.0217084>
- Naturalis Biodiversity Center (2025) Biodiversity Genomics Europe. <https://biodiversitygenomics.eu/>. Accessed on: 2025-8-06.
- Open JS Foundation (2024) Node.js Run JavaScript Everywhere. v22.x. Open JS Foundation. Release date: 2024-4-01. URL: <https://nodejs.org/en>
- Price B, et al. (2025) bold-library-curation. Release date: 2025-4-04. URL: <https://github.com/bge-barcoding/bold-library-curation>
- Rani V, Chauhan C, Sengar RS (2026) DNA barcoding markers: A comprehensive review and taxonomic classification across species. Computational Biology and Chemistry 122 <https://doi.org/10.1016/j.compbiolchem.2025.108872>
- Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). Molecular Ecology Notes 7 (3): 355-364. <https://doi.org/10.1111/j.1471-8286.2007.01678.x>

- Ratnasingham S, Hebert PDN (2013) A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. PloS one 8 (7): e66213. <https://doi.org/10.1371/journal.pone.0066213>
- Ratnasingham S, et al. (2024a) BOLD Systems. <https://boldsystems.org/>. Accessed on: 2025-8-05.
- Ratnasingham S, et al. (2024b) BOLD Systems / Data Packages. <https://boldsystems.org/data/data-packages/>. Accessed on: 2025-8-06.
- Vos R (2024) BOLD curated data and reference library. Zenodo <https://doi.org/10.5281/zenodo.11248591>
- Vos R, Deister F, Kuehbandner S, Hausmann A, Ekrem T, Stur E, Emerson B, Raupach M, Hollingsworth P, Price B, et al. (2026) BOLDetective: an automated analysis pipeline and the first large-scale analysis of European animal barcode data highlights the need for gap-filling and expert curation. Research Ideas and Outcomes (RIO) in preparation.

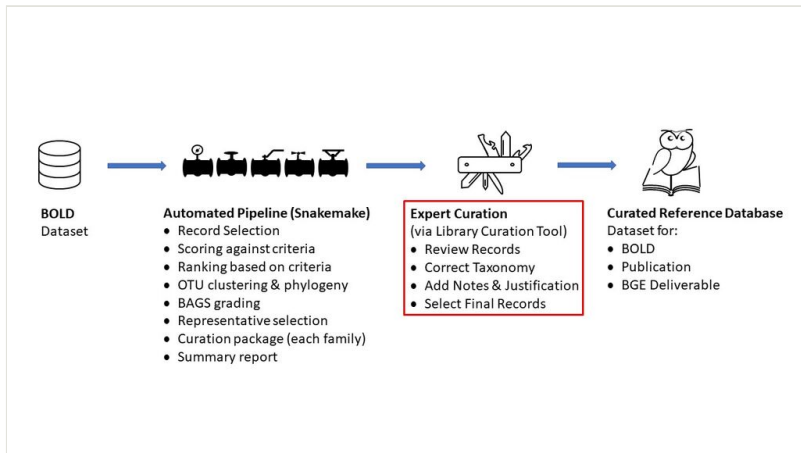


Figure 1.

The library curation workflow starts by mining data from BOLD, then processes this data with an automated curation pipeline, before following up with manual curation by taxonomic experts using the curation tool and publishing the curated reference library as a dataset on BOLD.

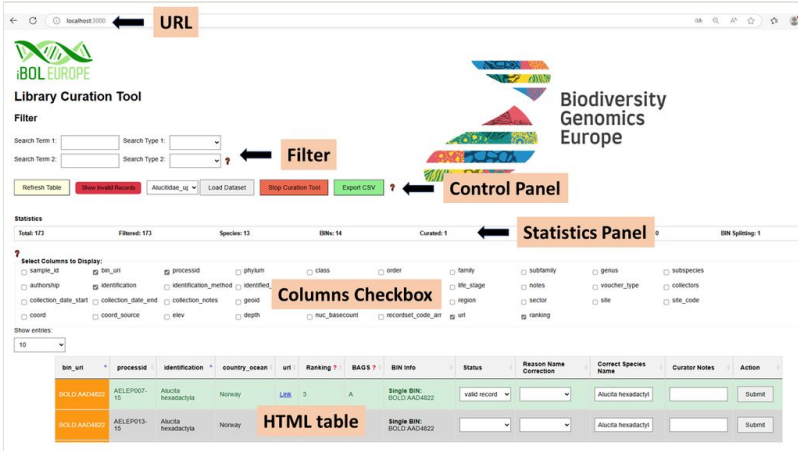


Figure 2. Main Interface of the Library Curation Tool with labels explaining the different sections.

