

Dec 11, 2024

Version 1

Workflow for Cleaning, Standardization, and Publication Biodiversity Data V.1

DOI

<https://dx.doi.org/10.17504/protocols.io.rm7vzkjyxvx1/v1>

Marcos Paulo Alves de Sousa¹, Nelson Nathan Lopes Maues Pinheiro¹, Ali Hassan Khalid¹

¹Museu Paraense Emílio Goeldi



Marcos Paulo Alves de Sousa

Museu Paraense Emílio Goeldi

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.rm7vzkjyxvx1/v1>

Protocol Citation: Marcos Paulo Alves de Sousa, Nelson Nathan Lopes Maues Pinheiro, Ali Hassan Khalid 2024. Workflow for Cleaning, Standardization, and Publication Biodiversity Data. **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.rm7vzkjyxvx1/v1>

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: December 11, 2024

Last Modified: December 11, 2024

Protocol Integer ID: 114726

Keywords: publication biodiversity data this workflow, processing biodiversity data, publication biodiversity data, quality biodiversity data, extracting biodiversity record, quality biodiversity data for research, biodiversity data, taxonomic cleaning, biodiversity records from excel spreadsheet, integrated publishing toolkit, data standardization, standardized dataset, gbif backbone taxonomy, taxonomic name, publication on gbif, reliable databases like catalogue, data preparation, geographical cleaning, darwin core, enabling robust data management, using reliable database, based tool, gbif, robust data management, standardization, global scale, interoperability across platform, conservation, coordinates against darwin core, publication, catalogue

Abstract

This workflow provides a structured approach for processing biodiversity data to ensure quality, consistency, and global accessibility. It begins with data preparation, extracting biodiversity records from Excel spreadsheets or CSV files, identifying and resolving missing, duplicate, or improperly formatted entries. Taxonomic cleaning utilizes Pytaxon, a Python-based tool, to verify and correct taxonomic names using reliable databases like Catalogue of Life and GBIF Backbone Taxonomy. Geographical cleaning follows, validating coordinates against Darwin Core (DwC) standards and correcting errors such as swapped latitudes and longitudes. Data standardization organizes fields to meet DwC requirements, ensuring interoperability across platforms. Integration into Specify involves importing validated and standardized datasets into its database, enabling robust data management. Finally, the workflow concludes with publication on GBIF via the Integrated Publishing Toolkit (IPT), making the data globally accessible. This semi-automated approach ensures high-quality biodiversity data for research, conservation, and informed decision-making on a global scale.

1. Data Preparation

- 1 The data preparation stage is essential to ensure the quality of information used in subsequent analyses. During this process, biodiversity data are extracted from Excel spreadsheets or CSV files provided by researchers. Initially, the overall structure of the data is reviewed to identify inconsistencies, such as missing values, duplicates, or improper formats. Key fields, such as taxonomic names, geographic coordinates, collection dates, and country codes (ISO-3166), are prioritized for review and organization.
- 2 Duplicate records are identified and removed using criteria based on identical values in key variables. Additionally, formatting errors, such as dates in multiple formats or ambiguously named columns, are corrected. This initial standardization ensures that the data comply with widely recognized standards, such as the Darwin Core (DwC), facilitating their use in specific analyses and systems.
- 3 Data preparation also includes identifying missing fields that may be critical for subsequent steps, enabling them to be added or flagged for future correction. The outcome is a well-structured dataset, ready to proceed through validation and cleaning stages. This step establishes a solid foundation for the entire workflow, minimizing errors and optimizing the performance of subsequent analyses.

2. Taxonomic Cleaning

- 4 Taxonomic cleaning will be conducted using pytaxon, a Python-based tool designed for identifying and correcting errors in taxonomic data. This process begins by reading Excel spreadsheets or CSV files containing biodiversity records. The pytaxon utilizes the Global Names Resolver (GNR) API to verify taxonomic names against reliable databases such as the Catalogue of Life, NCBI, and GBIF Backbone Taxonomy. The tool employs fuzzy matching techniques to identify and suggest corrections for misspellings and nomenclatural inconsistencies.
- 5 After the initial analysis, Pytaxon generates a correction spreadsheet containing the identified errors, the corresponding taxon type (e.g., genus, species), and suggested corrections. This spreadsheet allows researchers to review and accept or reject the proposed changes. Upon completion, the software updates the original spreadsheet with corrected names, ensuring a consistent and standardized dataset.
- 6 The implementation of Pytaxon not only facilitates the identification of errors in large datasets but also prevents the propagation of incorrect information in global biodiversity portals. This approach enhances the reliability of ecological studies and conservation decisions, ensuring high-quality taxonomic data.

3. Geographical Cleaning

- 7 Geographical cleaning is a critical step to ensure the accuracy and reliability of occurrence data. Initially, geographic coordinates (latitude and longitude) are verified for compliance with the standards established by the Darwin Core (DwC). These standards require values to be expressed in decimal degrees and follow the international coordinate representation format, with latitude ranging from -90 to 90 and longitude from -180 to 180.
- 8 Records with missing coordinates, out-of-range values, or locations in oceans are identified and flagged for review. Swapped coordinates, such as inversions between latitude and longitude, are detected and corrected. For records with coordinates near political or environmental boundaries, official maps like GADM are used to validate the location.
- 9 Additionally, the correspondence between coordinates and the fields country and countryCode is checked to ensure occurrences align with the declared geographic boundaries. Coordinates pointing to administrative areas, such as urban centers or institutional headquarters, are highlighted as "indicative" and not representative of the collection site.
- 10 Records that cannot be corrected are marked as low quality or removed, ensuring the final dataset is geographically accurate and suitable for scientific analysis and publication in biodiversity data systems.

4. Data Standardization

- 11 Data standardization and structuring are fundamental steps to ensure interoperability and efficient use of information in biodiversity systems. After taxonomic and geographical cleaning, the data are organized according to the standards established by the Darwin Core (DwC). These standards provide a widely accepted framework for describing species occurrence data, ensuring consistency and integration into global systems.
- 12 Initially, fields are renamed and reorganized to meet DwC requirements, such as scientificName, decimalLatitude, decimalLongitude, and eventDate. Dates are formatted in ISO 8601, ensuring uniformity and readability. Missing data in essential fields are flagged for completion, while inconsistent or redundant information is highlighted or removed.
- 13 Additionally, units of measurement and geographic coordinates are standardized to ensure consistency. Column names and categorical values are adjusted to eliminate ambiguity and facilitate future integration. Categorical data, such as basisOfRecord and occurrenceStatus, are mapped to values accepted in the DwC schema.

- 14 The result of this step is a structured dataset ready for integration into platforms like Specify or analysis systems. Standardization ensures data quality, interoperability, and reusability across various scientific and conservation contexts.

5. Integration into Specify

- 15 Data integration into Specify involves careful preparation and the use of its tools to efficiently import Excel spreadsheets. First, spreadsheets must be in Excel format (.xlsx or .xls), with clearly defined headers aligned with the fields expected by Specify, such as scientificName, decimalLatitude, decimalLongitude, eventDate, and catalogNumber. Following Darwin Core (DwC) guidelines for field naming and standardization is highly recommended.
- 16 After preparation, data are imported directly into the appropriate collection within Specify. To begin, navigate to the "Import Data" option in the collection menu and upload the Excel file. Specify provides an interactive interface to map spreadsheet columns to database fields, a critical step to ensure all data are correctly associated with their corresponding fields.
- 17 Once the fields are mapped, Specify performs an automatic data validation process, highlighting potential errors or inconsistencies, such as duplicates or missing values. These issues must be resolved before completing the import. After validation, click "Import" to transfer the data into the system.
- 18 Integration into Specify not only organizes biodiversity data but also prepares the dataset for advanced analyses and publication on platforms like GBIF, promoting global interoperability and accessibility.

6. Publication on GBIF

- 19 Publication on GBIF is conducted through the IPT (Integrated Publishing Toolkit), following a detailed workflow for preparing and validating data exported from Specify. First, the data must be exported in Excel format, ensuring the inclusion of essential fields such as scientificName, decimalLatitude, decimalLongitude, and eventDate. After export, data consistency is checked by aligning column headers with Darwin Core (DwC) standards and validating the accuracy of the information.
- 20 The data are uploaded to the IPT platform, where they are mapped to DwC terms. During this step, errors or warnings are reviewed and corrected before proceeding. The IPT also requires detailed metadata, including resource title, description, authors, and licensing information, which must be completed with care.
- 21 After reviewing and validating the data in IPT, publication is finalized by clicking "Publish." This makes the data globally accessible via the GBIF network. Regular reviews of the publication are recommended to ensure the data remain accurate and



relevant, contributing to the integration and accessibility of biodiversity information on a global scale. This process fosters the sharing of standardized data and supports international scientific collaboration.