bin_url	processid	identification	country_origin	url	Ranking	country_representative	BAGS	Bin Info	Status	Reason Name Correction	Correct Species Name	Curator Notes	Action
BOLD:ADY798	BOSNL511-23	<i>Alicata bidorsata</i>	Italy	L58	2	Yes	D	Single BIN: BOLD:ADY798			<i>Alicata bidorsata</i>		Submit
BOLD:AEI246	BOSNL512-23	<i>Alicata canariensis</i>	Spain	L58	2	Yes	C	BIN-splitting: BOLD:AEI246; BOLD:AAO5227			<i>Alicata canariensis</i>		Submit
BOLD:AEI246	BOSNL513-23	<i>Alicata canariensis</i>	Spain	L58	2	Yes	C	BIN-splitting: BOLD:AEI246; BOLD:AAO5227			<i>Alicata canariensis</i>		Submit
BOLD:AMQ527	GWOTV511-17	<i>Alicata cancellata</i>	Italy	L58	2	Yes	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527			<i>Alicata cancellata</i>		Submit
BOLD:AMQ527	GWOT5470-17	<i>Alicata cancellata</i>	Italy	L58	2	Yes	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527			<i>Alicata cancellata</i>		Submit
BOLD:AMQ527	LEAS1966-18	<i>Alicata cancellata</i>	France	L58	2	Yes	E	BIN-splitting: <i>Alicata cancellata</i> ; Single BIN: BOLD:AMQ527	invalid record	misidentified	<i>Alicata canariensis</i>		Submit
BOLD:AMQ527	PHLSA200-11	<i>Alicata cancellata</i>	Spain	L58	2	Yes	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527			<i>Alicata cancellata</i>		Submit
BOLD:AMQ527	LEAS1965-18	<i>Alicata cancellata</i>	France	L58	2	No	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527	valid record		<i>Alicata cancellata</i>		Submit
BOLD:AMQ527	GBLAZ707-15	<i>Alicata cancellata</i>	Italy	L58	2	No	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527			<i>Alicata cancellata</i>		Submit
BOLD:AMQ527	LEADY789-22	<i>Alicata cancellata</i>	France	L58	2	No	E	BIN-splitting: <i>Alicata canariensis</i> ; Single BIN: BOLD:AMQ527			<i>Alicata cancellata</i>		Submit

Showing 1 to 10 of 196 entries

Previous 1 2 3 4 5 ... 20 Next

Figure 3.

HTML table of the Library Curation Tool with taxonomic records from BOLD and additional curation specific metadata (like: url, Ranking, country\_representative, BAGS, Status, Reason Name Correction, Correct Species Name, Curator Notes). Records with grey background are pre-selected for reference library and need no action by the user for getting them added to the reference library. However, users can validate non-pre-selected records, which will get a green background or invalidate pre-selected records, which will have a red background.

```
BGE_library_curation_tool/  
├── data/           # .db files as input data  
├── logs/          # changes.log contains curated data  
├── node  
├── node_modules  
├── public/        # index.html (frontend)  
├── package.json  
├── server.js      # backend  
├── start_tool.bat # double-click starts the tool  
└── user_manual_windows.html # user manual for the windows version
```

Figure 4.

Curation Tool - main folder.

Table 1.  
BAGS - Barcode, Audit & Grade System.

GRADE	
A	>10 specimens in 1 BIN
B	3-10 specimens in 1 BIN
C	>1 BIN
D	<3 specimens in 1 BIN
E	BIN sharing (>1 species in single BIN)

Table 2.

Ranking system to pick representatives for each haplotype / species: Ranking 1-3 means records with good metadata quality (highlighted in grey), which will be pre-selected for the reference library and ranking 4-6 records with bad metadata quality, which are not pre-selected for the reference library. For "Public voucher" "✓(or)" means that only one of these criteria has to be fulfilled in order to meet the ranking for all criteria with this prefix. The same is true for "Collection".

Criteria	specimen rank					
	1	2	3	4	5	6
Species level ID	✓	✓	✓	✓	✓	✓
Type specimen	✓					
Good quality sequence		✓	✓	✓	✓	
Image(s) available		✓				
Collection country		✓	✓	✓		
ID identifier named		✓	✓			
ID method (method is not BIN match)		✓	✓			
Public voucher (has museum ID)		✓(or)	✓(or)			
Public voucher (agreed institution)		✓(or)	✓(or)			
Public voucher (agreed voucher type)		✓(or)	✓(or)			
Collection (Date)		✓				
Collection (Site)		✓(or)				
Collection (GPS coordinate)		✓(or)				
Collection (Sector)		✓(or)				
Collection (Region)		✓(or)				
Collector named		✓				

Table 3.

Associated GitHub repositories, respectively Zenodo archives

url / link	Description
<a href="https://github.com/bge-barcoding/BGE_library_curation_tool">https://github.com/bge-barcoding/BGE_library_curation_tool</a> <a href="https://doi.org/10.5281/zenodo.18877958">https://doi.org/10.5281/zenodo.18877958</a>	Curation Tool - Linux and macOS version
<a href="https://github.com/bge-barcoding/BGE_library_curation_tool_win">https://github.com/bge-barcoding/BGE_library_curation_tool_win</a> <a href="https://doi.org/10.5281/zenodo.18877909">https://doi.org/10.5281/zenodo.18877909</a>	Curation Tool - Windows version
<a href="https://github.com/bge-barcoding/bold-library-curation">https://github.com/bge-barcoding/bold-library-curation</a> <a href="https://zenodo.org/records/17495973">https://zenodo.org/records/17495973</a>	BOLD Library Curation Pipeline
<a href="https://bge-barcoding.github.io/manual-curation/">https://bge-barcoding.github.io/manual-curation/</a> <a href="https://doi.org/10.5281/zenodo.18877909">https://doi.org/10.5281/zenodo.18877909</a>	iBOL Europe BOLD Curation Datasets
<a href="https://github.com/FabianDeister/BGE_library_curation_tool_log_processing">https://github.com/FabianDeister/BGE_library_curation_tool_log_processing</a>	Curation Tool - Log Processing